

A REMARK ON CONTROLLABILITY FOR SYMMETRIC HYPERBOLIC SYSTEMS IN ONE SPACE DIMENSION*

N. WECK†

Abstract. In a recent survey paper [SIAM Rev., 20 (1978), pp. 639–739] D. L. Russell reports (among other things) on controllability results for symmetric hyperbolic systems in one space dimension. Russell explicitly gives two numbers T_0 and T_1 such that the system in question is exactly controllable in times $T \geq T_1$ but not even approximately controllable in times $T < T_0$. It is remarked that there must exist some T_c such that the same results hold if both T_0 and T_1 are replaced by T_c . In the present paper we want to show what this “critical time” T_c is.

1. Notation and formulation of the problem. Consider the symmetric hyperbolic system

$$(1) \quad \partial_t w(t, x) = A(x) \partial_x w(t, x) + B(x) w(t, x),$$

where

$$(t, x) \in R := [0, T] \times [0, 1],$$

$$w : R \rightarrow \mathbb{R}^N,$$

A, B are $N \times N$ -matrices.

Under mild assumptions (1) can be normalized such that

$$(2) \quad \begin{aligned} A(x) &= \text{diag}(\lambda_1^-, \dots, \lambda_K^-, \lambda_1^+, \dots, \lambda_L^+), \\ L + K &= N, \\ \lambda_K^-(x) &\leq \dots \leq \lambda_1^-(x) < 0 < \lambda_1^+(x) \leq \dots \leq \lambda_L^+(x), \end{aligned}$$

which we shall assume from now on. Accordingly we put

$$A^\pm := \text{diag}(\lambda_1^\pm, \dots, \lambda_{K \text{ resp. } L}^\pm)$$

and decompose

$$w = [w_1, \dots, w_N]^t \in \mathbb{R}^N$$

into

$$w^- := [w_1, \dots, w_K]^t \in \mathbb{R}^K,$$

$$w^+ := [w_{K+1}, \dots, w_{K+L}]^t \in \mathbb{R}^L,$$

thus identifying $w \in \mathbb{R}^N$ with $(w^-, w^+) \in \mathbb{R}^K \times \mathbb{R}^L$. The boundary conditions are governed by two matrices D_0 (of type $K \times L$) and D_1 (of type $L \times K$). Now we can formulate an initial boundary value problem

IBVP: Given $w_0 \in L^2(0, 1)$, $u \in L^2(0, T)$ find w solving (1) and satisfying

$$(3) \quad w(0, \cdot) = w_0,$$

$$(4) \quad \left. \begin{aligned} w^-(t, 0) &= D_0 w^+(t, 0) \\ w^+(t, 1) &= D_1 w^-(t, 1) + u(t) \end{aligned} \right\} \quad t \in [0, T].$$

* Received by the editors October 20, 1980, and in final form February 25, 1981.

† Universität Essen-Gesamthochschule, Postfach 6843, 4300 Essen 1, W. Germany.

Remark. We write $w_0 \in L^2(0, 1)$ instead of $w_0 \in L^2((0, 1), \mathbb{R}^N)$, etc., since it will be clear from the context what the appropriate number of components is.

The corresponding adjoint problem is a terminal boundary value problem of the same type

TBVP: Given $v_T \in L^2(0, 1)$ find v satisfying

$$(1') \quad \partial_t v(t, x) = A(x) \partial_x v(t, x) + \tilde{B}(x) v(t, x),$$

$$(3') \quad v(T, \cdot) = v_T,$$

$$(4') \quad v^+(t, 0) = C_0 v^-(t, 0) \left\{ \right. \\ (5') \quad v^-(t, 1) = C_1 v^+(t, 1) \left. \right\} \quad t \in [0, T],$$

where

$$\tilde{B} := -B' + \frac{d}{dx} A,$$

$$C_0 := -(A^+(0))^{-1} D_0' A^-(0),$$

$$C_1 := -(A^-(0))^{-1} D_1' A^+(1).$$

Existence and uniqueness of solutions to IBVP and TBVP in a suitably generalized sense are guaranteed by semigroup arguments (see [2] for details). The function $u \in L^2(0, T)$ in the IBVP is thought of as a boundary control. This explains

DEFINITION 1. The IBVP is said to be *exactly controllable* if given w_0 there exists u such that $w(T, \cdot) = 0$. The IBVP is said to be *approximately controllable* if given w_0 and $\varepsilon \in \mathbb{R}^+$ there exists u such that

$$\|w(T, \cdot)\|_{L^2(0, 1)} < \varepsilon.$$

DEFINITION 1'. The TBVP is said to be *observable* if there exists $\Gamma \in \mathbb{R}^+$ such that for all v^T we have

$$(6) \quad \Gamma \|v(0, \cdot)\|_{L^2(0, 1)} \leq \|v^+(\cdot, 1)\|_{L^2(0, T)}.$$

The TBVP is said to be *distinguishable* if, for all v_T ,

$$v^+|_{[0, T] \times \{1\}} = 0 \quad \Rightarrow \quad v|_{\{0\} \times [0, 1]} = 0.$$

Using the duality between controllability and observability [1] and considering the general case as a perturbation of the “diagonal case” (where \tilde{B} is a diagonal matrix) Russell obtains the following results:

- (i) Distinguishability \Leftrightarrow approximate controllability.
- (ii) Observability \Leftrightarrow exact controllability.
- (iii) Nondistinguishability in the diagonal case implies nonobservability in the general case.
- (iv) Observability in the diagonal case implies observability in the general case (provided the system generates a *group* and the perturbation is “small”).

These motivate the study of the observability problem for the simple case where \tilde{B} is diagonal.

2. Characteristics. Characteristic curves $c_k^-(t, x_0; \cdot)$ ($k = 1, \dots, K$) through $(t_0, x_0) \in \mathbb{R} \times [0, 1]$ are defined as the unique solutions (in $\mathbb{R} \times [0, 1]$) to the initial value problems

$$x'(t) = -\lambda_k^-(x(t)), \quad x(t_0) = x_0.$$

We define $c_l^+(t_0, x_0; \cdot)$ analogously. Let T_k^- , resp. T_l^- , be the unique positive numbers such that

$$c_k^-(0, 0; T_k^-) = 1,$$

resp.

$$c_l^+(0, 0; -T_l^+) = 1.$$

(T_k^- , resp. T_l^+ , is the time it takes the corresponding characteristic to cross the interval $[0, 1]$.) By (2) we see that

$$(7) \quad T_1^+ \geq \dots \geq T_L^+ > 0, \quad T_1^- \geq \dots \geq T_K^- > 0.$$

For fixed $T \in \mathbb{R}^+$, $k \in \{1, \dots, K\}$ and any $(t_0, x_0) \in R = [0, T] \times [0, 1]$ there exist unique $(t_I, x_I) \in \partial R$, $(t_F, x_F) \in \partial R$ such that

$$t_I \leq t_0 \leq t_F \quad \text{or} \quad t_I < t_0 \leq t_F,$$

$$c_k^-(t_0, x_0; t_I) = x_I,$$

$$c_k^-(t_0, x_0; t_F) = x_F.$$

It is convenient to introduce the maps

$$I_k^- : R \rightarrow \partial R$$

$$(t_0, x_0) \mapsto (t_I, x_I) \quad (\text{“initial point”}),$$

$$F_k^- : R \rightarrow \partial R$$

$$(t_0, x_0) \mapsto (t_F, x_F) \quad (\text{“final point”}).$$

I_l^+ and F_l^+ are defined analogously. From standard theorems on ordinary differential equations it is clear that these maps are continuous and even differentiable with bounded derivatives if differentiability is interpreted properly where boundary points or even corner points of R are involved. If \tilde{B} is a diagonal matrix and v is a solution of the TBVP, the functions $t \mapsto v_k^-(c_k^-(t_0, x_0; t))$ and $t \mapsto v_1^+(c_1^+(\cdot, \cdot; \cdot))$ satisfy homogeneous linear ordinary differential equations. It is pointed out in [2, p. 657] why for the following considerations we may as well assume that v_k^- and v_1^+ are constant along the corresponding characteristics (the case $\tilde{B} = 0$). So from now on we shall assume $\tilde{B} = 0$ (but formulate our results for diagonal matrices \tilde{B}).

3. The critical time. To determine T_c let us try to find solutions v of the TBVP such that $v|_{[0, T] \times \{1\}} = 0$. This can be achieved in two ways.

(a) “Activating” some characteristic $c_k^-(T, x_T; \cdot)$ or $c_l^+(T, x_T; \cdot)$ such that $I_k^-(T, x_T)$ or $I_l^+(T, x_T) \in \{0\} \times (0, 1)$ and putting all the other components of v equal to zero.

(b) “Activating” characteristics

$$c_\kappa^-(t_0, 0; \cdot) \quad \text{for } \kappa \in \{1, \dots, k\},$$

$$c_\lambda^+(t_0, 0; \cdot) \quad \text{for } \lambda \in \{1, \dots, l\}$$

such that

$$F_\lambda^-(t_0, 0) \in \{T\} \times [0, 1),$$

$$I_\kappa^+(t_0, 0) \in \{0\} \times [0, 1),$$

again putting all the other components of v equal to zero.

Whether method (b) is feasible for given k, l will be determined by the boundary condition at $(t_0, 0)$, i.e., by the matrix $C := C_0$. Let us define (with the convention $P_0 := 0, Q_L := 0$)

$$P_k : \mathbb{R}^K \rightarrow \mathbb{R}^K, \quad Q_l : \mathbb{R}^L \rightarrow \mathbb{R}^L,$$

$$\begin{bmatrix} v_1^- \\ \vdots \\ v_k^- \\ 0 \\ \vdots \\ 0 \end{bmatrix} \mapsto \begin{bmatrix} v_1^- \\ \vdots \\ v_k^- \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \begin{bmatrix} v_1^+ \\ \vdots \\ v_L^+ \end{bmatrix} \mapsto \begin{bmatrix} 0 \\ \vdots \\ 0 \\ v_{1+l}^+ \\ \vdots \\ v_L^+ \end{bmatrix}$$

For $k \in \{1, \dots, K\}$ let $l(k) \in \{1, \dots, L\}$ be the first integer such that the kernel of CP_k is different from the kernel of $Q_{l(k)}CP_k$, i.e.,

$$(8) \quad \ker(CP_k) = \dots = \ker(Q_{l(k)-1}CP_k) \subsetneq \ker(Q_{l(k)}CP_k),$$

and put

$$(9) \quad \tilde{T}_c := \max \{T_k^- + T_{l(k)}^+ \mid 1 \leq k \leq K\}.$$

$l(k)$ is not well defined by (8) if $CP_k = 0$. In this case we put $l(k) := L + 1$ and introduce an artificial transition time $T_{L+1}^+ := 0$ so that (9) still makes sense. (This corresponds to method (a) using a c^- -characteristic.) The other half of method (a) yields the largest time of nondistinguishability if the slowest c^+ -characteristic (i.e., c_1^+) is used. So we modify \tilde{T}_c and put

$$(9') \quad T_c := \max \{\tilde{T}_c, T_1^+\}.$$

THEOREM 1. *Let \tilde{B} be a diagonal matrix. Then for $T < T_c$ the TBVP is not distinguishable and the IBVP is not approximately controllable.*

Proof. Choose k such that $T < T_k^- + T_{l(k)}^+$ and $CP_k \neq 0$. (The special cases $T < T_1^+$ or $T < T_k^-$ and $CP_k = 0$ are treated similarly using method (a).) Choose $t_0 \in (0, T)$ such that

$$T_1^- \geq \dots \geq T_k^- > T - t_0,$$

$$T_1^+ \geq \dots \geq T_{l(k)}^+ > t_0$$

and therefore

$$0 < c_1^-(0, t_0; T) \leq \dots \leq c_k^-(0, t_0; T) < 1,$$

$$0 < c_1^+(0, t_0; 0) \leq \dots \leq c_{l(k)}^+(0, t_0; 0) < 1.$$

Hence, for sufficiently small ε and $\kappa \in \{1, \dots, k\}$, $\lambda \in \{1, \dots, l(k)\}$,

$$F_\kappa^- := F_\kappa^-((t_0 - \varepsilon, t_0 + \varepsilon) \times \{0\}) \subset \{T\} \times (0, 1),$$

$$I_\lambda^+ := I_\lambda^+((t_0 - \varepsilon, t_0 + \varepsilon) \times \{0\}) \subset \{0\} \times (0, 1).$$

Choose $v^- \in \ker(Q_{l(k)}CP_k) \setminus \ker(Q_{l(k)-1}CP_k)$ and define v_T by

$$v_T^+ := 0,$$

$$v_{T,\kappa}^-(x) := \begin{cases} v_\kappa^- & \text{if } \kappa \leq k \text{ and } (T, x) \in F_\kappa^-, \\ 0 & \text{otherwise.} \end{cases}$$

The corresponding solution v of the TBVP must then be zero on $[0, T] \times \{1\}$ and different from zero on $\{0\} \times I_{1(k)}^+$, which is nonvoid and relatively open in $\{0\} \times (0, 1)$. \square

For the corresponding observability result we shall need the following

LEMMA. *There exists $\gamma > 0$ such that, for each $k \in \{1, \dots, K\}$, $l \in \{1, \dots, l(k) - 1\}$ and each $(v^-, v^+) \in \mathbb{R}^K \times \mathbb{R}^L$ satisfying $v^+ = Cv^-$, we have*

$$\|v^+\|(\mathbb{R}^L) \leq \gamma \{ \|(I - P_k)v^-\|(\mathbb{R}^K) + \|Q_l v^+\|(\mathbb{R}^L) \}.$$

Proof. Assuming the contrary we get a sequence $(v^-(n), v^+(n))$ satisfying

$$(10) \quad v^+(n) = Cv^-(n),$$

$$(11) \quad \|v^+(n)\| > n \{ \|(I - P_k)v^-(n)\| + \|Q_l v^+(n)\| \}.$$

Decomposing

$$v^-(n) = w^-(n) + u^-(n) \in (\ker(CP_k))^\perp \oplus \ker(CP_k)$$

and replacing $v^-(n)$ by $P_k w^-(n) + (I - P_k)v^-(n)$ we see that (10) and (11) remain unchanged. Therefore without loss of generality we may assume

$$(12) \quad P_k v^-(n) \in (\ker(CP_k))^\perp.$$

Dividing by $\|v^-(n)\|$ and choosing convergent subsequences we get $(v^-, v^+) \in \mathbb{R}^K \times \mathbb{R}^L$ such that

$$(10') \quad v^+ = Cv^-,$$

$$(11') \quad (I - P_k)v^- = 0,$$

$$(11'') \quad Q_l v^+ = 0,$$

$$(12') \quad P_k v^- \in (\ker(CP_k))^\perp,$$

$$(13) \quad \|v^-\| = 1.$$

This implies

$$v^- \in \ker(Q_l CP_k) = \ker(CP_k),$$

contradicting (11'), (12') and (13). \square

THEOREM 2. *Let \tilde{B} be a diagonal matrix. Then for $T \geq T_c$ the TBVP is observable and the IBVP is exactly controllable.*

Proof. Let $l \in \{1, \dots, L\}$ be fixed. Then for $k \in \{1, \dots, K\}$ there are two possibilities:

Case 1. There exists a (unique) $x_k \in [0, 1)$ such that

$$F_k^-(F_l^+(0, x_k)) = (T, 1).$$

Case 2. For all $x \in [0, 1)$

$$F_k^-(F_l^+(0, x)) \in [0, T) \times \{1\}.$$

Defining $x_k := 1$ in case 2 we have in either case

$$(14) \quad F_k^-(F_l^+(0, x)) \in \begin{cases} \{T\} \times (0, 1) & \text{if } x > x_k, \\ [0, T) \times \{1\} & \text{if } x < x_k. \end{cases}$$

Furthermore because of (7)

$$0 =: x_0 \leq x_1 \leq x_2 \leq \dots \leq x_K \leq x_{K+1} := 1.$$

Therefore

$$\begin{aligned} \|v(0, \cdot)\|^2(L^2(0, 1)) &= \sum_{k=1}^K \|v_k^-(0, \cdot)\|^2 L^2(0, 1) + \sum_{l=1}^L \|v_l^+(0, \cdot)\|^2 L^2(0, 1) \\ &= \sum_{k=1}^K \|v_k^-(0, \cdot)\|^2 L^2(0, 1) + \sum_{l=1}^L \sum_{k=0}^K \|v_l^+(0, \cdot)\|^2 L^2(x_k, x_{k+1}). \end{aligned}$$

Hence to obtain the observability estimate (6) we have to estimate

$$\begin{aligned} \text{(a)} \quad & \|v_k^-(0, \cdot)\|^2 L^2(0, 1) \\ \text{(b)} \quad & \|v_l^+(0, \cdot)\|^2 L^2(0, x_1) \\ \text{(c)} \quad & \|v_l^+(0, \cdot)\|^2 L^2(x_k, x_{k+1}), \quad k = 1, \dots, K \end{aligned}$$

by $\|v^+(\cdot, 1)\|^2(L^2(0, T))$ or equivalently (because of (5')) by $\|v(\cdot, 1)\|^2(L^2(0, T))$.

Consider (c). Let $x \in (x_k, x_{k+1})$, $k \geq 1$. Then for $\lambda \geq l(k)$ we must have

$$(15) \quad I_\lambda^+(F_l^+(0, x)) \in (0, T) \times \{1\}.$$

For suppose to the contrary that

$$F_l^+(0, x) = (\tilde{t}, 0), \quad I_\lambda^+(\tilde{t}, 0) = (0, y) \in \{0\} \times [0, 1].$$

This implies $T_\lambda^+ \geq \tilde{t}$. We also have

$$F_k^-(F_l^+(0, x)) = F_k^-(\tilde{t}, 0) \in \{T\} \times (0, 1),$$

which implies $T_k^- > T - \tilde{t}$. Thus we reach a contradiction by

$$T_c \geq T_k^- + T_{l(k)}^+ \geq T_k^- + T_\lambda^+ > T.$$

Therefore (15) holds. Using this and (14) we have maps for $\lambda \in \{l(k), \dots, L\}$ and $\kappa \in \{k+1, \dots, K\}$:

$$\begin{aligned} g_\lambda &: (x_k, x_{k+1}) \rightarrow (0, T) \\ x &\mapsto t \quad (\text{if } I_\lambda^+(F_l^+(0, x)) = (t, 1)), \\ h_\kappa &: (x_k, x_{k+1}) \rightarrow (0, T) \\ x &\mapsto t \quad (\text{if } F_\kappa^-(F_l^+(0, x)) = (t, 1)). \end{aligned}$$

By what has been said in § 2, g_λ and h_λ are injective differentiable maps which together with their inverse maps $g_\lambda^{(-1)}$ and $h_\lambda^{(-1)}$ have uniformly bounded derivatives. For $x \in (x_k, x_{k+1})$ put $(t, 0) := F_l^+(0, x)$. Then by (4')

$$v^+(t, 0) = Cv^-(t, 0).$$

Therefore we may apply our lemma to get

$$|v_l^+(t, 0)|^2 \leq 2\gamma^2 \left\{ \sum_{\kappa=k+1}^K |v_\kappa^-(t, 0)|^2 + \sum_{\lambda=l(k)}^L |v_\lambda^+(t, 0)|^2 \right\}.$$

Since the components of v are constant along suitable characteristics, this implies

$$|v_l^+(0, x)|^2 \leq 2\gamma^2 \left\{ \sum_{\kappa=k+1}^K |v_\kappa^-(h_\kappa(x), 1)|^2 + \sum_{\lambda=l(k)}^L |v_\lambda^+(g_\lambda(x), 1)|^2 \right\}.$$

Thus (with $S = (x_k, x_{k+1})$)

$$\begin{aligned} \|v_l^+(0, \cdot)\|^2 L^2(S) &\leq 2\gamma^2 \left\{ \sum_{\kappa=k+1}^K \int_{h_\kappa(S)} |v_\kappa^-(t, 1)|^2 \left| \frac{d}{dt} h_\kappa^{(-1)}(t) \right| dt \right. \\ &\quad \left. + \sum_{\lambda=l(k)}^L \int_{g_\lambda(S)} |v_\lambda^+(t, 1)|^2 \left| \frac{d}{dt} g_\lambda^{(-1)}(t) \right| dt \right\} \\ &\leq \gamma_{l,k} \|v(\cdot, 1)\|^2 L^2(0, T) \end{aligned}$$

In order to estimate (a) and (b) we use the facts that for any k

$$F_k^-(0, x) \in (0, T) \times \{1\} \quad \text{if } x \in (0, 1),$$

and

$$F_k^-(F_l^+(0, x)) \in (0, T) \times \{1\} \quad \text{if } x \in (0, x_1),$$

and proceed similarly as in case (c).

4. Examples. Let the coupling at the boundary $[0, T] \times \{0\}$ be accomplished by

$$C := \begin{bmatrix} 1 & +1 \\ -1 & 1 \end{bmatrix}.$$

Then

$$CP_1 = \begin{bmatrix} 1 & 0 \\ -1 & 0 \end{bmatrix}, \quad \ker CP_1 = \text{lin} \{e^{(2)}\},$$

$$Q_1 CP_1 = \begin{bmatrix} 0 & 0 \\ -1 & 0 \end{bmatrix}, \quad \ker Q_1 CP_1 = \ker CP_1,$$

$$Q_2 CP_1 = 0, \quad \ker Q_2 CP_1 = \mathbb{R}^2 \neq \ker Q_1 CP_1,$$

$$l(1) = 2,$$

$$CP_2 = C = \begin{bmatrix} 1 & +1 \\ -1 & 1 \end{bmatrix}, \quad \ker CP_2 = \{0\},$$

$$Q_1 CP_2 = \begin{bmatrix} 0 & 0 \\ -1 & 1 \end{bmatrix}, \quad \ker Q_1 CP_2 = \text{lin} \{e^{(1)} + e^{(2)}\} \neq \ker CP_2,$$

$$l(2) = 1.$$

Therefore

$$T_c = \max \{T_1^- + T_2^+, T_2^- + T_1^+\}.$$

The times T_0, T_1 given in [2] would be

$$T_0 = T_1^- + T_2^+, \quad T_1 = T_1^+ + T_1^-.$$

(Note that in [2] the λ_k^- are in reverse order.) Thus if, e.g., $T_1^- = 2, T_2^- = T_2^+ = 1, T_1^+ = 3$, we find

$$T_0 = 3 < T_c = 4 < T_1 = 5.$$

This can be generalized: If $K = L = N/2$ and all $Q_{K-k}CP_k$ ($k = 1, \dots, K$) have maximal rank k , then

$$l(k) = K + 1 - k;$$

thus

$$T_c = \max \{T_k^- + T_{K+1-k}^+ | k = 1, \dots, K\};$$

whereas

$$T_0 = T_1^- + T_K^+, \quad T_1 = T_1^+ + T_1^-.$$

Acknowledgment. The author is indebted to the referee for a simplification in the proof of Theorem 2.

REFERENCES

- [1] S. DOLECKI AND D. L. RUSSELL, *A general theory of observation and control*, this Journal, 15 (1977), pp. 185–220.
- [2] D. L. RUSSELL, *Controllability and stabilizability theory for linear partial differential equations: Recent progress and open questions*, SIAM Rev. 20, (1978), pp. 639–739.

ADAPTIVE CONTROL WITH A COMPACT PARAMETER SET*

P. R. KUMAR†

Abstract. We consider the problem of the adaptive control of a Markov chain with unknown transition probabilities. We suppose the transition probabilities $\{p(i, j; u, \alpha)\}$ to be dependent on an unknown parameter α . At each time instant t , a maximum likelihood estimate $\hat{\alpha}_t$ of the unknown parameter is made, and a control input $u_t = \phi(x_t, \hat{\alpha}_t)$ is applied, where, for each α , $\phi(\cdot, \alpha)$ is a good feedback control law. It is shown that if α ranges over a compact set A , then $\{\hat{\alpha}_t\}$ may diverge with probability one. In the event, however, that the parameter estimates converge, or more generally just the control laws $\{\phi(\cdot, \hat{\alpha}_t)\}$ converge to some ψ , then under ψ the closed-loop transition probabilities for the true model are indistinguishable from those of any limit point of the parameter estimates.

1. Introduction. Let X and U , both *finite*, be respectively the state-space and control-set of a Markov chain. For each α belonging to a *compact metric space* A of possible models, we are given a set of transition probabilities $\{p(i, j; u, \alpha) : (i, j, u) \in X \times X \times U\}$, where each $p(i, j; u, \alpha)$ is the probability of transfer of the state of the system from i to j under the action of u in model α . We do not know the *true model* α^0 , which however belongs to A .

For each model $\alpha \in A$, we have a feedback control law $\phi(\cdot, \alpha) : X \rightarrow U$ which is "good" for model α . We analyze the behavior of the following simple adaptive control scheme. At each time t , we make a maximum-likelihood estimate

$$(1) \quad \hat{\alpha}_t := \arg \max_{\alpha \in A} \prod_{s=0}^{t-1} p(x_s, x_{s+1}; u_s, \alpha)$$

of the unknown parameter, and then apply the control input

$$(2) \quad u_t = \phi(x_t, \hat{\alpha}_t).$$

Here x_s and u_s are the values of the state and control at time s . To avoid ambiguity of choice in (1) we may impose some priority ordering of elements of A such that, if more than one element of A maximizes (1), then we can unambiguously choose the particular maximizer which is highest in the priority ordering. We assume

(i) $p(\cdot, \cdot; \cdot, \cdot)$ and $\phi(\cdot, \cdot)$ are continuous.

(ii) For each $(i, j) \in X \times X$,

$$(3) \quad \begin{array}{ll} \text{either} & p(i, j; u, \alpha) > 0 \quad \text{for all } (u, \alpha) \in U \times A \\ \text{or} & p(i, j; u, \alpha) = 0 \quad \text{for all } (u, \alpha) \in U \times A. \end{array}$$

(iii) For every $(i, j) \in X \times X$, there exists a sequence $i = i_0, i_1, \dots, i_r = j$ such that

$$p(i_{s-1}, i_s; u_s, \alpha) > 0 \quad \text{for all } s = 1, 2, \dots, r.$$

The problem considered is therefore more general than the one considered by Borkar and Varaiya [1], where A is restricted to be finite. In turn [1] is a relaxation of the identifiability condition of Mandl [2].

In § 2 we provide a counterexample to convergence of the parameter estimates. This shows that the results of [1] cannot be extended without further hypotheses. In § 3 we show that, if either the parameter estimates converge or more generally the

* Received by the editors December 26, 1979, and in final form March 11, 1981. The research reported here has been supported by the U.S. Army Research Office under contract DAAG-29-80-K0038.

† Department of Mathematics, University of Maryland Baltimore County, 5401 Wilkens Avenue, Baltimore, Maryland 21228.

control law $\phi(\cdot, \hat{\alpha}_t)$ converges to some ψ , then under ψ the closed-loop transition probabilities of α^0 coincide with those of every limit point α^* of $\{\hat{\alpha}_t\}$.

2. A counterexample to convergence. We now provide an example where the parameter estimates diverge with probability one.

Throughout, let $(\Omega, \mathcal{F}, \mathcal{P})$ be the underlying probability space, with \mathcal{P} the probability measure induced on the trajectories of the system by the adaptive control law (1)–(2). Define

$$l_t(\alpha, \omega) := \frac{p(x_t(\omega), x_{t+1}(\omega); u_t(\omega), \alpha)}{p(x_t(\omega), x_{t+1}(\omega); u_t(\omega), \alpha^0)},$$

$$L_t(\alpha, \omega) := \prod_{s=0}^{t-1} l_s(\alpha, \omega).$$

From (1),

$$(4) \quad L_t(\hat{\alpha}_t(\omega), \omega) \geq L_t(\alpha, \omega) \quad \text{for all } \alpha \in A.$$

Example 1.

$$X = \{1, 2\}, \quad U = \{a, b\},$$

$$A = \{-1, \dots, -\frac{31}{32}, -\frac{15}{16}, -\frac{7}{8}, -\frac{3}{4}, -\frac{1}{2}, \frac{1}{2}, \frac{3}{4}, \frac{7}{8}, \frac{15}{16}, \frac{31}{32}, \dots, 1\},$$

$$\phi(i, \alpha) = \begin{cases} a & \text{for all } \alpha < 0 \\ b & \text{for all } \alpha > 0 \end{cases} \quad \text{irrespective of } i,$$

$$p\left(1, 1; a, \frac{1-2^n}{2^n}\right) = 1 - p\left(1, 1; b, \frac{1-2^n}{2^n}\right)$$

$$= 1 - p\left(1, 1; a, \frac{2^n-1}{2^n}\right) = p\left(1, 1; b, \frac{2^n-1}{2^n}\right) = \frac{2^n-1}{2^n+1},$$

$$p(1, 1; a, -1) = 1 - p(1, 1; b, -1)$$

$$= 1 - p(1, 1; a, +1) = p(1, 1; b, +1) = \frac{1}{2},$$

$$p(2, 1; a, \alpha) = p(2, 1; b, \alpha) = 1 \quad \text{for all } \alpha \in A.$$

Let $\alpha^0 = 1$ be the true model and the priority ordering on A simply the natural ordering of the real line.

Define

$$n_a^i(t) := \sum_{s=0}^{t-1} 1(x_s = 1, x_{s+1} = i, u_s = a)$$

for $i = 1, 2$, $n_a(t) := n_a^1(t) + n_a^2(t)$ and similar quantities with subscript b . Using $p(1, 1; b, \alpha) = 1 - p(1, 1; a, \alpha)$ gives

$$L_t(\alpha, \omega) = \frac{p(1, 1; a, \alpha)^{n_a^1(t, \omega) + n_b^2(t, \omega)} [1 - p(1, 1; a, \alpha)]^{n_a^2(t, \omega) + n_b^1(t, \omega)}}{(0.5)^{n_a(t, \omega) + n_b(t, \omega)}}.$$

Clearly $\hat{\alpha}_t(\omega) > 0$ if and only if $n_a^1(t, \omega) + n_b^2(t, \omega) \geq n_a^2(t, \omega) + n_b^1(t, \omega)$.

Suppose that for some $\omega \hat{\alpha}_t(\omega)$ converges to a positive value; then $\hat{\alpha}_t(\omega) > 0$ for every t larger than some finite τ . But then $\phi(1, \hat{\alpha}_t(\omega)) = b$ and hence $n_a^i(t, \omega) = n_a^i(\tau, \omega)$ for $t \geq \tau$. As a consequence, $n_b^2(t, \omega) - n_b^1(t, \omega) \geq n_a^2(t, \omega) - n_a^1(t, \omega) = \text{constant}$ for $t \geq \tau$. But this is a zero probability event since $n_b^2(t, \omega) - n_b^1(t, \omega)$ denotes the position of a random walk since $p(1, 1; b, \alpha^0) = p(1, 2; b, \alpha^0) = \frac{1}{2}$. Similarly, convergence of $\hat{\alpha}_t$ to a negative value is also a zero probability event, thus showing the almost sure divergence of $\hat{\alpha}_t$.

Remark. Note that $\alpha = +1$ and $\alpha = -1$ are for all intents and purposes identical and $\hat{\alpha}_t \rightarrow \{+1, -1\}$. However identifying $\alpha = +1$ and $\alpha = -1$ as one single point would violate the continuity assumption on $\phi(\cdot, \alpha)$. More hypotheses are clearly needed before one can expect to extend the results of [1].

3. Asymptotic properties of estimates. Motivated by the example of the preceding section, we study the properties of the limit points of $\{\hat{\alpha}_t\}$.

DEFINITION. Let $\tau_n(i, \omega) := \inf \{t : t > \tau_{n-1}(i, \omega), x_t(\omega) = i\}$, where $\inf \phi := +\infty$. For fixed $\omega \in \Omega$, the *regulator* is said to *converge to ψ* if $\lim_{n \rightarrow \infty} u_{\tau_n(i, \omega)}(\omega) = \psi(i)$ for every $i \in X$ for which $x_t(\omega) = i$ infinitely often. ψ will be called the *limiting regulator* along ω .

The main result of this section is:

THEOREM 2. *There exists a null set $N \subset \Omega$, $\mathcal{P}(N) = 0$ such that, if for some $\omega \in N^c$ the regulator converges to ψ , then*

$$p(i, j; \psi(i), \alpha^*) = p(i, j; \psi(i), \alpha^0) \quad \text{for every } (i, j) \in X \times X$$

and every limit point α^* of $\{\hat{\alpha}_t(\omega)\}_{t=1}^\infty$.

Noting that if $\hat{\alpha}_t(\omega) \rightarrow \alpha^*$ then the regulator converges to $\phi(\cdot, \alpha^*)$ along ω , we obtain the following important corollary of Theorem 2.

COROLLARY 3. *There exists a null set $N \subset \Omega$, $\mathcal{P}(N) = 0$ such that, if for some $\omega \in N^c$ $\lim_{t \rightarrow \infty} \hat{\alpha}_t(\omega) = \alpha^*$, then*

$$p(i, j; \phi(i, \alpha^*), \alpha^*) = p(i, j; \phi(i, \alpha^*), \alpha^0) \quad \text{for every } (i, j) \in X \times X.$$

Remark. If A is a connected set, then the continuity assumption on $\phi(i, \cdot) : A \rightarrow U$ would render it a constant function since U is a finite (hence discrete) set, and so Theorem 2 would trivially follow. Our results are therefore addressed to those situations where A is not connected.

The proof of Theorem 2 rests on the following nonprobabilistic result.

LEMMA 4. *Let $B = \{b_1, \dots, b_m\} \subset (0, \infty)$ and suppose that for each i , $z_i \in B$ satisfies*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n z_i = 1 \quad \text{and} \quad \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \ln z_i = 0.$$

If

$$d_j := \liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n 1(z_i = b_j) > 0 \quad \text{for some } j \in \{1, \dots, m\},$$

then $b_j = 1$.

Proof. Let $d_j^n := (1/n) \sum_{i=1}^n 1(z_i = b_j)$ and note that $z_i = \sum_{j=1}^m 1(z_i = b_j) b_j$ and $\ln z_i = \sum_{j=1}^m 1(z_i = b_j) \ln b_j$. Now

$$\begin{aligned} 1 &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n z_i = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \left[\sum_{j=1}^m 1(z_i = b_j) b_j \right] \\ &= \lim_{n \rightarrow \infty} \sum_{j=1}^m \left[\frac{1}{n} \sum_{i=1}^n 1(z_i = b_j) \right] b_j \\ &= \lim_{n \rightarrow \infty} \sum_{j=1}^m d_j^n b_j \\ &= \lim_{n \rightarrow \infty} \sum_{j=1}^m d_j^n \exp(\ln b_j) \end{aligned}$$

$$\begin{aligned}
&\geq \limsup_{n \rightarrow \infty} \exp \left[\sum_{j=1}^m d_j^n \ln b_j \right] \\
&= \exp \left[\limsup_{n \rightarrow \infty} \sum_{j=1}^m d_j^n \ln b_j \right] \\
&= \exp \left[\limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^m 1(z_i = b_j) \ln b_j \right) \right] \\
&= \exp \left[\limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \ln z_i \right] \\
&= \exp(0) = 1.
\end{aligned}$$

The inequality above follows because $\sum_{j=1}^m d_j^n = 1$, $d_j^n \geq 0$ and the exponential function is convex. Since equality holds throughout,

$$(5) \quad 1 = \lim_{n \rightarrow \infty} \sum_{j=1}^m d_j^n b_j = \limsup_{n \rightarrow \infty} \left[\exp \left(\sum_{j=1}^m d_j^n \ln b_j \right) \right].$$

Now we can pick a common subsequence $\{n_k\}$ such that $\lim_{k \rightarrow \infty} [\sum_{j=1}^m d_j^{n_k} \ln b_j] = 0$ and, for each j , $d_j^{n_k}$ converges to some d_j . From (5) we get $0 = \ln \sum_{j=1}^m d_j b_j = \sum_{j=1}^m d_j \ln b_j$. But from the strict concavity of $\ln(\cdot)$, if $d_j > 0$ then $b_j = 1$, proving the lemma. \square

Now we can prove Theorem 2.

Proof of Theorem 2. By using the stability theorem of Loève [3] on the random variable $l_t(\alpha)$ for each α in a countable dense subset \tilde{A} of A , we conclude that

$$(6) \quad \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{s=0}^{t-1} l_s(\alpha, \omega) = 1 \quad \text{for all } \omega \in N^c, \quad \mathcal{P}(N) = 0$$

and all $\alpha \in \tilde{A}$. By using the Ascoli theorem on the equicontinuous family $\{(1/t) \sum_{s=0}^{t-1} l_s(\cdot, \omega)\}_{t=1}^{\infty}$ we can extend (6) to all $\alpha \in A$. By the concavity of $\ln(\cdot)$, it follows from (6) that

$$(7) \quad \limsup_{t \rightarrow \infty} \frac{1}{t} \sum_{s=0}^{t-1} \ln l_s(\alpha, \omega) \leq 0 \quad \text{for all } \omega \in N^c, \quad \alpha \in A.$$

Fix $\omega \in N^c$ and let α^* be a limit point of $\{\hat{\alpha}_t(\omega)\}_{t=1}^{\infty}$, obtained in particular by taking the limit along the subsequence $\{t_k\}$. Since $(1/t) \sum_{s=0}^{t-1} \ln l_s(\cdot, \omega)$ is continuous in α uniformly in t , for every $\varepsilon > 0$, there exists K , such that

$$\left| \frac{1}{t_k} \sum_{s=0}^{t_k-1} \ln l_s(\hat{\alpha}_{t_k}(\omega), \omega) - \sum_{s=0}^{t_k-1} \ln l_s(\alpha^*, \omega) \right| < \varepsilon \quad \text{for } k \geq K,$$

and hence from (4) that

$$(8) \quad \frac{1}{t_k} \sum_{s=0}^{t_k-1} \ln l_s(\alpha^*, \omega) \geq -\varepsilon \quad \text{for every } k \geq K.$$

From (7) and (8),

$$(9) \quad \limsup_{t \rightarrow \infty} \frac{1}{t} \sum_{s=0}^{t-1} \ln l_s(\alpha^*, \omega) = 0$$

for every $\omega \in N^c$ and every limit point α^* of $\{\hat{\alpha}_t(\omega)\}_{t=1}^{\infty}$. Additionally, by considering a hypothetical long-term average cost problem with one stage cost equal to $-1(x_s = i$,

$x_{s+1} = j$), we conclude from Theorem 4 of Mandl [1] that

$$(10) \quad \liminf_{t \rightarrow \infty} \frac{1}{t} \sum_{s=0}^{t-1} 1(x_s(\omega) = i, x_{s+1}(\omega) = j) > 0 \quad \text{for every } \omega \in N^c, \quad \mathcal{P}(N) = 0$$

and (i, j) such that $p(i, j; u, \alpha) > 0$ for all (u, α) . Now suppose $\omega \in N^c$ is such that the conditions of the theorem hold. By the finiteness of U and X , there exists $\tau < \infty$ such that $u_t(\omega) = \phi(x_t(\omega), \hat{\alpha}_t(\omega)) = \psi(x_t(\omega))$ for $t \geq \tau$. Hence $1(x_t(\omega) = i, x_{t+1}(\omega) = j) = 1(x_t(\omega) = i, x_{t+1}(\omega) = j, u_t(\omega) = \psi(i))$ for $t \geq \tau$, and so from (10)

$$\liminf_{t \rightarrow \infty} \frac{1}{t} \sum_{s=0}^{t-1} 1(x_s(\omega) = i, x_{s+1}(\omega) = j, u_s(\omega) = \psi(i)) > 0,$$

and therefore

$$(11) \quad \liminf_{t \rightarrow \infty} \frac{1}{t} \sum_{s=0}^{t-1} 1\left(l_s(\alpha^*, \omega) = \frac{p(i, j; \psi(i), \alpha^*)}{p(i, j; \psi(i), \alpha^0)}\right) > 0$$

for every (i, j) such that $p(i, j; u, \alpha) > 0$ for all (u, α) . Now let $B := \{l_t(\alpha^*, \omega) : t \geq 1\}$ and note that B is finite since X and U are. A comparison of (6), (9) and (11) with Lemma 4 shows that

$$\frac{p(i, j; \psi(i), \alpha^*)}{p(i, j; \psi(i), \alpha^0)} = 1 \quad \text{if } p(i, j; u, \alpha) > 0 \quad \text{for all } (u, \alpha).$$

On the other hand, from (3), if $p(i, j; u, \alpha) = 0$ for some (u, α) then again $p(i, j; \psi(i), \alpha^*) = p(i, j; \psi(i), \alpha^0)$, proving the theorem. \square

4. Concluding remarks. This paper clearly shows that a priority ordering of elements of A , even with the true parameter highest in the ordering, is not enough to eliminate oscillatory behavior of the parameter estimates when the parameter set A is compact. Hence additional hypotheses are needed to guarantee convergence of the parameter estimates. This makes the compact case different from the finite case treated in [1].

However, if the parameter estimates do converge or, more generally, even if only the control laws converge to some feedback law ψ , then under ψ the transition probabilities under any limit point of the parameter estimates are the true transition probabilities. Thus, subject to convergence, closed-loop identification takes place just as in the finite case [1].

Acknowledgments. The author is grateful to A. Becker, A. Pittenger and T. Seidman for useful discussions.

REFERENCES

- [1] V. BORKAR AND P. VARAIYA, *Adaptive control of Markov chains, I: Finite parameter set*, IEEE Trans on Automat Control, AC-24 (1979), pp. 953-958.
- [2] P. MANDL, *Estimation and control in Markov chains*, Adv. Appl. Prob., 6 (1974), pp. 40-60.
- [3] M. LOËVE, *Probability Theory*, Van Nostrand, New York, 1963.

HEDGING AND MAXMIN*

ELLIOT WINSTON†

Abstract. A continuous game is discussed in which two opposing players move in turn, and an algorithm is developed which finds an optimal strategy for the player moving first. The algorithm is characterized by a fictitious sequence of alternating plays in which the first player selects his strategies by hedging against all his opponent's strategies played up until that time.

1. Introduction. We consider a game consisting of a single play by each of two opposing players in turn. In particular, the x -player is assumed to select his strategy first, followed by the y -player who responds with full knowledge of his opponent's choice of play. This type of game is frequently encountered in models of military conflicts, which typically allow each opponent only one chance to select a strategy. Military planning, which is naturally conservative, seeks that strategy which provides the best overall protection against all opposing strategies. In this presentation, the role of the military planner is assumed by the x -player. The x -player tries to maximize the payoff function, $f(x, y)$, which the y -player tries to minimize. If $y(x)$ represents an optimal response by the y -player to an x -strategy, that is,

$$f(x, y(x)) = \min_y f(x, y),$$

then the x -player must try to determine a strategy, x^* , which satisfies

$$f(x^*, y(x^*)) = \max_x f(x, y(x)) = \max_x \min_y f(x, y).$$

If y' is any admissible y -strategy, then

$$f(x^*, y') \geq \min_y f(x^*, y) = f(x^*, y(x^*)) = \max_x \min_y f(x, y),$$

which is the usual definition of an optimal x -strategy. The order of the players' turns is important because no assumptions are made about the existence of a saddle point, thereby allowing the possibility that $\max_x \min_y f(x, y)$ is unequal to, and hence strictly less than, $\min_y \max_x f(x, y)$.

The algorithm developed to find an optimal strategy, x^* , is motivated by the notion of hedging. We visualize a series of plays of the game in which the players respond to each other alternately. Suppose the x -player initiates play with the strategy x_0 . The y -player responds optimally with $y_1 = y(x_0)$. The x -player then counters against y_1 with x_1 , determined by

$$f(x_1, y_1) = \max_x f(x, y_1),$$

and the y -player responds with $y_2 = y(x_1)$. Now the x -player hedges against y_1 and y_2 ; that is, he plays x_2 determined by

$$\min_{k=1,2} f(x_2, y_k) = \max_x \min_{k=1,2} f(x, y_k).$$

The play alternates in an analogous manner, with x_n and y_{n+1} being determined by

$$\min_{k=1,\dots,n} f(x_n, y_k) = \max_x \min_{k=1,\dots,n} f(x, y_k)$$

* Received by the editors April 17, 1980, and in revised form March 18, 1981.

† Naval Surface Weapons Center, Silver Spring, Maryland, 20910.

and

$$f(x_n, y_{n+1}) = \min_y f(x_n, y)$$

respectively. In the next section, we give conditions under which any accumulation point of $\{x_n\}$ is an optimal x -strategy.

We illustrate the algorithm with the simple “seesaw” example originally discussed by Danskin [2] in his work dealing with an alternate approach to the problem. The payoff function is defined by

$$f(x, y) = y \sin x$$

for $-\pi/2 \leq x \leq \pi/2$ and $-1 \leq y \leq 1$. The x -player chooses the angle the seesaw makes with the horizontal, and the y -player chooses a point somewhere between the left end, $y = -1$, and the right end, $y = 1$. The payoff is the height of the point chosen by the y -player above the horizontal. Suppose the x -player begins with some x_0 between 0 and $\pi/2$. Since the objective of the y -player is to minimize the payoff, his response is $y_1 = -1$. The x -player counters with $x_1 = -\pi/2$, and the y -player responds with $y_2 = 1$. In choosing his next strategy, the x -player hedges against y_1 and y_2 , and thus plays $x_2 = 0$, the point corresponding to the maximum of $\min_{k=1,2} f(x, y_k) = -|\sin x|$. See Fig. 1. The value of the game is now equal to 0, and remains equal to 0 regardless of how the y -player responds. Thus, the algorithm has converged, and $x^* = 0$. The same conclusion is obtained if $-\pi/2 \leq x_0 \leq 0$.

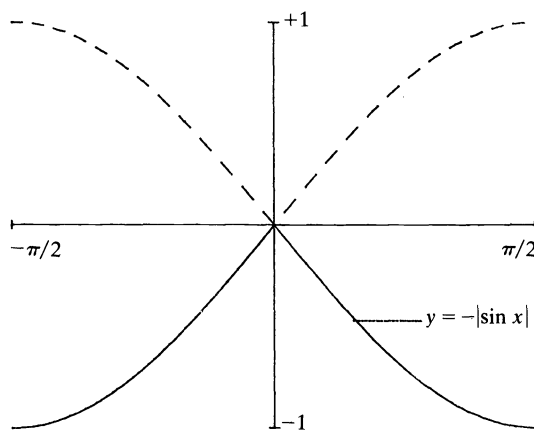


FIG. 1

The difficulty of calculating the y -responses $\{y_n\}$ and the hedged x -strategies $\{x_n\}$ varies with the particular game under consideration. In § 3, we discuss an example arising from an anti-submarine warfare (ASW) model which, happily, is endowed with a rich structure, thus enabling these sequences to be easily determined.

2. Results. X and Y are assumed to be subsets of a pair of arbitrary metric spaces throughout this section. The following lemma is a well-known, elementary result, and is therefore stated without proof.

LEMMA. *Let $f: X \times Y \rightarrow \mathbb{R}$ be a continuous real-valued function, where Y is compact. If $\{y_n\} \subset Y$, then $h(x) = \inf_n f(x, y_n)$ is continuous on X .*

THEOREM. *Let $f: X \times Y \rightarrow \mathbb{R}$ be a continuous real-valued function, where X and Y are compact. For $x \in X$, let $y(x)$ satisfy*

$$f(x, y(x)) = \min_y f(x, y).$$

Let $x_0 \in X$ and define the sequences $\{x_n\}$ and $\{y_n\}$ by $y_n = y(x_{n-1})$ and

$$h_n(x_n) = \max_x h_n(x),$$

where

$$h_n(x) = \min_{k=1, \dots, n} f(x, y_k).$$

If x^ is an accumulation point of $\{x_n\}$, then*

$$\lim_{n \rightarrow \infty} h_n(x_n) = f(x^*, y(x^*)) = \max_x f(x, y(x)).$$

Proof. We first note that $\{x_n\}$ exists because each $h_n(x)$ is continuous. Also, $\{h_n(x)\}$ is a monotonic decreasing sequence of functions because

$$h_{n+1}(x) = \min_{k=1, \dots, n+1} f(x, y_k) \leq \min_{k=1, \dots, n} f(x, y_k) = h_n(x).$$

If $h^*(x)$ denotes the limit function of $\{h_n(x)\}$, the lemma implies that $h^*(x)$ is continuous. By Dini's theorem [1], the convergence is uniform.

Let x^* be an accumulation point of $\{x_n\}$. Then there is a subsequence $\{x_{n_k}\}$ such that $x_{n_k} \rightarrow x^*$. We next make the estimate

$$|h_{n_k}(x_{n_k}) - h^*(x^*)| \leq |h_{n_k}(x_{n_k}) - h^*(x_{n_k})| + |h^*(x_{n_k}) - h^*(x^*)|.$$

The first difference on the right is small for large k by uniform convergence, and the second is small by the continuity of $h^*(x)$. Hence

$$\lim_{k \rightarrow \infty} h_{n_k}(x_{n_k}) = h^*(x^*).$$

However, $\{h_n(x_n)\}$ itself converges because it is monotone,

$$h_{n+1}(x_{n+1}) \leq h_n(x_{n+1}) \leq h_n(x_n).$$

Therefore

$$\lim_{n \rightarrow \infty} h_n(x_n) = h^*(x^*).$$

For all n ,

$$h_n(x) \leq h_n(x_n)$$

implies

$$h^*(x) \leq h^*(x^*),$$

that is,

$$h^*(x^*) = \max_x h^*(x).$$

Next we apply the lemma to obtain

$$\min_y f(x_{n_k}, y) \rightarrow \min_y f(x^*, y),$$

or

$$f(x_{n_k}, y(x_{n_k})) \rightarrow f(x^*, y(x^*)).$$

Since Y is compact, there exists a convergent subsequence of $\{y(x_{n_k})\}$, say $y(x_{n_{k_l}}) \rightarrow \bar{y}$. Thus

$$\begin{aligned} f(x_{n_{k_l}}, y(x_{n_{k_l}})) &\rightarrow f(x^*, \bar{y}) \\ &\Rightarrow f(x^*, y(x^*)) = f(x^*, \bar{y}), \end{aligned}$$

and

$$\begin{aligned} f(x^*, y(x_{n_{k_l}})) &\rightarrow f(x^*, \bar{y}) \\ &\Rightarrow \inf_n f(x^*, y(x_n)) \leq \inf_l f(x^*, y(x_{n_{k_l}})) \leq f(x^*, y(x^*)). \end{aligned}$$

However, the definition of $y(x^*)$ implies

$$f(x^*, y(x^*)) \leq f(x^*, y(x_n))$$

for all n , and thus

$$f(x^*, y(x^*)) \leq \inf_n f(x^*, y(x_n)).$$

Therefore

$$f(x^*, y(x^*)) = \inf_n f(x^*, y(x_n)) = h^*(x^*) = \lim_{n \rightarrow \infty} h_n(x_n).$$

Finally,

$$\begin{aligned} f(x, y(x)) &\leq f(x, y_n) \\ &\Rightarrow f(x, y(x)) \leq h^*(x) \\ &\Rightarrow \max_x f(x, y(x)) \leq \max_x h^*(x) = h^*(x^*) = f(x^*, y(x^*)), \end{aligned}$$

which implies that, in fact, equality holds, and the theorem is proved. \square

The pair (\bar{x}, \bar{y}) is a *saddle point* of $f(x, y)$ if

$$f(x, \bar{y}) \leq f(\bar{x}, \bar{y}) \leq f(\bar{x}, y)$$

for all $x \in X$ and all $y \in Y$. Note that, if (\bar{x}, \bar{y}) is a saddle point, it follows that

$$f(\bar{x}, \bar{y}) = \max_x \min_y f(x, y) = \min_y \max_x f(x, y).$$

Sufficient conditions for $f(x, y)$ to have a saddle point are that $f(x, y)$ is continuous, concave in x for fixed y , convex in y for fixed x , and that X and Y are convex (see [3, p. 28]).

COROLLARY. *If $f(x, y)$ has a saddle point and the y -response to x^* , $y(x^*)$, is unique, then $(x^*, y(x^*))$ is a saddle point.*

Proof. Let $v = \max \min f(x, y)$ and let (\bar{x}, \bar{y}) be a saddle point of $f(x, y)$. Then $f(x, \bar{y}) \leq v$ for all x . In particular, $f(x^*, \bar{y}) \leq v$. But $f(x^*, \bar{y}) \geq v$ because x^* is optimal. Hence $f(x^*, \bar{y}) = v$, which implies that \bar{y} is an optimal response to x^* . By uniqueness, $\bar{y} = y(x^*)$ and the result follows. \square

3. An application. We study the problem of optimally assigning submarines to patrol zones in such a way as to maximize the number surviving an attack by “hunter-killer” ASW aircraft. To begin, suppose a single submarine is patrolling an area A , and comes under attack by y aircraft at time $t = 0$. The aircraft are assumed to search the area randomly, and the submarine is destroyed if it is detected. The motion of the submarine is approximately stationary compared to that of the aircraft over a small instant of time Δt , so that the probability of detection over this interval is $Sy\Delta t/A$, where S is the area sweep rate of one aircraft. If $p(t)$ denotes the probability that the submarine survives at time t , then

$$\begin{aligned} p(t + \Delta t) &= p(t) \left(1 - \frac{Sy\Delta t}{A} \right) \Rightarrow \frac{p(t + \Delta t) - p(t)}{\Delta t} = -\frac{Sy}{A} p(t) \\ &\Rightarrow \dot{p}(t) = -\frac{Sy}{A} p(t) \\ &\Rightarrow p(t) = e^{-Sy t/A}. \end{aligned}$$

We now introduce the following notation:

N_j = number of patrol zones in ocean j , $j = 1, 2, \dots, J$;

X_j = number of submarines assigned to ocean j ;

Y = total number of aircraft;

A_{ij} = area of patrol zone i in ocean j , $i = 1, 2, \dots, N_j$;

T_{ij} = amount of time for search by aircraft in zone i , ocean j ;

B_{ij} = logistical bound on number of submarines allowed in zone i , ocean j ;

x_{ij} = number of submarines in zone i , ocean j ;

y_{ij} = number of aircraft in zone i , ocean j .

The quantity of interest is the expected number of submarines surviving the attack. The statement of the optimization problem is

$$\max_x \min_y \sum_{j=1}^J \sum_{i=1}^{N_j} x_{ij} e^{-a_{ij} y_{ij}},$$

where $a_{ij} = ST_{ij}/A_{ij}$, subject to the constraints

$$\begin{aligned} \sum_{i=1}^{N_j} x_{ij} &= X_j, & j &= 1, 2, \dots, J, \\ \sum_{j=1}^J \sum_{i=1}^{N_j} y_{ij} &= Y, \\ 0 &\leq x_{ij} \leq B_{ij}, & \text{all } i, j, \\ 0 &\leq y_{ij}, & \text{all } i, j. \end{aligned}$$

Since the payoff function is concave in x and convex in y , it has a saddle point. In fact, $f(x, y)$ is a strictly convex function of y defined on a convex set, and hence optimal responses are unique. Therefore, $(x^*, y(x^*))$ is a saddle point. The reader should keep in mind that our algorithm finds continuous solutions, although the model is, strictly speaking, discrete.

The Kuhn–Tucker conditions may be used to find the optimal ASW response, $y = y(x)$, to a given submarine distribution, x . Specifically, there exists a scalar λ such that

$$y_{ij}(a_{ij}x_{ij}e^{-a_{ij}y_{ij}} - \lambda) = 0.$$

Thus

$$y_{ij} > 0 \Rightarrow y_{ij} = \frac{1}{a_{ij}} \ln \frac{a_{ij}x_{ij}}{\lambda}.$$

Substitution into the conservation constraint for Y leads to consideration of the function

$$g(\lambda) = \sum_{\lambda < a_{ij} < x_{ij}} \frac{1}{a_{ij}} \ln \frac{a_{ij}x_{ij}}{\lambda}.$$

We note that $g(\lambda)$ is defined on $(0, \infty)$, continuous, strictly decreasing, $\lim_{\lambda \rightarrow 0^+} g(\lambda) = \infty$, and $\lim_{\lambda \rightarrow \infty} g(\lambda) = 0$. (In fact, $g(\lambda) = 0$ for $\lambda \geq \max_{i,j} a_{ij}x_{ij}$.) Hence, for any positive Y , $g(\lambda) = Y$ has a unique root. Moreover, $g(\lambda)$ is convex and differentiable except on a finite set of points, so that Newton's method may be applied to quickly calculate the root.

The problem of finding the hedged submarine distribution, x , against the first m ASW responses, $\{y^{(k)}\}_{k=1}^m$, may be formulated as the following linear program:

$$\max z$$

such that

$$\begin{aligned} \sum_{i=1}^{N_j} x_{ij} &= X_j, & j &= 1, 2, \dots, J, \\ z &\leq \sum_{j=1}^J \sum_{i=1}^{N_j} e^{-a_{ij}y_{ij}^{(k)}} x_{ij}, & k &= 1, 2, \dots, m, \\ 0 &\leq x_{ij} \leq B_{ij}. \end{aligned}$$

The dual simplex algorithm is the preferred method of solution because each successive hedge adds one additional constraint. It is interesting to note that this sequence of hedging problems resembles the sequence of linear programming problems associated with cutting plane algorithms.

Numerical experiments have been performed on problems consisting of two oceans, each containing ten zones. To maintain sufficient accuracy, the basis matrix of the linear program was reinverted every twenty pivots. The algorithm generally converged within 65 hedges and had an execution time of about 10 seconds on a CDC 6600.

REFERENCES

- [1] T. M. APOSTOL, *Mathematical Analysis*, Addison-Wesley, Reading, MA, 1964.
- [2] J. M. DANSKIN, *The Theory of Max–Min*, Springer-Verlag, New York, 1967.
- [3] S. KARLIN, *Mathematical Methods and Theory in Games, Programming, and Economics*, vol. I, Addison-Wesley, Reading, MA, 1959.

SECOND-ORDER SUFFICIENCY CONDITIONS FOR NONDIFFERENTIABLE PROGRAMMING PROBLEMS*

R. W. CHANEY†

Abstract. Second-order conditions are given which are sufficient for a point to be a local minimizer for a finite-dimensional nonlinear programming problem with a finite number of constraints. In the most general theorem, the functions which comprise the problem are required only to be locally Lipschitz. The sufficiency conditions are given in terms of Clarke generalized gradients. These conditions assume a somewhat more familiar form when the functions in the problem are assumed to be both semismooth and subdifferentially regular.

Introduction. We shall give sufficient conditions for optimality for nonlinear programming problems comprised of locally Lipschitz functions. Since the functions of the problem are not differentiable everywhere, we must have suitable replacements for gradients and for Hessians. We shall use the Clarke generalized gradients as the necessary replacement. The properties of generalized gradients are set forth in [3] and [2]; also, see [19] for a comprehensive account. The Clarke generalized gradient of a function f reduces to the subdifferential ∂f when f is convex and to the set consisting of the gradient ∇f alone when f is continuously differentiable. The definitions and results concerning generalized gradients which we shall use are given below.

We now state the problem of interest more explicitly. Let W be an open set in n -dimensional real Euclidean space R^n . Let $g_0, g_1, g_2, \dots, g_q$ be real-valued functions on R^n which are locally Lipschitz (as defined below) on W . Next, let

$$(1) \quad S = \bigcap_{i=1}^m \{x \in R^n: g_i(x) \leq 0\} \cap \bigcap_{i=m+1}^q \{x \in R^n: g_i(x) = 0\}.$$

In much of this paper, we shall consider the problem

P: Minimize $g_0(x)$ over x in $S \cap W$.

We shall also consider the unconstrained problem

P*: Minimize $F(x)$ over x in W ;

here, F is assumed to be a real-valued locally Lipschitz function on W .

Clarke [2, Thm. 1] has given first-order necessary conditions for a point x^* in $S \cap W$ to be optimal for problem P. Clarke's theorem will play an important role in the results presented here. Hiriart-Urruty [6, Thm. 6] has given a different result of comparable generality. The theorems of Clarke and Hiriart-Urruty are expressed in terms of generalized gradients. Ioffe [8] has also given first-order necessary conditions for certain nondifferentiable problems, stating them in terms of Levitin-Miljutin-Osmolovskii approximations. Much of the work of these authors has been carried out for problems in which R^n is replaced as the domain by a general Banach space. As noted above, we shall consider here only problems in R^n .

The classical second-order sufficiency conditions for problem P are of course given in terms of the gradients and Hessians of the functions g_i ; see, e.g., Hestenes [4, p. 37] or McCormick [14]. Extensions of these results to Banach spaces have been treated by a number of authors, including Borwein [1], Ioffe [9], Ioffe and Tikhomirov

* Received by the editors April 23, 1980.

† Department of Mathematics and Computer Science, Western Washington University, Bellingham, Washington 98225.

[10], and Maurer and Zowe [13]. In [9], Ioffe considers certain nondifferentiable problems of type P^* , where $F = g \circ G$, with g being sublinear and G being twice continuously differentiable.

The first sufficiency theorem for problem P in this paper is given in terms of auxiliary functions H and M . Thus, suppose that x^* belongs to $S \cap W$ and that $r > 0$ and $m^* \geq 0$. For x in W , we put

$$(2) \quad H(x) = \max \left\{ g_0(x) - g_0(x^*) + r \sum_{i=m+1}^q |g_i(x)|; g_i(x) \text{ for } 1 \leq i \leq m \right\}$$

and

$$(3) \quad M(x) = \max \left\{ g_0(x) - g_0(x^*) - (m^*/2)|x - x^*|^2 + r \sum_{i=m+1}^q |g_i(x)|; \right. \\ \left. g_i(x) \text{ for } i = 1, \dots, m \right\}.$$

Notice that the functions H and M depend upon x^* and r and that M also depends upon m^* . The function H is a slight variant of a function employed by Ioffe [9, Thms. 2, 7] through the use of his "reduction" theorem [7, Thm. 1].

The general second-order conditions given here are based on the functions H and M , and therefore make no direct reference to any specific Lagrange multiplier vector for problem P . As we shall see, the function h in Theorem 2.14 can be adapted to recover the classical theorem for a specific multiplier. And, we shall state one theorem directly in terms of a specific Lagrangian. But this theorem is awkward, partly because it still contains a reference to the function H . Nevertheless, it does extend one version of the classical result for problem P for twice continuously differentiable functions.

Finally, we must set some notation. If x and y belong to R^n and if $\delta > 0$, then $|x|$ denotes the Euclidean norm of x , $x \cdot y$ denotes the usual inner product of x and y , and we set $B(x, \delta) = \{z \in R^n: |x - z| \leq \delta\}$.

1. Generalized gradients. Let W be an open set in R^n . Suppose that f is a real-valued function defined on W . Then f is said to be locally Lipschitz on W in case each point x in W admits a neighborhood $V(x)$ and a number $K(x)$ such that $|f(z) - f(y)| \leq K(x)|z - y|$ whenever z and y belong to $V(x)$.

Throughout this paper we shall be dealing with locally Lipschitz functions on W .

DEFINITION 1.1 [3, p. 248]. Suppose that f is locally Lipschitz on W . According to Rademacher's theorem [20], f is differentiable a.e. on W . We denote by $\nabla f(x)$ the gradient of f at x (when it exists). Let E be the set of all points z in W for which f is differentiable at z . Now, suppose that x belongs to W . The generalized gradient of f at x , denoted by $\partial f(x)$, is the convex hull of the set of all limits of convergent sequences $\{\nabla f(x_k)\}$, where $\{x_k\}_{k=1}^\infty$ is a sequence in E convergent to x .

The generalized directional derivative of f at x in the direction d is defined by

$$f^0(x; d) = \limsup_{v \rightarrow 0; t \downarrow 0} \frac{f(x + v + td) - f(x + v)}{t}.$$

Remarks 1.2. We list here several facts about generalized gradients which we shall use in the sequel.

(a) [3, Prop. 1.4]. It is true that

$$f^0(x; d) = \max \{v \cdot d: v \in \partial f(x)\},$$

for all x in W and d in R^n ; in other words, $f^0(x; \cdot)$ is always the support function of the convex set $\partial f(x)$.

(b) [3, Cor. 1.9]. Given x in W , the function $f^0(x; \cdot)$ is convex on R^n .

(c) [2, Prop. 7]. The multifunction $x \rightarrow \partial f(x)$ is upper semicontinuous on W ; thus, if $\{x_k\}$ and $\{v_k\}$ converge respectively to x in W and v in R^n and if v_k is in $\partial f(x_k)$ for each k , then v belongs to $\partial f(x)$.

(d) Lebourg's mean value theorem (see [12] or [11, Thm. 1.7]). Suppose that x and y are in W . Suppose that the line segment L joining x and y lies in W . Then there exist z in L and v in $\partial f(z)$ such that $z \neq x$, $z \neq y$ and

$$f(x) - f(y) = v \cdot (x - y).$$

(e) [3, Cor. 1.10]. If for some v of R^n and for all d in R^n we have

$$v \cdot d \leq \limsup_{t \rightarrow 0} \frac{f(x + td) - f(x)}{t},$$

then v belongs to $\partial f(x)$.

DEFINITION 1.3. Let $\{x_k\}$ be a sequence in W which converges to x in W and suppose that $x_k \neq x$ for all k . Let d be a nonzero vector in R^n . Then $\{x_k\}$ converges to x in direction d in case it is true that $\{(x_k - x)/|x_k - x|\}$ converges to $d/|d|$.

DEFINITION 1.4. Let x be in W and let d be a nonzero vector in R^n . We define $\partial_d f(x)$ to be the set of all v in R^n for each of which there exist sequences $\{x_k\}$ in W and $\{v_k\}$ in R^n such that

- (a) $\{x_k\}$ converges to x in direction d ;
- (b) $\{v_k\}$ converges to v ;
- (c) v_k belongs to $\partial f(x_k)$ for each k .

(Observe that, in view of Remark 1.2(c), we have $\partial_d f(x) \subseteq \partial f(x)$. One may think of $\partial_d f(x)$ as the set of those generalized gradients at x which "arise" from the direction d .)

2. Sufficiency theorems. Let the functions $g_0, g_1, g_2, \dots, g_q$ be locally Lipschitz on W and form the set S as in (1). Next, we define

$$S_1 = \bigcap_{i=1}^m \{x \in W: g_i(x) \leq 0\}$$

and define a function f_0 on W by

$$f_0(x) = g_0(x) + \sum_{i=m+1}^q r |g_i(x)|, \quad x \in W.$$

THEOREM 2.1. Let x^* belong to the set $S \cap W$ and form the function M as in (3). Suppose that the following hypotheses hold:

- (a) We have $v \cdot d \geq 0$ whenever d is a nonzero vector in R^n and v is in $\partial_d M(x^*)$.
- (b) We have $\limsup v_k \cdot (x_k - x^*)/|x_k - x^*|^2 > 0$, provided the sequences $\{x_k\}$ and $\{v_k\}$ and the vector d in R^n satisfy the conditions
 - (i) $\{x_k\}$ converges to x^* in direction d with x_k in S_1 for each k ;
 - (ii) $\{v_k\}$ converges to 0 with v_k in $\partial M(x_k)$ for each k ;
 - (iii) there exists v_0 in $\partial_d f_0(x^*)$ such that $v_0 \cdot d \leq 0$.

Then there exists a positive number δ such that $g_0(x) \geq g_0(x^*) + (m^*/2)|x - x^*|^2$ whenever x belongs to $B(x^*, \delta) \cap S$.

Proof. Suppose that the conclusion is false. Choose a sequence $\{\delta_k\}$ of positive numbers decreasing to zero. Then, given k , there exists z_k in $B(x^*, \delta_k) \cap S$ such that

$g_0(z_k) - g_0(x^*) < (m^*/2)|z_k - x^*|^2$. Observe that $M(z_k) \leq 0 = M(x^*)$, and so M admits a minimizer x_k on $B(x^*, \delta_k)$ which is different from x^* . Let $t_k = |x_k - x^*|$ and define d_k by $d_k = (x_k - x^*)/t_k$. We may assume that $\{d_k\}$ converges to a unit vector d .

It follows from Clarke's theorem on first-order necessary conditions [2, Thm. 1] that there exists v_k in $\partial M(x_k)$ such that $-v_k$ belongs to the normal cone to the convex set $B(x^*, \delta_k)$ at x_k (see [17, p. 15]). But then we have $-v_k = c_k d_k$ for some $c_k \geq 0$, and so

$$(4) \quad v_k + c_k d_k = 0 \quad \text{for all } k.$$

We may assume that $\{v_k\}$ converges to v in $\partial_d M(x^*)$, in view of Remark 1.2(c) and Definition 1.4; and we may assume that $\{c_k\}$ converges to $c \geq 0$. We obtain $v + cd = 0$. Hence, $|v|^2 + c(d \cdot v) = 0$. By hypothesis (a), we have $d \cdot v \geq 0$, and so we infer $v = 0$.

By Lebourg's mean value theorem (Remark 1.2(d)), there exist z_{0k} and v_{0k} such that z_{0k} is in the "interior" of the line segment joining x_k and x^* , v_{0k} belongs to $\partial f_0(z_{0k})$ and $f_0(x_k) - f_0(x^*) = v_{0k} \cdot (x_k - x^*)$. As before, we may assume that $\{v_{0k}\}$ converges to v_0 in $\partial_d f_0(x^*)$. Since

$$\begin{aligned} f_0(x_k) - f_0(x^*) &\leq M(x_k) + \left(\frac{m^*}{2}\right)|x_k - x^*|^2 \\ &\leq M(x^*) + \left(\frac{m^*}{2}\right)|x_k - x^*|^2 = \left(\frac{m^*}{2}\right)|x_k - x^*|^2, \end{aligned}$$

we obtain $v_0 \cdot d \leq 0$.

Hence, it follows from hypothesis (b) that we may assume

$$(5) \quad \lim_{t_k} \frac{(v_k \cdot d_k)}{t_k} > 0;$$

here, we allow $+\infty$ as a possible value for the left side of (5). From (4), we get $v_k \cdot d_k + c_k |d_k|^2 = 0$ and so $v_k \cdot d_k \leq 0$ for all k . This last conclusion gives a contradiction of (5). The proof is complete.

We now obtain a sufficiency theorem for problem P^* as a corollary to Theorem 2.1.

COROLLARY 2.2. *Suppose that x^* belongs to W and that F is a real-valued locally Lipschitz function on W . Suppose*

- (a) $v \cdot d \geq 0$ whenever d is a nonzero vector in R^n and v is in $\partial_d F(x^*)$;
- (b) there exists $m^* \geq 0$ such that $\limsup v_k \cdot (x_k - x^*)/|x_k - x^*|^2 > m^*$ whenever $\{x_k\}$ is a sequence in W convergent to x^* for which $x_k \neq x^*$ for all k and $\{v_k\}$ is a sequence in R^n convergent to 0 with v_k in $\partial F(x_k)$ for all k .

Then there exists a positive number δ such that

$$F(x) - F(x^*) \geq \left(\frac{m^*}{2}\right)|x - x^*|^2 \quad \text{for all } x \text{ in } B(x^*, \delta).$$

Proof. Define a function F^* on W by $F^*(x) = F(x) - (m^*/2)|x - x^*|^2$ for x in W . We shall apply Theorem 2.1 to $M = F^*$. Observe that, by Definition 1.1, $\partial F^*(x) = \{v - m^*(x - x^*) : v \in \partial F(x)\}$, for x in W . Hence, it follows that $\partial F^*(x^*) = \partial F(x^*)$ and $\partial_d F^*(x^*) = \partial_d F(x^*)$ for each nonzero vector d in R^n . Furthermore, if x belongs to W and if $v^* = v - m^*(x - x^*)$ for some v in $\partial F(x)$, then we have

$$v^* \cdot (x - x^*) = v \cdot (x - x^*) - m^*|x - x^*|^2.$$

The corollary now follows from Theorem 2.1 and the observations just made.

Example 2.3. Define a function F on R^1 by setting $F(x) = \max \{x^2, 2x - x^2\}$ for all x in R^1 . It is clear that $x^* = 0$ minimizes $F(x)$ over all x in R^1 . Note that F is not convex on any neighborhood of 0 and that F fails to be differentiable at $x = 0$ and $x = 1$. We wish to show that $x^* = 0$ satisfies hypotheses (a) and (b) of Corollary 2.2.

For condition (a), note that if $d > 0$ then $v \cdot d = 2d > 0$ for all v in $\partial_d F(0)$, while if $d < 0$ then $v \cdot d = 0$ for all v in $\partial_d F(0)$. Next, suppose that $\{x_k\}$ converges to 0 with $x_k \neq 0$ for each k and that $\{v_k\}$ converges to 0 with v_k in $\partial F(x_k)$ for each k . Therefore, $x_k < 0$ for all large k and so $v_k \cdot x_k / |x_k|^2 = 2x_k \cdot x_k / |x_k|^2 = 2$ for all large k . Hence, hypothesis (b) of Corollary 2.2 is satisfied.

Remark 2.4. We wish now to examine hypothesis (a) of Theorem 2.1 in the presence of additional assumptions about the functions g_i . We shall assume that each g_i is both "semismooth" and "subdifferentiably regular" at x^* .

DEFINITION 2.5 [16, Def. 1]. Suppose that f is locally Lipschitz on W and suppose that x belongs to W . Then f is semismooth at x in case it is true that the sequence $\{v_k \cdot d\}$ has exactly one accumulation point whenever d is a nonzero vector in R^n and $\{\theta_k\}$ and $\{v_k\}$ are sequences in R^n , $\{t_k\}$ is a sequence in R^1 such that

- (i) $\{t_k\}$ decreases to 0;
- (ii) $\{\theta_k/t_k\}$ converges to 0 in R^n ;
- (iii) v_k is in $\partial f(x + t_k d + \theta_k)$ for each k .

Remark 2.6. Mifflin [16] has shown that, if f is semismooth at x , then, for each nonzero vector d in R^n , the directional derivative $f'(x; d)$ exists and equals $\lim v_k \cdot d$, where $\{v_k\}$ is any sequence chosen as in Definition 2.5 above. Mifflin [16] has also shown that convex functions are semismooth and that continuously differentiable functions are semismooth; moreover, so are certain marginal functions (i.e., functions which are pointwise maxima of certain continuously differentiable functions).

(Mifflin [15] has developed an algorithm for finding solutions to problems of type P, provided the functions are semismooth at the points of their domains.)

DEFINITION 2.7 [18]. Suppose that f is locally Lipschitz on W and suppose that x belongs to W . Then f is subdifferentiably regular at x in case the directional derivative $f'(x; d)$ exists for all d in R^n and $f^0(x; d) = f'(x; d)$ for all d .

(Actually, Rockafellar [18] gives a definition for a more general case. The definition is equivalent to the one just given when f is locally Lipschitz on W . Earlier, Clarke had used the term "regular" in place of "subdifferentiably regular" and Mifflin [16] has used the term "quasidifferentiable".)

Remark 2.8. Convex functions and continuously differentiable functions on W are subdifferentiably regular at all points of W , and Clarke's theorem on marginal functions [3, Thm. 2.1] shows that certain marginal functions on W are subdifferentiably regular at all points of W . Note that we always have $f^0(x; d) \geq f'(x; d)$ when the latter exists.

PROPOSITION 2.9. Let f and g be real-valued functions on W and let x belong to W . Let a be a number.

- (a) If f and g are semismooth at x , then so are $f + g$ and af .
- (b) If f and g are subdifferentiably regular at x , then so are $f + g$ and af (provided $a \geq 0$).

Proof. (a) The assertion for af is immediate from the fact that $\partial(af)(x) = a(\partial f(x))$ (see Definition 1.1). Suppose that f and g are semismooth at x and let $h = f + g$. Let d be a nonzero vector in R^n , and suppose that the sequences $\{v_k\}$, $\{\theta_k\}$ and $\{t_k\}$ are as in Definition 2.5 (i)–(iii), except with f replaced by h . According to [3, Prop. 1.12], $\partial h(y) \subseteq \partial f(y) + \partial g(y)$ for each y , and so there exist v_{1k} in $\partial f(x + t_k d + \theta_k)$ and v_{2k} in $\partial g(x + t_k d + \theta_k)$ such that $v_k = v_{1k} + v_{2k}$. By assumption, the sequences $\{v_{1k} \cdot d\}$ and $\{v_{2k} \cdot d\}$ both converge; hence so does $\{v_k \cdot d\}$.

(b) Suppose f and g are subdifferentially regular at x and let $h = f + g$. We always have $h' \leq h^0$ if the former limit exists; hence, by use of [3, Prop. 1.12] again, we have

$$\begin{aligned} h'(x; d) &\leq h^0(x; d) \leq f^0(x; d) + g^0(x; d) \\ &= f'(x; d) + g'(x; d) = h'(x; d). \end{aligned}$$

Next, if $a \geq 0$, we have $\partial(af)(x) = a(\partial f(x))$ and so $(af)' = (af)^0$ follows from Remark 1.2(a).

THEOREM 2.10. *Suppose that x^* is in $S \cap W$ and that H and M are defined as in (2) and (3). Suppose that the functions g_0, g_1, \dots, g_a are semismooth at x^* and that the functions $g_0, g_1, \dots, g_m, |g_{m+1}|, \dots, |g_a|$ are subdifferentially regular at x^* . Then hypothesis (a) of Theorem 2.1 holds if and only if $\partial H(x^*)$ contains the zero vector.*

Proof. It follows from Clarke's theorem on marginal functions [3, Thm. 2.1] that the maximum of a finite number of subdifferentially regular functions is subdifferentially regular. It follows from a theorem of Mifflin [16, Thm. 6] and [2, Prop. 9] that the maximum of a finite number of semismooth functions is again semismooth. Thus, since $|g_i| = \max\{g_i, -g_i\}$, it follows that $|g_i|$ is both subdifferentially regular and semismooth at x^* . It now follows from Proposition 2.9 that f_0 is both subdifferentially regular and semismooth at x^* . The remarks above show that H and M have the same properties. Furthermore, it follows from Clarke's theorem on marginal functions [3, Thm. 2.1] that $\partial H(x^*) = \partial M(x^*)$.

Now suppose that hypothesis (a) in Theorem 2.1 holds. By Remark 1.2(a) and Definition 1.4, we have $H^0(x^*; d) = M^0(x^*; d) \geq 0$ for all d in R^n , hence, by [17, Thm. 13.1], the zero vector belongs to $\partial H(x^*)$.

Conversely, suppose that 0 belongs to $\partial H(x^*)$. Since $\partial H(x^*) = \partial M(x^*)$, we have $M'(x^*; d) = M^0(x^*; d) = H^0(x^*; d) \geq 0$. It follows from Remark 2.6 that hypothesis (a) in Theorem 2.1 must hold in this case.

Remark. To guarantee that $|g_i|$ is subdifferentially regular at x^* , it is enough to assume that g_i is C^1 near x^* (since $|g_i| = \max\{g_i, -g_i\}$).

Example 2.11. Suppose now in Corollary 2.2 that the function F is twice continuously differentiable at x^* . We wish to show that hypotheses (a) and (b) of Corollary 2.2 reduce to the usual sufficiency conditions in this case.

First, note that, in view of Theorem 2.10, hypothesis (a) amounts to $\nabla F(x^*) = 0$; one must recall that the generalized gradient consists of the gradient alone in this situation. Now, let $\nabla^2 F(x)$ denote the Hessian of F at x . Hypothesis (b) becomes the statement that " $\limsup \nabla F(x_k) \cdot (x_k - x^*) / |x_k - x^*|^2 > m^*$ for every sequence $\{x_k\}$ convergent to x^* ." Since we have

$$\begin{aligned} \nabla F(x_k) \cdot (x_k - x^*) &= \{\nabla F(x_k) - \nabla F(x^*)\} \cdot (x_k - x^*) \\ &= (x_k - x^*) \cdot \nabla^2 F(x^*)(x_k - x^*) + o(|x_k - x^*|^2), \end{aligned}$$

we find that hypothesis (b) becomes the statement that " $\limsup (x_k - x^*) \cdot \nabla^2 F(x^*)(x_k - x^*) / |x_k - x^*|^2 > m^*$ for every sequence $\{x_k\}$ convergent to x^* (with $x_k \neq x^*$)." But this last condition is equivalent to the requirement that $d \cdot \nabla^2 F(x^*)d > m^*$ hold for all unit vectors d in R^n .

DEFINITION 2.12. Let x be in $S \cap W$. The tangent cone $T(S, x)$ of S at x is defined to be the set of all d in R^n for each of which there exist sequences $\{x_k\}$ in $S \cap W$ and $\{t_k\}$ of positive numbers such that $\{x_k\}$ converges to x and $\{(x_k - x)/t_k\}$ converges to d . (This notion of "tangent cone" is that used by Hestenes [4, p. 25].)

Remark 2.13. The theorem which follows applies to problem P, just as Theorem 2.1 does. It is of interest because the second-order conditions are given in terms of

an auxiliary function h rather than in terms of M . This can be an advantage when the structure of h is simpler than that of M . Several illustrations are given below.

THEOREM 2.14. *Suppose that x^* belongs to $S \cap W$ and that M and f_0 are defined as in Theorem 2.1. Suppose that h is a real-valued locally Lipschitz function defined on W for which*

$$(a) \quad h(x^*) = M(x^*);$$

$$(b) \quad h(x) \leq M(x) \text{ for all } x \text{ in } S \cap W;$$

(c) *we have $\limsup v_k \cdot (x_k - x^*)/|x_k - x^*|^2 > 0$ whenever $\{x_k\}$ is sequence in W which converges to x^* in direction d in $T(S, x^*)$ for which $v_0 \cdot d \leq 0$ for some v_0 in $\partial_d f_0(x^*)$ and $\{v_k\}$ is a sequence in R^n for which v_k is in $\partial h(x_k)$ for each k .*

Then there exists a positive number δ such that

$$g_0(x) - g_0(x^*) \geq \left(\frac{m^*}{2}\right) |x - x^*|^2 \quad \text{for all } x \text{ in } B(x^*, \delta) \cap S.$$

Proof. Suppose that the conclusion is false. Choose a sequence $\{\delta_k\}$ of positive numbers decreasing to 0. Then, given k , there exists z_k in $B(x^*, \delta_k) \cap S$ such that $g_0(z_k) - g_0(x^*) < (m^*/2)|z_k - x^*|^2$. Let $s_k = |z_k - x^*| > 0$ and let $d_k = (z_k - x^*)/s_k$ for each k . Now define C_k to be the smallest convex cone containing d_k . We may assume that the sequence $\{d_k\}$ converges to a unit vector d in $T(S, x^*)$. Notice also that $f_0(z_k) - f_0(x^*) < (m^*/2)|z_k - x^*|^2$ and so we may use Lebourg's theorem as in the proof of Theorem 2.1 to show that $v_0 \cdot d \leq 0$ for some v_0 in $\partial_d f_0(x^*)$.

By hypotheses (a) and (b), we have $h(z_k) \leq M(z_k) \leq 0 = M(x^*) = h(x^*)$ and so h attains its minimal value on the set $B(x^*, \delta_k) \cap (C_k + x^*)$ at a point x_k which is different from x^* . Let $t_k = |x_k - x^*| > 0$. Notice that $(x_k - x^*)/t_k = d_k$ for each k and so $\{x_k\}$ converges to x^* in direction d .

According to Clarke's theorem on necessary conditions [2, Thm. 1], there exists v_k in $\partial h(x_k)$ such that $-v_k$ is normal to the convex set $B(x^*, \delta_k) \cap (C_k + x^*)$ at x_k . We apply [17, Cor. 23.8.1] to obtain a nonnegative number c_k and a vector u_k normal to $C_k + x^*$ at x_k such that

$$(6) \quad v_k + c_k d_k + u_k = 0.$$

Since both $x_k + t_k d_k$ and $x_k - t_k d_k$ belong to $C_k + x^*$, we have $u_k \cdot d_k = 0$. Therefore we get from (6)

$$v_k \cdot d_k + c_k |d_k|^2 = 0.$$

From this we infer

$$(7) \quad \limsup v_k \cdot d_k / t_k \leq 0.$$

Since (7) contradicts hypothesis (c), the proof is complete.

Remark 2.15. Suppose x^* belongs to W . If F is any real-valued locally Lipschitz function on W , put

$$K(F) = \{d \in R^n : F^0(x^*; d) \leq 0\}.$$

It follows from Remarks 1.2(a) and 1.2(b) that $K(F)$ is a closed convex cone.

THEOREM 2.16. *Suppose x^* belongs to $S \cap W$ and that M and f_0 are as in Theorem 2.14. Suppose that the functions g_0, g_1, \dots, g_q are all semismooth at x^* and that the functions $g_0, |g_{m+1}|, \dots, |g_q|$ are all subdifferentially regular at x^* . Then it is true that $v_0 \cdot d \leq 0$ for some v_0 in $\partial_d f_0(x^*)$ if and only if d belongs to $K(f_0)$.*

Proof. We show that f_0 is semismooth and subdifferentially regular just as we did in the proof of Theorem 2.10. Thus, by Remark 2.6, we have $f_0^0(x^*; d) = f_0'(x^*; d) = v_0 \cdot d$, if v_0 belongs to $\partial_d f_0(x^*)$. The desired conclusion follows from this fact.

Example 2.17. Suppose that x^* belongs to $S \cap W$ and that the functions g_i are all twice continuously differentiable near x^* . We shall suppose that the standard second-order sufficiency conditions hold at x^* and we shall show that the hypotheses of Theorem 2.14 are then satisfied. Thus we assume that numbers y_1, y_2, \dots, y_q exist so that $y_i \geq 0$ and $y_i g_i(x^*) = 0$ for $i = 1 \dots m$, so that

$$(8) \quad \nabla g_0(x^*) + \sum_{i=1}^q y_i \nabla g_i(x^*) = 0,$$

and so that, if we set $L(x) = g_0(x) + y_1 g_1(x) + \dots + y_q g_q(x)$ for x in W , then we have $d \cdot \nabla^2 L(x^*) d > m^*$ whenever d is any unit vector in $T(S, x^*)$ for which $\nabla g_i(x^*) \cdot d = 0$ for all i in $I = \{i: 1 \leq i \leq m \text{ and } y_i > 0\}$.

To see that Theorem 2.14 applies in this situation, we must first choose $r > 0$. Let $r = 1$ if there are no nonzero numbers among $\{y_{m+1}, \dots, y_q\}$ and $r = \max \{|y_i|: i = m+1, \dots, q\}$ otherwise. Define h on W by $h(x) = L(x) - g_0(x^*) - (m^*/2)|x - x^*|^2$, $x \in W$. Then, conditions (a) and (b) of Theorem 2.14 hold, since $h(x^*) = 0$ and, for x in $S \cap W$,

$$\begin{aligned} h(x) &= L(x) - g_0(x^*) - \left(\frac{m^*}{2}\right)|x - x^*|^2 \\ &\leq g_0(x) - g_0(x^*) - \left(\frac{m^*}{2}\right)|x - x^*|^2 \leq M(x). \end{aligned}$$

Now suppose that $\{x_k\}$ is a sequence in W which converges to x^* in a direction d in $T(S, x^*)$, suppose that $v_0 \cdot d \leq 0$ for some v_0 in $\partial_d f_0(x^*)$, and suppose that v_k is in $\partial h(x_k)$ for each k . As in the proof of Theorem 2.10, we see that f_0 is both semismooth and subdifferentially regular at x^* . By Theorem 2.16, d belongs to $K(f_0)$. Therefore we have

$$0 \leq \lim_{t_k} \frac{\{f_0(x_k) - f_0(x^*)\}}{t_k} = \nabla g_0(x^*) \cdot d + \sum_{i=m+1}^q r |\nabla g_i(x^*) \cdot d|.$$

Since $\nabla L(x^*) = 0$ and since $\nabla g_i(x^*) \cdot d \leq 0$ for each i in I , we have

$$\begin{aligned} 0 &\leq \sum_{i=1}^m y_i \nabla g_i(x^*) \cdot d = -\nabla g_0(x^*) \cdot d - \sum_{i=m+1}^q y_i \nabla g_i(x^*) \cdot d \\ &\leq -\nabla g_0(x^*) \cdot d - \sum_{i=m+1}^q r |\nabla g_i(x^*) \cdot d| \leq 0. \end{aligned}$$

Since $\nabla g_i(x^*) \cdot d \leq 0$ and $y_i > 0$ for each i in I , we infer $\nabla g_i(x^*) \cdot d = 0$ for each i in I , and so we have $d \cdot \nabla^2 L(x^*) d > m^*$.

Now $v_k = \nabla L(x_k) - m^*(x_k - x^*)$ for each k . Set $t_k = |x_k - x^*|$ and $d_k = (x_k - x^*)/t_k$ for each k . We obtain

$$\begin{aligned} v_k \cdot d_k &= \nabla L(x_k) \cdot d_k - m^* t_k |d_k|^2 \\ &= \{\nabla L(x_k) - \nabla L(x^*)\} \cdot d_k - m^* t_k. \end{aligned}$$

Hence, $\limsup v_k \cdot d_k / t_k = d \cdot \nabla^2 L(x^*) d - m^* > 0$. Thus the hypotheses of Theorem 2.14 are fulfilled.

COROLLARY 2.18. Suppose that F is a real-valued locally Lipschitz function defined on W and that x^* belongs to W . Suppose that h is a real-valued locally Lipschitz

function defined on W for which

- (a) $h(x^*) = F(x^*)$;
- (b) $h(x) \leq F(x)$ for all x in W ;
- (c) there exists $m^* \geq 0$ such that $\limsup v_k \cdot (x_k - x^*) / |x_k - x^*|^2 > m^*$ whenever $\{x_k\}$ is a sequence in W which converges to x^* in a direction d for which $v_0 \cdot d \leq 0$ for some v_0 in $\partial_d F(x^*)$ and $\{v_k\}$ is a sequence in R^n for which v_k belongs to $\partial h(x_k)$ for each k .

Then there exists a positive number δ such that $F(x) - F(x^*) \geq (m^*/2)|x - x^*|^2$ for all x in $B(x^*, \delta)$.

Proof. We define functions h^* and F^* on W by $F^*(x) = F(x) - (m^*/2)|x - x^*|^2$ and $h^*(x) = h(x) - (m^*/2)|x - x^*|^2$ for x in W . We now proceed as in the proof of Corollary 2.2, applying Theorem 2.14 to F^* and h^* .

3. On Ioffe's sufficiency theorem.

Remark 3.1. Suppose that g is a sublinear function defined on R^m ; thus, g is convex, positively homogeneous, and subadditive on R^m . Suppose that the real-valued functions G_1, G_2, \dots, G_m are twice continuously differentiable on an open set $W \subseteq R^n$ containing the point x^* . Let G be defined by $G(x) = (G_1(x), G_2(x), \dots, G_m(x))$ for x in W . Let $f = g \circ G$ and consider the problem

$$P^{**}: \quad \text{Minimize } f(x) \text{ over } x \text{ in } W.$$

This is the finite-dimensional version of the problem over Banach spaces considered by Ioffe [9]. Since problem P^{**} is plainly a special case of problem P^* , it is natural to ask whether the sufficiency conditions given in Corollary 2.2 reduce to those of Ioffe [9, Thm. 2] when applied to problem P^{**} . A straightforward application of Corollary 2.2 to the function f produces a theorem which seems to be slightly weaker than that of Ioffe. But, we can get Ioffe's theorem from Corollary 2.18. Apparently, the latter approach takes proper advantage of the special structure of f .

Remark 3.2. The generalized gradient of f can be expressed in terms of the Jacobian J of G and the subdifferential ∂g of the convex function g . It follows from a theorem of Rockafellar [18] that f is subdifferentially regular at every point of W and that

$$(9) \quad \partial f(x) = \{J(x)^T y : y \in \partial g(G(x))\}, \quad x \in W;$$

here, the superscript T stands for transpose.

Mifflin [16, Thm. 5] has shown that the composite function of two semismooth functions is again semismooth. It follows from Remark 2.6 that f is semismooth at every point of W .

Several facts about g , which are consequences of its sublinearity (see [17, Cor. 13.2.2]), are listed below:

$$(10) \quad g(0) = 0.$$

$$(11) \quad \text{If } y \text{ belongs to } \partial g(u), \text{ then } g(u) = y \cdot u.$$

$$(12) \quad \text{For every } u \text{ in } R^m, \text{ we have } \partial g(u) \subseteq \partial g(0).$$

It follows from (10) and (12) that

$$(13) \quad \text{If } u \text{ belongs to } R^m \text{ and } y \text{ belongs to } \partial g(0), \text{ then } y \cdot u \leq g(u).$$

DEFINITION 3.3 [9]. Given f and x^* as above, put

$$\Omega(x^*) = \{y \in \partial g(G(x^*)) : J(x^*)^T y = 0\},$$

and

$$K_C = \{d \in R^n : g(G(x^*) + tJ(x^*)d) \leq g(G(x^*)) \text{ for some } t > 0\}.$$

Finally, we define a function L^* on $W \times R^m$ by

$$L^*(x, y) = y_1 G_1(x) + y_2 G_2(x) + \cdots + y_m G_m(x),$$

for x in W and y in R^m . (Ioffe [9] shows that K_C is a closed convex cone.)

Next, we state and prove Ioffe's sufficiency theorem.

THEOREM 3.4 [9, Thm. 2, finite-dimensional version]. *With f and x^* as above, suppose that the set $\Omega(x^*)$ is nonempty. Assume that positive numbers m_1 and m_2 exist so that*

(i) *given d in R^n , there exists z in K_C such that*

$$|d - z| \leq m_2 \{g(G(x^*) + J(x^*)d) - g(G(x^*))\},$$

and

(ii) *given d in K_C , there exists y in $\Omega(x^*)$ such that*

$$d \cdot \nabla_{xx}^2 L^*(x^*, y) d \geq m_1 |d|^2.$$

Then there exists a positive number δ such that

$$f(x) - f(x^*) \geq \left(\frac{m_1}{2}\right) |x - x^*|^2 \quad \text{for every } x \text{ in } B(x^*, \delta).$$

Proof. Since the set $\Omega(x^*)$ is nonempty, we can define h on W by

$$h(x) = \max \{y \cdot G(x) : y \in \Omega(x^*)\}, \quad x \in W.$$

We wish to apply Corollary 2.18. First, note that, since $\Omega(x^*)$ is nonempty, it follows from (9) that $\partial f(x^*)$ contains the zero vector. It follows from (11) and (13) that $h(x^*) = f(x^*)$ and $h(x) \leq f(x)$ for all x in W . By Clarke's theorem on marginal functions [3, Thm. 2.1], we have

$$\partial h(x) = \{J(x)^T y : y \in \Omega(x^*) \text{ and } h(x) = y \cdot G(x)\}.$$

(In particular, it is clear that $\partial h(x^*) = \{0\}$.)

In order to appeal to Corollary 2.18, it remains only to verify that hypothesis (c) of Corollary 2.18 holds. Thus, suppose that $\{x_k\}$ converges to x^* in direction d in $K(f)$ (here, we are using Theorem 2.16) and suppose that v_k belongs to $\partial h(x_k)$ for each k . For each k , there exists y_k^* in $\Omega(x^*)$ such that $v_k = J(x_k)^T y_k^*$ and $h(x_k) = y_k^* \cdot G(x_k)$. Since f is semismooth and subdifferentially regular at x^* , we have (with $t_k = |x_k - x^*|$ and $d_k = (x_k - x^*)/t_k$)

$$0 = f^0(x^*; d) \geq \limsup \{g(G(x_k)) - g(G(x^*))\}/t_k.$$

From Taylor's theorem, we get, since g is Lipschitz,

$$\lim_{t_k} \frac{\{g(G(x_k)) - g(G(x^*) + t_k J(x^*) d_k)\}}{t_k} = 0,$$

and so

$$0 \geq \limsup_{t_k} \frac{\{g(G(x^*) + t_k J(x^*) d_k) - g(G(x^*))\}}{t_k}.$$

According to hypothesis (i), there exists for each k a point z_k in K_C such that

$$|z_k - t_k d_k| \leq m_2 \{g(G(x^*) + t_k J(x^*) d_k) - g(G(x^*))\}.$$

Let $e_k = z_k/t_k$ for each k . It follows that the sequence $\{d_k - e_k\}$ converges to 0. We have

$$(14) \quad \begin{aligned} y_k^* \cdot G(x_k) - y_k^* \cdot G(x^*) &= y_k^* \cdot J(x^*) t_k d_k + \left(\frac{t_k^2}{2}\right) d_k \cdot \nabla^2 L^*(x^*, y_k^*) d_k + o(t_k^2) \\ &= \left(\frac{t_k^2}{2}\right) e_k \cdot \nabla^2 L^*(x^*, y_k^*) e_k + o(t_k^2). \end{aligned}$$

For each k , let y_k be a member of $\Omega(x^*)$ such that

$$e_k \cdot \nabla_{xx}^2 L^*(x^*, y_k) e_k \geq m_1 |e_k|^2.$$

Since $y_k \cdot G(x_k) \leq h(x_k) = y_k^* \cdot G(x_k)$ and $y_k \cdot G(x^*) = y_k^* \cdot G(x^*) = h(x^*)$, we have

$$(15) \quad \begin{aligned} y_k^* \cdot G(x_k) - y_k^* \cdot G(x^*) &\geq y_k \cdot G(x_k) - y_k \cdot G(x^*) \\ &= \left(\frac{t_k^2}{2}\right) e_k \cdot \nabla_{xx}^2 L^*(x^*, y_k) e_k + o(t_k^2) \\ &\geq \left(\frac{t_k^2}{2}\right) m_1 |e_k|^2 + o(t_k^2). \end{aligned}$$

Next, we observe that

$$\begin{aligned} v_k \cdot d_k &= J(x_k)^T y_k^* \cdot d_k \\ &= \{J(x_k)^T - J(x^*)^T\} y_k^* \cdot d_k \\ &= \nabla^2 L^*(x^*, y_k^*) t_k d_k \cdot d_k + o(t_k) \\ &= \nabla^2 L^*(x^*, y_k^*) t_k e_k \cdot e_k + o(t_k). \end{aligned}$$

From (14) and (15), we get

$$\frac{v_k \cdot d_k}{t_k} \geq m_1 |e_k|^2 + o(1),$$

and so

$$\limsup \frac{v_k \cdot d_k}{t_k} \geq m_1 |d|^2 = m_1.$$

4. A sufficiency theorem for Lagrangians. If the functions g_0, g_1, \dots, g_q in problem P are twice continuously differentiable at x^* , then the classical second-order sufficiency conditions for problem P (see [4, p. 37] or [14]) are given in terms of a Lagrangian. In view of [2, Prop. 9] and [3, Prop. 1.12], any generalized gradient of M does involve multipliers. But, there is no reference to any specific Lagrange multiplier.

We shall present a sufficiency theorem in terms of a particular set of multipliers. This effort is not wholly successful, because additional hypotheses are required on the g_i and because there is still a reference to H in the most general result.

Remark. In order to introduce a Lagrangian for problem P, we shall recall a special case of Clarke's theorem on necessary conditions [2, Thm. 1]; we have already used a different special case in the proofs of Theorems 2.1 and 2.14.

THEOREM 4.1. [Clarke, special case]. *Suppose that x^* in W is a local minimizer for problem P. Then there exist numbers a_i^* and vectors v_i^* for $i = 0, 1, \dots, q$ such that*

- (a) a_i^* is nonzero for at least one in $\{0, 1, \dots, q\}$;
 - (b) $a_i^* \geq 0$ for $i = 0, 1, \dots, m$;
 - (c) $a_i^* g_i(x^*) = 0$ for $i = 1, \dots, m$;
 - (d) v_i^* belongs to $\partial g_i(x^*)$ for all $i = 0, 1, \dots, q$;
 - (e) $0 = a_0^* v_0^* + a_1^* v_1^* + \dots + a_q^* v_q^*$.
- (We shall assume (a)–(e) below and we shall also assume
- (f) $a_0^* = 1$.

If, for example, there are no equality constraints and if the problem P is “calm” (see [2, p. 172]), then it is not a serious restriction to assume (f).)

Remark 4.2. Suppose that x^* belongs to W , that the functions g_i are all semismooth at x^* , that the functions $g_0, \dots, g_m, a_{m+1}^* g_{m+1}, \dots, a_q^* g_q$ are all subdifferentially regular at x^* , and that Theorem 4.1(a)–(f) hold. (We must emphasize that we do not assume that x^* is a local minimizer.)

Now we define a real-valued function L on W by

$$L(x) = \sum_{i=0}^q a_i^* g_i(x), \quad x \in W.$$

It follows from Proposition 2.9 that L is semismooth and subdifferentially regular at x^* . Rockafellar [18] has shown

$$(16) \quad \partial L(x^*) = \sum_{i=0}^q a_i^* \partial g_i(x^*).$$

It follows from Theorem 4.1(e) and (16) that 0 belongs to $\partial L(x^*)$. Proceeding as in Remark 2.15, we can introduce the closed convex cone

$$K = K(L) = \{d \in R^n : L^0(x^*; d) = 0\}.$$

THEOREM 4.3. *Suppose that x^* is in $S \cap W$, that the functions g_i are all semismooth at x^* , and that the functions $g_0, g_1, \dots, g_m, a_{m+1}^* g_{m+1}, \dots, a_q^* g_q$ are subdifferentially regular at x^* , where we are assuming that x^*, a^* , and $v_0^*, v_1^*, \dots, v_q^*$ satisfy Theorem 4.1(a)–(f). Define $I = \{i : 1 \leq i \leq m \text{ and } a_i^* > 0\}$. Next, suppose that $\limsup v_k \cdot (x_k - x^*)/|x_k - x^*|^2 > 0$ whenever the sequences $\{x_k\}$ and $\{v_k\}$ satisfy these conditions:*

- (i) $\{x_k\}$ converges to x^* in direction d with x_k in S for each k ;
- (ii) $\{v_k\}$ converges to 0 with v_k in $\partial L(x_k)$ for each k ;
- (iii) $g'_i(x^*; d) = 0$ for all i in I .

Then if C is a closed cone for which $C \subseteq \text{int}(K) \cup \{0\}$ there exists a positive number δ such that $g_0(x) > g_0(x^)$ whenever x belongs to $B(x^*, \delta) \cap (C + x^*) \cap S$ and $x \neq x^*$; here $\text{int}(K)$ denotes the interior of K .*

Proof. Suppose that the conclusion is false. Choose a sequence $\{\delta_k\}$ of positive numbers decreasing to 0. To each positive integer k , there corresponds x_k in $B(x^*, \delta_k) \cap S \cap (C + x^*)$ such that $x_k \neq x^*$ and

$$(17) \quad g_0(x_k) - g_0(x^*) \leq 0.$$

Let $t_k = |x_k - x^*| > 0$, and let $d_k = (x_k - x^*)/t_k$. Each d_k belongs to C and so we may assume that $\{d_k\}$ converges to a unit vector d in C . We apply Lebourg’s mean value theorem to obtain points z_k and z_{ki} in the interior of the line segment joining x_k to

x^* and vectors v_k in $\partial L(z_k)$ and v_{ki} in $\partial g_i(x_{ki})$ such that

$$(18) \quad L(x_k) - L(x^*) = v_k \cdot t_k d_k, \quad k \geq 1,$$

and

$$(19) \quad g_i(x_k) - g_i(x^*) = v_{ki} \cdot t_k d_k, \quad k \geq 1, \quad i = 0, 1, \dots, q.$$

As before, we may assume that $\{v_k\}$ converges to v^\sim in $\partial_d L(x^*)$, and that $\{v_{ki}\}$ converges to v_i^\sim in $\partial_d g_i(x^*)$ (for $i = 0, 1, \dots, q$).

Since each x_k belongs to S , we have, by Theorem 4.1(b), (c), $L(x_k) - L(x^*) \leq g_0(x_k) - g_0(x^*)$. It follows from (17), (18) and (19) that, for all k ,

$$(20) \quad \begin{aligned} v_k \cdot d_k &\leq v_{k0} \cdot d_k \leq 0, \\ v_{ki} \cdot d_k &\leq 0 \quad \text{for } i \text{ in } I, \\ v_{ki} \cdot d_k &= 0 \quad \text{for } m < i \leq q. \end{aligned}$$

From (20), we get $v^\sim \cdot d \leq 0$, $v_i^\sim \cdot d \leq 0$ for i in $I \cup \{0\}$, and $v_i^\sim \cdot d = 0$ for $m < i \leq q$. Since L is both semismooth and subdifferentiably regular at x^* , since v^\sim belongs to $\partial_d L(x^*)$ and 0 belongs to $\partial L(x^*)$, we have $v^\sim \cdot d = L'(x^*; d) = L^0(x^*; d) \geq 0$. We infer $v^\sim \cdot d = 0$.

Now we wish to show that $v^\sim = 0$. Since d is in C , d must be in $\text{int}(K)$. If v^\sim were nonzero, it would follow from Remarks 2.6 and 1.2(a) that, for positive t ,

$$L^0(x^*; d + tv^\sim) \geq v^\sim \cdot (d + tv^\sim) = t|v^\sim|^2 > 0.$$

It follows that $d + tv^\sim$ cannot belong to K (for $t > 0$). Hence, d cannot belong to $\text{int}(K)$, a contradiction. So, we have shown $v^\sim = 0$. In view of (20), we have

$$(21) \quad \limsup \frac{v_k \cdot d_k}{t_k} \leq 0.$$

Thus, if we can show that $g'_i(x^*; d) = 0$ for all i in I , then (21) will yield a contradiction.

Since $v^\sim = 0$, we have

$$0 = v^\sim \cdot d = \sum_{i=0}^q a_i^*(v_i^\sim \cdot d) = \sum_{i \in I} a_i^*(v_i^\sim \cdot d) + a_0^*(v_0^\sim \cdot d).$$

Inasmuch as $v_i^\sim \cdot d \leq 0$ for all i in $I \cup \{0\}$, we infer $v_i^\sim \cdot d = 0$ for all i in I . Because each g_i is semismooth and each v_i^\sim belongs to $\partial_d g_i(x^*)$, we get $0 = v_i^\sim \cdot d = g'_i(x^*; d)$ for all i in I .

Remark 4.4. Suppose that, in Theorem 4.3, the functions g_i are all continuously differentiable at x^* . Then Theorem 4.1(e) becomes $\nabla L(x^*) = 0$ and we have $K = \mathbb{R}^n$. In this case, we may choose $C = K = \mathbb{R}^n$ and thus obtain a sufficiency theorem.

If the functions g_i are all twice continuously differentiable near x^* , we can proceed as in Example 2.11 and recover from Theorem 4.3 one version of the classical second-order sufficiency conditions.

Remark 4.5. In some sense, Theorem 4.3 gives sufficiency conditions for certain directions (i.e., those directions which lie in C). A somewhat similar *first-order* theorem can be proved, where we allow directions d in a closed cone D ; here, D is required to be included in the union of the complement of K and the set $\{0\}$. This result can be regarded as a variant of [13, Thm. 5.3] or of [5, Chapt. 4, Thm. 6.3]. Using this first-order result and Theorem 4.3, we can employ the method of proof of Theorem 2.1 to obtain the following theorem:

THEOREM 4.6. *Suppose that the hypotheses of Theorem 4.3 are in force, except for the one concerning the cone C . Let H be as in (2) and let $b(K(L))$ be the boundary of the cone $K(L)$. Suppose also that*

$$\limsup \frac{v_k \cdot (x_k - x^*)}{|x_k - x^*|^2} > 0$$

whenever the sequences $\{x_k\}$ and $\{v_k\}$ satisfy not only conditions (i)–(iii) of Theorem 2.1, but also, the additional condition that d belong to the set $b(K(L))$.

Then there exists a positive number δ so that $g_0(x) > g_0(x^)$ whenever x belongs to $B(x^*, \delta) \cap S$ and $x \neq x^*$.*

We shall omit the proof of this theorem, because the theorem itself is unappealingly awkward and because we shall make no application of it.

REFERENCES

- [1] J. M. BORWEIN, *Optimization with respect to partial orderings*, Ph.D. Thesis, Jesus College, Oxford University, 1974.
- [2] F. H. CLARKE, *A new approach to Lagrange multipliers*, Math. Operations Res., 1 (1976), pp. 165–174.
- [3] ———, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.
- [4] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.
- [5] ———, *Optimization Theory: The Finite Dimensional Case*, Wiley-Interscience, New York, 1975.
- [6] J. B. HIRIART-URRUTY, *On optimality conditions in nondifferentiable programming*, Math. Prog., 14 (1978), pp. 73–86.
- [7] A. D. IOFFE, *Necessary and sufficient conditions for a local minimum. 1: A reduction theorem and first-order conditions*, this Journal, 17 (1979), pp. 245–250.
- [8] ———, *Necessary and sufficient conditions for a local minimum. 2: Conditions of the Levitin–Miljutin–Osmolovskii type*, this Journal, 17 (1979), pp. 251–265.
- [9] ———, *Necessary and sufficient conditions for a local minimum. 3: Second-order conditions and augmented duality*, this Journal, 17 (1979), pp. 266–288.
- [10] A. D. IOFFE AND W. M. TIKHOMIROV, *Theory of Extremal Problems*, North-Holland, Amsterdam, 1978.
- [11] G. LEBOURG, *Generic differentiability of Lipschitzian functions*, Trans. Amer. Math. Soc., 256 (1979), pp. 125–144.
- [12] ———, *Valeur moyenne pour gradient généralisé*, C.R. Acad. Sci. Paris Ser. A, 281 (1975), pp. 795–797.
- [13] H. MAURER AND J. ZOWE, *First- and second-order necessary and sufficient optimality conditions for infinite-dimensional programming problems*, Math. Prog., 16 (1979), pp. 98–110.
- [14] G. P. MCCORMICK, *Second order conditions for constrained minima*, SIAM J. Appl. Math., 15 (1967), pp. 641–652.
- [15] R. MIFFLIN, *An algorithm for constrained optimization with semismooth functions*, Math. Operations Res., 2 (1977), pp. 191–206.
- [16] ———, *Semismooth and semiconvex functions in constrained optimization*, this Journal, 15 (1977), pp. 959–972.
- [17] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [18] ———, *Directionally Lipschitzian functions and subdifferential calculus*, Proc. London Math. Soc., 39 (1979), pp. 331–355.
- [19] ———, *The theory of subgradients and its applications to problems of optimization*, Lecture Notes, University of Montreal, 1978.
- [20] E. M. STEIN, *Singular Integrals and Differentiability Properties of Functions*, Princeton Math. Ser., no. 30, Princeton University Press, Princeton, NJ, 1970.

A LEARNING MODEL FOR ROUTING IN TELEPHONE NETWORKS*

P. R. SRIKANTAKUMAR† AND K. S. NARENDRA†

Abstract. The aim of this paper is to develop a theory of adaptive routing in telephone networks using learning methods. A mathematical model of the network with slow-learning algorithms distributed at various nodes is presented. The algorithms update the routing probabilities on the basis of network feedback information (like call blocking or completion) only. Convergence of the routing strategies is established. Two linear updating algorithms, under certain conditions, are shown to have desirable equilibrium behavior like load equalization and minimum blocking probability for the entire network.

1. Introduction. The adaptive routing of traffic in large networks is currently a problem of great interest. While approaches to this problem through mathematical programming [12]–[14] exist, the one that has recently gained attention uses learning techniques. The suitability of learning methods has been demonstrated in the extensive simulation studies [1] of telephone traffic routing. The principal aim of this paper is to develop a mathematical model to explain the simulation results, and to provide a theory of network routing using learning methods.

A telephone network is a finite set of nodes connected by links or trunk groups of finite capacity. Calls (traffic) originate at the nodes of the network destined for other nodes. The routing control systems located at various nodes direct the arriving calls to subsequent nodes by attempting the call on alternate trunk groups at that node. If the call reaches the destination, it is connected (or set up) for communication, occupying a trunk on each link along the path it is routed; otherwise, the call is lost (or blocked) at some intermediate node failing to find a free trunk. Many paths for the connection of the call being generally available, the routing problem is to determine the decisions (e.g., routing probabilities for the various alternatives) that the nodal controllers have to make so that some network performance criterion like blocking probability is minimized.

The routing problem when the traffic and the network conditions are stationary and known has been well studied [9], [16]. Seldom do stationary situations prevail. Routing in the face of nonstationarities is the adaptive routing problem. In the approach to adaptive routing via learning methods, the traffic parameters are assumed to be fixed but unknown. The nodal controllers select the available alternatives to route a call with certain routing probabilities, and based upon only feedback (e.g., call block or completion) from the network update these probabilities. Thus, the controllers evolve to or “learn” some desired routing parameter. Being iterative, such routing algorithms provide adaptation to varying traffic and network conditions. (Any a priori knowledge of the traffic can be used to compute the initial routing parameters).

Central to the adaptive routing problem are questions concerning the adequacy of traffic models, the decision space of the routing controllers, the learning algorithms used by each controller and the network feedback information needed for updating. Questions related to convergence, speed, and adequacy of adaptation of the algorithms, and the overall performance of the network system are also important. While the investigation of the adaptive routing problem, in all its generality, is in its initial stages,

* Received by the editors November 16, 1979, and in revised form October 1, 1980. This work was supported by the National Science Foundation under grant 03664.

† Department of Engineering and Applied Science, Yale University, New Haven, Connecticut 06520.

an attempt is made in this paper to characterize and study some of its main aspects. Specifically, we formulate a mathematical model of the network using a class of learning algorithms, and study the convergence and asymptotic performance of the network and the nodal controllers. Conditions under which the algorithms employed are optimal are also discussed.

The learning algorithms for the nodal controllers have the following general form:

$$(1.1) \quad p(n+1) = p(n) + aU(p(n), \alpha(n), x(n)), \quad 0 < a < 1, \quad n = 0, 1, 2, \dots,$$

$p(n)$ is the vector of probabilities for the choice of various alternatives at n , $U(\cdot)$ is the updating function which depends upon the network state $x(n)$.

Algorithms of the form (1.1) have been studied in the context of learning models [3], [4]. In a typical learning model, an automaton (or algorithm) has a set of actions to choose from in an unknown stationary random environment, and at each state n it does so with certain action probabilities $p(n)$. For each action chosen the environment gives an output $x(n)$ (e.g., reward or penalty) from a stationary probability distribution. The automaton, then, updates the action probabilities depending upon the output at that stage, and evolves to some desired final behavior.

In the adaptive routing problem $\{x(n)\}_{n \geq 0}$ is nonstationary. In a recent paper by Narendra and Thathachar [2] an attempt was made to model this behavior. In this model when an action is performed its penalty probability increases by a fixed amount while the penalty probabilities of all others are decreased by fixed amounts. However, as indicated in [2] this model of a single automaton-environment combination could not be adequately analyzed using techniques normally used in the study of learning automata. To circumvent this difficulty an alternative approach was suggested by the authors in [15], wherein the penalty probabilities were assumed as functions of action probabilities. It is this feature which enables the resulting model to be extended to the adaptive routing problem. Extensions of some results of [15] are also in [2].

Norman [4] has studied "slow-learning" in a Markovian framework in terms of the asymptotic properties ($a \rightarrow 0, n \rightarrow \infty$) of the sequence $\{p(n)\}_{n \geq 0}$, where $\{x(n)\}_{n \geq 0}$ is a stationary random sequence. In view of the nonstationarity of $x(n)$ in the adaptive routing problem, the powerful methods of Norman cannot be directly applied. However, the specific assumption of the model suggested in [15] that $x(n)$ depends only on the probability iterates $p(n)$ enables slow-learning results to be extended, and incorporated into the routing problem.

In § 2 the network model comprising the traffic assumptions and the nodal routing schemes is developed. Two particular linear algorithms for $U(\cdot)$ are employed and are shown to result in desirable properties. We isolate, in § 3, one nodal algorithm to gain insight into its behavior in the network. An abstract nonstationary environment is postulated and the algorithms are analyzed in this environment. Section 4 contains some preliminaries about the network. In § 5.1, we prove the convergence in distribution of the network and the controllers by showing that the process $\{x(n), p(n)\}_{n \geq 0}$ on the product space has some random contraction properties. The steady state analysis is given in § 5.2, where results of § 3 are found to be useful. A quadratic approximation of the network blocking probability is derived in § 5.3 and it is shown that one of the linear algorithms, in steady state, is "near" the optimal for this approximation. The conclusions are stated in § 6.

2. Model of telephone network with learning routing.

Telephone network model. The telephone traffic system, for our purpose, has three constituents: (i) a network connecting source-destination pairs, (ii) a routing system, and (iii) a process for generating calls.

The network consists of a finite set of nodes \mathcal{N} : $\{1, 2, \dots, N\}$, and a set $\mathcal{L} \subseteq \mathcal{N} \times \mathcal{N}$ of trunk groups (or links). The link (i, j) has capacity l_{ij} trunks. Calls are generated randomly at source node i , destined for node j ($i, j \in \mathcal{N}$). Calls have a random holding time (or duration of communication of the call). The routing systems are distributed at various nodes. A call, in general, is processed through a sequence of nodal routing systems before it either reaches destination (completion) or gets blocked (loss); the time for this processing is the setup time. If a call is completed, it remains in the network for the duration of the holding time, after which the call is disconnected (or hung up) making free all the trunks it occupied.

A call with source i and destination j , when it arrives at node k is routed by the controller A_{ij}^k located at k . (If the calls are not distinguished with respect to their source¹, the controller is designated A_j^k). A_{ij}^k chooses a set of actions (alternatives) attempting to route the call to the subsequent node. The decision space determines the type of routing. For example, in the so called “fixed rule” routing [1] the call is attempted on a sequence of links $(k, l_1), (k, l_2), \dots, (k, l_r)$, i.e., only if all the links (k, l_i) , $i = 1, \dots, \alpha$ have no free trunk, the call is attempted (or overflows) on $(k, l_{\alpha+1})$, $\alpha = 1, \dots, r-1$. As is clear, the fixed rule has only a single choice and hence no scope for adaptation. In practice, the decision may take many forms, e.g., choosing from a set of links or from a set of sequences of links, or if all the links are busy placing the call in a waiting line. We restrict ourselves to the first two cases, i.e., either choosing from a set of links, or a set of sequence of links, at k . The alternatives are chosen with constant probabilities in fixed probabilistic routing; in adaptive routing these probabilities vary with time.

The following assumptions are made about the nature of the generated calls and their routing:

- (i) Calls from node i to node j follow an independent Poisson process with parameter λ_{ij} (i.e., λ_{ij} calls/unit time).
- (ii) The holding time of any call is an independent random variable with exponential distribution having mean time $1/\mu$. For convenience², we take $\mu = 1$.
- (iii) The setup time of each call is negligible.
- (iv) No call is deliberately blocked³.
- (v) The alternatives for A_{ij}^k are so restricted that routing is cycle-free (i.e., the call does not get trapped in a loop of nodes without either getting blocked or connected).

(Assumptions (i) and (ii) are quite standard in traffic theory.) A call is of type $t = (i, j) \in \mathcal{L}$, if it has origin i and destination j . We write A_t^k for A_{ij}^k (or A_j^k , if no source distinction).

Learning routing schemes. Controllers A_t^k are learning algorithms (2.1). A_t^k has r_{ik} choices. At instant n (see § 4), if a call arrives for A_t^k , the actions are chosen with probabilities $(p_t^{k,1}(n), p_t^{k,2}(n), \dots, p_t^{k,r_{ik}}(n)) = p_t^k(n)$, $\sum_{i=1}^{r_{ik}} p_t^{k,i}(n) = 1$. The call is either sent to the next node or blocked, and this depends upon the state of the network $x(n)$, at n , and the action chosen. The “state” of the network at n is the configuration of calls which are connected at n . The state gives the pattern of free and occupied trunks in the network (§ 4).

The general structure of updating for A_t^k is as follows:

$$(2.1) \quad p_t^k(n+1) = p_t^k(n) + aU(p_t^k(n), \alpha(n), x(n)),$$

¹ As in the AT & T network [1].

² All our results hold for the case of different μ_{ij} 's. For relaxation of (i) & (ii) see § 4.

³ i.e., making no attempt to send it to the next node. Such a policy has been studied for fixed routing in [9].

$0 < a < 1$. $\alpha(n)$ is the action at n . “ a ” is the step size. More generally, the updating could have the form

$$(2.2) \quad p_i^k(n+1) = p_i^k(n) + aU(p_i^k(n), z(n)),$$

where $z(n) \triangleq (\alpha(m), x(m))$, $0 \leq m \leq n$, is the history of A_i^k . For any call, A_i^k updates iff the call is of type t and is routed through node k .

Since $x(n)$ describes the state of the entire network, in large networks it is impractical to provide A_i^k with complete state information. The question of how much information is needed for optimal decision making, or the trade-off between state information and optimality, in general, appears to be difficult. In the linear updating algorithms (2.3) we employ, the updating depends only on the state of the path along which it is routed. Specifically, updating is only a function of the call blocking or completion, which depends only upon whether the call counters a blocking state or not. Such a simple overall routing scheme, wherein the nodal algorithms act independently without transferring any state information (except block/completion), has desirable and optimal (with respect to network blocking probability, defined in § 4) asymptotic behavior. While the techniques of the following sections are applicable to any algorithm of the form (2.1), attention is focussed on the linear algorithms (2.3). Algorithms which use average rather than instantaneous blockings and completions, and which have the form (2.2) are discussed towards the end of the paper.

Linear updating rule. At instant n let a call of type t arrive for A_i^k which has alternatives $\{\alpha_1, \alpha_2, \dots, \alpha_{r_{tk}}\}$. If the action chosen $\alpha(n)$ for routing this call is α_i the updating is given by the following.

Define $\Delta p_i^k(n) \triangleq p_i^k(n+1) - p_i^k(n)$. Let $\hat{x}(n)$ be the indicator of blocking, i.e., $\hat{x}(n) = 1$ if call is blocked. $0 < a, b < 1$.

$$(2.3) \quad \begin{aligned} j \neq i, \quad \Delta p_i^{k,j}(n) &= \begin{cases} -ap_i^{k,j}(n), & \hat{x}(n) = 0, \\ \frac{b}{r_{tk} - 1} - bp_i^{k,j}(n), & \hat{x}(n) = 1, \end{cases} \\ \Delta p_i^{k,i}(n) &= \begin{cases} a(1 - p_i^{k,i}(n)), & \hat{x}(n) = 0, \\ -bp_i^{k,i}(n), & \hat{x}(n) = 1. \end{cases} \end{aligned}$$

The probabilities are conserved at every instant. The following two cases will be the focus of study:

- (i) $a = b$, the symmetric case, named L_{R-P} algorithm [5].
- (ii)* $b = o(a)$, named $L_{R-\epsilon P}$ algorithm (also [6], [2])⁴.

The analysis of (2.3) follows in the subsequent sections. As a first step, we isolate one nodal algorithm for study of its probabilistic behavior. In § 3, the nonstationary environment abstracted from network considerations results in a novel automaton model. The asymptotic properties of the action probabilities in this model are analyzed. L_{R-P} attempts to equalize average penalty rates from its action, whereas $L_{R-\epsilon P}$ attempts to equalize average penalty probabilities from its actions [2]. The accuracy of this equalization is $O(a)$. As $a \rightarrow 0$, in steady state, action probability is Gaussian. Approximations to the asymptotic variance are derived. The analysis of the algorithm in the network is dealt with in § 4 and § 5. After showing the convergence of the algorithms and the network to a unique stationary measure, it is proved that the steady state behaviors of L_{R-P} and $L_{R-\epsilon P}$ resemble their behaviors in the automaton models of § 3.

⁴ $o(a)$ implies $o(a)/a \rightarrow 0$ as $a \rightarrow 0$. $O(a)$ is $\leq Ka$, where K is constant.

(Transient behaviors in the two situations are different.) L_{R-P} , thus, has a “load equalizing” effect, a desired attribute in traffic engineering. $L_{R-\varepsilon P}$, for small “ ε ”, is near optimal of a quadratic approximation of the network blocking probability.

3. New learning automata models. A learning automaton may be represented by a feedback loop containing an automaton and an unknown random environment. At stage n , the automaton (algorithm) chooses action $\alpha(n) = \alpha_i$ from a finite set $\{\alpha_1, \alpha_2, \dots, \alpha_r\}$ with probability $p_i(n)$. $\sum_{i=1}^r p_i(n) = 1$. In response to this action, the environment gives an output $x(n)$: reward (0) with probability $c_i(n)$, and penalty (1) with probability $d_i(n) \triangleq 1 - c_i(n)$.

$$c_i(n) = \Pr[x(n)/\alpha(n) = \alpha_i], \quad i = 1, \dots, r.$$

The automaton, then, updates its action probability according to the rule (2.1).

Norman [4] and others [3] have studied the behavior of different algorithms, but mostly in stationary environments, i.e., $c_i(n) = c_i$, a constant. Below, we postulate a novel nonstationary environment derived from network considerations, and analyze the updating scheme of (2.3) in this environment. In this environment $c_i(n)$ $i = 1, \dots, r$, are specific functions of $p(n) = (p_1(n), \dots, p_r(n))$. $\{p(n)\}_{n \geq 0}$ arising from the automaton-environment interaction is a homogeneous Markov process, and is ergodic. The steady state behavior, as will be shown in § 5, is similar to the behavior of the routing algorithm (2.3) in the network.

Environment. If $p(n) = p$, we write $c_i(p)$ for $c_i(n)$. The following are the assumptions on the environment.

A1. $c_i(p_1, p_2, \dots, p_r)$ is continuous in p_i , $i, j = 1, \dots, r$.

A2.

$$(3.1) \quad \frac{\partial c_i(\cdot)}{\partial p_i} > 0 \quad \forall i \quad \text{and} \quad \frac{\partial c_j(\cdot)}{\partial p_i} \ll \frac{\partial c_i(\cdot)}{\partial p_i} \quad \text{for } j \neq i.$$

A3. $c_i(\cdot)$ is continuously differentiable in all its arguments.

A4. $c_i(\cdot)$ and $\partial c_i(\cdot)/\partial p_i$ are Lipschitz functions of all their arguments.

While some of the assumptions are subsumed by others (e.g., A1 by A2, A3 and A4) we state them separately since not all the results stated later require the stricter assumptions.

Remarks. As will be seen in § 5, in the network $c_i(\cdot)$ corresponds to the blocking probability from action α_i . The blocking probability will, in general, satisfy the smoothness conditions (A1), (A3) and (A4). (see § 4). Assumption (A2) is motivated by the network situation: at a node, the amount of traffic attempted on different links is governed by the routing probability p , and increasing the traffic on a link increases the blocking probability along that link. The cross-derivatives being small in (A2) implies that the overflow (from one link to another at the same node, § 2) traffic is small. (In § 5.3, the cross-dependence results from merging traffic downchain, and could be large).

The probability with which the automaton receives a penalty at time n from action α_i is

$$p_i(n)c_i(p(n)) \triangleq f_i(p(n)).$$

We refer to $f_i(p)$ as the average penalty rate. (Note that $f_i(p)$ is the average rate of penalty seen by an outside observer, whereas $c_i(p)$ is the average rate of penalty given the action chosen is α_i .)

Algorithm.

- (3.2) The algorithm for updating is the same as in (2.3) without the indices t and k ; i.e., $p_i^{k,i} \equiv p_i$, $r_{tk} \equiv r$, \hat{x} of (2.3) is the binary output x of the environment.

Analysis. In this section we present the convergence and steady state analysis of both L_{R-P} and $L_{R-\varepsilon P}$ algorithms with r -actions in the environment (3.1).

THEOREM 1. *The Markov process $\{p(n)\}_{n \geq 0}$ is ergodic and converges in distribution as $n \rightarrow \infty$ to a unique stationary probability, for any distribution of $p(0)$. \square*

The proof of Theorem 1 follows from Norman's results [4] for Markov distance diminishing operators, and is in [15]. The proof is omitted here since it is not central to the theme of the paper.

Steady-state analysis. We will first consider the two-action case, which is illustrative of the type of results for r -action L_{R-P} and $L_{R-\varepsilon P}$ considered later.

Two-action L_{R-P} . In the analysis, the constant vector $p^* = (p_1^*, p_2^*)$, $p_1^* + p_2^* = 1$ satisfying $f_1(p^*) = f_2(p^*)$ is of importance. Assumptions (A1) and (A2) assure the existence and uniqueness respectively of p^* . The two penalty rates from the actions are equal at p^* .

PROPOSITION 1. *If (A1) is satisfied by $c_i(p)$, $i = 1, 2$, then there exists p^* such that*

$$f_1(p^*) = p_1^* c_1(p_1^*) = p_2^* c_2(p_2^*) = f_2(p^*).$$

Further, if (A2) holds, then p^ is unique.*

Proof.

$$f_2(p) - f_1(p) = \begin{cases} c_2(0, 1) & \text{if } p_1 = 0, \\ -c_1(1, 0) & \text{if } p_2 = 0 \text{ or } p_1 = 1. \end{cases}$$

$f_2(p) - f_1(p)$ is continuous in p_1 , and hence p_1^* and p_2^* exist such that $f_2(p^*) - f_1(p^*) = 0$. Uniqueness follows since the derivative with respect to p_1 is strictly decreasing. \square

Let $\Delta p_i(n) \triangleq p_i(n+1) - p_i(n)$. The conditional expectation of $\Delta p_1(n)$ in the algorithm (3.2) may be expressed as

$$(3.3) \quad E[\Delta p_1(n)/p(n) = p] = p_1 d_1(p)[ap_2] + p_2 d_2(p)[-ap_1] \\ - p_1 c_1(p)[bp_1] + p_2 c_2(p)[bp_2],$$

$d_i = 1 - c_i$. The four terms on the right-hand side of (3.3) are due to the four possible events at n .

Since, in the L_{R-P} algorithm, $a = b$, (3.3) becomes

$$(3.4) \quad E\left[\frac{\Delta p_1(n)}{a} \middle/ p(n) = p\right] = f_2(p) - f_1(p) \triangleq w(p)$$

and

$$E\left[\frac{\Delta p_2(n)}{a} \middle/ p(n) = p\right] = -w(p).$$

Let

$$(3.5a) \quad s_1(p) \triangleq E\left[\frac{\Delta p_1^2(n)}{a} \middle/ p(n) = p\right] = p_1 d_1(p) p_2^2 + p_1 c_1(p) p_1^2 + p_2 c_2(p) p_2^2 + p_2 d_2(p) p_1^2$$

and

$$(3.5b) \quad \tilde{s}_1(p) \triangleq s_1(p) - w^2(p).$$

Similarly, $s_2(p)$ and $\tilde{s}_2(p)$ are defined.

In the steady state

$$(3.6) \quad E[\Delta \bar{p}_i] = 0 \quad \text{or} \quad E[w(\bar{p})] = 0.$$

The zero of $w(p)$ is p^* and, in general, $E[w(\bar{p})] = 0$ need not yield p^* . However, if the parameter a is chosen sufficiently small, the difference between these two values may be made small, as indicated in the following theorem.

THEOREM 2. *Let $p(0)$ have stationary measure \bar{p} .*

(a) *Under assumptions (A1)–(A3)*

$$(3.7) \quad E[p_i(n) - p_i^*]^2 = O(a).$$

(b) *Under assumptions (A1)–(A4) if z_n^i is defined as*

$$z_n^i \triangleq \frac{p_i(n) - p_i^*}{\sqrt{a}} \quad \text{for } i = 1, 2$$

as $a \rightarrow 0$, z_n^i is normally distributed with mean zero and variance

$$\sigma_i^2 = \frac{\hat{s}_i(p^*)}{2 \left| \frac{\partial w(p^*)}{\partial p_i} \right|}.$$

Proof. See Appendix 2.

Comment. Since $[E(p_i(n) - p_i^*)]^2 \leq E[p_i(n) - p_i^*]^2$, part (a) of Theorem 2 implies that

$$(3.8) \quad E[p_i(n)] - p_i^* = O(\sqrt{a})$$

and further as $a \rightarrow 0$, $E[p_i(n) - p_i^*]^2 \rightarrow 0$ with $p_i(n) \rightarrow p_i^*$ with probability 1.

From (3.4) and (3.6) we have $E[w(p(n))] = E[f_2(p(n)) - f_1(p(n))] \rightarrow 0$ as $n \rightarrow \infty$, or

$$(3.9) \quad E[f_2(\bar{p})] = E[f_1(\bar{p})].$$

From this,

$$(3.10) \quad E[f_2(\bar{p})] - f_2(p^*) = E[f_1(\bar{p})] - f_1(p^*).$$

Since $f_i(\cdot)$ is Lipschitz bounded, from (3.10) and (3.8) we get

$$(3.11) \quad E[f_i(\bar{p})] - f_i(p^*) = O(\sqrt{a}).$$

Hence, for small values of “ a ”, it can be concluded from (3.9) and (3.11), that the asymptotic behavior of the L_{R-P} automaton can be approximated by

$$f_2(p^*) = f_1(p^*).$$

In other words, the two-action automaton asymptotically attempts to equalize the average penalty rates from the two actions.

Two-action $L_{R-\epsilon P}$ algorithm. In this case, $b = o(a)$. Equation (3.3) can now be written as

$$(3.12) \quad E\left[\frac{\Delta p_1(n)}{a} \middle| p(n) = p\right] = p_1 p_2 [c_2(p) - c_1(p)] + o(a) \triangleq w(p) + o(a).$$

If $c_i(p) = 0$ for $p_i = 0$, along with the assumption (A1) and (A2) we have a unique p^* such that $c_1(p^*) = c_2(p^*)$. (The proof is along the lines of Proposition 1). Otherwise, $w(p) > 0$ (or < 0) for all $p_1 \in (0, 1)$. If so, let $p_1^* = 1$ (or $p_1^* = 0$). With $s_i(p)$ and $\hat{s}_i(p)$ as

defined in (3.5a) and (3.5b), the statement of Theorem 2 holds for the $L_{R-\varepsilon P}$. Hence the two-action $L_{R-\varepsilon P}$ attempts to equalize the penalty probabilities from the two actions [2].

Multi-action case. We can write, for algorithm (3.2), the following.

$$\begin{aligned}
 E[(p_i(n+1) - p_i(n))/p(n) = p] &= E[\Delta p_i(n)/p(n) = p] \\
 (3.13) \quad &= p_i d_i(p) \left[a \sum_{j \neq i} p_j \right] + p_i c_i(p) \left[-b + b \sum_{j \neq i} p_j \right] \\
 &\quad + \sum_{j \neq i} p_j d_j(p) [-a p_i] + \sum_{j \neq i} p_j c_j(p) \left[\frac{b}{r-1} - b p_i \right].
 \end{aligned}$$

L_{R-P} automaton. Since $a = b$, we can write (3.13) as

$$\begin{aligned}
 E[\Delta p_i(n)/p(n) = p] &= -a \left[f_i(p) + \frac{1}{r-1} \sum_{j \neq i} f_j(p) \right] \\
 (3.14) \quad &\triangleq a w_i(p).
 \end{aligned}$$

Let $p^* = [p_1^*, p_2^*, \dots, p_r^*]$, $\sum_{i=1}^r p_i^* = 1$ and

$$f_i(p^*) = f_j(p^*), \quad i, j = 1, 2, \dots, r.$$

The existence of such a p^* is shown using Brouwer's fixed point theorem [7] in Appendix 1. The following proposition is a generalization of Proposition 1.

PROPOSITION 2. *If $c_i(p)$, $i = 1, \dots, r$ satisfy the assumption (A1), there exists a p^* such that*

$$f_1(p^*) = f_2(p^*) = \dots = f_r(p^*).$$

Further, if (A2) is satisfied then p^ is unique. \square*

Proof. Refer to Appendix 1.

Let

$$\begin{aligned}
 \Delta p(n) &= [\Delta p_1(n), \Delta p_2(n), \dots, \Delta p_r(n)], \\
 w(p) &= [w_1(p), w_2(p), \dots, w_r(p)],
 \end{aligned}$$

the conditional covariance matrix of $\Delta p(n)$

$$\begin{aligned}
 (3.15) \quad E \left[\left(\left[\frac{\Delta p(n)}{a} - w(p) \right] \left[\frac{\Delta p(n)}{a} - w(p) \right]' \right) / p(n) = p \right] &\triangleq \tilde{s}(p), \\
 z_n^i &\triangleq \frac{p_i(n) - p_i^*}{\sqrt{a}}, \quad z_n = [z_n^1, z_n^2, \dots, z_n^r],
 \end{aligned}$$

and let A be the $r \times r$ matrix, $A = (d/dp)w(p)|_{p=p^*}$.

The following theorem completes the generalization to the r -action case.

THEOREM 3. *Let $p(0)$ have stationary measure \bar{p} . Then*

(a) $E[\|p(n) - p^*\|^2] = O(a)$ ($\|\cdot\|$ denotes the norm) and $E[p(n)] - p^* = O(\sqrt{a})$.

(b) z_n , as $a \rightarrow 0$ is normal with mean zero and covariance Σ which is the unique solution of $A\Sigma + \Sigma A + \tilde{s}(p^*) = 0$. \square

The proof is in Appendix 2.

$L_{R-\varepsilon P}$ automaton. For the case $b = o(a)$, (3.13) reduces to

$$(3.16) \quad \begin{aligned} E[\Delta p_i(n)/p(n) = p] &= a \sum_{j \neq i} p_i p_j [c_j(p) - c_i(p)] + o(a) \\ &\triangleq a w_i(p) + o(a). \end{aligned}$$

If we assume that $c_i(p) = 0$ for $p_i = 0$, then along with assumptions (A1) and (A2) we have, through Proposition 2, a unique p^* such that

$$(3.17) \quad c_1(p^*) = c_2(p^*) = \cdots = c_r(p^*).$$

With $\Delta p(n)$, $w(p)$, A and z_n as in (3.15), and the conditional covariance of $\Delta p(n)$ defined as $\tilde{s}(p) + o(a)$, Theorem 3 holds for the case of $L_{R-\varepsilon P}$ automaton. However, if any of the penalty probabilities, $c_k(p)$, dominates (or is dominated by) others; i.e., if $c_k(p) > c_i(p)$ (or $< c_i(p)$) for all p , and for $i = 1, 2, \dots, r$, then $p_k(n)$ will be asymptotically concentrated near 1 (or 0). All this is summarized in Theorem 4.

THEOREM 4. *Let $p(0)$ have stationary measure \bar{p} .*

- (a) *If (3.17) holds, then $E[p(n)] - p^* = O(\sqrt{a})$, and z_n is normal with mean zero and variance Σ , the unique solution of $A\Sigma + \Sigma A + \tilde{s}(p^*) = 0$ as $a \rightarrow 0$.*
- (b) *If $c_k(p) > c_i(p)$ for all p , $i \neq k$,*

$$E[p_k(n) - 1] = O(\sqrt{a}),$$

or if $c_k(p) < c_i(p)$ for all p , $i \neq k$,

$$E[p_k(n)] = O(\sqrt{a}). \quad \square$$

Remarks. The proof (Appendix 2) to show that the variance of \bar{p} around p^* is $O(a)$, rests on demonstrating that $(p^* - p)w(p) > R > 0$ for almost all $p \neq p^*$, in all the cases of L_{R-P} and $L_{R-\varepsilon P}$. The normality result for the stationary measure follows as an application of Norman's results.

4. Network preliminaries. To understand the exact nature of the nonstationarity arising from the network in (2.3), and the dependence between the state and the routing probabilities, we first consider the case where A_t^k routes with *fixed* probabilities without any updating.

The state of the network is defined as the configuration of calls in progress (or currently connected). It can be characterized by a vector, each component of which gives the number of calls in progress on a particular type of a particular path. State transition occurs when a call in progress hangs up or a new call is connected. Blocking results in self transition. Due to assumptions (i)–(v) of § 2 the network can be described by a finite, homogeneous Markov chain [8]⁵.

Let

- $S \sim$ set of all possible states of network $\equiv \{1, 2, \dots, \nu\}$,
- $h_x \sim$ set of all states with one call hung up from state x ,
- $\alpha_{x,t}^\pi \sim$ set of all states with one additional call type t
connected along path π from state x ,
- $\eta_x \sim$ number of calls in progress in state x ,

$$\bar{r} \triangleq \sum_{t,k} r_t^k, \quad \bar{\lambda} = \sum_{i,j} \lambda_{ij}.$$

⁵ Assumptions (i) and (ii) of § 2 can be relaxed as long as the state evolves according to a homogeneous Markov chain.

The transition matrix Q can be described as follows [8]. Let $x(n)$ be the state immediately after the n th transition. Let $x(n) = x$, and $x(n+1) = y$. Let p be the \bar{r} -vector of the fixed routing probabilities: if $y \in h_x$, $q_{xy} = 1/(\bar{\lambda} + \eta_x)$. If $y \in \alpha_{x,t}^\pi$, $q_{xy} = p_t^\pi \lambda_{ij}/(\lambda + \eta_x)$, where p_t^π is the probability of call type $t = (i, j)$ being routed along π , upon its arrival (p_t^π can be computed from p), and if $y = x$,

$$q_{xx} = 1 - \Pr\{y \in h_x\} - \sum_{t, \pi} \Pr\{y \in \alpha_{x,t}^\pi\}.$$

If $\hat{p}_j(n) \triangleq \Pr\{x(n) = j \in S\}$, and $\hat{p}(n) = (\hat{p}_1(n), \dots, \hat{p}_\nu(n))$, the state evolves according to

$$(4.1) \quad \hat{p}(n+1) = Q(p)\hat{p}(n).$$

The transition matrix depends upon p , as well as upon λ_{ij} , μ and l_{ij} . $Q(p)$ is, in fact, an ergodic matrix (whose ergodic sets⁶ may be functions of p), since any state could be reached from every other state. As $n \rightarrow \infty$, $\hat{p}(n) \rightarrow \hat{p} = Q(p)\hat{p}$, a unique probability independent of $p(0)$; the limit \hat{p} itself depends upon p . (The chain does not have any cyclical subsets.)

Blocking probability. A call arriving at any link or a node either gets blocked or completed, and hence has a certain blocking probability. Different arrival processes in the network, i.e., at different nodes and links, have different blocking probabilities. Formally, we define the blocking probability P_b and completion probability P_c for an arbitrary arrival process. (Though these notions are the same as in the literature [8], [17], the development here which is convenient for our purposes is somewhat different.)

Let $m = 0, 1, 2, \dots$ be the arrival instants. Let $S_c \subseteq S$ be the set of states in which these arrivals can be completed; $S_b = S - S_c$, the set of blocking states. If $x(m^-)$ is the state of the network immediately preceding m ,

$$P_b \triangleq \lim_{m \rightarrow \infty} \Pr\{x(m^-) \in S_b\}, \quad P_c \triangleq \lim_{m \rightarrow \infty} \Pr\{x(m^-) \in S_c\}.$$

The above limits exist, and can be calculated from \hat{p} . Let $n = 0, 1, 2, \dots$ be the state transition instants of the chain (4.1), which are instants of connections, blockings and hangups. If $I_{x(n)}$ is the indicator of blocking of the arrival process,

$$\Pr\{I_{x(n)}\} = \sum_{x \in S_b} \Pr\{\text{arrival of the process}/x(n-1) = x\} \hat{p}_x(n-1),$$

and $P_b = \lim_{n \rightarrow \infty} \Pr\{I_{x(n)}\}$. For example, if blocking probability is calculated for calls with source i and destination j ,

$$(4.2) \quad P_b = \sum_{x \in S_b} \frac{\lambda_{ij}}{\lambda + \eta_x} \hat{p}_x \triangleq \sum_{x \in S_b} \tilde{p}_x \quad \text{and} \quad P_c = \sum_{x \in S_c} \hat{p}_x.$$

P_c and P_c are calculated from (4.2) with $P_b + P_c = \sum_{x \in S} \hat{p}_x = 1$.

Example. For the simplest case of a network with just one link of M trunks, and arrivals λ , $S = \{0, 1, 2, \dots, M\}$, and $S_b = \{M\}$. The vector \hat{p} satisfies $\hat{p}_0 = \hat{p}_1 \cdot 1/(\lambda + 1)$, $\hat{p}_n = \hat{p}_{n-1} \cdot \lambda/(\lambda + n - 1) + \hat{p}_n \cdot \lambda/(\lambda + n)$, and for $i = 1, \dots, n-1$, $\hat{p}_i = \hat{p}_{i-1} \cdot \lambda/(\lambda + i - 1) + \hat{p}_{i+1} \cdot i + 1/(\lambda + i + 1)$. With $\tilde{p}_x = [\lambda/(\lambda + x)]\hat{p}_x$, these equations reduce to $\tilde{p}_i \lambda = \tilde{p}_{i+1} \mu$, for $i = 1, \dots, n-1$. Since $\sum_{i=1}^n \tilde{p}_i = 1$, $P_b = \lambda^M/M! (\sum_{i=1}^M \lambda^i/i!)$, which is Erlang's loss formula [17].

⁶ $S' \subset S$, is an ergodic set for a process $\{x(n)\}_{n \geq 0}$, if $x(n) \in S' \Rightarrow x(m) \in S'$ w.p.l., for all $m \geq n$. E_1, E_2, \dots, E_d are cyclical subsets of some ergodic set if $x(n) \in E_\alpha \Rightarrow x(n+1) \in E_{\alpha+1}$ w.p.l. ($E_{d+1} \equiv E_1$).

If g is the average rate of arrivals at a link (or any process), and P_b its blocking probability, then we refer to $f \triangleq gP_b$ as the (average) blocking rate (as seen by an outside observer) at this link.

Blocking probabilities and blocking rates can be computed from \hat{p} , at various links and nodes, for various call types, and for the entire network.

5. Learning routing in the network. In this section, we first prove the convergence of the network together with all the learning schemes at various nodes (§ 5.1). Then we derive the equilibrium behavior of the linear routing algorithms, and show that they have optimal properties (§§ 5.2 and 5.3). Updating using averaged blockings and completions is discussed in § 5.4.

5.1. A convergence theorem. We consider here the convergence of learning algorithms (2.1) in the network. We show that for any “distance diminishing” updating scheme, the network system converges to a unique stationary measure. The convergence for the linear updating schemes (2.3) follows from its distance diminishing property.

DEFINITION. Let e denote an event—either a hangup or arrival accompanied by blocking or completion. Let us assume that algorithm (2.1) maps $p_i^k(n) = p$ to $p_i^k(n+1) = \bar{U}(p, e)$, under event e at instant n . The algorithm A_i^k is called *distance diminishing* if

$$|\bar{U}(p^1, e) - \bar{U}(p^2, e)| \leq \gamma |p^1 - p^2|$$

for all action probabilities p^1 and p^2 of A_i^k , with $\gamma \leq 1$ for all e , and $\gamma < 1$ for some e which has nonzero probability. (For a weaker notion see [4, § 2].)

Let $n = 0, 1, 2, \dots$ be the instants of network state transitions. Equations (2.1) and (4.1) together govern the evolution of the network and the nodal controllers. The learning schemes face the Markovian nonstationarity of the network. The routing probability of a single algorithm $\{p_i^k(n)\}_{n \geq 0}$ is not Markov, and hence Theorem 1 or the methods of its proof are not directly useful. Thus the convergence of the system cannot be established by showing the convergence of the individual algorithms. We show convergence of the entire network system by considering the process $\{y(n)\}_{n \geq 0}$

$$y(n) \triangleq (x(n), p(n))$$

($p(n)$ is the \bar{r} -vector of routing probabilities), which is a homogeneous Markov process on the product space $S \times [0, 1]^{\bar{r}}$. This process is a compact Markov process (definition and details in Appendix 3) which possesses random contraction properties when the nodal updating schemes are distance diminishing. The proof of Theorem 5 and the following corollary are in Appendix 3.

THEOREM 5. *If the nodal algorithms A_i^k are distance diminishing the process $\{y(n)\}_{n \geq 0}$ has a stationary distribution. If the algorithm is such that this process has a single ergodic set with no cyclical subsets, then $y(n)$ converges (exponentially fast) in distribution to a unique stationary probability \bar{y} , independent of the choice of $y(0)$.*

COROLLARY. *The linear scheme (2.3) is distance diminishing. Hence if A_i^k employs this linear scheme, $y(n)$ converges (exponentially fast) in distribution to a unique \bar{y} , for any $y(0)$.*

Remarks. The distance diminishing criterion is helpful in designing algorithms. If care is taken to avoid multiple ergodic sets and cyclical subsets, a unique statistical equilibrium is assured.

We now turn to characterizing the equilibrium behavior of the algorithms.

5.2. Equilibrium analysis of linear algorithms. Let \bar{p}_i^k be the marginal distribution of the routing probability of A_i^k evaluated from \bar{y} . Let \hat{p}_x be the marginal steady state probability of $x \in S$, again computed from \bar{y} . Attention is focused on one algorithm A_i^k , with $r_{ik} = 2$. Generalization to $r_{ik} > 2$ is straightforward. Let $I_x^{c,i}$ be the indicator of $x \in S$ a success (or call completion) state for action i ; and $I_x^{b,i}$, the indicator of blocking state x .

Let $m = 0, 1, 2, \dots$ be the arrival instants for A_i^k (m^- denotes immediately preceding m). Let A_i^k be the L_{R-P} algorithm. Then, if $p_i(m)$ is the probability for action α_i , and $\Delta p_i(m) \triangleq p_i(m+1) - p_i(m)$,

$$(5.1) \quad \begin{aligned} E[\Delta p_1(m)/p(m) = p, x(m^-) = x] \\ = p_1(ap_2)I_x^{c,1} + p_2(-ap_1)I_x^{c,2} + p_1(-ap_1)I_x^{b,1} + p_2(ap_2)I_x^{b,2}, \end{aligned}$$

since

$$(5.2) \quad \begin{aligned} E[\Delta p_1(m)/p(m) = p] &= EE_x[\Delta p_1(m)/p(m) = p, x(m^-) = x] \\ &= a[p_2P_b^2(p) - p_1P_b^1(p)], \end{aligned}$$

where $P_b^i(p) \triangleq \Pr\{I_{x(m)}^{b,i}\}$. Since $x(m)$ has distribution \hat{p}_x in steady state, $P_b^i(p)$ is the blocking probability of action α_i , for A_i^k . Equation (5.2) is the same as (3.4), of the two-action L_{R-P} in the environment of § 3. $P_b^i(\cdot)$ satisfies the smoothness conditions A3) and (A4), and increases with increase in p_i . If the assumption (A2) on cross-derivatives also holds, then the statement of Theorem 2 holds for the sequence $\{p_i^k(n)\}_{n \geq 0}$, with $p_i^k(0)$ having the stationary measure \bar{p}_i^k . Hence, if g_i^k is the average rate of arrivals for A_i^k , for small a , the blocking rates satisfy

$$(5.3) \quad g_i^k p_i P_b^i(p) = \text{constant for all } i,$$

under equilibrium conditions. Thus the L_{R-P} algorithms attempt to equalize the blocking rates at various parts of the network, achieving “load equalization”, a desired attribute in traffic engineering.

If $L_{R-\varepsilon P}$ algorithms are used for A_i^k , an equation similar to (3.12) can be derived following argument used earlier. $L_{R-\varepsilon P}$ algorithms, for small a , achieve in equilibrium

$$(5.4) \quad P_b^i(p) = \text{constant for all } i.$$

This behavior (5.4) of equalizing the blocking probabilities is shown in § 5.3 to be optimal for a quadratic approximation of the network blocking probability.

If p_i^{k*} is the routing vector which achieves (5.3) or (5.4), then $E[\|\bar{p}_i^k - p_i^{k*}\|^2] \leq Ka$, and $(\bar{p}_i^k - p_i^{k*})/\sqrt{a}$ is normal as $a \rightarrow 0$ for all t and k . The variance σ_i^k of this normal distribution is calculated (e.g., by Theorem 2) using the second order driving terms. Thus σ_i^k and K (Appendix 2) can be computed as functions of λ_{ij} , μ and l_{ij} . If these traffic and network parameters are known to lie within a certain set, the value of “ a ” for the worst case values of the parameters can be computed yielding specified asymptotic accuracy K and fluctuations σ_i^k . In general, the choice of a is crucial for proper adaptation. Larger values result in faster convergence but with larger fluctuations; smaller values in greater accuracy and slower speed of convergence. Different values can be chosen for different A_i^k , depending upon the traffic conditions in various parts of the network.

5.3. A quadratic approximation of blocking probability and optimality. In this section, we make assumptions about the link blocking rates and blocking probabilities as specific functions of the routing probabilities⁷. It will be shown that under these assumptions $L_{R-\epsilon P}$ algorithms operating at all nodes, for all types of calls, in a decentralized and independent manner minimize the global blocking rate for the entire network.

We assume that the blocking probability at a link is a linear function of the rate at which calls attempt that link. For call type $t = (i, j)$, if π is a path through which calls can arrive at node n , we also assume that the rate at which calls arrive at node n , attempting some link at n , through path π is

$$(5.5) \quad l_\pi \times \text{product of the routing probabilities along } \pi \times \lambda_{ij},$$

$1 \geq l_\pi \geq 0$. We take the probabilities of attempting the various available trunk groups (instead of the sequence of trunk groups)⁸ at any node as the routing probabilities at that node. These assumptions of linearity are reasonable under “low” and “average” traffic conditions.

The total rate g_t^n of arrivals for A_t^n is the sum over all possible paths π .

$$g_t^n = \sum_{\pi} l_\pi \times \text{product of routing probabilities along } \pi \times \lambda_{ij},$$

$f_{nl} = \sum_t g_t^n p_t^{nl}$ is the rate of attempts on link (n, l) at node n . Similarly we can compute f_{ln} ($\neq f_{nl}$), the rate of attempts on (n, l) at node l . $f_{nl} + f_{ln}$ is the total attempt rate for (n, l) . The blocking probability at (n, l) , then, is given by $K_{nl}[f_{nl} + f_{ln}]$, where $0 < K_{nl} < 1$ is a constant for the link (n, l) . The blocking rate at n for (n, l) is $f_{nl} \cdot K_{nl}[f_{nl} + f_{ln}]$. We focus attention on A_t^n , and derive the expression for the total network blocking probability and individual blocking probability of actions.

Let $\alpha_1, \alpha_2, \dots, \alpha_r$ be the actions of A_t^n ; p_i , the action probability of α_i . Let $p' = (p_1, \dots, p_r)$. We say node m is “reached” by α_i , if calls from α_i can get routed through m . Let N_i be the set of nodes reached by α_i .

From here on, we drop the scripts t and n , without ambiguity. Let

$$\mathcal{N}_1 = \{m : m \text{ is reached by some } \alpha_i\},$$

$$\mathcal{N}_2 = \{m : m \text{ is not reached by any } \alpha_i\}.$$

Let $l_{im} p_i g$ denote the rate of arrivals (of type t) at $m \in \mathcal{N}_1$, $0 \leq l_{im} \leq 1$. l_{im} can be computed from the l_π 's defined in (5.5). If $S_m = \{k : k \in \mathcal{N}_1 \text{ and } (m, k) \text{ is a link for } A_t^n\}$, then the total block rate, B_R^i , for action α_i is

$$(5.6) \quad B_R^i = \sum_{m \in N_i} \sum_{k \in S_m} h_{mk}^i K_{mk} \left[\left(\sum_{i=1}^r h_{mk}^i \right) + \hat{f}_{mk} \right],$$

where $h_{mk}^i \triangleq l_{im} p_i g p_t^{mk}$, is the rate at which calls through action α_i attempt (m, k) . \hat{f}_{mk} is the sum of the rates of all types of calls other than t which attempt (m, k) either at m or at k . Hence, the blocking probability $c_i(p)$ for α_i is,

$$(5.7) \quad c_i(p) = \sum_{m \in N_i} \sum_{k \in S_m} l_{im} p_i p_t^{mk} K_{mk} \left[\left(\sum_{i=1}^r h_{mk}^i \right) + \hat{f}_{mk} \right].$$

⁷ Exact expressions for block probabilities (§ 4) are complex, and approximations are frequently made.

⁸ Or if actions are sequences, then $\partial P_b^i / \partial p_i \approx 0$, $j \neq i$.

Let L be the $(r \times \mu_1)$ matrix of l_{im} ,

Q the $(\mu_2 \times \mu_1)$ matrix of p_i^{mk} , $m, k \in \mathcal{N}_1$,

and \bar{K} the diagonal $(\mu_2 \times \mu_2)$ matrix of K_{mk} , where μ_1 is the number of nodes in \mathcal{N}_1 , and μ_2 the number of links (m, k) such that $m, k \in \mathcal{N}_1$.

The total network blocking rate B_R can now be written⁹ as

$$(5.8) \quad B_R = [\hat{f} + gp' L Q] \bar{K} [g Q' L' p + \hat{f}] + \bar{f},$$

where \hat{f} is the vector of \hat{f}_{mk} , and \bar{f} is the total block rate in the subnetwork \mathcal{N}_2 . \hat{f} and \bar{f} are the functions of the routing probabilities other than p .

PROPOSITION. *The quadratic minimization problem*

$$\min B_R(p), \quad \text{subject to } \sum_{i=1}^r p_i \leq 1, \quad p_i \geq 0$$

has optimal solution p^* , and $\sum_{i=1}^r p_i^* = 1$.

Remark. $\sum_{i=1}^r p_i^* = 1$, says that no call should be “deliberately” blocked at any node of the network.

Proof of Proposition. $B_R(p)$ is continuous on a compact set, hence attains minimum p^* . From (5.8), we get

$$(5.9) \quad \frac{\partial B_R(p)}{\partial p} = 2gLQ\bar{K}[Q'L'pg + \hat{f}]$$

$$(5.10) \quad = 2gc(p),$$

where $c(p) = [c_1(p)c_2(p) \cdots c_r(p)]$, the vector of blocking probabilities (5.7).

The Kuhn–Tucker necessary conditions for optimality [10] yield

$$(5.11) \quad \begin{aligned} \frac{\partial B_R(p)}{\partial p_i} \Big|_{p_i=p^*} + u - v_i &= 0, \quad i = 1, \dots, r, \\ u \left(\sum_{i=1}^r p_i^* - 1 \right) &= 0, \quad v_i p_i^* = 0. \end{aligned}$$

Suppose $\sum_{i=1}^r p_i^* < 1$; we get $u = 0$. This gives either $\partial B_R(p)/\partial p_i|_{p_i=p^*} = 0$ for $p_i^* \neq 0$, or $p_i^* = 0$ for all i . The latter cannot be true if calls are to be accepted at all. The former also cannot be true: if $p_i \neq 0$, the block rate for α_i , $B_R^i(p)$, is quadratic in p_i , and $\partial B_R^i(p)/\partial p_i > 0$. Also $\partial B_R^i(p)/\partial p_i \geq 0$ for $j \neq i$. This implies $\partial B_R(p)/\partial p_i|_{p_i=p^*} > 0$ if $p_i^* \neq 0$. Hence we must have $\sum_{i=1}^r p_i^* = 1$. \square

Since $A \triangleq (gLQ)\bar{K}(gLQ)'$ is positive definite ($p'Ap$ is the block rate for A^n , when there are no calls which do not go through A^n ; $p'Ap > 0$, for $p \neq 0$), condition (5.11) is also sufficient for optimality.

⁹ denotes matrix transpose.

We now have the optimality condition translated as

$$\left. \frac{\partial B_R(p)}{\partial p_i} \right|_{p_i=p^*} = u \quad \text{for } p_i^* \neq 0, \\ \geq u \quad \text{for } p_i^* = 0,$$

or

$$(5.12) \quad c_i(p^*) = u \quad \text{for } p_i^* \neq 0, \\ \geq u \quad \text{for } p_i^* = 0.$$

If \hat{f} is assumed to be constant (see (5.9)), we get a unique p^* . However when we consider the simultaneous optimality of all automata A_i^n , p^* might be nonunique; p^* , in this case will be on a convex set, since it satisfies a set of linear inequalities like (5.12).

There is a dynamic programming nature of the minimization of B_R for each call type t which is clear from the following:

$$\min_{p, \tilde{p}} B_R(p, \tilde{p}) = \min_{\tilde{p}} \left\{ \bar{f} + \min_p \{ \hat{f} + gp' L Q \bar{K} [Q' L' p g + \hat{f}] \} \right\},$$

where \tilde{p} are the routing probabilities of automata in the subnetwork \mathcal{N}_2 . Every A_i^n must, for optimality, attempt to equalize $c_i(p)$ from its actions α_i . The proof of the following theorem is in Appendix 2.

THEOREM 6. *In the network with blocking probabilities as given by (5.7), for the $L_{R-\varepsilon P}$ routing schemes*

$$E[\|\bar{p} - p^*\|^2] = O(a),$$

where p^* satisfies (5.11), and \bar{p} the equilibrium routing probability. \square

5.4. Updating using averaging. The use of averaged completions and blockings, and its effect on the performance of the routing system is considered here. As an example, we discuss $L_{R-\varepsilon P}$ updating using its past history of completions averaged as

$$\hat{x}_i(n) = (1 - \beta)\hat{x}_i(n-1) + \beta I_{x(n)}^{b,i}, \quad 0 < \beta < 1.$$

Since $\hat{x}_i(n) \in [0, 1]$, algorithm (2.3) for $b = o(a)$ is modified as follows:

Let $\alpha(n) = \alpha_i$, and $1 - \hat{x}_i(n) = \alpha$.

$$(5.13) \quad \text{If } p_i(n) \neq 1, \text{ for } j \neq i, \Delta p_j(n+1) = -a\alpha p_j(n) \text{ and } \Delta p_i(n+1) = a\alpha[1 - p_i(n)].$$

If $p_i(n) = 1$, $p_i(n+1) = 1 - b$, and $p_j(n+1) = b/r - 1$, for $j \neq i$.

Algorithm (5.13) has the form of (2.2).

The convergence Theorem 5 holds for the above algorithm too, considering the process $y'(n) \equiv (x(n), \hat{x}(n), p(n))$. In equilibrium, for small a , this also equalizes the blocking probabilities, as in (5.4). However, the variances σ_i^k are different. In equilibrium,

$$\text{var } \hat{x}_i = (1 - \beta)^2 \text{var } \hat{x}_i + \beta^2 \text{var } I_x^{b,i},$$

or

$$\text{var } \hat{x}_i = \frac{\beta}{2 - \beta} \text{var } I_x^{b,i}.$$

Decreasing β results in smaller values of second order driving terms, and hence also of σ_i^k . K (Appendix 2) is also reduced resulting in increased accuracy.

6. Conclusions. A mathematical model of telephone traffic routing using learning algorithms at the various nodes is developed in this paper. It is shown that if the nodal algorithms are randomly contractive the network system attains statistical equilibrium. Equilibrium behavior is derived for slow-learning algorithms. Under certain assumptions it is demonstrated that the decentralized linear updating algorithms result in optimal routing in terms of the overall blocking probability.

The choice of the parameter a in the learning algorithms, inclusion of realistic constraints (e.g., desired node to node blocking probability), use of nonlinear algorithms and transfer of state information among nodes to improve performance and speed of convergence remain to be investigated in the future.

Appendix 1. Proof of Proposition 2.

1a. We notice that the following conditions hold:

- (i) $f_i(p_1, p_2, \dots, p_r) = 0$ for $p_i = 0$,
- (1) (ii) $f_i(p) \geq 0$ $p \in s^r$ {where s^r is the simplex in r -dimensions},
- (iii) $f_i(\cdot)$ is continuous.

Let

$$(2) \quad w_i(p) \triangleq -f_i(p) + \frac{1}{r-1} \sum_{j \neq i} f_j(p)$$

and

$$w_i^+(p) = \max \{w_i(p), 0\}.$$

We construct the following map $G: s^r \rightarrow s^r$. If $G(p) = p' = [p'_1 \dots p'_r]$,

$$(3) \quad p'_i = \frac{p_i + w_i^+(p)}{1 + \sum_{j=1}^r w_j^+(p)};$$

clearly $\sum_{i=1}^r p'_i = 1$, and G is a continuous map. Hence we have, from Brouwer's fixed-point theorem, $p^* \in s^r$, such that $G(p^*) = p^*$.

Since $\sum_{j=1}^r w_j(p^*) = 0$, from (2) there is at least one i , $1 \leq i \leq r$, such that $w_i^+(p^*) = 0$. From (3) we get

$$(4) \quad p_i^* \left[\sum_{j=1}^r w_j^+(p^*) \right] = 0.$$

$p_i^* = 0$ only if equality holds in condition (ii), in which case $f_i(p^*) = 0$ for all i . Otherwise from conditions (i) and (ii), and (2), $p_i^* \neq 0$. Hence we get $\sum_{j=1}^r w_j^+(p^*) = 0$. This along with $\sum_{j=1}^r w_j(p^*) = 0$ gives $w_i(p^*) = 0$, for all i . It immediately follows that $f_i(p^*) = f_j(p^*)$, for all i, j . \square

1b. Uniqueness. Let $p^*, q^* \in s^r$, $p^* \neq q^*$ such that $w_i(p^*) = w_i(q^*)$ for all i . Let $\hat{w}(p^*) = [w_1(p^*), \dots, w_{r-1}(p^*)]$, $\hat{p}^* = [p_1^*, \dots, p_{r-1}^*]$ and $\hat{q}^* = [q_1^*, q_2^* \dots q_{r-1}^*]$. Since $p_r^* = 1 - \sum_{i=1}^{r-1} p_i^*$, we write $\hat{w}_i(\hat{p})$ for $w_i(p)$, where $\hat{p} = [p_1, \dots, p_{r-1}]$. Considering the line $\hat{p}(\lambda) = \lambda \hat{p}^* + (1-\lambda) \hat{q}^*$ $\lambda \in [0, 1]$,

$$(5) \quad \frac{d\hat{w}(\hat{p}(\lambda))}{d\lambda} = \frac{d\hat{w}(\hat{p}(\lambda))}{d\hat{p}(\lambda)} [\hat{p}^* - \hat{q}^*],$$

where

$$A(p(\lambda)) \triangleq \frac{d\hat{w}(\hat{p}(\lambda))}{d\hat{p}(\lambda)}$$

is the $(r-1) \times (r-1)$ matrix of partial derivatives. From (2) and assumption (A2), the diagonal terms of $A(p(\lambda))$ are dominated by $-\partial f_i / \partial p_i - (1/r-1)(\partial f_r / \partial p_r)$, and the

off-diagonal terms, by $(1/(r-1))\partial f_i/\partial p_i - (1/(r-1))\partial f_r/\partial p_r$. Hence $A(p(\lambda))$ can be written as $BD + C$, where D is a diagonal matrix of $-\partial f_i/\partial p_i$, C is the matrix of identical terms $-(1/(r-1))\partial f_r/\partial p_r$, and B is the following $(r-1) \times (r-1)$ circulant matrix,

$$B = \begin{bmatrix} 1 & -\frac{1}{r-1} & -\frac{1}{r-1} & \cdots & -\frac{1}{r-1} \\ -\frac{1}{r-1} & 1 & \cdots & & -\frac{1}{r-1} \\ \vdots & & & & \\ -\frac{1}{r-1} & & \cdots & & 1 \end{bmatrix}$$

The $(r-1)$ eigenvalues of B are given by [11] $1 - (1/(r-1))[x + x^2 + \cdots + x^{r-2}]$, where x_i is a root of $x^{r-1} = 1$. Since $|x + x^2 + \cdots + x^{r-2}| < r-1$, and B is real and symmetric, the eigenvalues are real and positive. Hence, $A(p(\lambda)) = BD + C$ is negative definite, for $\partial f_i/\partial p_i > 0$.

We have from (5),

$$\int_0^1 \frac{d\hat{w}(\hat{p}(\lambda))}{d\lambda} d\lambda = \hat{w}(\hat{p}^*) - \hat{w}(\hat{q}^*) = 0 = A(\hat{p}^* - \hat{q}^*).$$

Since $A = \int_0^1 A(p(\lambda)) d\lambda$ is negative definite, $\hat{p}^* = \hat{q}^*$. This implies $p^* = q^*$, and uniqueness follows. \square

Appendix 2.

2a. Here, p^2 denotes $p'p$, if p is a vector. We write p_n for $p(n)$. K denotes a constant. $p_{n+1} = p_n + \Delta p_n$. $E[(\Delta p_n/a)/p_n = p] = w(p) + O(a)$. $E[(\Delta p_n^2/a^2)/p_n = p] = \hat{s}(p) + o(a)$.

LEMMA. Let $p(0)$ have stationary measure. If p^* is such that $w(p^*) = 0$, and

$$(1) \quad (p^* - p)w(p) > R(p - p^*)^2, \quad R > 0,$$

for all probabilities p , then $E[p(n) - p^*]^2 = O(a)$.

Proof of lemma. $(p_{n+1} - p^*)^2 = (p_n - p^*)^2 + 2(p_n - p^*)\Delta p_n + \Delta p_n^2$. Taking expectations on both sides, and cancelling $E(p_n - p_n^*)^2$,

$$0 = E[(p_n - p^*)\Delta p_n] + \frac{1}{2}E[\Delta p_n^2]$$

or

$$0 = E[(p_n - p^*)w(p_n)] + \frac{a}{2}E[\hat{s}(p_n)] + \frac{o(a)}{a}.$$

Since, we have only bounded random variables, $\hat{s}(p_n)$ is bounded and $o(a)/a \leq Ka$. Hence

$$(2) \quad E[(p^* - p_n)w(p_n)] \leq Ka.$$

The lemma follows from (1) and (2). \square

To derive the results (Theorems 2, 3, 4 and 6) of § 3 and § 5.3 for the two- and multi-action L_{R-P} and $L_{R-\varepsilon P}$ schemes we merely show that (1) holds. The proof for the normal approximation is given in Appendix 2b.

Proof of Theorem 2. (two-action L_{R-P}) Though this is a special case of Theorem 3, the proof is given here for the sake of continuity and follows [4, pp. 154]. Let $p = p_1$, and

$$g(p) = \frac{w(p)}{(p^* - p)}, \quad p \neq p^*,$$

$$= -\left. \frac{dw(p)}{dp} \right|_{p=p^*}, \quad p = p^*.$$

Because of assumption (A2), $w(p) < 0$ when $p > p^*$, and $w(p) > 0$ when $p < p^*$. So, $g(\cdot)$ is positive and continuous on $[0, 1]$. Hence there is a $R > 0$ such that $g(p) \geq R$. Thus,

$$(p^* - p_n)w(p_n) = (p^* - p_n)^2 g(p_n) \geq R(p_n - p^*)^2. \quad \square$$

Proof of Theorem 3. (r -action L_{R-P}). Let $p = (p_1, p_2, \dots, p_{r-1})$, and $w(p) = (w_1(p), \dots, w_{r-1}(p))$. $p_r = 1 - \sum_{i=1}^{r-1} p_i$. If $p(\lambda) = \lambda p + (1 - \lambda)p^*$, $\lambda \in [0, 1]$, going through the arguments of Appendix 1b, we have

$$(3) \quad (p^* - p_n)w(p_n) \geq (p^* - p_n)A(p_n - p^*)$$

$$\geq R(p_n - p^*)^2,$$

where A is negative definite, and R is the minimum eigenvalue of $-A$. Hence, we have

$$E[(p_n - p^*)^2] = O(a).$$

Also,

$$E[(p_i(n) - p_i^*)^2] \leq E[(p_n - p^*)^2] = O(a). \quad \square$$

Two-action $L_{R-\varepsilon P}$. Let $p = p_1$.

Case 1. Suppose $p^* \in (0, 1)$. $w(p) = 0$ for $p = 0$ or 1 . But for $p \neq 0$ or 1 , equation (4) holds for this situation. However, consider the regions $R = [\varepsilon, 1 - \varepsilon]$, $\bar{R} = [0, 1] - R$ with $0 < \varepsilon < 1$. If $E_R[(p_n - p^*)^2]$ denotes the integral over R , we have

$$E[(p_n - p^*)^2] = E_R[(p_n - p^*)^2] + E_{\bar{R}}[(p_n - p^*)^2],$$

$E_R[(p_n - p^*)^2] = O(a)$ from (4) and (3). For every a , we can have ε sufficiently small so that the stationary measure \bar{p} on \bar{R} is $O(a)$. This can be done because \bar{p} cannot assume a nonzero measure on a point set. (Notice that the algorithm moves away from any point with probability 1, in the next instant.)

Case 2. If $p^* = 1$, then $c_1(p) < c_2(p)$ for all p . Then $w(p) > 0$ for all p . If $p^* = 0$, then $w(p) < 0$ for all p . \square

Proof of Theorem 4. (r -action $L_{R-\varepsilon P}$). Let $p_i \in [\varepsilon, 1 - \varepsilon]$, for small $\varepsilon > 0$ as in the two-action $L_{R-\varepsilon P}$ case. $p = [p_1, p_2, \dots, p_r]$.

$$w_i(p) = p_i \sum_{j \neq i} p_j [c_j(p) - c_i(p)]$$

$$\geq \left[\min_{j \neq i} p_j \right] p_i (r-1) \bar{w}_i(p)$$

$$\leq \left[\max_{j \neq i} p_j \right] p_i (r-1) \bar{w}_i(p),$$

where

$$\bar{w}_i(p) = -c_i(p) + \frac{1}{r-1} \sum_{j \neq i} c_j(p).$$

Now, if $(p_i^* - p_i) > 0$, $(p_i^* - p_i)w_i(p) \geq m_1 p_i(r-1)[p_i^* - p_i]\bar{w}_i(p)$ and if $(p_i^* - p_i) < 0$, $(p_i^* - p_i)w_i(p) \geq m_2 p_i(r-1)[p_i^* - p_i]\bar{w}_i(p)$, where $\min_{i \neq 1} p_i = m_1 > 0$, $\max_{i \neq 1} p_i = m_2 > 0$, it follows that $\sum_{i=1}^r (p_i^* - p_i)w_i(p) > k_1 \sum_{i=1}^r (p_i^* - p_i)\bar{w}_i(p)$ with $k_1 > 0$. Since $c_i(p) = 0$ for $p_i = 0$, $c_i(\cdot)$ and $f_i(\cdot)$ [of Theorem 3] have similar properties. From (3),

$$\sum_{i=1}^r (p_i^* - p_i)\bar{w}_i(p) \geq R(p - p^*)^2, \quad R > 0. \quad \square$$

The cases of $c_k(p)(\geq)c_j(p)$ for $j \neq k$, is similar to Case 2 in the proof for two-action L_{R-EP} .

Proof of Theorem 6. (network). p_i , $i = 1, \dots, r$ is the action probability of A_i^n . Let $p = [p_1, p_2, \dots, p_{r-1}]$.

$$w(p) = -B[Ap + b],$$

where B is the positive definite circulant matrix defined earlier in the appendix, $A \triangleq gLQ\bar{K}\bar{Q}'L'g$ is positive definite, and $b = gLQ\bar{K}\hat{f}$. p^* is such that $Ap^* + b = u \cdot \mathbf{1}$ (u is a scalar constant, and $\mathbf{1}$ is a $(r-1)$ -vector of 1's), $w(p^*) = 0$.

$$w(p) = w(p) - w(p^*) = -BA(p - p^*).$$

Hence, $(p^* - p)w(p) \geq R(p - p^*)^2$, where $R > 0$ is the minimum eigenvalue of $(-BA)$.

p^* may not be unique. though there is a unique stationary measure, asymptotically the sequences have large correlation, for small " a ". This could be seen from

$$E(p'_{n+1}p_n) = \nu - \frac{a^2}{2}s(p_n),$$

where $\nu \triangleq E(p'_n p_n)$, and since $2E[\Delta p_n^2 p_n] = E[\Delta p_n^2] = a^2 s(p_n)$. Hence, every realization of the process asymptotically gets clustered at one of the p^* .

2b. In order to prove the normal approximation of the stationary measure, as $a \rightarrow 0$, (Theorems 2, 3 and 4 and § 5.2) we establish the following claim. This claim verifies the conditions of Lemma 2.1 [4, pp. 156] whose application yields the normal approximation. The claim is proved for the two-action case, here. Arguments for the multi-action case are similar and can be found in [15].

We write p_n for $p_1(n)$. Let

$$(1) \quad z_n \triangleq \frac{(p_n - p_1^*)}{\sqrt{a}}, \quad w'(p) = \frac{\partial w(p)}{\partial p}.$$

CLAIM. (1) $E[\Delta z_n / z_n] = aw'(p^*)z_n + o(a)$,

(2) $E[\Delta z_n^2 / z_n] = a\tilde{s}(p^*) + o(a)$,

(3) $E[|\Delta z_n|^3 / z_n] = o(a)$,

where $o(a)$ is such that $E[o(a)/a] \rightarrow 0$ as $a \rightarrow 0$.

Proof. Let

$$\zeta = E[\Delta z_n / z_n] / \sqrt{a} = E[z_{n+1} - z_n / z_n] / \sqrt{a} = E\left[\frac{\Delta p_n}{a} / z_n\right] = w(p_n).$$

we write

$$(2) \quad \zeta = w(p_n) - w(p^*)$$

so that

$$(3) \quad |\zeta| \leq K\sqrt{a}|z_n|,$$

since $w(\cdot)$ is Lipschitz (assumption (A4)). K denotes a constant.

Let

$$\eta = E[\Delta z_n^2 / z_n] / a = s(p_n) = \tilde{s}(p_n) + \zeta^2;$$

we have

$$(4) \quad \begin{aligned} |\eta - \tilde{s}(p^*)| &= |\tilde{s}(p_n) - \tilde{s}(p^*) + \zeta^2| \\ &\leq |\tilde{s}(p_n) - \tilde{s}(p^*)| + |\zeta^2| \\ &\leq K\sqrt{a}|z_n| + Ka|z_n|^2 \\ &\leq K|p_n - p^*| + K|p_n - p^*|^2, \end{aligned}$$

since $\tilde{s}(\cdot)$ is also Lipschitz (A4).

$E[p_n - p^*]^2 \leq Ca$ and $E[p_n - p^*] \leq C\sqrt{a}$. This and (4) imply that $|\eta - \tilde{s}(p^*)| \leq o(a)$. Hence claim (2) follows.

To prove claim (1), let

$$h(\lambda) = w(x + \lambda(y - x)), \quad \lambda \in [0, 1],$$

$$h'(\lambda) \triangleq \frac{\partial h(\lambda)}{\partial \lambda} = w'(x + \lambda(y - x))[y - x],$$

where h' is continuous (assumption (A4)). By the fundamental theorem of calculus,

$$w(y) - w(x) = h(1) - h(0) = \int_0^1 w'(x + \lambda(y - x)) d\lambda [y - x].$$

Hence,

$$w(y) - w(x) - w'(x)(y - x) = \int_0^1 [w'(x + \lambda(y - x)) - w'(x)] d\lambda [y - x]$$

so that

$$(5) \quad |w(y) - w(x) - w'(x)(y - x)| \leq \frac{\beta}{2(y - x)^2},$$

where $w(x)$ is Lipschitz bounded by β (A4). From (5), we have

$$(6) \quad |w(p_n) - w(p^*) - w'(p^*)(p_n - p^*)| \leq Ka|z_n|^2$$

Claim (2) follows from (6), (1) and (2).

The proof of claim (3) follows readily by observing that

$$E\left[\frac{|\Delta p_n|^3}{a} \middle| p_n = p\right] = \gamma(p) < \gamma < \infty,$$

and hence $E[|\Delta z_n|^3 / z_n] = E[(|\Delta p_n|^3 / \sqrt{a}) / z_n] < \gamma \cdot a^{3/2} \rightarrow 0$ as $a \rightarrow 0$. \square

Appendix 3. To prove Theorem 5, we first consider the case when there is only one algorithm with two-actions in the network. (Other nodal algorithms may be thought of as having fixed routing probabilities.) Extension to the general case is easy and is discussed later. Let $p_i(n)$ be the probability of action α_i , $i = 1, 2$. We focus attention on $p_1(n)$, and write p_n for $p_1(n)$. The equations governing the process are

$$(1) \quad \begin{aligned} \hat{p}_{n+1} &= Q(p_n) \hat{p}_n, \\ p_{n+1} &= p_n + aU(p_n, x_n) \triangleq \bar{U}(p_n, x_n). \end{aligned}$$

We can write the state transformation as:

$$(2) \quad y_{n+1} = \begin{bmatrix} x_{n+1} \\ p_{n+1} \end{bmatrix} \stackrel{W}{\leftarrow} \begin{bmatrix} x_n \\ p_n \end{bmatrix} = y_n, \quad T \triangleq \begin{bmatrix} W \\ \bar{U} \end{bmatrix},$$

$Y = S \times [0, 1]$ is the state set. The event set E has five objects $E \equiv \{h, (g, \alpha_i, \hat{x})\}$, which are hungup (h) and arrival (g) followed by action (α_i) and completion or blocking (\hat{x}).

Remarks. Each network state $x \in S$ can be viewed as a separate stationary environment for the algorithm. The algorithm faces these stationary environments which are switching in a Markov manner, the transitions being governed by $Q(p)$. The essence of the proof of Theorem 5 lies in the dependence of $Q(p)$ on p being smooth (which it is in the network), and in the distance diminishing of $\bar{U}(\cdot)$.

Before we present the proof, we give some preliminaries concerning compact Markov processes [4, § 3] and their convergence. Let Y be the state space, and E the event space; B and ζ Borel classes on Y and E respectively. Let Y be compact in metric d . Let the homogeneous Markov process $\{y_n\}_{n \geq 0}$ be generated by

$$y_{n+1} = T(y_n, e_{n+1}),$$

and for $A \in \zeta$

$$(3) \quad \Pr(e_{n+1} \in A / y_n, e_n, y_{n-1}, \dots) = \Pr(e_{n+1} \in A / y_n) \triangleq \Pr(e_{n+1}, y_n).$$

Define for $f: Y \rightarrow R$ (reals)

$$m(f) \triangleq \sup_{y_1 \neq y_2} \frac{|f(y_1) - f(y_2)|}{d(y_1, y_2)}.$$

DEFINITION. $\{y(n)\}_{n \geq p}$ is a *compact Markov process* if Y is compact in metric d , and if

$$(4) \quad m(E_1 f) \leq \gamma m(f) + R|f|, \quad \gamma < 1, \quad R < \infty,$$

where $E_1 f(y) \triangleq E[f(y_1) / y_0 = y]$ and f is any bounded Lipschitz function, $|f| \triangleq \sup_{y \in Y} |f(y)|$.

LEMMA. [4, § 3]. If $\{y_n\}_{n \geq 0}$ is compact Markov, then $\{y_n\}_{n \geq 0}$ has a stationary distribution. If in addition there is a single ergodic set with no cyclical classes, then y_n converges, exponentially fast, to a unique stationary distribution independent of the choice of y_0 .

Proof of Theorem 5. We write $y \in Y$ as $y = (x, p)$, $x \in S$ and $p \in [0, 1]$. Define the metric d as

$$d(y_1, y_2) \triangleq d_1(p_1, p_2) + d_2(x_1, x_2) \quad \text{for } y_i = (x_i, p_i), \quad i = 1, 2,$$

where $d_1(p_1, p_2) = |p_1 - p_2|$ and $d_2(x_1, x_2) = |J_{x_1} - J_{x_2}|$. J_{x_i} is real and $0 < J_{x_i} \neq J_{x_j}$ for all $i \neq j$. Y is compact in d . From here on, our goal is prove (4) for the process of equations (1) and (2).

Consider $y_1 \neq y_2$,

$$(5) \quad \begin{aligned} E_1 f(y_1) - E_1 f(y_2) &= \sum_{e \in E} \Pr(e, y_1) [f(T(y_1, e)) - f(T(y_2, e))] \\ &\quad + \sum_{e \in E} [\Pr(e, y_1) - \Pr(e, y_2)] f(T(y_2, e)). \end{aligned}$$

In (5), $\sum_{e \in E} \Pr(e, y_1) f(T(y_2, e))$ has been added and subtracted.

CLAIM 1. *Suppose*

$$(6) \quad |\Pr(e, y_1) - \Pr(e, y_2)| < K d(y_1, y_2), \quad K < \infty,$$

and

$$(7) \quad \left| \sum_{e \in E} \Pr(e, y_1) [f(T(y_1, e)) - f(T(y_2, e))] \right| \leq \gamma m(f) d_1(p_1, p_2) + R|f|,$$

where $\gamma > 1$ and $R > \infty$. Then (4) holds.

Proof of Claim 1. Taking $|\cdot|$ on both sides of (5), and dividing by $d(y_1, y_2)$, $y_1 \neq y_2$, we have

$$(8) \quad m(E_1 f) \leq \frac{\gamma m(f) d_1(p_1, p_2)}{d_1(p_1, p_2) + d_2(x_1, x_2)} + \frac{R|f|}{d_1(p_1, p_2) + d_2(x_1, x_2)}$$

$$(9) \quad \leq \gamma m(f) + R R_2 |f|.$$

$R_2 = 1/\min_{x_1 \neq x_2} d(x_1, x_2)$. Note that with $R R_2 < \infty$, (9) is the same as (4). Claim 1 is now proved. If we now show that (6) and (7) hold, the theorem is proved.

CLAIM 2. *Equation (6) holds.*

Proof of Claim 2. Note that matrix $Q(p)$ is continuously differentiable in $p \in (0, 1)$. (At the boundaries 0 and 1, the derivatives are the left and right limits.) Note also that $\Pr(e, y_1) = \Pr(e, x_1, p_1)$ is an element of the matrix $Q(p_1)$. Now

$$(10) \quad \begin{aligned} I &\triangleq |\Pr(e, x_1, p_1) - \Pr(e, x_2, p_2)| \\ &\leq |\Pr(e, x_1, p_1) - \Pr(e, x_1, p_2)| + |\Pr(e, x_2, p_2) - \Pr(e, x_1, p_2)|. \end{aligned}$$

The first term in (10) is $< K_1 d_1(p_1, p_2)$, since every element of $Q(p)$ is continuously differentiable. The second term is the difference of two elements of $Q(p_2)$ which is less than 1, and hence the second term $< K_2 d_2(x_2, x_1)$ since $x_1 \neq x_2$. ($x_1 - x_2$ implies second term = 0). Hence $I \leq \max(K_1, K_2) d(y_1, y_2)$ for all e . Claim 2 is proved.

CLAIM 3. *Equation (7) holds.*

Proof of Claim 3. We first note that

$$(11) \quad |f(x_1, p) - f(x_2, p)| \leq 2 \max_i |f(x_i, p)| \leq 2|f| R_2 d(x_1, x_2) \leq R_1 |f| d(x_1, x_2).$$

$R_1 = 2R_2 < \infty$. Going back to the left-hand side of (7), we have, (note $T = [W, \bar{U}]$)

$$(12) \quad \begin{aligned} I &\triangleq |f(T(y_1, e)) - f(T(y_2, e))| \\ &\leq |f[W(y_1, e), \bar{U}(y_1, e)] - f[W(y_2, e), \bar{U}(y_1, e)]| \\ &\quad + |f[W(y_2, e), \bar{U}(y_1, e)] - f[W(y_2, e), \bar{U}(y_2, e)]|. \end{aligned}$$

From (11), the first term in the right side of (12) is

$$\leq R_1 |f| d_1[W(y_1, e), W(y_2, e)] \leq R|f|,$$

where $R = R_1 \max_{x_i \neq x_j} d(x_i, x_j)$. The second term in the left side of (12) $\leq \gamma m(f) d_1(p_1, p_2)$, since $\bar{U}(\cdot)$ is a distance diminishing algorithm. Thus (12) holds for

every e . This proves Claim 3, and hence Theorem 5 holds for the case of one automaton in the network.

The proof for the general case is as follows. $y = (x, p)$, and now p is a \bar{r} -vector of p^t where $t = (m, n)$ and p^t is the routing probability of A_m^n . Let $y_i = (x_i, p_i) \in S \times [0, 1]^{\bar{r}}$, $i = 1, 2$. We define d as,

$$(13) \quad d(y_1, y_2) \triangleq d_2(x_1, x_2) + \sum_i d_1(p_1^i, p_2^i).$$

d_1 and d_2 are as defined earlier. It can be observed that for every $y \in Y$, there is a nonzero probability of some call arrival, and hence at least one of the algorithms gets updated, while others do not change. Hence the total routing system—all the algorithms taken together—is distance diminishing in metric (13). This fact along with the arguments given for the one algorithm case, proves Theorem 5 for the general case.

Proof of corollary. For the linear scheme (2.3), with two-actions

$$\begin{aligned} \frac{d(\bar{U}(p^1, e), \bar{U}(p^2, e))}{d(p^1, p^2)} &= 1 && \text{for } e = h, \\ &= 1 - a && \text{for } e = (g, \alpha_1, 0), \\ &= b && \text{for } e = (g, \alpha_1, 1), \\ &= a && \text{for } e = (g, \alpha_2, 0), \\ &= 1 - b && \text{for } e = (g, \alpha_2, 1), \end{aligned}$$

$\gamma \triangleq \max \{1 - a, b, a, 1 - b\} < 1$ for $e \neq h$, which has nonzero probability. Hence $\bar{U}(\cdot)$ is distance diminishing. (The proof for multi-action is similar with state space as simplex.) The corollary is proved after noting the fact that y_n process has only one ergodic set $S \times [0, 1]^{\bar{r}}$ which does not have cyclical subsets. (Note that the network chain and the algorithm, by themselves, do not have cyclical classes.) \square

Acknowledgment. The authors would like to thank Professor Thathachar for numerous discussions.

REFERENCES

- [1] K. S. NARENDRA, E. A. WRIGHT AND L. G. MASON, *Applications of learning automata to telephone traffic routing problems*, IEEE Trans. Systems, Man, and Cybernetics, SMC-7 (1977), pp. 785–792.
- [2] K. S. NARENDRA AND M. A. L. THATHACHAR, *On the behavior of a learning automaton in a changing environment with application to telephone traffic*, S & IS Report 7803, Yale University 1978, and IEEE Trans. Systems, Man, and Cybernetics, SMC-10 (1980), pp. 262–269.
- [3] ———, *Learning automata—a survey*, IEEE Trans. Systems, Man, and Cybernetics, SMC-4 (1974), pp. 323–334.
- [4] M. F. NORMAN, *Markov Processes and Learning Models*, Academic Press, New York, 1972.
- [5] V. I. VARSHAVSKII AND I. P. VORONTOVA, *On the behavior of stochastic automata with variable structure*, Automation and Remote Control, 24 (1963), pp. 327–333.
- [6] S. LAKSHMIVARAHAN AND K. S. NARENDRA, *Learning algorithm for two person zero-sum stochastic games with incomplete information*, S & IS Report 7908, Yale University (1979), to appear in Math. Ops. Res.
- [7] R. LARSEN, *Functional Analysis*, Marcel Dekker, New York, 1973.
- [8] V. BENES, *Mathematical Theory of Connecting Networks and Telephone Traffic*, Academic Press, New York, 1965.
- [9] R. L. FRANKS AND R. W. RISHL, *Optimum network call-carrying capacity*, Bell System Tech. J. (1973), pp. 1194–1214.
- [10] D. G. LUENBERGER, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1973.

- [11] R. BELLMAN, *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1960.
- [12] K. R. KRISHNAN, Bell Laboratories, private communication.
- [13] R. G. GALLAGER, *Distributed algorithms for data communication network routing*, IEEE Trans. Communic. 1 (1977), pp. 73–84.
- [14] T. E. STERN, *Relaxation methods for decentralized routing*, IEEE Trans. Communic. 10 (1977), pp. 1092–1102.
- [15] P. R. SRIKANTAKUMAR AND K. S. NARENDRA, *Learning algorithm model for routing in telephone networks*, S & IS Report, Yale University (1979).
- [16] V. BENEŠ, *Programming and control problems arising from optimal routing in telephone networks*, SIAM J. Control, 4 (1966), pp. 6–18.
- [17] T. L. SAATY, *Elements of Queueing Theory*, McGraw-Hill, New York, 1961.

OPTIMAL CONTROL OF STOCHASTIC INTEGRALS AND HAMILTON-JACOBI-BELLMAN EQUATIONS. I*

PIERRE-LOUIS LIONS† AND JOSÉ-LUIS MENALDI‡

Abstract. We consider the solution of a stochastic integral control problem and we study its regularity. In particular, we characterize the optimal cost as the maximum solution of

$$\begin{aligned} \forall v \in V, \quad A(v)u &\leq f(v) \quad \text{in } \mathcal{D}'(\mathcal{O}), \\ u &= 0 \quad \text{on } \partial\mathcal{O}, \quad u \in W^{1,\infty}(\mathcal{O}), \end{aligned}$$

where $A(v)$ is a uniformly elliptic second order operator and V is the set of the values of the control.

1. Introduction

1.1 General introduction. In this paper we are interested in the following problem. We consider a stochastic system governed by the stochastic differential equation

$$\begin{aligned} (1.1) \quad dy(t) &= \sigma(y(t), v(t)) dW_t + g(y(t), v(t)) dt, \quad t \geq 0, \\ y(0) &= x \in \mathbb{R}^N, \end{aligned}$$

where W_t is a Wiener process, g, σ , are given functions and $v(t)$ is a “continuous” control taking values in some set $V \subset \mathbb{R}^m$. We want to minimize the cost function.

$$(1.2) \quad J(x, v(\cdot)) = E \left\{ \int_0^\tau f(y(t), v(t)) \exp \left(- \int_0^t c(y(s), v(s)) ds \right) dt \right\}$$

over all admissible controls $v(t)$. In this formula f and c are known, given functions and τ is the exit time of the process $y(t)$ from a given domain $\bar{\mathcal{O}}$. Let us denote $u(x) = \inf_{v(\cdot)} J(x, v(\cdot))$.

At least formally, by the argument of dynamical programming, one can derive the following equation satisfied by u :

$$\begin{aligned} (1.3) \quad \sup_{v \in V} \{A(v)u(x) - f(x, v)\} &= 0 \quad \text{in } \mathcal{O}, \\ u &= 0 \quad \text{on } \partial\mathcal{O} = \Gamma, \end{aligned}$$

where $A(v) = -\frac{1}{2} \sigma \sigma^T(x, v) \cdot D^2 - g(x, v) \cdot D + c(x, v)^1$

Thus the initial stochastic control problem is connected to some nonlinear second order elliptic problem with Dirichlet boundary conditions; problem (1.3) is called the Dirichlet problem for Hamilton–Jacobi–Bellman equations.

In the following, we are going first to build a nonlinear semigroup whose generator is essentially the nonlinear operator defined by (1.3). The optimal cost function $u(x)$ appears then to be the unique fixed point of this semigroup: this fixed-point formulation can be viewed as a weak formulation of (1.3) or as the mathematical expression of dynamical programming. These results are in the spirit of those of M. Nisio [24].

* Received by the editors June 13, 1980, and in revised form January 30, 1981.

† Laboratoire d'Analyse Numérique, Université Pierre et Marie Curie, 4 Place Jussieu, 75230 Paris Cédex 05, France.

‡ INRIA, Domaine de Voluceau, Rocquencourt B.P. 105, 78153 Le Chesnay Cédex, France.

¹ σ^T , σ is the adjoint of σ .

Next we prove under very general assumptions that u lies in $W_0^{1,\infty}(\mathcal{O})$ and that u is the maximum element of functions $w \in W_0^{1,\infty}(\mathcal{O})$ satisfying $A(v)w \leq f(v)$ in $\mathcal{D}'(\mathcal{O})$ for all $v \in V$. Of course this is a characterization of u , and it seems very useful since in some degenerate cases it is known that (1.3) does not hold (cf. Genis and N. V. Krylov [10]).

Here in part I, after giving some general results in the construction of this nonlinear semigroup, we essentially treat the case of nondegenerate stochastic integrals ($A(v)$ is uniformly elliptic) under mild regularity assumptions. In Part II [26] (this issue, pp. 82–95) the general case is considered.

The main results of this study were announced in [21]; we also proved a result on the verification of (1.3) (including [21]) which was also proved by different methods at the same time by L. C. Evans and A. Friedman [6]. Concerning the verification of (1.3) more general results were obtained by P.-L. Lions [15], L. C. Evans and P.-L. Lions [7] (in the case of nondegenerate diffusions), P.-L. Lions [16], [17] (in the general case). Below we will recall briefly their main results. We emphasize that we give here a different characterization of the optimal cost, requiring less regularity of \mathcal{O} and of the coefficients and fewer assumptions on the nondegeneracy of $\sigma(x, u)$; this must be so for an approach to be valid while the verification of (1.3) is no longer true.

Finally, we recall that this kind of problem is introduced in the book of W. H. Fleming and R. Rishel [8], and that the first general results on this problem were obtained by N. Krylov [11], [12], [14].

1.2. Summary. Our results are organized in the following way:

Section 2 Construction of a nonlinear semigroup.

Section 3. A stochastic characterization of $u(x)$.

Section 4. An analytical characterization of $u(x)$.

In § 2, following some techniques of M. Nisio [23], we build a nonlinear semigroup whose generator is related to the operator appearing in (1.3). In § 3 we give a stochastic characterization of $u(x)$, the precise way to supply dynamical programming. Finally in § 4 we prove a characterization of $u(x)$, in terms of a maximum solution of inequalities. In § 4, we shall suppose that $\sigma(x, v)$ are nondegenerate matrices. The generalization to the case of degeneracy will be developed in Part II, together with results concerning other boundary conditions, the case of optimal stopping and the case of nonhomogeneous diffusions and parabolic equations.

1.3. Assumptions and notation. We now give notation and assumptions which will remain valid in §§ 2, 3 and 4.

Let \mathcal{O} be a domain in \mathbb{R}^N , and let V be a convex closed set in \mathbb{R}^m . We call an *admissible system* a set $\mathcal{A} = (\Omega, F, F_t, P, W_t, v(t), y_x(t))$, where (Ω, F, P) is a probability space, F_t is a nondecreasing right continuous family of sub σ -algebras F_t of F , W_t is a Wiener process with respect to F_t , $v(t)$ is a measurable adapted process taking values in some compact subset V_0 of V (V_0 of course may depend on $v(\cdot)$) and $y_x(t)$ is a solution of

$$\begin{aligned} dy_x(t) &= \sigma(y_x(t), v(t)) dW_t + g(y_x(t), v(t)) dt, \\ y_x(0) &= x. \end{aligned} \quad (1.4)$$

We suppose that σ, g satisfy

$$(1.5) \quad |\phi(x, v) - \phi(x', v')| \leq C|x - x'| + \rho(|v - v'|) \quad \forall x, x' \in \mathbb{R}^N, \quad \forall v, v' \in V,$$

where $\phi = \sigma_{ij}(1 \leq i, j \leq n)$, $g_i(1 \leq i \leq n)$ and ρ is a given continuous function from \mathbb{R}_+ into \mathbb{R}_+ with $\rho(0) = 0$.

We assume also that we have

$$(1.6) \quad |\sigma(x, v)| + |g(x, v)| \leq C \quad \forall x \in \mathbb{R}^N, \quad \forall v \in V.$$

Now for an admissible system \mathcal{A} we define a cost function

$$(1.7) \quad J(x, \mathcal{A}, t, h) = E \left\{ \int_0^{t \wedge \tau_x} f(y_x(s), v(s)) \exp \left(- \int_0^s c(y_x(\lambda), v(\lambda)) d\lambda \right) ds \right. \\ \left. + h(y_x(t \wedge \tau_x)) \cdot \exp \left(- \int_0^{t \wedge \tau_x} c(y_x(s), v(s)) ds \right) \right\},$$

where h is an arbitrary measurable bounded function, τ_x is the first exit time from $\bar{\mathcal{O}}$ of $y_x(t)$, and $f(x, v)$, $c(x, v)$ are given and are assumed to satisfy (1.5) with $\phi = c$, (1.6) and

$$(1.8) \quad |f(x, v) - f(x', v')| \leq \rho(|x - x'| + |v - v'|) \quad \forall x, x' \in \mathbb{R}^N, \quad \forall v, v' \in V,$$

$$(1.9) \quad c(x, v) \geq c_0 \geq 0 \quad \forall x \in \mathbb{R}^N, \quad \forall v \in V.$$

Finally we define for each h , an optimal cost function

$$(1.10) \quad Q(t)h(x) = \inf J(x, \mathcal{A}, t, h) \quad \forall 0 \leq t < +\infty.$$

Let us collect our assumptions:

$$(1.5) \quad |\phi(x, v) - \phi(x', v')| \leq C|x - x'| + \rho(|v - v'|) \quad \forall x, x' \in \mathbb{R}^N, \quad \forall v, v' \in V, \quad \forall \phi = \sigma_{ij}, g, c.$$

$$(1.6) \quad |\phi(x, v)| \leq C \quad \forall \phi = \sigma_{ij}, g, c, f, \quad \forall x \in \mathbb{R}^N, \quad \forall v \in V.$$

$$(1.8) \quad |f(x, v) - f(x', v')| \leq \rho(|x - x'| + |v - v'|) \quad \forall x, x' \in \mathbb{R}^N, \quad \forall v, v' \in V.$$

$$(1.9) \quad c(x, v) \geq c_0 \geq 0 \quad \forall x \in \mathbb{R}^N, \quad \forall v \in V.$$

We shall denote by B_s the set of bounded functions from $\bar{\mathcal{O}}$ into \mathbb{R} which are upper semicontinuous; B_s is a closed convex cone of the Banach space B of bounded measurable functions equipped with the supremum norm ($\|h\|_\infty = \sup |h(x)|$).

2. A nonlinear semigroup

2.1. The semigroup property. In this section we prove that $Q(t)$ acting on B_s is a nonlinear semigroup. This result generalizes [23] (cf. also [1]), where $\mathcal{O} = \mathbb{R}^N$. We need, in addition to (1.5-6-8-9), a technical assumption: the set of regular points is closed, i.e.,

$$(2.1) \quad \begin{aligned} &\forall \mathcal{A}, \text{ admissible } \Gamma_0(\mathcal{A}) = \{x \in \Gamma / P(\tau_x > 0) = 0\} \text{ is closed,} \\ &\forall x \in \bar{\mathcal{O}}, P[y_x(\tau_x) \in \Gamma_0(\mathcal{A})] = 1. \end{aligned}$$

We shall see below that in the nondegenerate case this assumption becomes obvious, and that in many cases one can give conditions for (2.1) to be satisfied.

THEOREM 2.1. Assume (1.5-6-8-9) and (2.1). Then $(Q(t), t \geq 0)$ satisfies:

$$(2.2) \quad Q(t) : B_s \rightarrow B_s, \quad Q(0) = I, \quad Q(t+s) = Q(t) \circ Q(s) = Q(s) \circ Q(t),$$

$$(2.3) \quad \|Q(t)h - Q(s)h\|_\infty \rightarrow 0 \text{ as } t \rightarrow s \text{ if } h \text{ is uniformly continuous on } \bar{\mathcal{O}},$$

$$(2.4) \quad \|Q(t)h_1 - Q(t)h_2\|_\infty \leq \|h_1 - h_2\|_\infty \quad \forall h_1, h_2 \in B_s, \quad \forall t \geq 0,$$

(2.5) $Q(t)h_1 \leq Q(t)h_2$ if $h_1 \leq h_2$.

Remark 2.1. We shall see below that, in the case of nondegenerate σ , $Q(t)$ leaves $C_b(\bar{O})$ invariant.

Remark 2.2. Let us give a heuristic justification of Theorem 2.1. By the dynamical programming argument $h(t) = Q(t)h$ is the “solution” of

$$\begin{aligned} \frac{dh}{dt}(s) + \sup_{v \in V} \{A(v)h(s, x) - f(x, v)\} &= 0 \quad \forall s \in [0, t], \quad \forall x \in O, \\ h(0) &= h, \quad h(s)|_{\Gamma_0} = h|_{\Gamma_0} \quad \forall s, \end{aligned}$$

where² $A(v) = -a_{ij} \partial^2 / \partial x_i \partial x_j + b_i \partial / \partial x_i + c$ and $a_{ij}(x, v) = \frac{1}{2} \sigma_{ik} \sigma_{jk}(x, v)$, $b_i(x, v) = -g_i(x, v)$.

Now (2.2) appears as a classical result for some Cauchy problem, and (2.4) and (2.5) are easy consequences of the maximum principle.

The proof will be divided in several parts. First we prove some lemmas.

LEMMA 2.1. *For all $h \in B_s$, we have*

$$(2.6) \quad Q(t)h(x) = \inf_{\mathcal{A}_{cl}} J(x, \mathcal{A}_{cl}, t, h) \quad (\text{resp.} = \inf_{\mathcal{A}_c} J(x, \mathcal{A}_c, t, h)),$$

where the infimum is taken over all admissible systems such that $v(t)$ is right continuous with left-hand limits (resp. is continuous).

Proof. Let \mathcal{A} be an admissible system. We define

$$(2.7) \quad v_k(t) = \frac{1}{k} \int_{(t-k)^+}^t v(\lambda) d\lambda + \left(1 - \frac{t}{k}\right)^+ v_0 \quad (\text{with } v_0 \in V)$$

and let \mathcal{A}_k be the same system as \mathcal{A} with $v(t)$ replaced by $v_k(t)$. Assuming Lemma 2.2 below for the moment,

$$J(x, \mathcal{A}_k, t, h) \rightarrow J(x, \mathcal{A}, t, h) \quad \text{as } k \rightarrow 0^+, \quad \forall h \in C_b(\bar{O}).$$

Thus the equality (2.6) is proved if h is continuous. But if $h \in B_s$, there exists $h_n \in C_b(\bar{O})$, $h_n(x) \downarrow h(x) \forall x \in \bar{O}$. As (2.6) is true for h_n and $Q(t)h_n(x) \downarrow Q(t)h(x)$, $\inf_{\mathcal{A}_{cl}} J(x, \mathcal{A}_{cl}, t, h_n) \downarrow \inf_{\mathcal{A}_{cl}} J(x, \mathcal{A}_{cl}, t, h)$ and $\inf_{\mathcal{A}_c} J(x, \mathcal{A}_c, t, h) \downarrow \inf_{\mathcal{A}_{cl}} J(x, \mathcal{A}, t, h)$, we deduce (2.6) for h .

LEMMA 2.2. *Let \mathcal{A} be an admissible system and let \mathcal{A}_k be the system defined above. We have*

$$\lim_{k \rightarrow 0^+} J(x, \mathcal{A}_k, t, h) = J(x, \mathcal{A}, t, h) \quad \forall h \in C_b(\bar{O}), \quad \forall x \in \bar{O}, \quad \forall t \geq 0.$$

Proof. Letting $y_k(t)$ be the solution of (1.4) corresponding to $v_k(t)$, we have

$$y_k(t) - y(t) = \int_0^t \{\sigma(y_k, v_k) - \sigma(y, v)\} dW_s + \int_0^t (g(y_k, v_k) - g(y, v)) ds.$$

Thus for all $0 \leq t \leq T$ there exists a C_T such that

$$E\{|y_k(t) - y(t)|^2\} \leq C_T E\left\{\int_0^t |y_k - y|^2 + \rho^2(|v_k - v|) ds\right\}.$$

² We shall always use the usual convention for sums.

By Gronwall's lemma and by a classical martingale technique, we deduce

$$(2.8) \quad E\left\{\sup_{0 \leq t \leq T} |y_k(t) - y(t)|^2\right\} \leq C_T^2 E\left\{\int_0^T \rho^2(|v_k - v|) ds\right\}.$$

But there is a $V_0 \subset V$, V_0 compact, such that $v(t, \omega) \in V_0$; thus $v_k(t, \omega) \in \text{conv}(V_0, v_0)$, which is also compact. Now $v_k \rightarrow v$ a.e. (t, ω) , and this implies

$$E\left\{\int_0^T \rho^2(|v_k - v|) ds\right\} \rightarrow 0 \quad \text{as } k \rightarrow 0_+;$$

from (2.8) we have

$$(2.8') \quad \lim_{k \rightarrow 0_+} E\left\{\sup_{0 \leq t \leq T} |y_k(t) - y(t)|^2\right\} = 0.$$

Finally, as in the proof of the Lemma 2.3 below, we have

$$(2.9) \quad \lim_{k \rightarrow 0_+} P\{|T \wedge \tau_k - T \wedge \tau| \geq \varepsilon\} = 0 \quad \forall \varepsilon > 0,$$

where τ_k is the exit time corresponding to the process $y_k(t)$; because of (2.8') we can extract a subsequence y_{k_n} , τ_{k_n} such that

$$y_{k_n}(t) \rightarrow y(t) \quad \text{in } C([0, T], \mathbb{R}^N) \quad \text{a.s.,}$$

$$T \wedge \tau_{k_n} \rightarrow T \wedge \tau \quad \text{a.s.}$$

Thus by the Lebesgue theorem we have proved the lemma. \square

LEMMA 2.3. *We have all admissible systems*

$$\lim_{\substack{x \rightarrow x_0 \\ x \in \bar{\mathcal{O}}}} P\{|T \wedge \tau_x - T \wedge \tau_{x_0}| \geq \varepsilon\} = 0 \quad \forall x_0 \in \bar{\mathcal{O}}, \quad \forall \varepsilon > 0, \quad \forall T > 0.$$

Proof. We define $\tau' = \tau'_x = \inf(t \geq 0, y_x(t) \notin \bar{\mathcal{O}} - \Gamma_0)$ and $N_x^T = \{\omega \in \Omega / \tau_x < T, y_x(\tau_x) \notin \Gamma_0\}$. By assumption (2.1), we have

$$(2.10) \quad P(N_x^T) = 0 \quad \forall x \in \bar{\mathcal{O}}, \quad \forall T > 0,$$

$$(2.11) \quad T \wedge \tau_x(\omega) = T \wedge \tau'_x(\omega) \quad \forall \omega \in \Omega - N_x^T.$$

The lemma is proved if we show that, for all $x_n \rightarrow x_0$ in $\bar{\mathcal{O}}$,

$$(2.12) \quad \begin{aligned} A &= \{\omega \in \Omega / \lim_n |T \wedge \tau_{x_n}(\omega)| > 0\} \\ &\subset B = \left(\bigcup_{n=1}^{\infty} N_{x_n}^T \right) \cup \{\omega \in \Omega / \lim_n \sup_{0 \leq t \leq T} |y_{x_n}(t, \omega) - y_{x_0}(t, \omega)| > 0\}, \end{aligned}$$

since from (2.10) and (2.8') (same proof) $P(B) = 0$.

In order to show (2.12), let $\omega \notin B$. First we prove $\lim_n T \wedge \tau_{x_n}(\omega) \leq T \wedge \tau_{x_0}(\omega)$. We can suppose $\tau_{x_0} < T$: For all $\delta > 0$ there is a $s_\delta < \tau_{x_0}(\omega) + \delta$ such that $y_{x_0}(s_\delta, \omega) \notin \bar{\mathcal{O}}$; hence $y_{x_n}(s_\delta, \omega) \notin \bar{\mathcal{O}}$ if n is large enough and $\tau_{x_n}(\omega) \leq s_\delta \leq \tau_{x_0}(\omega) + \delta$.

Next we prove $\lim_n T \wedge \tau'_{x_n}(\omega) \geq T \wedge \tau'_{x_0}(\omega)$. We may suppose $\tau'_{x_0}(\omega) > 0$, and we define, for $0 < \delta < \tau'_{x_0}(\omega)$, $K_\omega = \{y_{x_0}(t, \omega) / t \in [0, \tau'_{x_0}(\omega) - \delta]\}$. K_ω is a compact set such that $K_\omega \cap \Gamma_0 = \emptyset$. Now, by the choice of ω , we obtain for n large enough

$$K_\omega^n = \{y_{x_n}(t, \omega) / t \in [0, \tau'_{x_0}(\omega) - \delta]\} \cap \Gamma_0 = \emptyset,$$

and this implies $\tau'_{x_n}(\omega) \geq \tau'_{x_0}(\omega) = \delta$ for n large enough. \square

Proof of Theorem 2.1. We remark first that properties (2.4), (2.5) are immediate. The steps of the proof are the following:

- i) $Q(t)h \in B_s$ if $h \in B_s$.
- ii) Proof of (2.3).
- iii) $Q(t+s) = Q(t) \circ Q(s)$.

i) We begin by proving that if $h \in C_b(\bar{\mathcal{O}})$ then $Q(t)h \in B_s$. Indeed, Lemmas 2.2 and 2.3 imply that $J(x, \mathcal{A}, t, h) \in C_b(\bar{\mathcal{O}})$; thus

$$Q(t)h = \inf J(x, \mathcal{A}, t, h) \in B_s.$$

Furthermore, if $h \in B_s$, there exists $h_n \in C_b(\bar{\mathcal{O}})$, $h_n(x) \downarrow h(x)$ for all $x \in \bar{\mathcal{O}}$; therefore $Q(t)h_n(x) \downarrow Q(t)h(x)$ and $Q(t)h \in B_s$.

(ii) To prove (2.3), it is enough to prove that for all uniformly continuous

$$\sup_{\mathcal{A}} E\{|h(y_x(t \wedge \tau_x)) - h(y_x(s \wedge \tau_x))|\} \rightarrow 0 \quad (\text{as } t \rightarrow s) \text{ uniformly in } x.$$

First, remark we have $E\{|y_x(t \wedge \tau_x) - y_x(s \wedge \tau_x)|^2\} \leq C|t-s|$ (C is independent of \mathcal{A} and x); thus

$$P[|y_x(t \wedge \tau_x) - y_x(s \wedge \tau_x)| \geq \varepsilon] \leq C \frac{|t-s|}{\varepsilon^2} \quad \forall \varepsilon > 0.$$

Let $\mu > 0$. Then $\exists \varepsilon, \forall x, x' \in \bar{\mathcal{O}}, |x - x'| \leq \varepsilon \Rightarrow |h(x) - h(x')| \leq \mu$. We have

$$\sup_{\mathcal{A}} E\{|h(y_x(t \wedge \tau_x)) - h(y_x(s \wedge \tau_x))|\} \leq \frac{C\|h\|_{\infty}|t-s|}{\varepsilon^2} + \mu,$$

and the conclusion follows easily

iii) We want to prove the semigroup property $Q(t+s) = Q(t) \circ Q(s)$. Because of Lemma 2.1, we can restrict ourselves to admissible systems with continuous $v(t)$. We can also restrict our attention to admissible systems where (Ω, F, F_t) is the canonical space $\Omega = C([0, +\infty[, \mathbb{R}^{n+m})$ (just take image measures). But at this point the proof of this property is exactly the same as the one given in [2, Thm. 5.1]. The proof depends heavily on a theorem of regular conditional probabilities proved by D. W. Stroock–S. R. S. Varadhan [25] and N. V. Krylov [11]. \square

2.2. The generator of $Q(t)$. We are going to prove that the “generator” of $Q(t)$ is an extension of the operator $\phi \in C^2(\bar{\mathcal{O}}) \rightarrow \sup_{v \in V} \{A(v)\phi(x) - f(x, v)\}$.

THEOREM 2.2. *Under the assumptions of Theorem 2.1, we have for all $h \in C_b^2(\bar{\mathcal{O}})$*

$$(2.13) \quad \frac{1}{t} \{Q(t)h(x) - h(x)\} \rightarrow -\sup_{v \in V} \{A(v)h(x) - f(x, v)\} \quad \text{as } t \rightarrow 0_+ \quad \forall x \in \bar{\mathcal{O}}.$$

Moreover the convergence in (2.13) is uniform on compact subsets of $\bar{\mathcal{O}}$.

Proof. The proof is very similar to the proof of M. Nisio [23] (see also the presentation in [2, Thm. 5.2]). We define

$$K(x, \mathcal{A}, t, h) = \int_0^{t \wedge \tau_x} f(y_x(s), v(s)) - A(v(s)) h(y_x(s)) ds,$$

and we prove easily (see for example [1]) that

$$\begin{aligned} \forall \varepsilon > 0, \quad \exists \delta = \delta(\varepsilon, h) > 0, \quad \forall t \leq \delta, \quad \left| \frac{Q(t) h(x) - h(x)}{t} - \inf_{\mathcal{A}} E \left\{ \frac{1}{t} K(x, \mathcal{A}, t, h) \right\} \right| \leq \varepsilon, \\ \inf_{\mathcal{A}} E \left\{ \frac{1}{t} K(x, \mathcal{A}, t, h) - \inf_{v \in V} [f(x, v) - A(v) h(x)] \right\} \geq -C \left(1 - \inf_{\mathcal{A}} E \left\{ \frac{t \wedge \tau_x}{t} \right\} \right). \end{aligned}$$

On the other hand, if \mathcal{A}_0 is an admissible system corresponding to $v(t) = v_0 \in V$,

$$\begin{aligned} & \inf_{\mathcal{A}} E \left\{ \frac{1}{t} K(x, \mathcal{A}, t, h) - \inf_{v \in V} [f(x, v) - A(v) h(x)] \right\} \\ & \leq E \left\{ \frac{1}{t} K(x, \mathcal{A}_0, t, h) - \inf_{v \in V} [f(x, v) - A(v) h(x)] \right\} \\ & \leq C \left(1 - E \left\{ \frac{t \wedge \tau_x}{t} \right\} \right) \\ & \leq C \left(1 - \inf_{\mathcal{A}} E \left(\frac{t \wedge \tau_x}{t} \right) \right). \end{aligned}$$

Thus we have obtained

$$\begin{aligned} \forall t \leq \delta, \quad \left| \frac{Q(t) h(x) - h(x)}{t} - \inf_{v \in V} [f(x, v) - A(v) h(x)] \right| \\ \leq C \left\{ 1 - \inf_{\mathcal{A}} E \left(\frac{t \wedge \tau_x}{t} \right) \right\} + \varepsilon. \end{aligned}$$

To conclude, we just need to prove that if K is a compact subset of \mathcal{O} then

$$\sup_{\mathcal{A}, x \in K} P(\tau_x < t) \xrightarrow[t \rightarrow 0]{} 0.$$

Letting γ be $\gamma = d(K, \Gamma) > 0$, we have

$$\forall x \in \bar{\mathcal{O}}, \quad P[\tau_x < t] \leq P\left(\sup_{0 \leq s \leq t} |y_x(s) - x| \geq \gamma\right) \leq \frac{1}{\gamma^2} E \left\{ \sup_{0 \leq s \leq t} |y_x(s) - x|^2 \right\}.$$

Since $E\{\sup_{0 \leq s \leq t} |y_x(s) - x|^2\} \leq CE|y_x(t) - x|^2 \leq C_1 t + C_2 t^2$, where C, C_1, C_2 do not depend on \mathcal{A}, x and t , (2.13) is easily proved. \square

Remark 2.3. If we introduce

$$\Gamma_1 = \left\{ x \in \Gamma \limsup_{\varepsilon \rightarrow 0+} \sup_{\mathcal{A}} E \left(\frac{\varepsilon \wedge \tau_x}{\varepsilon} \right) = 0 \right\}, \quad \Gamma_2 = \left\{ x \in \Gamma \liminf_{\varepsilon \rightarrow 0} \inf_{\mathcal{A}} E \left(\frac{\varepsilon \wedge \tau_x}{\varepsilon} \right) = 1 \right\},$$

for $h \in C_b^2(\bar{\mathcal{O}})$ we have, as $t \rightarrow 0_+$,

$$\begin{aligned} \text{i)} \quad & \frac{Q(t)h(x) - h(x)}{t} \rightarrow 0 \quad \text{if } x \in \Gamma_1, \\ \text{ii)} \quad & \frac{Q(t)h(x) - h(x)}{t} \rightarrow -\sup_{v \in V} \{A(v)h(x) - f(x, v)\} \quad \text{if } x \in \Gamma_2. \end{aligned}$$

Remark that $\Gamma_0 \subset \Gamma_1$.

Remark 2.4. In the particular case of nondegeneracy, i.e.,

$$(2.14) \quad \exists \alpha > 0, \quad a_{ij}(x, v) \xi_i \xi_j \geq \alpha |\xi|^2 \quad \forall \xi \in \mathbb{R}^N, \quad \forall x \in \bar{\mathcal{O}}, \quad \forall v \in V,$$

we shall see that $\Gamma_0(\mathcal{A}) = \Gamma$ for all admissible systems (if some regularity condition on Γ is assumed); hence, for all $x \in \bar{\mathcal{O}}$, as $t \rightarrow 0$

$$\frac{Q(t)h(x) - h(x)}{t} \rightarrow -1_{\mathcal{O}}(x) \sup_{v \in V} \{A(v)h(x) - f(x, v)\}.$$

Remark 2.5. We shall see below a result more precise than Theorem 2.2.

2.3. The nondegenerate case. In this section in addition to (1.5-6-8-9), we assume (2.14) and \mathcal{O} has a uniform exterior sphere; i.e.,

$$(2.15) \quad \exists \rho > 0, \quad \forall x \in \Gamma, \quad \exists y \in \mathbb{R}^N - \mathcal{O}, \quad \{z/|y-z| \leq \rho\} \cap \bar{\mathcal{O}} = \{x\}.$$

We are going to prove that under these assumptions $Q(t)$ leaves X invariant, where $X = \{h \in C_b(\bar{\mathcal{O}}), h \text{ is uniformly continuous on } \bar{\mathcal{O}}\}$. Before doing so or even stating the precise result, we prove a lemma which will be useful.

LEMMA 2.4. Under assumptions (1.5-6-8-9) and (2.14-15), we have:

$$(2.16) \quad \text{If } \mathcal{O} \text{ is bounded, } \exists \mu > 0, \exists C > 0, \forall x \in \mathcal{O}, \forall \mathcal{A} \text{ admissible, } E[e^{\mu \tau_x}] \leq C;$$

$$(2.17) \quad \forall \mathcal{A} \text{ admissible, } \Gamma = \Gamma_0(\mathcal{A}).$$

Remark 2.6. It is clear that even if (2.14) is satisfied, \mathcal{O} has to be “smooth” in order to make (2.17) true. Indeed, if $N = 1$, $V = \{v_0\}$, $y_x(t) = x + W(t)$, $\sigma(v_0) = \sqrt{2}$, $\mathcal{O} =]0, 1[\cup]1, 2[$, we have $E[\tau_1] = \frac{1}{2}$, so $1 \in \Gamma - \Gamma_0$.

Proof of Lemma 2.4. First we consider $w(x) = 1 - \exp(-k|x|^2)$ (we may always assume that $0 \in \bar{\mathcal{O}}$). We have $A(v)w(x) \geq \{4a_{ii}(x, v)k^2x_ix_i - 2ka_{ii}(x, v) - 2kx_ib_i(x, v)\} \exp(-k|x|^2)$. Thus we can choose k large enough to insure that $A(v)w(x) \geq \alpha > 0$ for all $x \in \bar{\mathcal{O}}$ (because \mathcal{O} is bounded), where $\tilde{A} = A - c$.

Now we take $\mu = \alpha/2$, and we have

$$(2.18) \quad \tilde{A}(v)w - \mu w \geq \mu > 0 \quad \forall x \in \bar{\mathcal{O}}.$$

Using Ito's formula with w , it is easy to deduce (2.16) from (2.18).

Now we prove (2.17). We introduce

$$(2.15') \quad w(x, \xi) = \exp(-k\rho^2) - \exp(-k|x - \xi_1|^2),$$

where ρ is given by (2.15), $\xi \in \Gamma$ and ξ_1 is associated to ξ by (2.15), $x \in \bar{\mathcal{O}}$ and $k > 0$. By calculation similar to the above, one shows that for k large enough

$$(2.19) \quad A(v)w(x, \xi) \geq \alpha > 0 \quad \forall x \in \mathcal{O}.$$

Applying Ito's formula, we have

$$\begin{aligned} 0 = w(\xi, \xi) &= E \left\{ w(y_\xi(\tau_\xi)) + \int_0^{\tau_\xi} \alpha \exp \left(\int_0^t c(y_\xi(s), v(s)) ds \right) dt \right\} \\ &\cong \alpha E \left[\int_0^{\tau_\xi} e^{-ct} dt \right]; \end{aligned}$$

thus $P[\tau_\xi = 0] = 1$ and $\xi \in \Gamma_0(\mathcal{A})$ for all $\xi \in \Gamma$. \square

The first result concerning the regularity of $Q(t)h$ when h is smooth will be the following.

THEOREM 2.3. *We assume (1.5-6-8), (2.14-15) and*

$$(2.20) \quad |f(x, v) - f(x', v)| \leq C|x - x'| \quad \forall x, x' \in \mathcal{O},$$

$$(2.21) \quad c(x, v) \geq C > [\mu_0]^+,$$

where μ_0 is given by

$$(2.22) \quad \mu_0 = \sup_{\substack{x, x' \in \mathcal{O} \\ v \in V}} \left\{ \frac{1}{2} \text{Tr} \frac{(\sigma(x, v) - \sigma(x', v))(\sigma^T(x, v) - \sigma^T(x', v))}{|x - x'|^2} + \frac{(x - x') \cdot (g(x) - g(x'))}{|x - x'|^2} \right\}.$$

Then, if $h \in W^{2,\infty}(\mathcal{O})$, we have

$$(2.23) \quad |Q(t)h(x) - Q(t)h(x')| \leq C|x - x'| \quad \forall x, x' \in \mathcal{O},$$

where C is independent of t .

COROLLARY 2.1. *If we assume (1.5-6-8-9) and (2.14-15-21) then, for $h \in X$, $Q(t)h \in X$. Furthermore, $(Q(t)h, t \geq 0)$ is uniformly equicontinuous.*

Proof of Corollary 2.1. By a simple approximation (uniform in v) of the function $f(v)$, one can always assume that (2.20) is satisfied and that h belongs to $W^{2,\infty}(\mathcal{O})$; then the result is obvious in view of Theorem 2.3. \square

Remark 2.7. We shall see below (§ 3.1, Remark 3.5) that Corollary 2.1 is valid without assuming (2.21), and (§ 4.3) that Theorem 2.3 remains true without assuming (2.21).

Remark 2.8. If assumptions (2.14-15) are dropped, one can nevertheless prove Theorem 2.3 (and thus Corollary 2.1) with the same method if we assume

$$(2.24) \quad \begin{aligned} &\exists p_0 \in W^{1,\infty}(\mathcal{O}), \quad p_0|_{\Gamma_0} = 0, \quad \forall v \in V, \quad A(v)p_0 \in L^\infty(\mathcal{O}), \\ &\exists \alpha_0 > 0, \quad \forall v \in V, \quad A(v)p_0 \leq -\alpha_0 \quad \text{in } \mathcal{O}. \end{aligned}$$

For example suppose that $g = c = 0$, $\sigma(x, v) = \sigma(v)$ and that there exists $\beta_0 > 0$ such that $\det(\sigma(v)\sigma^T(v)) \geq \beta_0 > 0$. Furthermore, assume that $\mathcal{O} = \{p(x) < 0\}$ with $\partial\mathcal{O} = \{p(x) = 0\}$ and that $p \in W^{2,\infty}(\mathcal{O})$ and

$$\det \left(\frac{\partial^2 p}{\partial x_i \partial x_j} (x) \right) \geq \alpha_0 > 0 \quad \forall x \in \bar{\mathcal{O}}.$$

Then the results above remain true. This example generalizes a result of B. Gaveau [9].

Other generalizations to the case of degenerate σ are treated in Part II.

Remark 2.9. One can generalize Corollary 2.1 to the case where $\sup_{v \in V} |f(x, v)| \in L^N(\mathcal{O})$. Indeed, this comes easily from a result of N. V. Krylov [13].

Proof of Theorem 2.3. The proof is divided into several steps:

- 1) Construction of a subsolution.
- 2) Two lemmas.
- 3) Conclusion.

1) We consider the function $w(x, \xi)$ defined in Lemma 2.4, and we introduce $w(x) = \inf_{\xi \in \Gamma} w(x, \xi)$. Obviously $w(x) \in W^{1,\infty}(\mathcal{O})$, $w \geq 0$ in \mathcal{O} , $w = 0$ on Γ . Now applying Ito's formula to $w(x, \xi)$ for fixed ξ in Γ , we have (in the proof of this theorem, we shall take $c(x, v) \equiv c_0 > \mu_0$ for the sake of simplicity) that

$$w(y_x(t \wedge \tau_x), \xi) e^{-c_0 t \wedge \tau_x} + \alpha \int_0^{t \wedge \tau_x} e^{-c_0 s} ds$$

is a submartingale bounded and continuous.

Then, taking the infimum over all ξ in Γ , we have that

$$(2.25) \quad w(y_x(t \wedge \tau_x)) e^{-c_0 t \wedge \tau_x} + \alpha \int_0^{t \wedge \tau_x} e^{-c_0 s} ds$$

is a submartingale bounded and continuous.

2) LEMMA 2.5. *Under the assumptions of Theorem 2.3, we have*

$$(2.26) \quad E[|e^{-c_0 \tau_x} - e^{-c_0 \tau_{x'}}|] \leq \frac{2C_0}{\alpha} \|\nabla w\|_{\infty} |x - x'|.$$

Proof. Applying (2.25) between $\tau_x \wedge \tau_{x'}$ and τ_x , we have

$$E[w(y_x(\tau_x)) e^{-c_0 \tau_x} - w(y_x(\tau_x \wedge \tau_{x'})) e^{-c_0 \tau_x \wedge \tau_{x'}}] \geq -\alpha E\left[\int_{\tau_x \wedge \tau_{x'}}^{\tau_x} e^{-c_0 s} ds\right];$$

thus

$$\frac{\alpha}{c_0} E[e^{-c_0 \tau_x \wedge \tau_{x'}} - e^{-c_0 \tau_x}] \leq \|\nabla w\|_{\infty} E\{|y_x(\tau_x \wedge \tau_{x'}) - y_{x'}(\tau_x \wedge \tau_{x'})| e^{-c_0 \tau_x \wedge \tau_{x'}}}$$

and we deduce (2.26) from the following lemma. \square

LEMMA 2.6. *Under the assumptions of Theorem 2.3, we have for all stopping times θ*

$$(2.27) \quad E\{|y_x(\theta) - y_{x'}(\theta)|^2 e^{-2\mu_0 \theta}\} \leq |x - x'|^2.$$

Proof. We apply Ito's formula between 0 and $\theta \wedge T$ to the function $(\xi \rightarrow |\xi|^2)$ for the process $y_x(t) - y_{x'}(t)$, and obtain

$$\begin{aligned} & E\{|y_x(\theta \wedge T) - y_{x'}(\theta \wedge T)|^2 e^{-2\mu_0 \theta \wedge T}\} \\ &= |x - x'|^2 + E\left\{\int_0^{\theta \wedge T} \text{Tr}\{(\sigma(y_x(t)) - \sigma(y_{x'}(t))) \cdot (\sigma^T(y_x(t)) - \sigma^T(y_{x'}(t)))\} e^{-2\mu_0 t} \right. \\ &\quad \left. + 2(y_x(t) - y_{x'}(t)) \cdot (g(y_x(t)) - g(y_{x'}(t))) e^{-2\mu_0 t} dt \right. \\ &\quad \left. - 2\mu_0 \int_0^{\theta \wedge T} |y_x(t) - y_{x'}(t)|^2 e^{-2\mu_0 t} dt\right\}. \end{aligned}$$

Thus, by definition of μ_0 , we have

$$E\{|y_x(\theta \wedge T) - y_{x'}(\theta \wedge T)|^2 e^{-2\mu_0 \theta \wedge T}\} \leq |x - x'|^2.$$

3) *Conclusion.* Letting $x, x' \in \bar{\mathcal{O}}$, we have

$$|Q(t) h(x) - Q(t) h(x')| \leq I + J,$$

where

$$I = \sup_{\mathcal{A}} \left| E \left[\int_0^{\tau_x \wedge t} f(y_x(s), v(s)) e^{-c_0 s} ds \right] - E \left[\int_0^{\tau_{x'} \wedge t} f(y_{x'}(s), v(s)) e^{-c_0 s} ds \right] \right|$$

and

$$J = \sup_{\mathcal{A}} |E[h(y_x(\tau_x \wedge t)) e^{-c_0 \tau_x \wedge t} - h(y_{x'}(\tau_{x'} \wedge t)) e^{-c_0 \tau_{x'} \wedge t}]|.$$

First, because of Lemma 2.5 and (2.20), we easily have $I \leq C|x - x'|$.

Next,

$$\begin{aligned} J &\leq \sup_{\mathcal{A}} \{ |E\{h(y_x(t \wedge \tau_x)) e^{-c_0 t \wedge \tau_x} - h(y_x(t \wedge \tau_x \wedge \tau_{x'})) e^{-c_0 t \wedge \tau_x \wedge \tau_{x'}}\} \\ &\quad + |E\{h(y_{x'}(t \wedge \tau_{x'})) e^{-c_0 t \wedge \tau_{x'}} - h(y_{x'}(t \wedge \tau_x \wedge \tau_{x'})) e^{-c_0 t \wedge \tau_x \wedge \tau_{x'}}\}| \\ &\quad + |E\{h(y_x(t \wedge \tau_x \wedge \tau_{x'})) - h(y_{x'}(t \wedge \tau_x \wedge \tau_{x'}))| e^{-c_0 t \wedge \tau_x \wedge \tau_{x'}}\} \} \\ &\leq \sup_{v \in V} \|A(v) h\|_{\infty} \cdot \frac{2c_0}{\alpha} \|\nabla w\|_{\infty} |x - x'| + \|\nabla h\|_{\infty} |x - x'| \end{aligned}$$

(here we have applied Ito's formula and (2.26), (2.27)). \square

3. A stochastic interpretation of the minimum cost function

3.1. A stochastic control problem. We consider the optimal cost function

$$(3.1) \quad u(x) = \inf_{\mathcal{A}} E \left\{ \int_0^{\tau_x} f(y_x(t), v(t)) \exp \left(- \int_0^t c(y_x(s), v(s)) ds \right) dy \right\}.$$

We have the following;

THEOREM 3.1. *Under assumptions (1.5-6-8), (2.1) and*

$$(3.2) \quad c(x, v) \geq c_0 > 0 \quad \forall x \in \bar{\mathcal{O}}, \quad \forall v \in V,$$

or under assumptions (1.5-6-8-9), (2.14-15) if \mathcal{O} is bounded (the nondegenerate case), we have

$$(3.3) \quad u(x) = \lim_{t \rightarrow \infty} Q(t) h(x) \quad \text{in } B_s \quad \forall h \in B_s, h|_{\cup \Gamma_0(\mathcal{A})} = 0$$

(in the nondegenerate case $\forall h|_{\Gamma} = 0$),

$$u \in B_s, \quad Q(t)u = u \quad \forall t \geq 0.$$

Furthermore the equation of dynamical programming is satisfied:

$$(3.5) \quad \begin{aligned} u(x) = \inf_{\mathcal{A}} E \left\{ \int_0^{\theta \wedge \tau_x} f(y_x(t), v(t)) \exp \left(- \int_0^t c(y_x(s), v(s)) ds \right) dt \right. \\ \left. + u(y_x(\theta \wedge \tau_x)) \exp \left(- \int_0^{\theta \wedge \tau_x} c(y_x(t), v(t)) dt \right) \right\}, \end{aligned}$$

where θ is a stopping time with respect to F^t .

Finally, if $\Gamma_0(\mathcal{A})$ is independent of \mathcal{A} , $\Gamma_0(\mathcal{A}) = \Gamma_0$ for all \mathcal{A} admissible (in the nondegenerate case $\Gamma_0 = \Gamma$), then $u(x)$ is the unique solution of

$$(3.6) \quad u \in B_s, u|_{\Gamma_0} = 0, Q(t)u = u \quad \forall t \geq 0.$$

Remark 3.1. Equality (3.5) shows that the optimal cost function $u(x)$ satisfies in some general integral sense the Bellman equation: $\sup_{v \in V} \{A(v)u - f(v)\} = 0$ in \mathcal{O} .

Remark 3.2. i) If for all x and for all v , $f(x, v) \geq 0$ and $\Gamma_1 = \bigcup \Gamma_0(\mathcal{A})$, then it is easy to prove, by the same methods as those which follow, that $u(x)$ is the unique solution of

$$(3.6') \quad u \in B_s, u|_{\Gamma_1} = 0, Q(t)u = u \quad \forall t \geq 0.$$

Such a case will be considered in Part II.

ii) If we assume that for each \mathcal{A} , $\Gamma_0(\mathcal{A}) = \Gamma_0$, where Γ_0 is closed in Γ , then we can prove that $P[y_x(\tau_x) \in \Gamma_0] = 1$ for all $x \in \bar{\mathcal{O}}$.

COROLLARY 3.1. Under assumptions (1.5–6), (2.14–15–20–21), the optimal cost function belongs to $W_0^{1,\infty}(\mathcal{O})$.

Proof. Since $u(x) = \lim_{t \rightarrow \infty} Q(t)0(x)$ in B_s , and by Theorem 2.3 we have $|Q(t)0(x) - Q(t)0(x')| \leq C|x - x'|$, where C is independent of t , the result is immediate. \square

Remark 3.3. If we define (cf. Dynkin [5]) the closed subset B_0 of B_s ,

$$B_0 = \{h \in B_s | \forall x \in \bar{\mathcal{O}}, Q(t)h(x) \rightarrow Q(s)h(x) \text{ as } t \rightarrow s, h|_{\Gamma_0} = 0\},$$

we can consider instead of (3.6)

$$(3.6'') \quad u \in B_0, Q(t)u = u \quad \forall t \geq 0.$$

Remark 3.4. Let ϕ be given, where ϕ is the trace on Γ of some $\Phi \in B_s$; then we have $u_\phi(x) = Q(\infty)\Phi(x) = Q(\infty)h(x)$, $h \in B_s$ such that $h|_{\Gamma_0} = \phi$ (under the same hypotheses as in Theorem 3.1). Moreover, u_ϕ is the unique solution of the non-homogeneous problem $u_\phi \in B_s$, $u_\phi|_{\Gamma_0} = \phi$, $Q(t)u_\phi = u_\phi$ for all $t \geq 0$ and we also have the corresponding equation of dynamical programming.

Proof of Theorem 3.1. We prove (3.4) only for the case of nondegeneracy (hypothesis (2.14–15)) and (3.5); the other statements are obvious.

1) We know by Lemma 2.4 that there exists some $\mu > 0$ such that (\mathcal{O} is assumed to be bounded)

$$\exists C, \forall x, \forall \mathcal{A}, E[e^{\mu\tau_x}] \leq C;$$

thus

$$|Q(t)h(x) - u(x)| \leq \sup_{\mathcal{A}} E \left[\int_{t \wedge \tau_x}^{\tau_x} \sup_{v \in V} \|f(x, v)\|_\infty ds \right] + \sup_{\mathcal{A}} E[\|h\|_\infty 1_{(t < \tau_x)}].$$

But $\sup_{\mathcal{A}} P[\tau_x > t] \leq C e^{-\mu t}$ and $\sup_{\mathcal{A}} E[\tau_x - t \wedge \tau_x] \leq \sup_{\mathcal{A}} E[\tau_x 1_{(t < \tau_x)}] \leq C' e^{-\mu t}$.

2) In order to prove (3.5) we need only consider admissible systems such that $v(t)$ is a continuous process (cf. Lemma 2.1). Now we define, for fixed x in $\bar{\mathcal{O}}$,

$$\begin{aligned} \xi(t) = & \int_0^{t \wedge \tau_x} f(y_x(s), v(s)) \exp \left(- \int_0^s c(y_x(\lambda), v(\lambda)) d\lambda \right) dt \\ & + u(y_x(t \wedge \tau_x)) \exp \left(- \int_0^{t \wedge \tau_x} c(y_x(s), v(s)) ds \right). \end{aligned}$$

We want to prove that $\xi(t)$ is a F^t -submartingale satisfying to the property

$$(3.7) \quad \xi(\theta) \leq E\{\xi(\theta+t)/F^\theta\}, \quad \text{where } \theta \text{ is a stopping time and } t \geq 0.$$

But the proof of that fact is exactly the same as in \mathbb{R}^N (cf. [1, Thms. 5.1, 5.3]), from $u|_{\Gamma_0(\mathcal{A})} \leq 0$ and thus $P[u(y_x(\tau_x)) \leq 0] = 1$.

Therefore taking $t \rightarrow +\infty$ in (3.7) we prove that

$$\begin{aligned} E \left[\int_0^{\theta \wedge \tau_x} f(y_x(t), v(t)) \exp \left(- \int_0^t c(y_x(s), v(s)) ds \right) dt \right. \\ \left. + u(y_x(\theta \wedge \tau_x)) \exp \left(- \int_0^{\theta \wedge \tau_x} c(y_x(t), v(t)) dt \right) \right] \\ \leq E \int_0^{\tau_x} f(y_x(t), v(t)) \exp \left(- \int_0^t c(y_x(s), v(s)) ds \right) dt. \end{aligned}$$

To conclude, we have to prove that

$$\begin{aligned} u(x) \leq E \left[\int_0^{\theta \wedge \tau_x} f(y_x(t), v(t)) \exp \left(- \int_0^t c(y_x(s), v(s)) ds \right) dt \right. \\ \left. + u(y_x(\theta \wedge \tau_x)) \exp \left(- \int_0^{\theta \wedge \tau_x} c(y_x(t), v(t)) dt \right) \right]. \end{aligned}$$

But $\xi(t)$ is a submartingale and this inequality is satisfied if θ is replaced by θ_k a discrete approximation of θ such that $\theta_k \rightarrow \theta$ (a.s.) as $k \rightarrow \infty$.

Since u is upper semicontinuous, the inequality remains true for θ . \square

COROLLARY 3.2. *Under the assumptions of Theorem 3.1, we have for all $\lambda \geq 0$*

$$(3.8) \quad \begin{aligned} u(x) = \inf_{\mathcal{A}} E \left[\int_0^{\tau_x} \{f(y_x(t), v(t)) + \lambda u(y_x(t))\} x \right. \\ \left. \cdot \exp \left(- \int_0^t (c(y_x(s), v(s)) + \lambda) ds \right) dt \right]. \end{aligned}$$

Proof. The proof is immediate in view of the following lemma, due to N. V. Krylov [14].

LEMMA 3.1. *Let $z(s)$, $\xi(s)$ be two bounded measurable adapted processes and assume that $z(s) + \int_0^s \xi(r) dr$ is a submartingale. Then for all $\lambda \geq 0$ $z(s) e^{-\lambda s} + \int_0^s (\xi(r) + \lambda z(r)) e^{-\lambda r} dr$ is a submartingale.*

COROLLARY 3.3. *Under assumptions (1.5-6-8-9) and (2.14-15), $u(x)$ belongs to X : $\{h \in C_b(\bar{\mathcal{O}}), h \text{ is uniformly continuous}\}$.*

Proof. If we add the assumption (2.21), then by Corollary 3.1 $u(x) \in X$. Now let $\lambda > 0$ be such that $c(x, v) + \lambda \geq c_0 > \mu_0$ is given by (2.22), and let us consider the following application T defined on B_s : if $v \in B_s$, $w = Tv$ is given by

$$\begin{aligned} w(x) = \inf_{\mathcal{A}} E \left[\int_0^{\tau_x} \{f(y_x(t), v(t)) + \lambda v(y_x(t))\} \right. \\ \left. \cdot \exp \left(- \int_0^t (c(y_x(s), v(s)) + \lambda) ds \right) dt \right]. \end{aligned}$$

Then, by Corollary 3.2, u is a fixed point of T . To conclude, we just need to prove

that T is a strict contraction on B_s . But

$$\|Tv_1 - Tv_2\|_\infty \leq \sup_{\mathcal{A}} E[1 - e^{-\lambda \tau_x}] \|v_1 - v_2\|_\infty$$

and by Jensen's inequality

$$\|Tv_1 - Tv_2\|_\infty \leq (1 - e^{-\lambda C}) \|v_1 - v_2\|_\infty,$$

where $C = \sup_{\mathcal{A}} E[\tau_x] < +\infty$, by Lemma 2.4. \square

Remark 3.5. With the techniques developed above, it is easy to extend Corollary 2.1 to the case where (2.21) is replaced by (1.9) (i.e., $c(x, v) \geq 0$ instead of $c(x, v) \geq c_0 > \mu_0$).

3.2. Application to the generator of $Q(t)$. We now prove a local version of Theorem 2.2, concerning the generator of the nonlinear semigroup $Q(t)$.

THEOREM 3.2. *Under assumptions (1.5-6-8-9) and (2.1), if \mathcal{O}' is a bounded open set included in \mathcal{O} and if $h \in C^2(\mathcal{O}')$, then*

$$\frac{Q(t)h(x) - h(x)}{t} \xrightarrow[t \rightarrow 0]{v \in V} \sup_{v \in V} (A(v)h(x) - f(x, v)) \quad \forall x \in \mathcal{O}'$$

and the convergence is uniform on compact subsets of \mathcal{O}' .

Proof. Let B be an open ball strictly included in \mathcal{O}' . We consider two open balls B_1, B_2 such that $B_2 \subset \bar{B}_2 \subset B_1 \subset \bar{B}_1 \subset B \subset \bar{B} \subset \mathcal{O}'$ and we show the convergence in B_2 . We denote by τ_x^i the exit times of \bar{B}_i , $Q_i(t)$ the corresponding semigroups, $u_i(s, x) = Q_i(t-s)h(x)$ for $0 \leq s \leq t$. First, we remark that

$$(3.9) \quad \begin{aligned} u_t(s, x) = \inf_{\mathcal{A}} E \left\{ \int_0^{\sigma_{x,s}} f(y_x(r), v(r)) \exp \left(- \int_0^r c(y_x(\lambda), v(\lambda)) d\lambda \right) dr \right. \\ \left. + h(y_x(\sigma_{x,s})) \exp \left(- \int_0^{\sigma_{x,s}} c(y_x(r), v(r)) dr \right) \right\}, \end{aligned}$$

where $\sigma_{x,s}$ is the exit time of the set $\mathcal{O} \times]0, t[$ for the $(N+1)$ -dimensional process

$$z_{x,s}(r) = \begin{pmatrix} y_x(r) \\ r+s \end{pmatrix} \quad (r \geq 0).$$

Remark that $\Gamma'_0(\mathcal{A})$ for this process is $\Gamma_0(\mathcal{A})$ and that (2.1) is satisfied. Now by the equation of dynamical programming (3.5) we have

$$(3.9') \quad \begin{aligned} u_t(0, x) = \inf_{\mathcal{A}} E \left\{ \int_0^{\tau_x \wedge \theta \wedge t} f(y_x(s), v(s)) \exp \left(- \int_0^s c(y_x(\lambda), v(\lambda)) d\lambda \right) ds \right. \\ \left. + u_t(\tau_x \wedge \theta \wedge t, y_x(\tau_x \wedge \theta \wedge t)) \exp \left(- \int_0^{\tau_x \wedge \theta \wedge t} c(y_x(s), v(s)) ds \right) \right\}. \end{aligned}$$

Now we take $\theta = \tau_x^1$, and find

$$(3.9'') \quad \begin{aligned} Q(t)h(x) = \inf_{\mathcal{A}} E \left\{ \int_0^{\tau_x^1 \wedge t} f(y_x(s), v(s)) \exp \left(- \int_0^s c(y_x(\lambda), v(\lambda)) d\lambda \right) ds \right. \\ + 1_{(\tau_x^1 < t)} u_t(\tau_x^1, y_x(\tau_x^1)) \exp \left(- \int_0^{\tau_x^1} c(y_x(s), v(s)) ds \right) \\ \left. + 1_{(\tau_x^1 \geq t)} h(y_x(t)) \exp \left(- \int_0^t c(y_x(s), v(s)) ds \right) \right\}. \end{aligned}$$

Thus for all $x \in \bar{B}_2$, as $h \in C^2(\bar{B})$ we have (cf. proof of Theorem 2.2)

$$\begin{aligned} |Q(t)h(x) - h(x)| &\leq |Q(t)h(x) - Q^1(t)h(x)| + |Q^1(t)h(x) - h(x)| \\ &\leq \sup_{\mathcal{A}} E\{|u_t(\tau^1, y_x(\tau^1)) - h(y_x(\tau_x^1))| 1_{(\tau_x^1 < t)}\} + C_1 t \\ &\leq \sup_{0 \leq s \leq t} \|Q(s)h - h\|_{\infty, B_1} \cdot \sup_{\mathcal{A}} P(\tau^1 < t) + C_1 t. \end{aligned}$$

Now, as in the proof of Theorem 2.2, we can show that there exists $C_2 > 0$ such that for all $x \in \bar{B}_2$ $\sup_{\mathcal{A}} P(\tau_x^1 < t) \leq C_2 \sqrt{t}$.

Thus we have finally

$$\sup_{0 \leq s \leq t} \|Q(t)h - h\|_{\infty, \bar{B}_2} \leq C_2 \sqrt{t} \sup_{0 \leq s \leq t} \|Q(s)h - h\|_{\infty, \bar{B}_1} + C_1 t.$$

By a similar argument we have

$$\sup_{0 \leq s \leq t} \|Q(s)h - h\|_{\infty, \bar{B}_1} \leq C_2 \sqrt{t} \sup_{0 \leq s \leq t} \|Q(s)h - h\|_{\infty, \bar{B}} + C_3 t;$$

hence for $t \leq t_0$ we deduce

$$\sup_{0 \leq s \leq t} \|Q(s)h - h\|_{\infty, \bar{B}_2} \leq C_5 t.$$

Finally taking $\theta = \tau_x^2$ in (3.9'), we have

$$\begin{aligned} \forall x \in \bar{B}_2, \quad &\left| \frac{Q(t)h(x) - h(x)}{t} - \frac{Q_2(t)h(x) - h(x)}{t} \right| \\ &\leq \frac{1}{t} \sup_{0 \leq s \leq t} \|Q(s)h - h\|_{\infty, \bar{B}_2} \sup_{\mathcal{A}} P(\tau_x^2 < t) \end{aligned}$$

and we can conclude easily with the help of Theorem 2.2 and remarking that for all $x \in B_3$ a closed set $\subset B_2$, there exists C_6 such that $\sup_{\mathcal{A}} P[\tau_x^2 < t] \leq C_6 \sqrt{t}$. \square

4. Analytical interpretation of the optimal cost function and Hamilton–Jacobi–Bellman equations. In this section we shall always assume (1.5-6-8-9) and (2.14-15-20), i.e., the nondegenerate case, and that \mathcal{O} is a regular domain. In every statement in the following, we shall call this group of hypotheses assumption A.

The main result of this section is the following. Under assumption A, $u \in W_0^{1,\infty}(\mathcal{O})$ and u is the maximum element of the set $\{\tilde{u} \in W_0^{1,\infty}(\mathcal{O}), A(v)\tilde{u} \leq f(v) \text{ in } \mathcal{D}'(\mathcal{O}), \forall v \in V\}$.

We will also recall the main result concerning the solution of

$$(4.1) \quad \sup_{v \in V} \{A(v)u - f(v)\} = 0 \quad \text{a.e. in } \mathcal{O}, \quad u = 0 \text{ on } \Gamma.$$

This result is obtained in L. C. Evans and P.-L. Lions [7] (see also [15]) under more smoothness assumptions on σ, b, c, f and \mathcal{O} than A.

The results which we prove are organized in the following way.

- § 4.1. A first result of maximum solution.
- § 4.2. Approximation by systems of QVI.
- § 4.3. Final result for the maximum solution.
- § 4.4. Verification of H–J–B equation.

4.1. A first result of maximum solution.

THEOREM 4.1. *Under assumption A and if we assume in addition (see (2.21))*

$$c(x, v) \geq c \geq \mu_0, \quad \text{where } \mu_0 \text{ is given by (2.22),}$$

then the optimal cost function $u(x)$ belongs to $W_0^{1,\infty}(\mathcal{O})$ and is the maximum element of the set s ,

$$s = \{\tilde{u} \in W_0^{1,\infty}(\mathcal{O}), \forall v \in V, A(v)\tilde{u} \leq f(v) \text{ in } \mathcal{D}'(\mathcal{O})\}.$$

Remark 4.1. The optimal cost function $u(x)$ given by (see (3.1))

$$u(x) = \inf_{\mathcal{A}} E \left[\int_0^{\tau_x} f(y_x(t), v(t)) \exp \left(- \int_0^t c(y_x(s), v(s)) ds \right) dt \right]$$

appears to be the solution of (3.1) in some weak sense: $u(x)$ is the upper envelope of all subsolutions of (4.1). Of course $u(x)$ itself is a subsolution.

Proof. The proof will be divided into several steps:

- 1) $u(x)$ belongs to s .
- 2) A general lemma.
- 3) If $\tilde{u} \in s$ then $\tilde{u}(x) \leq u(x)$ for all $x \in \bar{\mathcal{O}}$.

1) In view of Corollary 3.1, we know that $u \in W_0^{1,\infty}(\mathcal{O})$. We have to prove that for all $v \in V$, $A(v)u \leq f(v)$ in $\mathcal{D}'(\mathcal{O})$. To do this, we use a technique due to N. V. Krylov [11] (see a simplified version in [1]). Let $v \in V$ and let us consider an admissible system corresponding to $v(t, \omega) \equiv v$; because of Corollary 3.2 we have

$$u(x) \leq E \left[\int_0^{\tau_x} \{f(y_x(s), v) + \lambda u(y_x(s))\} \exp \left(- \int_0^s c(y_x(t), v) dt - \lambda s \right) ds \right].$$

Now if we introduce u_λ , the solution of

$$A(v)u_\lambda + \lambda u_\lambda = u \quad \text{in } \mathcal{O}, \quad u_\lambda|_{\Gamma} = 0,$$

we know that

$$u_\lambda(x) = E \left[\int_0^{\tau_x} u(y_x(s), v) \exp \left(- \int_0^s c(y_x(t), v) dt - \lambda s \right) ds \right].$$

Thus

$$A(v)u_\lambda \leq E \left[\int_0^{\tau_x} f(y_x(s), v) \exp \left(- \int_0^s c(y_x(t), v) dt - \lambda s \right) ds \right] = f_\lambda(x)$$

or $A(v)(\lambda u_\lambda) \leq \lambda f_\lambda(x)$.

To conclude, we note that λu_λ is bounded in $L^\infty(\mathcal{O})$ and that $\lambda u_\lambda - u = A(v)u_\lambda = (1/\lambda) A(v)(\lambda u_\lambda) \rightarrow 0$, as $\lambda \rightarrow +\infty$, in $\mathcal{D}'(\mathcal{O})$; $\lambda f_\lambda \rightarrow f(v)$, as $\lambda \rightarrow +\infty$, (in fact for all $x \in \bar{\mathcal{O}}$ because f is continuous) and we have in conclusion that

$$\forall v \in V \quad A(v)u \leq f(v) \quad \text{in } \mathcal{D}'(\mathcal{O}).$$

Remark 4.2. Let us remark that even in the degenerate case (if we assume only (1.5-6-8) and (3.2)) the preceding proof remains valid, and thus we have

$$(4.2) \quad A(v)u \leq f(v) \quad \text{in } \mathcal{D}(\mathcal{O}) \quad \forall v \in V.$$

2) Let us make precise the notation of the following lemma. Let $y(t)$ be a continuous process on the canonical Wiener space (Ω, F, F_n, P, W_t) such that

$$(4.3) \quad 1_{[\theta_1(\omega), \theta_2(\omega)]}(t) y(t) = \left\{ \int_{\theta_1}^t \sigma(y(t)) dW_t + \int_{\theta_1}^t g(y(t)) dt \right\} 1_{[\theta_1(\omega), \theta_2(\omega)]}(t),$$

where $\theta_1 \leq \theta_2$ are two stopping times.

Let B be the differential operator

$$B = -\frac{1}{2} \sigma_{ik} \sigma_{jk} \frac{\partial^2}{\partial x_i \partial x_j} - g_i \frac{\partial}{\partial x_i} + c.$$

LEMMA 4.1. Assume that $\sigma, g, c \in W^{1,\infty}(\bar{\mathcal{O}})$, that c is nonnegative and σ is uniformly nondegenerate. Let $y(t)$ be a process satisfying (4.3), let $f \in C(\bar{\mathcal{O}})$ and let $\tilde{u} \in W_0^{1,\infty}(\mathcal{O})$ such that

$$B\tilde{u} \leq f \quad \text{in } \mathcal{D}'(\mathcal{O}).$$

Then if M belongs to F_{θ_1} , and if θ is a stopping time such that $\theta_1 \leq \theta \leq \theta_2$, we have for all $x \in \bar{\mathcal{O}}$

$$(4.4) \quad \begin{aligned} E \left\{ \left(\tilde{u}(y(\theta_1 \wedge \tau)) \exp \left(- \int_0^{\theta_1 \wedge \tau} c(y(t)) dt \right) \right. \right. \\ \left. \left. - \tilde{u}(y(\theta \wedge \tau)) \exp \left(- \int_0^{\theta \wedge \tau} c(y(t)) dt \right) \right) 1_M(\omega) \right\} \\ \leq E \left\{ 1_M(\omega) \int_{\theta_1 \wedge \tau}^{\theta \wedge \tau} f(y(t)) \exp \left(- \int_0^t c(y(s)) ds \right) dt \right\}, \end{aligned}$$

where τ is the exit time from $\bar{\mathcal{O}}$ for the process $y(t)$.

Proof of Lemma 4.1. We extend \tilde{u} , which is zero on $\mathbb{R}^N - \mathcal{O}$; then $B\tilde{u} \in W^{-1,p}(\mathbb{R}^N)$ for all $p < +\infty$. We introduce a regularizing positive convolution kernel $p_\varepsilon(\cdot) \in \mathcal{D}_+(\mathbb{R}^N)$ and we consider \tilde{u}_ε , a solution of

$$\begin{aligned} Bu_\varepsilon &= (p_\varepsilon * B\tilde{u})|_{\mathcal{O}} \quad \text{in } \mathcal{O}, \\ u_\varepsilon &= 0 \quad \text{on } \Gamma. \end{aligned}$$

Then $u_\varepsilon \in C^2(\bar{\mathcal{O}})$ and $u_\varepsilon \xrightarrow{W_0^{1,p}(\mathcal{O})} \tilde{u}$ for all $p < +\infty$; in particular, $u_\varepsilon \xrightarrow{C(\bar{\mathcal{O}})} \tilde{u}$.

Now if \mathcal{O}' is an open set such that $\mathcal{O}' \subset \bar{\mathcal{O}}' \subset \mathcal{O}$, the existence of an $\varepsilon \leq \varepsilon_0$ implies that $Au_\varepsilon \leq p_\varepsilon * f$ in \mathcal{O}' (indeed, if $\mathcal{O}' - \text{supp } p_\varepsilon \subset \mathcal{O}$, the inequality is true). Let τ' be the exit time of $\bar{\mathcal{O}}'$; then by Ito's formula we have (4.4) with \tilde{u} replaced by \tilde{u}_ε , τ by τ' and f by $p_\varepsilon * f$. Thus when $\varepsilon \rightarrow 0$, we have (4.4) with τ replaced by τ' . But \mathcal{O}' is arbitrary (with the condition $\mathcal{O}' \subset \mathcal{O}$); hence we deduce (4.4). \square

3) Let $\tilde{u} \in S$. By Lemma 1.1 it is sufficient to prove that $\tilde{u}(x) \leq J(x, \mathcal{A}, \infty, 0)$ for all admissible systems such that $v(t)$ is continuous. By taking image measure we can also assume that (Ω, F, F_t, P, W_t) is the canonical Wiener space. Let \mathcal{A} be such an admissible system. We introduce

$$\tilde{v}_n(t, \omega) = \sum_k v\left(\frac{k}{2^n}, \omega\right) 1_{[k/2^n, (k+1)/2^n]}(t),$$

$$\exists N, \quad P(N) = 0, \quad \forall \omega \notin N, \quad \forall t, \quad v_n(t, \omega) \rightarrow v(t, \omega) \quad \text{as } n \rightarrow \infty.$$

Now for k, n fixed $v((k/2^n), \omega) = \text{a.s.} \lim_{j \rightarrow \infty} v_j^{k,n} 1_{A_j}(\omega)$, where $v_j^{k,n} \in \mathbb{R}^m$, $A_j \in F_k/2^n$. Thus

there exists N^1 such that $P(N^1) = 0$ and

$$v(t, \omega) = \lim_n \bar{v}_n(t, \omega) \quad \forall \omega \notin N^1, \quad \forall t,$$

and

$$\bar{v}_n(t, \omega) = \sum_{j,k} v_{jk} 1_{A_{jk}}(\omega) 1_{[\theta_j, \theta_{j+1}[}(t),$$

where $\theta_j = j/2^n$, $\theta_{j+1} = (j+1)/2^n$, $v_{jk} \in \mathbb{R}^m$, $A_{jk} \in \mathcal{F}_{\theta_j}$ and, for fixed j , A_{jk} are disjoint sets.

On the other hand there is a V_0 compact $\subset V_0$ such that $v(t, \omega) \in V_0$. Let W_0 be the convex envelope of V_0 ; W_0 is convex compact included in V . Let P_{W_0} be the Euclidean projection onto W_0 , and let us finally consider

$$v^n(t, \omega) = \sum_{j,k} P_{W_0}(v_{jk}) 1_{A_{jk}}(\omega) 1_{[\theta_j, \theta_{j+1}[}(t) = P_{W_0}(v_n(t, \omega)).$$

Then

$$\omega \notin N^1, \quad \forall t, \quad v^n(t, \omega) \rightarrow v(t, \omega) \quad \text{as } n \rightarrow \infty, \quad v^n(t, \omega) \in W_0 \text{ compact of } V.$$

If we denote by $y_x^n(t)$ the process corresponding to $v^n(t)$, we have thus defined a sequence \mathcal{A}_n of admissible systems on the canonical Wiener space, and by Lemma 2.2 it is sufficient to prove that

$$u(x) \leq E \left[\int_0^{\tau_x} f(y_x^n(t), v^n(t)) \exp \left(- \int_0^t c(y_x^n(s), v^n(s)) ds \right) dt \right]$$

or

$$(4.5) \quad \begin{aligned} & \forall j, \quad \forall k, \quad E \left[1_{A_{jk}}(\omega) \tilde{u}(y_x^n(\theta_j \wedge \tau_x)) \exp \left(- \int_0^{\theta_j \wedge \tau_x} c(y_x^n(t), v^n(t)) dt \right) \right] \\ & \leq E \left[1_{A_{jk}}(\omega) \tilde{u}(y_x^n(\theta_{j+1} \wedge \tau_x)) \exp \left(- \int_0^{\theta_{j+1} \wedge \tau_x} c(y_x^n(t), v^n(t)) dt \right) \right. \\ & \quad \left. + \int_{\theta_j \wedge \tau_x}^{\theta_{j+1} \wedge \tau_x} f(y_x^n(t), v_{jk}) \exp \left(- \int_0^t c(y_x^n(s), v(s)) ds \right) dt \right]. \end{aligned}$$

But Lemma 4.1 implies this inequality and we conclude. \square

Remark 4.3. The preceding proof shows that if we do not assume (1.21), and if we know that $u(x) \in W_0^{1,\infty}(\mathcal{O})$, then u is the maximum element of \mathcal{S} .

4.2. Approximating systems of QVI. We are going to investigate in this section the approximation of (4.1) by different systems. Following an idea of L. Tartar, introduced independently in [6], we introduce the following penalized problem P_ε : Find u^1, \dots, u^n solutions of

$$(P_\varepsilon) \quad \begin{aligned} A_1 u^1 + \beta_\varepsilon(u^1 - u^2) &= f^1 \quad \text{in } \mathcal{O}, & u^1 &= 0 \quad \text{on } \Gamma, \\ A_2 u^2 + \beta_\varepsilon(u^2 - u^3) &= f^2 \quad \text{in } \mathcal{O}, & u^2 &= 0 \quad \text{on } \Gamma, \\ &\dots & & \\ A_n u^n + \beta_\varepsilon(u^n - u^1) &= f^n \quad \text{in } \mathcal{O}, & u^n &= 0 \quad \text{on } \Gamma, \end{aligned}$$

where $A_i = A(v_i)$, $f_i = f(v_i)$ and (v^1, \dots, v^n) is a fixed subset of V , and $\beta_\varepsilon(t) = \beta(t/\varepsilon)$. Here β is a continuous convex nondecreasing function on \mathbb{R} , such that $\beta(t) = 0$ if $t \leq 0$, $\beta(t) > 0$ if $t > 0$.

We also introduce the following system of quasivariational inequalities (in short QVI; see [2], [3], for example)

$$\begin{aligned}
 & A_1 u^1 \leq f_1, \quad u^1 \leq \varepsilon + u^2, \quad (A_1 u^1 - f_1)(u^1 - \varepsilon - u^2) = 0 \quad \text{in } \mathcal{O}, \\
 & u^1 = 0 \quad \text{on } \Gamma, \\
 & (Q_\varepsilon) \quad \dots \\
 & A_n u^n \leq f_n, \quad u^n \leq \varepsilon + u^1, \quad (A_n u^n - f_n)(u^n - \varepsilon - u^1) = 0 \quad \text{in } \mathcal{O}, \\
 & u^n = 0 \quad \text{on } \Gamma.
 \end{aligned}$$

In this section we solve problems (P_ε) , (Q_ε) (actually we shall prove just some obvious, nearly classical results which are sufficient for our goals) and we shall also give the stochastic interpretation of (Q_ε) . In the next section we are going to prove that $(u^1, \dots, u^n) \rightarrow u$, as $\varepsilon \rightarrow 0$, in $C(\mathcal{O})$ which is the optimal cost function.

THEOREM 4.2. *Under assumption A and if we assume in addition (see (3.2))*

$$\text{if } \mathcal{O} \text{ is unbounded, } c(x, v) \geq c_0 > 0 \quad \forall x, \quad \forall v,$$

and that Γ is regular, then there exists a unique solution (u^1, \dots, u^n) of (P_ε) in $C^{2,\alpha}(\mathcal{O})$ ($\forall \alpha < 1$) (resp. $C_{\text{loc}}^{2,\alpha}(\mathcal{O}) \cap C_b(\bar{\mathcal{O}})$ if \mathcal{O} is unbounded).

Proof. We prove just a priori estimates in the case of a bounded domain \mathcal{O} . First, we remark that $W^{2,p}(\mathcal{O})$ (and hence $C^{2,\alpha}$) estimates follow easily from $L^\infty(\mathcal{O})$ estimates. But $A_i u^i \leq f^i$, for all i , and this implies that $u^i \leq \text{const}$.

Now we consider $w(x) = w(x, \xi) = \exp(-k\rho^2) - \exp(k|x - \xi_1|^2)$, where ξ is fixed in Γ , ξ_1 is associated to ξ by (1.15) and $k > 0$. We have seen that for $k \geq k_0 > 0$ (see (2.19))

$$A(v) w(x) \geq \alpha > 0 \quad \forall x \in \mathcal{O} \quad \forall v \in V.$$

Thus, for λ large enough, we have

$$(4.6) \quad A_i(\lambda w(x)) < f^i \quad \forall x \in \mathcal{O}, \quad \forall i, \quad (-\lambda w)|_\Gamma \leq 0.$$

Let x_0 be in \mathcal{O} , i_0 be in $\{1, \dots, n\}$ such that

$$u_{i_0}(x_0) + \lambda w(x_0) = \min_{x, i} u_i(x) + \lambda w(x).$$

If $x_0 \in \Gamma$, $u_i(x) + \lambda w(x) \geq \lambda w(x_0)$, and we conclude that $u_i(x) \geq 0$.

If $x_0 \in \mathcal{O}$, by the maximum principle we have

$$A_{i_0}(u_{i_0}(x_0) + \lambda w(x_0)) \leq c_{i_0}(u_{i_0}(x_0) + \lambda w(x_0));$$

since one may assume $u_{i_0}(x_0) + \lambda w(x_0) < 0$ and $A_{i_0} u_{i_0}(x_0) = f_{i_0}(x_0)$, by (4.6) we have a contradiction and this contradiction gives the L^∞ estimate. Uniqueness is proved by similar arguments. \square

Remark 4.4. Actually uniqueness may be proved in the class $W_{\text{loc}}^{2,n}(\mathcal{O}) \cap C_b(\bar{\mathcal{O}})$.

Remark 4.5. If β_ε is smooth then u_i are smooth.

THEOREM 4.3. *Under assumption A and if \mathcal{O} is bounded, there exists a maximum weak solution of (Q_ε) in the following sense:*

$$\begin{aligned}
 & a_i(u^i, v - u^i) \geq (f, v - u^i), \quad v \in H_0^1(\mathcal{O}), \quad v \leq \varepsilon + u^{i+1}, \\
 & (Q_\varepsilon) \quad u^i \in H_0^1(\mathcal{O}), \quad u^i \leq \varepsilon + u^{i+1},
 \end{aligned}$$

where $u^{n+1} = u^1$, and $a_i(u, v) = \langle A_i u, v \rangle_{H^{-1} \times H_0^1}$.

Furthermore $u^i \in C(\bar{\mathcal{O}})$ and $u^i = \lim_{\eta \downarrow 0} u_\eta^i$, where (u_η^i) is the solution of

$$(R_{\varepsilon, \eta}) \quad A_i u_\eta^i + \beta_\eta (u_\eta^i - \varepsilon - u_\eta^{i+1}) = f^i \quad \text{in } \mathcal{O}, \quad u_\eta^i = 0 \quad \text{on } \Gamma.$$

Remark 4.6. The existence of (u_η^i) is obtained in the same way as the existence of the solution (u_i) of (P_ε) .

THEOREM 4.4. Under the assumptions of Theorem 4.3, we have

$$u^i(x) = \inf_{\theta} E \left\{ \int_0^{\tau_x} f(y_x(t), v(t)) \exp \left(- \int_0^t c(y_x(s), v(s)) ds \right) dt \right. \\ \left. + \varepsilon \sum_{n \geq 1} \exp \left(- \int_0^{\theta_n} c(y_x(s), v(s)) ds \right) \right\},$$

where $\theta = (\theta_n)_{n \in \mathbb{N}}$ is a sequence of stopping times such that $\theta_0 = 0 < \theta_1 < \theta_2 < \dots$, $v(t, w) = v_j$ $1_{(\theta_k(\omega) \leq t < \theta_{k+1}(\omega))}$, $j \equiv i + k - 1 \pmod{n}$, and $y_x(t)$ is the solution of

$$dy_x(t) = \sigma(y_x(t), v(t)) dW_t + g(y_x(t), v(t)) dt, \\ y_x(0) = x$$

(in the canonical Wiener space).

Proofs of Theorems 4.2 and 4.3. As these results are just variations of results given in [2], [3], we just give hints on the proofs.

Let $u^{i,m}$ be the solution of

$$A_i u^{i,m} \leq f^i, \quad u^{i,m} \leq \varepsilon + u^{i+1,m-1}, \quad (A_i u^{i,m} - f^i)(u^{i,m} - \varepsilon - u^{i+1,m-1}) = 0 \quad \text{in } \mathcal{O}, \\ u^{i,m}|_\Gamma = 0$$

(see [19] for the solution of this VI), and $u^{i,0}$ are given by $A_i u^{i,0} = f^i$ in \mathcal{O} , $u^{i,0} = 0$ on Γ .

One easily proves as in [2] that $u^{i,m} \downarrow_m$.

An argument similar to the one given in the proof of Theorem 4.2 gives

$$u^{i,m} \geq -\lambda w(x) \quad \forall i, \quad \forall m.$$

Thus $\|u^{i,m}\|_{L^\infty(\mathcal{O})} \leq \text{const.}$

Now, since there exists λ such that $a_i(u, u) + \lambda \|u\|_{L^2(\mathcal{O})}^2 \geq \nu \|u\|_{H^1_0(\mathcal{O})}^2$, we deduce easily from

$$a_i(u^{i,m}, -\lambda w - u^{i,m}) \geq (f^i, -\lambda w - u^{i,m})$$

that $\|u^{i,m}\|_{H^1_0(\mathcal{O})} \leq \text{const.}$

The proof of the first part of Theorem 4.3 follows the one given in [3], for example.

Next the proof of the continuity of u^i and of Theorem 4.4 is easily obtained by methods similar to those in [3] and in [22].

Finally, by a method similar to the one given in the proof of Theorem 4.2, we prove that

$$u_\eta^i \downarrow \quad \text{when } \mu \downarrow 0, \quad \|u_\eta^i\|_{L^\infty(\mathcal{O})} \leq \text{const.}, \quad \text{and} \\ u_\eta^i(x) \geq -\lambda w(x) \quad \forall i, \quad \forall \mu, \quad \forall x \in \mathcal{O}.$$

Then we prove easily that $u_\eta^i \downarrow \underline{u}^i$, which is a weak solution of (Q'_ε) , and thus $\underline{u}^i \leq u^i$. To conclude, we introduce $u_\eta^{i,m}$, the solution of

$$A_i u_\eta^{i,m} + \beta_\eta (u_\eta^{i,m} - \varepsilon - u_\eta^{i+1,m-1}) = f^i \quad \text{in } \mathcal{O}, \quad u_\eta^{i,m} = 0 \quad \text{on } \Gamma;$$

we have

$$u_{\eta}^{i,m} \downarrow_{\eta \downarrow 0} u^{i,m}, \quad u_{\eta}^{i,m} \downarrow_{n \uparrow \infty} u_{\eta}^i, \quad u_{\eta}^i \downarrow_{\eta \downarrow 0} \underline{u}^i, \quad u^{i,m} \downarrow_{m \uparrow \infty} u^i;$$

thus $u^i = \underline{u}^i$. \square

Remark 4.8. We have also that if u_{ε}^i is the solution of (Q_{ε}) , u^{i,r_1} is the solution of (P_{η}) , and $u_{\varepsilon}^{i,\eta}$ is the solution of $(R_{\varepsilon,\eta})$,

$$(4.7) \quad u_{\varepsilon}^i \leq u_{\varepsilon}^{i,\eta} \quad \forall \eta > 0, \quad u_{\varepsilon}^i = \lim_{\eta \downarrow 0} \downarrow u_{\varepsilon}^{i,\eta},$$

$$(4.8) \quad u^{i,\eta} \leq u_{\varepsilon}^{i,\eta} \quad \forall \varepsilon > 0, \quad u^{i,\eta} = \lim_{\varepsilon \downarrow 0} \downarrow u_{\varepsilon}^{i,\eta}. \quad \square$$

4.3. Final result for the maximum solution.

THEOREM 4.5. *Under assumption A, and if we assume (see (3.2))*

$$\text{if } \mathcal{O} \text{ is unbounded, } c(x, v) \geq c > 0 \quad \forall x \in \mathcal{O}, \quad \forall v \in V,$$

then the optimal cost function $u(x)$ belongs to $W_0^{1,\infty}(\mathcal{O})$ and is the maximum element of the set S .

Proof. The proof will be divided into several parts.

- 1) Lipschitz estimates on $u^{i,\eta}$.
- 2) $u^{i,\eta} \downarrow_{\eta \downarrow 0} u_n$, $u_n \downarrow_{n \uparrow +\infty} u$ if $c(x, v) \geq c_0 > \mu_0$.
- 3) Conclusion.

- 1) We prove that $\|u^{i,\eta}\|_{W^{1,\infty}(\mathcal{O})} \leq \text{const.}$ (independent of i, η).

- First, we remark that, if \mathcal{O} is bounded, we already know that $\|u^{i,\eta}\|_{L^{\infty}(\mathcal{O})} \leq \text{const.}$

In the case of an unbounded domain, one proves by a simple limiting process ($\mathcal{O}_n \rightarrow \mathcal{O}$, \mathcal{O}_n bounded) that

$$\|u^{i,\eta}\|_{L^{\infty}(\mathcal{O})} \leq \sup_i \frac{\|f^i\|_{L^{\infty}(\mathcal{O})}}{c_0}.$$

• Next we prove that $|u^i(x)| \leq \lambda |w(x, \xi)|$ for all $\xi \in \Gamma$ and for all $x \in B(\xi, p')$, where λ, p' do not depend on i, η, ξ , and $w(x, \xi)$ is given by (2.15'). The proof is immediate if we recall that, if k is large enough,

$$A_i w(x, \xi) \geq \alpha \exp -k|x - \xi_1|^2 \geq \beta > 0 \quad \text{on some } B(\xi, p') = B.$$

Now on $(\partial B) \cap \mathcal{O}$ $w \geq \gamma > 0$; thus there exists $\lambda > 0$ such that

$$A_i \lambda w(x, \xi) > \sup_i \|f_i\|_{L^{\infty}(\mathcal{O})} \quad \text{on } B,$$

$$\lambda w|_{(\partial B) \cap \mathcal{O}} > \max_{i,\eta} \|u^{i,\eta}\|_{L^{\infty}(\mathcal{O})}, \quad \lambda w|_{B \cap \partial \mathcal{O}} \geq 0.$$

From an application of the maximum principle similar to the one given in the proof of Theorem 4.2 we deduce

$$|u^{i,\eta}(x)| \leq \lambda |w(x, \xi)| \quad \forall x \in B(\xi, p'), \quad \forall \xi \in \Gamma,$$

and this implies $|\nabla u^{i,\eta}(\xi)| \leq \text{const.}$ for all $\xi \in \Gamma$.

• Finally we consider (as in [18]) the auxiliary function $w_i(x) = |\nabla u^{i,\eta}(x)|^2 + \lambda (C - u^{i,\eta}(x))^2$ (we shall forget about the η subscript in the following proof), where $\lambda > 0$ and $C \geq \max_{i,\eta,x} u^{i,\eta}(x)$. We shall assume in the proof to the

theorem that $\beta \in C^2(R)$; thus $u^i \in C^3(\mathcal{O})$. Differentiating (P_ε) with respect to x_j , we obtain (u_k will denote $\partial u / \partial x_k$)

$$\begin{aligned} & -a_{k1}^i(x) u_{klj}^i(x) + b_k^i u_{kj}^i(x) + c^i u_j^i + \beta'(u^i - u^{i+1})(u_j^i - u_j^{i+1}) \\ & = f_j^i(x) + a_{k1,j}^i(x) u_{k1}^i - b_{k,j}^i u_k^i - c_j^i u^i, \end{aligned}$$

and a simple calculation shows that for all i

$$\begin{aligned} & A_i w_i(x) + \beta'(u^i - u^{i+1}) 2(u_j^i u_j^i - u_j^{i+1} u_j^i) \\ & \leq -2\nu(u_{kj}^i)^2 (f_j^i + a_{k1,j}^i(x) u_{k1}^i - b_{k,j}^i u_k^i - c_j^i u^i) 2 u_j^i \\ & \quad + 2\lambda(C - u^i)[-f^i + \beta(u^i - u^{i+1})] + C_1 - 2\lambda\nu(u_j^i)^2. \end{aligned}$$

Thus we have, choosing λ large enough, for all i ,

$$\begin{aligned} & A_i w_i(x) + \beta'(u^i - u^{i+1}) 2(u_j^i u_j^i - u_j^{i+1} u_j^i) \\ & \quad - \beta(u^i - u^{i+1}) 2\lambda(C - u^i) \leq C_2 - \alpha w_i(x); \end{aligned}$$

as $(C - u^i) \geq 0$, $\beta(0) = 0$ and β is convex we have

$$-\beta(u^i - u^{i+1}) 2\lambda(C - u^i) \geq 2\lambda(C - u^i) \beta'(u^i - u^{i+1}) \{(C - u^i) - (C - u^{i+1})\}.$$

Finally suppose \mathcal{O} is bounded, and let $i_0 - x_0$ be such that $w_{i_0}(x_0) = \max_{i,x} w(x)$ if x_0 belongs to Γ ; we concluded that because of the above estimate if x_0 belongs to \mathcal{O} , at this point we have $A_{i_0} w_{i_0}(x_0) \geq 0$ and

$$\begin{aligned} & \beta'(u^i - u^{i+1}) 2(u_j^i u_j^i) 2(u_j^i u_j^i - u_j^{i+1} u_j^i) - \beta(u^i - u^{i+1}) 2\lambda(C - u^i) \\ & \geq \beta'(u^i - u^{i+1}) (w_i - w_{i+1}) \geq 0. \end{aligned}$$

Hence we deduce $w_i(x) \leq C_2/\alpha$.

The case of an unbounded domain is obtained by a limiting process, taking \mathcal{O}_n a sequence of domains converging to $\mathcal{O}(\mathcal{O}_n \uparrow \mathcal{O})$.

2) Next, we suppose that $c(x, v) \geq c_0 > \mu_0$ for all $x \in \mathcal{O}$ and all $v \in V$.

We know (by the preceding estimate) that $u^{i,\eta} \downarrow u_n \in W_0^{1,\infty}(\mathcal{O})$, as $\eta \rightarrow 0$

Furthermore for all $i \leq n$ $A_i u_n \leq f_i$ in $\mathcal{D}'(\mathcal{O})$. Now if we let n go to $+\infty$ such that $(v_i, i \in N)$ is dense in V , we see easily that $u^{i,\eta} \downarrow$ as $n \uparrow \infty$ we have $u_n \downarrow u \in W^{1,\infty}(\mathcal{O})$, as $n \rightarrow \infty$ (by the preceding estimate, which is independent of n) and for all $i \in N$

$$A_i u \leq f_i \quad \text{in } \mathcal{D}'(\mathcal{O}).$$

Thus

$$\forall v \in V, \quad A(v) u \leq f(v) \quad \text{in } \mathcal{D}'(\mathcal{O}).$$

Now if we suppose that $c(x, v) \geq c_0 > \mu_0$ then by Theorem 4.1, $\underline{u}(x) \leq u(x)$. But by remark 4.8 $u_\varepsilon^i \downarrow u_n$ as $\varepsilon \downarrow 0$, and from the stochastic interpretation of u_ε^i , we see that

$$\forall v \in V, \quad A(v) u \leq f(v), \quad \text{in } \mathcal{D}'(\mathcal{O}).$$

Hence, if we suppose $c(x, v) \geq c_0 > \mu_0$ $u(x) = \underline{u}(x)$, and in the general case $\underline{u}(x) \in W_0^{1,\infty}(\mathcal{O})$, belongs to S and $\underline{u}(x) \geq u(x)$ for all $x \in \mathcal{O}$.

3) In the general case, we consider $\lambda > 0$ such that $c(x, v) + \lambda \geq c_0 > \mu_0$, and we introduce a mapping $T_{\varepsilon,n}$ defined by: if $w \in C_b(\mathcal{O})$, $T_{\varepsilon,n} w = (T_{\varepsilon,n}^1 w)_i$ is the solution of (Q_ε) where A_i is replaced by $A_i + \lambda$, f^i by $f^i + \lambda w$.

From the stochastic interpretation, we have easily

$$\begin{aligned} \|T_{\varepsilon,n} w_1 - T_{\varepsilon,n} w_2\|_{L^\infty(\mathcal{O})} &\leq \frac{\lambda}{\lambda + c_0} \|w_1 - w_2\|_{L^\infty(\mathcal{O})} \quad \text{if } \mathcal{O} \text{ is unbounded,} \\ &\leq \frac{1}{\lambda} \sup_{\mathcal{A},x} E[1 - e^{-\lambda\tau_x}] \quad \text{if } \mathcal{O} \text{ is bounded,} \\ &\leq \frac{1 - e^{-\lambda C}}{\lambda} \quad \text{where } C > 0 \end{aligned}$$

by Jensen's inequality (cf. Lemma 2.4).

Now for any $w \in C_b(\bar{\mathcal{O}})$, $T_{\varepsilon,n} w \downarrow Tw \in C_b(\bar{\mathcal{O}})$, and by step 2)

$$Tw(x) = \inf_{\mathcal{A}} E \left[\int_0^{\tau_x} \{f(y_x(t), v(t)) + \lambda w(y_x(t))\} \exp \left(- \int_0^t c(y_x(s), v(s)) ds \right) dt \right].$$

From these two facts, we deduce that the fixed point of T_ε in $C_b(\bar{\mathcal{O}})$ converges to the fixed point of T , i.e., $u_\varepsilon^i \rightarrow u(x)$, in $C_b(\bar{\mathcal{O}})$. Thus $u \in W_0^{1,\infty}(\mathcal{O})$ and $u = \bar{u}$. To conclude, we remark that the proof of Theorem 4.1 now applies, and thus u is the maximum element of S . \square

COROLLARY 4.1. *Under the assumptions of Theorem 4.5, we have*

$$u(x) = \inf_{\mathcal{A}_\theta} E \left[\int_0^{\tau_x} f(y_x(t), v(t)) \exp \left(- \int_0^t c(y_x(s), v(s)) ds \right) dt \right],$$

where the infimum is taken over all admissible systems such that (Ω, F, F_t, P, W_t) is the canonical Wiener space, and there exists $\theta = (\theta_n)_{n \geq 0}$, a sequence of stopping times such that $\theta_0 = 0 < \theta_1 < \theta_2 < \dots < \theta_n \uparrow +\infty$ and $v(t, x) = v_j$ if $t \in [\theta_j(\omega), \theta_{j+1}(\omega)[$, where $(v_n)_{n \geq 0}$ is a sequence of elements of V .

Proof of Corollary 4.1. Immediate in view of Theorem 4.4 and the proof of Theorem 4.5. \square

4.4. Verification of H-J-B equations. We now recall a result due to L. C. Evans and P.-L. Lions [7] concerning the solution of (4.1). We will assume in this section that \mathcal{O} is smooth and we have

$$(4.9) \quad \phi(\cdot, v) \in W^{2,\infty}(\mathcal{O}) \quad \text{and} \quad \sup_{v \in V} \|\phi(\cdot, v)\|_{W^{2,\infty}(\mathcal{O})} < \infty \quad \forall \phi = \sigma, b, c, f.$$

THEOREM 4.6. *Under assumptions A and (4.9), we have that $u \in W^{2,\infty}(\mathcal{O})$ is the unique solution in $W^{2,\infty}(\mathcal{O})$ of (4.1):*

$$\sup_{v \in V} \{A(v)u - f(v)\} = 0 \quad \text{a.e. in } \mathcal{O}, \quad u = 0 \quad \text{on } \Gamma.$$

Remark 4.9. This result extends previous results due to H. Brezis and L. C. Evans [4], P.-L. Lions [20], L. C. Evans and A. Friedman [6], P.-L. Lions and J.-L. Menaldi [21], P.-L. Lions [15].

REFERENCES

- [1] A. BENSOUSSAN AND J.-L. LIONS, *Applications des inéquations variationnelles en contrôle stochastique*, Dunod, Paris, 1978.
- [2] ———, *Nouvelles méthodes en contrôle impulsionnel*, Appl. Math. Optim., 1 (1974), pp. 289–312.

- [3] ———, *Contrôle impulsif et systèmes d'inéquations quasi-variationnelles*, CRAS Paris, 278 (1974), pp. 747–751.
- [4] H. BREZIS AND L. C. EVANS, *A variational approach to the Bellman-Dirichlet problem for two uniformly elliptic operators*, Arch. Rat. Mech. Anal., 73 (1979), pp. 1–14.
- [5] E. B. DYNKIN, *Markov Processes*, Springer-Verlag, Berlin, 1965.
- [6] L. C. EVANS AND A. FRIEDMAN, *Optimal stochastic switching and the Dirichlet problem for the Bellman equation*, Trans. Am. Math. Soc., 253 (1979), pp. 365–389.
- [7] L. C. EVANS AND P.-L. LIONS, *Résolution des équations de Hamilton-Jacobi-Bellman pour des opérateurs uniformément elliptiques*, CRAS Paris, 290 (1980), pp. 1049–1052.
- [8] W. H. FLEMING AND R. RISHEL, *Optimal Deterministic and Stochastic Control*, Springer-Verlag, Berlin, 1975.
- [9] B. GAVEAU, *Méthodes de contrôle optimal en analyse complexe*, J. Résolution d'équation de Monge-Ampère., J. Funct., Anal. 25 (1977), pp. 381–411; cf. also CRAS Paris, 284 (1977), pp. 99–593.
- [10] I. L. GENIS AND N. V. KRYLOV, *An example of a one-dimensional controlled process*, Theory Prob. Appl., 21 (1976), pp. 148–152.
- [11] N. V. KRYLOV, *Control of a solution of a stochastic integral equation*, Theory. Prob. Appl., 17 (1972), pp. 114–131.
- [12] ———, *On control of the solution of a stochastic integral equation with degeneration*, Izv. Akad. Nauk. USSR, 36 (1972), pp. 249–262.
- [13] ———, *An inequality in the theory of stochastic integrals*, Theory Prob. Appl., 16 (1971), pp. 438–448.
- [14] ———, *Controlled Diffusion Processes*, Springer-Verlag, New York, 1980.
- [15] P.-L. LIONS, *Résolution analytique des problèmes de Bellman-Dirichlet*, Acta Mathematica (1981), to appear; cf. also CRAS Paris, 287 (1978), pp. 747–750.
- [16] ———, *Control of diffusion processes in \mathbb{R}^n* , Comm. Pure Appl. Math. (1981), to appear; cf. also CRAS Paris, 288 (1979), pp. 339–342.
- [17] ———, *Equations de Hamilton-Jacobi-Bellman dégénérées*, CRAS Paris, 289 (1979), pp. 329–332.
- [18] ———, *An estimate of the Lipschitz norm of solutions of variational inequalities*, in Variational Inequalities, Cottle, Gianessi, Lions, eds., John Wiley, New York, 1979.
- [19] ———, *Resolution de problèmes elliptiques du 2ème ordre sous forme divergence*, Proc. Roy. Soc. Edin., to appear.
- [20] ———, *Some problems related to the Bellman-Dirichlet equation for two elliptic operators*, Comm. P.D.E., to appear, cf. also thèse de 3ème cycle Paris VI, 1978.
- [21] P.-L. LIONS AND J.-L. MENALDI, *Problèmes de Bellman avec le contrôle dans le coefficients de plus haut degré*, CRAS Paris, 287 (1978), pp. 409–412.
- [22] J.-L. MENALDI, *On the optimal impulse control problem for degenerate diffusions*, this Journal, 18 (1980), pp. 722–739; cf. also CRAS Paris, 284 (1977), pp. 1400–1502.
- [23] M. NISIO, *Remarks on stochastic optimal controls*, Japan J. Math., 1, (1975), pp. 159–183.
- [24] ———, *On a nonlinear semi-group attached to stochastic optimal control*, Publ. Res. Inst. Math. Sci., 12 (1976–77), pp. 513–537.
- [25] D. W. STROOCK AND S. R. S. VARADHAN, *Diffusion processes with continuous coefficients*, I, II, Comm. Pure. Appl. Math., 22 (1969), pp. 345–400; pp. 479–530.
- [26] P.-L. LIONS AND J.-L. MENALDI, *Optimal control of stochastic integrals and Hamilton-Jacobi-Bellman equations*. II, this Journal, this issue, pp. 82–95.

OPTIMAL CONTROL OF STOCHASTIC INTEGRALS AND HAMILTON-JACOBI-BELLMAN EQUATIONS. II*

PIERRE-LOUIS LIONS† AND JOSÉ-LUIS MENALDI‡

Abstract. We consider the solution of a stochastic integral control problem, and we study its regularity. In particular, we characterize the optimal cost as the maximum solution of

$$\begin{aligned} \forall v \in V, \quad A(v)u &\leq f(v) \quad \text{in } \mathcal{D}'(\mathcal{O}), \\ u &= 0 \quad \text{on } \partial\mathcal{O}, \quad u \in W^{1,\infty}_0(\mathcal{O}), \end{aligned}$$

where $A(v)$ is a uniformly elliptic second order operator and V is the set of the values of the control.

1. Introduction

1.1. General introduction. In this paper, we extend the results of part I [14] (this Journal, this issue, pp. 58–81) to the degenerate case (see also [15]).

We consider a stochastic system governed by the stochastic differential equation

$$(1.1) \quad \begin{aligned} dy(t) &= \sigma(y(t), v(t)) dW_t + g(y(t), v(t)) dt, \quad t \geq 0, \\ y(0) &= x \in \mathbb{R}^N, \end{aligned}$$

where W_t is a Wiener process, g and σ are given functions, and $v(t)$ is a “continuous” control taking values in some set $V \subset \mathbb{R}^m$.

We want to minimize the cost function (with notational change from part I)

$$(1.2) \quad J_x(v) = E \left\{ \int_0^{\tau_x} f(y(t), v(t)) \exp \left(- \int_0^t c(y(s), v(s)) ds \right) dt \right\}$$

over all admissible controls $v(t)$. In this formula f and c are given functions and τ_x is the first exit time of the process $y(t)$ from a given domain $\bar{\mathcal{O}}$. Let us denote

$$(1.3) \quad u(x) = \inf \{ J_x(v) / v = v(\cdot) \text{ admissible control} \}$$

the optimal cost function.

In part I (see also [15]), under suitable assumptions containing an assumption of nondegeneracy,

$$(1.4) \quad \sigma \sigma^*(x, v) \geq \alpha > 0 \quad \forall x \in \bar{\mathcal{O}}, \quad \forall v \in V,^1$$

we proved that the function $u(x)$ is the maximum element of the set of functions \tilde{u} satisfying $\tilde{u} \in W^{1,\infty}_0(\mathcal{O})$ and

$$(1.5) \quad A(v)\tilde{u} \leq f(v) \quad \text{in } \mathcal{D}'(\mathcal{O}) \quad \forall v \in V,$$

where $A(v) = -\frac{1}{2} \text{Tr}(\sigma \sigma^*(x, v) D^2) - g(x, v) D + c(x, v)$.²

We give here results where (1.4) is relaxed and where nevertheless this approach may still be carried out.

* Received by the editors June 13, 1980, and in revised form January 30, 1981.

† Laboratoire d'Analyse Numérique, Université Pierre et Marie Curie, 4 Place Jussieu, 75230 Paris Cédex 05, France.

‡ INRIA, Domaine de Voluceau, Rocquencourt B.P. 105, 78153 Le Chesnay Cédex, France.

¹ σ^* denotes the adjoint of σ . The inequality has to be understood in the sense of symmetric matrices.

² D denotes the gradient operator (we will also use the notation ∇).

In view of the principles of dynamical programming, one could expect u to solve (in a convenient sense)

$$\sup_{v \in V} \{A(v)u - f(v)\} = 0 \quad \text{a.e. in } \mathcal{O}.$$

Results in this direction (with operators $A(v)$ eventually degenerate) are given in N. V. Krylov [5], [6], [7], M. V. Safonov [18], [19], P.-L. Lions [9], [10], [11], [12], [13] (in the nondegenerate case the most general results are given in L. C. Evans and P.-L. Lions [2], P.-L. Lions [8]).

But the counterexample of I. L. Genis and N. V. Krylov [3] shows that the equation may not be satisfied (even in the weakest sense); therefore it seems useful to have a different characterization of u . Our goal here is to propose as one such characterization the superior envelope of sub-solutions. We remark that some results in this direction, for the deterministic case, are given in R. Gonzalez [4].

1.2. Summary. The results are organized in the following way:

Section 2. The degenerate case.

Section 3. The Cauchy problem.

Section 4. The obstacle problem.

In § 2, using some techniques of N. V. Krylov [7] and [8] and M. Nisio [17], as in [14], we build a nonlinear semigroup whose generator is related to the operator appearing in (1.5). Next, we give a stochastic characterization of $u(x)$, which is the precise way to apply the dynamical programming argument. Finally, we prove a characterization of $u(x)$ in terms of the maximum subsolution.

In § 3, we briefly develop the parabolic case. In § 4, we consider the obstacle problem. The case without "continuous" control was studied in J.-L. Menaldi [16].

1.3. Assumptions and notation. We now give notation and assumptions which will remain valid in §§ 2, 3 and 4. Let \mathcal{O} be a domain of \mathbb{R}^N and let V be a convex closed set in \mathbb{R}^m . We call an *admissible system* a set $\mathcal{A} = (\Omega, F, F_t, P, W_t, v(t), y_x(t))$, where (Ω, F, P) is a probability space, F_t is a nondecreasing right-continuous family of complete sub σ -algebras of F , W_t is a Wiener process with respect to F_t , $v(t)$ is a measurable adapted process taking values in some compact subset V_0 of V (V_0 , of course, may depend on $v(\cdot)$) and $y_x(t)$ is a solution of Ito's equation

$$(1.6) \quad \begin{aligned} dy_x(t) &= \sigma(y_x(t), v(t)) dW_t + g(y_x(t), v(t)) dt, \quad t \geq 0, \\ y_x(0) &= x, \end{aligned}$$

where $\sigma(x, v)$ and $g(x, v)$ are uniformly continuous and bounded functions from $\mathbb{R}^N \times V$ into $\mathbb{R}^N \otimes \mathbb{R}^N$ and \mathbb{R}^N respectively which are uniformly Lipschitz continuous in x . This regularity and boundedness assumption will not be recalled in what follows (and may be relaxed in some of the results which follow).

Now, for an admissible system \mathcal{A} , we define a cost function

$$(1.7) \quad \begin{aligned} J(x, \mathcal{A}, t, h) &= E \left\{ \int_0^{t \wedge \tau_x} f(y_x(s), v(s)) \exp \left(- \int_0^s c(y_x(\lambda), v(\lambda)) d\lambda \right) ds \right. \\ &\quad \left. + h(y_x(t \wedge \tau_x)) \exp \left(- \int_0^{t \wedge \tau_x} c(y_x(s), v(s)) ds \right) \right\}, \end{aligned}$$

where h is an arbitrary measurable bounded function, τ_x is the first exit time of the process $y_x(t)$ from \mathcal{O} and $f(x, v)$, $c(x, v)$ are given uniformly continuous and bounded functions from $\mathbb{R}^N \times V$ into \mathbb{R} , \mathbb{R}_+ respectively.

Finally, we define, for each h , an optimal cost function

$$(1.8) \quad Q(t)h(x) = \inf_{\mathcal{A}} J(x, \mathcal{A}, t, h) \quad \forall 0 \leq t < \infty.$$

Let us collect our assumptions:

$$(1.9) \quad |\phi(x, v) - \phi(x', v')| \leq C|x - x'| + \rho(|v - v'|) \\ \forall x, x' \in \mathbb{R}^N, \quad \forall v, v' \in V, \quad \forall \phi = \sigma_{ij}, g_i, c, f.$$

$$(1.10) \quad |\phi(x, v)| \leq C \quad \forall x \in \mathbb{R}^N, \quad \forall v \in V, \quad \forall \phi = \sigma_{ij}, g_i, c, f, \\ c(x, v) \geq c_0 > 0 \quad \forall x \in \mathbb{R}^N, \quad \forall v \in V,$$

where ρ is a given continuous function from \mathbb{R}_+ into \mathbb{R}_+ such that $\rho(0) = 0$.

We shall denote by B_s the set of bounded functions from $\bar{\mathcal{O}}$ into \mathbb{R} which are upper semicontinuous and by B_s^+ the subset of B_s of nonnegative functions. B_s and B_s^+ are closed convex cones of the Banach space B of bounded measurable functions equipped with the supremum norm ($\|h\|_\infty = \sup \{|h(x)| : x \in \bar{\mathcal{O}}\}$).

Throughout this paper, we use an assumption which will replace the non-degeneracy assumption (1.4). We suppose that there exists a subsolution which is Lipschitz continuous, i.e.,

$$(1.11) \quad \text{there exists a } \bar{u} \in W_0^{1,\infty}(\mathcal{O}) \text{ such that for all } v \in V \text{ we have for} \\ \text{some } \varepsilon > 0 \ A(v)\bar{u} \leq -1 \text{ in } \mathcal{D}'(\mathcal{O}_\varepsilon), \ \bar{u} \leq -\alpha < 0 \text{ in } \mathcal{O} - \mathcal{O}_\varepsilon,$$

where $\mathcal{O}_\varepsilon = \{x \in \mathcal{O}, \text{dist}(x, \Gamma) > \varepsilon\}$, and the operator $A(v)$ is defined by

$$(1.12) \quad A(v) = -a_{ij} \frac{\partial^2}{\partial x_i \partial x_j} + b_i \frac{\partial}{\partial x_i} + c,^3$$

with $a_{ij}(x, v) = \frac{1}{2} \sigma_{ik} \sigma_{jk}(x, v)$, $b_i(x, v) = -g_i(x, v)$.

It is easy to prove, using barrier functions as in part I or [8], that if for some $\alpha > 0$ $\Gamma = \partial^\mathcal{O}$ satisfies (n is the unit exterior normal to Γ)

$$(1.13) \quad \Gamma = \{x \in \Gamma / \forall v \in v, |\sigma(x, v)n(x)| \geq \alpha\} \\ \cup \{x \in \Gamma / v \in V, 2rg(x, v)n(x) > \text{Tr}[\sigma\sigma^*(x, v)] + \alpha\},$$

r is the radius of the uniform exterior sphere associated with \mathcal{O} ,

and

$$(1.14) \quad \mathcal{O} \text{ is bounded, regular (i.e., the exterior uniform sphere property holds),}$$

then assumption (1.11) is satisfied for c_0 large enough.

We may also replace (1.11) by

$$(1.15) \quad f(x, v) \geq 0 \quad \forall x \in \mathbb{R}^N, \quad \forall v \in V,$$

$$(1.16) \quad \text{there exists } v(x) \text{ continuous on } \bar{\mathcal{O}} \text{ such that (1.13) is satisfied for } v = v(x).$$

2. Degenerate case. This section is divided into three sections. First, we study the nonlinear semigroup $Q(t)$. Next, we give a stochastic interpretation of the optimal cost. Finally, we establish an analytical interpretation.

³ We will always use the usual convention for sums.

2.1. Nonlinear semigroup. In this section, we first prove that $Q(t)$ acting on B_s or B_s^+ is a nonlinear semigroup. Next we consider the generator of $Q(t)$.

THEOREM 2.1. *Assume (1.9), (1.10) and (1.15). Then $(Q(t), t \geq 0)$ satisfies*

$$(2.1) \quad Q(t) : B_s^+ \rightarrow B_s^+, \quad Q(0) = I, \quad Q(t+s) = Q(t) \circ Q(s) = Q(s) \circ Q(t),$$

$$(2.2) \quad \|Q(t)h - Q(s)h\|_\infty \rightarrow 0 \quad \text{as } t \rightarrow s \quad \text{if } h \text{ is uniformly continuous in } \mathbb{R}^N,$$

$$(2.3) \quad \|Q(t)h_1 - Q(s)h_2\|_\infty \leq \|h_1 - h_2\|_\infty \quad \forall h_1, h_2 \in B_s^+, \quad \forall t \geq 0,$$

$$(2.4) \quad Q(t)h_1 \leq Q(t)h_2 \quad \text{if } h_1 \leq h_2.$$

Proof. We penalize the domain \mathcal{O} . Let $p(x)$ be the distance to \mathcal{O} , i.e.,

$$(2.5) \quad p(x) = \inf \{|y - x| : y \in \mathcal{O}\},$$

and consider the following operator ($\varepsilon > 0$):

$$(2.6) \quad \begin{aligned} & Q^\varepsilon(t)h(x) \\ &= \inf E \left\{ \int_0^t f(y_x(s), v(s)) \exp \left(- \int_0^s \left(c(y_x(\lambda), v(\lambda)) + \frac{1}{\varepsilon} p(y_x(\lambda)) \right) d\lambda \right) ds \right. \\ & \quad \left. + h(y_x(t)) \exp \left(- \int_0^t \left(c(y_x(s), v(s)) + \frac{1}{\varepsilon} p(y_x(s)) \right) ds \right) \right\}. \end{aligned}$$

Clearly, $Q^\varepsilon(t)$ leaves invariant the space $C_b(\mathbb{R}^N)$ of continuous and bounded functions. From Theorem 2.1 in part I, we obtain that $Q^\varepsilon(t)$ satisfies (2.1)–(2.4).

Finally, using the fact that, for all $t \geq 0$, for $x \in \mathbb{R}^N$ and $h \in B_s^+$,

$$(2.7) \quad Q^\varepsilon(t)h(x) \rightarrow Q(t)h(x) \quad \text{decreasing as } \varepsilon \downarrow 0,^4$$

it is easy to conclude. \square

Remark 2.1. Under assumptions (1.9), (1.10) and (1.11), the semigroup $(Q(t), t \geq 0)$ satisfies (2.1)–(2.4) with B_s instead of B_s^+ . Indeed, we need to observe that, defining

$$(2.8) \quad \Gamma_0(\mathcal{A}) = \{x \in \Gamma / P(\tau_x > 0) = 0\},$$

we deduce from (1.11) (using a lemma of [8])

$$(2.9) \quad \begin{aligned} & \Gamma_0(\mathcal{A}) = \Gamma \quad \forall \mathcal{A} \text{ admissible systems,} \\ & P(y_x(\tau_x) \in \Gamma_0(\mathcal{A}) \text{ if } \tau_x < \infty) = 1 \quad \forall x \in \bar{\mathcal{O}}, \end{aligned}$$

so we can use Theorem 2.1 in part I.

We set

$$(2.10) \quad X = \{h \in C_b(\bar{\mathcal{O}}), h \text{ is uniformly continuous}\}.$$

We have the following:

THEOREM 2.2. *If we assume (1.9), (1.10) and (1.11), then for each $h \in X$ $Q(t)h \in X$. Furthermore $(Q(t)h, t \geq 0)$ is uniformly equicontinuous.*

Proof. We first consider the case where $c(x, v)$ satisfy

$$(2.11) \quad c(x, v) \geq c_0 > \mu_0^+ \quad \forall x \in \mathcal{O}, \quad \forall v \in V,$$

⁴ We extend h by zero outside $\bar{\mathcal{O}}$.

where μ_0 is given by

$$(2.12) \quad \mu_0 = \sup \left\{ \frac{1}{2} \operatorname{Tr} \left[\frac{(\sigma(x, v) - \sigma(x', v))(\sigma(x, v) - \sigma(x', v))^*}{|x - x'|^2} \right] + \frac{(x - x') \cdot (g(x, v) - g(x', v))}{|x - x'|^2} \middle/ x, x' \in \mathcal{O}, v \in V \right\}.$$

By a density argument, it is enough to prove Theorem 2.2 for smooth $f(x, v)$ and $h(x)$. By the same argument as in part I, we only have to prove that $u(t, x) = Q(t)0 \in X$ (and is uniformly equicontinuous). Let us assume that under assumption (2.11) we have proved that $|u(t, x) - u(t, x')| \leq C|x - x'|$, for all x, x' in $\bar{\mathcal{O}}$ and all $t \geq 0$. We conclude remarking that, using the dynamical programming property as in [14], we have

$$(2.13) \quad u(t, x) = \inf_{\mathcal{A}} E \left\{ \int_0^{t \wedge \tau_x} [f(y_x(s), v(s)) + ku(s, y_x(s))] \cdot \exp \left(- \int_0^s c(y_x(\lambda), v(\lambda)) d\lambda - ks \right) ds \right\},$$

for all $k \geq 0$.

Thus $u(t, x)$ is a fixed point of the mapping which transforms $u(s, x)$ into the right-hand side of (2.13); but we have

$$(2.14) \quad \begin{aligned} Tu &\in W^{1, \infty} \quad \text{if } u \in W^{1, \infty}, \\ \|Tu - Tw\|_{\infty} &\leq \frac{k}{c_0 + k} \|u - w\|_{\infty}, \end{aligned}$$

choosing k large enough. This proves Theorem 2.2.

Now, there just remains to prove that under assumption (2.11), we have $|u(t, x) - u(t, x')| \leq C|x - x'|$. We first remark that in view of the arguments given in part I, if $c_0 > \mu_0$, then

$$\begin{aligned} E \left[\left| y_x(\theta) \exp \left(- \int_0^\theta c(y_x(s), v(s)) ds \right) \right. \right. \\ \left. \left. - y_{x'}(\theta) \exp \left(- \int_0^\theta c(y_{x'}(s), v(s)) ds \right) \right| \right] \leq C|x - x'|, \end{aligned}$$

for all x, x' in $\bar{\mathcal{O}}$, and all stopping times θ .

Let us assume for the moment that u satisfies

$$(2.15) \quad |u(t, x)| \leq C|\bar{u}(x)| \quad \forall x \text{ in } \bar{\mathcal{O}}.$$

Then, using the equation of dynamical programming as in the proof of Theorem 3.1 in part I, it is easy to deduce

$$(2.16) \quad \begin{aligned} |u(t, x) - u(t, x')| &\leq CE \left[\left| \bar{u}(y_x(\tau_x \wedge \tau_{x'})) \exp \left(- \int_0^{\tau_x \wedge \tau_{x'}} c(y_x(s), v(s)) ds \right) \right. \right. \\ &\quad \left. \left. - \bar{u}(y_{x'}(\tau_x \wedge \tau_{x'})) \exp \left(- \int_0^{\tau_x \wedge \tau_{x'}} c(y_{x'}(s), v(s)) ds \right) \right| \right] \\ &\leq C\|\nabla \bar{u}\|_{L^\infty(\mathcal{O})} |x - x'| + C|x - x'|, \end{aligned}$$

using the inequality above (see also part I); this argument will be detailed further on.

Now to prove (2.15), we argue as follows. As in part I, we denote $u_t(s, x) = Q(t-s)0$ (where t is fixed). Using the dynamical programming property, we obtain (see also [14])

$$u(t, x) = \inf_{\mathcal{A}} E \left\{ \int_0^{\tau_x \wedge \tau_x^e \wedge t} f(y_x(s), v(s)) \exp \left(- \int_0^s c(y_x(\lambda), v(\lambda)) d\lambda \right) ds \right. \\ \left. + u_t(\tau_x^e, y_x(\tau_x^e)) \exp \left(- \int_0^{\tau_x^e} c(y_x(s), v(s)) ds \right) 1_{(\tau_x^e < t)} \right\},$$

where τ_x^e is the first time $y_x(t)$ reaches $\mathcal{O} - \mathcal{O}_\varepsilon$.

For all x in \mathcal{O}_ε , we deduce that

$$|u(t, x)| \leq \sup_{\mathcal{A}} E \left\{ \int_0^{\tau_x \wedge \tau_x^e \wedge t} |f(y_x(s), v(s))| \exp \left(- \int_0^s c(y_x(\lambda), v(\lambda)) d\lambda \right) ds \right. \\ \left. + C(-\bar{u}(y_x(\tau_x^e))) \exp \left(- \int_0^{\tau_x^e} c(y_x(s), v(s)) ds \right) 1_{(\tau_x^e < t)} \right\}.$$

Now, using a method due to [13], we deduce from (1.13) that

$$(-\bar{u}(x)) \geq \sup_{\mathcal{A}} E \left[\int_0^{\tau_x^e \wedge t} \exp \left(- \int_0^t c(y_x(\lambda), v(\lambda)) d\lambda \right) ds \right. \\ \left. + (-\bar{u}(y_x(t_x^e))) \exp \left(- \int_0^{\tau_x^e} c(y_x(s), v(s)) ds \right) 1_{(\tau_x^e < t)} \right];$$

thus, if we choose C large enough, (2.15) is proved.

For the generator of $Q(t)$ we have

THEOREM 2.3. *Under assumptions (1.9), (1.10) and (1.11) (or (1.15)), we have for any $h \in C^2(\mathcal{O})$*

$$(2.17) \quad \frac{1}{t} [Q(t)h(x) - h(x)] \rightarrow -\sup_{v \in V} \{A(v)h(x) - f(v, x)\} \quad \text{as } t \rightarrow 0+, \quad \forall x \in \mathcal{O}.$$

Moreover, the convergence in (2.17) is uniform on compact subsets of \mathcal{O} .

Proof. It is similar to that of Theorem 2.2. in part I. \square

2. Stochastic interpretation. Let us consider the optimal cost function

$$(2.18) \quad u(x) = \inf_{\mathcal{A}} E \left\{ \int_0^{\tau_x} f(y_x(t), v(t)) \exp \left(- \int_0^t c(y_x(s), v(s)) ds \right) dt \right\}.$$

We set

$$(2.19) \quad \Gamma_0 = \{x \in \Gamma / \exists \mathcal{A} \text{ admissible such that } P(\tau_x > 0) = 0\}.$$

Remark that if we assume (1.11), then $\Gamma_0 = \Gamma$. We have the following:

THEOREM 2.4. *Under assumptions (1.9), (1.10) and (1.15) (resp. (1.11)) the function $u(x)$ defined by (2.18) is the unique solution of the problem*

$$(2.20) \quad u \in B_s^+ \text{ (resp. } B_s), \quad u|_{\Gamma} = 0, \quad Q(t)u = u, \quad t \geq 0.$$

Moreover, the equation of dynamical programming is satisfied:

$$(2.21) \quad u(x) = \inf_{\mathcal{A}} E \left\{ \int_0^{\theta \wedge \tau_x} f(y_x(t), v(t)) \exp \left(- \int_0^t c(y_x(s), v(s)) ds \right) dt \right. \\ \left. + u(y_x(\theta \wedge \tau_x)) \exp \left(- \int_0^{\theta \wedge \tau_x} c(y_x(t), v(t)) dt \right) \right\}$$

where θ is an arbitrary stopping time.

Furthermore, if we suppose (1.11) and (2.11), the optimal cost u defined by (2.18) belongs to $W_0^{1,\infty}(\mathcal{O})$.

Remark 2.3. Equations (2.20), (2.21) show that the optimal cost function $u(x)$ satisfies in some integral sense the Hamilton–Jacobi–Bellman equation: $\sup_{v \in V} \{A(v)u - f(v)\} = 0$ in \mathcal{O} .

Proof. The proof of the first part is similar to that of Theorem 3.1 in part I. We will prove that under assumptions (1.13) and (2.11), u belongs to $W_0^{1,\infty}(\mathcal{O})$. To simplify notation we assume $c(x, v) \equiv c_0$.

We first prove that there exists some constant $C > 0$ such that

$$|u(x)| \leq C |-\bar{u}(x)| \quad \forall x \in \mathcal{O}.$$

Indeed, if we choose C large enough, this inequality is obvious if $x \in \mathcal{O} - \mathcal{O}_\varepsilon$. Now if $x \in \mathcal{O}_\varepsilon$, writing (2.21) with $\theta = \tau_x^\varepsilon$ where τ_x^ε is the first time $y_x(t)$ reaches Γ_ε we deduce

$$|u(x)| \leq \sup_{\mathcal{A}} E \left[\int_0^{\tau_x^\varepsilon \wedge \tau_x} C e^{-c_0 t} dt + C(-\bar{u}(y_x(\tau_x^\varepsilon))) 1_{(\tau_x^\varepsilon < \tau_x)} e^{-c_0 \tau_x^\varepsilon} \right].$$

Now, using (1.13), we have

$$-\bar{u}(x) \geq \sup_{\mathcal{A}} E \left[\int_0^{\tau_x^\varepsilon \wedge \tau_x} e^{-c_0 t} dt + (-\bar{u}(y_x(\tau_x^\varepsilon))) 1_{(\tau_x^\varepsilon < \tau_x)} e^{-c_0 \tau_x^\varepsilon} \right],$$

and we conclude.

We are now able to prove that $u \in W_0^{1,\infty}(\mathcal{O})$. Let $x, x' \in \mathcal{O}$; we have, using (2.21) with $\theta = \tau_{x'} \wedge \tau_x$,

$$|u(x) - u(x')| \leq \sup_{\mathcal{A}} CE \left\{ \int_0^\infty |y_x(t) - y_{x'}(t)| e^{-c_0 t} dt \right\} \\ + \sup_{\mathcal{A}} E \{ |u(y_x(\tau_x \wedge \tau_{x'})) - u(y_{x'}(\tau_x \wedge \tau_{x'}))| e^{-c_0 \tau_x \wedge \tau_{x'}} \}.$$

Because of (2.11) the first term is bounded by $|x - x'|$, while the second term is bounded by

$$C \sup_{\mathcal{A}} [1_{(\tau_x \leq \tau_{x'})} |\bar{u}(y_{x'}(\tau_x))| e^{-c_0 \tau_x} + 1_{(\tau_x > \tau_{x'})} |\bar{u}(y_x(\tau_{x'}))| e^{-c_0 \tau_{x'}}].$$

Since $\bar{u} \in W_0^{1,\infty}(\mathcal{O})$, this quantity is less than

$$C \sup_{\mathcal{A}} E[|y_x(\tau_x \wedge \tau_{x'}) - y_{x'}(\tau_x \wedge \tau_{x'})| e^{-c_0 \tau_x \wedge \tau_{x'}}] \leq C|x - x'|,$$

and the theorem is proved. \square

We also have

THEOREM 2.5. Assume (1.9), (1.10), (1.15), (1.16) and that $\inf_{x,v} c(x, v)$ is large enough. Then the optimal cost function u given by (2.18) belongs to $W_0^{1,\infty}(\mathcal{O})$.

Proof. From (1.18), we may define $\bar{u}(x)$ by

$$(2.22) \quad \bar{u}(x) = E \left\{ \int_0^{\tau_x} \|f\|_{\infty} e^{-c_0 t} dt \right\},$$

where the admissible system considered \mathcal{A} is given by the feedback $v(x)$ appearing in (1.16), which by a density argument may be assumed to be Lipschitz continuous.

Using barrier functions as in [14] or [18], it is easy to prove that $C\bar{u}(x) \leq C \text{dist}(x, \partial\mathcal{O})$. Then, if c_0 is large enough, this implies by a proof similar to that of Theorem 2.4 that $\bar{u} \in W_0^{1,\infty}(\mathcal{O})$. We have

$$(2.23) \quad 0 \leq u(x) \leq \bar{u}(x).$$

Next, using dynamical programming, we have

$$u(x) - u(x') \leq C|x - x'| + \sup_{\mathcal{A}} E \{ |u(y_x(\tau_x \wedge \tau_{x'})) - u(y_{x'}(\tau_x \wedge \tau_{x'}))| e^{-c_0 \tau_x \wedge \tau_{x'}} \}.$$

Hence, (2.23) gives (since $u = \bar{u}$ on $\partial\mathcal{O}$)

$$(2.24) \quad u(x) - u(x') \leq C|x - x'| + \sup_{\mathcal{A}} E \{ |\bar{u}(y_x(\tau_x \wedge \tau_{x'})) - \bar{u}(y_{x'}(\tau_x \wedge \tau_{x'}))| e^{-c_0(\tau_x \wedge \tau_{x'})} \}$$

and since $\bar{u} \in W_0^{1,\infty}(\mathcal{O})$, we deduce the result. \square

COROLLARY 2.1. Assume (1.9), (1.10), (1.15) and (1.16). Then the optimal cost u given by (2.18) is uniformly continuous in $\bar{\mathcal{O}}$. Moreover, for each $h \in X$ (given by (2.10)) $Q(t)h \in X$, so $Q(t)$ is a semigroup acting on X .

Proof. It is similar to that of Theorem 2.2. \square

Remark 2.4. Clearly, under assumptions (1.15) and (1.16), we have $\Gamma_0 = \Gamma$.

Remark 2.5. Using Theorem 2.4, we can prove a local version of Theorem 2.3 as in part I.

2.3. Analytical interpretation. In all of what follows, u will be the optimal cost function defined by (2.18). We have already seen that, under some assumptions, u belongs to $W^{1,\infty}(\mathcal{O})$. Then, we are able to show that u is the maximum subsolution of (1.5), and that is u is the envelope (sup) of all w in $W_0^{1,\infty}(\mathcal{O})$ satisfying

$$(2.25) \quad A(v)w \leq f(v) \quad \text{in } \mathcal{D}'(\mathcal{O}).$$

This result may be viewed as a notion of a generalized solution of (1.5) (as is done for Monge–Ampère equations). We thus give the following result (generalizing our previous one in part I).

Throughout this section we assume

$$(2.26) \quad \psi(\cdot, v) \in W^{2,\infty}(\mathcal{O}), \text{ and } \psi(\cdot, v) \text{ remains bounded in } W^{2,\infty}(\mathcal{O}) \\ \text{as } v \in V \text{ for all } \psi = \sigma_{ij}, b_i, c, f.$$

THEOREM 2.6. Assume (1.9), (1.10), (2.26) and (1.11) (or (1.15)). Then, for all w satisfying $w \in W_{\text{loc}}^{1,\infty}(\mathcal{O}) \cap C(\bar{\mathcal{O}})$, $w|_{\Gamma} \leq 0$ and $A(v)w \leq f(v)$ in $\mathcal{D}'(\mathcal{O})$, for all v in V , we have $w \leq u$ in $\bar{\mathcal{O}}$.

COROLLARY 2.2. Assume (1.9), (1.10) and either (1.11) and (2.11) or (1.15), (1.16) and c_0 large enough. Then u is the maximum element of the set of functions w satisfying $w \in W_{\text{loc}}^{1,\infty}(\mathcal{O}) \cap C(\bar{\mathcal{O}})$, $w|_{\Gamma} \leq 0$ and

$$A(v)w \leq f(v) \quad \text{in } \mathcal{D}'(\mathcal{O}), \quad \forall v \text{ in } V.$$

Remark 2.6. If we assume that \mathcal{O} is regular and (1.13), (1.14) hold, Theorem 2.6 is still valid.

Proof of Theorem 2.6. The proof of Theorem 2.6 is very similar to the one given in part I, provided we use a lemma due to [9]. Indeed, if w satisfies the conditions listed in the above theorem, we have (using part I and [9])

$$w(x) \leq \inf_{\mathcal{A}} E \left[\int_0^{\tau_x^h} f(y_x(t), v(t)) \exp \left(- \int_0^t c(y_x(s), v(s)) ds \right) dt \right. \\ \left. + w(y_x(\tau_x^h)) \exp \left(- \int_0^{\tau_x^h} c(y_x(t), v(t)) dt \right) \right],$$

where τ_x^h is the first exit time of $\mathcal{O}^h = \{x \in \mathcal{O}, \text{dist}(x, \partial\mathcal{O}) \geq h\}$.

Then, if we take $h \rightarrow 0$, $\tau_x^h \uparrow \sigma_x$, where σ_x is the first exit time of \mathcal{O} . Thus,

$$w(x) \leq \inf_{\mathcal{A}} E \left[\int_0^{\sigma_x} f(y_x(t), v(t)) \exp \left(- \int_0^t c(y_x(s), v(s)) ds \right) dt \right].$$

Now, if we assume (1.11), $\sigma_x = \tau_x$ a.s. and we conclude.

On the other hand, if we assume (1.15), as $\sigma_x \leq \tau_x$ by definition, we also deduce $w \leq u$. \square

Corollary 2.2 is immediately deduced from Theorem 2.6 as in part I. \square

3. The Cauchy problem. We now consider the optimal control of time-dependent diffusions (or solutions of stochastic differential equations). We consider coefficients $\sigma_{ij}(x, t, v)$, $b_i(x, t, v)$, $c(x, t, v)$, $f(x, t, v)$ which, for the sake of simplicity, will be assumed to belong to $W^{2,1,\infty}(\mathcal{O} \times]0, T[)$ for some $T > 0$, and for all v in V . In addition $\phi(x, t, v)$ remains bounded in $W^{2,1,\infty}(\mathcal{O} \times]0, T[)$ as $v \in V$, and $\phi(x, t, v)$ is continuous in $v \in V$ uniformly in $(x, t) \in \bar{\mathcal{O}} \times [0, T]$. These assumptions may be considerably relaxed but we will not consider such generalizations here.

We will denote $Q = \mathcal{O} \times]0, T[$. We define the optimal cost function

$$(3.1) \quad u(t, x) = \inf_{\mathcal{A}} E \left[\int_t^{T \wedge \tau_{x,t}} f(y_{x,t}(s), s, v(s)) \exp \left(- \int_t^s c(y_{x,t}(\lambda), \lambda, v(\lambda)) d\lambda \right) ds \right. \\ \left. + u_0(y_{x,t}(T)) \exp \left(- \int_t^T c(y_{x,t}(s), s, v(s)) ds \right) 1_{(T < \tau_{x,t})} \right],$$

where the infimum is taken over all admissible systems \mathcal{A} , and where an admissible system is defined exactly as before except for $y_{x,t}$, which is the solution of:

$$(3.2) \quad dy_{x,t}(s) = \sigma(y_{x,t}(s), s, v(s)) dW_s - b(y_{x,t}(s), s, v(s)) ds, \quad s \in [t, T], \\ y_{x,t}(t) = x.$$

Obviously $\tau_{x,t}$ denotes the exit time from $\bar{\mathcal{O}}$ of the process $y_{x,t}(s)$, and u_0 is a given function in $W^{2,\infty}(\mathcal{O})$ satisfying $u = 0$ on $\partial\mathcal{O}$.

Of course this time-dependent problem may be reduced to the general case of degenerate stochastic integrals by looking at the “space-time” diffusion $(y_{x,t}(s), s)$ starting at the point (x, t) of \bar{Q} ; then $(\tau_{x,t} \wedge T)$ is just the first exit time from \bar{Q} of this process. Instead of considering both situations (time independent and time dependent) in a same general context (and defining in particular a set Γ_0 of regular points) we prefer to give just the case of time-independent stochastic integrals and to indicate how the preceding results may be adapted to the above situation.

We will not give any proofs in this section, since they are only trivial adaptations of the methods introduced above. We only give some examples of our results.

THEOREM 3.1. *Assume either*

$$(3.3) \quad \begin{aligned} & -\frac{\partial \bar{u}}{\partial t} + A(v)\bar{u} \leq -1 \quad \text{in } \mathcal{D}'(\mathcal{O}_\varepsilon \times]0, T[) \quad \forall v \in V, \\ & \bar{u} \in C(\bar{Q}), \quad \bar{u}(x, T) \leq u_0(x), \quad |\nabla_x \bar{u}(x, t)| \leq C \quad \forall (x, t) \in Q, \\ & u \leq -\alpha < 0 \quad \text{on } (\mathcal{O}_\varepsilon) \times [0, T]; \end{aligned}$$

or

$$(3.4) \quad \begin{aligned} & f(x, t, v) \geq 0 \quad \forall (x, t, v) \in \bar{\mathcal{O}} \times [0, T] \times V, \\ & \exists v(t, x) \text{ continuous on } \Gamma \times [0, T] \text{ such that } \exists \alpha > 0 \text{ such that} \\ & \Gamma \times [0, T] = \{(x, t) / |\sigma(x, t, v)n(x)| \geq \alpha\} \\ & \quad \cup \{(x, t) / -2rb(x, t, v) \cdot n(x) > \text{Tr}[\sigma\sigma^*(x, t, v)] + \alpha\}, \\ & \text{where } r \text{ is the radius of the uniform exterior sphere associated to } \mathcal{O}. \end{aligned}$$

Then

i) *we have the dynamical programming property:*

$$\begin{aligned} u(t, x) = \inf_{\mathcal{A}} E \Bigg[& \int_t^{\theta \wedge T \wedge \tau_{x,t}} f(y_{x,t}(s), s, v(s)) \exp \left(- \int_t^s c(y_{x,t}(\lambda), \lambda, v(\lambda)) d\lambda \right) ds \\ & + u_0(y_{x,t}(T)) \exp \left(- \int_t^T c(y_{x,t}(s), s, v(s)) ds \right) \cdot 1_{(T < \tau_{x,t} \wedge \theta)} \\ & + u(y_{x,t}(\theta)) \exp \left(- \int_t^\theta c(y_{x,t}(s), s, v(s)) ds \right) \cdot 1_{(\theta < \tau_x)} \Bigg], \end{aligned}$$

where θ is a stopping time.

ii)

$$\begin{aligned} & u \in W^{1,\infty}(Q), \quad u = 0 \quad \text{on } \Gamma \times [0, T], \quad u = u_0 \text{ on } \bar{\mathcal{O}} \times \{T\}, \\ & -\frac{\partial u}{\partial t} + A(v)u \leq f(v) \quad \text{in } \mathcal{D}'(Q) \quad \forall v \in V. \end{aligned}$$

iii) *u is the maximum element of the set of functions w satisfying*

$$\begin{aligned} & w \in C(\bar{Q}), \quad w \leq 0 \text{ on } \Gamma \times [0, T], \quad w \leq u_0 \text{ on } \mathcal{O} \times \{T\}, \quad \nabla_x w \in L^\infty(Q), \\ & -\frac{\partial w}{\partial t} + A(v)w \leq f(v) \quad \text{in } \mathcal{D}'(Q) \quad \forall v \in V. \end{aligned}$$

This result is only one example of how the results of the preceding sections adapt to this problem of control of time-dependent stochastic integrals and to this Cauchy problem for Hamilton–Jacobi–Bellman equations.

Let us also mention that a general result concerning the verification of H–J–B equations is given in P.-L. Lions [13].

4. The obstacle problem. This section is divided into two parts. First we give a stochastic interpretation of the optimal cost. Next, we establish an analytical interpretation.

4.1. Stochastic interpretation. Let $\Psi(x)$ be a function from \mathbb{R}^N into \mathbb{R} satisfying

$$(4.1) \quad \begin{aligned} |\Psi(x) - \Psi(x')| &\leq \rho(|x - x'|) \quad \forall x, x' \in \mathbb{R}^N, \\ \Psi(x) &\geq 0 \quad \forall x \in \Gamma_0 \quad (\text{given by (2.19)}), \\ |\Psi(x)| &\leq C \quad \forall x \in \mathbb{R}^N, \end{aligned}$$

where ρ is a given continuous function from \mathbb{R}^+ into \mathbb{R}^+ such that $\rho(0) = 0$.

In some results, we will also assume

$$(4.2) \quad \Psi(x) \geq 0 \quad \forall x \in \mathbb{R}^N.$$

Let us define the cost function

$$(4.3) \quad \begin{aligned} J_x(\mathcal{A}, \theta) = E \Big\{ &\int_0^{\theta \wedge \tau_x} f(y_x(t), v(t)) \exp \left(- \int_0^t c(y_x(s), v(s)) ds \right) dt \\ &+ 1_{\theta < \tau_x} \Psi(y_x(\theta \wedge \tau_x)) \exp \left(- \int_0^{\theta \wedge \tau_x} c(y_x(t), v(t)) dt \right) \Big\}, \end{aligned}$$

where \mathcal{A} is any admissible system and θ is a stopping time with respect to F_t .

The optimal cost function is given by

$$(4.4) \quad u(x) = \inf \{J_x(\mathcal{A}, \theta) / \mathcal{A}, \theta\}.$$

We have the following:

THEOREM 4.1. *Under assumptions (1.9), (1.10), (1.11) and (4.1) (resp. (1.9), (1.10), (1.15) and (4.2)) the function u defined by (4.4) is the maximum solution of the following problem:*

$$(4.5) \quad \begin{aligned} u &\in B_s \quad (\text{resp. } B_s^+), \quad u|_{\Gamma_0} = 0, \\ u &\leq \Psi \quad \text{in } \bar{O}, \\ u &\leq Q(t)u \quad \forall t \geq 0, \end{aligned}$$

where $Q(t)$ is the semigroup (1.8).

Proof. Let $\delta(t, \omega)$ be an adapted process such that $0 \leq \delta(t) \leq 1$ for all $t \geq 0$.

Let us define for $\varepsilon > 0$:

$$(4.6) \quad \begin{aligned} J_x^\varepsilon(\mathcal{A}, \delta) = E \Big\{ &\int_0^{\tau_x} \left[f(y_x(t), v(t)) + \frac{1}{\varepsilon} \delta(t) \Psi(y_x(t)) \right] \\ &\cdot \exp \left(- \int_0^t (c(y_x(s), v(s)) + \frac{1}{\varepsilon} \delta(s)) ds \right) dt \Big\} \end{aligned}$$

and

$$(4.7) \quad u_\varepsilon(x) = \inf \{J_x^\varepsilon(\mathcal{A}, \delta) / \mathcal{A}, \delta\}.$$

From Theorem 2.1 we have

$$(4.8) \quad u_\varepsilon \in B_s \quad (\text{resp. } B_s^+) \quad \text{and} \quad u_\varepsilon|_{\Gamma_0} = 0.$$

First, we prove that

$$(4.9) \quad \begin{aligned} u_\varepsilon(x) = \inf_{\mathcal{A}} E \Big\{ &\int_0^{\tau_x} \left[f(y_x(t), v(t)) - \frac{1}{\varepsilon} (u_\varepsilon - \Psi)^+(y_x(t)) \right] \\ &\cdot \exp \left(- \int_0^t c(y_x(s), v(s)) ds \right) dt \Big\}. \end{aligned}$$

Indeed, from (4.7) and the dynamical programming used for the function u_ε , we deduce that the process

$$\begin{aligned} \xi(t) = & \int_0^{t \wedge \tau} \left[f(y(s), v(s)) + \frac{1}{\varepsilon} \delta(s) \Psi(s) \Psi(y(s)) \right] \\ & \cdot \exp \left(- \int_0^s \left(c(y(\lambda), v(\lambda)) + \frac{1}{\varepsilon} \delta(\lambda) \right) d\lambda \right) ds \\ & + u_\varepsilon(y(t \wedge \tau)) \exp \left(- \int_0^{t \wedge \tau} \left(c(y(s), v(s)) + \frac{1}{\varepsilon} \delta(s) \right) ds \right) \end{aligned}$$

is a submartingale for each admissible system \mathcal{A} . Setting

$$\xi(\infty) = \int_0^\tau \left[f(y(s), v(s)) + \frac{1}{\varepsilon} \delta(s) \Psi(y(s)) \right] \exp \left(- \int_0^s \left(c(y(\lambda), v(\lambda)) + \frac{1}{\varepsilon} \delta(\lambda) \right) d\lambda \right) ds$$

we obtain from (4.8) for $\eta(t) = E^F \xi(\infty) - \xi(t)$

$$(4.10) \quad 0 \leq \eta(t) \leq C \exp \left[- \int_0^t \left(c_0 + \frac{1}{\varepsilon} \delta(s) \right) ds \right] \quad \forall t \geq 0.$$

The process $\xi(t) + \eta(t)$ is a F_t -martingale, so the process

$$\begin{aligned} Z(t) = & \eta(t) \exp \left(- \int_0^t \frac{1}{\varepsilon} \delta(s) ds \right) + u_\varepsilon(y(t \wedge \tau)) \exp \left(- \int_0^{t \wedge \tau} c(y(s), v(s)) ds \right) \\ & + \int_0^{t \wedge \tau} \left[f(y(s), v(s)) - \frac{1}{\varepsilon} \delta(s) (u_\varepsilon - \Psi)(y(s)) \right] \exp \left(- \int_0^s c(y(\lambda), v(\lambda)) d\lambda \right) ds \\ & - \frac{1}{\varepsilon} \int_0^t \eta(s) \delta(s) \exp \left(- \int_0^s \frac{1}{\varepsilon} \delta(\lambda) d\lambda \right) ds \end{aligned}$$

is a F_t -martingale too. Since $EZ(0) = EZ(t)$ and $\eta(t) \geq 0$, choosing

$$\delta(s) = \begin{cases} 1 & \text{if } \Psi(y(s)) < u_\varepsilon(y(s)), \\ 0 & \text{if } \Psi(y(s)) \geq u_\varepsilon(y(s)) \end{cases}$$

and taking the limit for $t \rightarrow \infty$ we deduce, using (4.10),

$$(4.11) \quad \begin{aligned} u_\varepsilon \leq & E \left\{ \int_0^\tau \left[f(y(s), v(s)) - \frac{1}{\varepsilon} (u_\varepsilon - \Psi)(y(s)) \right] \right. \\ & \cdot \exp \left(- \int_0^s c(y(\lambda), v(\lambda)) d\lambda \right) ds \Big\} \quad \forall \mathcal{A}. \end{aligned}$$

Next, given a constant $k > 0$, there exists $\mathcal{A} = \mathcal{A}_{k,x}$ such that

$$E\eta(0) \leq k.$$

Hence, using the fact that $EZ(0) = EZ(t)$, we have

$$\begin{aligned} u_\varepsilon \geq & -k \left(1 + \frac{1}{\varepsilon} t \right) - C e^{-c_0 t} + \inf_{\mathcal{A}} E \left\{ \int_0^\tau \left[f(y(t), v(t)) - \frac{1}{\varepsilon} (u_\varepsilon - \Psi)^+(y(t)) \right] \right. \\ & \cdot \exp \left(- \int_0^t c(y(s), v(s)) ds \right) dt \Big\}, \end{aligned}$$

where C is a constant independent of $t \geq 0$.

Thus, with $t \rightarrow \infty$ this proves (4.9).

Now, in a classical way (cf. A. Bensoussan–J.-L. Lions [1]), we deduce from (4.9) that

$$(4.12) \quad u_\varepsilon \rightarrow u \quad \text{as } \varepsilon \rightarrow 0 \quad \text{uniformly in } \bar{\mathcal{O}}$$

and

$$(4.13) \quad u(x) \leq u_\varepsilon(x) \leq E \left\{ \int_0^{\theta \wedge \tau_x} f(y_x(t), v(t)) \exp \left(- \int_0^t c(y_x(s), v(s)) ds \right) dt \right. \\ \left. + u_\varepsilon(y_x(\theta \wedge \tau_x)) \exp \left(- \int_0^{\theta \wedge \tau} c(y(t), v(y)) dt \right) \right\} \quad \forall \mathcal{A}, \theta.$$

Hence, we show that $u(x)$ is a solution of problem (4.5).

Finally, the same arguments as above prove that u is the maximal solution \square .

In order to obtain some results of regularity of the optimal cost $u(x)$ we assume

$$(4.14) \quad |\Psi(x) - \Psi(x')| \leq C|x - x'| \quad \forall x, x' \in \mathbb{R}^N,$$

we have

THEOREM 4.2. *Assume (1.9), (1.11), (2.11), (4.1), (4.14) and*

$$(4.15) \quad C\bar{u}(x) \leq \Psi(x) \quad \text{in } \bar{\mathcal{O}}, \quad \text{for some } C > 0,$$

or assume (1.9), (1.14), (1.15), (1.16), (4.1), (4.2), (4.14) and c_0 large enough in (1.10). Then the optimal cost function u belongs to $W_0^{1,\infty}(\mathcal{O})$.

Proof. As in [14] or [1] (and using the proof of Theorem 4.1) we have that u satisfies the dynamical programming property, i.e.,

$$(4.16) \quad u(x) = \inf_{\theta} E \left[\int_0^{\tau_x \wedge \theta} f(y_x(t), v(t)) \exp \left(- \int_0^t c(y_x(s), v(s)) ds \right) dt \right. \\ \left. + u(y_x(\tau)) \exp \left(- \int_0^{\tau} c(y_x(t), v(t)) dt \right) 1_{(\tau < \theta \wedge \tau_x)} \right. \\ \left. + \Psi(y_x(\theta)) \exp \left(+ \int_0^{\theta} c(y_x(t), v(t)) dt \right) 1_{(\theta < \tau \wedge \tau_x)} \right],$$

where τ is any stopping time.

Now using (4.15), we deduce as before (in similar situations)

$$|u(x) \leq C|\bar{u}(x)|.$$

Then the same methods as before give the Lipschitz character of u . \square

COROLLARY 4.1. *Under assumptions (1.9), (1.10), (1.11), (4.1) or (1.9), (1.10), (1.14), (1.15), (1.16) and (4.1), the optimal cost function u is uniformly continuous on $\bar{\mathcal{O}}$.*

The proof of this result uses the same argument as in Theorem 2.2.

4.2. Analytical interpretation. We will in this section just state some results which are proved with the same techniques as in § 2. These are examples of how our techniques apply to the obstacle problem.

We first prove that under fairly general conditions u is a “generalized solution” of

$$(4.17) \quad \sup_{v \in V} [\sup \{A(v)u - f(v)\}, u - \Psi] = 0 \quad \text{a.e. in } \mathcal{O}, \\ u = 0 \quad \text{in } \Gamma = \partial\mathcal{O}.$$

THEOREM 4.3. i) Under assumptions (1.9), (1.10), (2.26), (4.1) and (1.11) (or (1.15) and (4.2)), for all w satisfying $w \in W_{\text{loc}}^{1,\infty}(\mathcal{O}) \cap C(\bar{\mathcal{O}})$, $w|_{\Gamma} \leq 0$, $A(v)w \leq f(v)$ in $\mathcal{D}'(\mathcal{O})$, for all v in V , $w \leq \Psi$ in \mathcal{O} , we have

$$w \leq u \quad \text{in } \bar{\mathcal{O}}.$$

ii) Under the assumptions of Theorem 4.2 and if we assume in addition (2.26), then u is the maximum element of the set of functions w satisfying $w \in W_{\text{loc}}^{1,\infty}(\mathcal{O}) \cap C(\bar{\mathcal{O}})$; $w|_{\Gamma} \leq 0$ and

$$A(v)w \leq f(v) \quad \text{in } \mathcal{D}'(\mathcal{O}) \quad \forall v \in V, \quad w \leq \Psi \quad \text{in } \mathcal{O}.$$

REFERENCES

- [1] A. BENSOUSSAN AND J.-L. LIONS, *Applications des inéquations variationnelles en contrôle stochastique*, Dunod, Paris, 1978.
- [2] L. C. EVANS AND P.-L. LIONS, *Résolutions des equations de Hamilton-Jacobi-Bellman pour des opérateurs uniformément elliptiques*, CRAS Paris, 290 (1980), pp. 1049–1052.
- [3] I. L. GENIS AND N. V. KRYLOV, *An example of a one-dimensional controlled process*, Theory Prob. Appl., 21 (1976), pp. 148–152.
- [4] R. GONZALEZ, *Sur l'existence d'une solution maximale de l'équation de Hamilton-Jacobi*, CRAS Paris, 282 (1976), pp. 1287–1290.
- [5] N. V. KRYLOV, *Control of a solution of a stochastic integral equation*, Theory Prob. Appl., 17 (1972), pp. 114–131.
- [6] ———, *On control of the solution of a stochastic integral equation with degeneration*, Izv. Akad. Nauk. USSR, 36 (1972), pp. 249–262.
- [7] ———, *Controlled Diffusion Processes*, Springer-Verlag, New York, 1980.
- [8] P.-L. LIONS, *Résolution analytique des problèmes de Bellman-Dirichlet*, Acta Mathematica (1981), to appear; cf. also CRAS Paris, 287 (1978), pp. 747–750.
- [9] ———, *Control of diffusion processes in \mathbb{R}^N* , Comm. Pure Appl. Math. (1981), to appear; cf. also CRAS Paris, 288 (1979), pp. 339–342.
- [10] ———, *Some problems related to the Bellman-Dirichlet equation for two elliptic operators*, Comm. PDE, to appear; cf. also thèse de 3ème cycle Paris VI, 1978.
- [11] ———, *Equations de Hamilton-Jacobi-Bellman dégénérées*, CRAS Paris, 289 (1979), pp. 329–332.
- [12] ———, *Equations de Hamilton-Jacobi-Bellman*, in Séminaire Goulaumie-Schwartz, 1979–1980.
- [13] ———, *The Dirichlet problem for Hamilton-Jacobi-Bellman equations and optimal stochastic control*, to appear.
- [14] P.-L. LIONS AND J.-L. MENALDI, *Optimal control of stochastic integrals and Hamilton-Jacobi-Bellman equations*, I, this Journal, this issue, pp. 58–81.
- [15] ———, *Problèmes de Bellman avec le contrôle dans les coefficients de plus haut degré*, CRAS Paris, 287 (1978), pp. 409–412.
- [16] J.-L. MENALDI, *On the optimal stopping time problem for degenerate diffusions*, this Journal, 18 (1980), pp. 697–721; see also CRAS Paris, 284 (1977), pp. 1443–1446.
- [17] M. NISIO, *On a nonlinear semi-group attached to stochastic optimal control*, Publ. Res. Inst. Math. Sci., 12 (1976–77), pp. 513–537.
- [18] M. V. SAFONOV, *On the Dirichlet problem for Bellman's equation in a plane domain*, Math. USSR Sb., 31 (1977), pp. 231–248.
- [19] ———, *On the Dirichlet problem for Bellman's equation in a plane domain*, Math. USSR Sb., 34 (1978), pp. 521–526.

LINEAR SYSTEMS WITH INDIRECT CONTROLS: THE UNDERLYING MEASURES*

ZVI ARTSTEIN† AND GILEAD TADMOR†

Abstract. The control decisions in our model are functions defined on a measurable space. They affect the evolution of the linear system in an indirect and nonlocal manner. A delayed control is one example. We introduce the notion of an underlying measure of such a system. Mere existence of an underlying measure yields information concerning the set of attainability, bang-bang and optimal solutions. We study the underlying measures, provide tools to compute them and relate their properties to the structure of the control system.

1. Introduction. The action of the control function $\mathbf{u} = u(t)$ on the linear system

$$(1.1) \quad \dot{x} = A(t)x + B(t)u(t)$$

is direct in the sense that the local behavior of the evolution $x(t)$ is affected only by the local behavior of the control $u(t)$ at the time t . Natural applications give rise to mechanisms of indirect actions, where the decisions in the control function \mathbf{u} are shifted, twisted or combined before affecting the evolution. An example which is treated in the literature is the delayed action. An appropriate model for a finite number of pure delays is

$$(1.2) \quad \dot{x} = A(t)x + \sum_{i=1}^k B_i(t)u(t - h_i(t)),$$

with $h_i(t) \geq 0$. More general delayed terms can also be incorporated, e.g., in the model

$$(1.3) \quad \dot{x} = A(t)x + \int_{t-a}^t u(s) d_s H(t, s).$$

Here the integration is in the Stieltjes-Lebesgue sense with respect to the matrix-valued function of bounded variation $H(t, s)$. A systematic analysis of the systems (1.2) and (1.3) and of more general equations can be found in Banks and Jacobs [1970], Banks, Jacobs and Latina [1971], Chyung and Lee [1970], Lee [1968], [1969] and the references therein.

This paper contributes to the understanding of systems (1.2) and (1.3), but the analysis relates to a more general situation. The control decisions in our framework are functions \mathbf{u} which are defined on a space which might be different from the time scale. The entire choice \mathbf{u} then affects the evolution in a linear and clearly indirect mechanism. It is easy to construct meaningful examples, and one is given in Example 5.9. An abstract model which incorporates such behavior is

$$(*) \quad \dot{x} = A(t)x + \int_S u(s) d\mu_t(s),$$

where S is a measurable space and for each t the integration is done with respect to the measure μ_t . The technical assumptions of the model are presented in § 2. (A more general situation might be considered where \mathbf{u} is not necessarily a function. This case is not covered in the paper.)

* Received by the editors October 24, 1980, and in revised form March 18, 1981.

† Department of Theoretical Mathematics, Weizmann Institute of Science, Rehovot, Israel. This research was supported by a grant from the United States-Israel Binational Science Foundation, Jerusalem, Israel.

The following is a very useful observation concerning equation (1.1). Consider the bounded and measurable controls on an interval $[t_0, t_1]$ as elements of the Banach space $L_\infty(\lambda)$, λ being the Lebesgue measure. Suppose the initial condition $x(t_0)$ is fixed; then for any control \mathbf{u} there is a unique solution $x(t, \mathbf{u})$ of (1.1). The observation is that $x(t, \mathbf{u})$ is affine in \mathbf{u} , and continuous when the collection of controls \mathbf{u} is endowed with the weak* topology of $L_\infty(\lambda)$. The advantage of this observation is twofold. Bounded sets are precompact in the weak* topology. The continuity then yields existence results for optimal controls. The affine character of the mapping together with the continuity imply maximal principles and the existence of bang-bang controls; see Hermes and LaSalle [1969, Part II], or Lee and Markus [1967, Chapt. 2] (the latter text employs the weak topology of $L_2(\lambda)$ instead of the weak* topology of $L_\infty(\lambda)$). The second aspect of the continuity is that it indicates a concept of smallness for which small perturbations of the control \mathbf{u} result in only small perturbations of the outcome $x(t, \mathbf{u})$. In particular a change on a set of Lebesgue measure zero would not change the outcome at all.

The machinery just described is not available a priori when the systems (1.2), (1.3) or (*) are considered. It is trivial to show that $x(t, \mathbf{u})$ is an affine mapping, but, since for each t the action of \mathbf{u} on (*) is governed by a different measure, μ_t , there is no clear candidate for the role that the Lebesgue measure plays in (1.1). The main contribution of this paper is to prove the existence of such a measure and to supply means to compute it. We call such a measure an underlying measure for (*). We also prove existence of a minimal underlying measure, relative to the partial order of absolute continuity. A minimal underlying measure yields a coarser topological structure on the controls \mathbf{u} , hence stronger results. These subjects are dealt with in § 3.

Since a continuum of measures participates in (*) the only available, a priori, well-defined controls are the bounded ones. However, the construction of the minimal underlying measure in § 2 provides a larger class of admissible controls; namely all the controls integrable with respect to that underlying measure. This is discussed in § 4.

Various properties of the underlying measures, and a few examples, are listed in § 5. The particular properties of being atomless or being absolutely continuous with respect to a prescribed measure, are discussed in § 6.

The mere existence of the underlying measures enables us to derive a variety of results following techniques which are common in the analysis of the ordinary systems (1.1). However, this approach yields further results typical to indirect controls. Such is the bang-bang principle of § 7 which takes into account atomic points. In § 8 we analyze the set of attainability, which lays the foundation for the geometric approach to optimization. Further applications are due in a forthcoming paper.

The entire presentation of the text is with regard to the system (*). In particular the uncontrolled part is governed by an ordinary differential equation, this is done for simplicity of presentation. All the results can be applied to more general equations. In the final section we show how to handle delays in the states and discrete systems, and how to incorporate a nonlinearity into the model.

2. Notation, terminology and standing hypotheses. We set first our measure theoretic notation and terminology, then present the model and discuss some examples. A measure on a measurable space (S, Σ) is a countably additive set-function from Σ into a normed space. The variation of a measure (or a set-function) ν is an extended-real measure (or a set-function), denoted by $|\nu|$ and defined by $|\nu|(E) = \sup \sum |\nu(E_i)|$ where the supremum is taken over all disjoint unions $E = \cup E_i$, with $E_i \in \Sigma$. Here $|\nu(E_i)|$ is the norm of $\nu(E_i)$. Where we encounter matrix-valued measures, the norm of a matrix

will be arbitrarily specified. The spaces L_∞ and L_1 are the Banach spaces of equivalence classes of the essentially bounded functions and the integrable functions, respectively. Both are determined if the domain (S, Σ) , the measure ν on it and the range space of the functions, are specified. In the text we indicate which is which (e.g., write $L_1 = L_1(S, \nu, R^k)$) if necessary, but suppress either of the indicators when the meaning is transparent.

The control system

$$(*) \quad \dot{x} = A(t)x + \int_S u(s) d\mu_t(s)$$

is defined for $t \in [t_0, t_1]$, and solutions are assumed to satisfy a given initial condition $x(t_0) = x_0$. Here $x \in R^n$ the n -dimensional Euclidean space, and \dot{x} denotes the derivative with respect to time. The demand from the uncontrolled part is as described in the following assumption.

Assumption I. $A(t)$ is an $n \times n$ matrix-valued function integrable over $[t_0, t_1]$.

The control functions $\mathbf{u} = u(s)$ are defined on a measurable space (S, Σ) . An *admissible control* is a bounded measurable mapping from S into R^m . The action of an admissible control \mathbf{u} on $(*)$ is determined by μ_t , for $t \in [t_0, t_1]$. Each μ_t is an $n \times m$ matrix-valued measure. The integration $\int_S u(s) d\mu_t(s)$ produces a vector in R^n whose i th coordinate is $\sum_{j=1}^m \int_S u_j(s) d\mu_t^{ij}(s)$. Here u_j and μ_t^{ij} denote the j th and the (i, j) th entries, respectively. (A more rigorous notation would be $\int_S d\mu_t(s) u(s)$, but we are accustomed to the other way.) We have the following standing hypothesis (see also Remark 2.2):

Assumption II. For each $E \in \Sigma$ the mappings $t \rightarrow \mu_t(E)$ and $t \rightarrow |\mu_t|(E)$ are measurable, and $t \rightarrow |\mu_t|(S)$ is integrable.

In the three systems (1.1)–(1.3), the space (S, Σ) is the real line with its Borel or Lebesgue structures. The measures μ_t related to (1.1) are atomic measures, concentrated on $\{t\}$ and having the matrices $B(t)$ as values. Each of the measures μ_t related to (1.2) has k atoms, concentrated at $t - h_i(t)$, $i = 1, \dots, k$, and with values $B_i(t)$, respectively. For (1.3) the measures μ_t are determined by $H(t, s)$. It is clear how the regularity expressed in Assumption II translates into terms of (1.1), (1.2) or (1.3). In particular, if $H(t, s)$ is measurable in the two variables simultaneously (as assumed, e.g., by Banks and Jacobs [1970, page 463]), then the measurability part of Assumption II is met.

If Assumption II is fulfilled, then for every admissible control $\mathbf{u} = u(s)$ the expression $\int_S u(s) d\mu_t(s)$ is well defined and forms an integrable function from $[t_0, t_1]$ into R^n . Equation $(*)$ then has a unique solution which we denote by $x(t, \mathbf{u})$. It is clear that $x(t, \mathbf{u})$ is an affine function of \mathbf{u} , but so far no structure of equivalence classes is available for the family of bounded and measurable controls.

DEFINITION A. A nonnegative scalar measure μ on (S, Σ) is an *underlying measure* for $(*)$ if $x(t, \mathbf{u})$ is well defined on the equivalence classes $L_\infty(\mu, R^m)$ and furthermore, if for each fixed t , $x(t, \mathbf{u})$ is continuous in \mathbf{u} with respect to the weak* topology of $L_\infty(\mu, R^m)$. The underlying measure μ is *minimal* if it is absolutely continuous with respect to any other underlying measure.

The previous definition requires continuity of $x(t, \mathbf{u})$ for each t separately. It is sometimes desirable to have uniform continuity in t , i.e., continuity of the entire trajectory $x(\cdot, \mathbf{u})$ as a mapping into $C([t_0, t_1], R^n)$, the space of continuous functions with the sup norm. Simple examples, e.g., $\dot{x} = u$ for scalar x , show that as a mapping of \mathbf{u} with the weak* topology the solution is not uniformly continuous in t on the entire space of controls \mathbf{u} . The reason is that the range is infinite dimensional but

open sets in the weak* topology contain unbounded cones of finite codimensionality. However, if \mathbf{u} is restricted to bounded sets (a situation often encountered) the desired result holds, as follows.

PROPOSITION 2.1. *If μ is an underlying measure for (*) then on norm-bounded sets of controls $\mathbf{u} \in L_\infty(\mu)$ the continuity of $x(t, \mathbf{u})$ in \mathbf{u} (with respect to the weak* topology) is uniform in t .*

Proof. Solutions $x(t, \mathbf{u})$, for \mathbf{u} in a bounded set, are equicontinuous in t . This is easily seen from (*). Thus uniform continuity in t follows from separate continuity for each t .

Remark 2.2. We want to elaborate on the measurability condition of Assumption II. This is a measurability demand of the mapping $t \rightarrow \mu_t$, namely a mapping into the space of matrix-valued measures. The demand $t \rightarrow \mu_t(E)$ is measurable for every E , is weaker than the weak measurability, and certainly weaker than the strong measurability of the vector-valued mappings (see, e.g., Hille and Phillips [1957, p. 72]). Notice that the mapping $t \rightarrow \mu_t$ related to (1.1) is never strongly measurable. The measurability of $t \rightarrow |\mu_t|(E)$ does not follow in general from the measurability of $t \rightarrow \mu_t(E)$ for every E . Example 3.6 below provides a counterexample. If, however, Σ is countably generated then $t \rightarrow \mu_t(E)$ measurable for every E implies that $t \rightarrow |\mu_t|(E)$ is measurable. The proof of Lemma 3.5 below contains a comment on this matter.

3. The underlying measures. The following two theorems display our main results concerning existence and calculation of the underlying measures. The system is (*) with the assumptions of the previous section. For convenience, we choose the operator norm as a norm for the space of $n \times m$ matrices; that is, $|B| = \max \{|Bz| : |z| = 1\}$. The norms $|z|$ and $|Bz|$ on R^m and R^n are arbitrarily specified.

THEOREM A. *Consider the set-function M which associates with $E \in \Sigma$ the function on $[t_0, t_1]$ given by $M(E)(t) = \mu_t(E)$. Then M is a measure of bounded variation into the space of integrable matrix-valued functions on $[t_0, t_1]$ with the L_1 -norm. Its variation $|M|$ is a minimal underlying measure for (*).*

THEOREM B. *The set-function $\bar{\mu}(E) = \int_{t_0}^{t_1} |\mu_t|(E) dt$ is an underlying measure for (*). If Σ is countably generated then $\bar{\mu}$ is a minimal underlying measure.*

The proofs follow some observations which are needed later on.

LEMMA 3.1. *The set-function $\bar{\mu}$ is a real, nonnegative finite measure.*

Proof. $\bar{\mu}$ is well defined in view of Assumption II and the trivial inequality $|\mu_t|(E) \leq |\mu_t|(S)$. Additivity of $\bar{\mu}$ is obvious. If E_k is a sequence of mutually disjoint sets in Σ then $|\mu_t|(\bigcup_{i \leq k} E_i)$ converges monotonically to $|\mu_t|(\bigcup E_i)$ as $k \rightarrow \infty$. Since the latter is integrable it follows that the integrals $\int_{t_0}^{t_1} |\mu_t|(\bigcup_{i \leq k} E_i) dt$ converge to $\int_{t_0}^{t_1} |\mu_t|(\bigcup E_i) dt$. This establishes the countable additivity of $\bar{\mu}$.

LEMMA 3.2. *The set function M is countably additive and with bounded variation. Furthermore $|M|(E) \leq \bar{\mu}(E)$ for every $E \in \Sigma$.*

Proof. Additivity of M is obvious. The variation $|M|(E)$, with respect to the L_1 -norm in the range, is the supremum of the expressions

$$\sum_{i=1}^k \int_{t_0}^{t_1} |\mu_t(E_i)| dt,$$

taken over all disjoint unions $E = \bigcup_{i=1}^k E_i$. The displayed expression is less than or equal to

$$\sum_{i=1}^k \int_{t_0}^{t_1} |\mu_t|(E_i) dt$$

and the latter is equal to $\bar{\mu}(E)$. This proves that $|M|(E) \leq \bar{\mu}(E)$. Since by Lemma 3.1, $\bar{\mu}$ is a finite and countably additive nonnegative measure, it follows that M is also countably additive.

Once it is established that M is a vector measure with bounded variation into the space of integrable matrix-valued functions, the integral of a function $u: S \rightarrow R^m$ with respect to M can be performed; see, e.g., Dunford and Schwartz [1958, IV.10]. (The latter text treats integration of scalar functions; here we have a vector $\mathbf{u} = u(s)$ of scalar functions but there is no essential difference.) The value $\int_S u(s) dM(s)$ is a function in $L_1([t_0, t_1], R^n)$, which we denote by $M(\mathbf{u})$. Thus M can be viewed as a linear bounded operator from $L_1(S, M, R^m)$ into $L_1([t_0, t_1], R^n)$ with norm 1. In particular, M is now a linear bounded operator from $L_\infty(S, |M|, R^m)$ into $L_1([t_0, t_1], R^n)$.

LEMMA 3.3. $M(\mathbf{u})(t) = \int_S u(s) d\mu_t(s)$ for every admissible control $\mathbf{u} = u(s)$.

Proof. The equality for step functions follows immediately from the definition of M . The additivity of the two expressions together with Lemma 3.2 imply equality for any admissible control.

LEMMA 3.4. Consider the equation $\dot{x} = A(t)x + h(t)$, $x(t_0) = x_0$ with $\mathbf{h} = h(t) \in L_1([t_0, t_1], R^n)$. For a fixed t the solution $x(t, \mathbf{h})$ is continuous in \mathbf{h} with respect to the weak topology of L_1 .

Proof. For the ordinary differential equation in question, the result follows from the variation of parameters formula, see, e.g., Lee and Markus [1967, p. 65]. We shall provide an independent proof which generalizes easily to other equations. The solution $x(\tau, \mathbf{h})$ is the unique fixed point, in $C([t_0, t_1], R^n)$, of the operator

$$(3.1) \quad T(\varphi, \mathbf{h}) = x_0 + \int_{t_0}^t A(\tau)\varphi(\tau) d\tau + \int_{t_0}^t h(\tau) d\tau.$$

Let \mathbf{h}_k converge to \mathbf{h}_0 in the norm of L_1 . If φ ranges over a bounded set then all the values $T(\varphi, \mathbf{h}_i)$ are contained in a compact set of $C([t_0, t_1], R^n)$. It is easy to see then that if φ_i is a fixed point of (3.1) with $\mathbf{h} = \mathbf{h}_i$ then any cluster point of the φ_i is a fixed point of (3.1) for $\mathbf{h} = \mathbf{h}_0$. Since the solution is unique it follows that $x(\cdot, \mathbf{h}_i)$ converges in $C([t_0, t_1], R^n)$ to $x(\cdot, \mathbf{h}_0)$. In particular the affine mapping $x(t, \mathbf{h})$, for t fixed, is continuous with respect to the norm topology of L_1 . Since, for a fixed t , the mapping $x(t, \mathbf{u})$ is into a finite dimensional space, it follows that the continuity holds with respect to the weak topology on L_1 . This completes the proof.

Proof of Theorem A. Lemma 3.2 (actually the discussion following it) and Lemma 3.3 imply that, if \mathbf{u}_1 and \mathbf{u}_2 differ only on an $|M|$ -null set, then the controlled part of (*) is the same for \mathbf{u}_1 and \mathbf{u}_2 . This proves that $x(t, \mathbf{u})$ is indeed well defined on $L_\infty(|M|, R^m)$.

The operator $\mathbf{u} \rightarrow M(\mathbf{u})$ is bounded, hence continuous, from $L_1(|M|, R^m)$ into $L_1([t_0, t_1], R^n)$, hence continuous also with respect to the weak topologies in the two spaces (see, e.g., Hille and Phillips [1957, Thm. 2.11.11]). Convergence in the weak* topology of L_∞ implies convergence in the weak topology of L_1 , hence $\mathbf{u} \rightarrow M(\mathbf{u})$ is continuous from the weak* topology of $L_\infty(|M|, R^m)$ into the weak topology of $L_1([t_0, t_1], R^n)$. The desired continuity of $x(t, \mathbf{u})$ in \mathbf{u} follows now from Lemma 3.4.

It remains to check the minimality of $|M|$. Let ν be a real nonnegative measure such that $|M|$ is not absolutely continuous with respect to ν . Then an $E \in \Sigma$ exists with $\nu(E) = 0$, but $|M|(E) \neq 0$. For a certain subset E_1 of E we have $M(E_1) \neq 0$. Let u_0 be a vector in R^n such that $M(E_1)u_0 \neq 0$. Define \mathbf{u}_0 by $\mathbf{u}_0(s) = u_0$ if $s \in E_1$, and $\mathbf{u}_0(s) = 0$ otherwise. Then $M(\mathbf{u}_0)$ is not the zero function. In particular $x(t, \mathbf{u}_0)$ is not equal to

$x(t, 0)$ in spite of the fact that \mathbf{u}_0 differs from 0 only on a ν -null set. Therefore $x(t, \mathbf{u})$ is not well defined on $L_\infty(\nu)$, and ν is not an underlying measure. This completes the proof.

The following result is actually the main step in the proof of Theorem B. It is isolated here for future reference.

LEMMA 3.5. *If Σ is countably generated then $|M|(E) = \int_{t_0}^{t_1} |\mu_t|(E) dt$.*

Proof. The inequality $|M|(E) \leq \bar{\mu}(E)$ holds without the countability assumption, and is proved in Lemma 3.2. Let Σ_1 be a denumerable subfamily of Σ which generates Σ . Without loss of generality, Σ_1 is an algebra. The variation of any measure with bounded variation, say ν , is then obtained as follows. The number $|\nu|(E)$ is the supremum of all $\sum |\nu(E_i)|$, where $E = \bigcup E_i$ is a disjoint union and $E_i = F_i \cap E$ with $F_i \in \Sigma_1$. This follows for instance from Halmos [1950, p. 168]. We denote the partition E_1, \dots, E_k by \mathcal{E} and the value $\sum |\nu(E_i)|$ by $\nu(\mathcal{E})$. Since Σ_1 is denumerable there is a sequence \mathcal{E}_i of partitions such that $|\nu|(E) = \lim \nu(\mathcal{E}_i)$ no matter what ν is. In particular $|\mu_t|(E)$ is the pointwise limit of $\mu_t(\mathcal{E}_i)$ as $i \rightarrow \infty$. (This, incidentally, proves the claim in Remark 2.2.) For a fixed $\varepsilon > 0$ we can fix i_0 such that

$$\int_{t_0}^{t_1} |\mu_t|(E) dt \leq \int_{t_0}^{t_1} \mu_t(\mathcal{E}_i) dt + \varepsilon \quad \text{for any } i \geq i_0.$$

The right-hand side of the inequality is equal to $M(\mathcal{E}_i) + \varepsilon$, and as $i \rightarrow \infty$ it tends to $|M|(E) + \varepsilon$. Since ε was arbitrary, it follows that

$$\int_{t_0}^{t_1} |\mu_t|(E) dt \leq |M|(E),$$

and this completes the proof.

Proof of Theorem B. The inequality $|M|(E) \leq \bar{\mu}(E)$ (which implies the absolute continuity of $|M|$ with respect to $\bar{\mu}$) was established in Lemma 3.2, and this together with Theorem A implies that $\bar{\mu}$ is indeed an underlying measure. If Σ is countably generated, Lemma 3.5 states that $|M| = \bar{\mu}$, hence $\bar{\mu}$ is a minimal underlying measure.

Example 3.6. We wish to show that if Σ is not countably generated, then $\bar{\mu}$ might not be a minimal underlying measure. Let S be the interval $[-1, 1]$ but Σ be the σ -algebra of the sets which are either countable or the complement of a countable set. Let $[t_0, t_1] = [0, 1]$ and let the system (*) be $\dot{x} = u(t) - u(t-1)$. An admissible control must be constant, except for a denumerable number of points. In particular $M(\mathbf{u}) \equiv 0$. However $|\mu_t|(E)$ is the cardinality of $E \cap \{t, t-1\}$. Therefore $\bar{\mu}$ is exactly the Lebesgue measure. (If we consider the system $\dot{x} = \alpha(t)(u(t) - u(t-1))$ with $\alpha(t)$ not being Lebesgue measurable we get an example where $t \rightarrow \mu_t(E)$ is measurable for every $E \in \Sigma$, in fact it is the zero function, but where $t \rightarrow |\mu_t|(E)$ is not measurable. Such an example was promised in Remark 2.2.)

The σ -field in Example 3.6 is not as artificial as it might seem. The measurability with respect to the σ -field structure Σ represents, in general, information available when planning. If the structure is that of the σ -field in Example 3.6 then the planner is restricted to choosing constant controls with perturbations only at a countable number of points.

4. Integrable controls. The admissible controls were defined in § 2 as bounded measurable mappings $\mathbf{u}: S \rightarrow R^m$. A larger class, say mappings which are integrable with respect to a natural measure, could not be used since there was no such natural measure and since $\int_S u(s) d\mu_t(s)$ ought to be well defined for the continuum μ_t of measures. The construction of the underlying measure in the preceding section allows

us to extend the family of admissible controls. We summarize the information in the following theorem. The set-function M and its variation $|M|$ were defined and discussed in § 3.

THEOREM C. *If $\mathbf{u}: S \rightarrow R^m$ is measurable and $|M|$ -integrable then the expression $\int_S u(s) d\mu_t(s)$ defines an integrable R^n -valued mapping on $[t_0, t_1]$. The solution $x(t, \mathbf{u})$ of (*) is then defined, is affine in \mathbf{u} and for a fixed t , is continuous in \mathbf{u} with respect to the weak topology of $L_1(|M|)$. On subsets of L_1 which are sequentially compact in the weak topology, the continuity of $x(t, \mathbf{u})$ is uniform in t (i.e., there is continuity into the space $C([t_0, t_1], R^n)$).*

Proof. As is borne out by Lemma 3.2 and the discussion following it, the integral $M(\mathbf{u})$ is well defined for $|M|$ -integrable functions \mathbf{u} . Lemma 3.3, with the same proof, is also valid for $|M|$ -integrable controls. This verifies the first part of the statement. To check that $x(t, \mathbf{u})$ is an affine mapping in \mathbf{u} is trivial and the continuity for a fixed t follows from Lemma 3.4, since M is continuous from $L_1(S, |M|, R^m)$ into $L_1([t_0, t_1], R^n)$ with their weak topologies (e.g., Hille and Phillips [1957, Thm. 2.11.11]). By the continuity, the image by M of a weakly sequentially compact subset of controls is a weakly sequentially compact subset in $L_1([t_0, t_1], R^n)$. Therefore it is uniformly integrable (see e.g., Dunford and Schwartz [1958, p. 292]). It is easy to see then that the solutions $x(t, \mathbf{u})$, for \mathbf{u} in a weakly compact set, are equicontinuous. The uniform continuity in t follows now from the continuity in each t separately. This completes the proof.

It should be noted that the conclusion of Theorem C fails if $|M|$ is replaced there by an arbitrary underlying measure. Example 5.1 below will demonstrate this.

COROLLARY 4.1. *Let ν be an underlying measure for (*) such that the Radon-Nikodým derivative $d|M|/d\nu$ is bounded (in particular if $\nu = \bar{\mu}$). Then (*) is well defined for any ν -integrable control \mathbf{u} , and $x(t, \mathbf{u})$ is continuous when \mathbf{u} is endowed with the weak topology of $L_1(\nu)$.*

Proof. The Radon-Nikodým derivative is well defined since $|M|$ is absolutely continuous with respect to any underlying measure (Theorem A). If $d|M|/d\nu$ is bounded, then $L_1(\nu)$ is contained in $L_1(|M|)$ and the weak topology of $L_1(\nu)$ is finer than that of $L_1(|M|)$. The conclusion follows then, from Theorem C. The boundedness of $d|M|/d\bar{\mu}$ (in fact $d|M|/d\bar{\mu} \leq 1$) follows from Lemma 3.2.

It is worthwhile to note that the boundedness of $d|M|/d\nu$ is also necessary for the inclusion $L_1(\nu) \subset L_1(|M|)$. We leave out the details.

We finally note that integrability of \mathbf{u} with respect to $L_1(|M|)$ is not necessary for (*) to be well defined. Consider the extreme case where $m = 2$ and $n = 1$. Then μ_t is a 2-vector, say (μ_t^1, μ_t^2) , and we suppose that $\mu_t^1 = 0$. The first coordinate of \mathbf{u} would not make any difference in the evaluation of $\int_S u(s) d\mu_t(s)$.

5. Some examples and comments. We collect, not necessarily in a logical order, some examples demonstrating various aspects of the underlying measures, and some facts which ought to help in detecting the underlying measures and their properties. We employ, in some cases without references, the notation and results of previous sections.

Example 5.1. Consider the system

$$(5.1) \quad \dot{x} = A(t)x + B(t)u,$$

with (S, Σ) being the interval $[t_0, t_1]$ with Lebesgue measure λ . The minimal underlying measure $|M| = \bar{\mu}$ is clearly $\bar{\mu}(E) = \int_E |B(t)| dt$. The Lebesgue measure λ is an underlying measure, but unless $B(t) \neq 0$ for almost every t the measure λ is not minimal. A

$\bar{\mu}$ -integrable control \mathbf{u} is admissible in the sense of Theorem C. This amounts to $|B(t)|u(t)$ being λ -integrable. In general, a λ -integrable control might not yield a well defined system.

Example 5.2. Consider the system of type (1.2) given by

$$(5.2) \quad \dot{x} = A(t)x + B(t)u(t - h(t))$$

with $h(t) = t - a$. The minimal underlying measure $\bar{\mu}$ is concentrated at one point, namely a . The value $\bar{\mu}\{a\}$ is $\int_{t_0}^{t_1} |B(t)| dt$.

Example 5.3. Consider the real system of type (1.2)

$$(5.3) \quad \dot{x} = A(t)x + u(g(t))$$

defined on $[0, 1]$, with strictly increasing $g(t)$, and mapping a set of full measure in $[t_0, t_1]$ into a set of Lebesgue measure zero. (For instance, let $g(t)$ be the minimal element in $\{\theta: \psi(\theta) = t\}$, where ψ is the Cantor function. See Halmos [1950, p. 83].) (The function g can be made continuous.) The controls \mathbf{u} are defined on $[0, 1]$ with the Borel structure, but the minimal underlying measure $\bar{\mu}$ is atomless and concentrated on a set of Lebesgue measure zero, i.e., singular with respect to the Lebesgue measure. (For g defined above, $\bar{\mu}$ is generated by the Cantor function, i.e., $\bar{\mu}([\theta_1, \theta_2]) = \psi(\theta_2) - \psi(\theta_1)$.)

In Theorem A we showed that $|M|$ is absolutely continuous with respect to any underlying measure. The following is a converse. (As noted in § 4, the analogous statement does not hold if integrable controls are sought.)

PROPOSITION 5.4. *If ν is a real nonnegative measure and if $|M|$ is absolutely continuous with respect to ν , then ν is an underlying measure for (*).*

Proof. The absolute continuity implies that the expression $\int_S u(s) d\mu_t(s)$ defines a function in $L_1[t_0, t_1]$ for every control in $L_\infty(\nu)$. The continuity with respect to the weak* topology of $L_\infty(\nu)$ follows from the observation that convergence in the weak* topology of $L_\infty(\nu)$ implies convergence in the weak* topology of $L_\infty(|M|)$. (To see this, notice that the net \mathbf{u}_i converges in weak*- $L_\infty(|M|)$ exactly when $\int_S \varphi(s) u_i(s) d|M|$ converges for every $\varphi \in L_1(|M|, \mathcal{R}^m)$. This indeed holds if \mathbf{u}_i converges in weak*- $L_\infty(\nu)$, since the integral can be written as $\int_S \varphi(s) u_i(s) f(s) d\nu$, with $f(s) = d|M|/d\nu$ being the Radon-Nikodym derivative, and, by definition, $\varphi(s)f(s)$ is in $L_1(\nu, \mathcal{R}^m)$.)

On certain occasions the inhomogeneous term in (*) is given as a sum of several terms. For these circumstances the following additivity results might be useful.

PROPOSITION 5.5. *Suppose that S is a disjoint union $S = \bigcup_{i=1}^\infty S_i$. Then the restriction of an underlying measure of (*) to S_i is an underlying measure for the system*

$$(5.4) \quad \dot{x} = A(t)x + \int_{S_i} u(s) d\mu_t(s).$$

In particular, the restriction of $|M|$ to S_i is the minimal underlying measure of (5.4) as constructed in Theorem A. Conversely, if ν_i is an underlying measure for (5.4) and if ν , defined by $\nu(E) = \sum \nu_i(E \cap S_i)$, is a finite measure, then ν is an underlying measure for ().*

Proof. The statement concerning M follows from the construction of M . The other statements follow from Proposition 5.4.

PROPOSITION 5.6. *Suppose (*) has the form*

$$(5.5) \quad \dot{x} = A(t)x + \sum_{i=1}^\infty \int_S u(s) d\mu_t^i(s).$$

Let ν_i for $i = 1, 2, \dots$, be underlying measures for the systems generated, respectively, by $\int_S u(s) d\mu_i^i(s)$. If $\sum_{i=1}^{\infty} \nu_i$ is a finite measure, it is an underlying measure for (5.5). It might not be minimal even if each ν_i is minimal.

Proof. The measure M associated with (5.5) is the sum of the measures M_i associated with the system generated by $\int_S u(s) d\mu_i^i(s)$. Therefore $|M| \leq \sum_{i=1}^{\infty} |M_i|$. The result now follows from Proposition 5.4. An example for the last claim is provided by $\mu_i^1 = -\mu_i^2$.

The following two results relate properties of μ_i with properties of the underlying measures.

PROPOSITION 5.7. *If each μ_i is absolutely continuous with respect to a given real nonnegative measure ν , then ν is an underlying measure for (*).*

Proof. The proof is obvious.

Simple examples, e.g., Example 5.1, show that the extent to which the previous proposition can be applied is quite limited. It applies however to some examples. Chyung and Lee [1970] consider a delay system whose control part is essentially

$$(5.6) \quad \int_{t-\tau}^t B(t, s)u(s) ds + \sum_{i=0}^l B_i(t)u(t-h_i),$$

with h_i constants. Propositions 5.6 and 5.7 imply that the Lebesgue measure is an underlying measure for it. (Notice that for the left side term, μ_t is essentially the restriction of the Lebesgue measure to $[t-\tau, t]$.)

The support of the minimal underlying measure is certainly of interest; outside this support the values of the controls do not matter. (Recall that the support of ν is defined when S is topological with the Borel σ -field, and is the smallest closed set D such that $|\nu|$ vanishes on $S \setminus D$.) We denote the support of ν by $\text{supp } \nu$. If E is a set such that $|\mu_i|(E) = 0$ except for a set of t of Lebesgue measure zero, then $\bar{\mu}(E) = 0$. Hence, by Lemma 3.2, $|M|(E) = 0$ and if E is open, then $E \subset S \setminus \text{supp } |M|$. Therefore $\text{supp } |M|$ is contained in

$$(5.7) \quad C = \bigcap_N \text{closure} \bigcup_{t \in T \setminus N} \text{supp } |\mu_t|,$$

with $T = [t_0, t_1]$ and where N ranges over all subsets of T of Lebesgue measure zero. Example 3.5 shows that $\text{supp } |M|$ might be strictly contained in C .

PROPOSITION 5.8. *$C = \text{supp } \bar{\mu}$. If Σ is countably generated, say S is a separable metric space, then $\text{supp } |M| = C$.*

Proof. We shall show that $C = \text{supp } \bar{\mu}$. When the Borel σ -field Σ is countably generated it follows from Lemma 3.5 that $C = \text{supp } |M|$. The inclusion $\text{supp } \bar{\mu} \subset C$ was explained earlier. If $E \subset S \setminus \text{supp } \bar{\mu}$ is open, then clearly $|\mu_t|(E) = 0$ for almost every $t \in [t_0, t_1]$. This follows from the definition of $\bar{\mu}$. The construction of C then implies that $E \subset S \setminus C$. Hence $C \subset \text{supp } \bar{\mu}$. This completes the proof.

We turn now to the example promised in the introduction. It is a combination of a problem of growth and a problem of location, both treated extensively in the economic literature. For other models utilizing the measure theoretic aspects of location, see Faden [1977].

Example 5.9. A farmer owns a piece of land S ($\subset \mathbb{R}^2$). At the beginning of the season he decides what types of crops he wants, say out of m possibilities, and where to plant them. This decision cannot be reversed during the season. Let us denote the farmer's decision by \mathbf{u} ; here $\mathbf{u} = u(s) : S \rightarrow \mathbb{R}^m$, where $u_j(s)$ represents the density of the j th type of crop.

The revenues arrive during the season. In many practical situations the revenue obtained from one part of the field S does not depend on what happens in the other parts. This amounts to the mathematical assumption that revenue is an additive function of subsets of S . A less reasonable assumption is that earnings are linear in the decision \mathbf{u} . We shall dispense with this assumption shortly, but if both additivity and linearity are adopted, we can conclude that the rate at which income (or expenditure) arrives is

$$\int_S u(s) d\mu_t(s),$$

where μ_t is the appropriate vector of measures, which clearly depend on the time t . Suppose that $c(t)$ is the amount of cash our farmer has at time t . Then

$$(5.8) \quad \dot{c}(t) = rc(t) + \int_S u(s) d\mu_t(s),$$

with r being the interest rate, which might vary with time. We see that (5.8) is of the form (*). If the control decision \mathbf{u} is subject to some constraints we encounter an optimization problem, say how to maximize $c(t)$ at the end of the season.

A reasonable assumption concerning the system (5.6) is that the minimal underlying measure $|M|$ would be absolutely continuous with respect to the Lebesgue measure on S . Such an assumption expresses the fact that one cannot earn money from a piece of land of measure zero.

A possible constraint on the decision \mathbf{u} is that at every point s all but one coordinate of $u(s)$ are zero (i.e., two sorts cannot be planted at one point). The bang-bang principle § 7 (see Theorem D) addresses itself to such situations, and implies that whatever can be achieved with "mixed" plants can be achieved with pure plants as well.

A nonlinear dependence of revenue on the planted quantity $u(s)$ is more reasonable. To this end we may introduce a function $h(s, u(s))$ which indicates that the output which grows is not a linear function of the input $u(s)$, and might also depend on the point s in the field. (In our agricultural example $h(s, u)$ probably has a maximum at a finite u .) With this modification the model becomes

$$(5.9) \quad \dot{c}(t) = rc(t) + \int_S h(s, u(s)) d\mu_t(s).$$

We show in § 9 how such nonlinearities can be incorporated into the analysis.

6. Atomless and absolutely continuous underlying measures. For several reasons it is desirable to know whether the minimal underlying measure $|M|$ has atoms, or whether it is absolutely continuous with respect to a given measure. For instance, the bang-bang principle, which is given in the next section, is sharper if no atoms are present. In this section we provide some criteria for checking the two properties.

Recall that an atom of a measure ν on a space (S, Σ) is a set $E \in \Sigma$ such that $\nu(E) \neq 0$ and if $F \subset E$, $F \in \Sigma$ then either $\nu(E \setminus F) = 0$ or $\nu(F) = 0$. An atom E might or might not contain a singleton $\{e\}$ which is itself an atom. In some spaces, e.g., Euclidean spaces, such a singleton always exists. We shall formulate the condition for an atom which is a singleton, and then comment on the general case.

PROPOSITION 6.1. *A singleton $\{s\}$ is an atom of the minimal underlying measure $|M|$ if and only if the set $\{t: \mu_t\{s\} \neq 0\}$ has positive Lebesgue measure. In particular, if no μ_t has an atomic singleton then $|M|$ has no atomic singleton.*

Proof. $M\{s\}$ is not zero if and only if the set $\{t: \mu_t\{s\} \neq 0\}$ has positive Lebesgue measure. See the definition of M .

Remark 6.2. For an arbitrary atom the criteria is not as neat as for a singleton. In the system (1.1) the interval $[t_0, t_1]$ is an atom for every μ_t , yet it is not an atom for $|M|$, which might be the Lebesgue measure, see Example 5.1. If (S, Σ) is the interval $[-1, 1]$ with the σ -field generated by the denumerable sets (compare with Example 3.5) and if μ_t is concentrated on $-t$ and t with equal values, say 1, then S is an atom for $|M|$ but it is not an atom for either of the μ_t . A positive result is: Let $T \subset [t_0, t_1]$ be a set of positive Lebesgue measure and let $E \in \Sigma$ be such that $M(E) \neq 0$. If $|\mu_\tau|(E) = 0$ for $\tau \notin T$, and if whenever $F \subset E$ then either $|\mu_t|(F) = 0$ for all $t \in T$ or $|\mu_t|(E \setminus F) = 0$ for all $t \in T$ then E is an atom of $|M|$. This is easily verified. On the other hand it might happen that each μ_t is atomless but that $|M|$ has an atom. (The latter cannot be a singleton, in view of Proposition 6.1.) To construct such an example we modify slightly Example 3.5. Let S be the unit square $[0, 1] \times [0, 1]$ with the σ -algebra generated by the Borel sets which contain all but a denumerable family of intervals $\{(\xi, \eta): 0 \leq \eta \leq 1\}$. Let μ_t be the Lebesgue measure on $\{(t, \eta): 0 \leq \eta \leq 1\}$. Then each μ_t is atomless but S is an atom for $|M|$.

The following is an application.

COROLLARY 6.3. *Suppose that the structure of S is such that an atom always contains an atomic singleton. Suppose that $(*)$ has the form*

$$(6.1) \quad \dot{x} = A(t) + \sum_{i=1}^{\infty} B_i(t)u(\theta_i(t)) + \int_S u(s) d\nu_t(s),$$

with $\theta_i: [t_0, t_1] \rightarrow S$ measurable. Suppose that ν_t is atomless and that for every i the interval $[t_0, t_1]$ is a union of a countable number of sets, on each of them θ_i is one-to-one; then $|M|$ is atomless.

Proof. By the last statement of Proposition 6.1 the atomless part of (6.1) does not contribute (see Proposition 5.6) atoms for $|M|$, and according to the first statement of Proposition 6.1 the atomic parts of (6.1) do not generate atoms either. This completes the proof.

Notice that since any measure μ_t has at most a countable number of atoms, the system $(*)$ can always be written in the form (6.1). Without the extra assumption on S the nonatomicity does not follow. See the example in Remark 6.2.

A particular case of the previous condition is assumption Δ_2 in Banks and Jacobs [1970, p. 469]. The system there is a delay system, hence in (6.3) $\theta_i(t): [t_0, t_1] \rightarrow R$. The assumption of Banks and Jacobs is that each θ_i is strictly increasing.

A result concerning the absolute continuity of $|M|$ with respect to a given measure is presented in Proposition 5.7. The following is another result which copes also with the atomic part of the measures.

PROPOSITION 6.4. *Let ν be a given nonnegative measure on (S, Σ) . If for every null set E of ν the Lebesgue measure of $\{t: \mu_t(E) \neq 0\}$ is zero, then $|M|$ is absolutely continuous with respect to ν .*

Proof. The proof is obvious.

COROLLARY 6.5. *Suppose that (S, Σ) is the real line R with the Borel structure. Consider the system*

$$(6.2) \quad \dot{x} = A(t)x + \sum_{i=1}^{\infty} B_i(t)u(\theta_i(t)) + \int_S u(s) d\nu_t(s),$$

with $\theta_i: [t_0, t_1] \rightarrow R$ measurable. Suppose that each ν_t is absolutely continuous with respect

to the Lebesgue measure λ , and that each θ_i is continuously differentiable, with $\dot{\theta}_i(t) \neq 0$, except for a set of measure zero. Then $|M|$ is absolutely continuous with respect to the Lebesgue measure.

Proof. Let $E \subset R$ be of Lebesgue measure zero. We shall see that the measure of $\{t: \mu_t(E) \neq 0\}$ is zero, where $\mu_t = \nu_t + \sum_{i=1}^{\infty} B_i(t) \delta_{\theta_i(t)}$. (Here δ is the Dirac measure.) By the assumption, each $\nu_t(E) = 0$, hence it is enough to show for each index i that the Lebesgue measure of $\{t: \theta_i(t) \in E\}$ is zero. There is a closed set of measure zero on which $\dot{\theta}(t) = 0$ and the rest is a union of a countable number of open intervals on which $\dot{\theta}_i(t) \neq 0$. Since θ is a diffeomorphism on each closed subinterval of these open intervals the result follows.

A particular case of the previous example is studied by Chyung and Lee [1970, p. 400], where $\theta(t)$ is assumed to be differentiable with $\dot{\theta}(t) \geq \varepsilon$ for a certain $\varepsilon > 0$. The previous result fails if (S, Σ) is higher dimensional, as follows.

PROPOSITION 6.6. *Suppose that (S, Σ) is a Euclidean space R^k with $k \geq 2$. Consider the system*

$$(6.3) \quad \dot{x} = A(t)x + \sum_{i=1}^{\infty} B_i(t)u_i(\theta_i(t)),$$

with $\theta_i(t): [t_0, t_1] \rightarrow S$ continuously differentiable. Then the minimal underlying measure is singular with respect to the Lebesgue measure.

Proof. The image of $\theta_i(t)$ for $t \in [t_0, t_1]$ and $i = 1, 2, \dots$ is one-dimensional, hence has Lebesgue measure zero. The minimal underlying measure clearly vanishes outside this image.

7. Bang-bang for convex constraints. In this section we generalize the bang-bang principle (see, e.g., Hermes and LaSalle [1969]) to systems with indirect controls. To this end we restrict the admissible controls $\mathbf{u} = u(s)$ to those satisfying $u(s) \in \Omega(s)$ for a prescribed Ω . The following is a standing hypothesis for this section.

Assumption 7.1. For each s the set $\Omega(s) \subset R^m$ is closed and convex. The set-valued function $s \rightarrow \Omega(s)$ is measurable in the sense that $\Omega^-(B) = \{s: \Omega(s) \cap B \neq \emptyset\}$ is in Σ for every closed $B \subset R^m$. There exists an $|M|$ -integrable real function $h(s)$ such that $z \in \Omega(s)$ implies $|z| \leq h(s)$.

For any measurable $\mathbf{u} = u(s)$ satisfying $u(s) \in \Omega(s)$ the equation (*) is well defined (see § 4). If Ω is bounded (i.e., h in Assumption 7.1 is bounded) then any such measurable selection is admissible in the sense of § 2. In either case, we call such a control $\mathbf{u} = u(s)$ an Ω -admissible control.

The set of attainability $\mathcal{A}(t_1)$ is defined to consist of all the vectors $x(t_1, \mathbf{u})$ in R^n for Ω -admissible controls \mathbf{u} .

LEMMA 7.2. *The set $\mathcal{A}(t_1)$ is convex and compact.*

Proof. The family of Ω -admissible controls is convex and compact in the weak topology of $L_1(|M|)$. The set $\mathcal{A}(t_1)$ is a continuous image of this set under the affine mapping $x(t_1, \mathbf{u})$ (see Theorem C). This completes the proof.

THEOREM D. *For every z in $\mathcal{A}(t_1)$ there is an Ω -admissible control $\mathbf{v} = v(s)$ such that $x(t_1, \mathbf{v}) = z$ and $v(s)$ is an extreme point of $\Omega(s)$ for all points s , with the exception of at most n atoms of $|M|$. In particular, if $|M|$ is atomless $v(s)$ is an extreme point of $\Omega(s)$ for every s .*

Proof. We know from Theorem C that $x(t_1, \mathbf{u})$ is well defined for $\mathbf{u} \in L_1(|M|)$ and is affine and continuous with respect to the weak topology. Using the technique in Artstein [1980, Thms. 3.2 and 5.1] the result follows. For completeness we state here the main steps of the argument. Consider the set $L = \{\mathbf{u}: \mathbf{u} \text{ is } \Omega\text{-admissible and}$

$x(t_1, \mathbf{u}) = z\}$. Since $x(t_1, \mathbf{u})$ is a continuous affine mapping, the set L is convex and closed, hence compact. Let \mathbf{v} be an extreme point of L , (guaranteed by the Krein–Milman theorem). We claim that \mathbf{v} is the desired control. To see this, notice that the facial dimension of \mathbf{v} (namely the largest number of linearly independent controls $\mathbf{u}_1, \dots, \mathbf{u}_k$ such that $\mathbf{v} \pm \mathbf{u}_i$ is Ω -admissible) is at most n . Otherwise $x(t_1, \cdot)$ on their span is not one-to-one, which violates the extremality of \mathbf{v} . Finite facial dimensionality implies that $v(s)$ is an extreme point of $|M|$ for almost every s in the atomless part of $|M|$. Facial dimensionality n implies that $v(s)$ is not an extreme point of $\Omega(s)$ for at most n atoms.

8. Nonconvex constraints. In this section we examine the properties of the set of attainability for nonconvex constraints, and derive a simple existence theorem for optimization. The following will be our hypothesis.

ASSUMPTION 8.1. *For each s the set $\Omega(s) \subset R^m$ is closed. The set-valued mapping $s \rightarrow \Omega(s)$ is measurable and there exists an $|M|$ -integrable function $h(s)$ such that $z \in \Omega(s)$ implies $|z| \leq h(s)$.*

Notice that this assumption modifies that of § 7 only by dropping the convexity. The measurable functions $u : s \rightarrow \Omega(s)$ will still be called Ω -admissible controls and the set of attainability $\mathcal{A}(t_1)$ is as in the previous section.

THEOREM E. *The set $\mathcal{A}(t_1)$ is compact. If $|M|$ is atomless, then $\mathcal{A}(t_1)$ is also convex.*

The mapping $x(t_1, \mathbf{u})$ is affine and continuous in \mathbf{u} and therefore can be written as $x(t_1, \mathbf{u}) = z_0 + l(\mathbf{u})$, with $l(\mathbf{u})$ linear and continuous. If S_0, S_1, \dots is a partition of the measure space, and \mathbf{u}_i denotes the restriction of \mathbf{u} to S_i (i.e., $u_i(s) = u(s)$ for $s \in S_i$ and $u_i(s) = 0$ otherwise) then $l(\mathbf{u}) = \sum l(\mathbf{u}_i)$. Therefore, if $\mathcal{A}_i = \{l(\mathbf{u}) : u(s) = 0 \text{ if } s \notin S_i\}$ then we get that

$$\mathcal{A}(t_1) = z_0 + \sum \mathcal{A}_i = z_0 + \{\sum a_i : a_i \in \mathcal{A}_i\}.$$

The boundedness of $\mathcal{A}(t_1)$ implies that for any choice $a_i \in \mathcal{A}_i$ the summation converges. Moreover, if we succeed in showing that each of the \mathcal{A}_i is compact, then the compactness of $\mathcal{A}(t_1)$ follows. We choose now S_1, S_2, \dots to be the atoms of $|M|$, and S_0 to be the part of S on which $|M|$ is nonatomic. Now the compactness of \mathcal{A}_i for $i \geq 1$ follows, since $\Omega(s)$ is constant almost everywhere on atoms and closed; and the compactness of \mathcal{A}_0 follows from Lemma 7.2 and Theorem D. This completes the proof.

For delayed systems the theorem was proved by Banks and Jacobs [1970, Thms. 3.1, 3.2], where the convexity is derived from their assumption Δ_2 . As we saw in Corollary 6.4, this assumption indeed implies that the underlying measure $|M|$ is atomless. An even stronger condition implies the convexity of the set of attainability in Lee [1968].

It is worthwhile to note that the estimate n of the nonextreme values of \mathbf{u} in Theorem D, is false in the nonconvex case. It is easy to construct an example where, in order to reach a given $z \in \mathcal{A}(t_1)$, the values $u(s)$ are not extreme on any of the atoms of $|M|$.

The compactness of $\mathcal{A}(t_1)$ yields existence of optimal controls in some situations. For instance, if a continuous cost functional $c(z)$ ought to be minimized on $z \in \mathcal{A}(t_1)$ then there is always an optimal control.

9. Extensions

9.1. Delays in the states. The analysis in this paper employed very little of the specific structure of the ordinary differential equation (ODE) $\dot{x} = A(t)x$ in (*). In fact, the ODE appears in only two places. The first is in Lemma 3.4, where the property which is proved is: $\dot{x} = A(t)x + h(t)$, $x(t_0) = x_0$ has a unique solution which

is affine in \mathbf{h} and continuous with respect to the weak topology of L_1 . The proof of Lemma 3.4 applies to other systems as well, and at any rate whenever this property holds, the rest of our analysis applies. (Proposition 2.1 is the second place where the structure of the ODE is used. The conclusion there holds with greater generality, e.g., for the delay systems, and was not used throughout.)

We present here another example.

Example 9.1. Consider the delay equation

$$(9.1) \quad \dot{x}(t) = \int_{t-a}^t x(\tau) d\tau F(t, \tau) + \int_S u(s) d\mu_t(s).$$

An initial condition is a bounded measurable function $\varphi(\tau)$ on $[t_0 - a, t_0]$ and an initial vector $x(t_0) = x_0$.

Banks and Jacobs [1970], Banks, Jacobs and Latina [1971] and Chyung and Lee [1970] analyze a similar system with delayed controls, i.e., S is a real interval. A situation explored in these papers is that the values of $\varphi(\tau)$ can also be controlled.

Suppose that $F(t, \tau)$ is measurable in the two variables, with bounded variation in τ for a fixed t and that the variation is integrable in t . (Compare with Banks and Jacobs [1970, p. 463].) Then for a fixed φ and a given \mathbf{u} there is a unique solution to (9.1) and it moreover depends continuously on $h(t) = \int_S u(s) d\mu_t(s)$. Our methods are therefore applicable and all the results for the underlying measures and the consequences hold. When the initial function φ is also controlled then we simply regard $[t_0 - a, t_0]$ as part of the control space S . In particular, an underlying measure is then defined on $[t_0 - a, t_0]$, and all the consequences hold.

9.2. Discrete systems. The techniques and results of this paper can be applied also to discrete time systems of the form

$$(9.2) \quad x(j+1) = A(j)x(j) + \int_S u(s) d\mu_j(s).$$

Here $j = 0, 1, \dots, T$. The vectors $x(i)$ are in R^n , the control is a mapping $u(s): S \rightarrow R^m$ and the μ_j are $n \times m$ matrix-valued measures. We assume that a fixed initial condition $x(0) = x_0$ is satisfied.

An admissible control is a bounded and measurable function. It is clear that if \mathbf{u} is an admissible control then (9.2) has a unique solution $x(j, \mathbf{u})$. We use Definition A as stated to define an underlying measure for (9.2).

THEOREM F. *The measure $\bar{\mu} = \sum_{j=0}^T |\mu_j|$ is a minimal underlying measure for (9.2).*

Proof. It is clear that the vector $h_j = \int_S u(s) d\mu_j(s)$ for $j = 0, \dots, T$, is affine and depends continuously on \mathbf{u} when the latter is endowed with the weak* topology of $L_\infty(\bar{\mu})$. The solution $x(j, h)$ of $x(j+1) = A(j)x(j) + h_j$ depends continuously on and is affine in h . This implies that $\bar{\mu}$ is an underlying measure. If $\bar{\mu}$ is not absolutely continuous with respect to a certain measure ν then there is a ν -null set E such that $\mu_j(E) \neq 0$ for a certain $0 \leq j \leq T$. In particular for a certain \mathbf{u} which is zero outside E , $\int_S u(s) d\nu = 0$, while $\int_S u(s) d\mu_j \neq 0$. Then $x(j, 0) \neq x(j, \mathbf{u})$ although \mathbf{u} is equivalent to the zero function in $L_\infty(\nu)$. This prevents ν from being an underlying measure.

Once an underlying measure is identified for (9.2) the rest of the analysis of this paper follows almost automatically. We leave out the details. Notice however that the bang-bang principle (Theorem D) holds, and if $\sum_{j=0}^T |\mu_j|$ is atomless then a bang-bang control \mathbf{u} can be used—this in spite of the discreteness of the time variable.

9.3. Nonlinearities. A particular form of nonlinearity can be easily incorporated into our theory; namely if the control part in (*) (or in (9.1), (9.2)) has the form

$$(9.3) \quad \int_S h(s, u(s)) d\mu_t(s),$$

with h measurable. Such nonlinearities are desired in applications (see Example 5.9). The analysis in Banks and Jacobs [1970] with regard to the delay version of (9.1) incorporates such nonlinearities.

The method for coping with such nonlinearity goes back to Filippov. We shall sketch the idea. Although $\mathbf{u} = u(s)$ is the control variable, we pretend that we choose the value $y(s) = h(s, u(s))$. The term (9.3) is then linear in $\mathbf{y} = y(s)$ and the entire theory of underlying measures can be applied when the control part is $\int_S y(s) d\mu_t(s)$. Then the interpretation of the results in terms of \mathbf{u} can be done. The following is a useful conceptual observation.

PROPOSITION 9.1. *If ν is an underlying measure with respect to $y(s) = h(s, u(s))$ then $x(t, \mathbf{u})$ is well defined for $L_\infty(\nu)$.*

Proof. A change in \mathbf{u} on a ν -null set causes a change of $y(s) = h(s, u(s))$ on a ν -null set.

The solution $x(t, \mathbf{u})$ might not be continuous when \mathbf{u} is endowed with the weak* topology. This is due to the nonlinearity of h .

LEMMA 9.2. *Suppose that $h(s, u)$ is continuous in u . Let Ω be a constraint set satisfying Assumption 8.1. Let $y(s)$ be measurable and such that for every s the value $y(s)$ is in the range $h(s, \Omega(s))$. Then an Ω -admissible control $\mathbf{u} = u(s)$ exists such that $y(s) = h(s, u(s))$.*

This is the well-known Filippov lemma; see, e.g., Jacobs [1968].

We demonstrate now how to use this technique, and get the analogue of Theorem E for the nonlinear case. Let the system be

$$(9.4) \quad \dot{x} = A(t)x + \int_S h(s, u(s)) d\mu_t(s),$$

with h continuous in u . Let Ω satisfy Assumption 8.1 and suppose that the set-valued function $h(s, \Omega(s))$ is bounded by an $|M|$ -integrable function, where $|M|$ is the underlying measure for $\dot{x} = A(t)x + \int_S y(s) d\mu_t(s)$.

THEOREM G. *The attainable set $\mathcal{A}(t_1)$ of (9.4) is compact. If the underlying measure $|M|$ is atomless then $\mathcal{A}(t_1)$ is also convex.*

Proof. By Theorem E the attainable set related to $\dot{x} = A(t)x + \int_S y(s) d\mu_t(s)$ with the constraint $\Omega_1(s) = h(s, \Omega(s))$ is compact, and if $|M|$ is atomless then this set is also convex. This attainable set certainly contains $\mathcal{A}(t_1)$, but according to the Filippov lemma (Lemma 9.2) it is actually equal to $\mathcal{A}(t_1)$. This completes the proof.

Note added in proof.

We overlooked an important reference. It is the paper *Optimal controls with pseudodelays* by J. Warga, this Journal, 12 (1974), pp. 286–299. The goals, hence the technique and the results in this reference are different from ours; however, Warga sets and analyzes a model for indirect control action which overlaps with the present paper. Our terminology, namely indirect controls, was used earlier in a different technical context, but with the same spirit, for instance in connection with the Lurie problem (see, e.g., S. Lefschetz *Stability of Nonlinear Control Systems*, Academic Press, New York, 1965).

REFERENCES

- Z. ARTSTEIN, [1980], *Discrete and continuous bang-bang and facial dimensions, or: Look for the extreme points*, SIAM Rev., 22, pp. 172–185.
- H. T. BANKS AND M. Q. JACOBS, [1970], *The optimization of trajectories of linear functional differential equations*, this Journal 8, pp. 461–488.
- H. T. BANKS, M. Q. JACOBS AND M. R. LATINA, [1971], *The synthesis of optimal controls for linear time-optimal problems with retarded controls*, J. Optim. Theory Appl., 8, pp. 319–366.
- D. H. CHYUNG AND E. B. LEE, [1970], *Delayed action control problems*, Automatica, 6, pp. 395–400.
- N. DUNFORD AND J. T. SCHWARTZ, [1958], *Linear Operators, Part I*, Interscience, New York.
- A. M. FADEN, [1977], *Economics of Space and Time*, Iowa State University Press, Ames, Iowa.
- P. R. HALMOS, [1950], *Measure Theory*, Van Nostrand, New York.
- H. HERMES AND J. P. LASALLE, [1969], *Functional Analysis and Time Optimal Control*, Academic Press, New York.
- E. HILLE AND R. S. PHILLIPS, [1957], *Functional Analysis and Semi-Groups*, AMS Colloquium Publications XXXI, American Mathematical Society, Providence, RI.
- M. Q. JACOBS, [1968], *Measurable multivalued mappings and Lusin's theorem*, Trans. Amer. Math. Soc., 134, pp. 471–481.
- E. B. LEE, [1968], *Variational problems for systems having delay in the control action*, IEEE Trans. Automatic Control, AC-13, pp. 697–699.
- E. B. LEE, [1969], *Geometric theory of linear controlled systems*, Mathematical Systems Theory and Economics II, Kuhn and Szegö, ed., Lecture Notes in Operations Research and Economics 12, Springer-Verlag, Berlin, pp. 347–354.
- E. B. LEE AND L. MARKUS, [1967], *Foundations of Optimal Control Theory*, John Wiley, New York.

SYSTEMS OVER A PRINCIPAL IDEAL DOMAIN. A POLYNOMIAL MODEL APPROACH*

G. CONTE[†] AND A. M. PERDON[‡]

Abstract. The polynomial model approach to linear dynamical systems over a field was developed principally by P. A. Fuhrmann, starting in 1976 [J. Franklin Inst., 305 (1976), pp. 521–540]. In this paper an analogous approach is proposed for systems over a principal ideal domain.

When the concept of extended linear i/o map is introduced, fractional representations of transfer function matrices arise naturally in this theoretical framework. A correspondence between fractional representations of the transfer function matrix of a given i/o map and its reachable or observable realizations is established. The McMillan degree of a linear i/o map is proved to be equal to the degree of the determinant of the matrix appearing as “denominator” in a coprime fractional representation of the associated transfer function matrix.

Introduction. The polynomial model approach to linear systems over a field was principally developed by P. A. Fuhrmann in [11] and subsequently used in a series of papers [8], [12], [13], [14], [15], [16] by various authors.

The aim of this paper is to develop a polynomial model approach to linear systems over a principal ideal domain A . Systems over a p.i.d. arise as a natural generalization of systems over a field and their study is of interest in many cases (see [6], [9], [17], [21], [23], [27]).

The fundamental tool of the polynomial model approach is essentially the characterization of the quotient $A[z]$ -modules $\Omega U/M$ (dually, of the $A[z]$ -submodules of ΓY) which are free finitely generated over A . A complete characterization of these modules is given in Proposition 4.1 and Corollary 4.7. It generalizes the results already known when A is a field, but it cannot be obtained with the same techniques used in that case.

Recently a more general result in the same direction has been independently obtained in [22]. In that paper a characterization is given of the quotient $A[z]$ -modules $\Omega U/M$ which are finitely generated and projective over A , for A a commutative ring. This result specializes to ours when A is a p.i.d..

The paper is organized as follows. Section 1 contains notations and some results about the closure of a submodule. In § 2 systems and i/o maps are defined. Section 3 deals with realizations of i/o maps, in particular with free finite dimensional realizations. Some general results contained in [5 Chapt. 16] are recalled.

Reachable free realizations and observable free realizations of an i/o map f are determined respectively by quotient modules of the form $\Omega U/M$ such that $M \subset \text{Ker}(f)$ and by the submodules of ΓY containing $\text{Im}(f)$ which are free finitely generated over A . Such modules are characterized in § 4 and a correspondence between them and the regular matrices of appropriate size is established. This fact, with the techniques of [16] and [4], gives us a correspondence between right fractional representations of the transfer function matrix Z_f of f and reachable free realizations of f , and between left fractional representations and observable free realizations. The correspondence, as in the case where A is a field, is one-one for reachable realizations and right fractional representations. Moreover, in this case canonical realizations correspond to coprime representations. This is not true for observable realizations, more precisely,

* Received by the editors June 9, 1980, and in revised form November 20, 1980.

[†] Istituto di Matematica Applicata, Univ. di Padova, Via Belzoni 7, 35100 Padova, Italy.

[‡] Istituto di Matematica, Univ. di Genova, Via L. B. Alberti 4, 16132 Genova, Italy.

the correspondence is not one-one and coprimeness corresponds only to minimality. This led us to study i/o maps whose minimal realizations are canonical. We obtain that when $\text{Im}(f)$ is an A -closed submodule of ΓY , then any minimal realization of f is canonical (the converse also is true). In such a case observable free realizations are in one-one correspondence with left fractional representations.

Finally we prove that the McMillan degree of an i/o map f whose transfer function matrix has the coprime fractional representation $Z_f = ND^{-1}$ or $Z_f = D_1^{-1}N_1$ is equal to the degree of the determinant of D or D_1 . When A is a field this result has been obtained in [18] by making use of algebro-geometric techniques which the polynomial model approach allows us to avoid.

1. Algebraic preliminaries. Throughout this paper A will denote a commutative principal ideal ring (p.i.d.), and U and Y will denote respectively the free finitely generated A -modules A^m and A^p . We will assume that the reader is familiar with the theory of modules over a p.i.d. as can be found in [2].

We shall begin by introducing some notation. We denote by $A[z]$ the ring of polynomials in the indeterminate z with coefficients in A . We remark that $A[z]$ is a p.i.d. if and only if A is a field. We denote by $A[[z^{-1}]]$ the ring of formal power series in z^{-1} . The set $S = \{1, z^{-1}, z^{-2}, \dots\}$ is a multiplicatively closed subset of $A[[z^{-1}]]$. We denote by ΛA the ring of fractions $S^{-1}A[[z^{-1}]]$. It is the minimal ring containing $A[[z^{-1}]]$ in which z^{-1} is a unit. The ring ΛA will be referred to as the ring of generalized Laurent series. Any element $a \in \Lambda A$ is of the form $a = \sum_{t=t_0}^{\infty} a_t z^{-t}$ with $t \in \mathbb{Z}$, $a_t \in A$. $A[z]$ will be identified with the subring of ΛA consisting of all the elements $\sum_{t=t_0}^{\infty} a_t z^{-t}$ such that $a_t = 0$ for $t > 0$. We denote by ΓA the quotient $A[z]$ -module $\Lambda A/A[z]$. Any element $a \in \Gamma A$ is of the form $a = \sum_{t=1}^{\infty} a_t z^{-t}$ with $a_t \in A$. The $A[z]$ -module structure is given by

$$z \cdot \left(\sum_{t=1}^{\infty} a_t z^{-t} \right) = \sum_{t=1}^{\infty} a_{t+1} z^{-t}.$$

PROPOSITION 1.1. *Let $a = \sum_{t=t_0}^{\infty} a_t z^{-t} \in \Lambda A$. Then a is a unit in ΛA if and only if a_{t_0} is a unit in A . In particular, a polynomial is a unit in ΛA if and only if its leading coefficient is a unit in A .*

DEFINITION 1.2. Given the A -module $U = A^m$, we denote by

$$\Omega U \text{ the } A[z]\text{-module } U[z] = A^m[z],$$

$$\Lambda U \text{ the } \Lambda A\text{-module } S^{-1}U[[z^{-1}]] = S^{-1}A^m[[z^{-1}]].$$

We consider ΩU as an $A[z]$ -submodule of ΛU and we denote the inclusion by $j: \Omega U \rightarrow \Lambda U$. We remark that the natural basis e_1, \dots, e_m of U over A is also the natural basis of ΩU over $A[z]$ and of ΛU over ΛA . We denote by

$$\Gamma U \text{ the quotient } A[z]\text{-module } \Lambda U/\Omega U.$$

If we denote by $\pi: \Lambda U \rightarrow \Gamma U$ the canonical projection, then for any $y \in \Gamma U$ there exists a unique element in $\pi^{-1}(y)$ of the form $\sum_{t=1}^{\infty} u_t z^{-t}$. By abuse of notation we will write $y \in \Lambda U$, identifying y with $\sum_{t=1}^{\infty} u_t z^{-t}$. With the same convention, given $x \in \Lambda U$, $\pi(x) \in \Gamma U$ makes sense.

Denoting by $i: U \rightarrow \Omega U$ the inclusion and by $p: \Gamma U \rightarrow U$ the projection given by $p(\sum_{t=1}^{\infty} u_t z^{-t}) = u_1$, we have the following proposition.

PROPOSITION 1.3. *Let X be a $A[z]$ -module. For any A -morphism $g: U \rightarrow X$ (respectively $h: X \rightarrow U$) there exists a unique map $\tilde{g}: \Omega U \rightarrow X$ (respectively $\tilde{h}: X \rightarrow \Gamma U$)*

such that

- (i) $\tilde{g}(\tilde{h})$ is an $A[z]$ -morphism;
- (ii) $g = \tilde{g}i$ ($h = p\tilde{h}$).

Proof. Define $\tilde{g}: \Omega U \rightarrow X$ by $\tilde{g}(\sum_{t=0}^n u_t z^t) = \sum_{t=0}^n z^t \cdot g(u_t)$ (respectively define $\tilde{h}: X \rightarrow \Gamma U$ by $\tilde{h}(x) = \sum_{t=1}^{\infty} h(z^{t-1}x)z^{-t}$). In the above definition(s) the properties (i), (ii) are automatically built in. Uniqueness follows by (i), (ii).

Given the A -module $Y = A^p$, any $A[z]$ -morphism between ΩU and ΩY is given by a $p \times m$ polynomial matrix D . The same matrix defines a ΛA -morphism between ΛU and ΛY and an $A[z]$ -morphism between ΓU and ΓY such that the following row-exact diagram commutes:

$$\begin{array}{ccccc}
 \Omega U & \xrightarrow{j} & \Lambda U & \xrightarrow{\pi} & \Gamma U \\
 \downarrow D & & \downarrow D & & \downarrow D \\
 \Omega Y & \xrightarrow{j} & \Lambda Y & \xrightarrow{\pi} & \Gamma Y
 \end{array}$$

If x is an element of ΩU or of ΛU , the action of D is given by the product. If x belongs to ΓU , the action of D is given by the product followed by π . In general a ΛA -morphism between ΛU and ΛY is given by a $p \times m$ matrix Z with entries in ΛA . If Z is polynomial we say that the morphism is polynomial.

DEFINITION 1.4. A square polynomial matrix D is said to be *regular* if and only if $\det(D)$ is a unit in ΛA (by Proposition 1.1 this is to say that the leading coefficient of $\det(D)$ is a unit in A). D is said to be *unimodular* if and only if $\det(D)$ is a unit in A .

DEFINITION 1.5. Let Z be a matrix with entries in ΛA . Z is said to be *rational* if and only if there exists a pair (N, D) of polynomial matrices such that D is regular and either $Z = ND^{-1}$ or $Z = D^{-1}N$. Any pair (N, D) such that $Z = ND^{-1}$ (respectively $Z = D^{-1}N$) is called a right (resp. left) *fractional representation* of Z .

Given an A -module M we denote by $T_A(M)$ the torsion submodule of M , i.e., the submodule $T_A(M) = \{x \in M \text{ such that } ax = 0 \text{ for some nonzero } a \in A\}$. We recall that any f.g. A -module M is the direct sum of a free A -module and of $T_A(M)$.

DEFINITION 1.6. Let $M \subset P$ be A -modules. The A -closure of M in P is the A -module $\text{cl}_A(M, P) = \{x \in P \text{ such that } ax \in M \text{ for some nonzero } a \in A\}$. If no confusion arises we will denote the closure shortly by \bar{M} .

M is said to be A -closed in P if and only if $\bar{M} = M$ and M is said to be A -dense in P if and only if $\bar{M} = P$.

PROPOSITION 1.7. Let M, N, P , be A -modules such that $M \subset P$ and $N \subset P$. Then:

- (i) \bar{M} is A -closed in P and M is dense in its closure;
- (ii) if both M and N are A -closed in P , $M \cap N$ is A -closed in P ;
- (iii) $T_A(P/M) = 0$ if and only if M is A -closed in P ;
- (iv) if P is free f.g., M is A -closed in P if and only if it is a direct summand of P ;
- (v) if \bar{M} is free f.g., M and \bar{M} have the same dimension over A ;
- (vi) if \bar{M} is free f.g. and $M \subset N$, $\dim_A M = \dim_A N$ implies $N \subset \bar{M}$;

Proof. (i), (ii), (iii) are trivial; (iv) follows from (iii).

(v) Since A is a p.i.d. there exists a basis x_1, \dots, x_n of \bar{M} and a_1, \dots, a_k elements of A such that $a_1 x_1, \dots, a_k x_k$ is a basis of M . Moreover $k = n$; otherwise, for instance, ax_{k+1} cannot belong to M for any nonzero $a \in A$.

(vi) By tensoring the exact sequence $0 \rightarrow M \rightarrow N \rightarrow N/M \rightarrow 0$ with the quotient field K of A we get that N/M is a torsion A -module. Hence for any $x \in N$ there exists a nonzero $a \in A$ such that $ax \in M$ and so $N \subset \bar{M}$.

PROPOSITION 1.8. *Let $f: N \rightarrow P$ be an A -module morphism. If $M \subset P$ is A -closed, then $f^{-1}(M) \subset N$ is A -closed.*

Remark 1.9. We recall that, given an A -module homomorphism f , a factorization of f is a pair (p, m) of homomorphisms such that $f = mp$. Any f has a factorization (p, m) , where p is surjective and m is injective. Moreover, this factorization is unique up to isomorphisms. If p is a surjective homomorphism and $\text{Ker}(p) \subset \text{Ker}(f)$, then there exists a uniquely determined homomorphism m such that $f = mp$. Dually, if m is injective and $\text{Im}(m) \supset \text{Im}(f)$ there exists a uniquely determined p such that $f = mp$.

2. Systems over a p.i.d. and i/o maps. We give the definition of the systems to be studied.

DEFINITION 2.1. A constant, discrete time, (free) linear dynamical system Σ over A consists of a triple (F, G, H) where F, G and H are matrices of size $n \times n$, $n \times m$, $p \times n$ with entries in A .

We have the following dynamical interpretation in mind. Denote by X, U and Y respectively the free f.g. A -modules A^n, A^m, A^p . Then $F: X \rightarrow X, G: U \rightarrow X, H: X \rightarrow Y$ are A -module maps and Definition 2.1 defines a system whose state, input and output modules are given by X, U, Y and which evolves according to

$$\begin{aligned}x(t+1) &= Fx(t) + Gu(t), \\ y(t) &= Hx(t).\end{aligned}$$

Systems over p.i.d.'s are a natural generalization of systems over fields. The main illustration here will be given by systems over the ring \mathbb{Z} of ordinary integers (see [23] and [27]). Various motivations for the study of systems over p.i.d.'s or over more general rings are given in [20], [24], [26]. The main application areas we outline are the following.

As pointed out in [23], the theory of systems over \mathbb{Z} is of particular interest since one might hope for applications similar to the use of integer programming. A second motivation may be found in the study of 2D-systems. As is shown in [6], [7], [26], [28], it may be useful, in realization problems, to consider a 2D-system as a system over a p.i.d.. Moreover it is possible to see a system over a ring as a family of systems over the spectrum of the ring (see [17]). In this way one can study global properties of families of systems depending on a single parameter. It has to be remarked that our approach does not apply to the study of delay-differential systems over a field K . When all the delays are multiples of a fixed one denoted by σ , a system of this kind is defined (see [20], [21], [27]) by a triple of matrices with elements in $K[\sigma]$ (which is a p.i.d.), but the state, input and output space of the system are assumed to be finite dimensional K -vector spaces and so they cannot have (owing to the finiteness condition) a structure of free $K[\sigma]$ -modules. What is different from this point of view is the dynamical interpretation of the triple of matrices which defines the system. More precisely, using the terminology of [20, §1.3], a delay-differential system over K can be viewed as a "system over the ring of operators $K[\sigma]$ " (the delay σ acts on the function $x(t)$ with values in the state space) while the objects of our study are "systems over a ring of scalars" (the elements of the ring act on the elements of the state space).

So the realization of a delay-differential system defined by a triple of matrices with elements in $K[\sigma]$ must be done on K and not on $K[\sigma]$.

Following [29], Definition 2.1 allows us to carry out the following construction. Remark that $F: X \rightarrow X$ induces on X an $A[z]$ -module structure given by $zx = Fx$. Then any system $\Sigma = (F, G, H)$ gives rise to the following commutative diagram:

$$\begin{array}{ccccc}
 \Omega U & \xrightarrow{f} & \Gamma Y & & \\
 \uparrow i & \searrow \tilde{G} & \nearrow \tilde{H} & & \downarrow p \\
 U & \xrightarrow{G} & X_F & \xrightarrow{H} & Y
 \end{array}$$

where X_F means the A -module X provided with the $A[z]$ -module structure induced by F , \tilde{G} and \tilde{H} are given by 1.4 and $f = \tilde{H}\tilde{G}$.

DEFINITION 2.2. A map $f: \Omega U \rightarrow \Gamma Y$ is called a *linear input/output* (i/o) map if and only if it is $A[z]$ -linear. In the situation above we say that f is the i/o map of Σ .

Any $A[z]$ -linear map $f: \Omega U \rightarrow \Gamma Y$ is determined by $f(e_1), \dots, f(e_m)$, where e_1, \dots, e_m is the natural basis of U . Writing $f(e_i)$ as a column vector, we denote by Z_f the $p \times m$ matrix $f(e_1) \cdots f(e_m)$. By Definition 1.2, Z_f can be seen as a matrix with entries in ΛA . Then it defines a ΛA -morphism $\tilde{f}: \Lambda U \rightarrow \Lambda Y$ such that the following diagram commutes:

$$\begin{array}{ccc}
 \Lambda U & \xrightarrow{\tilde{f}} & \Lambda Y \\
 \uparrow j & & \downarrow \pi \\
 \Omega U & \xrightarrow{f} & \Gamma Y
 \end{array}$$

DEFINITION 2.3. A ΛA -linear map $\tilde{f}: \Lambda U \rightarrow \Lambda Y$ is called an *extended linear i/o* map if and only if all the entries of the matrix defining \tilde{f} are of the form $a = \sum_{t=t_0}^{\infty} a_t z^{-t}$, with $a_t = 0$ for $t \leq 0$. In the situation above we say that \tilde{f} is the extended i/o map associated to f and we call Z_f the transfer function matrix of f .

Remark 2.4. The definition of extended i/o map is built in such a way that it preserves the property of causality. A map $\tilde{f}: \Lambda U \rightarrow \Lambda Y$ which is ΛA -linear is said to be causal if and only if given two input sequences u and u' in ΛU such that $u_t = u'_t$ for $t \leq t_1$, then $\tilde{f}(u)_t = \tilde{f}(u')_t$ for $t \leq t_1 + 1$.

Defining, for $u \in \Lambda U$, $\text{ord}(u) = \min \{t \text{ such that } u_t \neq 0\}$ we may express causality by $\text{ord}(\tilde{f}(u)) < \text{ord}(u)$. It is easy to see that \tilde{f} verifies the above condition if and only if it is an extended i/o map in the sense of Definition 2.3 (see also [16 Def. 2.4]).

Summarizing, we have that any system $\Sigma = (F, G, H)$ gives rise to the following commutative diagram (see [29] and [16] for the case $A = \text{field}$):

$$\begin{array}{ccccc}
 \Lambda U & \xrightarrow{\tilde{f}} & \Lambda Y & & \\
 \uparrow j & & \downarrow \pi & & \\
 \Omega U & \xrightarrow{f} & \Gamma Y & & \\
 \uparrow i & \searrow \tilde{G} & \nearrow \tilde{H} & & \downarrow p \\
 U & \xrightarrow{G} & X_F & \xrightarrow{H} & Y
 \end{array}$$

3. Realizations of i/o maps. Assume that f is a given i/o map. To *realize* f means to find a system $\Sigma = (F, G, H)$ such that f is the i/o map of Σ .

We begin by giving the definition of an abstract realization. It will be shown in Remark 3.3 that abstract realizations which verify a certain finiteness condition actually define a concrete realization, namely a system Σ with the desired property (see also [5, Chapt. 16] and [24]).

DEFINITION 3.1 Let $f: \Omega U \rightarrow \Gamma Y$ be an i/o map. An *abstract realization* of f is a triple (X, g, h) where X is an $A[z]$ -module and $g: \Omega U \rightarrow X$ and $h: X \rightarrow \Gamma Y$ are $A[z]$ -morphisms such that $f = hg$.

DEFINITION 3.2 Let (X, g, h) be a realization of an i/o map f . (X, g, h) is said to be

- (i) *reachable* if and only if g is surjective;
- (ii) *observable* if and only if h is injective;
- (iii) *canonical* if and only if it is both reachable and observable;
- (iv) *finite dimensional* if and only if X is a f.g. A -module;
- (v) *free* if and only if it is f.d. and X is free over A ;
- (vi) *minimal* if and only if it is free and given any other free realization (X', g', h') then $\dim_A X \leq \dim_A X'$.

Remark 3.3. (i) Suppose that A is a field. Then finite dimensional realizations are also free. Moreover, minimal realization are canonical.

In general, R being a ring, canonical realizations of an i/o map over R are not free. This is true for f.d. canonical realizations over a p.i.d. A (see 3.7). In this case we have also that f.d. canonical realizations are minimal among free realizations. The converse is not true (see [27 § 3]).

(ii) Let (X, g, h) be a free realization of f . Then f is the i/o map of the system $\Sigma = (F, G, H)$, where $F: X \rightarrow X$ is defined by $Fx = zx$, $G: U \rightarrow X$ is defined by $G = gi$, $H: X \rightarrow Y$ is defined by $H = ph$.

DEFINITION 3.4. An i/o map f is said to be *realizable* if and only if it has a free realization (X, g, h) . In such a case the system $\Sigma = (F, G, H)$ defined by means of (X, g, h) as in (ii) above is said to realize f .

Remark 3.5. There is no difference between the concept of factorization reviewed in Remark 1.9 and the concept of realization given in Definition 3.1. The second is used instead of the first to point out explicitly the state module X .

We will say that two realizations (X, g, h) and (X', g', h') of the same f are isomorphic if and only if the corresponding factorizations are so; i.e., if and only if there exists an isomorphism $q: X' \rightarrow X$ such that $g = qg'$ and $h' = hq$.

PROPOSITION 3.6. Any i/o map $f: \Omega U \rightarrow \Gamma Y$ has a canonical realization which is unique up to isomorphism.

Proof. By Remarks 1.9 and 3.5. We remark that a canonical realization of f is given by $X = \Omega U / \text{Ker}(f) \cong \text{Im}(f)$, $g = \text{canonical projection}$, $h = \text{canonical inclusion}$.

PROPOSITION 3.7. Let (X, g, h) be an observable realization of the i/o map $f: \Omega U \rightarrow \Gamma Y$. Then $T_A(X) = 0$.

Proof. Since $T_A(\Gamma Y) = 0$ and since h is A -linear, we have $h(T_A(X)) = 0$. Hence, h being injective, $T_A(X) = 0$.

PROPOSITION 3.8. Assume that the i/o map $f: \Omega U \rightarrow \Gamma Y$ has a f.d. realization. Then f is realizable.

Proof. See [5, Chapt. 16, Prop. 3.1].

PROPOSITION 3.9. Assume that the i/o map $f: \Omega U \rightarrow \Gamma Y$ has a f.d. realization. Then f has a free canonical realization.

Proof. See [5, Chapt. 16, Prop. 5.2]. We remark that given the free realization (X, g, h) of f then $(\text{Im}(g)/(\text{Im}(g) \cap \text{Ker}(h)), g', h')$, where g' and h' are defined in the obvious way, is canonical.

4. $A[z]$ -modules which are free f.g. over A . Let $f: \Omega U \rightarrow \Gamma Y$ be an i/o map. Any reachable realization of (X, g, h) of f is uniquely determined by $g: \Omega U \rightarrow X$, and analogously any observable realization is uniquely determined by $h: X \rightarrow \Gamma Y$ (Remark 1.9). Conversely, from Remark 1.9 again, any surjective map $g: \Omega U \rightarrow X$ such that $\text{Ker}(g) \subset \text{Ker}(f)$ determines a unique reachable realization of f and any injective map $h: X \rightarrow \Gamma Y$ such that $\text{Im}(h) \supset \text{Im}(f)$ determines a unique observable realization of f .

Since we are interested in free realizations, this led us to study quotient modules $X = \Omega U / M$ and the submodules $X \subset \Gamma Y$ which are finitely generated and free over A . Such modules are characterized in Propositions, 4.1 and 4.7.

PROPOSITION 4.1. *Let M be an $A[z]$ -submodule of ΩU . Then $X = \Omega U / M$ is a free f.g. A -module if and only if $M = D\Omega U$ for a regular $m \times m$ matrix D .*

Proof. Sufficiency. Assume that $M = D\Omega U$ and let $d(z) = \det(D)$. Since e_1, \dots, e_m generates ΩU , every element x of X is a sum of terms $\alpha_i(z)[e_i]$. The leading coefficient of $d(z)$ being a unit, we can reduce each $\alpha_i(z)$ modulo $d(z)$ without altering the sum. Hence $x \in X$ is determined by m polynomials of degree less than $\deg(d(z))$ and X is generated over A by the $m \times \deg(d(z))$ elements $[e_i]$, $[ze_i]$, \dots , $[z^{(\deg(d(z))-1)}e_i]$ for $i = 1, \dots, m$.

Now let $u \in \Omega U$ and suppose that $au \in D\Omega U$ for some nonzero $a \in A$. This implies $au = Du'$ for some $u' \in \Omega U$ and, denoting by D^* the adjoint of D , we obtain $aD^*u = d(z)u'$. Since a cannot divide $d(z)$ unless it is a unit in A (in such a case $u \in D\Omega U$), a divides each component of u' and $u' = au''$. Then $u = Du'' \in D\Omega U$, $D\Omega U$ is A -closed in ΩU and X is free over A by Proposition 1.7 (iii).

Necessity. Let us first establish the following lemma:

LEMMA 4.2. *Let A be a p.i.d. Given the exact sequence of $A[z]$ -modules and morphisms $0 \rightarrow M \rightarrow A^m[z] \rightarrow X \rightarrow 0$ where X is free over A , then M is free over $A[z]$. Moreover, if X is also f.g. over A , then $\dim_{A[z]} M = m$.*

Proof of Lemma 4.2. A and $A[z]$ are Krull rings by [3, § 1, # 3, Example, and # 9, Proposition 13]. Let $\text{Ass}(X) = \{p, p \subset A[z] \text{ a prime ideal such that } p = \text{Ann}(x) \text{ for some } x \in X\}$, and let P be the set of prime ideals of $A[z]$ of height one, i.e., the set of the minimal nonzero prime ideals of $A[z]$. Our aim is to prove that $\text{Ass}(X) \subset P \cup \{0\}$. Let $p \in \text{Ass}(X)$ be a nonzero ideal and consider the extended ideal $p^e \subset K[z]$ where K is the quotient field of A . Since $K[z] = S^{-1}A[z]$ where $S = A - \{0\}$ and since X is free over A , we have $p \cap S = \emptyset$. Hence p^e is a prime ideal of $K[z]$ and $p = p^{ec} = p^e \cap A[z]$ by [1, Props. 3.11 and 1.17]. Therefore $p \in P$ by [3, § 1, # 9 Remark 2]. In the considered exact sequence, $A^m[z]$ is a reflexive module since it is free and, as we have shown, $\text{Ass}(X) \subset P \cup \{0\}$. Then [3, § 4, # 2, Prop. 7] implies that M is a reflexive $A[z]$ -module.

Moreover, M is f.g. and the global dimension of $A[z]$ is 2. Therefore M is projective f.g. over $A[z]$ (this result is due to Bass, see [10, Prop. 8.25B]) and hence it is free over $A[z]$ by [25].

Now assume that X is also f.g. over A . Denoting by $F: X \rightarrow X$ the A -linear map defined by $Fx = zx$ and by $d(z)$ the determinant of $(zI - F)$, we have $d(z)X = 0$. This implies that X is a torsion $A[z]$ -module. Tensoring the above exact sequence with the quotient field Q of $A[z]$ we have the exact sequence $0 \rightarrow M \otimes_{A[z]} Q \rightarrow Q^m \rightarrow 0$. Hence the vector spaces $M \otimes_{A[z]} Q$ and Q^m are isomorphic and $\dim_{A[z]} M = m$. By the lemma we have $M \cong \Omega U$ and $X = \Omega U / D\Omega U$ where D , by Definition 1.4, is an

$m \times m$ polynomial matrix. Moreover $d(z)e_i \in D\Omega U$ for $i = 1, \dots, m$. Then $d(z)I = DS$ for a suitable matrix S , and D must be regular.

We remark that when A is a field any nonsingular polynomial matrix is regular and Proposition 4.1 coincides with [11, Thm, 3.6].

Moreover any torsion f.g. $A[z]$ -module is automatically free f.g. over A when A is a field. In our case this is not true. For instance $\mathbb{Z}[z]/2\mathbb{Z}[z]$ is easily seen to be not free f.g. over \mathbb{Z} , since 2 is not a unit in \mathbb{Z} .

PROPOSITION 4.3. *Let D, D_1 be $m \times m$ regular matrices. Then $D\Omega U \subset D_1\Omega U$ if and only if there exists a regular $m \times m$ matrix S such that $D = D_1S$. Moreover $D\Omega U = D_1\Omega U$ if and only if S is unimodular.*

Proof. Obvious.

PROPOSITION 4.4. *Let X be an $A[z]$ -submodule of ΓY . Then the following conditions are equivalent:*

- (i) *There exists a $p \times p$ regular matrix D such that $X = \text{Ker}(D) \subset \Gamma Y$.*
- (ii) *X is free f.g. over A and X is A -closed in ΓY .*

Proof. Let D be regular. Applying the snake lemma to the row-exact diagram:

$$\begin{array}{ccccc} \Omega Y & \xrightarrow{id} & \Omega Y & \longrightarrow & 0 \\ \downarrow m(z) & & \downarrow m(z) & & \downarrow m(z) \\ \Omega Y & \xrightarrow{j} & \pi^{-1}(X) & \xrightarrow{\pi} & X \end{array},$$

we have that $\text{Ker}(D)$ is isomorphic to $\Omega Y/D\Omega Y$, which is free f.g. over A by Proposition 4.1. Moreover, if $y \in \Gamma Y$ and $a \in A$, $a \neq 0$, $D(ay) = a(Dy) = 0$ implies $y \in \text{Ker}(D)$ since ΓY has no torsion over A .

Conversely, let $F: X \rightarrow X$ be the A -linear map defined by $Fx = zx$ and denote by $m(z)$ the minimal polynomial of F . $m(z)$ is monic and $m(z)X = 0$, hence $m(z)(\pi^{-1}(X)) \subset \Omega Y$. Applying the snake lemma to the row-exact diagram:

$$\begin{array}{ccccc} \Omega Y & \xrightarrow{j} & \Lambda Y & \xrightarrow{\pi} & \Gamma Y \\ \downarrow D & & \downarrow D & & \downarrow D \\ \Omega Y & \xrightarrow{j} & \Lambda Y & \xrightarrow{\pi} & \Gamma Y \end{array},$$

we have the exact sequence $0 \rightarrow X \rightarrow \Omega Y/m(z)\Omega Y \rightarrow \Omega Y/m(z)(\pi^{-1}(X)) \rightarrow 0$. Since $\Omega Y/m(z)\Omega Y$ is f.g. over A , $\Omega Y/m(z)(\pi^{-1}(X))$ is also f.g. over A . Moreover, by Proposition 1.8 $\pi^{-1}(X)$ is A -closed in ΛY since X is A -closed in ΓY .

We claim that also $m(z)(\pi^{-1}(X))$ is A -closed in ΩY . In fact, let $u \in \Omega Y$, $a \in A$, $a \neq 0$ and $au = m(z)y$ for some $y \in \pi^{-1}(X)$. Since a cannot divide $m(z)$ unless it is a unit in A (in such a case $u \in m(z)(\pi^{-1}(X))$), a divides y in ΛY . Then $y = ay'$, $y' \in \pi^{-1}(X)$ and $u = m(z)y' \in m(z)(\pi^{-1}(X))$.

By Proposition 1.7 (iii), $\Omega Y/m(z)(\pi^{-1}(X))$ is free over A . Then there exists a $p \times p$ regular matrix D_1 such that $m(z)(\pi^{-1}(X)) = D_1\Omega Y$. So we have $\Omega Y \subset \pi^{-1}(X) = (m(z))^{-1}D_1\Omega Y$. It follows that there exists a $p \times p$ regular matrix D such that $I = (m(z))^{-1}D_1D$ and $\pi^{-1}(X) = D^{-1}\Omega Y$. Now it is easy to check that $X = \text{Ker}(D)$.

We remark that when A is a field any submodule of ΓY is A -closed and Proposition 4.4 coincides with [4, Prop. 1]. In our case this is not true. For instance, $2z^{-1}\mathbb{Z}$ is not \mathbb{Z} -closed in $\Gamma\mathbb{Z}$ since $2z^{-1} \in 2z^{-1}\mathbb{Z}$ but $z^{-1} \notin 2z^{-1}\mathbb{Z}$.

The proof of Proposition 4.4 is based on an alternative proof of [4, Prop. 1] suggested to the authors by M. Hautus in a personal communication.

PROPOSITION 4.5. *Let D, D_1 be $p \times p$ regular matrices. Then $\text{Ker}(D) \subset \text{Ker}(D_1) \subset \Gamma Y$ if and only if there exists a $p \times p$ regular matrix S such that $D_1 = SD$. Moreover, $\text{Ker}(D) = \text{Ker}(D_1)$ if and only if S is unimodular.*

Proof. Let $\text{Ker}(D) \subset \text{Ker}(D_1)$. Then $\pi^{-1}(\text{Ker}(D)) = D^{-1}\Omega Y \subset \pi^{-1}(\text{Ker}(D_1)) = D_1^{-1}\Omega Y$. Therefore there exists a matrix S such that $D^{-1} = D_1^{-1}S$. Then $D_1 = SD$ and S is regular. Conversely, if $D_1 = SD$ obviously $\text{Ker}(D) \subset \text{Ker}(D_1)$. The second part is obvious.

PROPOSITION 4.6. *Let X be an $A[z]$ -submodule of ΓY . If X is free f.g. over A , then $\text{cl}_A(X, \Gamma Y)$ is free f.g. over A .*

Proof. Denote by $F: X \rightarrow X$ the A -linear map defined by $Fx = zx$ and denote by $d(z)$ the determinant of $(zI - F)$. We have $X \subset \text{Ker}(d(z)I)$, which is free f.g. over A and A -closed in ΓY by Proposition 4.4. Hence $\text{cl}_A(X, \Gamma Y)$ is contained in $\text{Ker}(d(z)I)$ and as a consequence it is free f.g. over A .

COROLLARY 4.7. *Let X be an $A[z]$ -submodule of ΓY . X is free f.g. over A if and only if there exists a $\tilde{p} \times p$ regular matrix D such that $\bar{X} = \text{Ker}(D)$.*

Proof. Obvious by Propositions 4.4 and 4.6.

PROPOSITION 4.8. *Let $X \subset \Gamma Y$ be free f.g. over A and let $\text{Ker}(D) = \bar{X}$. If D_1 is a $p \times p$ regular matrix such that $X \subset \text{Ker}(D_1)$, then there exists a $p \times p$ regular matrix S such that $D_1 = SD$.*

Proof. By Proposition 4.4 $\text{Ker}(D_1)$ is A -closed; hence, by Proposition 1.7 (ii), $\bar{X} = \text{Ker}(D) \subset \text{Ker}(D_1)$ and the thesis follows from Proposition 4.5.

Example 4.9. Let $A = \mathbb{Z}$ and $X = 2z^{-1}\mathbb{Z}$. As we have already seen X is not \mathbb{Z} -closed in $\Gamma\mathbb{Z}$. We have $\bar{X} = z^{-1}\mathbb{Z} = \text{Ker}(z) \subset \Gamma\mathbb{Z}$. Moreover, any polynomial $p(z)$ such that $X \subset \text{Ker}(p(z))$ must have the constant term equal to zero. Therefore, according to Proposition 4.8, $p(z)$ is a multiple of z .

5. Realizations and fractional representation. In the previous section we proved that any quotient $A[z]$ -module $X = \Omega U / M$ (equivalently, any surjective homomorphism $g: \Omega U \rightarrow X$) determines, when X is free f.g. over A , a unique, up to unimodular factors, $m \times m$ regular matrix D_X defined by $D_X \Omega U = M$ (resp., $D_X \Omega U = \text{Ker}(g)$). Conversely, any regular $m \times m$ matrix D determines the free f.g. A -module $X_D = \Omega U / D \Omega U$. In this case we denote by $g_D: \Omega U \rightarrow X_D$ the canonical projection. More explicitly, we have a one-one correspondence between quotient modules of the previous kind and regular matrices (modulo unimodular factors).

On the other hand, any $A[z]$ -submodule X of ΓY (equivalently any injective homomorphism $h: X \rightarrow \Gamma Y$) determines, when X is free f.g. over A , a unique, up to unimodular factors, $p \times p$ regular matrix ${}_X D$ defined by $\text{Ker}({}_X D) = \text{cl}_A(X, \Gamma Y)$ (resp. $\text{Ker}({}_X D) = \text{cl}_A(\text{Im}(h), \Gamma Y)$). Conversely any $p \times p$ regular matrix D determines the free f.g. A -module ${}_D X = \text{Ker}(D)$ which is A -closed in ΓY . In this case we denote by ${}_D h: {}_D X \rightarrow \Gamma Y$ the canonical inclusion.

Here the correspondence is one-one only between A -closed submodules of the previous kind and regular matrices.

As we already remarked at the end of Proposition 4.4 the two situations are duals of each other when A is a field.

Remark 5.1. Quotient modules and submodules are defined up to isomorphisms. Since no confusion arises in our case, we will not take account of this fact and in the

following, given an injective homomorphism $h : X \rightarrow \Gamma Y$, we will identify, by abuse of notation, X with $\text{Im}(h)$.

The following propositions are stated with the notation above.

PROPOSITION 5.2. *Let $f : \Omega U \rightarrow \Gamma Y$ be an i/o map and let Z_f be its transfer function matrix. Any free reachable realization (X, g, h) of f determines a right fractional representation of Z_f given by $Z_f = ND_X^{-1}$. If (X, g, h) is canonical then N and D_X are left coprime (i.e., $N = N_1T$ and $D_X = D_1T$ implies that T is unimodular).*

Conversely, any right fractional representation $Z_f = ND^{-1}$ determines a free reachable realization (X_D, g_D, h_D) which is canonical if N and D are left coprime.

Proof. Assume that (X, g, h) is a free reachable realization. By $f = hg$ we have $D_X \Omega U = \text{Ker}(g) \subset \text{Ker}(f)$. This implies that $fD_X : \Omega U \rightarrow \Gamma Y$ is the zero map and that, as a consequence, $\tilde{f}D_X : \Lambda U \rightarrow \Lambda Y$ is a polynomial map. Denoting by N the polynomial matrix associated to $\tilde{f}D_X$, we have $Z_f D_X = N$. Hence $Z_f = ND_X^{-1}$.

Now assume that (X, g, h) is also canonical and let $N = N_1T$ and $D_X = D_1T$. We have $Z_f = N_1D_1^{-1}$; therefore $\tilde{f}D_1 : \Lambda U \rightarrow \Lambda Y$ is polynomial and, as a consequence, $D_1 \Omega U \subset \text{Ker}(f)$. By Proposition 3.6 X is isomorphic to $\Omega U / \text{Ker}(f)$, so $\text{Ker}(f) = D_X \Omega U$ and $D_1 = D_X S$ by Proposition 4.3. Now $D_X = D_X S T$ implies that T is unimodular.

Conversely, let $Z_f = ND^{-1}$ be a fractional representation. As above we have $D \Omega U \subset \text{Ker}(f)$ and defining $h_D : X_D \rightarrow \Gamma Y$ by means of Remark 1.9 we obtain a realization (X_D, g_D, h_D) of f which is free and reachable by construction.

Now assume that N and D are left coprime and consider the canonical realization of f determined by $X = \Omega U / \text{Ker}(f)$. Since (X_D, g_D, h_D) is in particular f.d., Proposition 3.9 assures that $\text{Ker}(f) = D_1 \Omega U$ for a regular matrix D_1 and by $D \Omega U \subset \text{Ker}(f)$ and Proposition 4.3 we have $D = D_1 T$. Denoting by N_1 the polynomial matrix associated to $\tilde{f}D_1 : \Lambda U \rightarrow \Lambda Y$, we obtain $N = N_1 T$. Hence T is unimodular and (X_D, g_D, h_D) is canonical.

PROPOSITION 5.3. *Let $f : \Omega U \rightarrow \Gamma Y$ be an i/o map and let Z_f be its transfer function matrix. Any free observable realization (X, g, h) of f determines a left fractional representation of Z_f given by $Z_f = {}_X D^{-1} N$. If (X, g, h) is minimal then N and ${}_X D$ are right coprime.*

Conversely any left fractional representation $Z_f = D^{-1} N$ determines a free observable realization $({}_D X, {}_D g, {}_D h)$ of f which is minimal if N and D are right coprime.

Proof. Assume that (X, g, h) is a free observable realization. By $f = hg$ we have $\text{Im}(f) \subset h(X) \subset \text{cl}_A(h(X), \Gamma Y) = \text{Ker}({}_X D)$. This implies that ${}_X D f : \Omega U \rightarrow \Gamma Y$ is the zero map and that ${}_X D \tilde{f} : \Lambda U \rightarrow \Lambda Y$ is a polynomial map. Denoting by N the polynomial matrix associated to ${}_X D f$, we have ${}_X D Z_f = N$. Hence $Z_f = {}_X D^{-1} N$.

Now assume that (X, g, h) is also minimal and let $N = T N_1$ and ${}_X D = T D_1$. We have $Z_f = D_1^{-1} N_1$; therefore $D_1 \tilde{f} : \Lambda U \rightarrow \Lambda Y$ is polynomial, which implies $\text{Im}(f) \subset \text{Ker}(D_1)$ and $\overline{\text{Im}}(f) \subset \text{Ker}(D_1)$. Consider the canonical realization of f determined by $\text{Im}(f)$, which is free and minimal by Proposition 3.9 and Remark 3.3(i). We have $\text{Im}(f) \subset h(X)$ and $\dim_A \text{Im}(f) = \dim_A h(X)$ by the minimality of (X, g, h) . By Proposition 1.7 (vi), $h(X) \subset \overline{\text{Im}}(f)$; therefore $\overline{h(X)} = \text{Ker}({}_X D) \subset \text{Im}(f) \subset \text{Ker}(D_1)$. Then by Proposition 4.8 $D_1 = S {}_X D$ and ${}_X D = T S {}_X D$ implies that T is unimodular.

Conversely, let $Z_f = D^{-1} N$ be a fractional representation. As above we have $\text{Im}(f) \subset \text{Ker}(D)$, and defining ${}_D g : \Omega U \rightarrow {}_D X$ by means of Remark 1.9 we obtain a realization $({}_D X, {}_D g, {}_D h)$ of f which is free and observable by construction.

Now assume that N and D are right coprime and consider the canonical realization of f determined by $\text{Im}(f)$. Since $({}_D X, {}_D g, {}_D h)$ is, in particular, f.d., Proposition 3.9 assured that $\overline{\text{Im}}(f) = \text{Ker}(D_1)$ for a regular matrix D_1 . Then $\overline{\text{Im}}(f) =$

$\text{Ker}(D_1) \subset \text{Ker}(D)$ and $D = TD_1$. Denoting by N_1 the polynomial matrix associated to $D_1 \tilde{f}: \Lambda U \rightarrow \Lambda Y$ we have $N = TN_1$. Hence T is unimodular and by Proposition 4.5 $\text{Ker}(D) = \text{Ker}(D_1)$. Therefore by Proposition 1.7 (v) $\dim_A({}_D X) = \dim_A(\text{Im}(f))$ and $({}_D X, {}_D g, {}_D h)$ is minimal.

Propositions 5.2 and 5.3 establish a correspondence between reachable or observable realizations of an i/o map and right or left fractional representations of its transfer function matrix. The analogous result for the case where A is a field was proved in [16] for reachable realizations and in [8] and [4], with different techniques, for observable realizations. A well-known consequence of this fact is the equivalence between rationality and the realizability property.

The correspondence is one-one in the case of free reachable realizations and right fractional representations. For free observable realizations and left fractional representations the situation is slightly different. Consider, for instance, the i/o map $f: \Omega \mathbb{Z} \rightarrow \Gamma \mathbb{Z}$ defined by $f(\sum_{i=0}^n a_i z^i) = 2a_0 z^{-1}$, whose transfer function matrix is $Z_f = [2z^{-1}]$. $\text{Im}(f)$ is the submodule $2z^{-1}\mathbb{Z}$ of $\Gamma \mathbb{Z}$ which we have already considered, and a canonical realization of f is given by (\mathbb{Z}, g, h) , where $g: \Omega \mathbb{Z} \rightarrow \mathbb{Z}$ is defined by $g(\sum_{i=0}^n a_i z^i) = a_0$ and $h: \mathbb{Z} \rightarrow \Gamma \mathbb{Z}$ by $h(a) = 2az^{-1}$ (equivalently by the system $\Sigma = (0, 1, 2)$). By Example 4.9 and Proposition 5.3, (\mathbb{Z}, g, h) determines the fractional representation $Z_f = [z]^{-1}[2]$, which in turn gives us the realization (\mathbb{Z}, g', h') where $g'(\sum_{i=0}^n a_i z^i) = 2a_0$ and $h'(a) = az^{-1}$ (equivalently, the system $\Sigma' = (0, 2, 1)$). The two realizations are not isomorphic since the second is minimal but not reachable.

To have a one-one correspondence also in the case of free observable realizations we must restrict ourselves to considering only free observable realizations (X, g, h) such that $h(X)$ is A -closed in ΓY . In this way we lose a large class of minimal realizations, except when $\text{Im}(f)$ is itself A -closed in ΓY . In this case, in fact, the following proposition holds:

PROPOSITION 5.4. *Let $f: \Omega U \rightarrow \Gamma Y$ be a realizable i/o map and assume that $\text{Im}(f)$ is A -closed in ΓY . Then any free minimal realization of f is canonical.*

Proof. Consider the canonical realization of f determined by $\text{Im}(f)$ and let (X, g, h) be a free minimal realization of f . We have $\dim_A X = \dim_A \text{Im}(f)$ by the minimality and, since $h(X)$ is free over A as it is f.g. and without torsion, also $\dim_A X \cong \dim_A h(X)$. Moreover, by $\text{Im}(f) \subset h(X)$, $\dim_A h(X) \geq \dim_A \text{Im}(f)$. Hence $\dim_A h(X) = \dim_A \text{Im}(f)$ and $h(X) = \text{Im}(f)$ by Proposition 1.7 (vi). Using again the equality $\dim_A X = \dim_A h(X)$, we have that h is an isomorphism onto its image. Now it is easy to check that h is actually an isomorphism between (X, g, h) and the canonical realization given by $\text{Im}(f)$.

For a p.i.d. A , in particular for a field, the state module X of a free canonical realization of a realizable i/o map f is determined up to A -isomorphisms by its dimension, as was proved in [19]. $\dim_A X$ is referred to as the McMillan degree of f . The following propositions relate the McMillan degree of f to the degree, as a polynomial in z , of the determinant of D , where D is a regular matrix in a coprime fractional representation of Z_f .

PROPOSITION 5.5. *Let D be a regular $m \times m$ matrix and let X_D and ${}_D X$ be as described at the beginning of this section. Assume that $n = \deg(\det(D))$; then $\dim_A(X_D) = \dim_A({}_D X) = n$.*

Proof. The first equality follows from the snake lemma isomorphism between X_D and ${}_D X$ obtained in the proof of Proposition 4.4, so we have only to prove that $\dim_A(X_D) = n$. For this purpose consider the exact sequence $0 \rightarrow \Omega U \xrightarrow{D} \Omega U \rightarrow X_D \rightarrow 0$ as a sequence of A -modules, and write it as $0 \rightarrow A^m[z] \xrightarrow{D} A^m[z] \rightarrow A^k \rightarrow 0$. Tensoring

with the quotient field K of A we get the exact sequence of K -vector spaces $0 \rightarrow K^m[z] \xrightarrow{D} K^m[z] \xrightarrow{p} K^k \rightarrow 0$. Now D is $K[z]$ -linear, hence this last sequence may be viewed as an exact sequence of $K[z]$ -modules, where K^k is endowed with the $K[z]$ -module structure induced by p . In this way, since D has not been changed and $\dim_A(X_D) = \dim_K(K^k)$, we have to prove our assertion only for the case $A = \text{field}$.

In this case $A[z]$ is itself a p.i.d. and we have $D = PD_1Q$, where D_1 is diagonal, $\det(D_1) = \det(D)$ up to units and P, Q are unimodular (see [2, § 4 # 5, Cor. 1]). Then by the commutativity of the diagram

$$\begin{array}{ccccccc}
 0 & \longrightarrow & \Omega U & \xrightarrow{D} & \Omega U & \longrightarrow & X_D \longrightarrow 0 \\
 & & \downarrow Q & & \downarrow p^{-1} & & \\
 0 & \longrightarrow & \Omega U & \xrightarrow{D_1} & \Omega U & \longrightarrow & X_1 \longrightarrow 0
 \end{array}$$

we have that there exists an isomorphism between X and X_1 . Hence $\dim_A(X) = \dim_A(X_1)$. Let $D_1 = \text{diag}(p_1(z), \dots, p_m(z))$ with $n_i = \deg(p_i(z))$; then X_1 is generated over A by $[e_1], [ze_1], \dots, [z^{n_1-1}e_1], \dots, [e_m], [ze_m], \dots, [z^{n_m-1}e_m]$, which are exactly $n_1 + n_2 + \dots + n_m = n$ linearly independent elements.

PROPOSITION 5.6. *Let f be an i/o map and let $Z_f = ND^{-1}$ (resp. $Z_f = D_1^{-1}N_1$) be a coprime fractional representation of its function matrix. If $n = \deg(\det(D))$ (resp. $n = \deg(\det(D_1))$), then the McMillan degree of f is n .*

Proof. By Propositions 5.2 (5.3) and 5.5.

That the McMillan degree is equal to $\deg(\det(D))$ was proved, in the case $A = \text{field}$, in [18], using algebro-geometric techniques. Our proof generalizes this result and it is based only on the possibility of reducing any matrix over a p.i.d. to diagonal form.

REFERENCES

- [1] M. F. ATIYAH AND I. G. MACDONALD, *Introduction to Commutative Algebra*, Addison-Wesley, Reading, MA, 1969.
- [2] N. BOURBAKI, *Algèbre*, Hermann, Paris, 1974, Chapter VII.
- [3] ———, *Algèbre commutative*, Hermann, Paris, 1965, Chapter VII.
- [4] G. CONTE AND A. M. PERDON, *Sur les représentations fractionnaires et les réalisations observables de dimension finie des relations linéaires entre-sortie*, CRAS 280 Sér. A (1980), pp. 515–517.
- [5] S. EILENBERG, *Automata, Languages and Machines*, vol. A, Academic Press, New York, 1974.
- [6] R. EISING, *Realization and stabilization of 2D-systems*, IEEE Trans. Automat. Control, AC-23, (1978), pp. 793–799.
- [7] ———, *2D-systems, an algebraic approach*, Mathematical Centre Tracts 125, Amsterdam, 1980.
- [8] E. EMRE AND M. L. J. HAUTUS, *A polynomial characterization of (A, B) -invariant and reachability subspaces*, Mem. COSOR 78/19 Eindhoven Univ. of Technology, Eindhoven, The Netherlands, 1978.
- [9] E. EMRE AND P. KHARGONEKAR, *Regulation of linear systems over rings: coefficient assignment and observers*, Center for Math. System Theory, Univ. of Florida, Gainesville, 1980.
- [10] C. FAITH, *Algebra: Rings, Modules and Categories*, Springer-Verlag, New York/Berlin, 1973.
- [11] P. A. FUHRMANN, *Algebraic system theory: an analyst's point of view* J. Franklin Inst., 30 (1976), pp. 521–540.
- [12] ———, *On strict system equivalence and similarity*, Internat. J. Control, 25 (1977), pp. 5–10.
- [13] ———, *Simulation of linear systems and factorization of matrix polynomials*, Internat. J. Control, 28 (1978), pp. 689–705.

- [14] ———, *Linear feedback via polynomial models*, Internat. J. Control, 30 (1979), pp. 363–377.
- [15] P. A. FUHRMANN AND J. C. WILLEMS, *A study of (A, B) -invariant subspaces via polynomial models*, Internat. J. Control, 31 (1980), pp. 467–494.
- [16] M. L. J. HAUTUS AND M. HEYMANN, *Linear feedback,—an algebraic approach*, this Journal, 16 (1978), pp. 83–105.
- [17] M. HAZEWINKEL, *A partial survey of the use of algebraic geometry in system and control theory*, Severi Centennial Conference INDAM, Rome, 1979.
- [18] C. MARTIN AND R. HERMANN, *Applications of algebraic geometry to systems theory: the McMillan degree and Kronecker indices of transfer functions as topological and holomorphic system invariants*, this Journal, 16 (1978), pp. 743–755.
- [19] R. E. KALMAN, *Irreducible realizations and the degree of a rational matrix*, J. Soc. Ind. Appl. Math., 13 (1965), pp. 520–544.
- [20] E. W. KAMEN, *On an algebraic theory of systems defined by convolution operators*, Math. System Theory, 9 (1975), pp. 57–74.
- [21] ———, *Lectures on algebraic system theory. Linear systems over rings*, NASA Contractor Report 3016, 1976.
- [22] P. KHARGONEKAR, *On matrix fraction representation for linear systems over commutative rings*, Center for Math. System Theory, Univ. of Florida, Gainesville, July, 1980.
- [23] Y. ROUCHALEAU AND B. WYMAN, *Linear dynamical systems over integral domains*, J. Comput. System Sci., 9 (1974), pp. 129–142.
- [24] Y. ROUCHALEAU, B. WYMAN AND R. E. KALMAN, *Algebraic structure of linear dynamical systems III. Realization theory over a commutative ring*, Proc. Nat. Acad. Sci., 69 (1972), pp. 3404–3406.
- [25] C. S. SESHADRI, *Triviality of vector bundles over the affine space K^2* , Proc. Nat. Acad. Sci., 44 (1958), pp. 456–458.
- [26] E. D. SONTAG, *On linear systems and noncommutative rings*, Math. System Theory, 9 (1975), pp. 327–344.
- [27] ———, *Linear systems over commutative rings. A survey*, Ricerche di Automatica, 7 (1976), pp. 1–34.
- [28] ———, *On first order equations for multidimensional filters*, IEEE Trans. Acoustics, Speech and Signal Proc., 26 (1978), pp. 480–482.
- [29] B. WYMAN, *Linear Systems over Commutative Rings*, Lecture Notes, Stanford University, Stanford, CA, 1972.

BOUNDARY CONTROL PROBLEMS WITH NONLINEAR STATE EQUATION*

VIOREL BARBU†

Abstract. First order necessary conditions of optimality are obtained for boundary control problems governed by parabolic equations with nonlinear boundary value conditions.

1. Introduction. We are concerned here with first order necessary conditions of optimality for convex control problems governed by nonlinear boundary-value problems of the form

$$(1.1) \quad \begin{aligned} y_t + Ay &= 0 && \text{in } Q = \Omega \times]0, T[, \\ \frac{\partial y}{\partial \nu} + \beta_i(y) &\ni B_i u_i + f_i && \text{in } \Sigma_i = \Gamma_i \times]0, T[, \quad i = 1, 2, \\ y(x, 0) &= y_0(x) && \text{in } \Omega. \end{aligned}$$

Here Ω is a bounded and open subset of the Euclidean space R^N , A is a second order elliptic and symmetric operator on Ω and β_i are maximal monotone graphs (in general multivalued) in $R \times R$. The controls u_i are taken from the Hilbert spaces U_i and B_i are linear continuous operators from U_i to $L^2(\Sigma_i)$, $i = 1, 2$. The functions y_0 and f_i are fixed in $L^2(\Omega)$ and $L^2(\Sigma_i)$, $i = 1, 2$, respectively.

The boundary Γ of Ω consists of two disjoint parts Γ_1 and Γ_2 , i.e., $\Gamma = \Gamma_1 \cup \Gamma_2$ and $\Gamma_1 \cap \Gamma_2 = \emptyset$.

To make what follows more meaningful, let us briefly describe some classical diffusion problems of the form (1.1) (see [7], [9] for further examples and complete references).

Newton's law of heat conduction is described by (1.1), where β_i are continuous and nondecreasing functions.

The Stefan-Boltzman heat radiation law. The functions β_i are of the following form:

$$(1.2) \quad \beta_i(r) = \begin{cases} a_i(r-c)^4 & \text{if } r \geq c, \\ 0 & \text{if } r < c, \end{cases}$$

where $a_i > 0$, $i = 1, 2$.

Natural convection.

$$(1.3) \quad \beta_i(r) = \begin{cases} ar^{5/4} & \text{if } r \geq 0, \\ 0 & \text{if } r < 0, \end{cases} \quad a > 0, \quad i = 1, 2.$$

Enzyme diffusion (the Michaelis-Menten law) is described by (1.1) where $\Gamma = \Gamma_1$ and $\beta = \beta_1$ is given by

$$(1.4) \quad \beta(r) = \begin{cases} \frac{r}{r+m} & \text{for } r > 0, \\]-\infty, 0] & \text{for } r = 0, \\ \emptyset & \text{for } r < 0. \end{cases}$$

* Received by the editors December 8, 1980, and in final form May 15, 1981.

† Faculty of Mathematics, University of Iași, Iași 6600, Romania.

The thermostat-control process. $\Gamma_1 = \Gamma$ and

$$(1.5) \quad \beta_1(r) = \begin{cases} \alpha_1(r - \theta_1) & \text{if } -\infty < r < \theta_1, \\ 0 & \text{if } \theta_1 \leq r \leq \theta_2, \\ \alpha_2(r - \theta_2) & \text{if } \theta_2 < r < \infty. \end{cases}$$

Here α_1 and α_2 are positive numbers.

The Signorini problem. The graph $\beta(\Gamma = \Gamma_1$ and $\beta_1 = \beta)$ is given by

$$(1.6) \quad \beta(r) = \begin{cases} 0 & \text{if } r > 0, \\]-\infty, 0] & \text{if } r = 0, \\ \emptyset & \text{if } r < 0. \end{cases}$$

The contents of this paper are outlined below. In § 2 we shall study existence and approximation of solutions for the boundary control system (1.1). In § 3 we give the main results, Theorems 1 and 2, which are concerned with necessary conditions for optimality in a control problem with a convex cost criterion governed by (1.1) in two typical cases: β_i locally Lipschitzian functions and the Signorini problem (1.6).

The proofs are presented in detail in §§ 5, 6. The main idea of our approach consists of approximating the control problem by a family of smooth problems for which the optimality equations are immediate, and then passing to a limit in the approximating equations. In § 4 we study the convergence of this approximating control process.

The results as well as the approach used here are similar to those in other efforts by the author [1], [2], [3]. For comparison with other literature on necessary conditions for boundary control problems the works [13], [14] are most closely related to the present paper. In particular, Theorem 1 includes and refines the results of [13].

The following notation will be used in the sequel. Given a real Banach space E , and $[0, T]$ a real interval, we shall denote by $L^p(0, T; E)$, $1 \leq p \leq \infty$ the space of all p -integrable E -valued functions on $]0, T[$ and by $C([0, T]; E)$ the Banach space of all continuous functions from $[0, T]$ to E . By $C_w([0, T]; E)$ we shall denote the space of all functions continuous from $[0, T]$ to the space E endowed with the weak topology.

Given a lower semicontinuous convex function $\varphi : E \rightarrow \bar{R} =]-\infty, +\infty]$ we shall denote by $\partial\varphi(x) \in E'$ (E' is the dual space of E) the set of all *subgradients* of φ at x , i.e.,

$$(1.7) \quad \partial\varphi(x) = \{x^* \in E'; \varphi(x) \leq \varphi(y) + (x^*, x - y) \text{ for all } y \in E\}.$$

If φ is Gâteaux differentiable at x , then $\partial\varphi(x)$ consists of a single element, namely the gradient $\nabla\varphi(x)$ of φ at x . The mapping $\partial\varphi : E \rightarrow E'$ is called the *subdifferential* of φ . If f is a locally Lipschitzian function on the real axis R , the *generalized gradient* of β , $\partial\beta$ (in the sense of Clarke [6]) is defined by

$$(1.8) \quad \partial\beta(r) = \text{conv} \{y \in R; y = \lim_{r_n \rightarrow r} \beta'(r_n)\}, \quad r \in R,$$

where $\nabla\beta = \beta'$ denotes the ordinary derivative of β . For other concepts and results in convex analysis relevant to this paper we refer the reader to [4], [5], [8], [12].

Let k, r, s be real numbers. We shall denote by $H^k(\Omega)$, $H^k(\Gamma)$, $H^{r,s}(Q)$ and $H^{r,s}(\Sigma)$ the usual Sobolev spaces on Ω , Γ , Q and Σ , respectively (see, e.g., [10, p. 14]). By $L^2(\Omega)$, $L^2(\Gamma)$, $L^2(Q)$ and $L^2(\Sigma)$ we shall denote the corresponding spaces of square integrable functions. Finally we shall denote by $W(Q)$ the space of all functions $y \in L^2(0, T; H^1(\Omega))$ such that $(d/dt)y \in L^2(0, T; (H^1(\Omega))')$. (Here $(H^1(\Omega))'$ is the dual space of $H^1(\Omega)$ and dy/dt denotes the derivative of $y(t)$ in the sense of $(H^1(\Omega))'$ -valued

distributions on $]0, T[$. $W(Q)$ is a Banach space with the natural norm

$$(1.9) \quad \|y\|_{W(Q)}^2 = (\|y\|_{L^2(0,T;H^1(\Omega))}^2 + \left\| \frac{dy}{dt} \right\|_{L^2(0,T;(H^1(\Omega))^*)}^2),$$

and it is well known that $W(Q) \subset C([0, T]; L^2(\Omega))$, algebraically and topologically.

2. The boundary control system. Let Ω be a bounded and open subset of R^N with a sufficiently smooth boundary Γ . We shall assume that Γ consists of two smooth and disjoint parts Γ_1 and Γ_2 , where $\text{meas}(\Gamma_1) > 0$ (except for the case $N = 1$ and $\Omega =]a, b[$ when $\Gamma_1 = \{a\}$ or $\Gamma_1 = \{a\} \cup \{b\}$).

Let A be a second order differential operator on Ω of the form

$$Ay = - \sum_{i,j=1}^N (a_{ij}(x)y_{x_i})_{x_j} + a(x)y,$$

where $a_{ij} \in C^1(\bar{\Omega})$, $a \in L^\infty(\Omega)$, $a_{ij} = a_{ji}$ for all i, j and, for some $\omega > 0$,

$$a \geq 0, \quad \sum_{i,j=1}^N a_{ij}\xi_i\xi_j \geq |\xi|^2 \quad \text{a.e. on } \Omega, \quad \xi \in R^n.$$

(Here y_{x_i} denotes the partial derivative of y with respect to x_i .)

For $y_0 \in L^2(\Omega)$ and $v_i \in L^2(\Sigma_i)$, $i = 1, 2$, consider the system

$$(2.1) \quad \begin{aligned} y_t + Ay &= 0 && \text{in } Q, \\ \frac{\partial y}{\partial \nu} + \beta_i(y) &\ni v_i && \text{in } \Sigma_i, \quad i = 1, 2, \\ y(x, 0) &= y_0(x) && x \in \Omega, \end{aligned}$$

where y_t stands for the partial derivative $\partial y / \partial t$ while $\partial y / \partial \nu$ is the outward normal derivative associated with A .

Here β_i , $i = 1, 2$ are two maximal monotone graphs in $R \times R$ which satisfy the conditions

$$(2.2) \quad \beta_i(0) \ni 0, \quad i = 1, 2.$$

Let us now give a precise meaning to system (2.1).

DEFINITION 1. A function $y \in W(Q)$ is a solution to (2.1) if there exist functions $w_i \in L^2(\Sigma_i)$, $i = 1, 2$, such that

$$(2.3) \quad w_i(\sigma, t) \in \beta_i(y(\sigma, t)) \quad \text{a.e. } (\sigma, t) \in \Sigma_i, \quad i = 1, 2$$

and

$$(2.4) \quad - \int_Q y \kappa_t dx dt + \int_0^T a(y, \kappa) dt + \sum_{i=1}^2 \int_{\Sigma_i} (w_i - v_i) \kappa d\sigma dt = \int_\Omega y_0(x) \kappa(x, 0) dx,$$

for all $\kappa \in W(Q)$ such that $\kappa(x, T) = 0$. Here $a : H^1(\Omega) \times H^1(\Omega) \rightarrow R$ is the bilinear functional

$$(2.5) \quad a(y, z) = \sum_{i,j=1}^N \int_\Omega (a_{ij} y_{x_i} z_{x_j} + a y z) dx, \quad y, z \in H^1(\Omega).$$

Condition (2.4) can be equivalently defined as

$$(2.6) \quad \begin{aligned} \frac{d}{dt}(y(t), \psi) + a(y(t), \psi) + \sum_{i=1}^2 \int_{\Gamma_i} (w_i - v_i) \psi \, d\sigma &= 0 \quad \text{a.e. } t \in]0, T[, \\ y(0) &= y_0, \end{aligned}$$

for all $\psi \in H^1(\Omega)$. Here (\cdot, \cdot) is the usual inner product in $L^2(\Omega)$, and will be also used to denote the pairing between $H^1(\Omega)$ and $(H^1(\Omega))'$.

Let ρ be a C_0^∞ -function on R satisfying $\rho(r) > 0$ for $r \in]-1, 1[$, $\rho(r) = 0$ for $|r| > 1$, $\rho(r) = \rho(-r)$ for all $r \in R$ and $\int_{-\infty}^{\infty} \rho(r) \, dr = 1$. We define, for $\varepsilon > 0$,

$$(2.7) \quad \beta_i^\varepsilon(r) = \int_{-\infty}^{\infty} \beta_{i\varepsilon}(r - \varepsilon\theta) \rho(\theta) \, d\theta, \quad i = 1, 2, \quad r \in R$$

where

$$(2.8) \quad \beta_{i\varepsilon}(r) = \varepsilon^{-1}(r - (1 + \varepsilon\beta_i)^{-1}r), \quad i = 1, 2.$$

It should be recalled that β_i^ε are monotonically increasing infinitely differentiable functions. Moreover, β_i^ε are Lipschitzian with Lipschitz constant ε^{-1} , and in a certain sense which will be explained below they approximate β_i for $\varepsilon \rightarrow 0$.

For each $\varepsilon > 0$ consider the approximating system

$$(2.9) \quad \begin{aligned} y_t + Ay &= 0 && \text{in } Q, \\ \frac{\partial y}{\partial \nu} + \beta_i^\varepsilon(y) &= v_i && \text{in } \Sigma_i, \quad i = 1, 2, \\ y(\cdot, 0) &= y_0 && \text{in } \Omega. \end{aligned}$$

Let $\mathcal{A}_\varepsilon : H^1(\Omega) \rightarrow (H^1(\Omega))'$ be the operator defined by

$$(2.10) \quad (\mathcal{A}_\varepsilon y, \psi) = a(y, \psi) + \sum_{i=1}^2 \int_{\Gamma_i} \beta_i^\varepsilon(y) \psi \, d\sigma, \quad y, \psi \in H^1(\Omega),$$

and let $f \in L^2(0, T; (H^1(\Omega))')$ be given by

$$(2.11) \quad (f(t), \psi) = \sum_{i=1}^2 \int_{\Gamma_i} v_i \psi \, d\sigma, \quad \psi \in H^1(\Omega).$$

Then, in the sense of Definition 1 (see (2.6)), (2.9) can be written as

$$(2.12) \quad \begin{aligned} \frac{dy}{dt} + \mathcal{A}_\varepsilon y &= f, && t \in [0, T], \\ y(0) &= y_0. \end{aligned}$$

Since \mathcal{A}_ε is continuous, monotone, coercive and sublinear from $H^1(\Omega)$ to $(H^1(\Omega))'$, according to a standard existence result due to Lions (see, for instance, [4, p. 64]), (2.12) (and therefore (2.9)) has a unique solution $y_\varepsilon \in W(Q)$.

Let $j_i : R \rightarrow R$, $i = 1, 2$ be two convex and lower semicontinuous functions such that $\partial j_i = \beta_i$ (it is well known that such functions always exist).

PROPOSITION 1. Let $y_0 \in L^2(\Omega)$ and $v_i \in L^2(\Sigma_i)$ be given such that $j_i(y_0) \in L^1(\Omega)$, $i = 1, 2$. Then the system (2.1) has a unique solution $y \in W(Q)$. Furthermore, for $\varepsilon \rightarrow 0$,

$$(2.13) \quad \begin{aligned} y_\varepsilon &\rightarrow y \quad \text{strongly in } C([0, T]; L^2(\Omega)) \cap L^2(0, T; H^1(\Omega)) \\ &\text{and weakly in } W(Q). \end{aligned}$$

There exists $C > 0$ independent of v_i such that

$$(2.14) \quad \|y\|_{W(Q)} + \sum_{i=1}^2 \|\beta_i(y)\|_{L^2(\Sigma_i)} \leq C \left(\sum_{i=1}^2 \|v_i\|_{L^2(\Sigma_i)} + 1 \right).$$

(If β_i are multi-valued we mean by $\beta_i(y)$ the single-valued section w_i which occurs in (2.3).)

Proof. We take the inner product of (2.12) (where $y = y_\varepsilon$) with y_ε , and integrate over $[0, t]$. By (2.10) and (2.11) it follows that

$$(2.15) \quad \|y_\varepsilon(t)\|_{L^2(\Omega)} + \int_0^t \|y_\varepsilon(s)\|_{H^1(\Omega)}^2 ds \leq C(\|v_1\|_{L^2(\Sigma_1)}^2 + \|v_2\|_{L^2(\Sigma_2)}^2 + 1), \quad t \in [0, T],$$

where C is independent of ε .

Next, we take the inner product of (2.12) with $\beta_i^\varepsilon(y_\varepsilon)$. Inasmuch as $(\psi, \beta_i^\varepsilon(\psi)) \geq 0$, for all $\psi \in H^1(\Omega)$, we find, after some calculations,

$$(2.16) \quad \int_\Omega j_i^\varepsilon(y_\varepsilon) dx + \sum_{j=1}^2 \int_{\Sigma_j} (\beta_j^\varepsilon(y_\varepsilon) - v_j) \beta_i^\varepsilon(y_\varepsilon) d\sigma dt \leq \int_\Omega j_i^\varepsilon(y_0) dx \quad \text{for } i = 1, 2,$$

where

$$j_i^\varepsilon(r) = \int_0^r \beta_i^\varepsilon(s) ds, \quad i = 1, 2.$$

Along with assumption (2.2), (2.16) yields

$$\sum_{i=1}^2 \|\beta_i^\varepsilon(y_\varepsilon)\|_{L^2(\Sigma_i)}^2 \leq C \left(\sum_{i=1}^2 \|v_i\|_{L^2(\Sigma_i)}^2 + 1 \right),$$

and by (2.12) and (2.15) we see that

$$(2.17) \quad \|y_\varepsilon\|_{W(Q)}^2 + \sum_{i=1}^2 \|\beta_i^\varepsilon(y_\varepsilon)\|_{L^2(\Sigma_i)}^2 \leq C \left(1 + \sum_{i=1}^2 \|v_i\|_{L^2(\Sigma_i)}^2 \right)$$

where C is independent of ε .

Now, using (2.12) once again, for $\varepsilon, \lambda > 0$ we get

$$\begin{aligned} &\|y_\varepsilon(t) - y_\lambda(t)\|_{L^2(\Omega)}^2 + \|y_\varepsilon - y_\lambda\|_{L^2(0, T; H^1(\Omega))}^2 \\ &+ C \sum_{i=1}^2 \int_{\Sigma_i} (\beta_i^\varepsilon(y_\varepsilon) - \beta_i^\lambda(y_\lambda))(y_\varepsilon - y_\lambda) d\sigma dt \leq 0. \end{aligned}$$

If we take into account (2.7), (2.8), (2.17) and the monotonicity of β_i , the latter implies by a standard procedure that

$$(2.18) \quad \|y_\varepsilon - y_\lambda\|_{C([0, T]; L^2(\Omega))}^2 + \|y_\varepsilon - y_\lambda\|_{L^2(0, T; H^1(\Omega))}^2 \leq C(\varepsilon + \lambda).$$

Hence $y = \lim_{\varepsilon \rightarrow 0} y_\varepsilon$ exists in the strong topology of $L^2(0, T; H^1(\Omega)) \cap C([0, T]; L^2(\Omega))$. In particular, this implies that

$$y_\varepsilon \rightarrow y \quad \text{strongly in } L^2(0, T; H^{1/2}(\Gamma)) \subset L^2(\Sigma).$$

and by (2.17) we may assume that

$$(2.19) \quad \beta_i^\varepsilon(y_\varepsilon) \rightarrow w_i \quad \text{weakly in } L^2(\Sigma_i), \quad i = 1, 2.$$

According to Definition 1, to prove that y is a solution to (2.1) it suffices to show that

$$(2.20) \quad w_i \in \beta_i(y) \quad \text{a.e. on } \Sigma_i, \quad i = 1, 2.$$

For this purpose, we set

$$z_\varepsilon^i = \beta_{i\varepsilon}(y_\varepsilon - \varepsilon\theta).$$

By (2.7) and (2.19) it follows that on some subsequence $\varepsilon \rightarrow 0$ we have

$$(2.21) \quad z_\varepsilon^i \rightarrow z^i \quad \text{weakly in } L^2(\Sigma_i \times]-1, 1[), \quad i = 1, 2.$$

On the other hand, since $z_\varepsilon^i \in \beta_i((1 + \varepsilon\beta_i)^{-1}(y_\varepsilon - \varepsilon\theta))$ and by (2.19), $(1 + \varepsilon\beta_i)^{-1}(y_\varepsilon - \varepsilon\theta)$ is strongly convergent to y in $L^2(\Sigma_i \times]-1, 1[)$ we may infer that

$$z^i(\sigma, t, \theta) \in \beta_i(y(\sigma, t)) \quad \text{a.e. on } \Sigma_i \times]-1, 1[.$$

Along with (2.7) and (2.21) the latter implies (2.20) as claimed. The uniqueness of y is immediate from Definition 1. To obtain (2.13) and (2.14) we let λ tend to zero in (2.18) and $\varepsilon \rightarrow 0$ in (2.17).

Let us denote by $K : L^2(\Sigma_1) \times L^2(\Sigma_2) \rightarrow W(Q)$ the operator defined by $y = K(v_1, v_2)$, where y is the solution to (2.1). By K_ε we shall denote the corresponding operator associated with (2.9).

PROPOSITION 2. *Under the conditions of Proposition 1, the operator K is weakly continuous from $L^2(\Sigma_1) \times L^2(\Sigma_2)$ to $W(Q)$ and compact from $L^2(\Sigma_1) \times L^2(\Sigma_2)$ to $L^2(Q)$. Furthermore, if for $\varepsilon \rightarrow 0$ the sequence $\{(v_1^\varepsilon, v_2^\varepsilon)\}$ is weakly convergent in $L^2(\Sigma_1) \times L^2(\Sigma_2)$ to (v_1, v_2) , then on some subsequence, again denoted ε , one has*

$$(2.22) \quad K_\varepsilon(v_1^\varepsilon, v_2^\varepsilon) \rightarrow K(v_1, v_2) \quad \text{weakly in } W(Q) \text{ and strongly in } L^2(Q).$$

If $(v_1^\varepsilon, v_2^\varepsilon) \rightarrow (v_1, v_2)$ strongly in $L^2(\Sigma_1) \times L^2(\Sigma_2)$ then

$$(2.23) \quad K_\varepsilon(v_1^\varepsilon, v_2^\varepsilon) \rightarrow K(v_1, v_2) \quad \text{strongly in } C([0, T]; L^2(\Omega)) \cap L^2(0, T; H^1(\Omega)).$$

Proof. Let $\{(v_1^n, v_2^n)\}$ be a sequence of $L^2(\Sigma_1) \times L^2(\Sigma_1)$ weakly convergent to (v_1, v_2) . By estimate (2.14) it follows that $\{y_n = K(v_1^n, v_2^n)\}$ is weakly compact in $W(Q)$.

Hence, for some subsequence again denoted y_n , we have

$$(2.24) \quad y_n \rightarrow y \quad \text{weakly in } W(Q) \text{ and strongly in } L^2(Q).$$

As a matter of fact, since $\{y_n\}$ is bounded in $L^2(0, T; H^1(\Omega))$ and $\{dy_n/dt\}$ in $L^2(0, T; (H^1(\Omega))')$ according to a well-known compactness theorem, $\{y_n\}$ is a pre-compact subset of some $L^2(0, T; H^\delta(\Omega))$ where $\frac{1}{2} < \delta < 1$.

Thus by the trace theorem we may conclude that $\{y_n\}$ is precompact in $L^2(\Sigma)$. Hence without loss of generality we may assume that

$$(2.25) \quad y_n \rightarrow y \quad \text{strongly in } L^2(\Sigma).$$

Selecting a further subsequence we have by (2.14) that

$$(2.26) \quad \beta_i(y_n) \rightarrow w_i \quad \text{weakly in } L^2(\Sigma_i), \quad i = 1, 2.$$

Since β_i are maximal monotone it follows by (2.25) and (2.26) that $w_i \in \beta_i(y)$ a.e. on Σ_i , $i = 1, 2$. Along with (2.24) this implies that $y = K(v_1, v_2)$ as claimed.

Now let $\{(v_1^\varepsilon, v_2^\varepsilon)\}$ be such that for $\varepsilon \rightarrow 0$

$$(2.27) \quad v_i^\varepsilon \rightarrow v_i \quad \text{weakly in } L^2(\Sigma_i), \quad i = 1, 2.$$

Then in virtue of estimate (2.17) we may assume that

$$(2.28) \quad \begin{aligned} \tilde{y}_\varepsilon = K_\varepsilon(v_1^\varepsilon, v_2^\varepsilon) &\rightarrow z \quad \text{weakly in } W(Q) \\ &\text{and strongly in } L^2(0, T; H^\delta(\Omega)), \quad \frac{1}{2} < \delta < 1 \end{aligned}$$

and

$$(2.29) \quad \beta_i^\varepsilon(\tilde{y}_\varepsilon) \rightarrow \tilde{w}_i \quad \text{weakly in } L^2(\Sigma_i), \quad i = 1, 2.$$

Since the sequence of traces of $\{K_\varepsilon(v_1^\varepsilon, v_2^\varepsilon)\}$ converges strongly in $L^2(\Sigma)$ to the trace of z , arguing as in the proof of Proposition 1 we may infer by (2.29) that $\tilde{w}_i \in \beta_i(z)$ a.e. on Σ_i , $i = 1, 2$. Hence z is a solution to (2.1) corresponding to v_1, v_2 and therefore $z = K(v_1, v_2)$.

If $(v_1^\varepsilon, v_2^\varepsilon) \rightarrow (v_1, v_2)$ strongly in $L^2(\Sigma_1) \times L^2(\Sigma_2)$, then arguing as in the proof of Proposition 1 we deduce (2.23). This completes the proof of Proposition 2.

Remark. It must be emphasized that more general systems of the form

$$(2.30) \quad \begin{aligned} y_t + Ay &= F \quad \text{in } Q, \\ \frac{\partial y}{\partial \nu} + \beta_i(y) &\ni v_i^0 \quad \text{in } \Sigma_i, \quad i = 1, 2, \\ y(0) &= y_0 \quad \text{in } \Omega, \end{aligned}$$

where $F \in L^2(Q)$ and $v_i^0 \in L^2(\Sigma_i)$, can be put into the form (2.1), where $v_i = v_i^0 - \partial z / \partial \nu$ and $z \in H^{2,1}(Q)$ is the solution to

$$(2.31) \quad \begin{aligned} z_t + Az &= F \quad \text{in } Q, \\ z &= 0 \quad \text{in } \Sigma, \\ z(0) &= 0 \quad \text{in } \Omega. \end{aligned}$$

Since Γ_i are smooth parts of Γ and $\partial z / \partial \nu \in L^2(\Sigma)$, it follows that the restrictions of $\partial z / \partial \nu$ to Γ_i belong to $L^2(\Sigma_i)$ and therefore $v_i \in L^2(\Sigma_i)$, $i = 1, 2$. Thus Propositions 1 and 2 are applicable and therefore their conclusions remain true for general systems (2.30).

3. The main results. We shall study the following control problem:

Minimize

$$(3.1) \quad \frac{1}{2} \int_Q h(x, t) |y(x, t) - y_d(x, t)|^2 dx dt + \psi_1(u_1) + \psi_2(u_2) + \varphi(y(T))$$

on the class of all $u_i \in U_i$, $i = 1, 2$ and $y \in W(Q)$ subject to state system (1.1).

We shall assume that the following conditions are satisfied.

1° U_i , $i = 1, 2$ are real Hilbert spaces with norms $\|\cdot\|_i$ and inner products $\langle \cdot, \cdot \rangle_i$.

2° The functions $\psi_i : U_i \rightarrow \bar{R} =]-\infty, +\infty]$, $i = 1, 2$ are convex, lower semicontinuous and $\neq +\infty$.

3° The function $\varphi : L^2(\Omega) \rightarrow R$ is convex and continuous on $L^2(\Omega)$.

4° $h \in L^\infty(Q)$ and $y_d \in L^2(Q)$ are given; $h \geq 0$ a.e. on Q .

As regards the control systems (1.1) we shall assume that

5° A is the elliptic symmetric operator presented in § 2 and $\beta_i, i = 1, 2$, are two maximal monotone graphs in $R \times R$ which satisfy condition (2.2).

6° $B_i : U_i \rightarrow L^2(\Sigma_i), i = 1, 2$, are linear continuous operators.

7° $f_i \in L^2(\Sigma_i), i = 1, 2$, and $y_0 \in L^2(\Omega)$ satisfies the assumptions of Proposition 1.

The solution to (1.1) is meant in the sense of Definition 1, and, according to Proposition 1, under our assumptions for every pair $(u_1, u_2) \in U_1 \times U_2$ the control system (1.1) has a unique solution $y \in W(Q)$.

We shall say that the state $y^* \in W(Q)$ and the controls $u_i^* \in U_i, i = 1, 2$ are optimal in problem (3.1) if the infimum of the functional (3.1) is attained for $y = y^*$ and $u_i = u_i^*$.

The first optimality result is given in the case in which β_i are single-valued and satisfy the following condition:

8° The functions $\beta_i, i = 1, 2$ are monotonically increasing and locally Lipschitzian on the real axis R . Moreover, there exists $C > 0$ such that

$$(3.2) \quad \beta'_i(r) \leq C(|\beta_i(r)| + |r| + 1) \quad \text{a.e. } r \in R, \quad i = 1, 2.$$

THEOREM 1. *Let $y^* \in W(Q)$ and $(u_1^*, u_2^*) \in U_1 \times U_2$ be optimal in problem (3.1). Assume that conditions 1°–8° are satisfied. Then there exists $p \in C_w([0, T]; L^2(\Omega)) \cap L^2(0, T; H^1(\Omega))$ with $\partial p / \partial \nu \in L^1(\Sigma)$ which satisfies, along with y^* and u_1^*, u_2^* , the system*

$$(3.3) \quad p_t - Ap = h(y^* - y_d) \quad \text{in } Q,$$

$$(3.4) \quad \frac{\partial p}{\partial \nu} + \partial \beta_i(y^*)p \ni 0 \quad \text{in } \Sigma_i, \quad i = 1, 2,$$

$$(3.5) \quad p(T) + \partial \varphi(y^*(T)) \ni 0 \quad \text{in } L^2(\Omega),$$

$$(3.6) \quad B_i^* p_i \in \partial \psi_i(u_i^*), \quad i = 1, 2.$$

Here we have denoted by $B_i^* : L^2(\Sigma_i) \rightarrow U_i$ the adjoint of the operator B_i and by $p_i \in L^2(\Sigma_i)$ the restriction of p to Σ_i . We have also denoted by $\partial \psi_i$ and $\partial \varphi$ the subdifferentials of ψ_i, φ and by $\partial \beta_i$ the generalized gradient of β_i (see (1.8)). The boundary value problem (3.3)–(3.5) must be interpreted in the following weak sense:

$$(3.7) \quad \int_Q p \kappa_t dx dt + \int_0^T a(p, \kappa) dt + \int_{\Sigma_1} \mu_1 \kappa d\sigma dt + \int_{\Sigma_2} \mu_2 \kappa d\sigma dt + \int_Q h(y^* - y_d) dx dt - \int_\Omega \zeta \kappa(x, T) dx = 0$$

for all $\kappa \in W(Q)$ satisfying $\kappa(x, 0) = 0$, a.e. $x \in \Omega$. Here the functions $\mu_i \in L^2(\Sigma_i), i = 1, 2$ and $\zeta \in L^2(\Omega)$ satisfy the equations

$$(3.8) \quad \mu_i(\sigma, t) \in \partial \beta_i(y^*(\sigma, t)) \quad \text{a.e. } (\sigma, t) \in \Sigma_i, \quad i = 1, 2,$$

$$(3.9) \quad \zeta(x) + \partial \varphi(y^*(\cdot, T))(x) \ni 0 \quad \text{a.e. } x \in \Omega.$$

It should be emphasized that Theorem 1 covers most of the physical problems presented in the Introduction. For instance, in the case $\Gamma_1 = \Gamma$ and $\beta_1 = \beta$ given by (1.5)

(the thermostat-control problem), (3.4) becomes

$$\frac{\partial p}{\partial \nu} = \begin{cases} -\alpha_1 p & \text{if } y^* < \theta_1, \\ -[0, \alpha_1] p & \text{if } y^* = \theta_1, \\ 0 & \text{if } \theta_1 < y^* < \theta_2 \text{ in } \Sigma, \\ -[0, \alpha_2] p & \text{if } y^* = \theta_2, \\ -\alpha_2 p & \text{if } y^* > \theta_2. \end{cases}$$

Now we shall consider the particular case of problem (3.1) where $\Gamma_1 = \Gamma$ and $\beta_1 = \beta$ is given by (1.6). In this case (1.1) reduces to the unilateral problem (see, e.g., [7])

$$(3.10) \quad \begin{aligned} y_t + Ay &= 0 \quad \text{in } Q, \\ y \left(\frac{\partial y}{\partial \nu} - B_1 u_1 - f_1 \right) &= 0, \quad y \geq 0, \quad \frac{\partial y}{\partial \nu} - B_1 u_1 - f_1 \geq 0 \quad \text{in } \Sigma, \\ y(0) &= y_0. \end{aligned}$$

We shall assume that conditions $1^\circ-7^\circ$ are satisfied (for $i = 1$), and note that in virtue of 7° we must assume that $y_0(x) \geq 0$ a.e. $x \in \Omega$.

Under these assumptions we shall prove the following optimality theorem.

THEOREM 2. *Let $y^* \in W(Q)$ and $u_1^* \in U_1$ be optimal in problem (3.1) with state system (3.10). Then there exists $p \in C_w([0, T]; L^2(\Omega)) \cap L^2(0, T; H^1(\Omega))$ with $\partial p / \partial \nu \in M(\Sigma)$ which satisfies, along with y^* and u_1^* , the system*

$$(3.11) \quad p_t - Ap = h(y^* - y_d) \quad \text{in } Q,$$

$$(3.12) \quad \left(\frac{\partial p}{\partial \nu} \right)_a = 0 \quad \text{a.e. in } \{(\sigma, t) \in \Sigma; y^*(\sigma, t) > 0\},$$

$$(3.13) \quad p = 0 \quad \text{a.e. } \{(\sigma, t) \in \Sigma; y^*(\sigma, t) = 0\} \cap \left\{ (\sigma, t) \in \Sigma, B_1 u_1^* - \frac{\partial y^*}{\partial \nu} - f_1 > 0 \right\},$$

$$(3.14) \quad p(T) + \partial \varphi(y^*(T)) \ni 0 \quad \text{in } L^2(\Omega),$$

$$(3.15) \quad B_1(\gamma_0 p) \in \partial \psi_1(u_1^*).$$

Here $(\partial p / \partial \nu)_a$ denotes the absolutely continuous part of the measure $\partial p / \partial \nu \in M(\Sigma)$ and $M(\Sigma)$ in the space of all bounded Radon measures on Σ . In (3.15) we have denoted by $\gamma_0 p \in L^2(\Sigma)$ the trace of p at Σ .

Postponing the proofs of these theorems until §§ 5 and 6 we shall discuss now a particular case of Theorem 1. We shall consider the following special case of problem (3.1):

$$(3.16) \quad h = 0, \quad U_i = L^2(0, T; \Sigma_i), \quad B_i = I, \quad i = 1, 2,$$

$$(3.17) \quad \psi_i(u_i) = \int_{\Sigma_i} g_i(\sigma, u_i(\sigma, t)) d\sigma dt, \quad u_i \in U_i, \quad i = 1, 2,$$

where $g_i: \Gamma_1 \times R \rightarrow R$ is defined by

$$g_1(\sigma, r) = \begin{cases} 0 & \text{if } |r| \leq \rho, \\ +\infty & \text{otherwise,} \end{cases}$$

and $g_2: \Gamma_2 \times R \rightarrow R$ is a normal convex integrand on $\Gamma_2 \times R$ [11].

In other words, we consider the following control problem:

Minimize

$$(3.18) \quad \int_{\Sigma_2} g_2(\sigma, u_2(\sigma, t)) d\sigma dt + \varphi(y(T))$$

on all $y \in W(Q)$ and $(u_1, u_2) \in L^2(\Sigma_1) \times L^2(\Sigma_2)$, subject to (1.1) (where U_i and B_i satisfy (3.16)) and to the control constraint

$$(3.19) \quad |u_1(\sigma, t)| \leq \rho \quad \text{a.e. } (\sigma, t) \in \Sigma_1.$$

In addition to the assumptions of Theorem 1, we shall suppose that the boundary Γ and the coefficients of the operator A are analytic and

$$(3.20) \quad 0 \in \partial\varphi(y^*(T)).$$

COROLLARY 1. *Let y^* and u_1^*, u_2^* be optimal in problem (3.18). Then u_1^* is a bang-bang control on Σ_1 ; i.e.,*

$$(3.21) \quad |u_1^*(\sigma, t)| = \rho \quad \text{a.e. } (\sigma, t) \in \Sigma_1.$$

Proof. Since Theorem 1 is applicable in the present situation, it follows by (3.6) (for $i = 1$) that

$$(3.22) \quad u_1^*(\sigma, t) \in \rho \operatorname{sgn} p(\sigma, t) \quad \text{a.e. } (\sigma, t) \in \Sigma_1.$$

where $\operatorname{sgn} r = r/|r|$ for $r \neq 0$ and $\operatorname{sgn} 0 = [-1, 1]$.

Now let

$$\Sigma_0 = \{(\sigma, t) \in \Sigma_1; p(\sigma, t) = 0\}.$$

By (3.4) we see that $\partial p / \partial \nu = 0$ a.e. on Σ_0 . Then by a well-known result due to Mizohata it follows that $m(\Sigma_0) = 0$ unless $p \equiv 0$ (m denotes Lebesgue measure). Since by (3.5) and (3.20) $p \not\equiv 0$, we may infer that $p \neq 0$ a.e. on Σ_1 . Then by (3.22) it follows that (3.21) holds, thereby completing the proof.

4. The approximating control process. Let $y^* \in W(Q)$ and $(u_1^*, u_2^*) \in U_1 \times U_2$ be optimal in problem (3.1).

For $\varepsilon > 0$ consider the following optimal control problem:

Minimize

$$(4.1) \quad \frac{1}{2} \int_Q h|y - y_d|^2 dx dt + \sum_{i=1}^2 (\psi_{i\varepsilon}(u_i) + \frac{1}{2} \|u_i^* - u_i\|_i^2 + \varphi_\varepsilon(y(T)))$$

over all $y \in W(Q)$ and $u_i \in U_i$, $i = 1, 2$, subject to the state system

$$(4.2) \quad \begin{aligned} yt + Ay &= 0 && \text{in } Q, \\ \frac{\partial y}{\partial \nu} + \beta_i^\varepsilon(y) &= B_i u_i + f_i && \text{in } \Sigma_i, \quad i = 1, 2, \\ y(0) &= y_0. \end{aligned}$$

Here $\psi_{i\varepsilon} : U_i \rightarrow \mathbb{R}$, $i = 1, 2$ and $\varphi_\varepsilon : L^2(\Omega) \rightarrow \mathbb{R}$ are defined by (see, e.g., [4 p. 107])

$$(4.3) \quad \psi_{i\varepsilon}(u) = \inf \left\{ \frac{\|u - v\|_i^2}{2\varepsilon} + \psi_i(v); v \in U_i \right\}, \quad i = 1, 2,$$

$$(4.4) \quad \varphi_\varepsilon(y) = \inf \left\{ \frac{\|y - z\|_{L^2(\Omega)}^2}{2\varepsilon} + \varphi(z); z \in L^2(\Omega) \right\}.$$

Now let

$$\begin{aligned}
 F_\varepsilon(u_1, u_2) = & \frac{1}{2} \int_Q h |K_\varepsilon(B_1 u_1 + f_1, B_2 u_2 + f_2) - y_d|^2 dx dt \\
 & + \sum_{i=1}^2 (\psi_{i\varepsilon}(u_i) + \frac{1}{2} \|u_i^* - u_i\|_i^2) \\
 & + \varphi_\varepsilon(K_\varepsilon(B_1 u_1 + f_1, B_2 u_2 + f_2)(T))
 \end{aligned}
 \tag{4.5}$$

and

$$\begin{aligned}
 F(u_1, u_2) = & \frac{1}{2} \int_Q h |K(B_1 u_1 + f_1, B_2 u_2 + f_2) - y_d|^2 dx dt \\
 & + \sum_{i=1}^2 \psi_{i\varepsilon}(u_i) + \varphi(K(B_1 u_1 + f_1, B_2 u_2 + f_2)(T)),
 \end{aligned}
 \tag{4.6}$$

where $K_\varepsilon : L^2(\Sigma_1) \times L^2(\Sigma_2) \rightarrow W(Q)$ and $K : L^2(\Sigma_1) \times L^2(\Sigma_2) \rightarrow W(Q)$ have been defined in § 2.

In terms of F_ε , problem (4.1) may be written as

$$(4.1)' \quad \min \{F_\varepsilon(u_1, u_2); u_1 \in U_1, u_2 \in U_2\},$$

while by (3.1) we have

$$(4.7) \quad F(u_1^*, u_2^*) = \min \{F(u_1, u_2); u_1 \in U_1, u_2 \in U_2\}.$$

Since the functions $\psi_{i\varepsilon}$ and φ_ε are weakly lower semicontinuous and by Proposition 2 the operator K_ε is weakly continuous, we may infer that the functional F_ε is weakly lower semicontinuous on $U_1 \times U_2$. Hence problem (4.1) (equivalently (4.1)') has at least one solution $(y_\varepsilon, u_{1\varepsilon}, u_{2\varepsilon}) \in W(Q) \times U_1 \times U_2$. On the other hand, since the functions $\psi_{i\varepsilon}$, $i = 1, 2$ and φ_ε are Fréchet differentiable on U_i and $L^2(\Omega)$ respectively (see, e.g., [4, p. 107]), it follows by a standard device that there exists some function $p_\varepsilon \in W(Q)$ which, along with y_ε and u_i^* satisfies the following system (the Euler-Lagrange system associated with problem (4.1)):

$$\begin{aligned}
 (p_\varepsilon)_t - A p_\varepsilon &= h(y_\varepsilon - y_d) && \text{in } Q, \\
 \frac{\partial p_\varepsilon}{\partial \nu} + (\beta_i^\varepsilon)'(y_\varepsilon) p_\varepsilon &= 0 && \text{in } \Sigma_i, \quad i = 1, 2, \\
 p_\varepsilon(T) + \partial \varphi_\varepsilon(y_\varepsilon(T)) &= 0 && \text{in } \Omega,
 \end{aligned}
 \tag{4.8}$$

$$(4.9) \quad B_i^* p_{\varepsilon,i} = \partial \psi_{i\varepsilon}(u_{i\varepsilon}^*) + u_{i\varepsilon} - u_i^* \quad \text{in } \Sigma_i, \quad i = 1, 2,$$

where $p_{\varepsilon i}$ is the restriction of p_ε to $L^2(\Sigma_i)$. The solution p_ε is meant in the sense of Definition 1, and the symbol $(p_\varepsilon)_t \in L^2(0, T; (H^1(\Omega))')$ is used for the derivative of p_ε in the sense of $(H^1(\Omega))'$ -valued distributions on $]0, T[$.

LEMMA 1. For $\varepsilon \rightarrow 0$ one has

$$(4.10) \quad u_{i\varepsilon} \rightarrow u_i^* \quad \text{strongly in } U_i, \quad i = 1, 2,$$

$$\begin{aligned}
 (4.11) \quad y_\varepsilon &\rightarrow y^* \quad \text{weakly in } W(Q) \\
 &\text{and strongly in } L^2(0, T; H^1(\Omega)) \cap C([0, T]; L^2(\Omega)),
 \end{aligned}$$

$$(4.12) \quad \beta_i^\varepsilon(y_\varepsilon) \rightarrow f_i - B_i u_i^* - \frac{\partial y^*}{\partial \nu} \quad \text{weakly in } L^2(\Sigma_i), \quad i = 1, 2.$$

Proof. We have

$$F_\varepsilon(u_{1\varepsilon}, u_{2\varepsilon}) \leq \frac{1}{2} \int_Q h|z_\varepsilon - y_d|^2 dx dt + \sum_{i=1}^2 \psi_{i\varepsilon}(u_i^*) + \varphi_\varepsilon(z_\varepsilon(T)),$$

where $z_\varepsilon = K_\varepsilon(B_1 u_1^* + f_1, B_2 u_2^* + f_2)$.

According to Proposition 1, we have

$$(4.13) \quad z_\varepsilon \rightarrow y^* \quad \text{strongly in } C([0, T]; L^2(\Omega)).$$

Since $\psi_{i\varepsilon} \leq \psi_i$ and $\varphi_\varepsilon \leq \varphi$, it follows that

$$(4.14) \quad \limsup_{\varepsilon \rightarrow 0} F_\varepsilon(u_{1\varepsilon}, u_{2\varepsilon}) \leq F(u_1^*, u_2^*).$$

In particular, it follows that $\{u_{i\varepsilon}\}$ are bounded in U_i , $i = 1, 2$. Thus without loss of generality we may assume that

$$(4.15) \quad u_{i\varepsilon} \rightarrow \tilde{u}_i^* \quad \text{weakly in } L^2(U_i), \quad i = 1, 2.$$

Then, according to Proposition 2, we have

$$(4.16) \quad y_\varepsilon \rightarrow \tilde{y}^* = K(B_1 \tilde{u}_1^* + f_1, B_2 \tilde{u}_2^* + f_2) \quad \begin{array}{l} \text{weakly in } W(Q) \\ \text{and strongly in } L^2(Q). \end{array}$$

On the other hand, since the functions ψ_i and φ are weakly lower semicontinuous, and (see [4, p. 107])

$$\begin{aligned} \psi_{i\varepsilon}(u_i) &= \varepsilon \|\partial \psi_{i\varepsilon}(u_i)\|_i^2 + \psi_i((1 + \varepsilon \partial \psi_i)^{-1} u_i), \\ \varphi_\varepsilon(y) &= \varepsilon \|\partial \varphi_\varepsilon(y)\|_{L^2(\Omega)}^2 + \varphi((1 + \varepsilon \partial \varphi)^{-1} y), \end{aligned}$$

it follows by (4.15) and (4.16) that

$$\liminf_{\varepsilon \rightarrow 0} \psi_{i\varepsilon}(u_{i\varepsilon}) \geq \psi_i(\tilde{u}_i^*), \quad i = 1, 2,$$

$$\liminf_{\varepsilon \rightarrow 0} \varphi_\varepsilon(y_\varepsilon(T)) \geq \varphi(\tilde{y}^*(T)).$$

Along with (4.7) and (4.14) the latter imply (4.10). Finally (4.11) and (4.12) follow from Proposition 2, thereby completing the proof.

LEMMA 2. *There exists $C > 0$ independent of ε such that*

$$(4.17) \quad \|p_\varepsilon(t)\|_{L^2(\Omega)} + \|p_\varepsilon\|_{L^2(0,T;H^1(\Omega))} \leq C, \quad t \in [0, T],$$

$$(4.18) \quad \int_{\Sigma_i} |(\beta_i^\varepsilon)'(y_\varepsilon) p_\varepsilon| d\sigma dt \leq C, \quad i = 1, 2,$$

$$(4.19) \quad \|(p_\varepsilon)_t\|_{L^2(0,T;(H^1(\Omega))')} \leq C.$$

Proof. Without loss of generality we may assume that p_ε is a regular solution to (4.8), i.e., $p_\varepsilon \in H^{1,2}(Q)$. Then multiplying equation (4.8) by p_ε and integrating on $Q_t = \Omega \times]t, T[$, by the Green's formula, we have

$$(4.20) \quad \begin{aligned} & \frac{1}{2} \|p_\varepsilon(t)\|_{L^2(\Omega)}^2 + \int_t^T a(p_\varepsilon, p_\varepsilon) ds \\ & \leq \frac{1}{2} \|p_\varepsilon(T)\|_{L^2(\Omega)}^2 + \int_{Q_t} h|y_\varepsilon - y_d| |p_\varepsilon| dx dt, \quad t \in [0, T]. \end{aligned}$$

Let ζ be a C^1 approximation of the function sgn . We multiply (4.8) by $\zeta(p_\varepsilon)$ and integrate over Q . Using once again the Green's formula and letting ζ tend to sgn , we get

$$(4.21) \quad \sum_{i=1}^2 \int_{\Sigma_i} |(\beta_i^\varepsilon)'(y_\varepsilon)p_\varepsilon| d\sigma dt \leq \int_Q h|y_\varepsilon - y_d| dx dt + \int_\Omega |p_\varepsilon(x, T)| dx.$$

On the other hand, since $y_\varepsilon(T) \rightarrow y^*(T)$ in $L^2(\Omega)$ and

$$\|\partial\varphi_\varepsilon(y)\|_{L^2(\Omega)} \leq \inf \{\|w\|_{L^2(\Omega)}; w \in \partial\varphi(y)\}$$

it follows by (4.8) that $\{p_\varepsilon(T)\}$ is bounded in $L^2(\Omega)$. (Here we have also used the fact that $\partial\varphi$ is locally bounded on $L^2(\Omega)$.) Then by (4.20) and (4.21) we get (4.17). Next by (4.8) we get estimate (4.19), thereby completing the proof of Lemma 2.

It follows by (4.17) and (4.18) that $\{p_\varepsilon\}$ is precompact in $L^2(0, T; H^\delta(\Omega))$, where $\frac{1}{2} < \delta < 1$. Hence there exists $p \in L^2(0, T; H^1(\Omega))$ with $p_t \in L^2(0, T; (H^1(\Omega))')$ such that for some sequence $\varepsilon \rightarrow 0$ one has

$$(4.22) \quad \begin{aligned} p_\varepsilon &\rightarrow p \quad \text{weakly in } L^2(0, T; H^1(\Omega)), \\ &\quad \text{strongly in } L^2(0, T; H^\delta(\Omega)), \\ &\quad \text{weak star in } L^\infty(0, T; L^2(\Omega)), \end{aligned}$$

$$(4.23) \quad (p_\varepsilon)_t \rightarrow p_t \quad \text{weakly in } L^2(0, T; H^{-1}(\Omega)).$$

Here p_t denotes the derivative of $p(t)$ in the sense of $(H^1(\Omega))'$ -valued distributions on $]0, T[$. Then it follows that $p(t)$ is absolutely continuous from $[0, T]$ to $H^{-1}(\Omega)$, and by (4.22) we see that $p(t)$ is weakly continuous from $[0, T]$ to $L^2(\Omega)$, i.e., $p \in C_w([0, T]; L^2(\Omega))$.

In particular, we may infer that

$$(4.24) \quad p_\varepsilon(t) \rightarrow p(t) \quad \text{weakly in } L^2(\Omega) \quad \text{for every } t \in [0, T].$$

Since $\{\partial\varphi_\varepsilon(y_\varepsilon(T))\}$ is bounded in $L^2(\Omega)$ and $y_\varepsilon(T) \rightarrow y^*(T)$ strongly in $L^2(\Omega)$, it follows by (4.8) that

$$p(T) + \partial\varphi(y^*(T)) \ni 0 \quad \text{in } L^2(\Omega).$$

Next by (4.22) and the trace theorem it follows that

$$(4.25) \quad p_\varepsilon \rightarrow p \quad \text{strongly in } L^2(\Sigma),$$

which along with (4.9) and (4.10) implies

$$(4.26) \quad B_i^* p_i \in \partial\psi_i(u_i^*), \quad i = 1, 2,$$

where p_i is the restriction of p to Σ_i , $i = 1, 2$. Finally, it follows by (4.18) that there exist two bounded Radon measures $\mu_p^i \in M(\Sigma_i)$ on $\bar{\Sigma}_i$, $i = 1, 2$, such that

$$(4.27) \quad (\beta_i^\varepsilon)'(y_\varepsilon)p_\varepsilon \rightarrow \mu_p^i \quad \text{weak star in } M(\Sigma_i), \quad i = 1, 2.$$

Thus, letting ε tend to zero in (4.8), we see that p is a solution of

$$(4.28) \quad \begin{aligned} p_t - Ap &= h(y^* - y_d) \quad \text{in } Q, \\ \frac{\partial p}{\partial \nu} + \mu_p^i &= 0 \quad \text{in } \Sigma_i \quad i = 1, 2, \\ p(T) + \partial\varphi(y^*(T)) &\ni 0 \quad \text{in } \Omega. \end{aligned}$$

Equation (4.28) must be interpreted of course in the following sense (see (3.7)):

$$(4.29) \quad \int_Q p \kappa_i dx dt + \int_0^T a(p, \kappa) dt + \sum_{i=1}^2 \int_{\Sigma_i} \mu_p^i \kappa d\sigma \\ + \int_Q h(y^* - y_d) \kappa dx dt - \int_{\Omega} \zeta \kappa(x, T) dx = 0.$$

for all $\kappa \in C^1(\bar{Q})$ such that $\kappa(x, 0) = 0$, $x \in \Omega$. Here ζ is an element of $L^2(\Omega)$ satisfying (3.9).

Summarizing, we have proved the following intermediate result:

PROPOSITION 3. *Let y^* , u_1^* , u_2^* be optimal in problem (3.1). Then under assumptions 1°–7° there exists a function*

$$p \in C_w([0, T]; L^2(\Omega)) \cap L^2(0, T; H^1(\Omega)),$$

with $p_t \in L^2(0, T; (H^1(\Omega))')$, which satisfies the system (4.28) and (4.26). Moreover, p is the limit in the sense of (4.22), (4.23), (4.24), (4.25) and (4.27) of the sequence $\{p_\varepsilon\}$ of solutions to (4.8).

5. Proof of Theorem 1. We begin with a technical result concerning generalized gradients. Let β be a locally Lipschitzian function on the real axis and let β^ε be the function defined by (2.7), i.e.,

$$(5.1) \quad \beta^\varepsilon(r) = \int_{-\infty}^{\infty} \beta_\varepsilon(r - \varepsilon\theta) \rho(\theta) d\theta, \quad r \in R, \quad \varepsilon > 0,$$

where $\beta_\varepsilon = \varepsilon^{-1}(1 - (1 + \varepsilon\beta)^{-1})$.

By $\partial\beta$ we shall denote the generalized gradient of β (see (1.8)).

LEMMA 3. *Let E be a locally compact space and let ν be a positive measure on E such that $\nu(E) < \infty$. Let $\{y_\varepsilon\} \subset L^1(E)$ be a sequence such that, for $\varepsilon \rightarrow 0$,*

$$(5.2) \quad y_\varepsilon \rightarrow y \quad \text{strongly in } L^1(E),$$

$$(5.3) \quad (\beta^\varepsilon)'(y_\varepsilon) \rightarrow g \quad \text{weakly in } L^1(E).$$

Then

$$(5.4) \quad g(x) \in \partial\beta(y(x)) \quad \nu\text{-a.e. } x \in E.$$

Proof. By $L^1(E)$ we have denoted the space of all real-valued ν -measurable functions $y(x)$, defined ν -a.e. on E , such that $|y(x)|$ is ν -integrable over E .

Selecting a subsequence of $\{y_\varepsilon\}$ we may assume that

$$(5.5) \quad y_\varepsilon(x) \rightarrow y(x) \quad \nu\text{-a.e. } x \in E.$$

Next, by (5.3) and the Mazur theorem, it follows that

$$(5.6) \quad g = \lim_{m \rightarrow \infty} g_m \quad \text{strongly in } L^1(E),$$

where $\{g_m\} \subset L^1(E)$ are of the form

$$(5.7) \quad g_m = \sum_{j \in I_m} \alpha_m^j (\beta^{\varepsilon_j})'(y_{\varepsilon_j}).$$

Here I_m is a finite subset of natural numbers in the interval $[m, \infty[$ and $\alpha_m^j \geq 0$, $\sum_{j \in I_m} \alpha_m^j = 1$. According to (5.6) we may also assume without loss of generality that

$$(5.8) \quad g_m(x) \rightarrow g(x) \quad \nu\text{-a.e. } x \in E.$$

We fix $x \in E$ such that (5.5) and (5.8) hold, and consider a sequence $\{z_n\}$ of real numbers such that $\beta'(z_n)$ exist and $z_n \rightarrow y(x)$ for $n \rightarrow \infty$. We set $y_j = y_{\varepsilon_j}(x)$ and notice that by (5.1) we have

$$(5.9) \quad (\beta^{\varepsilon_j})'(y_j) = \varepsilon_j^{-1} \int_{-\infty}^{\infty} \beta_{\varepsilon_j}(y_j - \varepsilon_j \theta) \rho'(\theta) d\theta.$$

On the other hand, we have

$$\begin{aligned} \beta(z_j) &= \beta((1 + \varepsilon_j \beta)^{-1}(y_j - \varepsilon_j \theta)) + \beta'(z_j)(z_j - (1 + \varepsilon_j \beta)^{-1}(y_j - \varepsilon_j \theta)) \\ &\quad + \omega_j(\theta)(z_j - (1 + \varepsilon_j \beta)^{-1}(y_j - \varepsilon_j \theta)), \end{aligned}$$

where $\omega_j(\theta) \rightarrow 0$ for $\delta_j = z_j - (1 + \varepsilon_j \beta)^{-1}(y_j - \varepsilon_j \theta) \rightarrow 0$. Along with (5.9), the latter yields

$$(5.10) \quad \begin{aligned} (\beta^{\varepsilon_j})'(y_j) &= \beta'(z_j) - \beta'(z_j) \int_{-\infty}^{\infty} \beta_{\varepsilon_j}(y_j - \varepsilon_j \theta) \rho'(\theta) d\theta \\ &\quad - \varepsilon_j^{-1} \int_{-\infty}^{\infty} \omega_j(\theta)(z_j - (1 + \varepsilon_j \beta)^{-1}(y_j - \varepsilon_j \theta)) \rho'(\theta) d\theta. \end{aligned}$$

Since β is locally Lipschitzian, it follows by (5.5) that $\beta_{\varepsilon_j}(y_j - \varepsilon_j \theta) \rightarrow \beta(y(x))$ uniformly in θ on $[-1, 1]$.

On the other hand, z_j can be chosen sufficiently close to y_j in such a way that

$$\frac{|y_j - z_j|}{\varepsilon_j} \rightarrow 0 \quad \text{for } j \rightarrow \infty.$$

Thus $\delta_j \rightarrow 0$ for $j \rightarrow \infty$, and (5.10) yields

$$|(\beta^{\varepsilon_j})'(y_j) - \beta'(z_j)| \rightarrow 0 \quad \text{for } j \rightarrow \infty.$$

Along with (5.7) and definition of $\partial\beta$, the latter yields $g(x) \in \partial\beta(y(x))$ as claimed.

Now we continue the proof of Theorem 1 by observing that, by condition (3.2), after some calculations involving (5.1), we have the estimate

$$(5.11) \quad (\beta^{\varepsilon_i})'(y) \leq C(|\beta^{\varepsilon_i}(y)| + |y| + 1), \quad i = 1, 2, \quad y \in \mathbb{R},$$

where $C > 0$ is independent of ε .

For each $\varepsilon > 0$ and natural number n , we set

$$E_n^{\varepsilon} = \{(\sigma, t) \in \Sigma; |y_{\varepsilon}(\sigma, t)| \leq n\},$$

where y_{ε} are defined as in § 4.

Let $\{y_{\varepsilon}\} \subset W(Q)$ and $\{p_{\varepsilon}\} \subset L^2(0, T; H^1(\Omega))$ be the sequences defined in § 4.

Since, as proved in Lemma 1, $\{\beta^{\varepsilon_i}(y_{\varepsilon})\}$ is weakly convergent in $L^2(\Sigma_i)$, $i = 1, 2$ and $\{p_{\varepsilon}\}$ is strongly convergent in $L^2(\Sigma)$, we may infer that the sequence $\{p_{\varepsilon} \beta^{\varepsilon_i}(y_{\varepsilon})\}$ is weakly convergent in $L^2(\Sigma_i)$ for $\varepsilon \rightarrow 0$. Inasmuch as, by (5.11),

$$(5.12) \quad |(\beta^{\varepsilon_i})'(y_{\varepsilon}) p_{\varepsilon}| \leq C(|\beta^{\varepsilon_i}(y_{\varepsilon})| + |y_{\varepsilon}| + |p_{\varepsilon}|) \quad \text{a.e. on } \Sigma_i, \quad i = 1, 2$$

we may infer that the family $\{(\beta^{\varepsilon_i})'(y_{\varepsilon}) p_{\varepsilon}\}$ is equicontinuous and bounded in $L^1(\Sigma_i)$, $i = 1, 2$. Thus by the Dunford-Pettis criterion $\{(\beta^{\varepsilon_i})'(y_{\varepsilon}) p_{\varepsilon}\}$ is weakly compact in $L^1(\Sigma_i)$.

Then by (4.27) we see that $\mu_p^i \in L^2(\Sigma_i)$ and

$$(5.13) \quad (\beta^{\varepsilon_i})'(y_{\varepsilon}) p_{\varepsilon} \rightarrow \mu_p^i \quad \text{weakly in } L^1(\Sigma_i).$$

On the other hand, it follows by (4.11) that $\{y_{\varepsilon}\}$ is strongly convergent in $L^2(0, T; H^{1/2}(\Gamma))$ for $\varepsilon \rightarrow 0$.

Thus, selecting a subsequence if necessary, we have

$$(5.14) \quad y_\varepsilon(x, t) \rightarrow y^*(x, t) \quad \text{a.e. } (x, t) \in \Sigma_i, \quad i = 1, 2$$

and, by Egorov's theorem, for each $\eta > 0$ there exists a measurable subset $E_\eta^i \subset \Sigma_i$ such that $m(\Sigma_i \setminus E_\eta^i) \leq \eta$, $\{y_\varepsilon\}$ is bounded on E_η^i and

$$(5.15) \quad y_\varepsilon(x, t) \rightarrow y^*(x, t) \quad \text{uniformly on } E_\eta^i, \quad i = 1, 2.$$

Next, since $\{(\beta_i^\varepsilon)'(y_\varepsilon)\}$ are uniformly bounded on E_η^i , we may assume (extracting a further subsequence) that

$$(5.16) \quad (\beta_i^\varepsilon)'(y_\varepsilon) \rightarrow g_i \quad \text{weakly in } L^1(E_\eta^i)$$

(actually weak star in $L^\infty(E_\eta^i)$). Then by Lemma 3 it follows that

$$g_i(x, t) \in \partial\beta_i(y^*(x, t)) \quad \text{a.e. } (x, t) \in E_\eta^i, \quad i = 1, 2.$$

Now by (4.22) and Egorov's theorem we may assume that $p_\varepsilon \rightarrow p$ uniformly on E_η^i . Along with (5.15) and (5.16) the latter implies that $\mu_p^i = g_i p$ on E_η^i . Hence

$$\mu_p^i(x, t) \in p(x, t) \partial\beta_i(y^*(x, t)) \quad \text{a.e. } (x, t) \in E_\eta^i.$$

Since $m(\Sigma_i \setminus E_\eta^i)$ can be made arbitrarily small, we conclude that

$$(5.17) \quad \mu_p^i(x, t) \in p(x, t) \partial\beta_i(y^*(x, t)) \quad \text{a.e. } (x, t) \in \Sigma_i, \quad i = 1, 2.$$

Thus the conclusions of Theorem 1 follows by Proposition 3.

Proof of Theorem 2. If β is the graph defined by (1.6) then $\beta_\varepsilon(r) = -\varepsilon^{-1}r^- = \varepsilon^{-1} \inf(r, 0)$ for $r \in R$, and

$$\beta^\varepsilon(r) = \varepsilon^{-1} \int_{\varepsilon^{-1}r}^{\infty} (r - \varepsilon\theta) \rho(\theta) d\theta, \quad r \in R,$$

respectively (we set $\dot{\beta}^\varepsilon = (\beta^\varepsilon)'$),

$$\dot{\beta}^\varepsilon(r) = \varepsilon^{-1} \int_{\varepsilon^{-1}r}^{\infty} \rho(\theta) d\theta.$$

Hence

$$\begin{aligned} |y_\varepsilon \dot{\beta}^\varepsilon(y_\varepsilon) p_\varepsilon - p_\varepsilon \beta^\varepsilon(y_\varepsilon)| &= \left| p_\varepsilon \int_{\varepsilon^{-1}y_\varepsilon}^{\infty} \rho(\theta) d\theta \right| \\ &\leq \varepsilon |\dot{\beta}^\varepsilon(y_\varepsilon) p_\varepsilon|. \end{aligned}$$

On the other hand, arguing as in [1], [2] we find that

$$(6.2) \quad |p_\varepsilon \beta^\varepsilon(y_\varepsilon)| \leq 2\varepsilon |\dot{\beta}^\varepsilon(y_\varepsilon) p_\varepsilon| (\zeta_\varepsilon + \varepsilon^{-1} |y_\varepsilon| \gamma_\varepsilon) \quad \text{a.e. on } \Sigma,$$

where

$$\zeta_\varepsilon(\sigma, t) = \begin{cases} 0 & \text{if } |y_\varepsilon(\sigma, t)| > \varepsilon, \\ 1 & \text{if } |y_\varepsilon(\sigma, t)| \leq \varepsilon, \end{cases}$$

and

$$\gamma_\varepsilon(\sigma, t) = \begin{cases} 0 & \text{if } y_\varepsilon(\sigma, t) > -\varepsilon, \\ 1 & \text{if } y_\varepsilon(\sigma, t) \leq -\varepsilon. \end{cases}$$

Inasmuch as by (4.12), $\{\beta^\varepsilon(y_\varepsilon)\}$ is bounded in $L^2(\Sigma)$, and by Lemma 2 $\{\dot{\beta}^\varepsilon(y_\varepsilon) p_\varepsilon\}$ is

bounded in $L^1(\Sigma)$, we see by (6.2) that on some subsequence $\varepsilon \rightarrow 0$ we have

$$(6.3) \quad p_\varepsilon \beta^\varepsilon(y_\varepsilon) \rightarrow 0 \quad \text{a.e. on } \Sigma.$$

On the other hand, we know that $\beta^\varepsilon(y_\varepsilon) \rightarrow B_1 u_1^* - f_1 - \partial y^*/\partial \nu$ weakly in $L^2(\Sigma)$ and $p_\varepsilon \rightarrow p$ strongly in $L^2(\Sigma)$. This implies that the sequence $\{p_\varepsilon \beta^\varepsilon(y_\varepsilon)\}$ is weakly convergent in $L^1(\Sigma)$ to $p(B_1 u_1^* - \partial y^*/\partial \nu - f_1)$. Now by (6.3) it follows that

$$(6.4) \quad p \left(B_1 u_1^* - \frac{\partial y^*}{\partial \nu} - f_1 \right) = 0 \quad \text{a.e. on } \Sigma,$$

and therefore

$$p_\varepsilon \beta^\varepsilon(y_\varepsilon) \rightarrow 0 \quad \text{strongly in } L^1(\Sigma).$$

Then by (6.1) we see that

$$(6.5) \quad y_\varepsilon \dot{\beta}^\varepsilon(y_\varepsilon) p_\varepsilon \rightarrow 0 \quad \text{strongly in } L^1(\Sigma).$$

Next, by Egorov's theorem, for each $\eta > 0$ there exists E_η a measurable subset of Σ such that $m(\Sigma \setminus E_\eta) \leq \eta$, $y_\varepsilon \rightarrow y^*$ uniformly on E_η and y^* is continuous on E_η . Along with (6.5) the latter yields

$$(6.6) \quad \lim_{\varepsilon \rightarrow 0} y^* \dot{\beta}^\varepsilon(y_\varepsilon) p_\varepsilon = 0 \quad \text{strongly in } L^1(E_\eta).$$

Denote by $E_{\eta,\delta}$ the following subset of Σ :

$$E_{\eta,\delta} = \{(\sigma, t) \in E_\eta; |y^*(\sigma, t)| \geq \delta\}.$$

Next, by Proposition 3, it follows that there exists a measure $\mu \in M(\Sigma)$ such that (see (4.27))

$$(6.7) \quad p_\varepsilon \dot{\beta}^\varepsilon(y_\varepsilon) \rightarrow \mu \quad \text{weak star in } M(\Sigma).$$

Let $\mu = (\mu)_a + (\mu)_s$ be the Lebesgue decomposition of μ into the absolutely continuous part $(\mu)_a$ and the singular part $(\mu)_s$. By (6.6) and (6.7) we see that $\mu = 0$ on $E_{\eta,\delta}$. By the definition of singular part we deduce that the support of $(\mu)_s$ is concentrated in $E_\eta \cap \{(\sigma, t) \in \Sigma; y^*(\sigma, t) > 0\}$. Since $m(\Sigma \setminus E_\eta) \rightarrow 0$ for $\eta \rightarrow 0$, we may conclude that

$$(\mu)_a = 0 \quad \text{on } \{(\sigma, t); y^*(\sigma, t) > 0\}.$$

Along with (4.26), (4.28) and (6.4), the latter completes the proof of Theorem 2.

Remark. Let us consider problem (3.1) with the state system

$$(6.8) \quad \begin{aligned} y_t + Ay &= 0 \quad \text{in } Q, \\ y \left(\frac{\partial y}{\partial \nu} + \beta_0(y) - B_1 u_1 - f_1 \right) &= 0, \quad y \geq 0, \\ \frac{\partial y}{\partial \nu} + \beta_0(y) - B_1 u_1 - f_1 &\geq 0 \quad \text{in } \Sigma, \\ y(0) &= y_0 \end{aligned}$$

where $u_1 \in U_1$ and $A, B_1: U_1 \rightarrow L^2(\Sigma)$, f_1, y_0 satisfy conditions 1°–7°. Here β_0 is a differentiable, monotonically increasing Lipschitzian function on R . System (6.8) can be

written in the form (1.1) where $\Gamma_1 = \Gamma$ and

$$(6.9) \quad \beta_1(r) = \begin{cases} \beta_0(r) & \text{if } r > 0, \\]-\infty, 0] & \text{if } r = 0, \\ \emptyset & \text{if } r < 0. \end{cases}$$

The prototype of this problem is the enzyme diffusion problem (1.4). In order to obtain necessary conditions for optimality in this case, it is more convenient to replace the approximating system (4.2) by

$$(6.10) \quad \begin{aligned} y_t + Ay &= 0 && \text{in } Q, \\ \frac{\partial y}{\partial \nu} + \beta_0(y) + \beta^\varepsilon(y) &= B_1 u_1 + f_1 && \text{in } \Sigma, \\ y(0) &= y_0. \end{aligned}$$

Obviously, Lemmas 1, 2 as well as Proposition 3 remain valid and p_ε is in the case the solution to

$$(6.11) \quad \begin{aligned} (p_\varepsilon)_t - Ap_\varepsilon &= h(y_\varepsilon - y_d) && \text{in } Q, \\ \frac{\partial p_\varepsilon}{\partial \nu} + \beta'_0(y_\varepsilon)p_\varepsilon + (\beta^\varepsilon)'(y_\varepsilon)p_\varepsilon &= 0 && \text{in } \Sigma, \\ p_\varepsilon(T) + \partial\varphi_\varepsilon(y_\varepsilon(T)) &= 0. \end{aligned}$$

Then passing to the limit we have from the preceding proof that p is the solution to

$$(6.12) \quad \begin{aligned} p_t - Ap &= h(y^* - y_d) && \text{in } Q, \\ \left(\frac{\partial p}{\partial \nu}\right)_a + \beta'_0(y^*)p &= 0 && \text{a.e. in } \{y^* > 0\}, \\ p &= 0 && \text{on } \{y^* = 0\} \cap \left\{B_1 u_1^* - \frac{\partial y^*}{\partial \nu} - \beta'_0(y^*)p - f_1 > 0\right\}. \end{aligned}$$

with (3.14) and (3.15) still satisfied.

REFERENCES

- [1] V. BARBU, *Necessary conditions for nonconvex distributed control problems governed by elliptic variational inequalities*, J. Math. Anal. Appl., 80 (1981), pp. 566–597.
- [2] ———, *Necessary conditions for distributed control problems governed by parabolic variational inequalities*, SIAM J. Control Optim., 19 (1981), pp. 64–86.
- [3] ———, *Necessary conditions for boundary control problems governed by parabolic variational inequalities*, An. St. Univ. Al. I. Cuza, T XXVI fasc. 1 (1980), pp. 47–66.
- [4] V. BARBU AND TH. PRECUPANU, *Convexity and Optimization in Banach Spaces*, Sijthoff & Noordhoff–Publishing House of Romanian Academy, 1978.
- [5] H. BRÉZIS, *Opérateurs maximaux monotones et semigroupes de contractions dans les espaces Hilbert*, Math. Studies 5, North-Holland, Amsterdam, 1973.
- [6] H. F. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.
- [7] G. DUVAUT AND J. L. LIONS, *Inequalities in Mechanics and Physics*, Springer-Verlag, Berlin–Heidelberg–New York, 1976.
- [8] I. EKELAND AND R. TEMAM, *Analyse convexe et problèmes variationnelles*, Dunod, Gauthier-Villars, Paris, 1974.
- [9] B. KAWOHL, *On nonlinear parabolic equations with abruptly changing nonlinear boundary conditions*, preprint, Technische Hochschule Darmstadt, June, 1980.
- [10] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, vol. II, Springer-Verlag, Berlin–Heidelberg–New York, 1980.

- [11] R. T. ROCKAFELLAR, *Integral functionals, normal integrands and measurable selections, nonlinear operators and measurable selections*, in *Nonlinear Operators and the Calculus of Variations*, J. P. Gossez, ed., *Lecture Notes in Mathematics* 543, Springer-Verlag, New York, 1976.
- [12] ———, *La théorie des sousgradients et ses applications à l'optimization*, Les Presses de l'Université de Montréal, 1978.
- [13] E. SACHS, *A parabolic control problem with a boundary condition of the Stefan-Boltzman type*, *ZAMM* 58 (1978), pp. 443–450.
- [14] CH. SAGUEZ, *Conditions nécessaires d'optimalité pour des problèmes de contrôle optimale associées à des inéquations variationnelles* (to appear).

THE CIRCLE CRITERION AND THE L^p STABILITY OF FEEDBACK SYSTEMS*

S. MOSSAHEB†

Abstract. It is shown that under certain mild assumptions the circle criterion of Sandberg and Zames implies the L^p stability of a broad class of linear time-varying feedback systems for $1 \leq p \leq \infty$. Moreover, it is proved that if the input to the feedback configuration tends to zero then so does the output. The results are extended to the case of certain open-loop unstable systems as well as sector bounded nonlinear feedback gains. Finally it is pointed out that all the results apply to any system whose open-loop transmittance is given by a finite dimensional linear time-invariant element with a strictly proper rational transfer function.

1. Introduction. Nonlinear equations of the form

$$(1) \quad x(t) = f(t) - \int_0^t g(t-s)n(s, x(s)) ds$$

arise in the study of many control systems. For example, the input-output equation defining the feedback connection of a linear time-invariant system with impulse response g and a time-varying nonlinearity with characteristic $n(t, x(t))$ may be written as

$$(2) \quad x(t) = (g * u)(t) - \int_0^t g(t-s)n(s, x(s)) ds,$$

which is a particular form of (1).

A celebrated L^2 stability theorem for (1) is the circle theorem of Sandberg and Zames [7], [11], [12]. Using the technique of exponential weighting Zames has obtained an L^∞ stability result in the case when $e^{at}g(t) \in L^1 \cap L^2$ for some $a > 0$ and has given a criterion based on the shifted Nyquist diagram of g [13]. On the other hand, Sandberg has shown that if $(1+t)^2g(t) \in L^1 \cap L^2$ then the circle criterion in itself gives the L^∞ stability of (1) and the shifted Nyquist diagram of g need not be used. He has also shown that under the same conditions if f is bounded and tends to zero then so does any solution of (1) [8], [9].

Our aim is to prove some L^p stability theorems for (1) under somewhat weaker conditions. Using the notion of the resolvent kernel of Volterra integral equations, we prove that if in (1) $n(s, x) = k(s)x$, i.e., the case of linear time-varying feedback gains, then the assumption $(1+t)g(t) \in L^1 \cap L^2$ together with the circle condition implies L^p stability of (1) for all p , $1 \leq p \leq \infty$. Moreover, if $p \geq 2$ and $f \in L^p$ tends to zero then so does x . We then use this result to prove that under the same assumption on g the satisfaction of the circle criterion implies the L^p stability of the nonlinear equation (1) for all $p \geq 2$ and of (2) for all $1 \leq p \leq \infty$. If in addition $p \geq 2$ and $f \in L^p$ tends to zero, then so does any solution x of (1). The above results will be extended to accommodate a class of open-loop unstable systems which includes all finite dimensional systems with strictly proper transfer functions.

2. Notation and background material.

2.1. The Laplace transform of a function f on $[0, \infty)$ is denoted by \hat{f} . We shall write A_1 for the set of functions g on $[0, \infty)$ such that $(1+t)g(t) \in L^1 \cap L^2$. Similarly,

* Received by the editors September 10, 1980, and in revised form March 24, 1981.

† Postgraduate School of Control Engineering, University of Bradford, Bradford, BD7 1DP, England.

A_2 is the set of those functions g on $[0, \infty)$ such that $g = g_1 + g_2$, where $g_1 \in A_1$ and g_2 is such that its Laplace transform is a strictly proper rational function. We write N for the set of functions $n(t, x)$ defined and measurable on $\mathbb{R}_+ \times \mathbb{R}$ such that n is continuous in x for almost all t . $N(a, b)$ is the subset of N consisting of those n such that $a \leq x^{-1}n(t, x) \leq b$ for $x \neq 0$ and almost all t .

2.2. For ease of reference we state the following version of the circle criterion [6, p. 302]. Let $g \in L^1(0, \infty)$ and $n \in N(a, b)$. Let

$$(3) \quad \gamma = \frac{b-a}{2} \sup_w \left| \hat{g}(jw) \left(1 + \frac{b+a}{2} \hat{g}(jw) \right)^{-1} \right|.$$

If $\gamma < 1$, then whenever f is locally square integrable so is any solution x of (1) and for any $T > 0$ $\|x\| \leq c(\gamma)\|f\|$, where $c(\gamma)$ depends only on γ and $\|\cdot\|$ is the norm in $L^2(0, T)$. In particular, if $f \in L^2(0, \infty)$ then any solution x of (1) is in L^2 and $\|x\|_2 \leq c(\gamma)\|f\|_2$.

2.3. Resolvent kernels. For a concise account of the following results the reader is referred to [6, Ch. iv]. Let $a(t, s)$ be a locally integrable function on $\mathbb{R}_+ \times \mathbb{R}_+$ such that $a(t, s) = 0$ whenever $s > t$. The formal resolvent equation associated with $a(t, s)$ is

$$(4) \quad r(t, s) = -a(t, s) + \int_s^t r(t, u)a(u, s) du.$$

It is easily seen that $r(t, s) = 0$ if $s > t$. It may be proved that, when $a(t, s) = g(t-s)k(s)$ with $g \in L^1(0, \infty)$ and $k \in L^\infty(0, \infty)$, the above equation has a unique solution $r(t, s)$ which is locally integrable and also satisfies

$$(5) \quad r(t, s) = -a(t, s) + \int_s^t a(t, u)r(u, s) du.$$

The function r above is called the resolvent kernel of $a(t, s)$. If $a(t, s)$ has a unique locally integrable resolvent $r(t, s)$, then the equation

$$x(t) = f(t) + \int_0^t a(t, s)x(s) ds$$

has a unique solution in x which is given by

$$(6) \quad x(t) = f(t) - \int_0^t r(t, s)f(s) ds.$$

The above property of the resolvent is reminiscent of that of the state transition matrix $R(t, s)$ of the differential equation $\dot{x} = A(t)x$ in the sense that the solution of $\dot{x} = A(t)x + f(t)$, $x(t_0) = x_0$ is

$$x(t) = R(t, t_0)x_0 + \int_{t_0}^t R(t, s)f(s) ds$$

(see, e.g., [1, Thm. 1, p. 40]).

2.4. We shall need the following result, known as Young's inequality [10, p. 178]. Let p, q and r be not less than 1 and $1/q = 1/p + 1/r - 1$. Then if $g \in L^r$ and $f \in L^p$ then $g * f \in L^q$ and $\|g * f\|_q \leq \|g\|_r \|f\|_p$. In particular, if g is integrable then $g * f \in L^p$ for all $1 \leq p \leq \infty$ and $f \in L^p$ and $\|g * f\|_p \leq \|g\|_1 \|f\|_p$.

2.5 [5, p. 141]. Let $\phi(t)$ be a real-valued measurable function on $(0, \infty)$ such that $\phi(0) = 1$ and $0 \leq \phi(t_1 + t_2) \leq \phi(t_1)\phi(t_2)$. Let $L(\phi) = \{f: f\phi \in L^1\}$. Then, with convolution as multiplication and under the norm $\|f\| = \int_0^\infty |f(t)|\phi(t) dt$, $L(\phi)$ is a commutative Banach algebra. Let $A(\phi)$ be the algebra obtained from $L(\phi)$ by joining a unit to it. Every element of $A(\phi)$ may be written in the form $c\delta + f$, where δ is the Dirac unit mass at the origin, c is a real number and $f \in L(\phi)$. Let $w_0 = \lim_{t \rightarrow \infty} t^{-1} \log \phi(t)$. Then w_0 exists, and if $h \in A(\phi)$ a necessary and sufficient condition for the existence of an inverse in $A(\phi)$ for h is that $\inf \{|\hat{h}(z)|: \operatorname{Re} z \geq w_0\} > 0$.

3. L^p stability of linear time-varying systems. Consider (1) with the following assumptions.

(H1) $n(s, x) = k(s)x$, where $k \in L^\infty(0, \infty)$ and there exist two constants a and b such that $a \leq k(s) \leq b$ for almost all s .

(H2) $g \in A_1$ and with γ as in (3) we have $\gamma < 1$. Note that in view of (H2) the conditions of the circle criterion are satisfied. We have the following theorem.

THEOREM 1. *If in (1) (H1) and (H2) hold, then for any p , $1 \leq p \leq \infty$ the linear map $f \rightarrow x$ is bounded on L^p so that there exists a constant c_p independent of f and x such that $\|x\|_p \leq c_p \|f\|_p$.*

Proof. Let $a(t, s) = -g(t-s)k(s)$ and let $r(t, s)$ be the resolvent kernel of $a(t, s)$. By § 2.3, $r(t, s)$ exists as a unique locally integrable function on $\mathbb{R}_+ \times \mathbb{R}_+$ and (1) has a unique solution given by (6). It follows, therefore, that the map $f \rightarrow x$ is a well-defined linear map, and it suffices to show that the map $(Tf)(t) = \int_0^t r(t, s)f(s) ds$ is bounded on L^p . Now, as is well known [2, pp. 109–112], T is bounded on L^1 if and only if $c_1 = \sup_s \int_0^\infty |r(t, s)| dt < \infty$, and T is bounded on L^∞ if and only if $c_\infty = \sup_t \int_0^\infty |r(t, s)| ds < \infty$. Suppose for the moment that both c_1 and c_∞ are finite. By a well-known interpolation theorem of Marcinkiewicz [10, p. 183], it follows that T is bounded on L^p for all p , $1 \leq p \leq \infty$ and the result follows (cf. [2, pp. 109–112]). The proof of the finiteness of c_1 and c_∞ is given in the following two steps.

Step 1. $c_1 < \infty$. In (5) let $a(t, s) = -g(t-s)k(s)$ and $t = s + v$. Since $g(v)$ and $k(s)$ vanish for $v < 0$ and $s < 0$ respectively, we have $r(s+v, s) = 0$ for $v < 0$ or $s < 0$. Thus, without loss of generality we may assume that $v \geq 0$ and $s \geq 0$. A simple change of variable in (5) gives

$$r(s+v, s) = g(v)k(s) - \int_0^v g(v-u)k(s+u)r(s+u, s) du.$$

Fix $s \geq 0$ and let $R(v) = r(s+v, s)$, $k_0 = k(s)$ and $K(u) = k(s+u)$. From the above equation we have

$$(7) \quad R(v) = k_0 g(v) - \int_0^v g(v-u)K(u)R(u) du.$$

From (H1), (H2) and § 2.2 we have $R \in L^2(0, \infty)$ and $\|R\|_2 \leq c(\gamma)\|k_0\|\|g\|_2$, where $c(\gamma)$ is independent of s . Let $S(v) = (1+v)R(v)$. On multiplying (7) by $1+v$ and noting that $1+v = 1+u+v-u$, we have

$$(8) \quad S(v) = h(v) - \int_0^v g(v-u)K(u)S(u) du,$$

where

$$(9) \quad h(v) = k_0(1+v)g(v) - \int_0^v (v-u)g(v-u)K(u)R(u) du.$$

By (H2) $(1+v)g(v) \in L^2$. Moreover, by the same assumption $tg(t) \in L^1$, and since $R \in L^2$ and K is bounded the integral in (9) is the convolution of an integrable and a square integrable function. Using § 2.4 a simple calculation gives

$$\|h\|_2 \leq d(\|(1+t)g(t)\|_2 + \|R\|_2 \|tg(t)\|_1),$$

where $d = \max(|a|, |b|)$. From the estimate for $\|R\|_2$ it follows that $\|h\|_2 \leq M$ for a constant M , depending on d , $c(\gamma)$ and g only and independent of s . Now apply the circle criterion to (8) to obtain $S \in L^2$ and $\|S\|_2 \leq c(\gamma)M$. From the Cauchy-Schwarz inequality we now have

$$\int_0^\infty |R(v)| dv = \int_0^\infty |(1+v)^{-1}S(v)| dv \leq \|S\|_2 \leq c(\gamma)M.$$

Thus, $R \in L^1$ and $\|R\|_1 \leq C$ for some constant C independent of s . Since $r(t, s) = 0$ for $s > t$, we have

$$\int_0^\infty |r(t, s)| dt = \int_0^\infty |r(s+v, s)| dv \leq C,$$

and hence $c_1 < \infty$.

Step 2. $c_\infty < \infty$. This follows in much the same way as in step 1 except that we use the representation (4) for $r(t, s)$. Thus, for each t fixed and for $0 \leq v \leq t$, putting $R_1(v) = r(t, t-v)$ we obtain

$$R_1(v) = g(v)k(t-v) - k(t-v) \int_0^v g(v-u)R_1(u) du.$$

Similarly, putting $S_1(v) = (1+v)R_1(v)$ we have

$$(10) \quad S_1(v) = h_1(v) - k(t-v) \int_0^v g(v-u)S_1(u) du,$$

where

$$h_1(v) = (1+v)g(v)k(t-v) - k(t-v) \int_0^v (v-u)g(v-u)R_1(u) du.$$

Now trivial modifications of the arguments in step 1 give a bound C_1 for the norm of R_1 in $L^1(0, t)$ with C_1 independent of t . Since $r(t, s) = 0$ for $s > t$ we have

$$\int_0^\infty |r(t, s)| ds = \int_0^t |r(t, s)| ds = \int_0^t |r(t, t-v)| dv \leq C_1,$$

so that $c_\infty \leq C_1$. This completes the proof of the theorem.

THEOREM 2. *With the assumptions of Theorem 1 let $p \geq 2$, $f \in L^p$ and suppose that $\lim_{t \rightarrow \infty} f(t) = 0$. Then $\lim_{t \rightarrow \infty} x(t) = 0$.*

Proof. By (6) of § 2.3 it suffices to show that $I = \int_0^t r(t, s)f(s) ds$ tends to zero as $t \rightarrow \infty$. Let $\varepsilon > 0$ be given and choose $M > 0$ such that $|f(t)| \leq \varepsilon$ whenever $t \geq M$. Keep M fixed and for $t \geq M$ let $I = I_1 + I_2$, where

$$I_1 = \int_0^M r(t, s)f(s) ds, \quad I_2 = \int_M^t r(t, s)f(s) ds.$$

With the notation of Theorem 1 $|I_2| \leq c_\infty \varepsilon$, so that it suffices to show that $I_1 \rightarrow 0$ as $t \rightarrow \infty$. To this end consider $r(t, s)$. By examining the proof of step 2 of Theorem 1,

we can find two constants D_1 and D_2 independent of t such that $\|R_1\| \leq D_1$ and $\|S_1\| \leq D_2$ where the norms are taken in $L^2(0, t)$. Using these estimates and the Cauchy-Schwarz inequality in the equation defining h_1 , we have

$$|h_1(v)| \leq d|(1+v)g(v)| + dD_1\|tg(t)\|_2,$$

where $d = \max(|a|, |b|)$. The same argument applied to (10) with the above estimate for h_1 gives

$$|S_1(v)| \leq d|(1+v)g(v)| + dD_1\|tg(t)\|_2 + dD_2\|g\|_2.$$

From the definition of S_1 it follows that

$$(11) \quad |R_1(v)| \leq d|g(v)| + D_3(1+v)^{-1}$$

for some constant D_3 independent of t .

Going back to I_1 , we have

$$I_1 = \int_0^M r(t, s)f(s) ds = \int_{t-M}^t r(t, t-v)f(t-v) dv = \int_{t-M}^t R_1(v)f(t-v) dv.$$

Hence, from (11),

$$|I_1| \leq d \int_{t-M}^t |g(v)||f(t-v)| dv + D_3 \int_{t-M}^t (1+v)^{-1}|f(t-v)| dv.$$

If $p \geq 2$ and q is its conjugate, i.e., $1/p + 1/q = 1$, then $1 \leq q \leq 2$. Since $g \in L^1 \cap L^2$ by [4, Thm. 13.19, p. 196], $g \in L^q$ for all $1 \leq q \leq 2$. Hence, by Hölder's inequality

$$|I_1| \leq d\|g\|_q'\|f\|_p + D_3\|(1+v)^{-1}\|_q'\|f\|_p,$$

where $\|\cdot\|_q'$ denotes the norm in $L^q(t-M, t)$. As far as g is concerned we have

$$\|g\|_q' \leq \left(\int_{t-M}^\infty |g(v)|^q dv \right)^{1/q} = o(1) \quad \text{as } t \rightarrow \infty, \quad 1 \leq q \leq 2.$$

The same sort of estimate applies to $\|(1+v)^{-1}\|_q'$ if $1 < q \leq 2$, while if $q = 1$ we have $\|(1+v)^{-1}\|_1' = \log(1+t)/(1+t-M) = o(1)$ as $t \rightarrow \infty$. Thus, all the $\|\cdot\|_q'$ -norms appearing in the final estimate for I_1 tend to zero as $t \rightarrow \infty$, so that $I_1 \rightarrow 0$. This completes the proof.

Remarks. (i) The idea of multiplying (7) by $1+v$ and applying the circle theorem twice is due to Sandberg [8]. (ii) In step 1 of Theorem 1 it was observed that by the Cauchy-Schwarz inequality integrability of R follows at once from $(1+t)R(t) \in L^2$. This simple observation has been used to great effect in feedback theory in the form that if $\hat{R}(j\omega)$ and its derivative are in L^2 then R is integrable. For an interesting example in assessing stability via multipliers see [3, Lemma 2].

We now extend Theorem 1 to a class of equations with unstable g . In doing so we shall need the following proposition which is similar to the usual stability criterion for linear time-invariant systems.

PROPOSITION 1. *Let $g \in A_2$ and suppose that $\hat{g}(z) = \hat{g}_1(z) + p(z)/q(z)$, where $g_1 \in A_1$, p and q are relatively prime polynomials in z and $\deg p < \deg q$. A necessary and sufficient condition for the existence of a real number c and an $f \in A_1$ such that $(c\delta + f) * (\delta + g) = \delta$ is that $\inf \{|1 + \hat{g}(z)| : \operatorname{Re} z \geq 0\} > 0$. If this condition holds then $c = 1$ and $(\delta + f) * g \in A_1$.*

Proof. If such c and f exist then, since $c + \hat{f}(z)$ is a bounded continuous function for $\operatorname{Re} z \geq 0$, the stated condition follows on taking Laplace transforms. To prove the

converse let $n = \deg q$ and let q_0 be the coefficient of z^n in $q(z)$. Write $(1 + \hat{g}(z))^{-1} = h_1(z)(h_2(z))^{-1}$, where $h_1(z) = q(z)(z+1)^{-n} = q_0 + Q(z)(z+1)^{-n}$, with $Q(z)$ a polynomial of degree at most $n-1$ and

$$\begin{aligned} h_2(z) &= (p(z) + q(z) + q(z)\hat{g}_1(z))(z+1)^{-n} \\ &= q_0 + (p(z) + Q(z))(z+1)^{-n} + h_1(z)\hat{g}_1(z). \end{aligned}$$

Since p and q are relatively prime and $g_1 \in L^1$, it follows immediately from the assumption that $\inf \{|h_2(z)|: \operatorname{Re} z \geq 0\} > 0$. In § 2.5 let $\phi(t) = 1+t$. By elementary properties of Laplace transforms and since $L(1+t)$ is an algebra, $h_2(z)$ is the Laplace transform of an element of $A(1+t)$, and since $\lim_{t \rightarrow \infty} (1/t) \log(1+t) = 0$ the inversion theorem of § 2.5 leads to the existence of a real number c_1 and $\phi_1 \in L(1+t)$ such that $(h_2(z))^{-1} = c_1 + \hat{\phi}_1(z)$. Again by elementary properties of Laplace transforms $h_1(z) = q_0 + \hat{\phi}_2(z)$, where $\phi_2 \in L(1+t)$. Thus, $(1 + \hat{g}(z))^{-1} = c_1 q_0 + \hat{f}(z)$, where $f = \phi_1 + \phi_2 + \phi_1 * \phi_2 \in L(1+t)$. By the uniqueness of the Laplace transform it follows that $(c_1 q_0 \delta + f) * (\delta + g) = \delta$, and since $(c_1 q_0 + \hat{f}(z))(1 + \hat{g}(z)) \rightarrow c_1 q_0$ as $\operatorname{Re} z \rightarrow \infty$ we have $c_1 q_0 = 1$.

To complete the proof it has to be verified that $(1+t)f(t) \in L^2$ and $(\delta + f) * g \in A_1$. From $(\delta + f) * (\delta + g) = \delta$ we have $(\delta + f) * g = -f$, so that it suffices to show that $(\delta + f) * g \in A_1$.

In general, let $x \in L(1+t)$ and $y \in A_1$. Since $L(1+t)$ is an algebra under convolution we have $x * y \in L(1+t)$. Moreover,

$$(1+t)(x * y)(t) = \int_0^t (t-s)x(t-s)y(s) ds + \int_0^t x(t-s)(1+s)y(s) ds.$$

Each of the above integrals is the convolution of an L^1 and an L^2 function, so that by § 2.4 $(1+t)(x * y)(t) \in L^2$, and thus $x * y \in A_1$. Going back to $(\delta + f) * g$ we have

$$((\delta + f) * g)^\wedge(z) = \hat{g}(z)(1 + \hat{g}(z))^{-1} = h_3(z)(h_2(z))^{-1},$$

where

$$h_3(z) = \frac{p(z)}{(z+1)^n} + \frac{q_0 \hat{g}_1(z)}{(z+1)^n} + \frac{Q(z)\hat{g}_1(z)}{(z+1)^n}.$$

The first two terms in $h_3(z)$ are the Laplace transforms of two elements of A_1 , while the last term is that of the convolution of two elements of A_1 . Hence, $h_3(z) = \hat{\phi}_3(z)$ for some $\phi_3 \in A_1$. Thus, $(\delta + f) * g = \phi_3 * (c_1 \delta + \phi_1) = c_1 \phi_3 + \phi_3 * \phi_1$ and since $\phi_3 \in A_1$ and $\phi_1 \in L(1+t)$ we have $(\delta + f) * g \in A_1$ as required.

Using the above proposition we can now give an extension of Theorems 1 and 2. (H3) Let $g \in A_2$, and with γ as in (3) we have $\gamma < 1$.

THEOREM 3. *Suppose that in (1) (H1) and (H3) hold with the further assumptions that if \hat{g} has any pole z with $\operatorname{Re} z \geq 0$ then $(a+b)/2 \neq 0$ and $\inf \{|1 + ((a+b)/2)\hat{g}(z)|: \operatorname{Re} z \geq 0\} > 0$. Then the conclusions of Theorems 1 and 2 hold.*

Proof. If \hat{g} has no poles in the closed right-half plane, then $g \in A_1$ and by Theorems 1 and 2 there is nothing to prove. Otherwise, let $c = (a+b)/2$ and $K(s) = k(s) - c$, and write (1) as

$$x(t) + c \int_0^t g(t-s)x(s) ds = f(t) - \int_0^t g(t-s)K(s)x(s) ds.$$

By Proposition 1 there exists $h \in A_1$ such that $(\delta + h) * (\delta + cg) = \delta$ and $m =$

$(\delta + h) * g \in A_1$. Hence,

$$(12) \quad x(t) = (\delta + h) * f - \int_0^t m(t-s)K(s)x(s) ds.$$

Since $h \in A_1$, we have $(\delta + h) * f \in L^p$ whenever $f \in L^p$, $1 \leq p \leq \infty$. It is now trivial to verify that the hypothesis of Theorem 1 applies to (12), so that the conclusions of the said theorem hold. To complete the proof we need only verify that if $p \geq 2$, $f \in L^p$ and $\lim_{t \rightarrow \infty} f(t) = 0$ then $\lim_{t \rightarrow \infty} f_1(t) = 0$, where $f_1 = (\delta + h) * f$. To this end note that since $h \in A_1$ we have $h(t) = (1+t)^{-1}n(t)$ with $n \in L^1 \cap L^2$. Given $\varepsilon > 0$ choose $M > 0$ such that $|f(t)| < \varepsilon$ for $t \geq M$. Then

$$h * f(t) = \int_0^M (1+t-s)n(t-s)f(s) ds + \int_M^t h(t-s)f(s) ds,$$

so that

$$|h * f(t)| \leq (1+t-M)^{-1} \int_0^M |n(t-s)f(s)| ds + \varepsilon \|h\|_1.$$

As in Theorem 2, since $n \in L^1 \cap L^2$ we have $n \in L^q$ with $1/q + 1/p = 1$, so that by Hölder's inequality

$$|h * f(t)| \leq (1+t-M)^{-1} \|n\|_q \|f\|_p + \varepsilon \|h\|_1.$$

Thus, $h * f(t)$ and so $f_1(t)$ tend to zero as $t \rightarrow \infty$ and the result follows.

4. Application to nonlinear equations. Using the above theorems we can now prove our promised L^p stability results for nonlinear equations of the form (1) and (2). Assume that the following hypothesis holds.

$$(H4) \quad n(t, x) \in N(a, b).$$

THEOREM 4. *In (1) suppose that (H2) and (H4) are satisfied. Then for any $p \geq 2$ and any $f \in L^p$ any solution x of (1) is in L^p and there exists a constant c_p independent of f and x such that $\|x\|_p \leq c_p \|f\|_p$. Moreover, if $\lim_{t \rightarrow \infty} f(t) = 0$ then $\lim_{t \rightarrow \infty} x(t) = 0$.*

Proof. Since $p \geq 2$, by [4, Thm. 13.17, p. 196] if $f \in L^p$ then f is locally square integrable. The hypothesis implies that § 2.2 is applicable, and hence any solution of (1) is locally square integrable. Let x be an arbitrary solution. By the above remark the function $k(t) = n(t, x(t))/x(t)$ if $x(t) \neq 0$ and $k(t) = 0$ if $x(t) = 0$ is well defined, and in view of (H4) $k(t)$ satisfies (H1). Thus,

$$(13) \quad x(t) = f(t) - \int_0^t g(t-s)k(s)x(s) ds.$$

Clearly, the assumptions of Theorem 1 hold for (13), so that it has a unique solution which is x itself. The conclusions of the theorem now follow from Theorems 1 and 2.

The above proof does not apply in the case $1 \leq p < 2$, for then we have no a priori knowledge of the existence of local solutions to (1) and so cannot define $k(t)$ as in the proof of Theorem 4. If the existence of solutions as locally integrable functions could be assumed, then the same proof would extend to this case. On the other hand, we can obtain an L^p stability result for a class of nonlinear equations which arise in the study of feedback systems. Thus, if g is the impulse response of a linear time-invariant system and $n(t, x)$ is the characteristic of a nonlinearity, then the input-output equation for the feedback connection of these two elements is given by

$$x(t) = (g * e)(t), \quad e(t) = u(t) - n(t, x(t)),$$

where e is the input to the linear system and x is the output of the feedback system. It follows that x satisfies an equation of the form (2). Using Young's inequality we can partially extend Theorem 4 to (2) for $1 \leq p \leq \infty$.

THEOREM 5. *In (2) suppose (H2) and (H4) are satisfied. Then for any $1 \leq p \leq \infty$ and $u \in L^p$ any solution x of (2) is in L^p and there exists a constant c_p such that $\|x\|_p \leq c_p \|u\|_p$. Moreover, if $p \geq 2$ and $\lim_{t \rightarrow \infty} u(t) = 0$ then $\lim_{t \rightarrow \infty} x(t) = 0$.*

Proof. If $p \geq 2$, then $g \in L^1$ and $u \in L^p$ imply that $g * u \in L^p$, $\|g * u\|_p \leq \|g\|_1 \|u\|_p$ and Theorem 4 applies. Suppose, therefore, that $1 \leq p < 2$. Since $g \in L^1 \cap L^2$ by [4, Thm. 13.19, p. 196] for any r , $1 \leq r \leq 2$, we have $g \in L^r$. The equation $\frac{1}{2} = 1/p + 1/r - 1$ has the solution $r = 2p/(3p-2)$ which lies between 1 and 2. Thus, by Young's inequality, § 2.4, $g * u \in L^2$ and $\|g * u\|_2 \leq \|g\|_r \|u\|_p$. The proof of Theorem 4 may now be repeated word for word to give the existence of a constant d_p such that $\|x\|_p \leq d_p \|g * u\|_p \leq d_p \|g\|_1 \|u\|_p$ for any solution x of (2). Hence, the theorem holds with $c_p = d_p \|g\|_1$.

Each of Theorems 4 and 5 has an extension to the class of unstable systems considered in Theorem 3. The precise statements are given in the following corollary.

COROLLARY 1. *Suppose that (H3) and (H4) are satisfied together with the further assumptions of Theorem 3. Then the conclusions of Theorems 4 and 5 hold.*

Proof. If \hat{g} has no poles in the closed right-half plane, by Theorems 4 and 5 there is nothing to prove. Otherwise, with the same notation as in Theorem 3, transform (1) and (2) to

$$x(t) = (\delta + h) * f(t) - \int_0^t m(t-s)N(s, x(s)) ds$$

and

$$x(t) = m * u(t) - \int_0^t m(t-s)N(s, x(s)) ds$$

respectively, where $N(s, x) = n(s, x) - ((a+b)/2)x$. The required result follows from Theorems 4 and 5 as in the proof of Theorem 3.

Concluding remarks. (i) The above stability results have the same graphical interpretation of the circle criterion in terms of the Nyquist diagram of g and the encirclements of a critical disk [2, p. 141].

(ii) When g is the impulse response of a finite dimensional linear time-invariant system with a strictly proper transfer function, then by elementary properties of the Laplace transform $g \in A_2$. Thus, all the above theorems apply to such systems.

Acknowledgment. We are grateful to the reviewer of the paper for his valuable comments which have lead to a clearer presentation.

REFERENCES

- [1] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.
- [2] C. A. DESOER AND M. VIDYASAGAR, *Feedback Systems: Input-Output Properties*, Academic Press, New York, 1975.
- [3] M. FREEDMAN AND G. ZAMES, *Logarithmic variation criteria for the stability of systems with time-varying gains*, SIAM J. Control, 6 (1968) pp. 487-507.
- [4] E. HEWITT AND K. STROMBERG, *Real and Abstract Analysis*, Springer-Verlag, New York, 1965.
- [5] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, AMS Colloquium Publications 31, American Mathematical Society, Providence, RI, 1957.

- [6] R. K. MILLER, *Nonlinear Volterra Integral Equations*, W. A. Benjamin, San Francisco, 1971.
- [7] I. W. SANDBERG, *A frequency domain condition for stability of feedback systems containing a single time-varying nonlinear element*, Bell Sys. Tech. J., 43 (1964), pp. 1601–1608.
- [8] ———, *On the boundedness of solutions of nonlinear integral equations*, Bell Sys. Tech. J., 44 (1965), pp. 439–453.
- [9] ———, *Some results on the theory of physical systems governed by nonlinear functional equations*, Bell Sys. Tech. J., 44 (1965), pp. 871–898.
- [10] E. M. STEIN AND G. WEISS, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton University Press, Princeton, NJ, 1971.
- [11] G. ZAMES, *On the stability of nonlinear, time-varying feedback systems*, Proc. National Electronics Conference, 20, 1964, pp. 725–730.
- [12] ———, *On the input-output stability of nonlinear time-varying feedback systems, Pt. I and II*, IEEE Trans. Automat. Control, AC-11 (1966), pp. 228–238; pp. 465–477.
- [13] ———, *Nonlinear time-varying feedback systems—condition for L_∞ boundedness derived using conic operators on exponentially weighted spaces*, Proc. 3rd Allerton Conference, Oct. 1965.

CORRIGENDUM: THE OPTIMAL STRATEGY IN THE CONTROL PROBLEM ASSOCIATED WITH THE HAMILTON-JACOBI-BELLMAN EQUATION*

AVNER FRIEDMAN† AND PIERRE-LOUIS LIONS‡

It was pointed out to us by Professor L. C. Evans that in this paper, there is a mistake in the proof of Theorem 1.2 (i.e., the derivation of (2.22) is not justified). We give here an alternate proof based on the following representation of a solution of the H-J-B equation (1.5):

Let

$$\begin{aligned}\phi_\varepsilon(t) &= 0 \quad \text{if } t \leq 0, & \phi'_\varepsilon(t) &= 1 \quad \text{if } t \geq \varepsilon, \\ 0 &\leq \phi'_\varepsilon(t) \leq 1, & \phi_\varepsilon &\in C^\infty \text{ and convex in } t.\end{aligned}$$

Let

$$\begin{aligned}F_\varepsilon(t_1, t_2) &= t_1 + \phi_\varepsilon(t_2 - t_1), \\ F_\varepsilon^N(t_1, t_2, \dots, t_N) &= F_\varepsilon(t_1, F_\varepsilon^{N-1}(t_2, \dots, t_N)).\end{aligned}$$

Obviously

$$\begin{aligned}0 &\leq \frac{\partial F^N}{\partial t_i} \leq 1, & \sum \frac{\partial F_\varepsilon^N}{\partial t_i} &= 1, \\ F_\varepsilon^N &\text{ is convex and increasing in } t_i.\end{aligned}$$

Consider

$$(1) \quad F_\varepsilon^N(A_1 u^\varepsilon - f_1, A_2 u^\varepsilon - f_2, \dots, A_N u^\varepsilon - f_N) = 0,$$

with A_i uniformly elliptic. By a recent result of P. L. Lions (Comm. Pure Appl. Math., 34 (1981), pp. 121-147) and L. C. Evans (to appear) there exists a unique solution 0042

$$(2) \quad u^\varepsilon \rightarrow u, \quad A_1 u^\varepsilon \rightarrow A_1 u$$

uniformly in compact subsets.

To prove Theorem 1.2 we may assume without loss of generality (as before) that A_i are uniformly elliptic, and

$$(3) \quad A_1 f_m \geq c > 0 \quad \text{if } |x| > R.$$

(We take $f_1 \equiv 0$; otherwise consider $u - u_1$, where $Au_1 = f_1$ in R^n .) Suppose there exists an x_0 ,

$$(4) \quad A_1 u(x_0) < 0, \quad |x_0| > R_1.$$

We wish to derive a contradiction with R_1 sufficiently large. From (2), (4) we get

$$(5) \quad A_1 u(x_0) \leq -\alpha < 0$$

for some small ε and large N , α independent of ε , N . Applying A_1 to (1) we obtain

$$(6) \quad \sum F_i(\cdot) [A_i(A_1 u) - A_1 f_i] \geq 0.$$

* This Journal, 18 (1980), pp. 191-198.

† Department of Mathematics, Northwestern University, Evanston, Illinois 60201

‡ Laboratoire d'Analyse Numérique, Université de Paris VI, 4 Place Jussieu, Paris 5^e, France.

Let

$$z(x) = A_1 u(x) + \gamma |x - x_0|^2$$

and choose $\gamma > 0$ small enough that

$$g_i \equiv \gamma A_i |x - x_0|^2 + A_1 f_i \geq \frac{c}{2} \quad \text{for } |x| \geq R, \quad |x_0| \geq R,$$

ρ large enough that $z \geq 0$ on $\partial B_\rho(x_0)$, and R_1 large enough that $B_\rho(x_0) \subset \{|x| > R\}$.

Let

$$z(y_0) = \min_{B_\rho(x_0)} z(y), \quad y_0 \in B_\rho(x_0).$$

Since $z(x_0) < 0$, also $z(y_0) < 0$, and therefore $y_0 \in B_\rho(x_0)$ and

$$A_1 u(y_0) \leq A_1 u(x_0) \leq -\alpha.$$

Notice that

$$(7) \quad \frac{\partial F}{\partial t_1} = 1 - \phi'_\varepsilon(F^{N-1}(t_2, \dots, t_N) - t_1) = 0 \quad \text{if } F^{N-1}(t_2, \dots, t_N) - t_1 \geq \varepsilon.$$

Now, from (1) and (5),

$$\phi_\varepsilon(F^{N-1}(A_2 u - f_2, \dots) - A_1 u) = -A_1 u \geq \alpha \quad \text{at } y_0,$$

and thus for ε small enough

$$F^{N-1}(A_2 u - f_2, \dots) - A_1 u \geq \varepsilon \quad \text{at } y_0.$$

It follows from (7) that

$$\frac{\partial F_1}{\partial t_1}(A_1 u(y_0), F^{N-1}(A_2 u(y_0) - f_2(y_0), \dots)) = 0.$$

Hence (6) gives

$$(8) \quad \sum_{i \geq 2} F'_i(\cdot)(A_i z - g_i) \geq 0 \quad \text{at } y_0.$$

Since

$$A_i z(y_0) \leq c^i z(y_0) < 0, \quad -g_i(y_0) \leq -\frac{c}{2} < 0,$$

the left-hand side of (8) is < 0 , a contradiction.

ON NECESSARY AND SUFFICIENT CONDITIONS FOR REGULATION OF LINEAR SYSTEMS OVER RINGS*

E. EMRE†

Abstract. Necessary and sufficient conditions are given for regulation of linear systems over rings using observers and causal dynamic state feedback systems with the polynomial fractional representation property. The results are then used to obtain stabilizability conditions for systems over integers, delay-differential systems, systems over polynomial rings, and to obtain conditions to make a 2-D system nonrecursive.

1. Introduction. Regulation of linear systems over rings has been considered by several authors. (See Pandolfi [1975], Morse [1976], Sontag [1976], Byrnes [1978], [1979], Kamen and Green [1980], Emre and Khargonekar [1980], Hautus and Sontag [1980] and the references therein.) The first solution to the regulation problem using observers and causal dynamic state feedback for finite free split linear systems over arbitrary commutative rings was given in Emre and Khargonekar [1980], where a theory of observers and coefficient assignment by causal dynamic state feedback was developed. Although the split condition is necessary for regulation via coefficient assignment, it is not necessary for regulation.

The purpose of this paper is to replace the split condition by stabilizability and detectability, which are (as we will show) necessary and sufficient conditions for regulation by observers and causal dynamic state feedback systems satisfying the fractional representation property (a system (F, G, H, J) is said to satisfy the fractional representation property if and only if its transfer matrix can be expressed as PQ^{-1} , where P, Q are polynomial matrices such that $\det Q = \det(zI - F)$).

Recently the concept of detectability has been extended to systems over finitely generated algebras by Hautus and Sontag [1980]. In § 2 of this paper we extend the concepts of stabilizability and detectability to linear systems defined over arbitrary commutative rings and prove that these are necessary and sufficient conditions for regulation by using observers and causal dynamic state feedback systems satisfying the polynomial fractional representation property. (For details of this scheme the reader is referred to Emre and Khargonekar [1980] and also to § 2.) Then in § 3, we use the results of § 2 to obtain stabilizability (also detectability) conditions for systems over polynomial rings, delay-differential systems and systems over integers, we also obtain conditions which make a 2-D system nonrecursive. We also discuss the fact that for the first two cases our detectability result is (essentially) the same as that of Hautus and Sontag [1980].

For general properties and formulations concerning linear systems over commutative rings, the reader is referred to the survey papers Sontag [1976] and Kamen [1978].

2. Stabilizability and detectability. In this section we introduce some notation and other preliminaries, and then give necessary and sufficient conditions for regulation using observers and dynamic causal state feedback systems satisfying the polynomial fractional representation property, namely, stabilizability and detectability. We will

* Received by the editors October 9, 1980, and in revised form January 9, 1981.

† Department of Electrical Engineering, Texas Tech University, Lubbock, TX 79409. The initial part of this research was done when the author was at the Center for Mathematical System Theory, University of Florida, Gainesville, Florida 32611, and was supported in part by the U.S. Army Research Office under grant DAA29-77-G-0225 and the U.S. Air Force under grant AFOSR 76-3034 Mod. B through the Center for Mathematical System Theory, University of Florida.

first assume that the state is available and concentrate on causal dynamic state feedback. Then we will explain how the case where the state is not available can be solved using these results (which concern stabilizability) and using observers (detectability).

Throughout the paper, k denotes an arbitrary but fixed commutative ring with identity. k^p denotes vectors of size p with entries in k . For a given set S , $S[z]$ denotes polynomials in z with coefficients in S . $S^{p \times q}$ denotes $p \times q$ matrices over S . $S((z^{-1}))$ denotes formal power series of the form

$$\sum_{i=l}^{\infty} a_i z^{-i},$$

where l is an integer and a_i is in S . A power series a in $S((z^{-1}))$ is *causal* (strictly causal) if and only if $l \geq 0$ ($l > 0$). $z^{-1}S[[z^{-1}]]$ denotes the set of strictly causal power series with coefficients in S . For a $p \times p$ nonsingular matrix Q over $k[z]$, k_Q is defined to be the k -linear module of polynomial vectors x in $k^p[z]$ such that $Q^{-1}x$ is strictly causal (as a power series).

The k -linear maps Π and Π_Q are defined as follows:

$$\begin{aligned} \Pi: k^p((z^{-1})) &\rightarrow z^{-1}k^p[[z^{-1}]], & x &\mapsto \text{the strictly causal part of } x, \\ \Pi_Q: k^p[z] &\rightarrow k_Q, & x &\mapsto Q\Pi(Q^{-1}x). \end{aligned}$$

For a $p \times r$ polynomial matrix Φ with the i th column φ_i , we define $\Pi_Q(\Phi)$ to be the $p \times r$ matrix whose i th column is $\Pi_Q(\varphi_i)$. For a k -linear map $M: X_1 \rightarrow X_2$, where X_1, X_2 are k -linear modules, $\text{Im } M$ denotes the image of X_1 under M as a k -linear module, and $\ker M$ denotes the kernel of M . If P is a $p \times m$ polynomial matrix whose i th column is expressed as

$$p_i = \sum_{j=0}^{v_i} a_{ij} z^j,$$

where $a_{iv_i} \neq 0$, we say that P is *column proper* if and only if $a_{1v_1}, \dots, a_{mv_m}$ is a set of generators for the k -module k^p . P is *row proper* if and only if its transpose is column proper.

Throughout the paper we assume that there exists a multiplicatively closed set of monic polynomials P_s in $k[z]$, called the set of *stable polynomials*. A rational function p/q , where p, q are in $k[z]$, is *stable* if and only if q is in P_s . A rational matrix is *stable* if and only if all of its entries are stable.

A finite free linear system over a commutative ring k is the triple (F, G, H) , where F is in $k^{n \times n}$, G is in $k^{n \times m}$, and H is in $k^{p \times n}$. Throughout the paper we will be concerned with such systems only. An equivalent representation is in terms of k -linear maps with a finite free state module. For a detailed introduction to linear systems over rings we refer to the survey papers Sontag [1976] and Kamen [1978].

For a given pair (F, G) , we define

$$W_i := \text{Im } G + \dots + \text{Im } F^i G, \quad i = 0, 1, \dots$$

For a matrix A , $\det A$ denotes the determinant of the matrix A .

A system (F_1, G_1, H_1, J_1) is said to have the *polynomial fractional representation property* if and only if its transfer matrix $H_1(zI - F_1)^{-1}G_1 + J_1$ can be expressed as

$$P_c Q_c^{-1},$$

where P_c, Q_c are polynomial matrices (over $k[z]$) such that $\det Q_c = \det(zI - F_1)$.

For a matrix A , $\text{Sp}_k A$ denotes the k -linear module generated by the columns of A . For a polynomial matrix P , $\delta_{ci}(P)$ denotes the degree of the i th column of P .

Now we state the main results of this section.

DEFINITION 2.1. Let F be in $k^{n \times n}$, and let G be in $k^{n \times m}$. Then (F, G) is *stabilizable* if and only if there exist stable rational matrices V_1, V_2 such that

$$(2.2) \quad [zI - F, G] \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = I_n.$$

Remark 2.3. We will call (H, F) *detectable* if and only if (F', H') is stabilizable. (For a matrix A , A' denotes the transpose of A .)

THEOREM 2.4. There exist polynomial matrices P_c, Q_c such that

- (i) Q_c is column proper;
- (ii) $P_c Q_c^{-1}$ is well defined as a power series, and is causal, and has a realization (F_1, G_1, H_1, J_1) such that $\det Q_c = \det(zI - F_1)$; and
- (iii) the determinant of

$$(2.5) \quad \Phi := (zI - F)Q_c + GP_c$$

is a stable polynomial if and only if (F, G) is stabilizable.

Proof. Necessity. Postmultiply both sides of (2.5) by Φ^{-1} .

Sufficiency. If (F, G) is stabilizable, then there exist stable polynomial matrices V_1, V_2 satisfying (2.2). Express V_1, V_2 as

$$V_1 =: N_1(d \cdot I)^{-1}, \quad V_2 =: N_2(d \cdot I)^{-1},$$

where N_1, N_2 are polynomial matrices and d is a stable monic common multiple of the denominators of the entries of V_1 and V_2 . (Such a d exists as V_1 and V_2 are both stable.) Then we have

$$(zI - F)N_1 + GN_2 = d \cdot I.$$

Let v be the smallest integer such that $W_{v-1} = W_{n-1}$. Let r be the degree of d . Let γ_i be the smallest integer such that

$$(2.6) \quad \gamma_i := lr \geq v$$

for some integer $l \geq 1$. Define

$$(2.7) \quad d_1 := d^l, \quad \bar{N}_1 := d^{l-1} \cdot N_1, \quad \bar{N}_2 := d^{l-1} \cdot N_2.$$

Then we have

$$(2.8) \quad (zI - F)\bar{N}_1 + G\bar{N}_2 = d_1 \cdot I.$$

Note here that, as P_s is multiplicatively closed, d_1 is stable.

Equation (2.8) implies that

$$(2.9) \quad \Pi_{(zI-F)}(d_1 \cdot I) \subset W_{n-1}.$$

Then, by Emre [1980, Thm. 3.1], there exist polynomial matrices P_c, Q_c such that:

- (i) Q_c is column proper with the i th column degree $\gamma_i - 1$, and with the highest degree column coefficient matrix I (which ensures that Q_c^{-1} is well defined).
- (ii) $P_c Q_c^{-1}$ is well defined and causal, and has a realization (F_1, G_1, H_1, J_1) such that $\det Q_c = \det(zI - F_1)$. (The fact that this system has the polynomial fractional representation property is seen from the results of Emre and Khargonekar [1980].)

(iii)

$$(zI - F)Q_c + GP_c = d_1 \cdot I.$$

As $d_1^n = \det(d_1 \cdot I)$ is stable, the proof is complete. \square

The next theorem shows that stabilizability is a necessary and sufficient condition for regulation of the system (F, G, I) by causal dynamic feedback with the polynomial fractional representation property.

THEOREM 2.10. *Let F, G be given. Then there exists a finite free dynamic feedback system (F_1, G_1, H_1, J_1) over k such that*

(i) $H_1(zI - F_1)^{-1}G_1 + J_1$ can be expressed as $P_c Q_c^{-1}$ for some polynomial matrices P_c, Q_c with the property that $\det Q_c = \det(zI - F_1)$, and

(ii) the characteristic polynomial of the closed loop system obtained by taking the state as the external direct sum of the states of the open loop system and the feedback system is stable if and only if (F, G) is stabilizable.

Proof. Under the hypotheses of the theorem the characteristic polynomial of the closed loop system can be easily shown to be equal to

$$\det((zI - F)Q_c + GP_c).$$

The rest follows from Theorem 2.4. \square

The next theorem provides a criterion to determine the stabilizability of (F, G) in terms of $[zI - F, G]$.

We consider $k[z, z_1]$, and its maximal ideals which we denote as $\{m_\lambda\}_{\lambda \in \Lambda}$ for some index set Λ .

For a matrix $A = (a_{ij})$ over $k[z, z_1]$, A_λ denotes the matrix which is obtained from A by replacing a_{ij} with the residue class of a_{ij} modulo m_λ . For a detailed description of these concepts, the reader is referred to an algebra book (e.g., Bourbaki [1972, Chapt. 2]). Let $\{m_{\bar{\lambda}}\}_{\bar{\lambda} \in \bar{\Lambda}}$ be the set of maximal ideals of $k[z, z_1]$ such that

$$(2.11) \quad \text{rank}[zI - F, G]_{\bar{\lambda}} < n.$$

After these preliminaries, we have:

THEOREM 2.12. *(F, G) is stabilizable if and only if there exists a stable polynomial q such that*

$$q_{\bar{\lambda}} = 0_{\bar{\lambda}}$$

for each $m_{\bar{\lambda}}$.

Proof. Necessity. If (F, G) is stabilizable, by Theorem 2.4 there exist polynomial matrices P_c, Q_c, Φ , with $\det \Phi$ stable, such that (2.5) is satisfied. Then evaluating both sides of (2.5) at each m_λ , we have from (2.11) that $\det \Phi$ evaluated at each $m_{\bar{\lambda}}$ must be zero.

Sufficiency. It follows from Bourbaki [1972, Chapt. 2] that a matrix M over $k[z, z_1]$ is right invertible over $k[z, z_1]$ if and only if M_λ is right invertible over $k[z, z_1]/m_\lambda$ for each maximal ideal m_λ . Now define

$$M := [zI - F, G, (z_1q - 1) \cdot I].$$

From (2.11), the only maximal ideals such that rank of M_λ can possibly be less than n are $\{m_{\bar{\lambda}}\}$. But, for each $m_{\bar{\lambda}}$, we have $q_{\bar{\lambda}} = 0$. Hence for each $m_{\bar{\lambda}}$, we have

$$\text{rank } M_{\bar{\lambda}} = n.$$

Thus M is right invertible over $k[z, z_1]$. That is, there exist polynomial matrices $M_1(z, z_1), M_2(z, z_1), M_3(z, z_1)$ such that

$$(zI - F)M_1 + GM_2 + (z_1q - 1)M_3 = I.$$

But then, letting $z_1 = 1/q$, we obtain

$$[zI - F, G] \begin{bmatrix} M_1(z, 1/q) \\ M_2(z, 1/q) \end{bmatrix} = I.$$

As q is stable, by definition, (F, G) must be stabilizable. \square

Remark 2.13. If $k = K[s_1, \dots, s_N]$, where K is a field, evaluations at the maximal ideals of $k[z, z_1]$ become evaluations of the polynomials at the points $(s_1^*, \dots, s_N^*, z^*, z_1^*)$ of \bar{K}^{N+2} , where \bar{K} is the algebraic closure of K . For a detailed discussion of this the reader is referred to Hautus and Sontag [1980], and for further details to Bourbaki [1972, Chapt. 2]. In this case our definition of detectability becomes the same (essentially) as the one developed in Hautus and Sontag [1980]. We should note here that Theorem 2.12 remains valid when $zI - F$ and G are replaced by arbitrary polynomial matrices of compatible dimensions.

Based on Theorems 2.4, 2.12, we obtain the following corollary:

COROLLARY 2.14. *If $k = K[s_1, \dots, s_N]$, then (F, G) is stabilizable if and only if there exists a stable polynomial q in $k[z]$ which vanishes at the points of \bar{K}^{N+1} , where $[zI - F, G]$ loses rank.*

Proof. If we note that evaluating a polynomial in $k[z]$ at the points of \bar{K}^{N+2} is the same as evaluating it at the points of \bar{K}^{N+1} , the result follows from Theorem 2.12 and Corollary 2.14. \square

REMARK 2.15. If a system is given in the form (F, G, H) (i.e., the state is not available), then one can use observers and dynamic feedback compensators together, as shown in Emre and Khargonekar [1980], to achieve regulation. It is seen from the formulations given in that paper, and in Hautus and Sontag [1980], that an observer exists if and only if (H, F) is detectable. Furthermore, in such a scheme, the characteristic polynomial of the closed loop system is the product of the characteristic polynomial of the observer and $\det((zI - F)Q_c + GP_c)$, where Q_c, P_c are as defined in this paper. Hence, regulation can be achieved by using observers and causal dynamic feedback systems having the polynomial fractional representation property if and only if (F, G) is stabilizable and (H, F) is detectable, and this result is valid for systems over arbitrary commutative rings with our definitions here.

As for the polynomial fractional representation requirement of the feedback systems, this is not a big restriction as far as known results are concerned because, for example, nondynamic (constant) state feedback satisfies this property trivially. One advantage of this property is that it allows the consideration of internal stability in terms of the polynomial equations arising in stabilizability and detectability and immediately guarantees the realizability of the feedback system. For a natural realization that can be used to implement $P_c Q_c^{-1}$, the reader is referred to Kalman, Falb and Arbib [1969], Fuhrmann [1976] and Emre [1980b].

3. Stabilizability of some specific classes of systems. In this section, using the results of § 2, we obtain stabilizability (detectability) criteria for certain specific classes of linear systems over rings.

1) *Systems over integers.* These systems are discrete time systems (F, G, H) over integers. The problem is to construct dynamic compensators with integer coefficients such that the closed-loop system is regulated. In this case the set of stable polynomials is of the form z^r for some $r \geq 0$. From the definition, (F, G) is stabilizable if and only if $[zI - F, G]$ has a right inverse whose entries have denominators of the form z^r for some integer $r \geq 0$, or if and only if there exist polynomial matrices N_1, N_2 with integer

coefficients and some integer $r \geq 0$ such that

$$(3.1) \quad (zI - F)N_1 + GN_2 = z^r \cdot I.$$

We see that (3.1) is possible if and only if

$$(3.2) \quad F^r \subset \text{Im } G + \cdots + \text{Im } F^{n-1}G,$$

for some $r \geq 0$.

2) *2-D systems*. In this case, k = the ring of proper rational functions over a field. Here the problem is to find a causal dynamic feedback system such that the characteristic polynomial of the closed loop system becomes z^r for some $r \geq 0$. This problem will have a solution if and only if (F, G) and (F', H') satisfy the condition (3.2).

3) *Systems over a polynomial ring* $K[s_1, \dots, s_N]$. In this case we obtain the following theorem.

THEOREM 3.3. (F, G) is stabilizable if and only if at every point $(s_1^*, \dots, s_N^*, z^*)$ of \bar{K}^{N+1} where $[zI - F, G]$ loses rank, the real part of z^* is negative.

4) *Delay-differential systems*. This is the same as systems over polynomial rings except that the set of stable polynomials is different. We have Corollary 2.14.

REFERENCES

- N. BOURBAKI [1972], *Commutative Algebra*, Addison-Wesley, Reading, MA.
- C. I. BYRNES [1978], *On the control of certain infinite dimensional systems by algebro-geometric techniques*, Amer. J. Math., 100, pp. 1333–1381.
- , [1979], *On the stabilizability of linear control systems depending on parameters*, Proc. IEEE Conf. on Decision and Control, Florida.
- P. A. FUHRMANN [1976], *Algebraic system theory: an analyst's point of view*, J. Franklin Inst., 301, pp. 521–540.
- E. EMRE [1980a], *The polynomial equation $QQ_c + RP_c = \Phi$ with application to dynamic feedback*, this Journal, 18, pp. 611–620.
- , [1980b], *On a natural realization of matrix fraction descriptions*, IEEE Trans. Automat. Control, AC-25, pp. 288–289.
- E. EMRE AND P. P. KHARGONEKAR [1980], *Regulation of split systems over rings: coefficient assignment and observers*, IEEE Conference on Decision and Control, New Mexico, IEEE Transactions on Autom. Control, to appear.
- M. L. J. HAUTUS AND E. D. SONTAG [1980], *An approach to detectability and observers*, Dept. of Math., Memo-COSOR 79-08, Eindhoven University of Technology, Eindhoven, the Netherlands, in Algebraic and Geometric Methods in Linear Systems Theory, C. I. Byrnes and C. F. Martin, eds., American Mathematical Society, Providence, RI, to appear.
- R. E. KALMAN, P. L. FALB AND M. A. ARBIB [1969], *Topics in Mathematical System Theory*, McGraw-Hill, New York.
- E. W. KAMEN [1978], *Lectures on algebraic system theory: Linear systems over rings*, NASA Contractor Rep. 3016.
- E. W. KAMEN AND W. L. GREEN [198-], *Asymptotic stability of linear difference equations defined over a commutative Banach algebra*, J. Math. Anal. Appl., to appear.
- P. P. KHARGONEKAR [1982], *On matrix fraction representations for linear systems over commutative rings*, this Journal, this issue, pp. 172–197.
- A. S. MORSE [1976], *Ring models for delay-differential systems*, Automatica, 12, pp. 529–539.
- L. PANDOLFI [1975], *On feedback stabilization of functional differential equations*, Bull. Un. Mat. Ital., 11, pp. 626–635.
- E. D. SONTAG [1976], *Linear systems over commutative rings: a survey*, Recherche di Automatica, 7, pp. 1–34.

ON THE LOCAL CONVERGENCE OF QUASI-NEWTON METHODS FOR CONSTRAINED OPTIMIZATION*

PAUL T. BOGGS[†], JON W. TOLLE[‡] AND PYNG WANG[§]

Abstract. We consider the application of a general class of quasi-Newton methods to the solution of the classical equality constrained nonlinear optimization problem. Specifically, we develop necessary and sufficient conditions for the Q -superlinear convergence of such methods and present a companion linear convergence theorem. The essential conditions relate to the manner in which the Hessian of the Lagrangian function is approximated.

1. Introduction. In this paper we consider means of solving the equality constrained nonlinear optimization problem

$$\begin{aligned} & \text{Minimize } f(x) \\ \text{(NLP)} \quad & \text{subject to } g(x) = 0, \end{aligned}$$

where it is assumed that $f: \mathbb{R}^n \rightarrow \mathbb{R}$ and $g: \mathbb{R}^n \rightarrow \mathbb{R}^m$ are smooth functions.

After two decades of experimentation with penalty function techniques, augmented Lagrangian functions, gradient projection methods and other procedures, research on numerical methods for solving NLP has recently centered on implementing some form of a quasi-Newton technique for this constrained problem. The preeminence of quasi-Newton methods for solving unconstrained nonlinear problems and good experimental results to date lead one to believe that this approach is sound. However, there remain numerous questions concerning convergence, rates of convergence, update formulas, and implementation that are as yet unanswered. It is the purpose of this paper to shed light on some of these questions, in particular, on the local and Q -superlinear convergence of these methods.

We define a quasi-Newton method for NLP as an iterative scheme which generates sequences $\{x^k\}$, $\{\lambda^k\}$, and $\{B_k\}$ from formulas

$$(1.1) \quad \lambda^{k+1} = \Lambda(x^k, \lambda^k, B_k),$$

$$(1.2) \quad B_k \delta_x^k = -l_x(x^k, \lambda^{k+1}),$$

$$(1.3) \quad x^{k+1} = x^k + \alpha^k \delta_x^k,$$

$$(1.4) \quad B_{k+1} = \mathcal{B}(x^k, x^{k+1}, \lambda^k, \lambda^{k+1}, B_k),$$

where x^0 , λ^0 and B_0 are given, Λ and \mathcal{B} are appropriate update functions and $l(x, \lambda) = f(x) + \lambda^T g(x)$ is the standard Lagrangian function. The step lengths α^k are obviously important, but for local convergence theory $\alpha^k = 1$ is the optimal choice and α^k will be taken to have this value throughout.

* Received by the editors February 6, 1980 and in final revised form March 27, 1981. This work was supported in part by the U.S. Army Research Office under grant DAAG29-79-G0014.

[†] U.S. Army Research Office, Research Triangle Park, North Carolina 27709 and Curriculum in Operations Research and Systems Analysis, University of North Carolina, Chapel Hill, North Carolina 27514.

[‡] Department of Mathematics and Curriculum in Operations Research and Systems Analysis, University of North Carolina, Chapel Hill, North Carolina 27514.

[§] Bell Laboratories, Whippany, New Jersey 07981.

Much of the recent work on quasi-Newton methods for NLP can be put into this framework. Powell [8], [9], following the work of Han [6], obtains δ_x^k by solving a quadratic program

$$\begin{aligned} &\text{Minimize } \nabla f(x^k)^T \delta_x^k + \frac{1}{2} \delta_x^k{}^T B_k \delta_x^k \\ &\text{subject to } \nabla g(x^k)^T \delta_x^k = -g(x^k), \end{aligned}$$

and chooses λ^{k+1} to be the optimal multiplier vector for this program. B_k is a standard rank two update approximation to l_{xx} with a modification which assures that the B_k remain positive definite.

Tapia [10], [11] shows that Powell's choices of δ_x^k and λ^{k+1} can be obtained by applying a structured quasi-Newton method to the system

$$(1.5) \quad l_x(x, \lambda) = 0, \quad l_\lambda(x, \lambda) = 0.$$

That is, x^{k+1} and λ^{k+1} are obtained from the equations

$$(1.6) \quad \begin{bmatrix} B_k & \nabla g(x^k) \\ \nabla g(x^k)^T & 0 \end{bmatrix} \begin{bmatrix} \delta_x^k \\ \delta_\lambda^k \end{bmatrix} = \begin{bmatrix} -l_x(x^k, \lambda^k) \\ -g(x^k) \end{bmatrix},$$

$$(1.7) \quad \lambda^{k+1} = \lambda^k + \delta_\lambda^k,$$

with x^{k+1} given by (1.3). Here again B_k is an approximation to l_{xx} , so that the $(n+m) \times (n+m)$ matrix in (1.6) is a structured approximation to the Jacobian matrix for the system (1.5). It is easily seen that the solutions to (1.6), (1.7), given by

$$(1.8) \quad \lambda^{k+1} = (\nabla g(x^k)^T B_k^{-1} \nabla g(x^k))^{-1} \{g(x^k) - \nabla g(x^k)^T B_k^{-1} \nabla f(x^k)\},$$

$$(1.9) \quad \begin{aligned} \delta_x^k = & -B_k^{-1} \{I - \nabla g(x^k)(\nabla g(x^k)^T B_k^{-1} \nabla g(x^k))^{-1} \nabla g(x^k)^T B_k^{-1}\} \nabla f(x^k) \\ & - B_k^{-1} \nabla g(x^k)(\nabla g(x^k)^T B_k^{-1} \nabla g(x^k))^{-1} g(x^k), \end{aligned}$$

also satisfy (1.1) and (1.2). In addition to the formula (1.8), Tapia presents a number of other possible updates for λ , preferring, for theoretical reasons, a double update of λ .

In [1] the authors have considered a variation of the system (1.5) in which the Lagrangian function $l(x, \lambda)$ is replaced by a more general Lagrangian $M(x, \lambda)$ which is quadratic in λ . The purpose for introducing this generalization was to obtain better convergence from poor starting points. Locally, however, the quasi-Newton equations derived from $M(x, \lambda)$ are nearly identical to those of (1.6).

The local convergence of these methods has been investigated by a number of authors. Before reviewing their results, we point out the distinction between Q -superlinear and R -superlinear rates of convergence and the difference between the convergence rates of the vector $\{(x^k, \lambda^k)\}$ and its component $\{x^k\}$. Recall that a vector sequence $\{v^k\}$ converges R -superlinearly to v^* if and only if the sequence $\{v^k - v^*\}$ is bounded by a sequence which converges Q -superlinearly to zero. Because an R -superlinearly convergent sequence need not be even Q -linearly convergent, R -superlinear convergence by itself is computationally meaningless. It is also the case that the Q -superlinear convergence of $\{v^k\}$ implies only the R -superlinear convergence of its components. (See Tapia [10, section 8] for a more detailed discussion.) Since λ^{k+1} depends only on x^k and not on λ^k , to be most effective the structured quasi-Newton method should yield Q -superlinear convergence of the sequence $\{x^k\}$.

The major convergence analyses center on how well and in what sense the B_k generated by (1.4) approximate the Hessian of the Lagrangian function at (x^*, λ^*) ,

the optimal solution pair. These analyses are based on similar studies of the unconstrained problem. In the latter case extensive use is made of the Broyden–Dennis–Moré analysis of the quasi-Newton update formulas and the Dennis–Moré characterization of Q -superlinear convergence. (See Dennis and Moré [4] for a survey of these results.) This characterization shows that Q -superlinear convergence in the unconstrained case occurs if and only if

$$(1.10) \quad \frac{|(\bar{B}_k - \nabla^2 f(x^*))\delta^k|}{|\delta^k|} \rightarrow 0,$$

where $\nabla^2 f(x^*)$ is the Hessian of the function to be minimized, δ^k is the step generated, and \bar{B}_k is the approximate Hessian.

For the constrained case Powell [9] develops a procedure for updating the B_k and shows that, under the second order sufficiency conditions, the resulting method is at least R -superlinearly convergent in x . He also provides a condition related to (1.10) which is sufficient for “2-step” superlinear convergence. In particular, for the projection matrix

$$P(x) = I - \nabla g(x)(\nabla g(x)^T \nabla g(x))^{-1} \nabla g(x)^T,$$

Powell shows that

$$(1.11) \quad \frac{|P(x^k)(B_k - l_{xx}(x^*, \lambda^*))P(x^k)\delta_x^k|}{|\delta_x^k|} \rightarrow 0$$

is sufficient for

$$\frac{|x^{k+1} - x^*|}{|x^{k-1} - x^*|} \rightarrow 0.$$

Powell was not able to show that his method satisfies this condition, however.

Also under the second order sufficiency conditions, Han [6] demonstrates the Q -superlinear convergence of $\{(x^k, \lambda^k)\}$ when a form of Greenstadt’s update is used in (1.4). However, Han requires the stronger assumption that $l_{xx}(x^*, \lambda^*)$ be positive definite in order to obtain the Q -superlinear convergence in (x, λ) for the BFGS update. It should be noted that Greenstadt’s method is not computationally attractive, since it almost always performs poorly in spite of its theoretical properties. To guarantee that $l_{xx}(x^*, \lambda^*)$ is positive definite requires the addition of a penalty term to the Lagrangian, a computationally unattractive option.

Tapia [10], [11] and Glad [5] obtain Q -superlinear convergence in (x, λ) for $l_{xx}(x^*, \lambda^*)$ positive definite. Tapia [10] obtains the stronger result of Q -superlinear convergence in x but at the cost of an additional update of λ at each step.

In this paper we first characterize Q -superlinear convergence in x for these methods (Theorem 3.1). The characterization is a natural generalization of the Dennis–Moré result (1.10). Simply put, it states that Q -superlinear convergence in x occurs if and only if

$$(1.12) \quad \frac{|P(x^k)(B_k - l_{xx}(x^*, \lambda^*))\delta_x^k|}{|\delta_x^k|} \rightarrow 0,$$

where $P(x)$ is the projection matrix given above. Note that (1.12) does not contain a post-multiplication of $B_k - l_{xx}(x^*, \lambda^*)$ by $P(x^k)$ as does (1.11), and hence, it takes into account the action of $B_k - l_{xx}(x^*, \lambda^*)$ off of the null space of $\nabla g(x^k)^T$, which (1.11) does not.

Using the characterization (1.11), we then show (Theorem 3.2) that Q -superlinear convergence in x is obtained when $l_{xx}(x^*, \lambda^*)$ is positive definite. This is a slightly stronger result than those previously published and reviewed above. Finally, a sufficient condition for Q -linear convergence in x is established (Theorem 3.3). This theorem also makes use of the matrices $P(x^k)(B_k - l_{xx}(x^*, \lambda^*))$, requiring that they be small in norm for all k . Hence it provides a complementary result to Theorem 3.1.

2. Basic notation and assumptions. For the problem NLP considered here we assume f and g are at least three times continuously differentiable and that the gradient $\nabla g(x)$ has full rank for all x . In addition we assume that NLP has a (local) solution x^* at which the second order sufficiency conditions hold. That is, there exists a unique vector $\lambda^* \in \mathbb{R}^m$ such that

- (i) $l_x(x^*, \lambda^*) = 0$,
- (ii) $\nabla g(x^*)^T y = 0, y \neq 0$ implies $y^T l_{xx}(x^*, \lambda^*) y > 0$.

For functions $h: \mathbb{R}^n \rightarrow \mathbb{R}^q$, we denote the Jacobian and Hessian matrices by $\nabla h(x)$ and $\nabla^2 h(x)$, respectively. Here, for notational convenience, $\nabla h(x)$ is always written as an $n \times q$ matrix. For functions of x and λ , we denote derivatives with respect to x or λ by subscripts; hence, $l_x(x, \lambda) = \nabla f(x) + \nabla g(x)\lambda$, $l_{x\lambda}(x, \lambda) = \nabla g(x)$, etc.

Vectors are always column vectors unless transposed, the transposition operation for vectors and matrices being indicated by a superscript T .

In the theory of constrained minimization, the projection of vectors onto the tangent space of the level sets of the constraints plays an important role. For a given \hat{x} , the matrix

$$P(\hat{x}) = [I - \nabla g(\hat{x})(\nabla g(\hat{x})^T \nabla g(\hat{x}))^{-1} \nabla g(\hat{x})^T]$$

projects vectors onto the tangent space of the smooth manifold

$$\{x: g(x) = g(\hat{x})\}$$

at $x = \hat{x}$. The projection onto the orthogonal complement of this tangent space will be denoted by $Q(\hat{x})$. Thus

$$Q(\hat{x}) = I - P(\hat{x}).$$

$\|\cdot\|$ will everywhere denote the l_2 -norm. In § 3, it is necessary to use the Frobenius norm for matrices. The Frobenius norm weighted by the matrix M is denoted by $\|\cdot\|_M$.

3. Necessary and sufficient conditions for superlinear convergence. We consider the algorithm obtained by applying a structured quasi-Newton method to the system (1.5), thus obtaining the formulas given in (1.6) and (1.7) with solutions (1.8) and (1.9). This algorithm has the important property that δ_x^k satisfies the linearized constraints, i.e.,

$$g(x^k) + \nabla g(x^k)^T \delta_x^k = 0.$$

Extending the analysis by Powell [9], we obtain a necessary and sufficient condition for Q -superlinear convergence in x , given linear convergence and a few basic assumptions on the approximating matrices B_k . The essential condition is that the matrix B_k must approximate the Hessian matrix $l_{xx}(x^*, \lambda^*)$ in the sense of Dennis and Moré [3] but only when projected onto the tangent hyperplane to the surface $\{z: g(z) = g(x^k)\}$.

We assume in the remainder of this section that the B_k are symmetric, nonsingular, and uniformly bounded. In addition, we assume that the matrices B_k are uniformly positive definite on the null space of $\nabla g(x^*)^T$. That is, there exists a $\beta > 0$ such that

whenever $y \neq 0$ and $\nabla g(x^*)^T y = 0$,

$$y^T B_k y \geq \beta |y|^2$$

for every k . Thus we require the B_k to satisfy the second order sufficiency condition satisfied by $l_{xx}(x^*, \lambda^*)$ (see § 2). This assumption is slightly weaker than that of Powell who assumes that the B_k are positive definite on \mathbb{R}^n and uniformly positive definite on the null space of $\nabla g(x^*)^T$.

Note that with the above assumptions on the matrices B_k , the matrices $\nabla g(x^*)^T B_k^{-1} \nabla g(x^*)$ are nonsingular. For if $\nabla g(x^*)^T B_k^{-1} \nabla g(x^*) z = 0$, then $B_k^{-1} \nabla g(x^*) z$ is in the null space of $\nabla g(x^*)^T$, and hence, $z \neq 0$ implies the contradiction $0 < (B_k^{-1} \nabla g(x^*) z)^T B_k (B_k^{-1} \nabla g(x^*) z) = z^T \nabla g(x^*)^T B_k^{-1} \nabla g(x^*) z = 0$. It follows that (1.8) and (1.9) are well-defined for x^k sufficiently near to x^* .

Our first result is a generalization of a result of Powell [9].

LEMMA 1. *The value of δ_x^k is invariant under the transformation*

$$B_k \rightarrow B_k + \nabla g(x^k) U^T \equiv \hat{B}_k,$$

where U is any $n \times m$ matrix such that both \hat{B}_k and $\nabla g(x^k)^T \hat{B}_k^{-1} \nabla g(x^k)$ are nonsingular.

Proof. It follows from (1.9) that the lemma is true if the matrices

$$A_1 = \hat{B}_k^{-1} \nabla g(x^k) [\nabla g(x^k)^T \hat{B}_k^{-1} \nabla g(x^k)]^{-1}$$

and

$$A_2 = \hat{B}_k^{-1} - \hat{B}_k^{-1} \nabla g(x^k) [\nabla g(x^k)^T \hat{B}_k^{-1} \nabla g(x^k)]^{-1} \nabla g(x^k)^T \hat{B}_k^{-1}$$

are independent of U . The assumptions on U allow the use of the Sherman–Morrison–Woodbury formula (see, e.g., Ortega and Rheinboldt [7, p. 50]) to express \hat{B}_k^{-1} as

$$\hat{B}_k^{-1} = B_k^{-1} - B_k^{-1} \nabla g(x^k) [I - U^T B_k^{-1} \nabla g(x^k)]^{-1} U^T B_k^{-1}.$$

Substitution of this expression into A_1 yields

$$A_1 = B_k^{-1} \nabla g(x^k) [\nabla g(x^k)^T B_k^{-1} \nabla g(x^k)]^{-1},$$

which establishes the result for A_1 . For A_2 , note that

$$A_2 = [I - A_1 \nabla g(x^k)] \hat{B}_k^{-1}.$$

Again using the expression for \hat{B}_k^{-1} yields the desired result.

It follows from the assumptions made on the B_k that if $U = \gamma \nabla g(x^k)$, where γ is a sufficiently large positive constant, then the hypotheses of the lemma are satisfied. It should also be noted that the value of λ^{k+1} is not invariant under the given transformation in B_k . Thus, a variety of choices of λ^{k+1} give rise to the same value of δ_x^k (as demonstrated by Tapia in [11]). However, it is easily seen that the first order necessary conditions and equation (1.8) imply that if $\{x^k\} \rightarrow x^*$ then $\{\lambda^{k+1}\} \rightarrow \lambda^*$.

For convenience, we now write (1.9) in the form

$$(3.1) \quad -B_k \delta_x^k = V_k \nabla f(x^k) + W_k g(x^k).$$

We note that the two vectors on the right-hand side are conjugate with respect to B_k^{-1} ; in fact, $V_k^T B_k^{-1} W_k = 0$. Letting P_k be the projection matrix at x^k defined in § 2, we see that

$$(3.2a) \quad P_k V_k = P_k,$$

$$(3.2b) \quad V_k P_k = V_k,$$

$$(3.2c) \quad P_k W_k = 0.$$

The next lemma is also a modification of the results of Powell. Two positive sequences, $\{s^k\}$ and $\{r^k\}$, which converge to zero are said to be of the same order if there exist positive constants c_1 and c_2 such that for k sufficiently large,

$$c_1 \leq \frac{|r^k|}{|s^k|} \leq c_2.$$

LEMMA 2. Suppose $\{x^k\} \rightarrow \{x^*\}$ with a linear rate of convergence. Then the sequences $\{|\delta_x^k|\}$, $\{|x^k - x^*|\}$, and $\{|g(x^k)| + |P_k \nabla f(x^k)|\}$ converge to zero and are of the same order.

Proof. Using Lemma 1 and the properties of the B_k , we see that by choosing γ sufficiently large we may replace the B_k by \hat{B}_k for which \hat{B}_k and \hat{B}_k^{-1} are uniformly bounded and positive definite. The change does not affect the value of δ_x^k or the relations (3.2). Now using (3.2b) and (3.1) there exists an $\alpha_1 > 0$ such that

$$|\delta_x^k| \leq \alpha_1 \{|g(x^k)| + |P_k \nabla f(x^k)|\}.$$

(3.1), (3.2a), (3.2c), and the linearized constraint equation yield the existence of an $\alpha_2 > 0$ such that

$$\{|g(x^k)| + |P_k \nabla f(x^k)|\} \leq \alpha_2 |\delta_x^k|.$$

Thus, $\{|\delta_x^k|\}$ and $\{|g(x^k)| + |P_k \nabla f(x^k)|\}$ are of equivalent order. That $\{|\delta_x^k|\}$ and $\{|x^k - x^*|\}$ are of the same order follows from the consequence of linear convergence

$$1 - r \leq \frac{|\delta_x^k|}{|x^k - x^*|} \leq 1 + r,$$

where $r < 1$. This completes the proof.

Now let $\{G_k\}$ be any sequence of matrices satisfying

- (i) $G_k \delta_x^k = l_x(x^{k+1}, \lambda^{k+1}) - l_x(x^k, \lambda^{k+1}),$
- (ii) $G_k \rightarrow l_{xx}(x^*, \lambda^*).$

For example, G_k could be chosen as

$$G_k = \int_0^1 l_{xx}(x^k + t\delta_x^k, \lambda^{k+1}) dt.$$

LEMMA 3. Assume $\{x^k\} \rightarrow x^*$ linearly. Then there exists an $\alpha > 0$ such that

$$|x^{k+1} - x^*| \leq \alpha [|\delta_x^k|^2 + |P_k(G_k - B_k)\delta_x^k|].$$

Proof. By Lemma 2 there exists an $\eta > 0$ such that

$$(3.3) \quad |x^{k+1} - x^*| \leq \eta \{|g(x^{k+1})| + |P_{k+1} \nabla f(x^{k+1})|\}.$$

Now

$$(3.4) \quad g(x^{k+1}) = g(x^k) + \nabla g(x^k)^T \delta_x^k + O(|\delta_x^k|^2) = O(|\delta_x^k|^2).$$

From (i) above and (1.3),

$$(G_k - B_k)\delta_x^k = l_x(x^{k+1}, \lambda^{k+1}).$$

Using the fact that $P_{k+1} \nabla g(x^{k+1}) = 0$, we obtain the identity

$$(3.5) \quad (P_{k+1} - P_k)(G_k - B_k)\delta_x^k + P_k(G_k - B_k)\delta_x^k = P_{k+1} \nabla f(x^{k+1}).$$

The smoothness assumptions on $g(x)$ assure that

$$(3.6) \quad P_{k+1} - P_k = O(|\delta_x^k|)$$

and the lemma follows from (3.3)–(3.6) and the uniform boundedness of the G_k and the B_k .

We can now state and prove the necessary and sufficient conditions for Q -superlinear convergence for the structured quasi-Newton method as applied to the system (1.5).

THEOREM 3.1. *Let $(\delta_x^k, \delta_\lambda^k)$ satisfy (1.6) where the matrices B_k satisfy the conditions stated at the beginning of the section. Suppose $\{x^k\} \rightarrow x^*$ linearly. Then $\{x^k\} \rightarrow x^*$ Q -superlinearly if and only if*

$$(3.7) \quad \lim_{k \rightarrow \infty} \frac{|P_k(B_k - l_{xx}(x^*, \lambda^*))\delta_x^k|}{|\delta_x^k|} = 0,$$

where $P_k = I - \nabla g(x^k)(\nabla g(x^k)^T \nabla g(x^k))^{-1} \nabla g(x^k)^T$.

Proof. Let $\{G_k\}$ be a sequence of approximations to $l_{xx}(x^*, \lambda^*)$ as defined above. Clearly G_k can replace $l_{xx}(x^*, \lambda^*)$ in (3.7). Now suppose (3.7) holds. Then by Lemma 3

$$|x^{k+1} - x^*| = o(|\delta_x^k|).$$

But by Lemma 2 $\{|\delta_x^k|\}$ and $\{|x^k - x^*|\}$ are of the same order; hence there is a constant $\alpha > 0$ such that

$$\frac{|x^{k+1} - x^*|}{|x^k - x^*|} \leq \alpha \cdot \frac{|x^{k+1} - x^*|}{|\delta_x^k|} = \alpha \cdot \frac{o(|\delta_x^k|)}{|\delta_x^k|},$$

which demonstrates Q -superlinear convergence.

For the converse, suppose $\{x^k\} \rightarrow x^*$ Q -superlinearly. Using Lemma 2 and (3.5), we have that, for some $\eta > 0$,

$$|P_k(B_k - G_k)\delta_x^k + (P_{k+1} - P_k)(B_k - G_k)\delta_x^k| + |g(x^{k+1})| \leq \eta |x^{k+1} - x^*|,$$

which, together with (3.4) and (3.6), imply that

$$\frac{|P_k(B_k - G_k)\delta_x^k|}{|\delta_x^k|} \leq \eta \cdot \frac{|x^{k+1} - x^*|}{|\delta_x^k|} + O(|\delta_x^k|).$$

Again using Lemma 2, we have

$$\frac{|P_k(B_k - G_k)\delta_x^k|}{|\delta_x^k|} \leq \eta \cdot \frac{|x^{k+1} - x^*|}{|\delta_x^k|} + O(|\delta_x^k|) \leq \hat{\eta} \cdot \frac{|x^{k+1} - x^*|}{|x^k - x^*|} + O(|\delta_x^k|).$$

Letting $k \rightarrow \infty$ (and hence $|\delta_x^k| \rightarrow 0$) gives the desired results.

We note that if $f(x)$ is augmented by the penalty term $cg(x)^T g(x)$ with c a large positive constant, then the second order sufficiency conditions imply that the Hessian of the augmented Lagrangian is positive definite at (x^*, λ^*) . Moreover, it is easily shown that the formula (1.9) is unchanged by this added term; thus the only effect is in the update formula (1.4). If, as is common, the B_k are chosen to approximate $l_{xx}(x^*, \lambda^*)$ in the sense that

$$(3.8) \quad B_{k+1}\delta_x^k = y^k \equiv l_x(x^{k+1}, \lambda^{k+1}) - l_x(x^k, \lambda^{k+1}),$$

then the assumption that $l_{xx}(x^*, \lambda^*)$ is positive definite makes the update formulas which preserve positive definiteness, such as the DFP or BFGS, natural candidates

for use in this scheme. The next theorem shows that Q -superlinear convergence is achieved in these cases (cf. Han [6], Tapia [11], and Glad [5]). The following lemma is important for our proof.

LEMMA 4. *Let B_{k+1} be derived from B_k by either the DFP or the BFGS update with y^k given by (3.8). Assume that $l_{xx}(x^*, \lambda^*)$ is positive definite and $\{x^k\}$ converges linearly to x^* . Let B_{k+1}^* be generated from B_k using the same update formula as for B_{k+1} but with y^k replaced by y^* , where*

$$y^* \equiv l_x(x^{k+1}, \lambda^*) - l_x(x^k, \lambda^*).$$

Then

$$(3.9) \quad |B_{k+1} - B_{k+1}^*| \leq \alpha \sigma((x^{k+1}, \lambda^{k+1}), (x^k, \lambda^k)),$$

where α is a constant independent of k and

$$\sigma((x^{k+1}, \lambda^{k+1}), (x^k, \lambda^k)) = \max \{ |(x^j, \lambda^j) - (x^*, \lambda^*)| : j = k, k+1 \}.$$

Proof. We prove the lemma for the BFGS update. The proof for the DFP update is similar but more laborious. From the definitions of y^k and y^* we have

$$y^k - y^* = (\nabla g(x^{k+1}) - \nabla g(x^k))(\lambda^{k+1} - \lambda^*),$$

and thus, there is a constant β_1 such that

$$|y^k - y^*| \leq \beta_1 |\delta_x^k| |\lambda^{k+1} - \lambda^*|.$$

From the assumptions there exist positive constants η_1, η_2 such that for large k

$$\begin{aligned} (y^k)^T \delta_x^k &\geq \eta_1 |\delta_x^k|^2, & |y^k| &\leq \eta_2 |\delta_x^k|, \\ (y^*)^T \delta_x^k &\geq \eta_1 |\delta_x^k|^2, & |y^*| &\leq \eta_2 |\delta_x^k|. \end{aligned}$$

For the BFGS update,

$$B_{k+1} - B_{k+1}^* = \frac{((y^*)^T \delta_x^k) y^k (y^k)^T - ((y^k)^T \delta_x^k) y^* (y^*)^T}{((y^*)^T \delta_x^k)((y^k)^T \delta_x^k)},$$

from which it follows that

$$|B_{k+1} - B_{k+1}^*| \leq \frac{3|y^k||y^*||\delta_x^k||y^k - y^*|}{|(y^*)^T \delta_x^k| |(y^k)^T \delta_x^k|} \leq 3\beta_1 (\eta_2/\eta_1)^2 |\lambda^{k+1} - \lambda^*|.$$

Inequality (3.9) follows immediately.

THEOREM 3.2. *Assume $l_{xx}(x^*, \lambda^*)$ is positive definite. If the B_k are obtained by either the DFP or BFGS formulas with y^k defined by (3.8) and if $\{x^k\}$ converges to x^* linearly, then the convergence is Q -superlinear.*

Proof. The proof follows the lines of argument used in unconstrained optimization. Let y^* and B_{k+1}^* be as defined in Lemma 4. From our assumptions x^* is an unconstrained minimum of $l(x, \lambda^*)$ and hence the results of Broyden, Dennis and Moré [2] for the unconstrained case can be applied to obtain the fundamental inequality:

$$(3.10) \quad \|B_{k+1}^* - l_{xx}(x^*, \lambda^*)\|_M \leq \{(1 - c\theta_k^2)^{1/2} + \alpha_1 \hat{\sigma}(x^{k+1}, x^k)\} \|B_k - l_{xx}(x^*, \lambda^*)\|_M \\ + \alpha_2 \hat{\sigma}(x^{k+1}, x^k),$$

where α_1 and α_2 are constants independent of k ,

$$\begin{aligned}\hat{\sigma}(x^{k+1}, x^k) &= \max\{|x^{k+1} - x^*|, |x^k - x^*|\}, \\ \theta_k &= \frac{|M(B_k - l_{xx}(x^*, \lambda^*))\delta_x^k|}{\|B_k - l_{xx}(x^*, \lambda^*)\|_M |M^{-1}\delta_x^k|}, \\ M &= l_{xx}(x^*, \lambda^*)^{-1/2},\end{aligned}$$

and the M -norm, $\|Q\|_M$, stands for the Frobenius norm of the matrix MQM . The triangle inequality can now be used with (3.9) and (3.10) to establish

$$(3.11) \quad \|B_{k+1} - l_{xx}(x^*, \lambda^*)\|_M \leq \{(1 - c\theta_k^2)^{1/2} + \alpha_3\hat{\sigma}(x^{k+1}, x^k)\}\|B_k - l_{xx}(x^*, \lambda^*)\|_M \\ + \alpha_4\sigma((x^{k+1}, \lambda^{k+1}), (x^k, \lambda^k)),$$

where B_{k+1} is the DFP and BFGS update of B_k and σ is defined as in Lemma 4. Since $\{x^k\}$ converges to x^* (and hence $\{\lambda^k\}$ converges to λ^*) it follows that $\{\|B_k - l_{xx}(x^*, \lambda^*)\|_M\}$ has a limit (Dennis and Moré [3]). If the limit is not zero, then (3.11) implies $\theta_k \rightarrow 0$; if the limit is zero then $\|B_k - l_{xx}(x^*, \lambda^*)\|_M \rightarrow 0$. In either case we have

$$\lim_{k \rightarrow \infty} \frac{|(B_k - l_{xx}(x^*, \lambda^*))\delta_x^k|}{|\delta_x^k|} = 0.$$

Since the projection matrices $P(x^k)$ are bounded, Theorem 3.1 can be applied to establish the Q -superlinear convergence.

Theorem 3.2 rests heavily on the assumption that $l_{xx}(x^*, \lambda^*)$ is positive definite. In theory, $l_{xx}(x^*, \lambda^*)$ need only be positive definite on the null space of $\nabla g(x^*)^T$. Nevertheless, most implementations of the quasi-Newton approach use updates (such as BFGS) which maintain positive definiteness of the B_k (with some ad hoc scheme to assure that $(y^k)^T \delta_x^k$ is positive). It remains an open question as to whether Q -superlinear convergence can be guaranteed with these approaches.

In the previous theorems, linear convergence of the $\{x^k\}$ is assumed. However, if the bounded deterioration inequality (3.11) holds, then linear convergence can be achieved by requiring $|x^0 - x^*|$ and $|B_0 - l_{xx}(x^*, \lambda^*)|$ to be sufficiently small. As shown above, (3.11) holds when $l_{xx}(x^*, \lambda^*)$ is positive definite. Without the positive definite assumptions the usual conditions for linear convergence require that $|B_k - l_{xx}(x^*, \lambda^*)|$ be small for all k . (See Han [6] and Tapia [10] for the relevant results.) In the next theorem we relax this restriction by showing linear convergence under the requirement that $|P(x^k)(B_k - l_{xx}(x^*, \lambda^*))|$ be small for all k . This theorem further illustrates the significance of the projection operator in the quasi-Newton theory for constrained minimization.

THEOREM 3.3. *Let the B_k satisfy*

$$|B_k^{-1}| \leq \eta$$

for some $\eta > 0$. Then there exist positive constants ε and ξ such that if

- (i) $|x^0 - x^*| < \xi$,
- (ii) $|P(x^k)(B_k - l_{xx}(x^*, \lambda^*))| < \varepsilon$ for all $k \geq 0$,

then the sequence $\{x^k\}$ generated by

$$(3.12) \quad x^{k+1} = x^k - B_k^{-1}l_x(x^k, \Lambda_k(x^k)),$$

where

$$(3.13) \quad \Lambda_k(x) = (\nabla g(x)^T B_k^{-1} \nabla g(x))^{-1} (g(x) - \nabla g(x)^T B_k^{-1} \nabla f(x)),$$

is well defined and converges linearly to x^* .

Remark. The iteration (3.12)–(3.13) is equivalent to (1.6)–(1.7), but this form makes the proof easier.

Proof. As demonstrated earlier, it follows from the assumptions that for some $\hat{\xi} > 0$ and $|x - x^*| < \hat{\xi}$, $(\nabla g(x)^T B_k^{-1} \nabla g(x))^{-1}$ exists and is uniformly bounded. Thus, for $|x^0 - x^*| < \hat{\xi}$, x^1 is well defined. Since $\Lambda_k(x^*) = \lambda^*$ for all k , we have

$$\begin{aligned} x^1 - x^* &= x^0 - x^* - B_0^{-1} l_x(x^0, \Lambda_0(x^0)) \\ &= B_0^{-1} \{B_0 - l_{xx}(x^*, \lambda^*) - \nabla g(x^*)^T \nabla \Lambda_0(x^*)^T\} (x^0 - x^*) + h^0(x^0), \end{aligned}$$

where $\nabla \Lambda_0(x^*)$ denotes the Jacobian of Λ_0 at $x = x^*$ and $|h^0(x^0)| \leq \alpha^0 |x^0 - x^*|^2$, α^0 constant. From (3.13) we see that

$$\nabla \Lambda_0(x^*)^T = (\nabla g(x^*)^T B_0^{-1} \nabla g(x^*))^{-1} \nabla g(x^*)^T B_0^{-1} (B_0 - l_{xx}(x^*, \lambda^*)).$$

Therefore,

$$\begin{aligned} |x^1 - x^*| &\leq |B_0^{-1}| \cdot |I - \nabla g(x^*)^T B_0^{-1} \nabla g(x^*)| \cdot |x^0 - x^*| + \alpha^0 |x^0 - x^*|^2 \\ &\quad \cdot |B_0 - l_{xx}(x^*, \lambda^*)|. \end{aligned}$$

Let $V_k^* = I - \nabla g(x^*)^T B_k^{-1} \nabla g(x^*)$ and note that as in (3.2b), $V_k^* P(x^*) = V_k^*$. Thus

$$|x^1 - x^*| \leq |B_0^{-1}| \cdot |V_0^*| \cdot |P(x^*)(B_0 - l_{xx}(x^*, \lambda^*))| \cdot |x^0 - x^*| + \alpha^0 |x^0 - x^*|^2.$$

From our assumptions, it now follows that the $|V_k^*|$ will be uniformly bounded by, say, $\hat{\beta} > 0$. We now choose ε and ξ small enough so that $\eta \hat{\beta} \varepsilon + \alpha^0 \xi \leq \rho < 1$, and therefore, $|x^1 - x^*| \leq \rho |x^0 - x^*|$. The desired result can now be proven by induction since the sequence $\{\alpha^k\}$ can be uniformly bounded.

We observe that in the above theorem, condition (ii) could be replaced by

$$|P(x^k)(B_k - l_{xx}(x^*, \lambda^*))| < \varepsilon,$$

which is consistent with the form in Theorem 3.1.

Acknowledgment. The authors would like to thank R. A. Tapia for his helpful comments.

REFERENCES

- [1] P. T. BOGGS AND J. W. TOLLE, *Augmented Lagrangians which are quadratic in the multiplier*, J. Opt. Theory Appl., 31 (1980), pp. 17–26.
- [2] G. BROYDEN, J. DENNIS AND J. MORÉ, *On the local and superlinear convergence of quasi-Newton methods*, J. Inst. Math. Appl., 12 (1973), pp. 223–246.
- [3] J. DENNIS AND J. MORÉ, *A characterization of superlinear convergence and its application to quasi-Newton Methods*, Math. Comp., 28 (1974), pp. 549–560.
- [4] ———, *Quasi-Newton methods, motivation and theory*, SIAM Review, 19 (1977), pp. 46–89.
- [5] S. GLAD, *Properties of updating methods for the multipliers in augmented Lagrangians*, J. Opt. Theory Appl., 28 (1979), pp. 135–156.
- [6] S. P. HAN, *Dual variable metric algorithms for constrained optimization*, this Journal, 15 (1977), pp. 546–565.
- [7] J. ORTEGA AND W. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

- [8] M. J. D. POWELL, *A fast algorithm for nonlinearly constrained optimization calculations*, 1977 Dundee Conference on Numerical Analysis, June, 1977.
- [9] ———, *The convergence of variable metric methods for nonlinearly constrained optimization calculations*, Nonlinear Programming 3, O. Mangasarian, R. Meyer and S. Robinson, eds., Academic Press, New York, 1978, pp. 27–63.
- [10] R. A. TAPIA, *Diagonalized multiplier methods and quasi-Newton methods for constrained optimization*, J. Opt. Theory Appl., 22 (1977), pp. 135–194.
- [11] ———, *Quasi-Newton methods for equality constrained optimization: Equivalence of existing methods and a new implementation*, Nonlinear Programming 3, O. Mangasarian, R. Meyer and S. Robinson, eds., Academic Press, New York, 1978, pp. 125–164.

ON MATRIX FRACTION REPRESENTATIONS FOR LINEAR SYSTEMS OVER COMMUTATIVE RINGS*

P. P. KHARGONEKAR†

Abstract. This paper deals with the problem of existence of polynomial matrix fraction representations for transfer matrices of linear systems over rings as well as the related realization theory. These representations are then used in establishing new results for various classes of systems including split systems. The relevance of these results to the regulation of various “nonclassical” classes of linear systems (for example, delay-differential systems, two-dimensional systems, etc.) is also discussed.

1. Introduction. Linear systems over arbitrary fields (and in particular, over the field of real numbers) have been investigated extensively over the last 20 years. As a natural generalization of linear systems over fields, linear systems over commutative rings were first considered by Rouchaleau [1972] and Rouchaleau, Wyman and Kalman [1972]. It was shown by Kamen [1975] that a large class of infinite dimensional continuous-time systems (including delay-differential systems) can be represented as vector differential equations over a ring of operators. Some other applications of linear systems over commutative rings are in coding theory (see Johnston [1973]), two-dimensional digital systems viewed as linear systems over the ring of proper rational functions (see Sontag [1978a], Eising [1978], and Eising [1979]), families of systems (see Byrnes [1979], Hazewinkel [1980]), discretized partial differential equations (see Brockett and Willems [1974] and Sontag [1976]), large scale systems (see Kamen [1978]), etc. The reader is referred to Eilenberg [1974, Chapt. 16], Sontag [1976], Kamen [1978] and the references given there for further details of linear systems over commutative rings. The reader is referred to Byrnes [1978] for an algebro-geometric approach to a class of linear systems over rings.

Consider for example a delay-differential system Σ defined by

$$\begin{aligned}\dot{x}_1(t) &= 2x_2(t-1) + u(t), \\ \dot{x}_2(t) &= x_1(t-1) + x_2(t-\pi) + u(t-1), \\ y(t) &= x_1(t).\end{aligned}$$

If we introduce the delay operators σ_1 and σ_2 defined by

$$\sigma_1(x)(t) := x(t-1), \quad \sigma_2(x)(t) := x(t-\pi),$$

we can rewrite the delay-differential system Σ in matrix form as

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 2\sigma_1 \\ \sigma_1 & \sigma_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ \sigma_1 \end{bmatrix} u, \quad y = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

We thus see that the delay-differential system Σ can be expressed in a form very similar to the ordinary finite dimensional constant linear systems, except that the entries of the coefficient matrices belong to the ring of polynomials $\mathbb{R}[\sigma_1, \sigma_2]$. The reader is referred to Sontag [1976] and Kamen [1978] for several other examples.

* Received by the editors August 18, 1980, and in final form April 18, 1981.

† Center for Mathematical System Theory, University of Florida, Gainesville, Florida 32611. This research was supported in part by the U.S. Army Research Office under grant DAAAG29-80 C-0050 and the U.S. Air Force under grant AFOSR 76-3034 Mod. D through the Center for Mathematical System Theory, University of Florida.

It is well known that the polynomial matrix approach to linear systems is very useful in studying several system and control theoretic problems such as realization theory, dynamic compensation, output regulation in the presence of disturbances, etc. The reader is referred to Rosenbrock [1970], Wolovich [1974], Fuhrmann [1976], Rosenbrock and Hayton [1977], Cheng and Pearson [1978], Antsaklis [1979], Desoer et al. [1979], Emre [1980a], Emre and Silverman [1981] and the references given there for applications of polynomial matrix methods to various system and control theoretic problems.

In this paper we consider, for the first time, the existence and the realization theory of matrix fraction representations for linear systems over commutative rings. In § 2 we give definitions and some preliminary results that we will need in the later sections. In § 3 we obtain polynomial matrix representations for finitely generated projective modules with endomorphisms. This also leads to a natural definition of characteristic polynomial of an endomorphism of a finitely generated projective module over an arbitrary commutative ring. Our results generalize several results obtained by Fuhrmann [1976] and [1977] for fields, to arbitrary commutative rings.

In § 4 we establish a correspondence between realizations of an input-output map with a finitely generated projective state module (referred to as *projective realizations*) and matrix fractions representations of the transfer matrix in the form $PQ^{-1}R$. We also obtain conditions for strict system equivalence which are similar to those obtained by Fuhrmann [1977]. We then establish a one-to-one correspondence between reachable projective realizations (modulo state module isomorphism) of an input-output map and matrix fraction representations of the transfer matrix of the form PQ^{-1} (modulo multiplying P and Q on the right by unimodular matrices). This result generalizes a similar result in Hautus and Heymann [1978, Remark 4.8] proved for the field case.

It is well known that if f is a realizable input-output map over a field, and Z_f is the transfer matrix of f , then there exist right coprime polynomial matrices P and Q , and left coprime polynomial matrices Q_1 and R_1 such that

$$(1.1) \quad Z_f = PQ^{-1} = Q_1^{-1}R_1.$$

Also, in the field case, the coprimeness conditions on P , Q and Q_1 , R_1 are well known to be equivalent to the existence of polynomial matrices Y_1 , Y_2 , Y_3 , and Y_4 such that

$$(1.2) \quad Y_1P + Y_2Q = I,$$

and

$$(1.3) \quad Q_1Y_4 + R_1Y_3 = I.$$

However, for polynomial matrices over an arbitrary commutative ring coprimeness conditions and (1.2), (1.3) are not equivalent in general. Two polynomial matrices P and Q (Q_1 and R_1) are said to be *right (left) Bezout* if and only if there exist polynomial matrices Y_1 and Y_2 (Y_3 and Y_4) such that (1.2) ((1.3)) is satisfied. A factorization $Z_f = PQ^{-1}$ ($Z_f = Q_1^{-1}R_1$) where P and Q are right Bezout (Q_1 and R_1 are left Bezout) will be called a *right (left) Bezout factorization*.

In § 5, based on some results developed by Sontag [1978], we will establish a necessary and sufficient condition on an input-output map f for the associated transfer matrix Z_f to have a right or left Bezout factorization. In particular, it will be seen that the following conditions are equivalent:

- 1) Z_f admits a left Bezout factorization.
- 2) Z_f admits a right Bezout factorization.
- 3) f is a split input-output map. (See § 5 for the definition of a split input-output map.)

We will also show how the existence of Bezout factorizations can be checked using some tests on the associated behavior matrix of the input-output map f . Existence of Bezout factorization has been assumed in approaches to some control theoretic problems. (See, for example, Antsaklis [1979], Desoer et al. [1980], Emre [1980b].) We remark, however, that the dynamic compensators obtained by Antsaklis [1979] and Desoer et al. [1980] are not necessarily proper. For finite-free split systems, (i.e., the canonical state module is finitely generated and free and the system is split), over arbitrary commutative rings, proper dynamic compensators can be obtained using the results of Emre and Khargonekar [1980]. (In fact, the results in Emre and Khargonekar [1980] can be used for detectable and reachable systems as well. For the concept of detectability, see Hautus and Sontag [1980].) Our results on existence of Bezout factorizations establish a clear connection between the approaches taken by Emre and Khargonekar [1980] and Antsaklis [1979] and Desoer et al. [1980].

In this paper we restrict our attention to polynomial matrix fraction representations. For the case of matrix fraction representations over the rings of stable rational functions and stable proper rational functions, the reader is referred to Khargonekar and Sontag [1981].

The framework developed in this paper can now be used to attack some of the control problems for linear systems over rings. In fact, some of the results of § 5 have been used by Emre [1980b] in an approach to the output regulation problem in the presence of disturbances for linear systems over commutative rings.

The split condition given in this paper is not very restrictive in general. For example, in the case of delay-differential systems, recent results of Lee and Olbort [1980] indicate that the system is generically split if the number of input and output channels each exceed the number of noncommensurate delays.

2. Definitions and preliminary results. In this section we will give some definitions, notation, and preliminary results that we will need in the later sections.

Let A denote an arbitrary but fixed commutative ring with identity. We will denote the tensor product of A -modules by \otimes , and direct sum of A -modules by \oplus . If A is an integral domain then K will denote the quotient field of A . Let $A[z]$ denote the ring of polynomials in the indeterminate z , and let $A(z)$ denote the ring of formal Laurent series in z^{-1} , with coefficients in A (i.e., the elements of $A(z)$ are of the form $\sum_{j=-n}^{\infty} a_j z^{-j}$, where a_j is in A for each j). Similarly, for an A -module X , let $X[z]$ denote the natural $A[z]$ -module of polynomials in z with coefficients in X , and let $X(z)$ denote the natural $A(z)$ -module of formal Laurent series in z^{-1} with coefficients in X . Define a projection map

$$\pi: X(z) \rightarrow X(z): \sum_{j=-n}^{\infty} x_j z^{-j} \mapsto \sum_{j=1}^{\infty} x_j z^{-j}.$$

We also denote $\pi(x)$ by $(x)_-$ for any x in $X(z)$. Define the set

$$z^{-1}X[[z^{-1}]] := \pi(X(z)).$$

For an element x in $X(z)$, x is said to be *strictly proper* if and only if $\pi(x) = x$, $\pi(x)$ is called the *strictly proper part* of x , and x is said to be *proper* if and only if x is of the form $x = \sum_{j=0}^{\infty} x_j z^{-j}$. Define $(x)_n$ to be the coefficient of z^{-n} in the formal Laurent series representation of x in $X(z)$.

If p in $A[z]$ is such that the leading coefficient of p is invertible in A , then we define p^{-1} to be the (unique) formal Laurent series q in $A(z)$ such that $pq = 1$. (q is unique since the leading coefficient of p is invertible in A .) p^{-1} can be obtained, for example, by

carrying our formal division of 1 by p . For x in $X(z)$ and p in $A[z]$ with invertible leading coefficient, we define $p^{-1}x := qx$, where q in $A(z)$ is such that $pq = 1$.

Let S be a set. We denote by S^p and $S^{p \times m}$ respectively the set of p length column vectors, and the set of $p \times m$ matrices with entries in S . If $f: S^p \rightarrow T$ is a function, and Q is in $S^{p \times m}$, then $f(Q)$ denotes the m length row vector with $f(q_i)$ as its i th entry, where q_i is the i th column of Q .

DEFINITION 2.1. An m input, p output linear, time-invariant, causal input-output map f is an $A[z]$ -module homomorphism

$$f: A^m(z) \rightarrow A^p(z),$$

such that for $u = \sum_{j=-n}^{\infty} u_j z^{-j}$ in $A^m(z)$

$$(f(u))_k = 0, \quad k \leq -n.$$

Let I_m be the $m \times m$ identity matrix. Then the $p \times m$ strictly proper matrix

$$Z_f := f(I_m) = \sum_{j=1}^{\infty} A_j z^{-j}$$

is called the *transfer matrix* associated with f , and $(A_i)_{i=1}^{\infty}$ is called the *impulse response sequence* of f . A linear system $\Sigma = (F, G, H, X)$ over A consists of a finitely generated A -module X , and A -module homomorphisms $F: X \rightarrow X$, $G: A^m \rightarrow X$, and $H: X \rightarrow A^p$. A system $\Sigma = (F, G, H, X)$ is called a *realization* of an input-output map f if and only if

$$A_i = HF^{i-1}G,$$

for all positive integers i . An input-output map f is said to be *realizable* if and only if there exists a least one realization of f .

By a *discrete-time dynamical interpretation* of a system $\Sigma = (F, G, H, X)$ we mean the equations:

$$x_{t+1} = Fx_t + Gu_t, \quad y_t = Hx_t,$$

where t is an integer, the states $x(t)$ are in X , the inputs $u(t)$ are in A^m and the outputs $y(t)$ are in A^p . For further details concerning these definitions, the reader is referred to Kalman, Falb, and Arbib [1969, Chapt. 10], Eilenberg [1974, Chapt. 16], and Wyman [1972].

Let f be a given input-output map with the transfer matrix Z_f . The restricted input-output map induced by f is the $A[z]$ -module homomorphism

$$f^*: A^m[z] \rightarrow z^{-1}A^p[[z^{-1}]]: u \mapsto \pi(Z_f u).$$

Let $\Sigma = (F, G, H, X)$ be a realization of f . Define an $A[z]$ -module structure on X by $z \cdot x := Fx$, for all x in X . Further, if $A[z]$ -module homomorphisms g and h are defined as

$$g: A^m[z] \rightarrow X: \sum_{t=0}^n u_t z^t \mapsto \sum_{t=0}^n F^t G u_t,$$

$$h: X \rightarrow z^{-1}A^p[[z^{-1}]]: x \mapsto \sum_{t=1}^{\infty} HF^{t-1} x z^{-t},$$

then the $A[z]$ -module homomorphism f^* can be written as $f^* = hg$. Conversely, given an $A[z]$ -module X and $A[z]$ -module homomorphisms

$$g: A^m[z] \rightarrow X, \quad h: X \rightarrow z^{-1}A^p[[z^{-1}]]$$

such that $f^* = hg$, then we can obtain a realization $\Sigma = (F, G, H, X)$ of f by defining

$$F: X \rightarrow X: x \mapsto z \cdot x, \quad G: A^m \rightarrow X: u \mapsto g(u), \quad H: X \rightarrow A^p: x \mapsto (h(x))_{-1}.$$

Thus, there is a one-to-one correspondence between realizations of f and $A[z]$ -module factorizations of the map f^* as $f^* = hg$. A system $\Sigma = (F, G, H, X)$ is said to be *reachable* if and only if g is surjective, is said to be *observable* if and only if h is one-to-one, and is said to be *canonical* if and only if it is reachable and observable.

We will now give some definitions and results that will be useful in the following sections for examining matrix fraction representations for linear systems over commutative rings.

Let Q be an $r \times r$ polynomial matrix over A . (i.e., Q is in $A^{r \times r}[z]$.) Q is said to be *admissible* if and only if there exists a polynomial s and a monic polynomial p such that

$$(2.2) \quad p = |Q|s,$$

where $|Q|$ denotes the determinant of Q . Note that $|Q|$ is not a zero divisor since p is monic. In particular, if A is an integral domain then Q is admissible if and only if the leading coefficient of $|Q|$ is invertible in A . It is obvious that if A is a field then Q is admissible if and only if Q is nonsingular. For an admissible $r \times r$ matrix Q , we define

$$(2.3) \quad Q^{-1} := p^{-1}s \operatorname{adj} Q,$$

where p and s are as in (2.2) and $\operatorname{adj} Q$ denotes the adjoint of Q . Note that Q^{-1} is well defined, since if p_1, p_2 are monic polynomials and s_1, s_2 are polynomials such that

$$p_1 = |Q|s_1, \quad p_2 = |Q|s_2,$$

then

$$|Q|(p_1^{-1}s_1 - p_2^{-1}s_2) = 0.$$

As $|Q|$ is not a zero divisor, it follows that $p_1^{-1}s_1 = p_2^{-1}s_2$. Thus for an admissible polynomial matrix Q , Q^{-1} is well defined. It is easy to verify using classical formulas concerning adjoints that

$$Q^{-1}Q = QQ^{-1} = I.$$

It is not difficult to prove that Q is invertible as a formal Laurent series if and only if it is admissible.

Let Q be an $r \times r$ admissible polynomial matrix. Define an A -module

$$A_Q := \{x \text{ in } A^r[z]: Q^{-1}x \text{ is strictly proper}\}.$$

Define an A -linear projection map

$$\pi_Q: A^r[z] \rightarrow A_Q: x \mapsto Q\pi(Q^{-1}x).$$

Clearly, π_Q is surjective. Further, if p is a monic polynomial such that $Q^{-1} = p^{-1}S$ for some polynomial matrix S , then

$$\pi_Q(px) = Q\pi(Q^{-1}px) = Q\pi(Sx) = 0.$$

Let n be the degree of p . Let e_i be the i th column of the $r \times r$ identity matrix. Then it is easy to verify that a set of generators for the A -module A_Q is given by

$$\{\pi_Q(z^j e_i): j = 0, 1, \dots, n-1; i = 1, 2, \dots, r\}.$$

Thus, A_Q is a finitely generated A -module.

Let f be an input-output map and P, Q, R be polynomial matrices such that Q is admissible, and

$$Z_f = PQ^{-1}R.$$

The following lemma, first proved in Fuhrmann [1976] for the case $A = \text{a field}$, gives a natural realization of f in terms of the polynomial matrices P, Q , and R . It can be easily checked that the proof given by Fuhrmann [1976] remains valid for arbitrary commutative rings with identity. The reader is also referred to Emre [1980c] for a simple proof of this result.

LEMMA 2.4. *Let P, Q and R be $p \times r, r \times r$, and $r \times m$ polynomial matrices such that Q is admissible. Let f be an input-output map such that the associated transfer matrix*

$$Z_f = PQ^{-1}R.$$

Then $\Sigma_Q = (F_Q, G_Q, H_Q, A_Q)$ with

$$F_Q: A_Q \rightarrow A_Q: x \mapsto \pi_Q(zx),$$

$$G_Q: A^m \rightarrow A_Q: u \mapsto \pi_Q(Ru),$$

$$H_Q: A_Q \rightarrow A^p: x \mapsto (PQ^{-1}x)_{-1}$$

is a realization of f . Further, Σ_Q corresponds to the factorization

$$f^* = h_Q g_Q$$

of the restricted input-output map, where for x in A_Q and u in $A^m[z]$ we have

$$z \cdot x = \pi_Q(zx), \quad g_Q(u) = \pi_Q(Ru), \quad h_Q(x) = \pi(PQ^{-1}x).$$

Finally, we have some standard definitions from commutative algebra. For details concerning these definitions the reader is referred to Rotman [1968], Knight [1971], and Bourbaki [1972].

DEFINITION 2.5. An A -module X is said to be *projective* if and only if one of the following (equivalent) conditions holds:

(i) If Z is an A -module and $\chi: Z \rightarrow X$ is a surjective A -module homomorphism then there exists an A -module homomorphism $\tau: X \rightarrow Z$ such that $\chi\tau$ is the identity map on X .

(ii) There exists an A -module Y such that $X \oplus Y$ is a free A -module.

A system $\Sigma = (F, G, H, X)$ is said to be a *free (projective) system* if the state module X is a free (projective) A -module.

DEFINITION 2.6. Let X be an A -module. The *dual* X' of X is the A -module consisting of all A -module homomorphisms $X \rightarrow A$. If $\chi: X \rightarrow Y$ is an A -module homomorphism then $\chi': Y' \rightarrow X'$ is the A -module homomorphism given by $\chi': \tau \mapsto \tau\chi$ for each $\tau: Y \rightarrow A$. χ' is called the *dual* of χ . For a system $\Sigma = (F, G, H, X)$, the dual system Σ' is defined as $\Sigma' := (F', H', G', X')$.

Let $\beta: A^m \rightarrow A^p$ be an A -linear map. Then there exists a unique $p \times m$ matrix B such that

$$\beta: A^m \rightarrow A^p: x \mapsto Bx.$$

It is well known (see, e.g., MacLane and Birkhoff [1967]) that β' can be identified with the map

$$\beta': A^p \rightarrow A^m: x \mapsto B'x,$$

where B' represents the transpose of B . Now, if $\Sigma = (F, G, H, X)$ is a given system with

input-output map f , then the input-output map of $\Sigma' = (F', G', H', X')$ is denoted by f' . It is not difficult to see that

$$Z_{f'} = Z'_f.$$

3. Polynomial representations for finitely generated projective modules. In this section we will establish polynomial representations for finitely generated projective modules with an endomorphism. In particular, we will show that for any finitely generated projective module X with an endomorphism $F: X \rightarrow X$, there exists an admissible polynomial matrix Q and an A -module isomorphism $\chi: A_Q \rightarrow X$, such that $\chi F_Q = F\chi$. (i.e., χ is an intertwining map.) We will also obtain necessary and sufficient conditions on the given admissible polynomial matrices Q and \hat{Q} for A_Q and $A_{\hat{Q}}$ to be isomorphic. Our results also lead to a natural definition of the characteristic polynomial of an endomorphism of a finitely generated projective module. These results will be used in the following sections in examining matrix fraction representations for linear systems over commutative rings.

The following simple lemmas will be useful in establishing the main results of this section.

LEMMA 3.1. *Let X be a finitely generated projective A -module. Then $X[z]$ is a finitely generated projective $A[z]$ -module.*

Proof. Since X is a finitely generated projective A -module, there exists a free A -module W of rank r , and an A -module Y such that $W = X \oplus Y$. Since W is a free module of rank r , $W[z]$ is a free $A[z]$ -module of rank r . Further, it is not difficult to see that

$$W[z] = X[z] \oplus Y[z].$$

Hence $X[z]$ is a finitely generated projective $A[z]$ -module. \square

LEMMA 3.2. *Let Q be an $r \times r$ admissible polynomial matrix. Then*

$$A'[z] = A_Q \oplus QA'[z]$$

as A -modules.

Proof. For any u in $A'[z]$, there exists a unique p in $A'[z]$ and q in $z^{-1}A'[[z^{-1}]]$ such that

$$Q^{-1}u = p + q.$$

Since $Qq = u - Qp$ is polynomial, it follows that Qq is in A_Q . Hence,

$$u = Qp + Qq$$

is in $A_Q + QA'[z]$. Further, if u in $A'[z]$ is in A_Q and $QA'[z]$, then there exists p in $A'[z]$ and q in $z^{-1}A'[[z^{-1}]]$ such that

$$u = Qp = Qq.$$

As Q is admissible, $u = 0$. Thus

$$A'[z] = A_Q \oplus QA'[z]. \quad \square$$

The following theorem is one of the fundamental results of this paper.

THEOREM 3.3. *Let X be a finitely generated projective A -module with an endomorphism $F: X \rightarrow X$. Let*

$$g: A'[z] \rightarrow X: \sum_{t=0}^n u_t z^t \mapsto \sum_{t=0}^n F^t g(u_t)$$

be a surjective A -linear map. Then there exists an $r \times r$ admissible polynomial matrix Q

such that

$$\ker g = QA'[z].$$

Further, if χ is the restriction of g to A_Q then $\chi: A_Q \rightarrow X$ is an A -module isomorphism such that

$$\chi F_Q = F\chi.$$

Proof. Let g_1 be the $A[z]$ -module homomorphism

$$g_1: A'[z] \rightarrow X[z]: \sum_{t=0}^n u_t z^t \mapsto \sum_{t=0}^n z^t g(u_t).$$

Consider the $A[z]$ -module homomorphism

$$\psi: A'[z] \oplus X[z] \rightarrow X[z]: u + x \mapsto zx - F(x) - g_1(u),$$

where

$$F\left(\sum_{t=0}^n x_t z^t\right) = \sum_{t=0}^n z^t F(x_t).$$

We will first prove that ψ is surjective. For this, it is enough to prove that X is contained in the image of ψ . Since g is surjective, for any x in X , there exists $u = u_n z^n + \cdots + u_0$ in $A'[z]$ such that

$$x = g(u) = \sum_{t=0}^n F^t g(u_t).$$

We now have

$$\psi(-u + 0) = \sum_{t=0}^n z^t g(u_t) = x + \sum_{t=0}^n (z^t - F^t) g(u_t).$$

For any $j \geq 1$ we also have

$$(z^j - F^j)g(u_j) = (z - F)(z^{j-1} + z^{j-2}F + \cdots + F^{j-1})g(u_j) = zx_j - Fx_j,$$

where

$$x_j = (z^{j-1} + z^{j-2}F + \cdots + F^{j-1})g(u_j).$$

Let us define

$$\hat{x} := -(x_1 + x_2 + \cdots + x_n).$$

It now follows that

$$\psi(-u + \hat{x}) + \sum_{t=0}^n (z^t - F^t)g(u_t) + (z - F)\hat{x} = x = g(u).$$

Thus ψ is surjective. By Lemma 3.1, $X[z]$ is a finitely generated projective $A[z]$ -module. It follows that $A'[z] \oplus X[z]$ is $A[z]$ -module isomorphic to

$$\ker \psi \oplus X[z].$$

We will now show that $\ker \psi$ and $\ker g$ are $A[z]$ -module isomorphic. Consider the $A[z]$ -module homomorphism

$$\theta: \ker \psi \rightarrow A'[z]: u + x \mapsto u.$$

To show that θ is one-to-one, it is enough to show that if $0+x$ belongs to $\ker \theta$ then $x=0$. Now, if

$$\psi(0+x) = zx - F(x) = 0,$$

then by comparing coefficients it is easy to see that $x=0$. So, we need to show that image of θ is $\ker g$. If u is in $\ker g$, as in the proof of surjectivity of ψ we can find \hat{x} in $X[z]$ such that

$$\psi(u + \hat{x}) = g(-u) = 0,$$

and

$$\theta(u + \hat{x}) = u.$$

Thus $\ker g$ is contained in image of θ . Conversely, let $u = u_n z^n + \cdots + u_0$ in $A^r[z]$ and $x = x_m z^m + \cdots + x_0$ in $X[z]$ be such that

$$\psi(u+x) = zx - F(x) - g_1(u) = z \sum_{t=0}^n x_t z^t - F\left(\sum_{t=0}^n x_t z^t\right) - g_1\left(\sum_{t=0}^m u_t z^t\right) = 0.$$

Since $\psi(u+x)$ is the zero polynomial in $X[z]$, it is easy to see that

$$F \sum_{t=0}^n F^t x_t - F \sum_{t=0}^n F^t x_t - \sum_{t=0}^m F^t g(u_t) = \sum_{t=0}^m F^t g(u_t) = 0.$$

Therefore,

$$g(u) = \sum_{t=0}^m F^t g(u_t) = 0.$$

Hence $\text{im } \theta = \ker g$.

It now follows that $\ker g$ and $\ker \psi$ are $A[z]$ -module isomorphic, and $A^r[z] \oplus X[z]$ and $\ker g \oplus X[z]$ are $A[z]$ -module isomorphic. Since $X[z]$ is finitely generated, projective $A[z]$ -module, there exists an $A[z]$ -module Y such that

$$X[z] \oplus Y = A^n[z]$$

for some n . Therefore, $A^r[z] \oplus X[z] \oplus Y$ is $A[z]$ -module isomorphic to $\ker g \oplus X[z] \oplus Y$. Consequently, $\ker g \oplus A^n[z]$ and $A^r[z] \oplus A^n[z]$ are $A[z]$ -module isomorphic. Hence, $\ker g$ is a stably free (and, of course, projective) $A[z]$ -module. As $\ker g$ is a quotient of a finitely generated module it must be finitely generated. Thus $\ker g$ is a finitely generated projective $A[z]$ -module.

We will now show that $\ker g$ is a free $A[z]$ -module. Since X is a finitely generated A -module, it follows from Cayley-Hamilton theorem (see Atiyah and MacDonald [1969, Chapt. 2, Prop. 2.3]) that there exists a monic polynomial

$$\alpha = z^n + \alpha_{n-1} z^{n-1} + \cdots + \alpha_0,$$

such that

$$(F^n + \alpha_{n-1} F^{n-1} + \cdots + \alpha_0)x = 0$$

for all x in X . If e_i represents the i th column of the $r \times r$ identity matrix then αe_i belongs to $\ker g$. Let $A_R(z)$ denote the localization of $A[z]$ at the multiplicative set consisting of all monic polynomials in $A[z]$ (i.e., $A_R(z)$ is the ring of *rational Laurent series* with coefficients in A). Now, since αe_i belongs to $\ker g$, and α is a monic polynomial, e_i belongs to $A_R(z) \otimes \ker g$ for each i . Hence, $A_R(z) \otimes \ker g$ is isomorphic to $A_R^r(z)$. As

$\ker g$ is a finitely generated projective module, and since $A_R(z) \otimes \ker g$ is isomorphic to $A'_R(z)$, and hence is a free $A_R(z)$ -module, it follows from the global affine Horrocks theorem (see Lam [1978, Chapt. 5, Suppl. 2.3]) that $\ker g$ is a free $A[z]$ -module. Further, as $A_R(z) \otimes \ker g$ is isomorphic to $A'_R(z)$, it follows by the invariant basis property of commutative rings (see Lam [1978, p. 25]) that $\ker g$ is isomorphic to $A'[z]$. Thus $\ker g$ has r generators. Let q_1, q_2, \dots, q_r in $A'[z]$ be a set of generators of $\ker g$. Let Q be the $r \times r$ polynomial matrix with q_i as its i th column. Then we have

$$\ker g = QA'[z].$$

We will now show that Q is admissible. Since αe_i belongs to $\ker g$ for each i , there exists an $r \times r$ polynomial matrix S such that $\alpha I_r = QS$, where I_r is the $r \times r$ identity matrix. Multiplying both sides by $\text{adj } Q$ we get

$$\alpha \text{adj } Q = |Q|S.$$

Thus, Q is admissible.

Since g is surjective, $\ker g = QA'[z]$, and by Lemma (3.2)

$$A'[z] = A_Q \oplus QA'[z],$$

it follows that

$$\chi(A_Q) = X.$$

Also, as $A_Q \cap QA'[z] = 0$, χ is one-to-one. Finally, for any u in A_Q we have

$$F_Q(u) = \pi(zu) = Q\pi(Q^{-1}zu) = (zu - Qc)$$

for some c in $A'[z]$. Therefore

$$(\chi F_Q)(u) = \chi(zu - Qc) = g(zu - Qc) = g(zu) = (F\chi)(u).$$

Thus $\chi F_Q = F\chi$. \square

The following result is an immediate consequence of Theorem 3.3. This result provides polynomial representations for arbitrary finitely generated projective modules with an endomorphism. This result will be very useful in developing existence and realization theory for matrix fraction representations for linear systems over commutative rings.

THEOREM 3.4. *Let X be a finitely generated projective A -module with an endomorphism $F: X \rightarrow X$. Then there exists an integer r , an $r \times r$ admissible polynomial matrix Q , and an A -module isomorphism $\chi: A_Q \rightarrow X$ such that $\chi F_Q = F\chi$. Conversely, if Q is an $r \times r$ admissible polynomial matrix then A_Q is a finitely generated projective A -module.*

Proof. Since X is finitely generated, there exists an integer r and a surjective A -module homomorphism

$$g: A^r[z] \rightarrow X: \sum_{t=0}^n u_t z^t \mapsto \sum_{t=0}^n F^t g(u_t).$$

Then by Theorem 3.3 there exists an $r \times r$ admissible polynomial matrix Q and an A -module isomorphism $\chi: A_Q \rightarrow X$ such that

$$\chi F_Q = F\chi.$$

Conversely, for any $r \times r$ admissible polynomial matrix Q ,

$$A_Q \oplus QA'[z] = A'[z].$$

As $A'[z]$ is a free A -module, A_Q is a projective module. Also, from the remarks made in § 2, A_Q is finitely generated. \square

Remark 3.5. Let us define an $A[z]$ -module structure on X by $z \cdot x = Fx$, for all x in X . Similarly, let an $A[z]$ -module structure on A_Q be defined by

$$z \cdot u = F_Q(u) = \pi_Q(zu).$$

Then $\chi: A_Q \rightarrow X$ is an $A[z]$ -module isomorphism. Also, with this module structure on A_Q , it follows that

$$\pi_Q: A'[z] \rightarrow A_Q: u \mapsto \pi_Q(u)$$

is a surjective $A[z]$ -module homomorphism with $QA'[z]$ as its kernel. Thus A_Q is $A[z]$ -module isomorphic to the quotient module $A'[z]/QA'[z]$, and the induced map

$$\hat{\pi}: \frac{A'[z]}{QA'[z]} \rightarrow A_Q: u + QA'[z] \mapsto \pi_Q(u)$$

is an $A[z]$ -module isomorphism. In what follows, we will assume that X and A_Q have the $A[z]$ -module structure mentioned above. The reader is referred to Fuhrmann [1976] for further details in the field case.

Let Q and \hat{Q} be any $r \times r$ and $l \times l$ admissible polynomial matrices. We will now give necessary and sufficient conditions on Q and \hat{Q} for A_Q and $A_{\hat{Q}}$ to be $A[z]$ -module isomorphic. In case A is a field, results on this problem were obtained by Fuhrmann [1976, Thm. 4.7]. It will be seen that our results constitute a natural generalization of the results obtained by Fuhrmann [1976]. We remark that certain polynomial matrix equations that arise here also arise in certain system and control theoretic problems. Our results indicate that there is a close relation between certain control theoretic problems and $A[z]$ -module isomorphisms. (See Remark 3.17 and § 5.)

THEOREM 3.6. *Let Q and \hat{Q} be $r \times r$ and $l \times l$ admissible polynomial matrices. A map $\chi: A_Q \rightarrow A_{\hat{Q}}$ is an $A[z]$ -module isomorphism if and only if there exists $l \times r$ polynomial matrices C and D , and there exist polynomial matrices Y_1, Y_2, Y_3 , and Y_4 such that for any f in A_Q*

$$(3.7) \quad \chi(f) = \pi_{\hat{Q}}(Cf),$$

$$(3.8) \quad CQ = \hat{Q}D,$$

$$(3.9) \quad CY_1 + \hat{Q}Y_2 = I,$$

$$(3.10) \quad Y_3D + Y_4Q = I.$$

Proof. Suppose that $\chi: A_Q \rightarrow A_{\hat{Q}}$ is an $A[z]$ -module isomorphism. It follows from Remark 3.5 that χ induces an $A[z]$ -module isomorphism

$$\hat{\chi}: \frac{A'[z]}{QA'[z]} \rightarrow \frac{A'^l[z]}{\hat{Q}A'^l[z]},$$

such that $\chi(f) = \pi_{\hat{Q}}(g)$, where g is such that

$$\hat{\chi}(f + QA'[z]) = g + \hat{Q}A'^l[z].$$

Let e_i be the i th column of I_r and let c_i in $A'^l[z]$ be such that

$$\hat{\chi}(e_i + QA'[z]) = c_i + \hat{Q}A'^l[z].$$

Let C be the $l \times r$ polynomial matrix with c_i as its i th column. Then we have

$$\hat{\chi}(f + QA'[z]) = Cf + \hat{Q}A'[z]$$

for f in $A'[z]$. Consequently, for any f in A_O

$$(3.7) \quad \chi(f) = \pi\hat{Q}(Cf).$$

Let \hat{e}_i be the i th column of the $l \times l$ identity matrix. Since $\hat{\chi}$ is surjective, there exists y_i in $A'[z]$ such that

$$\hat{\chi}(y_i + QA'[z]) = \hat{e}_i + \hat{Q}A'[z].$$

But we also have

$$\hat{\chi}(y_i + QA'[z]) = Cy_i + \hat{Q}A'[z].$$

Let Y_1 be the $r \times l$ polynomial matrix with y_i as its i th column. Then there exists a polynomial matrix Y_2 such that

$$(3.9) \quad CY_1 + \hat{Q}Y_2 = I.$$

We will now show the existence of D such that (3.8) holds. Let q_i be the i th column of Q . Then

$$\hat{\chi}(q_i + QA'[z]) = 0.$$

But we also have

$$\hat{\chi}(q_i + QA'[z]) = Cq_i + \hat{Q}A'[z].$$

Therefore, there exists a polynomial matrix D such that

$$(3.8) \quad CQ = \hat{Q}D.$$

Now, as $\hat{\chi}$ is an isomorphism, $\hat{\chi}^{-1}$ is also an $A[z]$ -module isomorphism. Therefore

$$\hat{\chi}^{-1}(\hat{e}_i + \hat{Q}A'[z]) = y_i + QA'[z].$$

Consequently, for any f in $A'[z]$, we have

$$\hat{\chi}^{-1}(f + \hat{Q}A'[z]) = Y_1f + QA'[z].$$

Further as

$$\hat{\chi}(e_i + QA'[z]) = c_i + \hat{Q}A'[z],$$

it follows that

$$\hat{\chi}^{-1}(c_i + \hat{Q}A'[z]) = Y_1c_i + QA'[z] = e_i + QA'[z].$$

Therefore, there exists a polynomial matrix Y_4 such that

$$(3.11) \quad Y_1C + QY_4 = I.$$

Furthermore, if \hat{q}_i represents the i th column of \hat{Q} then we have

$$\hat{\chi}^{-1}(\hat{q}_i + \hat{Q}A'[z]) = 0 + QA'[z],$$

and also

$$\hat{\chi}^{-1}(\hat{q}_i + \hat{Q}A'[z]) = Y_1\hat{q}_i + QA'[z].$$

Consequently, there exists a polynomial matrix Y_3 such that

$$(3.12) \quad Y_1\hat{Q} = QY_3.$$

Multiplying both sides of (3.8) by Y_1 , we get

$$Y_1 C Q = Y_1 \hat{Q} D.$$

Now using (3.11) and (3.12), it follows that

$$Q Y_3 D = Y_1 C Q = Q - Q Y_4 Q.$$

As Q is admissible, it is clear that

$$(3.13) \quad Y_3 D + Y_4 Q = I.$$

Conversely, suppose (3.7), (3.8), (3.9), and (3.10) hold. It is clear that χ is an $A[z]$ -module homomorphism. We will first prove that χ is surjective. Let f be in $A_{\hat{Q}}$. Now using (3.7) we have

$$C Y_1 f + \hat{Q} Y_2 f = f.$$

Let f_1, f_2 in $A'[z]$ be such that

$$\pi_Q(Y_1 f) = f_1 = Y_1 f - Q f_2.$$

Consequently

$$\chi(f_1) = \pi_{\hat{Q}}(C f_1) = \pi_{\hat{Q}}(C Y_1 f - C Q f_2) = \pi_{\hat{Q}}(f - \hat{Q} Y_2 f - \hat{Q} D f_2) = f.$$

Thus χ is surjective.

To show that χ is one-to-one, we will first show that there exists a polynomial matrix Y_5 such that

$$Y_1 \hat{Q} = Q Y_5.$$

Equations (3.8), (3.9), and (3.10) can be rewritten as

$$\begin{bmatrix} C & \hat{Q} \\ Y_3 & -Y_4 \end{bmatrix} \begin{bmatrix} Y_1 & Q \\ Y_2 & -D \end{bmatrix} = \begin{bmatrix} I & 0 \\ Y_6 & I \end{bmatrix},$$

for some polynomial matrix Y_6 . Define

$$Y_7 := Y_3 - Y_6 C, \quad Y_5 := Y_4 + Y_6 \hat{Q}.$$

We now have

$$\begin{bmatrix} C & \hat{Q} \\ Y_7 & -Y_5 \end{bmatrix} \begin{bmatrix} Y_1 & Q \\ Y_2 & -D \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

It now follows that

$$\begin{bmatrix} Y_1 & Q \\ Y_2 & -D \end{bmatrix} \begin{bmatrix} C & \hat{Q} \\ Y_7 & -Y_5 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

Rewriting the above equations we get

$$(3.14) \quad Y_1 \hat{Q} = Q Y_5,$$

$$(3.15) \quad Y_1 C + Q Y_7 = I.$$

Now suppose $\chi(f) = 0$, for some f in $A_{\hat{Q}}$. Then $\pi_{\hat{Q}}(C f) = 0$. Consequently, there exists f_2 in $A[z]$ such that

$$C f = \hat{Q} f_2.$$

Multiplying on the left by Y_1 we obtain

$$Y_1 C f = Y \hat{Q} f_2 = Q Y_5 f_2.$$

But we also have

$$Y_1 C f = f - Q Y_7 f.$$

Thus

$$f = Q(Y_7 f + Y_5 f_2).$$

As f is in A_Q , $f = 0$. \square

Remark 3.16. In the result by Fuhrmann [1976, Thm. 4.7] for the field case, the equations (3.9) and (3.10) are replaced by the conditions that C and \hat{Q} be left coprime, and D and Q be right coprime. In case A = a field, it is well known (see, for example, MacDuffee [1957, Chapt. 3, Thm. 23.1]) that (3.9) and (3.10) are equivalent to the coprimeness conditions on C and Q , and D and \hat{Q} . However, in case of polynomial matrices over arbitrary commutative rings, the coprimeness conditions are not necessarily equivalent to the Bezout conditions (3.9) and (3.10). Our results thus constitute a natural generalization of the results obtained by Fuhrmann [1976, Thm. 4.7] for the field case. Furthermore, our results also show that Bezout conditions play an important role in $A[z]$ -module isomorphism.

Remark 3.17. Polynomial matrix equations (3.8) through (3.13) arise frequently in several system and control theoretic problems such as dynamic compensation, stochastic control, output regulation and stabilization, etc. Polynomial matrices C and Q that satisfy (3.11) are called internally skew prime. The reader is referred to Roth [1952], Kucera [1975], Rosenbrock and Hayton [1977], Wolovich [1978], Cheng and Pearson [1978], Antsaklis [1979], Emre [1980a], Emre and Silverman [1981], Desoer et al. [1980], etc., for other results concerning these equations. Our results show that there is a close relation between these polynomial matrix equations and $A[z]$ -module isomorphism. In § 5, we will relate these equations to factorizations of transfer matrices and a class of linear systems known as split systems.

The next result is an immediate corollary of Theorem 3.6 and gives a relation between $|Q|$ and $|\hat{Q}|$ whenever A_Q and $A_{\hat{Q}}$ are $A[z]$ -module isomorphic.

COROLLARY 3.18. *Let Q and \hat{Q} be $r \times r$ and $l \times l$ polynomial matrices such that A_Q and $A_{\hat{Q}}$ are $A[z]$ -module isomorphic. Then there exists an invertible element a in $A[z]$ such that*

$$|\hat{Q}| = |Q|a.$$

Proof. We follow the notation of Theorem 3.6. Equation (3.13) implies that

$$a := \left| \begin{bmatrix} C & \hat{Q} \\ Y_7 & -Y_5 \end{bmatrix} \right|$$

is invertible in $A[z]$. Furthermore,

$$\begin{bmatrix} C & \hat{Q} \\ Y_7 & -Y_5 \end{bmatrix} \begin{bmatrix} 0 & Q \\ I & -D \end{bmatrix} = \begin{bmatrix} \hat{Q} & 0 \\ -Y_5 & I \end{bmatrix}.$$

Therefore we have

$$a|Q| = |\hat{Q}|. \quad \square$$

Some of the ideas in the proof of Corollary 3.16 are adapted from Antsaklis [1979]. Theorems 3.3, 3.6, and Corollaries 3.4 and 3.18 constitute our polynomial characterization of finitely generated projective modules with an endomorphism. These

results will be used in the later sections in examining matrix fraction representations for linear systems over rings.

Remark 3.19. Let X be a finitely generated projective A -module with an endomorphism $F: X \rightarrow X$. Then by Theorem 3.2, there exists an $r \times r$ admissible polynomial matrix Q such that X and A_Q are $A[z]$ -module isomorphic. We can now define the *characteristic polynomial* of F to be $|Q|$. This definition is unambiguous in the sense that if \hat{Q} is another admissible polynomial matrix such that $A_{\hat{Q}}$ and X are $A[z]$ -module isomorphic, then $|Q|$ and $|\hat{Q}|$ are related by a unit in $A[z]$ (i.e., $|Q|$ and $|\hat{Q}|$ are associates in $A[z]$). In case A is a field, it was shown by Fuhrmann [1976, Thm. 4.8] that $|Q|$ is the characteristic polynomial of F . In case Q is row proper, which implies that A_Q is free, it was proved in Emre and Khargonekar [1980] that $|Q|$ is equal to $|zI - F|$ for a matrix representation of F_Q . In general, if X is a free module then we can find a matrix F_1 such that X and $A_{(zI - F_1)}$ are $A[z]$ -module isomorphic. If Q is another admissible polynomial matrix such that A_Q and X are $A[z]$ -module isomorphic, then $|Q|$ and $|zI - F_1|$ are associates in $A[z]$. Further, if A is an integral domain, then the invertible elements of $A[z]$ are the same as the invertible elements in A . Then if Q and \hat{Q} are such that A_Q and $A_{\hat{Q}}$ are $A[z]$ -module isomorphic, then the degrees of $|Q|$ and $|\hat{Q}|$ are the same, and $|Q|$ and $|\hat{Q}|$ are the same up to a unit in A . Thus, our results naturally lead to a definition of the characteristic polynomial of an endomorphism of a finitely generated projective module over an arbitrary commutative ring with identity.

4. Matrix fraction descriptions for linear systems over commutative rings. In this section we will consider existence and realization theory of matrix fraction representations for linear systems over commutative rings. Matrix fraction representations have been found very useful in solving several control and system theoretic problems such as dynamic compensation, observers, stochastic control, output regulation and tracking, etc. The reader is referred to the references given in Remark 3.17. In this section, we will show that there is a correspondence between projective realizations of an input-output map f and matrix fraction representations of the transfer matrix Z_f . We will also show that there is a one-to-one correspondence between reachable projective realizations (up to state-module isomorphism) and right matrix fraction representations (up to multiplication by unimodular matrices). Finally, we will also derive necessary and sufficient conditions for Q -realizations to be canonical for integral domains. These conditions are a natural generalization of previously known results for the field case obtained by Fuhrmann [1976].

The following theorem establishes a correspondence between projective realizations and matrix fraction representations.

THEOREM 4.1. *Let f be an input-output map and Z_f be the $p \times m$ transfer matrix of f .*

(a) *If $\Sigma = (F, G, H, X)$ is a projective realization of f , then there exist $p \times r$, $r \times r$, and $r \times m$ polynomial matrices P , Q , and R such that Q is admissible, $Z_f = PQ^{-1}R$, and $\Sigma_Q = (F_Q, G_Q, H_Q, A_Q)$ is isomorphic to $\Sigma = (F, G, H, X)$.*

(b) *If P , Q , and R are $p \times r$, $r \times r$, and $r \times m$ polynomial matrices such that Q is admissible, and $Z_f = PQ^{-1}R$, then $\Sigma_Q = (F_Q, G_Q, H_Q, A_Q)$ is a projective realization of f .*

Proof. Since X is a finitely generated projective A -module with the A -linear map $F: X \rightarrow X$, there exists an integer r and a surjective A -module homomorphism $G_1: A^r \rightarrow X$. Let

$$g: A^r[z] \rightarrow X: \sum_{i=0}^n u_i z^i \mapsto \sum_{i=0}^n F^i G_1(u_i)$$

be the associated A -module surjective homomorphism. We are now in the setup of

Theorem 3.3. Hence, there exists an $r \times r$ admissible polynomial matrix Q such that

$$\ker g = QA'[z].$$

Also, the restriction of g to A_Q ,

$$\chi: A_Q \rightarrow X: u \mapsto g(u),$$

is an A -module isomorphism such that $\chi F_Q = F\chi$.

Let us define a $p \times r$ input-output map f_1 by the impulse response sequence

$$f_1 := (HF^{j-1}G_1)_{j=1}^\infty.$$

Let

$$Z_1 = \sum_{j=1}^\infty (HF^{j-1}G_1)z^{-j}$$

be the transfer matrix of f_1 . Then, we have a factorization of the restricted input-output map f_1^* through X as $f_1^* = hg$, where h is the output map associated with the pair (H, F) . Now, since the columns of Q generate $\ker g$ as an $A[z]$ -module, it follows that $f_1^*(Q) = 0$. Therefore, there exists a $p \times r$ polynomial matrix P such that $Z_1Q = P$, or equivalently, $Z_1 = PQ^{-1}$.

Let e_i be the i th column of I_m . Since G_1 is surjective, there exists r_i in A^r such that $G_1(r_i) = G(e_i)$. Let R be the $r \times m$ matrix with r_i as its i th column. It is clear that

$$G_1(R) = G(I_m).$$

We now have

$$PQ^{-1}R = Z_1R = \sum_{j=1}^\infty (HF^{j-1}G_1)z^{-j}(R) = \sum_{j=1}^\infty (HF^{j-1}G)z^{-j}(I_m) = Z_f.$$

Further, $\Sigma_Q = (F_Q, G_Q, H_Q, A_Q)$ is a realization of f , and $\chi: A_Q \rightarrow X$ is such that $\chi F_Q = F\chi$. Also, for any x in A_Q , we have

$$H_Q(x) = (PQ^{-1}x)_{-1} = (Hg_1(x)) = (H\chi)(x).$$

Therefore $H_Q = H\chi$. Finally, for any u in A^m , we have

$$G(u) = G_1(Ru) = g_1(Ru) = (g_1\pi_Q)(Ru) = (\chi G_Q)(Ru),$$

since $QA'[z] = \ker g_1$. Thus $\chi G_Q = G$. It follows that Σ_Q and Σ are isomorphic realizations of f .

Suppose P , Q , and R are such that Q is admissible, and

$$Z_f = PQ^{-1}R.$$

Then, by Lemma 2.2, $\Sigma_Q = (F_Q, G_Q, H_Q, A_Q)$ is a realization of f . Also, by Lemma 3.2, A_Q is a finitely generated projective A -module. Thus, Σ_Q is a projective realization of f . \square

Remark 4.2. Theorem 4.1 shows that there is a correspondence between projective realizations of f and matrix fraction descriptions of Z_f . In general, there may exist several matrix fraction representations of Z_f having isomorphic Q -realizations. In particular, if

$$Z_f = PQ^{-1}R = \hat{P}\hat{Q}^{-1}\hat{R}$$

are such that Q and \hat{Q} -realizations of f are isomorphic, then Q, \hat{Q} satisfy conditions (3.7) to (3.11). This is closely related to the problem of strict system equivalence. The

reader is referred to Rosenbrock [1970] and Fuhrmann [1977] where the problem of strict system equivalence is considered for the field case. In fact, using Theorem 3.6, the result obtained by Fuhrmann [1977, Thm. 4.1] immediately generalizes to the following theorem.

THEOREM 4.3. *Let Z_f be a strictly proper transfer matrix. Let $P, Q, R, \hat{P}, \hat{Q}, \hat{R}$ be polynomial matrices such that Q and \hat{Q} are admissible, and*

$$Z_f = PQ^{-1}R = \hat{P}\hat{Q}^{-1}\hat{R}.$$

Then Q and \hat{Q} -realizations of Z_f are isomorphic if and only if there exist polynomial matrices $C, D, Y_1, Y_2, Y_3, Y_4, Y_5$, and Y_6 such that

$$\begin{bmatrix} C & 0 \\ Y_5 & I \end{bmatrix} \begin{bmatrix} Q & R \\ -P & 0 \end{bmatrix} = \begin{bmatrix} \hat{Q} & \hat{R} \\ -\hat{P} & 0 \end{bmatrix} \begin{bmatrix} D & -Y_6 \\ 0 & I \end{bmatrix},$$

$$CY_1 + \hat{Q}Y_2 = I, \quad Y_3D + Y_4Q = I.$$

Proof. Using Theorem 3.6, it can be easily checked that the proof of Theorem 4.1 by Fuhrmann [1977] remains valid for our case, with coprimeness of C, Q and \hat{Q}, D replaced by the existence of Y_1, Y_2, Y_3 , and Y_4 satisfying the above conditions. \square

Let f be a given input-output map. In case A is a field, it is shown in Hautus and Heymann [1978, Remark 4.8] that there is a one-to-one correspondence between reachable realizations (up to state-space isomorphism) of f and factorizations of Z_f of the form PQ^{-1} (modulo multiplication of P and Q on the right by unimodular matrices). We will now generalize this result to arbitrary commutative rings with identity in the following theorem.

THEOREM 4.4 *Let f be an input-output map and Z_f be the $p \times m$ transfer matrix of f .*

(a) *If $\Sigma = (F, G, H, X)$ is a reachable and projective realization of f , then there exist $p \times m$ and $m \times m$ polynomial matrices P and Q such that Q is admissible, $Z_f = PQ^{-1}$, and the Q -realization $\Sigma_Q = (F_Q, G_Q, H_Q, A_Q)$ of PQ^{-1} is isomorphic to $\Sigma = (F, G, H, X)$. Further, if \hat{P}, \hat{Q} are $p \times m, m \times m$ polynomial matrices such that \hat{Q} is admissible, $Z_f = \hat{P}\hat{Q}^{-1}$, and the \hat{Q} -realization $\Sigma_{\hat{Q}} = (F_{\hat{Q}}, G_{\hat{Q}}, H_{\hat{Q}}, A_{\hat{Q}})$ of $\hat{P}\hat{Q}^{-1}$ is isomorphic to $\Sigma = (F, G, H, X)$, then there exists an $m \times m$ polynomial matrix N such that $|N|$ is an invertible element of $A[z]$ and*

$$QN = \hat{Q}, \quad PN = \hat{P}.$$

(b) *If P and Q are $p \times m$ and $m \times m$ polynomial matrices such that Q is admissible, and $Z_f = PQ^{-1}$, then the Q -realization $\Sigma_Q = (F_Q, G_Q, H_Q, A_Q)$ is a reachable and projective realization of f .*

Proof. Since $\Sigma = (F, G, H, X)$ is reachable, it follows that the input map

$$g: A^m[z] \rightarrow X: \sum_{i=0}^n u_i z^i \mapsto \sum_{i=0}^n F^i G(u_i)$$

is surjective. Since X is a finitely generated projective A -module, it follows from Theorem 3.2 that there exists an $m \times m$ admissible polynomial matrix Q such that

$$\ker g = QA^m[z],$$

and the restriction χ of g to A_Q , $\chi: A_Q \rightarrow X$, is an A -module isomorphism such that $\chi F_Q = F\chi$. Since $f^* = hg$, and the columns of Q generate $\ker g$, it follows that $f_1^*(Q) = 0$. Hence, there exist a $p \times m$ polynomial matrix P such that $P = Z_f Q$, or equivalently, $Z_f = PQ^{-1}$. As in the proof of Theorem 4.1, it can be easily checked that the Q -realization $\Sigma_Q = (F_Q, G_Q, H_Q, A_Q)$ of Z_f is isomorphic to $\Sigma = (F, G, H, X)$.

Let \hat{P} and \hat{Q} be as in (a). Since $\Sigma_{\hat{Q}} = (F_{\hat{Q}}, G_{\hat{Q}}, H_{\hat{Q}}, A_{\hat{Q}})$ is isomorphic to $\Sigma = (F, G, H, X)$, it follows that

$$\ker g = \ker \pi_{\hat{Q}}.$$

Further, $\ker \pi_{\hat{Q}} = \hat{Q}A^m[z]$, and $\ker g = QA^m[z]$. Therefore, there exist $m \times m$ polynomial matrices N and \hat{N} such that

$$\hat{Q} = QN, \quad Q = \hat{Q}\hat{N}.$$

Consequently,

$$Q = \hat{Q}\hat{N} = QN\hat{N}.$$

Since Q is admissible, $N\hat{N} = I$. Therefore, $|N|$ is an invertible element of $A[z]$. Also,

$$\hat{P} = Z_f \hat{Q} = Z_f QN = PN.$$

To prove (b), we note that $\Sigma_Q = (F_Q, G_Q, H_Q, A_Q)$ is a projective realization of f . Since $\pi_Q: A^m[z] \rightarrow A_Q$ is surjective, Σ_Q is reachable as well. Thus Σ_Q is a reachable and projective realization of f . \square

Thus, we see that there is a one-to-one correspondence between reachable realizations (up to state-module isomorphism) of f and factorizations of Z_f and PQ^{-1} (up to multiplication by unimodular matrices). This result also constitutes a natural generalization of the corresponding result in the field case. (See Hautus and Heymann [1978, Remark 4.8].)

We will now consider conditions under which a given Q -realization is canonical. These results generalize the corresponding results in the field case obtained in Fuhrmann [1976].

THEOREM 4.5. *Let f be an input-output map with the $p \times m$ transfer matrix Z_f . Let P , Q , and R be $p \times r$, $r \times r$, and $r \times m$ polynomial matrices such that Q is admissible, and*

$$Z_f = PQ^{-1}R.$$

Then $\Sigma_Q = (F_Q, G_Q, H_Q, A_Q)$ is reachable if and only if there exist polynomial matrices Y_1 and Y_2 such that

$$RY_1 + QY_2 = I.$$

If A is an integral domain, then Σ_Q is observable if and only if there exist Y_3 in $K^{r \times p}[z]$ and Y_4 in $K^{r \times r}[z]$ such that

$$Y_3P + Y_4Q = I.$$

Proof. By definition Σ_Q is reachable if and only if

$$g_Q: A^m[z] \rightarrow A_Q: u \mapsto \pi_Q(Ru)$$

is surjective. Let Y_5 be a polynomial matrix such that $\pi_Q(I) = Y_5$. Then there exists a polynomial matrix Y_6 such that $I = Y_5 + QY_6$. Now, if g_Q is surjective, then there exists u_1, u_2, \dots, u_m in $A^m[z]$, such that $g_Q(u_j)$ is the j th column of Y_5 . Consequently, if Y_1 is the $m \times m$ polynomial matrix with u_j as its j th column then

$$\pi_Q(RY_1) = Y_5 = I - QY_6.$$

There exists a polynomial matrix Y_7 such that

$$RY_1 = I - QY_6 - QY_7.$$

If we define $Y_2 := Y_6 + Y_7$, then

$$QY_2 + RY_1 = I.$$

Then for any x in A_Q ,

$$x = RY_1x + QY_2x.$$

It is clear that

$$g_Q(Y_1x) = \pi_Q(RY_1x) = x.$$

Thus, g_Q is surjective.

Now, let A be an integral domain. Σ_Q is observable if and only if the map

$$h_Q: A_Q \rightarrow z^{-1}A^p[[z^{-1}]]: x \mapsto \pi(PQ^{-1}x)$$

is one-to-one. Consider the map

$$\hat{h}_Q: K_Q \rightarrow z^{-1}K^p[[z^{-1}]]: x \mapsto \pi(PQ^{-1}x).$$

Since K is the quotient field of A , it is clear that if \hat{h}_Q is one-to-one then h_Q is also one-to-one. Conversely, if \hat{h}_Q is not one-to-one, then there exists x in K_Q such that $\pi(PQ^{-1}x) = 0$. By clearing denominators, we can write x as αx_1 where α is in K and x_1 is in A_Q . Then $\hat{h}_Q(\alpha x_1) = 0$. Since \hat{h}_Q is K -linear, it follows that $\hat{h}_Q(x_1) = 0$. Thus, h_Q is one-to-one if and only if \hat{h}_Q is one-to-one. It is known (see Fuhrmann [1976, Thm. 6.1]) that \hat{h}_Q is one-to-one if and only if P and Q are left coprime over $K[z]$. But, this is equivalent to the existence of Y_3 in $K^{r \times p}[z]$ and Y_4 in $K^{r \times r}[z]$ such that

$$Y_3P + Y_4Q = I. \quad \square$$

Remark 4.6. Conte and Perdon [1982] have (independently and simultaneously) obtained somewhat similar results for the particular case of systems over a principal ideal domain. Our results are considerably more general and apply to arbitrary commutative rings with identity. Further, the concepts of Bezout factorizations and split systems considered in § 5 have not been examined by Conte and Perdon [1981].

Remark 4.7. We have restricted our attention to matrix fraction descriptions of the form $PQ^{-1}R$. However, it is easy to check that most of the results can be appropriately extended to descriptions of the form $PQ^{-1}R + S$.

5. Bezout factorizations and split systems. Let f be a realizable input-output map with the transfer matrix Z_f . In case A is a field, it is well known (see, e.g., Rosenbrock [1970]) that there exist left coprime polynomial matrices Q_1 and R_1 and right coprime polynomial matrices P and Q such that the transfer matrix

$$(5.1) \quad Z_f = Q_1^{-1}R_1 = PQ^{-1}.$$

Furthermore, the coprimeness conditions on the polynomial matrices P , Q and Q_1 , R_1 are well-known to be equivalent to the existence of polynomial matrices Y_1 , Y_2 , Y_3 and Y_4 such that

$$(5.2) \quad Y_1P + Y_2Q = I,$$

$$(5.3) \quad R_1Y_3 + Q_1Y_4 = I.$$

The existence of these coprime factorizations and (5.2) and (5.3) play an important role in the polynomial matrix approach to several system and control theoretic problems. The reader is referred to the references given in Remark 3.17. (Also, see Remark 5.10.)

In this section we will establish a necessary and sufficient condition on a realizable input-output map f over an arbitrary but fixed commutative ring A with identity for the associated transfer matrix Z_f to admit Bezout factorizations of the form (5.1). Recall that a factorization $Z_f = PQ^{-1}$ (respectively, $Z_f = Q_1^{-1}R_1$) is said to be *right* (respectively, *left*) Bezout if and only if there exist polynomial matrices Y_1, Y_2 (respectively, Y_3, Y_4) satisfying (5.2) (respectively, (5.3)). We will show that Z_f admits right and/or left Bezout factorizations if and only if the input-output map f is split. The concept of a split input-output map was first introduced by Sontag [1978, Def. 1.8] which is as follows.

DEFINITION 5.4 (Sontag [1978, Def. 1.8]). A system $\Sigma = (F, G, H, X)$ is said to be *split* if and only if the following conditions hold:

- (a) X is a (finitely generated) projective module,
- (b) Σ is reachable,
- (c) $\Sigma' = (F', H', G', X')$ is reachable.

An input-output map f and the associated transfer matrix Z_f are said to be *split* if and only if f can be realized by a split system.

The concept of a split system was introduced for application to the problem of regulation of linear systems over commutative rings. A complete solution to the regulator problem for free split systems (i.e., the state-module is free), has been given by Emre and Khargonekar [1980] using a construction of observer and dynamic state-feedback. The existence of Bezout factorizations of transfer matrices has been assumed by Desoer et al. [1980] in their study of some control theoretic problems. Our results in this section establish a link between the approaches developed by Emre and Khargonekar [1980] and Desoer et al. [1980].

Using the results developed by Sontag [1978] we will also obtain some alternative criteria for checking the existence of Bezout factorizations.

We will first establish two simple results which will be useful in establishing the main results of this section.

LEMMA 5.5. *An input-output map f is split if and only if f' is split.*

Proof. Suppose f is split. Let $\Sigma = (F, G, H, X)$ be a split realization of f . Since X is a finitely generated projective A -module, it follows from Jans [1964, Chapt. 5, p. 66] that X' is a finitely generated projective A -module, and there exists an isomorphism $\chi: X \rightarrow X''$ such that $\chi F = F''\chi$, and $\chi G = G''$. Since Σ is reachable, there exists an integer n such that the n -step reachability map

$$g_n: A^{nm} \rightarrow X: (u_0, u_1, \dots, u_{n-1}) \mapsto \sum_{j=0}^{n-1} F^j G(u_j)$$

is surjective. Further, $(A^{nm})''$ is isomorphic to A^{nm} . Also, we have

$$g_n'': A^{nm} \rightarrow X'': (u_0, u_1, \dots, u_{n-1}) \mapsto \sum_{j=0}^{n-1} (F'')^j G''(u_j).$$

Since

$$\sum_{j=0}^{n-1} (F'')^j G''(u_j) = \chi \left(\sum_{j=0}^{n-1} F^j G(u_j) \right),$$

and g_n is surjective, it follows that g_n'' is surjective. By assumption, $\Sigma' = (F', G', H', X')$ is reachable. Since g_n'' is surjective, $\Sigma'' = (F'', G'', H'', X'')$ is also reachable. Thus, Σ' is split. Now, since Σ' is a realization of f' , f' is split.

Conversely, if f' is split then, by the first part of the proof, f'' is split. But the input-output map f'' is the same as f . Therefore, f is split. \square

We will now establish another characterization of reachability of Σ' . In what follows, for a system $\Sigma = (F, G, H, X)$, let h_n denote the n -step observability map

$$h_n : X \rightarrow A^{np} : x \mapsto \begin{bmatrix} Hx \\ HFx \\ \vdots \\ HF^{n-1}x \end{bmatrix}.$$

LEMMA 5.6. *Let $\Sigma = (F, G, H, X)$ be a projective system. Then Σ' is reachable if and only if there exists an integer n and an A -linear map*

$$\varphi : A^{np} \rightarrow X$$

such that φh_n is the identity map on X .

Proof. Suppose Σ' is reachable. Then, there exists an integer n such that

$$h'_n : A^{np} \rightarrow X' : (u_0, u_1, \dots, u_{n-1}) \mapsto \sum_{j=0}^{n-1} (F')^j H'(u_j)$$

is surjective. It now follows from Bourbaki [1962, Chapt. 2, § 2, Prop. 12] that $h_n : X \rightarrow A^{np}$ splits; i.e., there exists an A -linear map $\varphi : A^{np} \rightarrow X$ such that φh_n is the identity map on X .

Conversely, suppose there exists an A -linear map $\varphi : A^{np} \rightarrow X$ such that φh_n is the identity map on X . Then, $h'_n \varphi'$ is the identity map on X' . Hence, $h'_n : A^{np} \rightarrow X'$ is surjective. But h'_n is the n -step reachability map of Σ' . Thus, Σ' is reachable. \square

We now have the main result of this section.

THEOREM 5.7. *Let f be an input-output map, and $Z_f = \sum_{j=1}^{\infty} A_j z^{-j}$ be the $p \times m$ transfer matrix of f . Then the following statements are equivalent.*

(a) *There exist polynomial matrices P, Q, Y_1 , and Y_2 such that Q is admissible, and*

$$Z_f = PQ^{-1}, \quad Y_1 P + Y_2 Q = I$$

(i.e., Z_f admits a right Bezout factorization).

(b) *There exist polynomial matrices Q_1, R_1, Y_3 , and Y_4 such that Q_1 is admissible, and*

$$Z_f = Q_1^{-1} R_1, \quad R_1 Y_3 + Q_1 Y_4 = I$$

(i.e., Z_f admits a left Bezout factorization).

(c) *F is a split input-output map.*

Proof. We will show that (a) and (c) are equivalent, and (b) and (c) are equivalent.

(a) *implies* (c). Let $\Sigma_Q = (F_Q, G_Q, H_Q, A_Q)$ be the Q -realization of PQ^{-1} . By Theorem 4.3, it follows that Σ_Q is a reachable, projective realization of f . We will now show that Σ'_Q is reachable. Consider the map

$$\varphi_1 : z^{-1} A^p[[z^{-1}]] \rightarrow A_Q : q \mapsto \pi_Q((Q\pi(Y_1 q))_+),$$

where, for x in $A^p(z)$,

$$(x)_+ := x - \pi(x).$$

Since φ_1 is composition of A -linear maps, φ_1 is A -linear. We have

$$Y_1 P + Y_2 Q = I,$$

which can be rewritten as

$$Y_1 P Q^{-1} + Y_2 = Q^{-1}.$$

It follows that for any x in A_Q , we have

$$Y_1 P Q^{-1} x + Y_2 x = Q^{-1} x.$$

Since x is in A_Q and Y_2 is a polynomial matrix,

$$Q^{-1} x = \pi(Y_1 P Q^{-1} x).$$

Therefore, we have

$$\begin{aligned} x &= Q Q^{-1} x = Q \pi(Y_1 P Q^{-1} x), \\ (\varphi_1 h_Q)(x) &= \pi_Q((Q \pi(Y_1 \pi(P Q^{-1} x)))_+) = \pi_Q((Q \pi(Y_1 P Q^{-1} x))_+). \end{aligned}$$

Thus

$$x = Q \pi(Y_1 P Q^{-1} x) = \pi_Q((Q \pi(Y_1 P Q^{-1} x))_+) = (\varphi_1 h_Q)(x).$$

It is thus clear that $\varphi_1 h_Q$ is the identity map X . Since A_Q is a finitely generated A -module, it follows from Atiyah and MacDonald [1969, Chapt. 2, Prop. 2.3] that there exist a monic polynomial α such that $\alpha(F_Q) = 0$. Let n be the degree of α . Now, if h_n represents the n -step observability map of Σ_Q , then it is easy to see there exists an A -linear map

$$\psi: A^{np} \rightarrow z^{-1} A^p[[z^{-1}]]$$

such that $h_Q = \psi h_n$. Let $\varphi := \varphi_1 \psi$. Then φh_n is the identity map on X . Now, by Lemma 5.6, Σ'_Q is reachable. Thus, Σ_Q is a split realization of f . Therefore, f is a split input-output map.

(c) *implies* (a). Let $\Sigma = (F, G, H, X)$ be a split realization of f . Then, Σ is a reachable projective realization of f . By Theorem 4.3, there exist polynomial matrices P and Q such that Q is admissible, $Z_f = P Q^{-1}$, and the Q -realization of $\Sigma_Q = (F_Q, G_Q, H_Q, A_Q)$ of $P Q^{-1}$ is isomorphic to Σ . Hence Σ_Q is split. It follows that Σ'_Q is reachable. Now by Lemma 5.6, there exists an integer n , and an A -linear map $\varphi: A^{np} \rightarrow A_Q$, such that if h_n represents the n -step observability map of Σ_Q , then φh_n is the identity map on A_Q . Let θ be the A -linear map

$$\theta: A_Q \rightarrow A^m: x \mapsto (Q^{-1} x)_{-1}.$$

Then, $(\theta \varphi h_n) = \theta$. Furthermore, $\theta \varphi: A^{np} \rightarrow A^m$ is A -linear. Therefore, there exist V_0, V_1, \dots, V_{n-1} in $A^{m \times p}$ such that

$$(\theta \varphi) \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1} \end{bmatrix} = \sum_{j=0}^{n-1} V_j y_j,$$

where each y_j is in A^p . We also have

$$h_n(\pi_Q(z^l I)) = \begin{bmatrix} (P Q^{-1} z^l)_{-1} \\ (P Q^{-1} z^l)_{-2} \\ \vdots \\ (P Q^{-1} z^l)_{-n} \end{bmatrix} = \begin{bmatrix} A_{l+1} \\ A_{l+2} \\ \vdots \\ A_{l+n} \end{bmatrix}, \quad l \geq 0.$$

Further,

$$\theta(\pi_Q(z^l I)) = (z^l Q^{-1})_{-1}.$$

It now follows that

$$(\theta\varphi h_n)(\pi_Q(z^l I)) = \sum_{j=0}^{n-1} V_j A_{l+j+1} = \theta(\pi_Q(z^l I)).$$

It is clear that

$$\pi(Q^{-1}) = \sum_{l=0}^{\infty} (z^l Q^{-1})_{-1} z^{-(l+1)}.$$

Therefore

$$\begin{aligned} \pi(Q^{-1}) &= V_0 \sum_{l=1}^{\infty} A_l z^{-l} + V_1 \sum_{l=1}^{\infty} A_{l+1} z^{-l} + \cdots + V_{n-1} \sum_{l=1}^{\infty} A_{l+n-1} z^{-l} \\ &= V_0 Z_f + V_1 \pi(z Z_f) + \cdots + V_{n-1} \pi(z^{n-1} Z_f) \\ &= \pi((V_0 + V_1 z + \cdots + V_{n-1} z^{n-1}) Z_f). \end{aligned}$$

Let us define

$$Y_3 := \sum_{i=0}^{n-1} V_i z^i.$$

Then there exists a polynomial matrix Y_4 such that

$$Y_3 Z_f = Q^{-1} - Y_4.$$

Since $Z_f = P Q^{-1}$, it follows that

$$Y_3 P + Y_4 Q = I.$$

Thus, we have proved that (a) and (c) are equivalent.

(c) *implies* (b). Since f is split, by Lemma 5.5, f' is split. Also, the transfer matrix of f' is $Z_{f'}$. As (c) implies (a), it follows that there exist polynomial matrices Q'_1, R'_1, Y'_3 , and Y'_4 such that Q'_1 is admissible, and

$$Z_{f'} = R'_1 (Q'_1)^{-1}, \quad Y'_3 R'_1 + Y'_4 Q'_1 = I.$$

(b) follows immediately by taking matrix transposes.

(b) *implies* (c). By taking transposes we have

$$Z'_f = R'_1 (Q'_1)^{-1}, \quad Y'_3 R'_1 + Y'_4 Q'_1 = I.$$

As (a) implies (c) and Z'_f is the transfer matrix of f'_1, f' is split. By Lemma 5.5, f is split. Thus (b) and (c) are equivalent. \square

Thus, it is seen that the existence of Bezout factorizations of Z_f is equivalent to the condition that the input-output map that f be split.

If A is a Noetherian integral domain, a useful condition for checking whether f is split is given by Sontag [1978, Thm. 2.1]. In what follows, let H_n be the $n \times n$ block behavior (Hankel) matrix

$$H_n := \begin{bmatrix} A_1 & A_2 & \cdots & A_n \\ A_2 & A_3 & & A_{n+1} \\ \vdots & & & \vdots \\ A_n & A_{n+1} & \cdots & A_{2n-1} \end{bmatrix}$$

of the input-output map f .

THEOREM 5.8 (Sontag [1978, Thm. 2.1]). *Let A be a Noetherian integral domain and K be its quotient field. Let f be an input-output map over A and let n be the dimension of a canonical realization of f over K . Then f is split if and only if the ideal generated by all $(n \times n)$ minors of the $n \times n$ block behavior matrix H_n of f is A .*

We will now state a classical result from commutative algebra which gives a necessary and sufficient condition for a finitely generated ideal in a polynomial ring A to be A .

THEOREM 5.9 (the Hilbert Nullstellensatz, Zariski and Samuel [1958, Chapt. 7, Thm. 14]). *Let L be a field and A be $L[T_1, T_2, \dots, T_n]$, the ring of polynomials in the indeterminates T_1, T_2, \dots, T_n over the field L . Then p_1, p_2, \dots, p_m in A generate the ideal A if and only if there do not exist c_1, c_2, \dots, c_n in an algebraic closure of L such that*

$$p_j(c_1, c_2, \dots, c_n) = 0, \quad j = 1, 2, \dots, m$$

(i.e., the polynomials p_1, p_2, \dots, p_m do not have a common zero in an algebraic closure of L).

We note that the results from classical elimination theory can be used to check whether a given finite set of polynomials in several unknowns have a common zero in an algebraic closure of L . The reader is referred to Hodge and Pedoe [1968, Chapt. 4, § 7] for a discussion of these techniques. Thus, using Theorems 5.8, 5.9 and results from elimination theory, we can check the existence of Bezout factorizations for linear systems over polynomial rings in several variables. This has immediate applications to delay-differential systems with a finite number of noncommensurate delays. Kamen [1975] established that such a delay-differential system can be viewed as a system over a polynomial ring of delay operators. These results are also applicable to systems with parameters where the parameter variations are modeled as polynomials.

Another special case arises if A is a principal ideal domain. Then it follows from Theorem 5.8 that a given input-output map f is split if and only if a greatest common divisor of all $n \times n$ minors of the (block) $n \times n$ behavior matrix H_n is the multiplicative identity of A .

Remark 5.10. Let f be a split input-output map. Then (5.1)–(5.3) hold. In particular, (5.1) can be rewritten as

$$(5.11) \quad R_1 Q = Q_1 P.$$

Now, (5.11), (5.2), and (5.3) are exactly the equations in the set up of Theorem 3.6. With the above notation, (3.13) becomes

$$(5.12) \quad Y_3 R_1 + Q Y_7 = I, \quad P Y_5 + Y_4 Q_1 = I,$$

for some polynomial matrices Y_5 and Y_7 . These equations are closely related to the concept of skew prime polynomial matrices, which plays an important role in the polynomial matrix approach to certain control-theoretic problems such as output regulation with internal stability, (see Wolovich and Ferreira [1979]), and stochastic control, (see Kucera [1975]). Two polynomial matrices $Q_1, P(Q, R_1)$ are said to be *externally (internally) skew prime* if and only if there exist polynomial matrices Y_4 and Y_5 (Y_3 and Y_7) such that

$$(5.13) \quad P Y_5 + Y_4 Q_1 = I, \quad (Y_3 R_1 + Q Y_7 = I).$$

Equation (5.13) has been considered in several papers. The reader is referred to Roth [1952], Wolovich [1978], Gustafson [1979], Emre and Silverman [1981] and the references given therein. It has been shown by Wolovich [1978] that in case A is a field, the polynomial matrices Q_1, P are externally skew prime if and only if there exist

polynomial matrices R_1 , Q , Y_1 , Y_2 , Y_3 and Y_4 such that (5.11), (5.2) and (5.3) are satisfied. Our results show that for arbitrary commutative rings with identity, if R_1 , Q , Y_1 , Y_2 , Y_3 and Y_4 exist such that (5.11), (5.2) and (5.3) are satisfied, then Q_1 , $P(Q, R_1)$ are externally (internally) skew prime. Using the results of Theorem 5.7, skew-primeness and related questions have been considered in Emre [1980, Lemma 3.4] for arbitrary commutative rings with identity. Our results indicate that there is a close relation between split systems and these polynomial equations.

Remark 5.14. In a personal communication, C. Byrnes has indicated that the algebro-geometric methods developed by Byrnes [1978] can be used to prove Theorem 5.6 for systems over a ring of polynomials in several variables over the field of complex numbers. Our results, however, are for arbitrary commutative rings with identity and are obtained by a purely algebraic approach. It may be interesting to examine in detail the connections between the results of this paper and the algebro-geometric approach to the theory of linear systems.

Acknowledgment. The author wishes to thank Dr. E. D. Sontag for several valuable discussions concerning the research reported here. The author is also grateful to the reviewer for his constructive comments.

REFERENCES

- P. J. ANTSAKLIS [1979], *Some relations satisfied by prime polynomial matrices and their role in linear multivariable system theory*, IEEE Trans. Automatic Control, AC-24, pp. 611–616.
- M. F. ATIYAH AND I. G. MACDONALD [1969], *Introduction to Commutative Algebra*, Addison-Wesley, Reading, MA.
- N. BOURBAKI [1962], *Algebre*, Hermann, Paris.
- [1972], *Commutative Algebra*, Addison-Wesley, Reading, MA.
- R. W. BROCKETT AND J. L. WILLEMS [1974], *Discretized partial differential equations: examples of control systems defined over modules*, Automatica, 10, pp. 507–515.
- C. I. BYRNES [1978], *On the control of certain deterministic infinite-dimensional systems by algebro-geometric techniques*, Amer. J. Math., 100, pp. 1333–1381.
- [1979], *On the stabilizability of linear control systems depending on parameters*, Proc. IEEE Conference on Decision and Control, Florida.
- L. CHENG AND J. B. PEARSON, JR. [1978], *Frequency domain synthesis of multivariable linear regulators*, IEEE Trans. Automatic Control, AC-23, pp. 3–15.
- G. CONTE AND A. M. PERDON [1982], *Systems over a principal ideal domain. A polynomial model approach*, Facolta' di Ingegneria, Universita' di Padova, 1980, this Journal, 20, pp. 112–124.
- 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. Automatic Control, AC-25, pp. 399–412.
- S. EILENBERG [1974], *Automata, Languages, and Machines*, Vol. A, Academic Press, New York.
- R. EISING [1978], *Realization and stabilization of 2-D systems*, IEEE Trans. Automatic Control, AC-23, pp. 793–799.
- [1979], *2-D Systems, an algebraic approach*, Mathematisch Centrum, Amsterdam.
- E. EMRE [1980a], *The polynomial equation $QQ_c + RP_c = \Phi$ with applications to dynamic feedback*, this Journal, 18, pp. 611–620.
- [1980b], *Output regulation in the presence of disturbances of linear systems over rings*, Center for Mathematical System Theory, University of Florida.
- [1980c], *On a natural realization of matrix fraction descriptions*, IEEE Trans. Automatic Control, AC-25, pp. 288–289.
- E. EMRE AND P. P. KHARGONEKAR [1980], *Regulation of linear systems over rings: coefficient assignment and observers*, Center for Mathematical System Theory, University of Florida.
- E. EMRE AND L. M. SILVERMAN [1981], *The equation $XR + QY = \Phi$: a characterization of solutions*, Center for Mathematical System Theory, University of Florida, this Journal, 19, pp. 33–38.
- P. A. FUHRMANN [1976], *Algebraic system theory: an analyst's point of view*, J. Franklin Inst., 301, pp. 521–540.

- [1977], *On strict system equivalence and similarity*, Internat. J. Control, 25, pp. 5–10.
- W. H. GUSTAFSON [1979], *Roth's Theorem over commutative rings*, Linear Algebra and Appl., 23, pp. 245–251.
- M. L. J. HAUTUS AND M. HEYMANN [1978], *Linear feedback—an algebraic approach*, this Journal, 16, pp. 83–105.
- M. L. J. HAUTUS AND E. D. SONTAG [1980], *An approach to detectability and observers*, Dept. of Mathematics, Memo. COSOR 70-08, Eindhoven University of Technology, Eindhoven, the Netherlands, also to appear in Algebraic and Geometric Methods in Linear Systems Theory, C. I. Byrnes and C. F. Martin, eds., American Mathematical Society, Providence, RI.
- M. HAZEWINKEL [1980], *A partial survey of the uses of algebraic geometry in systems and control theory*, in Sym. Math. INOAM (Severi Centennial Conference), Academic Press, New York.
- W. V. D. HODGE AND D. PEDOE [1968], *Methods of Algebraic Geometry*, vol. 1, Cambridge University Press, Cambridge.
- J. P. JANS [1964], *Rings and Homology*, Holt, Rinehart, and Winston, New York.
- R. JOHNSTON [1973], *Linear systems over various rings*, Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA.
- R. E. KALMAN, P. L. FALB AND M. A. ARBIB [1969], *Topics in Mathematical System Theory*, McGraw-Hill, New York.
- E. W. KAMEN [1975], *On an algebraic theory of systems defined by convolution operators*, Math. Systems Theory, 9, pp. 57–74.
- [1978], *Lectures on algebraic system theory: linear systems over rings*, NASA Contractor Report 3016.
- P. P. KHARGONEKAR AND E. D. SONTAG [1981], *On the relation between stable matrix fraction factorizations and regulable realizations of linear systems over rings*, Department of Mathematics, Rutgers University, New Brunswick, NJ.
- J. T. KNIGHT [1971], *Commutative Algebra*, London Mathematical Society, Lecture Note Series 5, Cambridge University Press, Cambridge.
- W. KUCERA [1975], *Algebraic approach to discrete stochastic control*, Kybernetika, 11, pp. 114–119.
- T. Y. LAM [1978], *Serre's Conjecture*, Lecture Notes in Mathematics 635, Springer-Verlag, Berlin.
- E. B. LEE AND A. W. OLBOROT [1980], *On reachability over polynomial rings and a related genericity problem*, Proc. 17th Conference on Information Science and Systems, Princeton, NJ.
- C. C. MACDUFFEE [1956], *The Theory of Matrices*, Chelsea, New York.
- S. MACLANE AND G. BIRKHOFF [1967], *Algebra*, Macmillan, New York.
- H. H. ROSENBRACK [1970], *State Space and Multivariable Theory*, John Wiley, New York.
- H. H. ROSENBRACK AND G. E. HAYTON [1977], *The general problem of pole assignment*, Control Systems Center Report no. 288, University of Manchester, England.
- W. E. ROTH [1952], *The equations $AX - YB = C$ and $AX - XB = C$ in matrices*, Proc. Amer. Math. Soc., 3, pp. 392–396.
- Y. ROUCHALEAU [1972], *Linear, discrete-time, finite-dimensional dynamical systems over some classes of commutative rings*, Ph.D. dissertation, Stanford Univ., Stanford, CA.
- Y. ROUCHALEAU, B. F. WYMAN AND R. E. KALMAN [1972], *Algebraic structure of linear dynamical systems. III. Realization theory over a commutative ring*, Proc. Nat. Acad. Sci. USA, 69, pp. 3404–3406.
- E. D. SONTAG [1976], *Linear systems over commutative rings: a survey*, Recherche di Automatica, 7, pp. 1–34.
- [1978], *On split realizations of response maps over rings*, Information and Control, 37, pp. 23–33.
- [1978a], *On first-order equations for multi-dimensional filters*, IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP-26, pp. 480–482.
- W. A. WOLOVICH [1974], *Linear Multivariable Systems*, Springer, New York.
- [1978], *Skew prime polynomial matrices*, IEEE Trans. Automatic Control, AC-23, pp. 880–887.
- W. A. WOLOVICH AND P. FERREIRA [1979], *Output regulations and tracking in linear multivariable systems*, IEEE Trans. Automatic Control, AC-24, pp. 460–465.
- B. F. WYMAN [1972], *Linear systems over commutative rings*, Lecture notes, Stanford Univ., Stanford, CA.
- O. ZARISKI AND P. SAMUEL [1958], *Commutative Algebra*, Vol. 2, Van Nostrand, New York.

LOWER CLOSURE PROBLEMS WITH WEAK CONVERGENCE CONDITIONS IN A NEW PERSPECTIVE*

E. J. BALDER†

Abstract. A transparent approach is taken towards the class of (lower) closure problems with weak convergence conditions for the “derivatives”. A deparametrization procedure is formulated for the abstract lower closure problem of this class, as well as for its variant in a control problem of Lagrange type. It is shown that, if one follows this procedure, the solution of the lower closure problem merely lies in proving that a certain “modified Lagrangian” is a normal integrand. A lower closure result is obtained that generalizes, in itself, all comparable results of [2a], [6e-f], [7a], [21]. For control problems of Lagrange type with uniform boundedness conditions on the “controls,” a special approximate Lagrangian can be formulated by which the deparametrization procedure yields results that are superior to those obtained previously [2b], [7b-c]. An important novelty in deriving the measurability properties of Lagrangians is also introduced here: it consists of the employment of the Kunugui–Novikov–Stechegolkov projection theorem [4].

1. Introduction. Consider the following so-called *deparametrization procedure*, an approach to solve a lower closure or closure problem with weak convergence conditions for the “derivatives”:

- (i) Construct a suitable so-called *modified Lagrangian* from the basic components of the problem and show it to be a normal integrand.
- (ii) Apply a standard lower semicontinuity theorem on the integral functional having the modified Lagrangian as its integrand, with “trajectories” and “derivatives” as its arguments.
- (iii) Apply a standard measurable implicit function theorem (only when dealing with optimal control problems).

Because steps (ii), (iii) of this procedure are routine, the solution of the (lower) closure problem by means of the above procedure lies in demonstrating the measurability and semicontinuity properties of the modified Lagrangian. As will be made clear, this simplifies and standardizes the solution method to a considerable extent.

It would seem that the above deparametrization procedure can only be successful for the class of closure problems with weak convergence conditions. In contrast with the situation at present, this class was for some time overshadowed in importance by those problems where weak convergence conditions are avoided by strengthening the degree of seminormality (for instance by requiring Cesari’s property (Q) to hold); this is evident from L. Cesari’s earlier work (cf. [6a-d]). That could possibly account for the undeserved absence of the deparametrization procedure from the literature on (lower) closure problems thus far (although the procedure was mentioned in passing in a remark [18, p. 15]).

As indicative of this situation one can observe—and we shall cite only two instances—that the results of [2a] can be obtained, straightaway and under essentially weaker conditions, from the lower semicontinuity result in [2c] by the deparametrization procedure. Also, [6f, Statements (5i), (6i)] can be shown to follow directly (and at least Statement (5i) under significantly weaker conditions) from the semicontinuity result of [6f, Statement (1iii)] by the same procedure.

* Received by the editors March 5, 1980 and in revised form April 6, 1981.

† Mathematical Institute, University of Utrecht, Budapestlaan 6, P.O. Box 80 010, 3508 TA Utrecht, the Netherlands.

On the other hand, for some optimal control problems of Lagrange type whose compactness conditions (e.g., compactness of the space of control points) completely overshadow the importance of seminormality, the deparametrization procedure has been used quite successfully; cf. [9], [10], [14], [19]. As mentioned above, the procedure has been discussed very briefly by C. Olech in [18], but his remarks go without proof, bear only on the case where the upper set property (π) holds, and do not apply to lower closure problems without seminormality conditions, such as considered in [7b–c]. (It would seem that Olech did not return to the subject elsewhere.) The purpose of this paper is to present a systematic exposition of the use of the deparametrization procedure for (lower) closure problems with at most weak seminormality conditions. At the same time, quite new lower closure results are obtained for control problems of Lagrange type with uniform boundedness conditions on the “controls” (e.g., L_1 -bounds [2b], [7c]). For these special problems we shall introduce another Lagrangian, the *approximate Lagrangian*, whose semicontinuity property is proven in line with the proofs of [9], [10], [14] for the classical Lagrangian.

The organization of this paper is simple. The technical aspects of the first step receive due consideration in § 2. After this, the deparametrization procedure is followed through for abstract (lower) closure problems in § 3 and for optimal control problems of Lagrange type in § 4.

2. Lagrangians and normal integrands. The first step of the deparametrization procedure is studied in this section. Along the way, a large part of the basic framework of this paper is described in notation similar to that used in [6a–f]. However, we shall allow the setting to be substantially more abstract than in the above references.

Let (G, \mathcal{G}, μ) be a finite measure space whose σ -algebra \mathcal{G} is either countably generated or the completion of a countably generated σ -algebra. Let X be a metrizable Lusin space [8] (also known under the name of standard Borel space) with given metric d , equipped with the Borel σ -algebra. Let B be a measurable closed-valued multifunction from G into X and denote its graph by \mathcal{B} ; this graph is a measurable subset of $G \times X$ [5, III.10, III.13]. Further, let r be a prescribed natural number and let \tilde{Q} be a multifunction from \mathcal{B} into \mathbb{R}^{r+1} with closed values and measurable graph. (From now on, every Euclidean space is supposed to be equipped with its Borel σ -algebra, unless stated otherwise.) We shall say that \tilde{Q} has *property (K) with respect to $B(t)$* at a point $(t, x^0) \in \mathcal{B}$ if

$$(1) \quad \tilde{Q}(t, x^0) = \bigcap_{\gamma > 0} \text{cl} \bigcup \{ \tilde{Q}(t, x) : x \in B(t), d(x, x^0) < \gamma \},$$

where cl denotes closure. This property is classical in the existence theory for optimal control problems (cf. [6], [7], [11]).

Let us also agree to say that \tilde{Q} has *property (K) with respect to \mathcal{B}* if \tilde{Q} has property (K) with respect to $B(t)$ at every $(t, x^0) \in \mathcal{B}$.

Remark 1. In case the set $B(t)$ is a singleton for every $t \in G$, property (K) with respect to \mathcal{B} will hold vacuously.

Remark 2. Let us say that \tilde{Q} has property $(K)_{\text{loc}}$ with respect to $B(t)$ at $(t, x^0) \in \mathcal{B}$ if there exists $N_0 > 0$ such that for every $N \geq N_0$ the multifunction \tilde{Q}_N satisfies (1), where $\tilde{Q}_N(t, x)$ denotes the intersection of $\tilde{Q}(t, x)$ with the closed ball with radius N around 0 in \mathbb{R}^{r+1} , $(t, x) \in \mathcal{B}$. Then it is not hard to see that at every point $(t, x^0) \in \mathcal{B}$, the properties $(K)_{\text{loc}}$ and (K) with respect to $B(t)$ are equivalent.

Remark 3. It should be noted that we do not follow [6], [7] in the custom of introducing additional exceptional μ -null sets where the assumptions are allowed not

to hold. For instance, one could introduce a μ -null set N such that $B(t)$, introduced above, is allowed to be nonclosed whenever $t \in N$. It is easy to see that the inclusion of such provisions does not lead to more general results, since our results will always be such that they hold modulo μ -null sets.

Consider the *modified Lagrangian* l on $G \times X \times \mathbb{R}^{r+1}$ defined by

$$(2) \quad l(t, x, \xi, \lambda) \equiv \begin{cases} \inf \{ \eta \in \mathbb{R} : (\xi, \eta) \in \tilde{Q}(t, x), \eta \geq \lambda \} & \text{if } (t, x) \in \mathcal{B}, \\ +\infty & \text{otherwise,} \end{cases}$$

it being understood that the infimum over the empty set equals $+\infty$, $(t, x, \xi, \lambda) \in G \times X \times \mathbb{R}^{r+1}$. Note that the classical Lagrangian corresponds to the improper choice $\lambda = -\infty$ and is well known [6c], [9], [10], [14], [18], [19]. A very important aspect of this definition is the following.

Remark 4. Observe that for every $(t, x, \xi, \lambda) \in G \times X \times \mathbb{R}^{r+1}$, finiteness of the number $l(t, x, \xi, \lambda)$ implies that $(t, x) \in \mathcal{B}$ and that the infimum in (2) is *attained*; the latter fact follows immediately from the closedness of the set $\tilde{Q}(t, x)$.

We need additional conventions. The multifunction \tilde{Q} is said to have the (*upper set*) *property* (π) if for every $(t, x) \in \mathcal{B}$, $(\xi, \eta) \in \tilde{Q}(t, x)$ and $\eta' \geq \eta$ we have that $(\xi, \eta') \in \tilde{Q}(t, x)$; cf. [6], [7]. Also, we shall say that *condition* (M) holds if G is a metrizable Lusin space having \mathcal{G} as its Borel σ -algebra. A version of the following result was announced in [18, p. 15]. Cf. the appendix for a different approach.

THEOREM 1. *Suppose that condition (M) holds or that the multifunction \tilde{Q} has property (π) . Then*

- (i) *the modified Lagrangian l is measurable on $G \times X \times \mathbb{R}^{r+1}$,*
- (ii) *the modified Lagrangian l is a normal integrand on $G \times (X \times \mathbb{R}^{r+1})$ if and only if \tilde{Q} has property (K) with respect to \mathcal{B} .*

Proof. (i) To begin with, suppose that G is metrizable Lusin having \mathcal{G} as its Borel σ -algebra. For a given $\gamma \in \mathbb{R}$, define F to be the set of those $(t, x, \xi, \lambda) \in G \times X \times \mathbb{R}^{r+1}$ such that $l(t, x, \xi, \lambda) \leq \gamma$.

By Remark 4, $(t, x, \xi, \lambda) \in F$ if and only if $\lambda \leq \gamma$, $(t, x) \in \mathcal{B}$ and $(\xi, \eta) \in \tilde{Q}(t, x)$ for some $\eta \in \mathbb{R}$, $\lambda \leq \eta \leq \gamma$. Consider the set E of all $(t, x, \xi, \lambda, \eta) \in G \times X \times \mathbb{R}^{r+2}$ such that $(t, x) \in \mathcal{B}$, $(\xi, \eta) \in \tilde{Q}(t, x)$ and $\lambda \leq \eta \leq \gamma$. Since the graph of \tilde{Q} is measurable, E is obviously measurable. At every $(t, x, \xi, \lambda) \in G \times X \times \mathbb{R}^{r+1}$, the section of E , i.e., the set $\{ \eta \in [\lambda, \gamma] : (\xi, \eta) \in \tilde{Q}(t, x) \}$, is compact. In view of the suppositions, we may apply the Kunugui–Novikov–Stchegolkov projection theorem [4, Theorem 1, Remark 2], [8, III.21b] to conclude that the projection F of E onto $G \times X \times \mathbb{R}^{r+1}$ is measurable.

Alternatively, suppose that \tilde{Q} has property (π) . Let $\gamma \in \mathbb{R}$ be arbitrary and define E, F as above. With property (π) we now have that $(t, x, \xi, \lambda) \in F$ if and only if $\lambda \leq \gamma$, $(t, x) \in \mathcal{B}$ and $(\xi, \gamma) \in \tilde{Q}(t, x)$. Thus F , being now the section of E at γ , is measurable.

(ii) *If.* The question of measurability has been settled in (i). Fix $t \in G$ and let $\{(x^k, \xi^k, \lambda^k)\}_0^\infty$ be an arbitrary sequence in $X \times \mathbb{R}^{r+1}$ such that $x^k \rightarrow x^0$, $\xi^k \rightarrow \xi^0$, $\lambda^k \rightarrow \lambda^0$. To prove lower semicontinuity of $l(t, \cdot, \cdot, \cdot, \cdot)$ it will be enough to consider only the nontrivial case where $\beta \equiv \liminf_k l(t, x^k, \xi^k, \lambda^k)$ is finite. Rather than extracting a suitable subsequence, we may suppose w.l.o.g. that the numbers $\eta^k \equiv l(t, x^k, \xi^k, \lambda^k)$ are all finite and that $\eta^k \rightarrow \beta$. By Remark 4 we now have $(\xi^k, \eta^k) \in \tilde{Q}(t, x^k)$, $x^k \in B(t)$ and $\eta^k \geq \lambda^k$ for every $k \in \mathbb{N}$. By closedness of $B(t)$, $x^0 \in B(t)$ it follows from the above that for every $\gamma > 0$, there exists $k_0 \in \mathbb{N}$ such that $(\xi^k, \eta^k) \in \bigcup \{ \tilde{Q}(t, x) : x \in B(t), d(x, x^0) < \gamma \}$ for every $k \geq k_0$. But then (ξ^0, β) will lie in the closure of the above union. Our supposition implies that $(\xi^0, \beta) \in \tilde{Q}(t, x^0)$. Also $\beta = \lim_k \eta^k \geq \lim_k \lambda^k = \lambda^0$. It is now clear from (2) that $l(t, x^0, \xi^0, \lambda^0) \leq \beta$, so lower semicontinuity of $l(t, \cdot, \cdot, \cdot, \cdot)$ has been proven.

Only if. Fix $(t, x^0) \in \mathcal{B}$ arbitrarily. Only the nontrivial inclusion in (1) deserves proof. Suppose $(\xi^0, \eta^0) \in \mathbb{R}^{r+1}$ belongs to the right side of (1). For every $k \in \mathbb{N}$ the ball with radius k^{-1} around (ξ^0, η^0) contains, for some $x^k \in B(t)$ with $d(x^k, x^0) < k^{-1}$, an element $(\xi^k, \eta^k) \in \tilde{Q}(t, x^k)$. Clearly, $(x^k, \xi^k, \eta^k) \rightarrow (x^0, \xi^0, \eta^0)$. By (2), $\eta^k \geq l(t, x^k, \xi^k, \eta^k)$ for every $k \in \mathbb{N}$. Thus, the supposition yields $\eta^0 \geq l(t, x^0, \xi^0, \eta^0) \geq \eta^0$. By Remark 4 we have now $(\xi^0, \eta^0) \in \tilde{Q}(t, x^0)$. This finishes the proof.

The usual representation of \tilde{Q} to be found in optimal control problems of Lagrange type is as follows. Let Ω be a measurable closed-valued multifunction from \mathcal{B} into a σ -compact metric space U , equipped with its Borel σ -algebra. Also, let $\tilde{f} \equiv (f, f^0)$ be a measurable function from $H \equiv \{(t, x, u) \in \mathcal{B} \times U : u \in \Omega(t, x)\}$ into \mathbb{R}^{r+1} and such that, for every $t \in G$, $\tilde{f}(t, \cdot, \cdot)$ is continuous on the section of H at t . Suppose now that for every $(t, x) \in \mathcal{B}$,

$$(3) \quad \tilde{Q}(t, x) = \{(f(t, x, u), \eta) \in \mathbb{R}^{r+1} : \eta \geq f^0(t, x, u), u \in \Omega(t, x)\},$$

with the implicit understanding that the set on the right-hand side is closed.

LEMMA 2. Suppose that \tilde{Q} allows the representation (3). Then \tilde{Q} is measurable, closed-valued and has property (π) .

Proof. Because Ω is measurable and has complete values, there exists a sequence $\{\omega_i\}_1^\infty$ of measurable functions from \mathcal{B} into U such that for every $(t, x) \in \mathcal{B}$ $\Omega(t, x) = \text{cl} \{\omega_i(t, x)\}_1^\infty$ whenever $\Omega(t, x)$ is nonempty (Castaing representation, [5, III.9]). Fix $(t, x) \in \mathcal{B}$ and suppose $\Omega(t, x)$ is nonempty. Then by the continuity property of \tilde{f}

$$(4) \quad \tilde{Q}(t, x) \subset \text{cl} \{(f(t, x, \omega_i(t, x)), f^0(t, x, \omega_i(t, x))) + r_j : i, j \in \mathbb{N}\}.$$

Here $\{r_j\}_1^\infty$ denotes a denumeration of the nonnegative rationals. By supposition $\tilde{Q}(t, x)$ is closed, so the converse inclusion in (4) holds too. Because \tilde{f} is measurable on H , a Castaing representation of \tilde{Q} has been obtained, which is equivalent to saying that \tilde{Q} is measurable [5, III.9]. The other statements are trivially true.

Note that if \tilde{Q} allows representation (3), the modified Lagrangian can be specified by

$$(5) \quad l(t, x, \xi, \lambda) = \inf \{\max \{f^0(t, x, u), \lambda\} : u \in \Omega(t, x), f(t, x, u) = \xi\},$$

for $(t, x) \in \mathcal{B}$, $(\xi, \lambda) \in \mathbb{R}^{r+1}$. In analogy to Remark 4 we have the following.

Remark 5. In case \tilde{Q} has representation (3), for every $(t, x, \xi, \lambda) \in G \times X \times \mathbb{R}^{r+1}$, finiteness of $l(t, x, \xi, \lambda)$ implies that $(t, x) \in \mathcal{B}$ and that the infimum in (5) is attained.

In control problems with uniform boundedness conditions on the controls, the following alternative Lagrangian will prove to be very useful. Let h be a nonnegative measurable functional on $G \times U$ such that $h(t, \cdot)$ is inf-compact on U for every $t \in G$. Define for every $\varepsilon \geq 0$ the approximate Lagrangian l_ε on $G \times X \times \mathbb{R}^{r+1}$ by

$$(6) \quad l_\varepsilon(t, x, \xi, \lambda) = \inf \{\max \{f^0(t, x, u), \lambda\} + \varepsilon h(t, u) : u \in \Omega(t, x), f(t, x, u) = \xi\},$$

for $(t, x) \in \mathcal{B}$, $(\xi, \lambda) \in \mathbb{R}^{r+1}$. Elsewhere, set l_ε equal to $+\infty$.

Note that the modified Lagrangian l of (5) is equal to l_0 . Note also that convexity of the set $\tilde{Q}(t, x)$, $(t, x) \in \mathcal{B}$, in (3) does not imply that $l_\varepsilon(t, x, \cdot, \cdot)$ is convex, unless $\varepsilon = 0$ and $h(t, \cdot)$ is convex.

Let us say that \tilde{Q} has property (K) with respect to \mathcal{B} if (1) holds at every $(t, x^0) \in \mathcal{B}$, with \tilde{Q} replaced by Ω .

In anticipation of what is to follow we urge the reader to distinguish carefully between normal integrands on $G \times (X \times \mathbb{R}^{r+1})$ and normal integrands on $(G \times X) \times \mathbb{R}^{r+1}$; cf. [5, VII].

LEMMA 3. Suppose that condition (M) holds and that \tilde{Q} has representation (3). The following statements hold for every $\varepsilon > 0$:

- (i) The approximate Lagrangian l_ε is measurable on $G \times X \times \mathbb{R}^{r+1}$.
- (ii) The approximate Lagrangian l_ε is a normal integrand on $G \times (X \times \mathbb{R}^{r+1})$ if Ω has property (K) with respect to \mathcal{B} .

Also, for $\varepsilon = 0$ (cf. Theorem 1(ii))

- (iii) The modified Lagrangian l is a normal integrand on $(G \times X) \times \mathbb{R}^{r+1}$.

Proof. (i) Fix $\varepsilon > 0$ arbitrarily. To start with, note that for every $(t, x, \xi, \lambda) \in G \times X \times \mathbb{R}^{r+1}$ finiteness of $l_\varepsilon(t, x, \xi, \lambda)$ implies that $(t, x) \in \mathcal{B}$ and that the infimum is attained in (6) (by inf-compactness of $h(t, \cdot)$, continuity of $\tilde{f}(t, x, \cdot)$ and closedness of $\Omega(t, x)$).

Let $\gamma \in \mathbb{R}$ be given. It is easy to check that for every $(t, x, \xi, \lambda) \in G \times X \times \mathbb{R}^{r+1}$ $l_\varepsilon(t, x, \xi, \lambda) \leq \gamma$ if and only if $\lambda \leq \gamma$, $(t, x) \in \mathcal{B}$ and there exists $u \in \Omega(t, x)$ such that $h(t, u) \leq \varepsilon^{-1}(\gamma - \lambda)$, $f(t, x, u) = \xi$ and $\max(f^0(x, t, u), \lambda) + \varepsilon h(t, u) \leq \gamma$; observe that the set of $u \in \Omega(t, x)$ satisfying these three relations is compact. The rest of the argument runs entirely parallel to that used in proving Theorem 1(i) by means of the Kunugui-Novikov-Stchegolkov theorem. We shall omit the details.

(ii) Let $\varepsilon > 0$ be fixed. From (i) we know that l_ε is measurable. Let $t \in G$ be arbitrary. We will show that $l_\varepsilon(t, \cdot, \cdot, \cdot)$ is lower semicontinuous. Consider an arbitrary given sequence $\{(x^k, \xi^k, \lambda^k)\}_0^\infty$ such that $x^k \rightarrow x^0$, $\xi^k \rightarrow \xi^0$, $\lambda^k \rightarrow \lambda^0$. Without loss of generality we may assume that $\beta \equiv \liminf_k l_\varepsilon(t, x^k, \xi^k, \lambda^k)$ is finite and that $\eta^k \rightarrow \beta$, where $\eta^k \equiv l_\varepsilon(t, x^k, \xi^k, \lambda^k)$ is finite, $k \in \mathbb{N}$. Thus, for every $k \in \mathbb{N}$ there exists $u^k \in \Omega(t, x^k)$ such that

$$\eta^k = \max(f^0(t, x^k, u^k), \lambda^k) + \varepsilon h(t, u^k), \quad f(t, x^k, u^k) = \xi^k.$$

Because $h(t, u^k) \leq \varepsilon^{-1}(\eta^k - \lambda^k)$, it is easy to see that a subsequence of $\{u^k\}_1^\infty$ will converge to some $\bar{u} \in U$. Without loss of generality we may assume $u^k \rightarrow \bar{u}$. Since $\bar{u} \in \text{cl } \bigcup_1^\infty \Omega(t, x^k)$ we conclude from the supposition that $\bar{u} \in \Omega(t, x^0)$. Also, by continuity of $\tilde{f}(t, \cdot, \cdot)$, we find $f(t, x^0, \bar{u}) = \xi^0$ and $\beta \geq \max(f^0(t, x^0, \bar{u}), \lambda^0) + \varepsilon h(t, \bar{u})$. In all, we find now that $l_\varepsilon(t, x^0, \xi^0, \lambda^0) \leq \beta$.

(iii) In view of Lemma 2, measurability of l is assured by Theorem 1(i). Let the element $(t, x) \in \mathcal{B}$ and the sequence $\{(\xi^k, \lambda^k)\}_0^\infty$ be arbitrary, $\xi^k \rightarrow \xi^0$, $\lambda^k \rightarrow \lambda^0$. As usual we may suppose w.l.o.g. that all numbers $\eta^k \equiv l(t, x, \xi^k, \lambda^k)$ are finite and converge to the finite number $\beta \equiv \liminf_k \eta^k$. Rather than using the representation (5) for l , we shall keep to the original definition given in (2). By Remark 4 we have that for every $k \in \mathbb{N}$ $(\xi^k, \eta^k) \in \tilde{Q}(t, x)$, $\eta^k \geq \lambda^k$. By closedness of $\tilde{Q}(t, x)$ we thus have that $(\xi^0, \beta) \in \tilde{Q}(t, x)$. Also $\beta \geq \lambda^0$, so from (2) we conclude that $l(t, x, \xi^0, \lambda^0) \leq \beta$. This finishes the proof.

We now turn to the description of the only tool used in the second step of the deparametrization procedure: a lower semicontinuity result for integral functionals. Before doing so, we again introduce a part of the basic framework, namely a sequence $\{x_k\}_0^\infty$ of measurable functions from G into X such that

$$(7) \quad x_k \xrightarrow{\mu} x_0,$$

where $\xrightarrow{\mu}$ denotes convergence in measure μ . In this section we shall also consider a sequence $\{\tilde{\xi}_k\}_0^\infty$ belonging to the collection $L_1(G)^{r+1}$ of measurable, μ -integrable functions from G into \mathbb{R}^{r+1} such that

$$\tilde{\xi}_k \xrightarrow{\sigma} \tilde{\xi}_0,$$

where $\xrightarrow{\sigma}$ denotes convergence in the weak $\sigma(L_1(G)^{r+1}, L_\infty(G)^{r+1})$ -topology. A lower

semicontinuity result for integral functionals will now be stated. Variants of it appear in [2c], [6f], [10], [14], [18], to mention just a few references of this venerable result.

THEOREM 4. *Suppose m is a normal integrand on $G \times (X \times \mathbb{R}^{r+1})$ such that*

- (a) *the sequence $\{m^-(\cdot, x_k(\cdot), \tilde{\xi}_k(\cdot))\}_1^\infty$ is uniformly integrable,*
- (b) *for every $t \in G$ the functional $m(t, x_0(t), \cdot)$ is convex on \mathbb{R}^{r+1} .*

Then

$$\liminf_k \int_G m(t, x_k(t), \tilde{\xi}_k(t)) \mu(dt) \geq \int_G m(t, x_0(t), \tilde{\xi}_0(t)) \mu(dt).$$

Proof. For finite-dimensional G, X the result is practically contained in [13, Theorem 1]. Stated as is, the result follows directly from combining [16, Theorem 5, Remark 2, Case 1].

Remark 6. Note that the inf-compactness result in [19] follows immediately—and under weaker conditions—from Theorem 4. (The so-called “basic growth condition” imposed upon the Hamiltonian in [19] implies the weak convergence of the derivatives in a far more direct way than might appear from that paper. It is enough to note that the growth property for the Lagrangian implied by the basic growth condition (cf. [19, p. 321]) is such that the derivatives involved are equi-absolutely continuous.)

3. The abstract (lower) closure problem. The deparametrization procedure introduced in § 1 will now be implemented. Thus we follow through with the first step, worked out in the previous section. We obtain a quite general lower closure and closure result. We shall also discuss how these results and those of the following section generalize comparable results in [2a–b], [6e–f], [7a–c], [21].

Let $G, X, r, \{x_k\}_0^\infty$ be as in the previous section (with (7) holding). The multifunction B will be specialized as follows from now on:

$$(8) \quad B(t) = \text{cl} \{x_k(t)\}, \quad t \in G.$$

with this proviso, \mathcal{B}, \tilde{Q} and l are supposed to be as before. Also, a sequence $\{\xi_k\}_0^\infty$ in $L_1(G)^r$, with

$$(9) \quad \xi_k \xrightarrow{\sigma} \xi_0,$$

will be part of the basic framework from now on. The main lower closure result is now as follows.

THEOREM 5. *Suppose that condition (M) holds or that \tilde{Q} has property (π) . Also, suppose*

$$(10) \quad \text{for every } t \in G \text{ the set } \tilde{Q}(t, x_0(t)) \text{ is convex,}$$

$$(11) \quad \tilde{Q} \text{ has property (K) with respect to } \mathcal{B}.$$

Let $\{\eta_k\}_1^\infty$ be a sequence of measurable functionals on G and suppose there exists a sequence $\{(\delta_k, \delta_k^0)\}_1^\infty$ of measurable functions from G into \mathbb{R}^{r+1} , converging in measure μ to zero, such that

$$(12) \quad \text{for every } t \in G, k \in \mathbb{N}, \quad (\xi_k(t) + \delta_k(t), \eta_k(t) + \delta_k^0(t)) \in \tilde{Q}(t, x_k(t)).$$

Suppose further that there exists a uniformly integrable sequence $\{\lambda_k\}_1^\infty$ in $L_1(G)$ such that

$$(13) \quad \text{for every } k \in \mathbb{N}, \quad \eta_k \geq \lambda_k.$$

Then, if $i \equiv \liminf_k \int_G \eta_k d\mu < +\infty$, there exists $\eta^ \in L_1(G)$ such that*

- (i) *for μ -a.e. t , $(\xi_0(t), \eta^*(t)) \in \tilde{Q}(t, x_0(t))$, $x_0(t) \in B(t)$,*
- (ii) *$\int_G \eta^* d\mu \leq i$.*

Proof. Instead of taking a suitable subsequence twice, we may suppose w.l.o.g. that $i = \lim_k \int_G \eta_k d\mu$ and that there exists $\lambda_0 \in L_1(G)$ such that $\lambda_k \xrightarrow{g} \lambda_0$ (by the Dunford–Pettis compactness criterion [8, II.22, 25], [10, VII.1.7]). We claim that $\eta^* \equiv l(\cdot, x_0(\cdot), \xi_0(\cdot), \lambda_0(\cdot))$ satisfies (i), (ii), where l denotes the modified Lagrangian defined in (2). We have for every $t \in G$, $k \in \mathbb{N}$,

$$(14) \quad l(t, x_k(t), (\xi_k + \delta_k)(t), (\lambda_k + \delta_k^0)(t)) - \delta_k^0(t) \geq \lambda_k(t).$$

Because of the initial assumption and (11), l is a normal integrand on $G \times (X \times \mathbb{R}^{r+1})$ by Theorem 1. Therefore

$$m(t, x, \delta, \delta^0, \varepsilon, \xi, \lambda) \equiv l(t, x, \xi + \delta, \lambda + \delta^0) - \varepsilon,$$

defined for $t \in G$, $x \in X$, $\delta, \xi \in \mathbb{R}^r$, $\delta^0, \varepsilon, \lambda \in \mathbb{R}$, gives a normal integrand m on $G \times (X \times \mathbb{R}^{2r+3})$. From (2) and (10) it follows directly that $l(t, x_0(t), \cdot, \cdot)$ is convex on \mathbb{R}^{r+1} for every $t \in G$. Thus, as for every $t \in G$, $(\xi, \lambda) \in \mathbb{R}^{r+1}$,

$$(15) \quad m(t, x_0(t), 0, 0, 0, \xi, \lambda) = l(t, x_0(t), \xi, \lambda),$$

m satisfies condition (b) imposed in Theorem 4. Further, by (14) we have for every $k \in \mathbb{N}$

$$m_k \geq \lambda_k,$$

where $m_k(t)$ is defined to be the left side of (14), $t \in G$. Hence, the uniform integrability of the sequence $\{\lambda_k\}^\infty$ implies the uniform integrability of the sequence $\{\max(-m_k, 0)\}_1^\infty$. So condition (a) of Theorem 4 holds as well. If we apply Theorem 4, and take into account (15), we find easily

$$\liminf_k \int_G m_k d\mu \geq \int_G \eta^* d\mu.$$

This implies (ii), since for every $k \in \mathbb{N}$, $m_k \leq \eta_k$ by (2), (8), (12) and (13). Clearly, η^* is measurable. In view of (2), we also have $\eta^* \geq \lambda_0$. Since $i < +\infty$, we conclude that $\eta^* \in L_1(G)$. Since a fortiori η^* is finite μ -a.e., (i) follows by definition of η^* and Remark 4.

Remark 7. It deserves attention that Theorem 5 remains valid in case the assumption $\delta_k^0 \xrightarrow{\mu} 0$ is replaced by $\{\delta_k^0\}_1^\infty \subset L_1(G)$ and $\delta_k^0 \xrightarrow{g} 0$. This is seen by redefining $\lambda'_k \equiv \lambda_k + \delta_k^0$, $\eta'_k \equiv \eta_k + \delta_k^0$, $k \in \mathbb{N}$ in Theorem 6 and observing that $\lambda'_k \xrightarrow{g} \lambda_0$, $i = \liminf_k \int_G \eta'_k d\mu$ and $\eta'_k \geq \lambda'_k$ for every $k \in \mathbb{N}$. Mutatis mutandis, the same can be said for $\{\delta_k\}_0^\infty$. Conversely, the assumption $\xi_k \xrightarrow{g} \xi_0$ can be replaced by $\xi_k \xrightarrow{\mu} \xi_0$, in which case assumption (10) can be omitted.

The main closure result follows immediately from Theorem 5. Let Q be a multifunction from \mathcal{B} into \mathbb{R}^r with closed values and measurable graph.

We shall say that Q has property (K) with respect to \mathcal{B} if (1), with Q substituted for \tilde{Q} , holds at every point $(t, x^0) \in \mathcal{B}$.

THEOREM 6. *Suppose that*

$$(16) \quad \text{for every } t \in G \text{ the set } Q(t, x_0(t)) \text{ is convex,}$$

$$(17) \quad Q \text{ has property (K) with respect to } \mathcal{B}.$$

Also, suppose that there exists a sequence $\{\delta_k\}_1^\infty$ of measurable functions from G into \mathbb{R}^r , converging in measure μ to zero, such that

$$(18) \quad \text{for every } t \in G, k \in \mathbb{N}, \quad \xi_k(t) + \delta_k(t) \in Q(t, x_k(t)).$$

Then for μ -a.e. t , $\xi_0(t) \in Q(t, x_0(t))$ and $x_0(t) \in B(t)$.

Proof. Define the multifunction \tilde{Q} by

$$\tilde{Q}(t, x) \equiv Q(t, x) \times \mathbb{R}^+, \quad (t, x) \in \mathcal{B}.$$

It is easy to see that \tilde{Q} inherits measurability, closedness of its values, (10) and (11) from Q (by (16), (17)). The result now follows from applying Theorem 5 with $\lambda_k \equiv 0$, $\eta_k \equiv 0$, $\delta_k^0 \equiv 0$ for all $k \in \mathbb{N}$.

Remark 8. As was explained in Remark 7, the assumption $\delta_k \xrightarrow{\mu} 0$ can be replaced by $\{\delta_k\}_1^\infty \subset L_1(G')$ and $\delta_k \xrightarrow{\sigma} 0$.

Remark 9. A simple auxiliary Lagrangian can be formulated for the closure problem itself by defining for $(t, x, \xi) \in G \times X \times \mathbb{R}^r$

$$l'(t, x, \xi) \equiv \text{dist}(\xi, Q(t, x)),$$

where dist denotes Euclidean distance. One could paraphrase Theorem 1 for l' and Q .

Let us compare the abstract lower closure results of [6e–f], [7c], [21] with Theorem 5. In all of these references, G and X are finite dimensional, \mathcal{G} is the Lebesgue σ -algebra and μ the Lebesgue measure. A remarkable aspect of the framework used there is that the multifunction \tilde{Q} is not required to be measurable. (The arguments involving this multifunction are held pointwise as a rule.) Of course, if one keeps the nature of the applications in mind, the loss of generality involved in our taking \tilde{Q} to be measurable, is negligible indeed. The following statements all follow from Theorem 5 above; for each result we briefly sketch the principal differences with Theorem 5. In [6e, Statement (5i)] property (Q) with respect to x is required; further, $\delta_k = 0$, $\delta_k^0 = 0$. Also, Statement (5ii) follows by Remark 7. In [6f, Statement (6i)], boundedness and convexity conditions imposed on the classical Lagrangian $l(\cdot, \cdot, \cdot, -\infty)$ imply directly that $\eta_k \geq \alpha + \beta|\xi_k|$ for certain constants α, β [6f, p. 392], so that (13) certainly holds. Also, property (π) has been imposed and $\delta_k = 0$, $\delta_k^0 = 0$. In [7c, Theorem 4.1] one has $x_k = x_0$, so that (11) holds vacuously by (8) and Remark 1. In [21, Theorem 3.1], property (Q) with respect to x is avoided by using a condition whose unnaturalness is already evident from [21, Remarks 3.1, 3.2]. Besides, property (π) has been used and the convergence $\lambda_k \rightarrow \lambda_0$ takes place in L_1 -norm (cf. [21, Remark 3.5]). Also, $\delta_k = 0$, $\delta_k^0 = 0$.

In the same vein we could elaborate on the claim that Theorem 6 generalizes [6e, Statements (4i–iii)], [6f, Statement (5i)], [7c, Theorem 3.1] and [21, Theorem 3.2]. This will be left to the reader, however.

4. Closure problems of Lagrange and Mayer type. In this section we shall deal with the lower closure problem in case the representation (3) is valid, as happens in optimal control problems of Lagrange type. Let $G, X, r, \{x_k\}_0^\infty, B, \xi_0$ be as before (with (7), (8) holding). From now on we shall assume that the representation (3) holds for \tilde{Q} . Also, we shall include in the framework a sequence $\{u_k\}_1^\infty$ of measurable functions from G into U such that

$$(19) \quad \begin{aligned} &\text{for } \mu\text{-a.e. } t, u_k(t) \in \Omega(t, x_k(t)), \\ &\{f(\cdot, x_k(\cdot), u_k(\cdot))\}_1^\infty \subset L_1(G)', \quad f(\cdot, x_k(\cdot), u_k(\cdot)) \xrightarrow{\sigma} \xi_0. \end{aligned}$$

We shall state two immediate consequences of Theorem 5.

THEOREM 7. *Suppose assumptions (10), (11) regarding \tilde{Q} are valid. Also, suppose that*

$$(20) \quad \text{for every } t \in G, k \in \mathbb{N}, \quad u_k(t) \in \Omega(t, x_k(t)),$$

and that there exists a uniformly integrable sequence $\{\lambda_k\}_1^\infty$ in $L_1(G)$ such that

$$(21) \quad \text{for every } k \in \mathbb{N}, \quad f^0(\cdot, x_k(\cdot), u_k(\cdot)) \geq \lambda_k.$$

Then, if $i \equiv \liminf_k \int_G f^0(t, x_k(t), u_k(t)) \mu(dt) < +\infty$, there exists a measurable function u^* from G into U such that $f^0(\cdot, x_0(\cdot), u^*(\cdot)) \in L_1(G)$ and

- (i) for μ -a.e. t , $x_0(t) \in B(t)$, $u^*(t) \in \Omega(t, x_0(t))$, $\xi_0(t) = f(t, x_0(t), u^*(t))$,
- (ii) $\int_G f^0(t, x_0(t), u^*(t)) \mu(dt) \leq i$.

Proof. By Lemma 2 we may apply Theorem 5 with $(\xi_k, \eta_k) \equiv \tilde{f}(\cdot, x_k(\cdot), u_k(\cdot))$, $\delta_k = 0$, $\delta_k^0 = 0$, $k \in \mathbb{N}$, because assumption (12) holds by (3), (20). So there exists $\eta^* \in L_1(G)$ for which conclusions (i), (ii) of Theorem 5 hold. In view of Remark 5, the result now follows from invoking a standard measurable implicit function result [12, Theorem 7.1] (cf. [22, p. 876]).

THEOREM 8. Suppose that assumptions (10) and (20) hold, that

$$(22) \quad \text{for every } t \in G, k \in \mathbb{N}, \quad \Omega(t, x_k(t)) = \Omega(t, x_0(t)),$$

and that

$$(23) \quad \tilde{\delta}_k \xrightarrow{\mu} 0,$$

where $\tilde{\delta}_k(t) \equiv \tilde{f}(t, x_0(t), u_k(t)) - \tilde{f}(t, x_k(t), u_k(t))$, $t \in G$, $k \in \mathbb{N}$.

Suppose further that there exists a uniformly integrable sequence $\{\lambda_k\}_1^\infty$ in $L_1(G)$ such that (21) holds.

Then the conclusions of Theorem 7 are true.

Proof. Let \bar{x}_k denote x_k as used in Theorem 5 and set $\bar{x}_k = x_0$, $k \in \mathbb{N}$. So assumption (9) of Theorem 5 is satisfied (by Remark 1). By (22), (23) assumption (12) will hold with $(\xi_k, \eta_k) \equiv \tilde{f}(\cdot, x_k(\cdot), u_k(\cdot))$. Now apply Theorem 5 and finish the proof exactly as described in the previous proof.

Remark 10. Regarding the mode of convergence of $\{\tilde{\delta}_k\}_1^\infty$, the same can be said as in Remark 7.

In [7b] (cf. [2a, 20]) a number of conditions of Lipschitz and growth-type were formulated, all of which guarantee that $\tilde{\delta}_k \rightarrow 0$ in the L_1 -norm (hence certainly imply (23)). Later, it was recognized that all conditions of [7b] involving a uniform L_1 -bound for $\{u_k\}_1^\infty$ can be relaxed substantially [2b, 7c].

We shall say more on this, but not before having stated an obvious relaxation of the only conditions in [7b] that do not impose a uniform L_1 -bound on $\{u_k\}_1^\infty$, viz. conditions (F_p) , $1 \leq p \leq \infty$.

PROPOSITION 9. Suppose that \tilde{f} has the following uniform continuity property (F):

$$\begin{aligned} &\text{for every } t \in G, \text{ every sequence } \{x^k\}_0^\infty \text{ in } B(t), x^k \rightarrow x^0, \\ &\tilde{f}(t, x^k, u) - \tilde{f}(t, x^0, u) \rightarrow 0 \text{ uniformly in } u \in \Omega(t, x_0(t)). \end{aligned}$$

Then we may suppose w.l.o.g. in Theorem 8 that (23) holds.

Proof. Because of (7), it is enough to extract a subsequence of $\{x_k\}_1^\infty$ which converges to x_0 μ -a.e., and to invoke property (F). It is obvious from the proof of Theorem 5 that the argument now runs as if (23) were valid for the whole sequence $\{\tilde{\delta}_k\}_1^\infty$.

Let us next turn to conditions involving uniform bounds for $\{u_k\}_1^\infty$. Rather than taking the winding path of formulating more and more intricate analogues of ‘‘Nemitsky’s theorem’’ [7c, Theorem 2.2], [16, Lemma 2.1], we shall follow the avenue lined by the approximate Lagrangian l_ϵ of § 2. Thus, we can reach substantial generalizations of the results in [2b], [7c].

Let h be as in § 2, i.e. a nonnegative, measurable functional on $G \times U$ such that $h(t, \cdot)$ is inf-compact on U for every $t \in G$. The result below can replace Theorem 7 in the presence of uniform bounds.

THEOREM 10. *Suppose that condition (M) holds and that assumptions (10) and (20) are valid. Also, suppose that*

$$(24) \quad \text{there exists } \gamma \geq 0 \text{ such that } \int_G h(t, u_k(t)) \mu(dt) \leq \gamma \text{ for every } k \in \mathbb{N},$$

$$(25) \quad \Omega \text{ has property (K) with respect to } \mathcal{B}.$$

and that there exists a uniformly integrable sequence $\{\lambda_k\}_1^\infty$ in $L_1(G)$ such that assumption (21) holds.

Then the conclusions of Theorem 7 are true.

Proof. Let $\varepsilon > 0$ be arbitrary and l_ε be as in (6). By (6), (20), (21), (24) we find for every $t \in G$, $k \in \mathbb{N}$

$$(26) \quad \int_G f^0(t, x_k(t), u_k(t)) \mu(dt) + \varepsilon \gamma \geq \int_G l_\varepsilon(t, x_k(t), \xi_k(t), \lambda_k(t)) \mu(dt),$$

where $\xi_k(t) \equiv f(t, x_k(t), u_k(t))$. Also, for μ -a.e. t and every $k \in \mathbb{N}$, by nonnegativity of h and (2).

$$(27) \quad l_\varepsilon(t, x_k(t), \xi_k(t), \lambda_k(t)) \geq l(t, x_k(t), \xi_k(t), \lambda_k(t)) \geq \lambda_k(t).$$

By assumption (25) and Lemma 3(ii) we have that l_ε is a normal integrand on $G \times (X \times \mathbb{R}^{r+1})$. By Lemma 3(iii), l is a normal integrand on $(G \times X) \times \mathbb{R}^{r+1}$. Let us define for $(t, x, \xi, \lambda) \in G \times X \times \mathbb{R}^{r+1}$

$$m(t, x, \xi, \lambda) \equiv \begin{cases} l_\varepsilon(t, x, \xi, \lambda) & \text{if } x \neq x_0(t), \\ l(t, x_0(t), \xi, \lambda) & \text{otherwise.} \end{cases}$$

Since $l_\varepsilon \geq l$, it follows from the above that m is a normal integrand on $G \times (X \times \mathbb{R}^{r+1})$. Assumption (a) of Theorem 4 is valid by (27) and assumption (b) holds by (10), in view of (2) and the definition of m . By (7), (19) we may apply Theorem 4. In conjunction with (26), (27) this gives

$$(28) \quad \liminf_k \int_G f^0(t, x_k(t), u_k(t)) \mu(dt) \geq \int_G l(t, x_0(t), \xi_0(t), \lambda_0(t)) \mu(dt) - \varepsilon \gamma.$$

In view of the arbitrary choice of ε , (28) holds with $\varepsilon = 0$. We can now proceed exactly as in Theorems 5, 7, making use of Remark 5 and the cited implicit measurable function result.

Remark 11. The above result shows that, under the uniform boundedness assumption (24), the seminormality condition (11) for \tilde{Q} in Theorem 7 can be replaced by a seminormality condition for Ω . In [2b], [7c] and related work, the following assumption is always imposed:

$$\text{for every } t \in G, x, x' \in B(t), \quad \Omega(t, x) = \Omega(t, x').$$

Under this extra assumption and the boundedness assumption, Theorem 10 also entails Theorem 8. Indeed, (20) follows now trivially from (22), and the extra assumption implies (25), so actually assumption (23) is then dispensable.

The lower closure results for optimal control problems of Lagrange type derived in this section can automatically be turned into closure results for optimal control

problems of Mayer type by setting $f^0 = 0$, $\lambda_k = 0$, $k \in \mathbb{N}$. We shall not spell out this conversion of statements.

Let us compare the above theorems with similar results obtained in [2a–b], [6e–f], [7b–c] (these include all relevant results recapitulated in [7a]). Modulo a μ -null set, the basic condition (C) used in these references is easily seen to imply our basic conditions involving \tilde{f} and Ω . (Invoke the Scorza–Dragoni theorem and use [5, III.30]; cf. Remark 3.) An exception to this is made by [2a–b], where f^0 is supposed to be lower semicontinuous and f continuous. In that case Lemmas 2, 3 can be shown to hold by making use of the σ -compactness of U ; the details of this are left to the reader. [6e, Statements (7.i), (7.ii)] follow from Theorem 7. Note that property (Q) with respect to x is used in (7.i); cf. Remark 7. Also [6f, Statement (6.ii)] follows from Theorem 7 by the observations we made when discussing [6f, Statement (6.i)]. [2a, Theorem 1] also follows from Theorem 7. There assumption (21) is satisfied by the expression [2a, (5.3), p. 33]. Theorem 2 of that reference follows from Theorem 8 and Proposition 9, since \tilde{f} has property (F) there. [2b, Theorem 3.1] is easily seen to be a consequence of Theorem 10. Also, the various substatements in [7b, Theorem 8.1] all follow from Theorem 8 together with Proposition 9 or from Theorem 10.

Finally, [7c, Theorem 4.1] practically coincides, for the simpler setting of [7c], with Theorem 8.

Along the same lines, these comments could be repeated for the closure results in control problems of Mayer type, whose presence here can be considered implicit by what was said above.

5. Epilogue. The main tenet of this paper states that (lower) closure results for the class of closure problems with weak convergence conditions upon the derivatives are basically all the manifestation of a single lower semicontinuity result, viz. Theorem 4. In [1b] it was demonstrated that Theorem 4 can be obtained in a very direct way by a compactification-relaxation approach, entirely in the spirit of L. C. Young's work; cf. [1a], [3], [17], [23]. (In the light of these references, observe that, by de la Vallée Poussin's theorem, the weak convergence condition characterizing the class of closure problems studied here, involves an implicit compactness condition.) In view of this, we hope that the above efforts may contribute to the unification of the existence theory for optimal control problems.

Appendix. In case the measure space (G, \mathcal{G}, μ) is complete, it is possible to take another approach to proving that modified and approximate Lagrangians are normal integrands. By now this approach, mainly due to R. T. Rockafellar, is entirely standard (e.g., cf. [19] and its references). It gives the following alternatives to Theorem 1 and Lemma 3:

THEOREM 1'. *Suppose that (G, \mathcal{G}, μ) is complete, that X is a Polish space and that \tilde{Q} has property (K) with respect to \mathcal{B} . Then the modified Lagrangian l is a normal integrand on $G \times (X \times \mathbb{R}^{r+1})$.*

LEMMA 3'. *Suppose that (G, \mathcal{G}, μ) is complete, that X is a Polish space, that \tilde{Q} has representation (3) and that Ω has property (K) with respect to \mathcal{B} . Then for every $\varepsilon > 0$ the approximate Lagrangian l_ε is a normal integrand on $G \times (X \times \mathbb{R}^{r+1})$. Also, the modified Lagrangian l is a normal integrand on $(G \times X) \times \mathbb{R}^{r+1}$.*

We shall only sketch a proof of Theorem 1', since that of Lemma 3' is quite similar and equally standard (e.g., cf. [19, p. 317]).

Consider the functional n on $G \times X \times \mathbb{R}^{r+2}$ defined by

$$n(t, x, \xi, \lambda, \eta) \equiv \begin{cases} \eta & \text{if } (t, x) \in \mathcal{B}, \quad (\xi, \eta) \in \tilde{Q}(t, x), \quad \eta \geq \lambda, \\ +\infty & \text{otherwise,} \end{cases}$$

for $(t, x, \xi, \lambda, \eta) \in G \times X \times \mathbb{R}^{r+2}$. It is easy to see that n is a normal integrand on $G \times (X \times \mathbb{R}^{r+2})$ by our suppositions; cf. the proof of Theorem 1(ii). By the completeness of (G, \mathcal{G}, μ) , our notion of a normal integrand on $G \times (X \times \mathbb{R}^{r+2})$ coincides with the one used by Rockafellar. That is to say that the epigraphic multifunction of n from G into $X \times \mathbb{R}^{r+2}$ is measurable; cf. [19, Proposition 2] and note that this result continues to hold in our more general framework. Now in

$$l(t, x, \xi, \lambda) = \inf \{n(t, x, \xi, \lambda, \eta) : \eta \in \mathbb{R}\}$$

the infimum is attained when it is finite (Remark 4). Hence, for every $t \in G$ the (possibly empty) epigraph of $l(t, \cdot, \cdot, \cdot)$ is the projection of the epigraph of $n(t, \cdot, \cdot, \cdot, \cdot)$ onto $X \times \mathbb{R}^{r+1}$. It follows directly from the Castaing representation of the epigraphic multifunction of n that the epigraphic multifunction of l is measurable. Hence, l is a normal integrand in the sense of Rockafellar. This means that l is a normal integrand in the sense used here.

The above comment implies that Theorems 5 and 10 continue to hold if we interpret condition (M) there as follows: either G is metrizable Lusin with Borel σ -algebra \mathcal{G} or X is Polish and \mathcal{G} is the μ -completion of a countably generated σ -algebra.

When reading the comparison of our results with those available in the literature (§§ 3, 4) these observations should be kept in mind.

Acknowledgment. The author is indebted to an unknown referee for suggesting the use of the functional m in the proof of Theorem 10. (The original proof went by way of relaxation.)

REFERENCES

- [1a] E. J. BALDER, *On a useful compactification for optimal control problems*, J. Math. Anal. Appl., 72 (1979), pp. 391–398.
- [1b] ———, *Lower semicontinuity of integral functionals with nonconvex integrands by relaxation-compactification*, this Journal, 19 (1981), pp. 533–542.
- [2a] L. D. BERKOVITZ, *Existence and lower closure theorems for abstract control problems*, this Journal, 12 (1974), pp. 27–42.
- [2b] ———, *A lower closure theorem for abstract control problems with L_p -bounded controls*, J. Optimization Theory Appl., 14 (1974), pp. 521–528.
- [2c] ———, *Lower semicontinuity of integral functionals*, Trans. Amer. Math. Soc., 192 (1974), pp. 51–57.
- [2d] ———, *Existence theory for optimal control problems*, in Optimal Control and Differential Equations, A. B. Schwarzkopf, W. G. Kelley and S. B. Eliason, eds., Academic Press, New York, 1978, pp. 107–130.
- [3] H. BERLIOCCI AND J.-M. LASRY, *Intégrales normales et mesures paramétrées en calcul des variations*, Bull. Soc. Math. France, 131 (1973), pp. 129–184.
- [4] L. D. BROWN AND R. PURVES, *Measurable selections of extrema*, Ann. Stat., 1 (1973), pp. 902–912.
- [5] C. CASTAING AND M. VALADIER, *Convex analysis and measurable multifunctions*, Lecture Notes in Mathematics, 580, Springer, Berlin, 1977.
- [6a] L. CESARI, *Existence theorems for weak and usual optimal solutions in Lagrange problems with unilateral constraints I*, Trans. Amer. Math. Soc., 124 (1966), pp. 369–412.
- [6b] ———, *Existence theorems for optimal controls of the Mayer type*, this Journal, 6 (1968), pp. 517–552.
- [6c] ———, *Seminormality and upper semicontinuity in optimal control*, J. Optimization Theory Appl., 6 (1970), pp. 114–137.
- [6d] ———, *Closure, lower closure and semicontinuity theorems in optimal control*, this Journal, 9 (1971), pp. 287–315.
- [6e] ———, *Closure theorems for orientor fields and weak convergence*, Arch. Rational Mech. Anal., 55 (1974), pp. 332–356.
- [6f] ———, *Lower semicontinuity and lower closure theorems without seminormality conditions*, Ann. Mat. Pura Appl., 98(1974), pp. 381–397.

- [7a] L. CESARI AND M. B. SURYANARAYANA, *Convexity and property (Q) in optimal control theory*, this Journal, 12 (1974), pp. 705–720.
- [7b] ———, *Closure theorems without seminormality conditions*, J. Optimization Theory Appl., 15 (1975), pp. 441–465.
- [7c] ———, *Nemitsky's operators and lower closure theorems*, J. Optimization Theory Appl., 19 (1976), pp. 165–183.
- [8] C. DELLACHERIE AND P. A. MEYER, *Probabilités et Potentiel*, Hermann, Paris, 1975.
- [9] I. EKELAND, *Sur le contrôle optimal de systèmes gouvernés par des équations elliptiques*, J. Functional Analysis, 9 (1972), pp. 1–62.
- [10] I. EKELAND AND R. TEMAM, *Convex Analysis and Variational Problems*, Dunod, Paris, 1972; English transl., North-Holland, Amsterdam, 1976.
- [11] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, Vestnik, Moscow, Univ. Ser. Math. Mech. Astr., 2 (1959), pp. 25–32, English transl., this Journal, 1 (1962), pp. 76–84.
- [12] C. J. HIMMELBERG, *Measurable relations*, Fund. Math., 87 (1975), pp. 53–72.
- [13] A. D. IOFFE, *On lower semicontinuity of integral functions I*, this Journal, 15 (1977), pp. 521–538.
- [14] A. D. IOFFE AND V. M. TICHOMIROV, *Theorie der Extremalaufgaben*, Nauka, Moscow, 1974; German transl., Deutscher Verlag der Wissenschaften, Berlin, 1979.
- [15] V. KLEE AND C. OLECH, *Characterizations of a class of convex sets*, Math. Scand., 20 (1967), pp. 290–296.
- [16] M. A. KRASNOSEL'SKII, P. R. ZABREIKO, E. I. PUSTYL'NIK AND P. W. SOBOLEVSKII, *Integral Operators in Spaces of Summable Functions*, Nauka, Moscow, 1966; English transl., Noordhoff, Leyden, 1976.
- [17] E. J. MCSHANE, *Relaxed controls and variational problems*, this Journal, 5 (1967), pp. 435–485.
- [18] C. OLECH, *Weak lower semicontinuity of integral functions*, J. Optimization Theory Appl., 19 (1976), pp. 3–16.
- [19] R. T. ROCKAFELLAR, *Existence theorems for general control problems of Bolza and Lagrange*, Advances in Math., 15 (1975), pp. 312–333.
- [20] E. H. ROTHE, *An existence theorem in the calculus of variations*, Arch. Rational Mech. Anal., 21 (1966), pp. 151–162.
- [21] M. B. SURYANARAYANA, *Remarks on lower semicontinuity and lower closure*, J. Optimization Theory Appl., 19 (1976), pp. 125–140.
- [22] D. H. WAGNER, *Survey of measurable selection theorems*, this Journal, 15 (1977), pp. 859–903.
- [23] L. C. YOUNG, *Lectures on the Calculus of Variations and Optimal Control Theory*, Saunders, Philadelphia, 1969.

ON LOCAL CONTROLLABILITY*

HENRY HERMES†

Abstract. Let X, Y be real analytic vector fields on a real analytic n -dimensional manifold M . Consider a control system on M with dynamics described by

$$(1) \quad dx/dt = X(x(t)) + u(t)Y(x(t)), \quad x(0) = p,$$

where an admissible control is Lebesgue measurable with values $|u(t)| \leq 1$. The relationship between the map

$$(2) \quad (s_1, \dots, s_n) \rightarrow (\exp s_1 Y) \circ \dots \circ (\exp s_n (\operatorname{ad}^{n-1} X, Y)) \circ (\exp tX)(p)$$

being one-one on a neighborhood of $0 \in \mathbb{R}^n$ for each $0 < t < \varepsilon$ and other known necessary conditions for local controllability are studied. In dimension $n = 2$, many necessary conditions are equivalent, and also sufficient. For $n \geq 3$, the map (2) being locally one-one implies many necessary conditions are satisfied, but these need not be sufficient. Examples which illustrate what occurs geometrically are given.

Introduction. Let M be a real analytic n -dimensional manifold and X, Y real analytic vector fields on M . We study the nonlinear control system $(\dot{x} = dx/dt)$,

$$(1) \quad \dot{x}(t) = X(x(t)) + u(t)Y(x(t)), \quad x(0) = p \in M,$$

where an admissible control u , is Lebesgue measurable with values $|u(t)| \leq 1$. The attainable set at time t , where $t \geq 0$, is denoted by $\mathcal{A}(t, p)$ and defined as the set of all points in M which can be attained at time t by solutions of (1) corresponding to all admissible controls. The solution of (1), at time t , corresponding to control $u \equiv 0$ is denoted $(\exp tX)(p)$. Our goal is to find necessary and sufficient conditions to insure that $(\exp tX)(p)$ belongs to the interior of $\mathcal{A}(t, p)$, denoted $\operatorname{int} \mathcal{A}(t, p)$, for all $t > 0$. When this holds, the system (1) is said to be locally controllable along the reference solution $(\exp tX)(p)$ at p .

Let $V(M)$ denote the vector space of all real analytic vector fields on M considered as a real Lie algebra with product the Lie product $[X, Y]$. For $\mathcal{C} \subset V(M)$ let $L(\mathcal{C})$ denote the Lie subalgebra generated by \mathcal{C} , TM_p the tangent space to M at p and $\mathcal{C}(p) = \{W(p) \in TM_p : W \in \mathcal{C}\}$. Let $(\operatorname{ad}^0 X, Y) = Y$, inductively $(\operatorname{ad}^{k+1} X, Y) = [X, (\operatorname{ad}^k X, Y)]$, and

$$\mathcal{S}^1 = \{(\operatorname{ad}^\nu X, Y) : \nu = 0, 1, \dots\}.$$

A necessary and sufficient condition [1], [2] that $\operatorname{int} \mathcal{A}(t, p) \neq \emptyset$ for all $t > 0$ is that $\dim L(\mathcal{S}^1)(p) = n$. A sufficient condition (the first order, or linear, test) that $(\exp tX)(p) \in \operatorname{int} \mathcal{A}(t, p)$ for all $t > 0$ is that $\dim \operatorname{span} \mathcal{S}^1(p) = n$. (See [3].) If $\dim \operatorname{span} \mathcal{S}^1(p) < n$, one may show ([2, Prop. 2.6]) that for small $t > 0$, $\dim \operatorname{span} \mathcal{S}^1((\exp tX) \cdot (p)) < n$ and $(\exp tX)(p)$ is called a singular solution of (1). In [2], [4], [5], higher order tests were given to determine when, in the singular case, one still has local controllability. These sufficient conditions essentially depend on a Taylor series expansion in which the coefficients involve elements of $L(\mathcal{S}^1)(p)$ other than those of $\mathcal{S}^1(p)$. The complexity of these coefficients make a general result from this approach seem unlikely. A different type of sufficient condition was given in [6]. This condition was derived by a method which depended on the dimension n being two and can loosely, geometrically, be stated as follows.

* Received by the editors November 25, 1980, and in final form April 21, 1981.

† Department of Mathematics, University of Colorado, Boulder, Colorado 80309. This work was supported in part by the National Science Foundation under grant MCS-79-26316 at the University of Colorado and grant MCS 77-18723 A03 at the Institute for Advanced Study, Princeton, New Jersey.

THEOREM 1. *Let $n = 2$ and $X(p), Y(p)$ be linearly independent. Then a necessary and sufficient condition that $(\exp tX)(p) \in \text{int } \mathcal{A}(t, p)$ for all $t > 0$ is that for some $\varepsilon > 0$ and all $\tau \in (0, \varepsilon)$ the integral curves of Y and $[X, Y]$ through $(\exp \tau X)(p)$ cross each other.*

Note that if $Y(p)$ and $[X, Y](p)$ are linearly independent, their integral curves will cross at p , but we will also have $\dim \text{span } \mathcal{S}^1(p) = 2$; hence the linear test applies. The interesting case is when $(\exp tX)(p)$ is a singular solution, so the integral curves of Y and $[X, Y]$ through $(\exp \tau X)(p)$ are tangent at $(\exp \tau X)(p)$, and the linear test is inclusive.

The geometric condition that the integral curves of Y and $[X, Y]$ cross can be replaced by requiring that the map

$$(2) \quad (s_1, s_2) \rightarrow (\exp s_1 Y) \circ (\exp s_2 [X, Y]) \circ (\exp tX)(p)$$

be one-one on some neighborhood of $0 \in \mathbb{R}^2$. (Then, by the invariance of domain theorem, this map will take every neighborhood of $0 \in \mathbb{R}^2$ onto a neighborhood of $(\exp tX)(p)$.) The natural extension of the two-dimensional condition to n dimensions would consider the map

$$(3) \quad (s_1, \dots, s_n) \rightarrow (\exp s_1 Y) \circ \dots \circ (\exp s_n (\text{ad}^{n-1} X, Y)) \circ (\exp tX)(p).$$

Example 2 will show that for $n = 3$ one may have the map (3) being one-one for $s = (s_1, \dots, s_n)$ in some neighborhood of $0 \in \mathbb{R}^n$ and each $t \in (0, \varepsilon)$, $\varepsilon > 0$, yet *not* have $(\exp tX)(p) \in \text{int } \mathcal{A}(t, p)$ for all $t > 0$. On the other hand, the map (3) being one-one will be shown to imply that many of the standard necessary conditions for higher order tests of local controllability are satisfied. The geometry involved is intriguing. Indeed, Example 2 and the geometric implications of the necessary conditions not being sufficient may be considered the main results in this paper.

1. The geometry of local controllability. We begin by showing the relationship among the various results known in dimension $n = 2$. The methods motivate those of the n -dimensional case.

The proof of Theorem 1, as given in [6], used the Green's theorem techniques; i.e., one examines the sign of a two-form, ω , in a neighborhood of the reference solution $(\exp tX)(p)$. In particular, if $(\exp tX)(p)$ is a singular solution on some interval $[0, t_1]$, then $\omega((\exp tX)(p)) = 0$ for $0 \leq t \leq t_1$. The stated necessary and sufficient condition that $(\exp tX)(p) \in \text{int } \mathcal{A}(t, p)$ for all $t > 0$ follows if for some $\varepsilon > 0$, ω has the same sign on "both sides" of $(\exp tX)(p)$, $0 < t < \varepsilon$.

Next, with the assumptions of Theorem 1, for each $t > 0$ the map

$$(4) \quad s \rightarrow q(s, t) = (\exp sY) \circ (\exp tX)(p)$$

takes a neighborhood of $0 \in \mathbb{R}^1$ to a one-dimensional submanifold (the orbit of Y through $(\exp tX)(p)$) which we denote by L_t . The "leaves" L_t , $0 \leq t \leq t_1$, form a foliation of codimension one for a neighborhood of the reference solution $(\exp tX)(p)$, $0 \leq t \leq t_1$. Thus, given x in this neighborhood, it belongs to a unique leaf L_t ; i.e., we have a real-valued function f defined on M by $f(x) = t$ if $x \in L_t$. This function has the following properties:

- (i) $f((\exp tX)(p)) = t$, so $(d/dt)f((\exp tX)(p)) \equiv (Xf)((\exp tX)(p)) = 1$.
- (ii) If ψ is any solution of (1) with $\psi(\tau) \in L_t$ (note $t \neq \tau$ in general), then $(d/dt)f(\psi(t))|_{t=\tau} = (Xf)(\psi(\tau))$.

This second property follows from the fact that the vector field Y is tangent to each leaf L_t ; i.e., f is constant on integral curves of Y . Geometrically, $(Xf)(q(s, t)(p))$ measures the "speed" with which a solution of (1) can cross L_t at $q(s, t)(p)$. Let $t_1, \delta > 0, p^1 = (\exp t_1 X)(p)$ and $\mathcal{N} = \{(s, t) : |s| < \delta, 0 \leq t \leq t_1\}$. Suppose $(Xf)(q(s, t)(p)) \geq 1$ (or ≤ 1) for $(s, t) \in \mathcal{N}$. Then any solution ψ of (1) which joins p to p^1 , say $\psi(t_2) = p^1$, and has orbit in $q(\mathcal{N})(p)$ satisfies

$$t_1 = f(\psi(t_2)) = \int_0^{t_2} (Xf)(\psi(t)) dt;$$

hence $t_2 \leq t_1$ (or $t_2 \geq t_1$) showing $(\exp tX)(p)$ extremizes time to reach p^1 ; i.e., $(\exp tX)(p)$ belongs to the boundary of $\mathcal{A}(t, p)$ for $0 \leq t \leq t_1$. This shows the following.

PROPOSITION 1. *A necessary condition that $(\exp tX)(p) \in \text{int } \mathcal{A}(t, p)$ for all $t > 0$ is that for some $\varepsilon > 0$, $((Xf)(q(s, t)(p)) - 1)$ changes sign as a function of s for each $t \in (0, \varepsilon)$.*

Finally, the third condition for local controllability in dimension $n = 2$ is that for some $\varepsilon > 0$ the integral curves of Y and $[X, Y]$ through $(\exp tX)(p)$ cross at $(\exp tX)(p)$ for $0 < t < \varepsilon$. Assume a Riemannian metric on M , and let $\langle V(p), W(p) \rangle$ denote the inner product relative to this metric. Then for any vector field W , $(Wf)(q(s, t)(p)) = \langle \text{grad } f(q(s, t)(p)), W(q(s, t)(p)) \rangle$ where $\text{grad } f$ is, by construction, orthogonal to Y . Thus the crossing of the integral curves of Y and $[X, Y]$ is equivalent to that

$$(5) \quad \langle \text{grad } f(q(s, t)(p)), [X, Y](q(s, t)(p)) \rangle$$

does not change sign as a function of s in some neighborhood of $0 \in \mathbb{R}^1$ for each $0 < t < \varepsilon$.

Our first goal is to relate the signs of the two-form $\omega, (Xf - 1), \langle \text{grad } f, [X, Y] \rangle$ and the map (2) being locally one-one.

PROPOSITION 2. *Let the dimension $n = 2, X(p), Y(p)$ be linearly independent, and for any $\varepsilon, \delta > 0, \mathcal{N} = \{|s| < \delta, 0 < t < \varepsilon\}$. The following are equivalent.*

(a) *There exists a neighborhood \mathcal{N} such that for each $t \in (0, \varepsilon)$ the map (2) is one-one for $|s| < \delta$.*

(b) *There exists a neighborhood \mathcal{N} in which $\langle \text{grad } f(q(s, t)(p)), [X, Y](q(s, t)(p)) \rangle$ does not change sign and is zero at most when $s = 0$.*

(c) *There exists a neighborhood in which $\omega(q(s, t)(p))$ does not change sign and is zero at most when $s = 0$.*

(d) *$((Xf)(q(s, t)(p)) - 1)$ has opposite signs for $s > 0$ and $s < 0$ in \mathcal{N} .*

Any one of the above is a necessary and sufficient condition that $(\exp tX)(p) \in \text{int } \mathcal{A}(t, p)$ for all $t > 0$.

Proof. It is easiest to show (a) \Leftrightarrow (b) \Leftrightarrow (c) and (d) \Leftrightarrow (b).

For (a) \Leftrightarrow (b) one notes that if $\langle \text{grad } f(q(s, t)(p)), [X, Y](q(s, t)(p)) \rangle$ changes sign, with s , in every neighborhood of zero, the integral curve $(\exp s_2[X, Y])$ through $(\exp tX)(p)$ is tangent to the integral curve of Y at $(\exp tX)(p)$ and remains, for small $|s_2|$, on one side of this latter curve. Thus the map (2) is not one-one. The converse (b) \Rightarrow (a) is straightforward.

For the relationship between (b) and (c), choose local coordinates such that $Y = (1, 0)$ and, say, $X(x) = (a_1(x), a_2(x))$. Then $X(p), Y(p)$ linearly independent implies $a_2(x) \neq 0$ near p . In these coordinates, $\text{grad } f(x) = (0, \alpha(x))$ with $\alpha(x) \neq 0, a_2^2(x)\omega(x) = -\partial a_2(x)/\partial x_1$, while $[X, Y](x) = (\partial a_1/\partial x_1, \partial a_2/\partial x_2)$. Thus $-\alpha(x)a_2^2(x)\omega(x) = \langle \text{grad } f(x), [X, Y](x) \rangle$, and the desired result follows.

The equivalence (d) \Leftrightarrow (b) is more subtle. Let $(\exp sY)_*$ denote the tangent space isomorphism induced by the map $p \rightarrow (\exp sY)(p)$. Since $f(q(s, t)(p)) \equiv t$,

$$(6) \quad \begin{aligned} 1 &= \frac{d}{dt} f(q(s, t)(p)) = (\exp sY)_* X((\exp -sY) \circ q(s, t)(p)) \\ &= \sum_{\nu=0}^{\infty} \frac{(-s)^\nu}{\nu!} (\text{ad}^\nu Y, X)(q(s, t)(p)). \end{aligned}$$

Thus (see [7, p. 108])

$$(7) \quad \begin{aligned} ((Xf)(q(s, t)(p)) - 1) &= s \left\langle \text{grad } f(q(s, t)(p)), \sum_{\nu=0}^{\infty} \frac{(-s)^\nu}{(\nu+1)!} (\text{ad}^\nu Y, [X, Y])(q(s, t)(p)) \right\rangle \\ &= s \left\langle \text{grad } f(q(s, t)(p)), \frac{d}{d\sigma} ((\exp(-sY)) \circ (\exp(sY - \sigma[X, Y]))(q(s, t)(p)))|_{\sigma=0} \right\rangle. \end{aligned}$$

A moment's reflection shows that the integral curves of $[X, Y]$ crossing those of Y for all s in a neighborhood of zero; i.e., $\langle \text{grad } f(q(s, t)(p)), [X, Y](q(s, t)(p)) \rangle$ not changing sign, is equivalent to

$$\left\langle \text{grad } f(q(s, t)(p)), \frac{d}{d\sigma} ((\exp(-sY)) \circ (\exp(sY - \sigma[X, Y]))(q(s, t)(p)))|_{\sigma=0} \right\rangle$$

not changing sign (i.e., $(\exp(-sY)) \circ (\exp(sY - \sigma[X, Y]))(q)$ will be on the "same side" of the orbit of Y through q for s in a neighborhood of zero and, say, $\sigma > 0$). The factor s , in the right side of (7), now gives the sign change of $(Xf - 1)$.

The final statement in Proposition 2 is now a consequence of Theorem 1 as stated in the introduction. \square

Remark 1. If one uses (6) directly in the expression $Xf - 1$, one obtains $((Xf)(q(s, t)(p)) - 1) = \langle \text{grad } f(q(s, t)(p)), \sum_{\nu=1}^{\infty} (-s)^\nu / \nu! (\text{ad}^\nu Y, X)(q(s, t)(p)) \rangle$. It is not at all clear from this that the term for $\nu = 1$ determines the sign of $Xf - 1$.

Remark 2. Let $\mathcal{A}^\alpha(t, p)$ denote the attainable set for (1) when the control values are bounded by $\alpha > 0$, i.e., $|u(t)| \leq \alpha$. It is possible to have a two-dimensional system for which given any $t_1 > 0$ there is an $\alpha > 0$ such that $(\exp t_1 X)(p) \in \text{int } \mathcal{A}^\alpha(t_1, p)$, yet the system is not locally controllable along $(\exp tX)(p)$ at p .

Remark 3. In Proposition 2, part (a), one cannot replace the map (2) with a map $(s_1, s_2) \rightarrow (\exp s_1 Y) \circ (\exp s_2 (\text{ad}^k X, Y)) \circ (\exp tX)(p)$ for integer $k \geq 2$. The need for $k = 1$ in the proof occurs at (6), (7). Specifically, in (6) one writes $(\text{ad}^\nu Y, X) = -(\text{ad}^{\nu-1} Y, [X, Y])$, and here it is essential that $[X, Y]$ be the element whose exponential is used in the map (2). This, together with Remark 2, is illustrated in the following.

Example 1. (All vectors will be written as row vectors for printing ease.) Let $M = \mathbb{R}^2$, $X(x) = (1 - x_2^2, x_2)$, $Y(x) = (x_2^2, 1)$ and p be the origin. Since $\dot{x}_1(t) = 1 - x_2^2(1 - u(t))$, for $|u(t)| \leq 1$ certainly $\dot{x}_1(t) \leq 1$, and $(\exp tX)(p) = (t, 0)$ belongs to the boundary of $\mathcal{A}(t, p)$ for all $t > 0$. Computing gives $(\text{ad } X, Y)(x) = (-2x_2^2 - 2x_2, 1)$, $(\text{ad}^2 X, Y)(x) = (4x_2^2, 1)$ while $\dim \text{span } \mathcal{G}^1(p) = 1$; so $(\exp tX)(p)$ is a singular solution. The map $(s_1, s_2) \rightarrow (\exp s_1 Y) \circ (\exp s_2 (\text{ad } X, Y)) \circ (\exp tX)(p) = (((s_1 + s_2)^3/2) - s_2^3 - s_2^2 + t, s_1 + s_2)$ is not one-one on any neighborhood of $0 \in \mathbb{R}^2$. The integral curves of Y and $(\text{ad } X, Y)$ through a point $(\exp tX)(p)$ do not cross (see Fig. 1). On the other hand, the map $(s_1, s_3) \rightarrow (\exp s_1 Y) \circ (\exp s_3 (\text{ad}^2 X, Y)) \circ (\exp tX)(p) = (t + s_3^3 + (s_1 + s_3)^3/3, s_1 + s_3)$ is locally one-one; the integral curves of Y and $(\text{ad}^2 X, Y)$

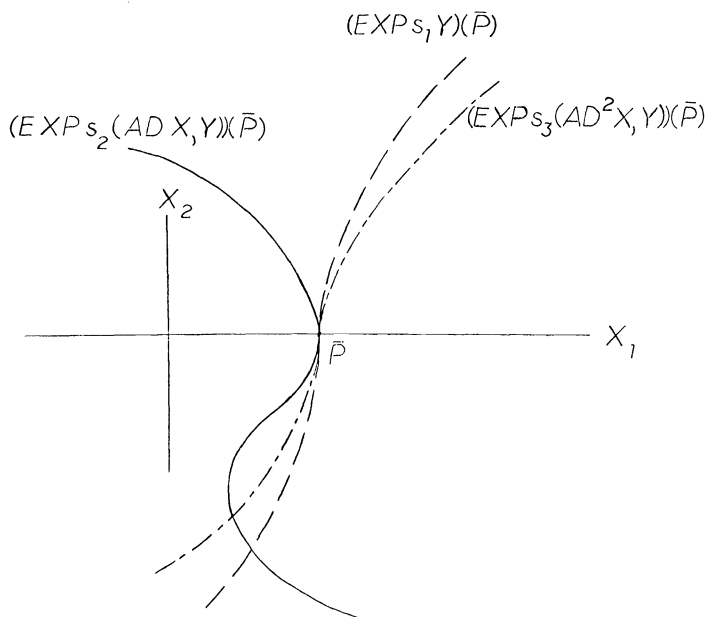


FIG. 1

through $(\exp tX)(p)$ do cross at $(\exp tX)(p)$, but as noted in Remark 3 this is not a sufficient condition for local controllability along $(\exp tX)(p)$ at p .

In this example, the two-form ω of the Green's theorem approach is $\omega(x) = -x_2(3x_2 + 2)/(x_2^3 + x_2^2 - 1)$ which, as expected, does change sign on opposite sides of the reference solution. However, it again changes sign at $x_2 = -\frac{2}{3}$, and one may easily show that given any $t_1 > 0$ there is an $\alpha > 0$ such that with $|u(t)| \leq \alpha$, $(\exp t_1 X)(p) \in \text{int } \mathcal{A}^\alpha(t_1, p)$.

We next proceed to n dimensions and compare, mainly, the condition that the map (3) be locally one-one with a necessary condition of the form " $Xf - 1$ must change sign." One could as well replace the map (3) with

$$(s_1, \dots, s_n) \rightarrow (\exp s_1 Y) \circ (\exp s_2(\text{ad}^{k_2} X, Y)) \circ \dots \circ (\exp s_n \text{ad}^{k_n} X, Y) \circ (\exp tX)(p)$$

where (k_2, \dots, k_n) is any permutation of the integers $(1, \dots, n-1)$. For the sake of exposition we will consider only the identity permutation, i.e., the map (3).

Let $\varepsilon > 0$ be given, and assume that for each $t \in (0, \varepsilon)$ the map (3) is one-one for s in some neighborhood of $0 \in \mathbb{R}^n$. As remarked previously, if $\dim \text{span } \mathcal{S}^1(p) = n$, the first order theory yields $(\exp tX)(p) \in \text{int } \mathcal{A}(t, p)$ for all $t > 0$. Thus we assume $\dim \text{span } \mathcal{S}^1(p) < n$ and that $X(p)$ is linearly independent of $\text{span } \mathcal{S}^1(p)$. (This is equivalent to $X(p), Y(p)$ linearly independent in dimension two.) For each integer $i = 2, \dots, n$ define the map $(s_1, \dots, s_{i-1}, s_{i+1}, \dots, s_n) \rightarrow q_i(s, t)(p)$ where

$$(8) \quad q_i(s, t)(p) = (\exp s_1 Y) \circ \dots \circ (\exp s_{i-1}(\text{ad}^{i-2} X, Y)) \\ \circ (\exp s_{i+1}(\text{ad}^i X, Y)) \circ \dots \circ (\exp s_n(\text{ad}^{n-1} X, Y)) \circ (\exp tX)(p).$$

The assumption that the map (3) is locally one-one implies that each of the maps $s \rightarrow q_i(s, t)(p)$ is locally one-one for s in some neighborhood of $0 \in \mathbb{R}^{n-1}$. Let the image of such a neighborhood be denoted L_i^t . The L_i^t are leaves of a codimension one foliation of a neighborhood, \mathcal{N} , of $\{(\exp tX)(p): 0 < t < \varepsilon\}$ since the reference solution is transverse to the L_i^t for small $t > 0$. For any $x \in \mathcal{N}$ and each $i = 2, \dots, n$, we may define a real

valued map f^i on \mathcal{N} by $f^i(x) = t$ if $x \in L_t^i$. As in the case of dimension two, we have

$$(i') \quad f^i((\exp tX)(p)) \equiv t \text{ or } (d/dt)f^i((\exp tX)(p)) = (Xf^i)((\exp tX)(p)) = 1.$$

$$(ii') \quad \text{If } \psi \text{ is any solution of (1) with } \psi(\tau) \in L_\tau^i, \text{ then } (d/dt)f^i(\psi(t))|_{t=\tau} = (Xf^i)(\psi(\tau)).$$

Again, (ii') follows since $(\exp s_1 Y)$ was the "leftmost" factor defining each of the maps q_i , thereby assuring f^i is constant on trajectories of Y . Analogous to Proposition 1, we have

PROPOSITION 3. *A necessary condition that $(\exp tX)(p) \in \text{int } \mathcal{A}(t, p)$ for all $t > 0$ is that for some $\varepsilon > 0$ and all $t \in (0, \varepsilon)$, $i = 2, \dots, n$, $((Xf^i)(q_i(s, t)(p)) - 1)$ change sign as a function of s in every neighborhood of $0 \in \mathbb{R}^{n-1}$.*

THEOREM 2. *Assume that for some $\varepsilon > 0$ and all $t \in (0, \varepsilon)$ the map (3) is locally one-one. Then $\text{int } \mathcal{A}(t, p) \neq \emptyset$ for all $t > 0$. If $(\exp tX)(p)$ is a singular solution and $X(p)$ is linearly independent of $\text{span } \mathcal{S}^1(p)$, the necessary condition of Proposition 3 is satisfied for each $i \in \{2, \dots, n\}$ for which $(\text{ad}^{i-1} X, Y)$ is not identically tangent to, or point to opposite sides of, L_t^i along the integral curve of $(\text{ad}^{i-2} X, Y)$ through $(\exp tX)(p)$.*

Proof. If the map (3) is locally one-one for any $t \in (0, \varepsilon)$, the vector fields $Y, \dots, (\text{ad}^{n-1} X, Y)$ cannot have an integral manifold of dimension less than n through any point $(\exp tX)(p)$, [8]. This implies $\dim L(\mathcal{S}^1)(\exp tX)(p) = n$ for every $t > 0$, which, by the Sussmann–Jurdjevic theorem [1, Th. 3.2], gives $\text{int } \mathcal{A}(t, p) \neq \emptyset$ for all $t > 0$.

Next, let $i \in \{2, \dots, n\}$, define

$$\begin{aligned} k^i(s)(\bar{p}) &= (\exp s_1 Y) \circ \dots \circ (\exp s_{i-1}(\text{ad}^{i-2} X, Y)) \circ (\exp s_{i+1}(\text{ad}^i X, Y)) \\ &\quad \circ \dots \circ (\exp s_n(\text{ad}^{n-1} X, Y))(\bar{p}), \end{aligned}$$

so $q_i(s, t)(p) = k^i(s) \circ (\exp tX)(p)$, and let $k_*^i(s)$ denote the tangent space isomorphism induced by the map $\bar{p} \rightarrow k^i(s)(\bar{p})$. Since $f^i(k^i(s) \circ (\exp tX)(p)) \equiv t$, we have the identity

$$(9) \quad \langle \text{grad } f^i(q_i(s, t)(p)), k_*^i(s)X((k^i(s))^{-1}(q_i(s, t)(p))) \rangle \equiv 1.$$

Thus

$$(10) \quad \begin{aligned} & (Xf^i)(q_i(s, t)(p)) - 1 \\ &= \langle \text{grad } f^i(q_i(s, t)(p)), X(q_i(s, t)(p)) - k_*^i(s)X((k^i(s))^{-1}(q_i(s, t)(p))) \rangle. \end{aligned}$$

Now $s = (s_1, \dots, s_{i-1}, s_{i+1}, \dots, s_n)$; set all $s_j = 0$ except s_{i-1} , and for ease of notation, let $q(s_{i-1}) = q_i(0, \dots, s_{i-1}, 0, \dots, 0, t)(p)$. Then, in (10),

$$\begin{aligned} & (Xf^i)(q(s_{i-1})) - 1 \\ &= \langle \text{grad } f^i(q(s_{i-1})), X(q(s_{i-1})) \\ &\quad - (\exp s_{i-1}(\text{ad}^{i-2} X, Y))_* X((\exp -s_{i-1}(\text{ad}^{i-2} X, Y))(q(s_{i-1}))) \rangle \\ &= \left\langle \text{grad } f^i(q(s_{i-1})), \sum_{\nu=1}^{\infty} \frac{(-s_{i-1})^\nu}{\nu!} (\text{ad}^\nu(\text{ad}^{i-2} X, Y), X)(q(s_{i-1})) \right\rangle \\ (11) \quad &= s_{i-1} \left\langle \text{grad } f^i(q(s_{i-1})), \sum_{\nu=0}^{\infty} \frac{(-s_{i-1})^\nu}{(\nu+1)!} (\text{ad}^\nu(\text{ad}^{i-2} X, Y), (\text{ad}^{i-1} X, Y))(q(s_{i-1})) \right\rangle \\ &= s_{i-1} \left\langle \text{grad } f^i(q(s_{i-1})), \frac{d}{d\sigma} ((\exp (-s_{i-1}(\text{ad}^{i-2} X, Y))) \right. \\ &\quad \left. \circ (\exp (s_{i-1}(\text{ad}^{i-2} X, Y) - \sigma(\text{ad}^{i-1} X, Y)))(q(s_{i-1}))) \right|_{\sigma=0} \rangle. \end{aligned}$$

From our hypothesis, $(\text{ad}^{i-1} X, Y)$ is not identically tangent to and does not point to opposite sides of the leaf L_i^i along the solution of $(\text{ad}^{i-2} X, Y)$ thru $(\exp tX)(p)$. This means $\langle \text{grad } f^i(q(s_{i-1})), (\text{ad}^{i-1} X, Y)(q(s_{i-1})) \rangle$ is not identically zero and does not change sign, as a function of s_{i-1} in some neighborhood of zero. Thus we may argue that $(Xf^i)(q(s_{i-1})) - 1$ changes sign with s_{i-1} in precisely the same manner as in the proof of Proposition 2. \square

Remark 4. The condition that $(\text{ad}^{i-1} X, Y)$ is not identically tangent to and does not point to opposite sides of the leaf L_i^i along the solution of $(\text{ad}^{i-2} X, Y)$ through $(\exp tX)(p)$ is merely convenient, but not necessary, to conclude a sign change of $(Xf^i - 1)$, as will be illustrated in Examples 2, 3. Indeed, if this condition holds for $i = 2$, $((Xf^2)(q(s_1)) - 1)$ changes sign with s_1 along the integral curve of Y through $(\exp tX) \cdot (p)$, and I would conjecture that this provides a sufficient condition that $(\exp tX)(p) \in \text{int } \mathcal{A}(t, p)$ for all $t > 0$.

Another, seemingly reasonable, conjecture is that the map (3) being locally one-one implies $(Xf^i(q_i(s, t)(p)) - 1)$ does change sign, as a function of s , in every neighborhood of $0 \in \mathbb{R}^{n-1}$ (but not necessarily does this change sign with s_{i-1}). I have been unable to show this.

We now turn to the heart of the problem in dimension $n \geq 3$; i.e., the necessary conditions of Proposition 3 are no longer sufficient. In Example 2, we show that for $n = 3$, the map (3) can be locally one-one, the necessary conditions $(Xf^i - 1)$ changing sign for $i = 2, 3$ are satisfied, but $(\exp tX)(p)$ is not in the interior of $\mathcal{A}(t, p)$ for small $t > 0$.

With reference to Remark 4, for $i = 3$ the hypothesis of Theorem 2 is satisfied and $((Xf^3)(q_3(s_2)) - 1)$ does change sign with s_2 ; however, for $i = 2$ the hypothesis of Theorem 2 is no longer satisfied and $(Xf^2)(q_2(s_1) - 1)$ does not change sign although $((Xf^2)(q_2(s, t)(p)) - 1)$ does change sign for $s = (s_1, s_3)$ in every neighborhood of $0 \in \mathbb{R}^2$.

Example 2. Let $M = \mathbb{R}^3$, $X(x) = (0, x_1 + x_2, 1 + x_2^3 + x_1^2)$, $Y = (1, 0, 0)$ and p be the origin. Then $(\exp tX)(p) = (0, 0, t)$ while $(\text{ad } X, Y)(x) = (0, 1, 2x_1)$, $(\text{ad}^2 X, Y)(x) = (0, 1, 3x_2^2)$ and $\dim \text{span } \mathcal{S}^1(p) = 2$, so $(\exp tX)(p)$ is a singular solution. We also have $X(p)$ linearly independent of $\text{span } \mathcal{S}^1(p)$. The map (3) is $(s_1, s_2, s_3) \rightarrow (\exp s_1 Y) \circ (\exp s_2 (\text{ad } X, Y)) \circ (\exp s_3 (\text{ad}^2 X, Y)) \circ (\exp tX)(p) = (s_1, s_2 + s_3, s_3^3 + t)$, which for any $t_1 > 0$ and $t \in (0, t_1)$ is locally one-one for s in a neighborhood of $0 \in \mathbb{R}^3$.

We first show, by direct calculation, that $(\exp tX)(p)$ is not in $\text{int } \mathcal{A}(t, p)$ for small $t > 0$. For any admissible control u let $U(t) = \int_0^t u(\sigma) d\sigma$. Then letting $x(t, u)$ denote the solution of our system at time t for control u , we have

$$\begin{aligned} x_1(t, u) &= U(t), & x_2(t, u) &= \int_0^t e^{t-\sigma} U(\sigma) d\sigma, \\ x_3(t, u) &= t + \int_0^t \left(\int_0^\tau e^{\tau-\sigma} U(\sigma) d\sigma \right)^3 d\tau + \int_0^t U^2(\sigma) d\sigma. \end{aligned}$$

We will show $x_3(t, u) \geq t$ for small $t > 0$. Using, in order, the Cauchy-Schwarz inequality, a change of order of integration and elementary estimates for small $t > 0$, yields

$$\begin{aligned} \left| \int_0^t \left(\int_0^\tau e^{\tau-\sigma} U(\sigma) d\sigma \right)^3 d\tau \right| &= \int_0^t e^{3\tau} \left| \int_0^\tau e^{-\sigma} U(\sigma) d\sigma \right|^3 d\tau \\ &\leq \int_0^t e^{3\tau} \left(\int_0^\tau e^{-2\sigma} d\sigma \right)^{3/2} \left(\int_0^\tau U^2(\sigma) d\sigma \right)^{3/2} d\tau \end{aligned}$$

$$\begin{aligned} &\leq \int_0^t \tau^{3/2} e^{3\tau} \left(\int_0^\tau U^2(\sigma) d\sigma \right)^{3/2} d\tau \\ &\leq \int_0^t U^2(\sigma) d\sigma \left(\int_0^t \tau^{3/2} e^{3\tau} d\tau \right) \leq \frac{2}{5} e^{3t} t^{5/2} \int_0^t U^2(\sigma) d\sigma. \end{aligned}$$

The latter expression is less than or equal to $\int_0^t U^2(\sigma) d\sigma$ if $t \geq 0$ is small. This shows $x_3(t) \geq t$; hence $(\exp tX)(p)$ is on the boundary of $\mathcal{A}(t, p)$ for small $t > 0$.

We next show, geometrically, why the necessary conditions $(Xf^i - 1)$ changing sign need no longer be sufficient.

First, $q_3(s, t)(p) = (\exp s_1 Y) \circ (\exp s_2 (\text{ad } X, Y)) \circ (\exp tX)(p) = (s_1, s_2, t)$; hence the leaves L_t^3 are planes parallel to the plane $x_3 = 0$. (See Fig. 2.) Here $f^3(x) = x_3$, so $(Xf^3)(q_3(s, t)(p)) - 1 = s_2^3 + s_1^2$ while $(\text{ad}^2 X, Y)(q_3(s, t)(p)) = (0, 1, 3s_2^2)$ which is tangent to a leaf L_t^3 along an integral curve of Y but not along an integral curve of $(\text{ad } X, Y)$. As Theorem 2 implies, and the direct computation above verifies, $(Xf^3)(q_3(0, s_2, t)(p)) - 1$ does change sign with s_2 at $s_2 = 0$.

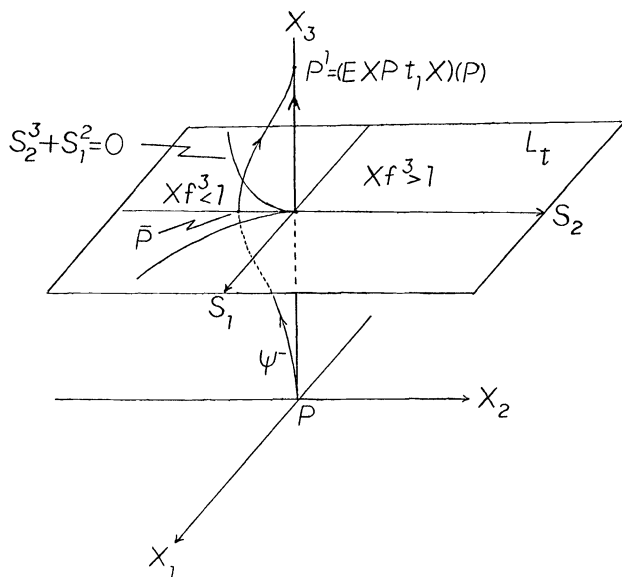


FIG. 2

If one could construct a solution ψ^- with orbit remaining in the region $Xf^3 < 1$ as pictured, one could prove $(\exp t_1 X)(p) \in \text{int } \mathcal{A}(t_1, p)$ for any $t_1 > 0$. Indeed, since $Y((\exp tX)(p))$, $[X, Y]((\exp tX)(p))$ are linearly independent, linear theory gives, for each t , the existence of a two manifold, \mathcal{M}_t^2 , contained in $\mathcal{A}(t, p)$, having $(\exp tX)(p)$ as a relative interior point and tangent space at $(\exp tX)(p)$ spanned by $Y((\exp tX)(p))$ and $[X, Y]((\exp tX)(p))$. Thus if one can show that for some $\delta > 0$ an interval of the form $\{(\exp tX)(p) : |t - t_1| < \delta\} \subset \mathcal{A}(t_1, p)$, it would follow that $(\exp t_1 X)(p) \in \text{int } \mathcal{A}(t_1, p)$. The existence of such an interval would be straightforward if one could construct solutions ψ^+ , ψ^- joining p to points of the reference solution other than p , with orbits, respectively, in regions where $Xf^3 > 1$, $Xf^3 < 1$. The point is that no such solution ψ^- exists! Indeed, the cusp of the curve $s_2^3 + s_1^2 = 0$ is sufficiently sharp to require the orbit of any solution ψ^- , such that $\psi^-(\tau)$ is a point of the form $\bar{p} = (\exp(-s_2(\text{ad } X, Y))) \circ (\exp tX)(p) \in L_t^3$ with $s_2 > 0$, be in the region $Xf^3 > 1$ enough so

that $\tau < t$. This must be considered somewhat unexpected since in [9, Lemma 3] or [4] it is shown that for any integer $m = 0, 1, \dots$ one can construct controls $\pm u$ for (1) having corresponding solutions $\psi(\cdot, \pm u)$ satisfying

$$(12)_0 \quad \psi(t, \pm u) = (\exp t(X + c_0 Y))(p), \quad c_0 = 1,$$

while for $m = 1, 2, \dots$,

$$(12)_m \quad \begin{aligned} \psi(t, \pm u) &= \exp(tX \pm c_m t^{2m}(\operatorname{ad}^m X, Y) + o(t^{2m}))(p) \\ &= \exp(\pm c_m t^{2m}(\operatorname{ad}^m X, Y) + (t^{2m})) \circ (\exp tX)(p), \end{aligned}$$

where $c_m = 2^{-m(1+3m)/2}$. Thus one can certainly reach a point such as $\bar{p} \in L_t$, but not necessarily in time t ; however, the orbit of a solution which does this cannot remain "inside" the cusp of $s_2^2 + s_1^2 = 0$.

Finally, we compute the second necessary condition, i.e., that $(Xf^2 - 1)$ must change sign, and show a similar cusp is involved. Here $q_2(s, t)(p) = (\exp s_1 Y) \circ (\exp s_3(\operatorname{ad}^2 X, Y)) \circ (\exp tX)(p) = (s_1, s_3, s_3^3 + t)$. If $x = (x_1, x_2, x_3) \in L_t^2$ one needs $x_3 - x_2^2 = t$ so $f^2(x) = x_3 - x_2^3$ and $(Xf^2)(q_2(s, t)(p)) - 1 = \langle \operatorname{grad} f^2(q_2(s, t)(p)), X(q_2(s, t)(p)) \rangle - 1 = -s_3^2(3s_1 + 2s_3) + s_1^2$. Thus $(Xf^2 - 1)$ does change sign in every neighborhood of zero, indeed $(Xf^2(q_2(0, s_3, t)(p)) - 1)$ changes sign with s_3 ; however, $(\operatorname{ad} X, Y)(q_2(s_1)) = (0, 1, 2s_1)$ which points to opposite sides of L_t^2 for s_1 positive and negative. Thus the conditions of Theorem 2 are not satisfied for $i = 2$ and indeed $(Xf^2(q_2(s_1)) - 1) = ((Xf^2)(q_2(s_1, 0, t)(p)) - 1) = s_1^2$ does not change sign with s_1 .

Again, in any leaf L_t^2 , the curve $s_3^2(3s_1 + 2s_3) = s_1^2$ which determines the sign change of $(Xf^2 - 1)$ has a cusp. See Fig. 3.

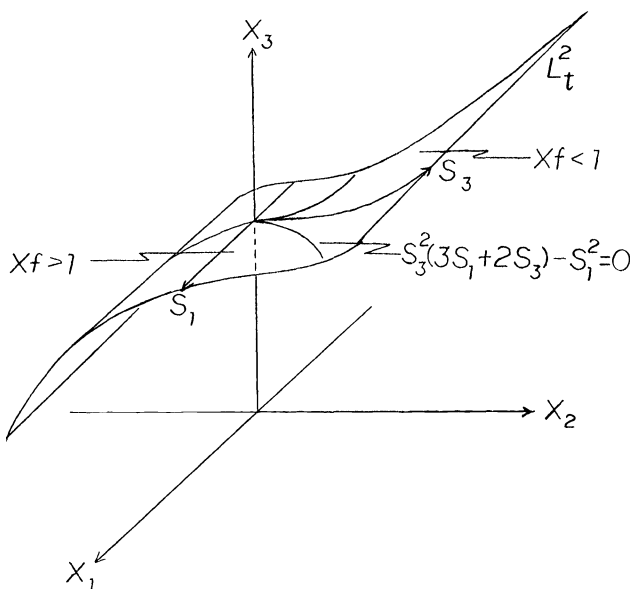


FIG. 3

Example 3. The purpose of this example is twofold. First, we again show the conditions of Theorem 2 are not necessary for a sign change in $(Xf^i - 1)$. Secondly, we conjecture that if the sign change of $(Xf^i - 1)$ on L_t^i occurs on a smooth curve through $(\exp tX)(p)$, then the necessary condition of Proposition 3 is sufficient.

Let $M = \mathbb{R}^3$, $X(x) = (x_2, x_1, 1 + x_2x_1^2 + 2x_2^3/3)$, $Y = (1, 0, 0)$ and p be the origin. Then $(\text{ad } X, Y)(x) = (0, 1, 2x_1x_2)$, $(\text{ad}^2 X, Y)(x) = (1, 0, -x_1^2)$, $\dim \text{span } \mathcal{S}^1(p) = 2$, so $(\exp tX)(p) = (0, 0, t)$ is a singular solution, and $X(p)$ is linearly independent of $\text{span } \mathcal{S}^1(p)$. Computing the map (3) gives $(\exp s_1 Y) \circ (\exp s_2 (\text{ad } X, Y)) \circ (\exp s_3 (\text{ad}^2 X, Y)) \circ (\exp tX)(p) = (s_1 + s_3, s_2, s_3s_2^2 - s_3^3/3 + t)$, which is locally one-one for each $t \geq 0$. $q_3(s, t)(p) = (\exp s_1 Y) \circ (\exp s_2 (\text{ad } X, Y)) \circ (\exp tX)(p) = (s_1, s_2, t)$, so the leaves L_i^3 are planes parallel to the coordinate plane $x_3 = 0$. Here $(\text{ad}^2 X, Y)(q_3(s, t) \cdot (p)) = (1, 0, -s_1^2)$ is tangent to a leaf L_i^3 along the integral curve $s_2 \rightarrow (\exp s_2 (\text{ad } X, Y)) \circ (\exp tX)(p) = (0, s_2, t)$; i.e., the method used in (7) to determine the sign of $(Xf^3 - 1)$ does not suffice. On the other hand, direct calculation shows $((Xf^3) \cdot (q_3(s, t)(p)) - 1) = s_2(s_1^2 + s_2^2/3)$. On each leaf, $((Xf^3) - 1)$ changes sign along the curve $s_2 = 0$, which is smooth (without cusps). It seems reasonable to conjecture that in this case solutions ψ^+ , ψ^- with orbits (respectively) remaining in the regions $Xf^3 > 1$ and $Xf^3 < 1$ can be constructed, and local controllability should follow by the method discussed in Example 2.

REFERENCES

- [1] H. J. SUSSMAN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.
- [2] H. HERMES, *Local controllability and sufficient conditions in singular problems*, J. Differential Equations, 20 (1976), pp. 213–232.
- [3] ———, *On local and global controllability*, this Journal 12 (1974), pp. 43–52.
- [4] A. J. KRENER, *The high order maximal principle and its application to singular extremals*, this Journal, 15 (1977), pp. 256–293.
- [5] H. HERMES, *Controlled stability*, Annali di Matematica Pura ed Appl., CXIV (1977), pp. 103–119.
- [6] ———, *On necessary and sufficient conditions for local controllability along a reference trajectory*, Geometric Methods in Systems Theory, D. Mayne and R. Brockett, eds., Reidel, Dordrecht, the Netherlands, 1974.
- [7] V. S. VARADARAJAN, *Lie Groups, Lie Algebras and Their Representations*, Prentice Hall, Englewood Cliffs, NJ, 1974.
- [8] T. NAGANO, *Linear differential systems with singularities and an application of transitive Lie algebras*, J. Math. Soc. Japan, 18 (1966), pp. 398–404.
- [9] H. HERMES, *On the synthesis of a stabilizing feedback control via Lie algebraic methods*, this Journal, 18 (1980), pp. 352–361.

PROJECTED NEWTON METHODS FOR OPTIMIZATION PROBLEMS WITH SIMPLE CONSTRAINTS*

DIMITRI P. BERTSEKAS†

Abstract. We consider the problem $\min \{f(x) | x \geq 0\}$, and propose algorithms of the form $x_{k+1} = [x_k - \alpha_k D_k \nabla f(x_k)]^+$, where $[\cdot]^+$ denotes projection on the positive orthant, α_k is a stepsize chosen by an Armijo-like rule and D_k is a positive definite symmetric matrix which is partly diagonal. We show that D_k can be calculated simply on the basis of second derivatives of f so that the resulting Newton-like algorithm has a typically superlinear rate of convergence. With other choices of D_k convergence at a typically linear rate is obtained. The algorithms are almost as simple as their unconstrained counterparts. They are well suited for problems of large dimension such as those arising in optimal control while being competitive with existing methods for low-dimensional problems. The effectiveness of the Newton-like algorithm is demonstrated via computational examples involving as many as 10,000 variables. Extensions to general linearly constrained problems are also provided. These extensions utilize a notion of an active generalized rectangle patterned after the notion of an active manifold used in manifold suboptimization methods. By contrast with these methods, many constraints can be added or subtracted from the binding set at each iteration without the need to solve a quadratic programming problem.

1. Introduction. We consider the problem

$$(1) \quad \begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } x \geq 0, \end{aligned}$$

where $f: R^n \rightarrow R$ is a continuously differentiable function, and the vector inequality $x \geq 0$ is meant to be componentwise (i.e., for $x = (x^1, x^2, \dots, x^n) \in R^n$, we write $x \geq 0$ if $x^i \geq 0$ for all $i = 1, \dots, n$). This type of problem arises very often in applications; for example, when f is a dual functional relative to an original inequality constrained primal problem and x represents a vector of nonnegative Lagrange multipliers corresponding to the inequality constraints, and when f represents an augmented Lagrangian or exact penalty function taking into account other possibly nonlinear equality and inequality constraints. The analysis and algorithms that follow apply also with minor modifications to problems with rectangle constraints such as

$$(2) \quad \begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } b_1 \leq x \leq b_2, \end{aligned}$$

where b_1 and b_2 are given vectors. Problems (1) and (2) are referred to as *simply constrained problems*, and their algorithmic solution is the primary subject of this paper.

In view of the simplicity of the constraints, one would expect that solution of problem (1) is almost as easy as unconstrained minimization of f . This expectation is partly justified in that the first order necessary condition for a vector $\bar{x} = (\bar{x}^1, \dots, \bar{x}^n)$ to be a local minimum of problem (1) takes the simple form

$$(3a) \quad \frac{\partial f(\bar{x})}{\partial x^i} \geq 0 \quad \forall i = 1, \dots, n,$$

* Received by the editors August 26, 1980, and in final revised form April 22, 1981.

† Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. This work was supported in part by the National Science Foundation under grant ENG-79-06332 and in part by Alphatech Inc., through Department of Energy contract DE-AC01-79ET29243.

$$(3b) \quad \frac{\partial f(\bar{x})}{\partial x^i} = 0 \quad \text{if } x^i > 0, \quad \forall i = 1, \dots, n.$$

Furthermore the direct analog of the method of steepest descent takes the simple form

$$(4) \quad x_{k+1} = [x_k - \alpha_k \nabla f(x_k)]^+, \quad k = 0, 1, \dots,$$

where α_k is a positive scalar stepsize and for any vector $z = (z^1, \dots, z^n) \in \mathbb{R}^n$ we denote

$$[z]^+ = \begin{bmatrix} \max\{0, z^1\} \\ \vdots \\ \max\{0, z^n\} \end{bmatrix}.$$

The stepwise α_k may be chosen in a number of ways. In the original proposal of Goldstein [1] and Levitin and Poljak [2], α_k is taken to be a constant $\bar{\alpha}$ (i.e., $\alpha_k \equiv \bar{\alpha}$, for all k), and a convergence result is shown under the assumption that $\bar{\alpha}$ is sufficiently small and ∇f is Lipschitz continuous. In general a proper value for $\bar{\alpha}$ can be found only through experimentation. An alternative suggested by McCormick [3] is to choose α_k by function minimization along the arc of points $x_k(\alpha)$, $\alpha \geq 0$, where

$$(5) \quad x_k(\alpha) = [x_k - \alpha \nabla f(x_k)]^+, \quad \alpha \geq 0.$$

Thus α_k is chosen so that

$$(6) \quad f[x_k(\alpha_k)] = \min_{\alpha \geq 0} f[x_k(\alpha)].$$

Unfortunately the minimization above is very difficult to carry out, particularly for problems of large dimension, since $f[x_k(\alpha)]$ need not be differentiable, convex, or unimodal as a function of α even if f is convex. For most problems we prefer the Armijo-like stepsize rule, first proposed in Bertsekas [4], whereby α_k is given by

$$(7a) \quad \alpha_k = \beta^{m_k} s,$$

where m_k is the first nonnegative integer m satisfying

$$(7b) \quad f(x_k) - f[x_k(\beta^m s)] \geq \sigma \nabla f(x_k)' [x_k - x_k(\beta^m s)].$$

Here the scalars s , β and σ are fixed and are chosen so that $s > 0$, $\beta \in (0, 1)$ and $\sigma \in (0, \frac{1}{2})$. In addition to being easily implementable and convergent, the algorithm (4), (7) has the advantage that when it converges to a local minimum x^* satisfying the standard second order sufficiency conditions for optimality (including strict complementarity) it identifies the binding constraints at x^* in a finite number of iterations in the sense that there exists \bar{k} such that

$$(8) \quad B(x^*) = B(x_k) \quad \forall k > \bar{k},$$

where, for every $x \in \mathbb{R}^n$, $B(x)$ denotes the set of indices of binding constraints at x ,

$$(9) \quad B(x) = \{i | x^i = 0, i = 1, \dots, n\}.$$

Minor modifications of the proofs given in [4] show that the results stated above hold also for the algorithm

$$(10) \quad x_{k+1} = [x_k - \alpha_k D_k \nabla f(x_k)]^+,$$

where D_k is a *diagonal* positive definite matrix, and α_k is chosen by (7) where now $x_k(\alpha)$ is given by

$$(11) \quad x_k(\alpha) = [x_k - \alpha D_k \nabla f(x_k)]^+.$$

For this it is necessary to assume that the diagonal elements d_k^i , $i = 1, \dots, n$ of the matrices D_k satisfy

$$\underline{d} \leq d_k^i \leq \bar{d} \quad \forall i = 1, \dots, n, \quad k = 0, 1, \dots,$$

where \underline{d} and \bar{d} are some positive scalars.

While it is often possible to achieve substantial computational savings by proper diagonal scaling of ∇f as in (10), the resulting algorithm is typically characterized by linear convergence rate [4], [22]. Any attempt to construct a superlinearly convergent algorithm must by necessity involve a nondiagonal scaling matrix D_k which is an adequate approximation of the inverse Hessian $\nabla^2 f(x_k)^{-1}$, at least along a suitable subspace. At this point we find that the algorithms available at present are far more complicated than their unconstrained counterparts, particularly when the problem has large dimension. Thus the most straightforward extension of Newton's method is given by

$$(12) \quad x_{k+1} = x_k + \alpha_k (\bar{x}_k - x_k),$$

where \bar{x}_k is a solution of the quadratic program

$$(13) \quad \begin{aligned} &\text{minimize } \nabla f(x_k)'(x - x_k) + \frac{1}{2}(x - x_k)'\nabla^2 f(x_k)(x - x_k) \\ &\text{subject to } x \geq 0, \end{aligned}$$

and α_k is a stepsize parameter. There are convergence and superlinear rate of convergence results in the literature regarding this type of method (Levitin and Poljak [2], Dunn [5]) and its quasi-Newton versions (Garcia-Palomares and Mangasarian [6]); however, its effectiveness is strongly dependent upon the computational requirements of solving the quadratic program (13). For problems of small dimension problem (13) can be solved rather quickly by standard pivoting or manifold suboptimization methods, but for large-dimensional problems the solution of the quadratic program (13) by standard methods can be very time consuming. Indeed there are large-scale quadratic programming problems arising in optimal control, the solution of which by pivoting methods is unthinkable. In any case the facility or lack thereof of solving the quadratic program (13) must be accounted for when comparing method (12) against other alternatives.

Another possible approach for constructing superlinearly convergent algorithms for solving problem (1) stems from the original gradient projection proposal of Rosen [7] and is based on manifold suboptimization and active set strategies as in Gill and Murray [8], Goldfarb [9], Luenberger [10] and other sources, (see Lenard [11] for up-to-date performance evaluation of various alternatives). Methods of this type are quite efficient for problems of relatively small dimension, but are typically unattractive for large-scale problems with a large number of constraints binding at a solution. The main reason is that typically at most one constraint can be added to the active set at each iteration, so if, for example, 1,000 constraints are binding at the point of convergence and an interior starting point is selected, then the method will require at least 1,000 iterations (and possibly many more) to converge. While several authors [8], [10] have alluded to the possibility of bending the direction of search along the constraint boundary, the only specific proposal known to the author that has been made in the context of the manifold suboptimization approach is the one of McCormick [12] and it does not seem particularly attractive for large-scale problems. (The quasi-Newton methods proposed by Brayton and Cullum [13] incorporate bending but simultaneously require the solution of quadratic programming subproblems.)

Manifold suboptimization methods require also additional computation overhead in deciding which constraint to drop from the currently active set. For the apparently most successful strategies (Lenard [11]) which attempt to drop as many constraints as possible this overhead can be significant and must be taken into account when comparing the manifold suboptimization approach with other alternatives.

The algorithms proposed in this paper attempt to combine the basic simplicity of the steepest descent iteration (4), (7) with the sophistication and fast convergence of the constrained Newton's method (12), (13). They do not involve solution of a quadratic program thereby avoiding the associated computational overhead, and there is no bound to the number of constraints that can be added to the currently active set thereby bypassing a serious inherent limitation of manifold suboptimization methods. The basic form of the method is

$$(14) \quad x_{k+1} = x_k(\alpha_k),$$

where

$$(15) \quad x_k(\alpha) = [x_k - \alpha D_k \nabla f(x_k)]^+ \quad \forall \alpha \geq 0.$$

D_k is a positive definite symmetric matrix which is partly diagonal, and α_k is a stepsize determined by an Armijo-like rule similar to (1) that will be described later. The convergence and rate of convergence properties of this method are discussed in § 2. A key property of the method is that under mild assumptions *it identifies the manifold of binding constraints at a solution in a finite number of iterations* in the sense of (8). This means that eventually the method is reduced to an unconstrained method on this manifold and brings to bear the extensive methodology and analysis relating to unconstrained minimization algorithms.

In § 3 we discuss how the method (14), (15) can form the basis for constructing algorithms for general linearly constrained problems of the form

$$(16) \quad \begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } b_1 \leq Ax \leq b_2. \end{aligned}$$

The main idea here is to view problem (16) *locally* as a simply constrained problem via a transformation of variables. For example, if the matrix A is square and invertible problem (16) is equivalent to the problem

$$\begin{aligned} &\text{minimize } h(y) \triangleq f(A^{-1}y) \\ &\text{subject to } b_1 \leq y \leq b_2, \end{aligned}$$

via the transformation

$$y = Ax.$$

A similar approach based on an active set strategy is employed when A is not square and invertible. The ideas are similar to those involved in manifold suboptimization methods where a linear manifold is selected as a "local universe" for the purposes of the current iteration. In our algorithms we take a suitably chosen rectangle (i.e., a set described by upper and lower bounds on the variables) as a "local universe" instead of a manifold.

Finally in § 4 we provide results of computational experiments with large scale optimal control problems some of which involve several thousand variables.

Throughout the paper we emphasize Newton-like methods as prototypes for broad classes of superlinearly converging algorithms that fit the framework of the

paper. We often make positive definiteness assumptions on the Hessian matrix of f in order to avoid getting bogged down in technical details relating to modifications of Newton's method such as those employed in unconstrained minimization [14]–[16] to account for the possibility that $\nabla^2 f$ is not positive definite. Quasi-Newton, approximate Newton and conjugate gradient versions of the Newton-like methods presented are possible but the discussion of specific implementations is beyond the scope of the paper. More generally it may be said that the nature of the algorithms proposed is such that almost every useful idea from unconstrained minimization can be fruitfully adapted within the constrained minimization framework considered here; however, the precise details of how this should be done may involve considerable further research and experimentation.

The notation employed throughout the paper is as follows. All vectors are considered to be column vectors. A prime denotes transposition. The standard norm in R^n is denoted by $|\cdot|$, i.e., for $x = (x^1, \dots, x^n)$ we write $|x| = [\sum_{i=1}^n (x^i)^2]^{1/2}$. The gradient and Hessian of a function $f: R^n \rightarrow R$ are denoted by ∇f and $\nabla^2 f$ respectively.

2. Algorithms for minimization subject to simple constraints. We consider first the problem $\min \{f(x) | x \geq 0\}$ of (1). Any vector $\bar{x} \geq 0$ satisfying the first order necessary condition (3) will be referred to as a *critical point with respect to problem* (1). We focus attention at iterations of the form

$$x_{k+1} = [x_k - \alpha_k D_k \nabla f(x_k)]^+,$$

where D_k is a positive definite symmetric matrix and α_k is chosen by search along the arc of points

$$x_k(\alpha) = [x_k - \alpha D_k \nabla f(x_k)]^+, \quad \alpha \geq 0.$$

It is easy to construct examples (see Fig. 1) where an arbitrary positive definite choice of the matrix D_k leads to situations where it is impossible to reduce the value of the objective by suitable choice of the stepsize α (i.e., $f[x_k(\alpha)] \geq f(x_k)$, $\forall \alpha \geq 0$). The

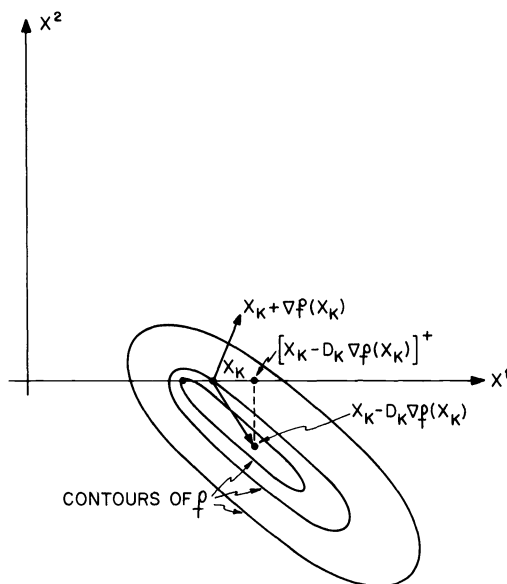


FIG. 1

following proposition identifies a class of matrices D_k for which an objective function reduction is possible. Define for all $x \geq 0$,

$$(17) \quad I^+(x) = \left\{ i \mid x^i = 0, \frac{\partial f(x)}{\partial x^i} > 0 \right\}.$$

We say that a symmetric $n \times n$ matrix D with elements d^{ij} is *diagonal with respect to a subset of indices* $I \subset \{1, 2, \dots, n\}$ if

$$(18) \quad d^{ij} = 0 \quad \forall i \in I, \quad j = 1, 2, \dots, n, \quad j \neq i.$$

PROPOSITION 1. *Let $x \geq 0$ and D be a positive definite symmetric matrix which is diagonal with respect to $I^+(x)$ and denote*

$$(19) \quad x(\alpha) = [x - \alpha D \nabla f(x)]^+ \quad \forall \alpha \geq 0.$$

(a) *The vector x is a critical point with respect to problem (1) if and only if*

$$x = x(\alpha) \quad \forall \alpha \geq 0.$$

(b) *If x is not a critical point with respect to problem (1) there exists a scalar $\bar{\alpha} > 0$ such that*

$$(20) \quad f[x(\alpha)] < f(x) \quad \forall \alpha \in (0, \bar{\alpha}).$$

Proof. Assume without loss of generality that for some integer r we have

$$I^+(x) = \{r+1, \dots, n\}.$$

Then D has the form

$$(21) \quad D = \begin{bmatrix} \bar{D} & & & 0 \\ - & d^{r+1} & & 0 \\ 0 & & \cdot & \\ & 0 & & d^n \end{bmatrix},$$

where \bar{D} is positive definite and $d^i > 0, i = r+1, \dots, n$. Denote

$$(22) \quad p = D \nabla f(x).$$

(a) Assume x is a critical point. Then using (3), (17)

$$\begin{aligned} \frac{\partial f(x)}{\partial x^i} &= 0 \quad \forall i = 1, \dots, r, \\ \frac{\partial f(x)}{\partial x^i} &> 0 \quad \text{if } x^i = 0, \quad \forall i = r+1, \dots, n. \end{aligned}$$

These relations and the positivity of $d^i, i = r+1, \dots, n$ imply that

$$\begin{aligned} p^i &= 0 \quad \forall i = 1, \dots, r, \\ p^i &> 0 \quad \forall i = r+1, \dots, n. \end{aligned}$$

Since $x^i(\alpha) = [x^i - \alpha p^i]^+$ and $x^i = 0$ for $i = r+1, \dots, n$ it follows that $x^i(\alpha) = x^i$ for all i , and $\alpha \geq 0$.

Conversely assume that $x = x(\alpha)$ for all $\alpha \geq 0$. Then we must have

$$\begin{aligned} p^i &= 0 \quad \forall i = 1, \dots, n \quad \text{with } x^i > 0, \\ p^i &\geq 0 \quad \forall i = 1, \dots, n \quad \text{with } x^i = 0. \end{aligned}$$

Now by definition of $I^+(x)$ we have that if $x^i = 0$ and $i \notin I^+(x)$ then $\partial f(x)/\partial x^i \leq 0$. This together with the relations above imply

$$\sum_{i=1}^r p^i \frac{\partial f(x)}{\partial x^i} \leq 0.$$

Since by (21), (22),

$$\begin{bmatrix} p_1 \\ \vdots \\ p_r \end{bmatrix} = \begin{bmatrix} \frac{\partial f(x)}{\partial x^1} \\ \vdots \\ \frac{\partial f(x)}{\partial x^r} \end{bmatrix}$$

and \bar{D} is positive definite we have $\sum_{i=1}^r p^i \partial f(x)/\partial x^i \geq 0$, and it follows that

$$p^i = \frac{\partial f(x)}{\partial x^i} = 0 \quad \forall i = 1, \dots, r.$$

Since for $i = r+1, \dots, n$, $\partial f(x)/\partial x^i > 0$ and $x^i = 0$, we obtain that x is a critical point.

(b) For $i = r+1, \dots, n$ we have $\partial f(x)/\partial x^i > 0$, $x^i = 0$, and, from (21), (22), $p^i > 0$. Since $x^i(\alpha) = [x^i - \alpha p^i]^+$ we obtain

$$(23) \quad x^i = x^i(\alpha) = 0 \quad \forall \alpha \geq 0, \quad i = r+1, \dots, n.$$

Consider the sets of indices

$$(24) \quad I_1 = \{i | x^i > 0 \text{ or } x^i = 0 \text{ and } p^i < 0, i = 1, \dots, r\},$$

$$(25) \quad I_2 = \{i | x^i = 0 \text{ and } p^i \geq 0, i = 1, \dots, r\}.$$

Let

$$(26) \quad \alpha_1 = \sup \{\alpha | x^i - \alpha p^i \geq 0, \forall i \in I_1\}.$$

Note that, in view of the definition of I_1 , α_1 is either positive or $+\infty$. Define the vector \bar{p} with coordinates

$$(27) \quad \bar{p}^i = \begin{cases} p^i & \text{if } i \in I_2, \\ 0 & \text{if } i \in I_2 \text{ or } i \in I^+(x) \end{cases}$$

In view of (23)–(27), we have

$$(28) \quad x(\alpha) = x - \alpha \bar{p} \quad \forall \alpha \in (0, \alpha_1).$$

In view of (25) and the definition of $I^+(x)$, we have

$$(29) \quad \frac{\partial f(x)}{\partial x^i} \leq 0 \quad \forall i \in I_2,$$

and hence

$$(30) \quad \sum_{i \in I_2} \frac{\partial f(x)}{\partial x^i} p^i \leq 0.$$

Now using (27) and (30), we have

$$(31) \quad \nabla f(x)' \bar{p} = \sum_{i \in I_1} \frac{\partial f(x)}{\partial x^i} p^i \geq \sum_{i=1}^r \frac{\partial f(x)}{\partial x^i} p^i.$$

Since x is not a critical point, by part (a) and (28), we must have $x \neq x(\alpha)$ for some $\text{HH}\alpha > 0$ and hence also, in view of (23), $p^i \neq 0$ for some $i \in \{1, \dots, r\}$. In view of the positive definiteness of \bar{D} and (21), (22) it follows that

$$\sum_{i=1}^r \frac{\partial f(x)}{\partial x^i} p^i > 0.$$

It follows from (31) that

$$\nabla f(x)' \bar{p} > 0.$$

Combining this relation with (28) and the fact $\alpha_1 > 0$ yields that \bar{p} is a feasible descent direction at x and there exists a scalar $\bar{\alpha} > 0$ for which the desired relation (20) is satisfied. Q.E.D.

Based on Proposition 1 we are led to the conclusion that the matrix D_k in the iteration

$$x_{k+1} = [x_k - \alpha_k D_k \nabla f(x_k)]^+$$

should be chosen diagonal with respect to a subset of indices that contains

$$I^+(x_k) = \left\{ i \mid x_k^i = 0, \frac{\partial f(x_k)}{\partial x^i} > 0 \right\}.$$

Unfortunately the set $I^+(x_k)$ exhibits an undesirable discontinuity at the boundary of the constraint set, whereby given a sequence $\{x_k\}$ of interior points that converges to a boundary point \bar{x} the set $I^+(x_k)$ may be strictly smaller than the set $I^+(\bar{x})$. This causes difficulties in proving convergence of the algorithm and may have an adverse effect on its rate of convergence. (This phenomenon is quite common in feasible direction algorithms and is referred to as zigzagging or jamming.) For this reason we will employ certain enlargements of the sets $I^+(x_k)$ with the aim of bypassing these difficulties.

The algorithm that we describe utilizes a scalar $\varepsilon > 0$ (typically small), a fixed¹ diagonal positive definite matrix M (for example the identity), and two parameters $\beta \in (0, 1)$ and $\sigma \in (0, \frac{1}{2})$ that will be used in connection with an Armijo-like stepsize rule. An initial vector $x_0 \geq 0$ is chosen and at the k th iteration of the algorithm we have a vector $x_k \geq 0$. Denote

$$w_k = |x_k - [x_k - M \nabla f(x_k)]^+|, \quad \varepsilon_k = \min \{\varepsilon, w_k\}.$$

(k + 1)st iteration of the Algorithm. We select a positive definite symmetric matrix D_k which is diagonal with respect to the set I_k^+ given by

$$(32) \quad I_k^+ = \left\{ i \mid 0 \leq x_k^i \leq \varepsilon_k, \frac{\partial f(x_k)}{\partial x^i} > 0 \right\}.$$

Denote

$$(33) \quad p_k = D_k \nabla f(x_k),$$

$$(34) \quad x_k(\alpha) = [x_k - \alpha p_k]^+ \quad \forall \alpha \geq 0.$$

Then x_{k+1} is given by

$$(35) \quad x_{k+1} = x_k(\alpha_k),$$

¹ Actually the results that follow can be shown also for the case where M is changed from one iteration to the next in a way that its diagonal elements are bounded above and away from zero.

where

$$(36) \quad \alpha_k = \beta^{m_k}$$

and m_k is the first nonnegative integer m such that²

$$(37) \quad f(x_k) - f[x_k(\beta^m)] \geq \sigma \left\{ \beta^m \sum_{i \in I_k^+} \frac{\partial f(x_k)}{\partial x^i} p_k^i + \sum_{i \in I_k^+} \frac{\partial f(x_k)}{\partial x^i} [x_k^i - x_k^i(\beta^m)] \right\}.$$

The stepsize rule (36), (37) (see Fig. 2) may be viewed as a combination of the Armijo-like rule (7) and the Armijo rule usually employed in unconstrained minimization (see, e.g., Polak [18]). When I_k^+ is empty the right-hand side of (37) becomes $\sigma \beta^m \nabla f(x_k) p_k$ and is identical to the corresponding expression of the Armijo

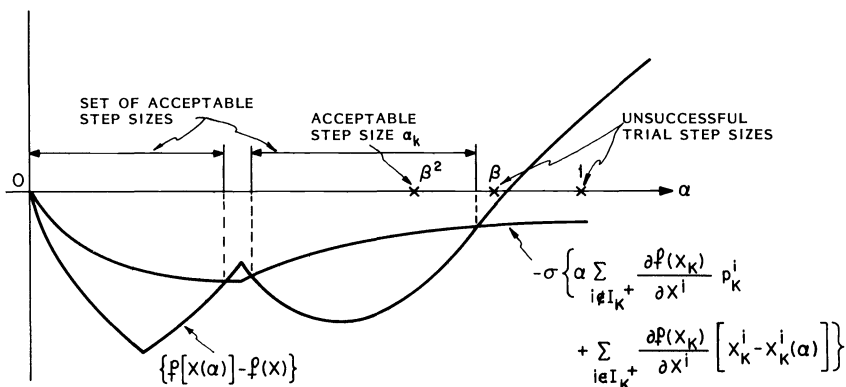


FIG. 2

rule in unconstrained optimization, while if $I_k^+ = \{1, 2, \dots, n\}$ then inequality (37) is identical with (7). Note that for all k we have

$$I_k^+ \supset I^+(x_k)$$

so the matrix D_k is diagonal with respect to $I^+(x_k)$. It is possible to show that for all $m \geq 0$ the right-hand side of (37) is nonnegative, and it is positive if and only if x_k is not a critical point. Indeed since D_k is positive definite and diagonal with respect to I_k^+ we have

$$(38) \quad \sum_{i \in I_k^+} \frac{\partial f(x_k)}{\partial x^i} p_k^i \geq 0 \quad \forall k = 0, 1, \dots,$$

while for all $i \in I_k^+$, in view of the fact $\partial f(x^k)/\partial x^i > 0$, we have $p_k^i > 0$ and hence

$$x_k^i - x_k^i(\alpha) \geq 0 \quad \forall \alpha \geq 0, \quad i \in I_k^+, \quad k = 0, 1, \dots,$$

$$(39) \quad \frac{\partial f(x_k)}{\partial x^i} [x_k^i - x_k^i(\alpha)] \geq 0 \quad \forall \alpha \geq 0, \quad i \in I_k^+, \quad k = 0, 1, \dots.$$

This shows that the right side of (37) is nonnegative. If x_k is not critical then it is easily seen (compare also with the proof of Proposition 1(b)) that one of the inequalities

² The results that follow can also be proved if $\sum_{i \in I_k^+} (\partial f(x_k)/\partial x^i) p_k^i$ is replaced in (37) by $\gamma_k \sum_{i \in I_k^+} (\partial f(x_k)/\partial x^i) p_k^i$, where $\gamma_k = \min \{1, \bar{\alpha}_k\}$ and $\bar{\alpha}_k = \sup \{\alpha | x_k^i - \alpha p_k^i \geq 0, \forall i \in I_k^+\}$. This modification makes (37) easier to satisfy.

(38) or (39) is strict for $\alpha > 0$ so the right side of (37) is positive for all $m \geq 0$. A slight modification of the proof of Proposition 1(b) also shows that if x_k is not a critical point then (37) will be satisfied for all m sufficiently large so the stepsize α_k is well defined and can be determined via a finite number of arithmetic operations. If x_k is a critical point then, by Proposition 1(a), we have $x_k = x_k(\alpha)$ for all $\alpha \geq 0$. Furthermore the argument given in the proof of Proposition 1(a) shows that

$$\sum_{i=1}^r \frac{\partial f(x_k)}{\partial x^i} p_k^i = 0$$

so both terms in the right side of (37) are zero. Since also $x_k = x_k(\alpha)$ for all $\alpha \geq 0$ it follows that (37) is satisfied for $m = 0$ thereby implying that

$$x_{k+1} = x_k(1) = x_k \quad \text{if } x_k \text{ is critical.}$$

In conclusion, the algorithm is well defined, decreases the value of the objective function at each iteration k for which x_k is not a critical point, and essentially terminates if x_k is critical. We proceed to analyze its convergence and rate of convergence properties. To this end we will make use of the following two assumptions:

(A) *The gradient ∇f is Lipschitz continuous on each bounded set of R^n ; i.e., given any bounded set $S \subset R^n$ there exists a scalar L (depending on S) such that*

$$(40) \quad |\nabla f(x) - \nabla f(y)| \leq L|x - y| \quad \forall x, y \in S.$$

(B) *There exist positive scalars λ_1, λ_2 and nonnegative integers q_1, q_2 such that*

$$(41) \quad \lambda_1 w_k^{q_1} |z|^2 \leq z' D_k z \leq \lambda_2 w_k^{q_2} |z|^2 \quad \forall z \in R^n, \quad k = 0, 1, \dots,$$

where

$$w_k = |x_k - [x_k - M \nabla f(x_k)]^+|.$$

Assumption (A) is not essential for the result of Proposition 2 that follows, but simplifies its proof. It is satisfied for just about every problem likely to appear in practice. For example it is satisfied when f is twice differentiable as well as when f is an augmented Lagrangian of the type used for inequality constrained problems involving twice differentiable functions. Assumption (B) is a condition of the type commonly utilized in connection with unconstrained minimization algorithms. When $q_1 = q_2 = 0$, relation (41) takes the form

$$(42) \quad \lambda_1 |z|^2 \leq z' D_k z \leq \lambda_2 |z|^2 \quad \forall z \in R^n, \quad k = 0, 1, \dots$$

and simply says that the eigenvalues of D_k are uniformly bounded above and away from zero.

PROPOSITION 2. *Under Assumptions (A) and (B) above, every limit point of a sequence $\{x_k\}$ generated by iteration (35) is a critical point with respect to problem (1).*

Proof. Assume the contrary, i.e., that there exists a subsequence $\{x_k\}_K$ converging to a vector \bar{x} which is not critical. Since $\{f(x_k)\}$ is decreasing and f is continuous it follows that $\{f(x_k)\}$ converges to $f(\bar{x})$ and therefore

$$[f(x_k) - f(x_{k+1})] \rightarrow 0.$$

Since each of the sums in the right-hand side of (37) is nonnegative (cf. (38), (39)), we must have

$$(43) \quad \alpha_k \sum_{i \in I_k^+} \frac{\partial f(x_k)}{\partial x^i} p_k^i \rightarrow 0,$$

$$(44) \quad \sum_{i \in I_k^+} \frac{\partial f(x_k)}{\partial x^i} [x_k^i - x_k^i(\alpha_k)] \rightarrow 0.$$

Also since \bar{x} is not critical and M is diagonal we have $|\bar{x} - [\bar{x} - M \nabla f(\bar{x})]| \neq 0$, so (41) implies that the eigenvalues of $\{D_k\}_K$ are uniformly bounded above and away from zero. In view of the fact that D_k is diagonal with respect to I_k^+ , it follows that there exist positive scalars $\bar{\lambda}_1, \bar{\lambda}_2$ such that for all $k \in K$ that are sufficiently large

$$(45) \quad 0 < \bar{\lambda}_1 \frac{\partial f(x_k)}{\partial x^i} \leq p_k^i \leq \bar{\lambda}_2 \frac{\partial f(x_k)}{\partial x^i} \quad \forall i \in I_k^+,$$

$$(46) \quad \bar{\lambda}_1 \sum_{i \in I_k^+} \left| \frac{\partial f(x_k)}{\partial x^i} \right|^2 \leq \sum_{i \in I_k^+} p_k^i \frac{\partial f(x_k)}{\partial x^i} \leq \bar{\lambda}_2 \sum_{i \in I_k^+} \left| \frac{\partial f(x_k)}{\partial x^i} \right|^2.$$

We will show that our hypotheses so far lead to the conclusion that

$$(47) \quad \lim_{\substack{k \rightarrow \infty \\ k \in K}} \inf \alpha_k = 0.$$

Indeed, since \bar{x} is not a critical point there must exist an index i such that either

$$(48) \quad \bar{x}^i > 0 \quad \text{and} \quad \frac{\partial f(\bar{x})}{\partial x^i} \neq 0$$

or

$$(49) \quad \bar{x}^i = 0 \quad \text{and} \quad \frac{\partial f(\bar{x})}{\partial x^i} < 0.$$

If $i \notin I_k^+$ for an infinite number of indices $k \in K$ then (47) follows from (43), (46), (48) and (49). If $i \in I_k^+$ for an infinite number of indices $k \in K$ then for all those indices we must have $\partial f(x_k)/\partial x^i > 0$ so (49) cannot hold. Therefore from (48)

$$(50) \quad \bar{x}^i > 0 \quad \text{and} \quad \frac{\partial f(\bar{x})}{\partial x^i} > 0.$$

Since we have [cf. (39)] for all $k \in K$ for which $i \in I_k^+$

$$\sum_{i \in I_k^+} \frac{\partial f(x_k)}{\partial x^i} [x_k^i - x_k^i(\alpha_k)] \geq \frac{\partial f(x_k)}{\partial x^i} [x_k^i - x_k^i(\alpha_k)] \geq 0,$$

it follows from (44) and (50) that

$$\lim_{\substack{k \rightarrow \infty \\ k \in K}} [x_k^i - x_k^i(\alpha_k)] = 0.$$

Using the above relation, (45) and (50), we obtain (47).

We will complete the proof by showing that $\{\alpha_k\}_K$ is bounded away from zero thereby contradicting (47). Indeed in view of (46) the subsequences $\{x_k\}_K$, $\{p_k\}_K$ and $\{x_k(\alpha)\}_K$, $\alpha \in [0, 1]$ are uniformly bounded so by Assumption (A) there exists a scalar $L > 0$ such that for all $t \in [0, 1]$, $\alpha \in [0, 1]$ and $k \in K$ we have

$$(51) \quad |\nabla f(x_k) - \nabla f[x_k - t(x_k - x_k(\alpha))]| \leq tL|x_k - x_k(\alpha)|.$$

We have for all $k \in K$ and $\alpha \in [0, 1]$

$$\begin{aligned} f[x_k(\alpha)] &= f(x_k) + \nabla f(x_k)'[x_k(\alpha) - x_k] \\ &\quad + \int_0^1 \{\nabla f(x_k) - \nabla f[x_k - t(x_k - x_k(\alpha))]\}' dt [x_k - x_k(\alpha)], \end{aligned}$$

so

$$\begin{aligned}
 & f(x_k) - f[x_k(\alpha)] \\
 &= \nabla f(x_k)'[x_k - x_k(\alpha)] + \int_0^1 \{\nabla f[x_k - t[x_k - x_k(\alpha)]] - \nabla f(x_k)\}' dt [x_k - x_k(\alpha)] \\
 &\geq \nabla f(x_k)'[x_k - x_k(\alpha)] - \int_0^1 |\nabla f[x_k - t[x_k - x_k(\alpha)]] - \nabla f(x_k)| dt |x_k - x_k(\alpha)|,
 \end{aligned}$$

and finally by using (51)

$$(52) \quad f(x_k) - f[x_k(\alpha)] \geq \nabla f(x_k)'[x_k - x_k(\alpha)] - \frac{L}{2} |x_k - x_k(\alpha)|^2.$$

For $i \in I_k^+$ we have $x_k^i(\alpha) = [x_k^i - \alpha p_k^i]^+ \geq x_k^i - \alpha p_k^i$ and $p_k^i > 0$, so $0 \leq x_k^i - x_k^i(\alpha) \leq \alpha p_k^i$. It follows using (45) that

$$(53) \quad \sum_{i \in I_k^+} |x_k^i - x_k^i(\alpha)|^2 \leq \alpha \sum_{i \in I_k^+} p_k^i [x_k^i - x_k^i(\alpha)] \leq \alpha \bar{\lambda}_2 \sum_{i \in I_k^+} \frac{\partial f(x_k)}{\partial x^i} [x_k^i - x_k^i(\alpha)].$$

Consider the sets

$$I_{1,k} = \left\{ i \mid \frac{\partial f(x_k)}{\partial x^i} > 0, i \notin I_k^+ \right\}, \quad I_{2,k} = \left\{ i \mid \frac{\partial f(x_k)}{\partial x^i} \leq 0, i \notin I_k^+ \right\}.$$

For all $i \in I_{1,k}$ we must have $x_k^i > \varepsilon_k$ for otherwise we would have $i \in I_k^+$. Since $|\bar{x} - [\bar{x} - M \nabla f(\bar{x})]^+| \neq 0$ we must have $\lim_{k \rightarrow 0, k \in K} \inf \varepsilon_k > 0$ and $\varepsilon_k > 0$ for all k . Let $\bar{\varepsilon} > 0$ be such that $\bar{\varepsilon} \leq \varepsilon_k$ for all $k \in K$, and let B be such that $|p_k^i| \leq B$ for all i and $k \in K$. Then for all $\alpha \in [0, \bar{\varepsilon}/B]$ we have $x_k^i(\alpha) = x_k^i - \alpha p_k^i$ for all $i \in I_{1,k}$ so it follows that

$$(54) \quad \sum_{i \in I_{1,k}} \frac{\partial f(x_k)}{\partial x^i} [x_k^i - x_k^i(\alpha)] = \alpha \sum_{i \in I_{1,k}} \frac{\partial f(x_k)}{\partial x^i} p_k^i \quad \forall \alpha \in \left[0, \frac{\bar{\varepsilon}}{B}\right].$$

Also for all $\alpha \geq 0$ we have $x_k^i - x_k^i(\alpha) \leq \alpha p_k^i$, and since $\partial f(x_k)/\partial x^i \leq 0$ for all $i \in I_{2,k}$, we obtain

$$(55) \quad \sum_{i \in I_{2,k}} \frac{\partial f(x_k)}{\partial x^i} [x_k^i - x_k^i(\alpha)] \geq \alpha \sum_{i \in I_{2,k}} \frac{\partial f(x_k)}{\partial x^i} p_k^i.$$

Combining (54) and (55), we obtain

$$(56) \quad \sum_{i \notin I_k^+} \frac{\partial f(x_k)}{\partial x^i} [x_k^i - x_k^i(\alpha)] \geq \alpha \sum_{i \notin I_k^+} \frac{\partial f(x_k)}{\partial x^i} p_k^i \quad \forall \alpha \in \left[0, \frac{\bar{\varepsilon}}{B}\right].$$

For all $\alpha \geq 0$ we also have

$$|x_k^i - x_k^i(\alpha)| \leq \alpha |p_k^i| \quad \forall i = 1, \dots, n.$$

Furthermore it is easily seen using Assumption (B) that there exists $\lambda > 0$ such that

$$\sum_{i \notin I_k^+} (p_k^i)^2 \leq \lambda \sum_{i \notin I_k^+} \frac{\partial f(x_k)}{\partial x^i} p_k^i \quad \forall k \in K.$$

Combining the last two relations we obtain for all $\alpha \geq 0$

$$(57) \quad \sum_{i \notin I_k^+} |x_k^i - x_k^i(\alpha)|^2 \leq \alpha^2 \lambda \sum_{i \notin I_k^+} \frac{\partial f(x_k)}{\partial x^i} p_k^i \quad \forall k \in K.$$

We now combine (52), (53), (56) and (57) to obtain for all $\alpha \in [0, (\bar{\varepsilon}/B)]$ and $k \in K$

$$(58) \quad f(x_k) - f[x_k(\alpha)] \geq \left(\alpha - \frac{\alpha^2 \lambda L}{2} \right) \sum_{i \in I_k^+} \frac{\partial f(x_k)}{\partial x^i} p_k^i + \left(1 - \frac{\alpha \bar{\lambda}_2 L}{2} \right) \sum_{i \in I_k^+} \frac{\partial f(x_k)}{\partial x^i} [x_k^i - x_k^i(\alpha)].$$

Suppose α is chosen so that

$$(59) \quad 0 \leq \alpha \leq \frac{\bar{\varepsilon}}{B}, \quad 1 - \frac{\alpha \lambda L}{2} \geq \sigma, \quad 1 - \frac{\alpha \bar{\lambda}_2 L}{2} \geq \sigma, \quad \alpha \leq 1$$

or equivalently

$$(60) \quad 0 \leq \alpha \leq \min \left\{ \frac{\bar{\varepsilon}}{B}, \frac{2(1-\sigma)}{\lambda L}, \frac{2(1-\sigma)}{\bar{\lambda}_2 L}, 1 \right\}.$$

Then we have from (58), (59) for all $k \in K$

$$f(x_k) - f[x_k(\alpha)] \geq \sigma \left\{ \alpha \sum_{i \in I_k^+} \frac{\partial f(x_k)}{\partial x^i} p_k^i + \sum_{i \in I_k^+} \frac{\partial f(x_k)}{\partial x^i} [x_k^i - x_k^i(\alpha)] \right\}.$$

This means that if (60) is satisfied with $\beta^m = \alpha$, then the inequality (37) of the Armijo-like rule will be satisfied. It follows from the way the stepsize is reduced that α_k satisfies

$$(61) \quad \alpha_k \geq \beta \min \left\{ \frac{\bar{\varepsilon}}{B}, \frac{2(1-\sigma)}{\lambda L}, \frac{2(1-\sigma)}{\bar{\lambda}_2 L}, 1 \right\} \quad \forall k \in K.$$

This contradicts (47) and proves the proposition. Q.E.D.

It is interesting to note that the argument of the last part of the proof above shows that if the level set $\{x | f(x) \leq f(x_0), x \geq 0\}$ is bounded, then there exists a scalar $\bar{\alpha} > 0$ such that, for every $\alpha \in (0, \bar{\alpha}]$, the *constant* stepsize algorithm $x_{k+1} = x_k(\alpha)$ generates sequences $\{x_k\}$ the limit points of which are critical points with respect to problem (1).

We now focus attention at a local minimum x^* satisfying the following second order sufficiency conditions. For all $x \geq 0$ we denote by $B(x)$ the set of indices of binding constraints at x , i.e.,

$$(62) \quad B(x) = \{i | x^i = 0\} \quad \forall x \geq 0.$$

(C) *The local minimum x^* of problem (1) is such that for some $\delta > 0$, f is twice continuously differentiable in the open sphere $\{x | \|x - x^*\| < \delta\}$, and there exist positive scalars m_1, m_2 such that*

$$(63) \quad m_1 |z|^2 \leq z' \nabla^2 f(x) z \leq m_2 |z|^2$$

$\forall x$ such that $\|x - x^*\| < \delta$ and $z \neq 0$ such that $z^i = 0, \forall i \in B(x^*)$.

Furthermore

$$(64) \quad \frac{\partial f(x^*)}{\partial x^i} > 0 \quad \forall i \in B(x^*).$$

The following proposition demonstrates an important property of the algorithm, namely that under mild conditions it is attracted by a local minimum x^* satisfying Assumption (C) and identifies the set of active constraints at x^* in a finite number of iterations. Thus if the algorithm converges to x^* then after a finite number of iterations it is equivalent to an unconstrained optimization method restricted on the subspace of

binding constraints at x^ .* This property is instrumental in proving superlinear convergence of the algorithm when the portion of D_k corresponding to the indices $i \notin I_k^+$ is chosen in a way that approximates the inverse of the portion of the Hessian of f corresponding to these same indices.

PROPOSITION 3. *Let x^* be a local minimum of problem (1) satisfying Assumption (C). Assume also that (B) holds in the stronger form whereby, in addition to (41), the diagonal elements d_{ii}^k of the matrices D_k satisfy for some scalar $\bar{\lambda}_1 > 0$*

$$(65) \quad \bar{\lambda}_1 \leq d_{ii}^k \quad \forall k = 0, 1, \dots, \quad i \in I_k^+.$$

Then there exists a scalar $\bar{\delta} > 0$ such that if $\{x_k\}$ is a sequence generated by iteration (35) and for some index \bar{k} we have

$$(66) \quad |x_{\bar{k}} - x^*| \leq \bar{\delta},$$

then $\{x_k\}$ converges to x^ and we have*

$$(67) \quad I_k^+ = B(x_k) = B(x^*) \quad \forall k \geq \bar{k} + 1.$$

Proof. Since f is twice differentiable on $\{x | |x - x^*| < \delta\}$, it follows that there exist scalars $L > 0$ and $\delta_1 \in (0, \delta]$ such that for all x, y with $|x - x^*| \leq \delta, |y - x^*| \leq \delta_1$ we have

$$|\nabla f(x) - \nabla f(y)| \leq L|x - y|.$$

Also for x_k sufficiently close to x^* the scalar

$$w_k = |x_k - [x_k - M\nabla f(x_k)]^+|$$

is arbitrarily close to zero while, in view of (64), we have

$$\left[x_k^i - \mu^i \frac{\partial f(x_k)}{\partial x^i} \right]^+ = 0 \quad \forall i \in B(x^*),$$

where μ^i is the i th diagonal element of M . It follows that for x_k sufficiently close to x^* we have

$$x_k^i \leq w_k = \varepsilon_k \quad \forall i \in B(x^*),$$

while

$$x_k^i > \varepsilon_k \quad \forall i \notin B(x^*).$$

This implies that there exists $\delta_2 \in (0, \delta_1]$ such that

$$(68) \quad B(x^*) = I_k^+ \quad \forall k \text{ such that } |x_k - x^*| \leq \delta_2.$$

Also there exist scalars $\bar{\varepsilon} > 0$ and $\delta_3 \in (0, \delta_2]$ such that

$$x_k^i > \bar{\varepsilon} \quad \forall i \notin B(x^*) \text{ and } k \text{ such that } |x_k - x^*| \leq \delta_3.$$

By essentially repeating the argument in the proof of Proposition 2 that led to (61), we find that there exists a scalar $\bar{\alpha} > 0$ such that

$$(69) \quad \alpha_k \geq \bar{\alpha} \quad \forall k \text{ such that } |x_k - x^*| \leq \delta_3.$$

By using (65) and (68) it follows that

$$(70) \quad 0 < \bar{\lambda}_1 \frac{\partial f(x_k)}{\partial x^i} \leq p_k^i \quad \forall i \in B(x^*) \text{ and } k \text{ such that } |x_k - x^*| \leq \delta_3,$$

while there exists a scalar $\lambda > 0$ such that

$$(71) \quad \sum_{i \notin B(x^*)} |p_k^i|^2 \leq \lambda \sum_{i \notin B(x^*)} \left| \frac{\partial f(x_k)}{\partial x^i} \right|^2 \quad \forall k \text{ such that } |x_k - x^*| \leq \delta_3.$$

Since $\partial f(x^*)/\partial x^i > 0$ for all $i \in B(x^*)$ and $\partial f(x^*)/\partial x^i = 0$ for all $i \notin B(x^*)$ it follows from (68)–(71) that there exists a scalar $\delta_4 \in (0, \delta_3]$ such that

$$(72) \quad B(x^*) = B(x_{k+1}) \quad \forall k \text{ such that } |x_k - x^*| \leq \delta_4$$

and

$$(73) \quad |x_{k+1} - x^*| \leq \delta_3 \quad \forall k \text{ such that } |x_k - x^*| \leq \delta_4.$$

In view of (68) we obtain from (72), (73)

$$(74) \quad B(x^*) = B(x_{k+1}) = I_{k+1}^+ \quad \forall k \text{ such that } |x_k - x^*| \leq \delta_4.$$

Thus when $|x_k - x^*| \leq \delta_4$ we have $|x_{k+1} - x^*| \leq \delta_3$, $B(x^*) = B(x_{k+1})$, and the $(k+1)$ st iteration of the algorithm reduces to an iteration of an unconstrained minimization algorithm on the subspace of binding constraints at x^* . From known results on unconstrained minimization algorithms (cf. [19, Proposition 1.12]) and Assumption (C) it follows that there exists an (open) neighborhood $N(x^*)$ of x^* such that $|x - x^*| < \delta_4$ for all $x \in N(x^*)$ and with the property that if $x_{k+1} \in N(x^*)$ and $B(x_{k+1}) = B(x^*)$, then $x_{k+2} \in N(x^*)$ and, by (74), $B(x_{k+2}) = B(x^*)$. This argument can be repeated and shows that if for some \bar{k} we have

$$x_{\bar{k}} \in N(x^*), \quad B(x_{\bar{k}}) = B(x^*),$$

then $\{x_k\} \rightarrow x^*$ and

$$x_k \in N(x^*), \quad B(x_k) = B(x^*) \quad \forall k \geq \bar{k}.$$

To complete the proof it is sufficient to show that there exists $\bar{\delta} > 0$ such that if $|x_k - x^*| \leq \bar{\delta}$ then $x_{k+1} \in N(x^*)$ and $B(x_{k+1}) = B(x^*)$. Indeed by repeating the argument that led to (73) and (74), we find that given any $\tilde{\delta} > 0$ there exists a $\bar{\delta} > 0$ such that if $|x_k - x^*| \leq \bar{\delta}$ then

$$|x_{k+1} - x^*| \leq \tilde{\delta}, \quad I_{k+1}^+ = B(x_{k+1}) = B(x^*).$$

By taking $\tilde{\delta}$ sufficiently small so that

$$\{x | |x - x^*| \leq \tilde{\delta}\} \subset N(x^*)$$

the proof is complete. Q.E.D

Under the assumptions of Proposition 3 we see that if the algorithm converges to a local minimum x^* satisfying Assumption (C) then it reduces eventually to an unconstrained minimization method restricted to the subspace

$$(75) \quad T = \{x | x^i = 0, \forall i \in B(x^*)\}$$

Furthermore, as shown in Proposition 3, for some index \bar{k} we will have

$$(76) \quad I_k^+ = B(x^*) \quad \forall k \geq \bar{k}.$$

This shows that if the portion of the matrix D_k corresponding to the indices $i \notin I_k^+$ is chosen to be the inverse of the Hessian of f with respect to these indices then the algorithm eventually reduces to Newton's method restricted on the subspace T .

More specifically, by rearranging indices if necessary, assume without loss of generality that

$$(77) \quad I_k^+ = \{r_k + 1, \dots, n\},$$

where r_k is some integer. Then D_k has the form

$$(78) \quad D_k = \begin{bmatrix} \bar{D}_k & \cdots & 0 \\ 0 & d_k^{r_k+1} & \cdots & 0 \\ & 0 & \ddots & \\ & & & d_k^n \end{bmatrix},$$

where $d_k^i > 0$, $i = r_k + 1, \dots, n$ and \bar{D}_k can be an arbitrary positive definite matrix. Suppose we choose \bar{D}_k to be the inverse of the Hessian of f with respect to the indices $i = 1, \dots, r_k$, i.e., the elements $[\bar{D}_k^{-1}]_{ij}$ are

$$(79) \quad [\bar{D}_k^{-1}]_{ij} = \frac{\partial^2 f(x_k)}{\partial x^i \partial x^j} \quad \forall i, j \notin I_k^+.$$

By Assumption (C), $\nabla^2 f(x^*)$ is positive definite on T so it follows from (76) that this choice is well defined and satisfies the assumption of Proposition 3 for k sufficiently large. Since the conclusion of this proposition asserts that the method eventually reduces to Newton's method restricted on the subspace T a superlinear convergence rate result follows. This type of argument can be used to construct a number of Newton-like and quasi-Newton methods and prove corresponding convergence and rate of convergence results. We state one of the simplest such results regarding a Newton-like algorithm which is well suited for problems where f is strictly convex and twice differentiable. Its proof follows simply from the preceding discussion and standard results on the unconstrained form of Newton's method so it is left to the reader.

PROPOSITION 4. *Let f be convex and twice continuously differentiable. Assume that problem (1) has a unique optimal solution x^* satisfying Assumption (C), and that there exist positive scalars m_1, m_2 such that*

$$m_1 |z|^2 \leq z' \nabla^2 f(x) z \leq m_2 |z|^2, \quad \forall z \in \{x | f(x) \leq f(x_0)\}.$$

Assume also that in the algorithm (32)–(37), the matrix D_k is given by

$$D_k = H_k^{-1},$$

where H_k is the matrix with elements H_k^{ij} given by

$$H_k^{ij} = \begin{cases} 0 & \text{if } i \neq j, \text{ and either } i \in I_k^+ \text{ or } j \in I_k^+, \\ \frac{\partial^2 f(x_k)}{\partial x^i \partial x^j} & \text{otherwise.} \end{cases}$$

Then the sequence $\{x_k\}$ generated by iteration (35) converges to x^ and the rate of convergence of $\{x_k - x^*\}$ is superlinear (at least quadratic if $\nabla^2 f$ is Lipschitz continuous in a neighborhood of x^*).*

Note that by making use of the result of Proposition 3 it follows that when f is a positive definite quadratic function, the algorithm of Proposition 4 solves problem (1) in a finite number of iterations provided the unique solution x^* satisfies Assumption (C).

The algorithm of Proposition 4 also has the property that, for all k sufficiently large, the initial unity stepsize will be accepted by the Armijo rule. Our computational experience suggests that the unity stepsize is also acceptable for the great majority

of iterations even before the binding constraints at the solution are identified. We did observe however some cases where it was necessary to reduce the initial unity stepsize several times before a sufficient reduction in the objective function value was effected. The most typical situation where such a phenomenon can occur is when the scalar $\hat{\gamma}_k$ defined by

$$\hat{\gamma}_k = \min \{1, \hat{\alpha}_k\}, \quad \hat{\alpha}_k = \sup \{\alpha | x_k^i - \alpha p_k^i \geq 0, x_k^i > 0, i \in I_k^+\}$$

is small relative to unity. Under these circumstances some nonbinding constraint, that was not taken into account when forming the index set I_k^+ , is encountered after a small movement along the arc $\{x_k(\alpha) | \alpha > 0\}$. As a result it may occur that the objective function value increases as α is increased from $\hat{\gamma}_k$. A reasonable heuristic device to avoid a large number of function evaluations in such cases is to modify the line search so that if at any iteration a fixed number r of trial stepsizes $1, \beta, \dots, \beta^{r-1}$ fail to pass the Armijo rule test then $\hat{\gamma}_k$ is computed and used as the next trial stepsize.

There is another (infrequent) situation where a unity initial stepsize may be inappropriate when far from convergence, and the Armijo rule may need a large number of stepsize reductions before determining an acceptable stepsize. This situation can arise when the sets of indices $\{i | x_k^i = 0, i \in I_k^+\}$ and $\{i | x_k^i = 0, p_k^i < 0, i \in I_k^+\}$ are not equal, and as a result the initial direction of motion along the arc $\{x_k(\alpha) | \alpha \geq 0\}$ is not a Newton direction along any subspace. A difficulty of this type can be easily detected and can be typically corrected by combining the Armijo rule with some form of a line minimization rule.

Extension to upper and lower bounds. The algorithm (32)–(37) described so far in this section can be easily extended to handle problems of the form

$$(80) \quad \begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } b_1 \leq x \leq b_2, \end{aligned}$$

where b_1 and b_2 are given vectors of lower and upper bounds with $b_1 \leq b_2$. The set I_k^+ is replaced by

$$(81) \quad I_k^\# = \left\{ i \mid b_1^i \leq x_k^i \leq b_1^i + \varepsilon_k \text{ and } \frac{\partial f(x_k)}{\partial x^i} > 0 \text{ or } b_2^i - \varepsilon_k \leq x_k^i \leq b_2^i \text{ and } \frac{\partial f(x_k)}{\partial x^i} < 0 \right\},$$

and the definition of $x_k(\alpha)$ is changed to

$$(82) \quad x_k(\alpha) = [x_k - \alpha D_k \nabla f(x_k)]^\#,$$

where for all $z \in R^n$ we denote by $[z]^\#$ the vector with coordinates

$$(83) \quad [z]^\# = \begin{cases} b_2^i & \text{if } b_2^i \leq z^i, \\ z^i & \text{if } b_1^i < z^i < b_2^i, \\ b_1^i & \text{if } z^i \leq b_1^i. \end{cases}$$

The scalar ε_k is given by $\varepsilon_k = \min \{\varepsilon, |x_k - [x_k - M \nabla f(x_k)]^\#|\}$. The matrix D_k is positive definite and diagonal with respect to $I_k^\#$, and M is a fixed diagonal positive definite matrix. The iteration is given by

$$(84) \quad x_{k+1} = x_k(\alpha_k),$$

where α_k is chosen by the Armijo rule (36), (37), with $[x_k^i - x_k^i(\beta^m)]^+$ replaced by $[x_k^i - x_k^i(\beta^m)]^\#$.

The preceding algorithm also makes sense if some of the upper bounds b_2^i equal $+\infty$ and some of the lower bounds b_1^i equal $-\infty$. This covers the case where only some of the variables x^i are simply constrained by upper and/or lower bounds.

3. Extensions to general linear constraints. In this section we discuss briefly how the algorithms of the previous section can form the basis for constructing methods for solving the problem

$$(85) \quad \begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } b_1 \leq Ax \leq b_2, \end{aligned}$$

where $f: R^n \rightarrow R$ is a continuously differentiable function, A is an $m \times n$ matrix and b_1, b_2 are given vectors. We denote by $a'_j, j = 1, \dots, m$ the rows of A and by $b_{1,j}, b_{2,j}, j = 1, \dots, m$ the coordinates of b_1 and b_2 , respectively, so the constraint set is represented by the m inequality constraints

$$(86) \quad b_{1,j} \leq a'_j x \leq b_{2,j} \quad j = 1, \dots, m.$$

By slight abuse of standard mathematical notation we allow the possibilities $b_{1,j} = -\infty$ and $b_{2,j} = +\infty$. In this way each of the inequalities (86) may represent a two-sided inequality constraint ($-\infty < b_{1,j} \leq b_{2,j} < +\infty$), a one-sided inequality constraint ($-\infty = b_{1,j} < b_{2,j} < +\infty$ or $-\infty < b_{1,j} < b_{2,j} = +\infty$) or no constraint at all ($b_{1,j} = -\infty, b_{2,j} = +\infty$). When $b_{1,j} = b_{2,j}$ then (86) represents an equality constraint. We assume that problem (85) has at least one feasible solution. We denote for every feasible x

$$(87) \quad B(x) = \{j \mid b_{1,j} = a'_j x \text{ or } a'_j x = b_{2,j}\}.$$

We assume that for every feasible x the set of vectors

$$(88) \quad \{a_j \mid j \in B(x)\}$$

is linearly independent. This is essentially a nondegeneracy assumption. It can be dispensed with at the expense of technical complications which are beyond the scope of the paper. In order to simplify the statement of the algorithm that follows we assume that the set of inequality constraints (86) includes the trivial inequalities

$$(89) \quad -\infty \leq x^i \leq +\infty, \quad i = 1, \dots, n,$$

for which a_j is a unit coordinate vector and $b_{1,j} = -\infty, b_{2,j} = +\infty$. The value of this somewhat unorthodox device will become apparent shortly.

In the algorithm to be described, given a feasible vector x_k obtained at iteration k , we select a subset $B_k \subset \{1, \dots, m\}$ containing exactly n indices and satisfying the following two conditions

(a) $B(x_k) \subset B_k$.

(b) The set of vectors $\{a_j \mid j \in B_k\}$ is linearly independent.

Such a choice is always possible since the set of vectors $\{a_j \mid j \in B(x_k)\}$ is linearly independent by earlier assumption, and it is always possible to supplement the set $B(x_k)$ with a suitable subset of indices corresponding to the trivial constraints (89) so as to form a set B_k satisfying (a) and (b) above. However this may not be the only possibility and the manner in which the set B_k is formed is left open at this point. The set

$$(90) \quad X_k = \{x \mid b_{1,j} \leq a'_j x \leq b_{2,j}, j \in B_k\}$$

is referred to as the *active generalized rectangle at iteration k* . It plays a role similar to the one of the manifold of active constraints in manifold suboptimization methods.

By rearranging indices if necessary, we assume without loss of generality that B_k consists of the first n indices, i.e. $B_k = \{1, 2, \dots, n\}$. Then A is written as

$$A = \begin{bmatrix} A_k^+ \\ A_k^- \end{bmatrix},$$

where A_k^+ is the $n \times n$ invertible matrix having $a'_j, j \in B_k$, as its rows. We partition similarly the vectors b_1, b_2 ;

$$b_1 = \begin{bmatrix} b_{1,k}^+ \\ b_{1,k}^- \end{bmatrix}, \quad b_2 = \begin{bmatrix} b_{2,k}^+ \\ b_{2,k}^- \end{bmatrix}.$$

The idea of the algorithm is to consider at the $(k+1)$ st iteration the transformation of variables

$$(91) \quad y = A_k^+ x,$$

by means of which the active generalized rectangle X_k of (90) is transformed into the (ordinary) rectangle

$$(92) \quad Y_k = \{y | b_{1,k}^+ \leq y \leq b_{2,k}^+\},$$

while problem (85) is transformed into the problem

$$(93) \quad \begin{aligned} &\text{minimize } h_k(y) \triangleq f[(A_k^+)^{-1}y] \\ &\text{subject to } y \in Y_k, b_{1,k}^- \leq A_k^-(A_k^+)^{-1}y \leq b_{2,k}^-. \end{aligned}$$

Let $y_k = A_k^+ x_k$. By construction we have that the constraints

$$(94) \quad b_{1,k}^- \leq A_k^-(A_k^+)^{-1}y \leq b_{2,k}^-$$

are not binding at y_k , so we temporarily ignore them and carry out an iteration of the method of the previous section in the space of variables y . It takes the form (cf. (81)–(84))

$$(95a) \quad y_{k+1} = y_k(\alpha_k),$$

where

$$(95b) \quad y_k(\alpha) = [y_k - \alpha D_k \nabla h_k(y_k)]^\# \quad \forall \alpha \geq 0;$$

D_k is a positive definite matrix which is diagonal with respect to the appropriate set of indices, and $[\cdot]^\#$ denotes projection on the rectangle Y_k of (92). The stepsize α_k is selected by means of the Armijo-like rule of the previous section subject, however, to the additional restriction that it belongs to the set of stepsizes

$$\{\alpha | b_{1,k}^- \leq A_k^-(A_k^+)^{-1}y_k(\alpha) \leq b_{2,k}^-\}$$

that do not lead to violation of the nonbinding constraints (94). Since this set contains an interval of the form $[0, \bar{\alpha}]$, where $\bar{\alpha} > 0$, it is clear that the Armijo-like rule will yield a stepsize after a finite number of arithmetic operations. Taking into account the fact that the gradient of the transformed objective function is

$$\nabla h_k(y_k) = [(A_k^+)^{-1}]^T \nabla f(x_k)$$

and making use of (91) we can finally write iteration (95) in terms of the original variables as

$$(96) \quad x_{k+1} = (A_k^+)^{-1} [A_k^+ x_k - \alpha_k D_k [(A_k^+)^{-1}]^T \nabla f(x_k)]^\#,$$

where $x_{k+1} = (A_k^+)^{-1} y_{k+1}$. The algorithmic process by means of which x_{k+1} is obtained is illustrated in Fig. 3.

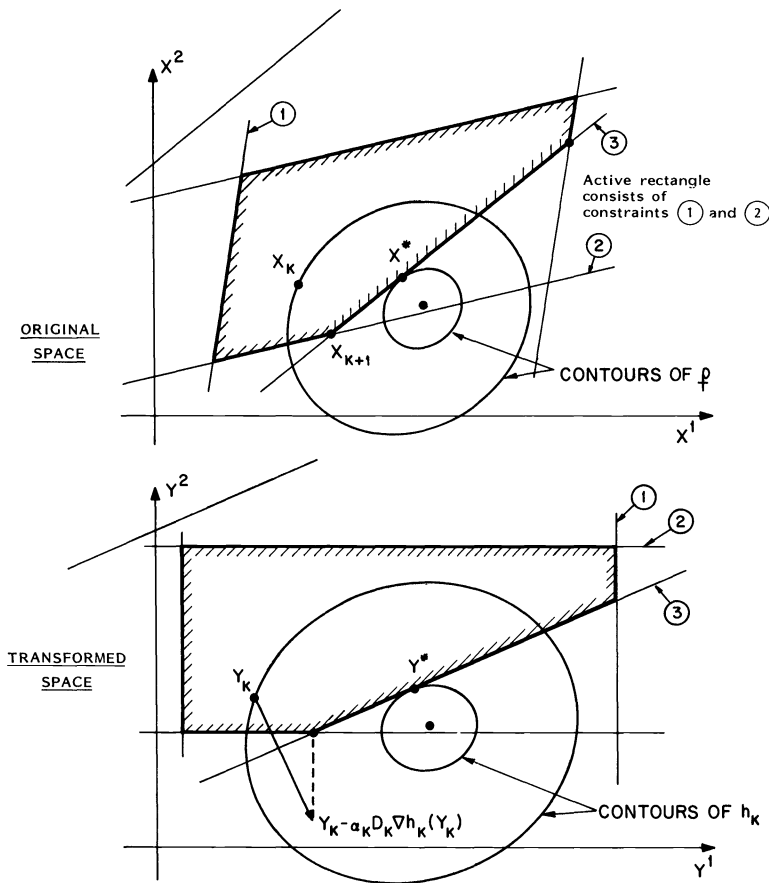


FIG. 3

It may appear that iteration (96) involves excessive computational overhead in view of the presence of the inverse $(A_k^+)^{-1}$. However in many problems with special structure it is possible to compute this inverse very efficiently. For other problems we note that it is possible to organize the algorithm so that most of the indices in the sets B_k and B_{k+1} are common. In fact at each iteration k typically at most one nonbinding inequality not belonging to the current active rectangle B_k will become binding at the next iteration, the exception being the unlikely situation where more than one of the constraints (94) will become simultaneously binding at y_{k+1} . In this case the matrices A_k^+ and A_{k+1}^+ need only differ by at most one row and as a result the inverse $(A_{k+1}^+)^{-1}$ can be obtained from $(A_k^+)^{-1}$ by the Householder modification rule involving only $O(n^2)$ arithmetic operations (see Gill and Murray [8, p. 59]). Note also that if a number (say n_k) of the trivial constraints (89) participate in the formation of the active rectangle (90) then the inverse $(A_k^+)^{-1}$ can be formed by matrix inversion of order $(n - n_k)$.

The reader who is familiar with manifold suboptimization methods, as described for example in Gill and Murray [8], will notice a strong similarity between the transformation process involved in these methods and the one employed above. The only essential difference is that in our method we use the active generalized rectangle X_k in place of the manifold of active constraints. The main advantage that algorithm (96) offers over manifold suboptimization alternatives is that as many as n new

constraints may become binding in a single iteration while considerable flexibility is afforded in changing the active set of constraints. By contrast, in manifold suboptimization methods, barring exceptional circumstances, at most *one* new constraint will become binding in any single iteration while dropping currently active constraints must be carefully controlled. Thus a *fundamental limitation of these methods is substantially overcome, the capability of attaining superlinear convergence is maintained and there is no need to solve a quadratic programming subproblem at each iteration.*

There are many issues relating to convergence, rate of convergence, active rectangle selection and implementation of the algorithm described in this section but their discussion properly belongs to a separate paper. We provide instead a specific superlinearly convergent Newton-like implementation of algorithm (96) for the case where the constraint set is a simplex.

Example (minimization on a simplex). Consider the problem

$$(97) \quad \begin{aligned} & \text{minimize } f(x) \\ & \text{subject to } x \geq 0, \sum_{i=1}^n x^i = 1, \end{aligned}$$

where we assume that the function $f: R^n \rightarrow R$ is convex, twice continuously differentiable with everywhere positive definite Hessian matrix. Given a feasible vector x_k let

$$\bar{i} = \arg \max \{x_k^i | i = 1, \dots, n\}.$$

We consider the transformation of variables defined by

$$y^i = x^i \quad \forall i \neq \bar{i}, \quad y^{\bar{i}} = \sum_{i=1}^n x^i,$$

thus implicitly forming an active rectangle consisting of the equation $\sum_{i=1}^n x_i = 1$ and the inequalities $x^i \geq 0, i \neq \bar{i}$. The inverse transformation is

$$x^i = y^i \quad \forall i \neq \bar{i}, \quad x^{\bar{i}} = y^{\bar{i}} - \sum_{i \neq \bar{i}} y^i.$$

If we write this transformation as $x = T_k y$, where T_k is the appropriate matrix, the problem is transformed into

$$\begin{aligned} & \text{minimize } h_k(y) \triangleq f(T_k y) \\ & \text{subject to } y^i \geq 0 \quad \forall i \neq \bar{i}, \\ & \quad y^{\bar{i}} = 1, \\ & \quad y^{\bar{i}} - \sum_{i \neq \bar{i}} y^i \geq 0. \end{aligned}$$

The last constraint is (by construction) inactive at the point $y_k = T_k x_k$, so it will be ignored in the iteration of the Newton-like method of § 2.

The first and second derivatives of the transformed objective function h_k with respect to the variables y^i at y_k are given by

$$\begin{aligned} \frac{\partial h_k(y_k)}{\partial y^i} &= \frac{\partial f(x_k)}{\partial x^i} - \frac{\partial f(x_k)}{\partial x^{\bar{i}}} \quad \forall i \neq \bar{i}, \\ \frac{\partial h_k(y_k)}{\partial y^{\bar{i}}} &= \frac{\partial f(x_k)}{\partial x^{\bar{i}}}, \end{aligned}$$

$$\begin{aligned}\frac{\partial^2 h_k(y_k)}{\partial y^i \partial y^{\bar{i}}} &= \frac{\partial^2 f(x_k)}{\partial x^i \partial x^{\bar{i}}} - \frac{\partial^2 f(x_k)}{\partial x^i \partial x^{\bar{i}}} - \frac{\partial^2 f(x_k)}{\partial x^j \partial x^{\bar{i}}} + \frac{\partial^2 f(x_k)}{(\partial x^{\bar{i}})^2} \quad \forall i \neq \bar{i}, \quad j \neq \bar{i}, \\ \frac{\partial^2 h_k(y_k)}{\partial y^i \partial y^{\bar{i}}} &= \frac{\partial^2 f(x_k)}{\partial x^i \partial x^{\bar{i}}} - \frac{\partial^2 f(x_k)}{(\partial x^{\bar{i}})^2} \quad \forall i \neq \bar{i}, \\ \frac{\partial^2 h_k(y_k)}{(\partial y^{\bar{i}})^2} &= \frac{\partial^2 f(x_k)}{(\partial x^{\bar{i}})^2}.\end{aligned}$$

The Newton-like iteration to be performed in the space of variables y is a slight variation of the one of § 2 (cf. Proposition 3) to account for the presence of the constraint $y^{\bar{i}} = 1$. It takes the form described below. Let

$$w_k = \left\{ \sum_{i \neq \bar{i}} \left(y_k^i - \left[y_k^i - \mu_k^i \frac{\partial h_k(y_k)}{\partial y^i} \right]^+ \right)^2 \right\}^{1/2},$$

where $\mu_k^i = [\partial^2 h_k(y_k) / (\partial y^i)^2]^{-1}$. Let also

$$\varepsilon_k = \min \{ \varepsilon, w_k \}, \quad I_k^+ = \left\{ i \mid 0 \leq y_k^i \leq \varepsilon_k, \frac{\partial h_k(y_k)}{\partial y^i} > 0 \right\},$$

and form the matrix H_k with elements H_k^{ij} given by

$$H_k^{ij} = \begin{cases} 0 & \text{if } i \neq j \text{ and either } i \in I_k^+ \text{ or } j \in I_k^+, \\ \frac{\partial^2 h_k(y_k)}{\partial y^i \partial y^j} & \text{otherwise.} \end{cases}$$

Let

$$(98) \quad p_k = H_k^{-1} \nabla h_k(y_k).$$

Then $y_{k+1} = y_k(\alpha_k)$, where for all $\alpha \geq 0$

$$y_k^i(\alpha) = [y_k^i - \alpha p_k^i]^+ \quad \forall i \neq \bar{i}, \quad y_k^{\bar{i}}(\alpha) = 1.$$

The stepsize α_k is given by $\alpha_k = \beta^{m_k}$, where m_k is the first nonnegative integer m such that

$$h_k(y_k) - h_k[y_k(\beta^m)] \geq \sigma \left\{ \beta^m \sum_{i \notin I_k^+} \frac{\partial h_k(y_k)}{\partial y^i} p_k^i + \sum_{i \in I_k^+} \frac{\partial h_k(y_k)}{\partial y^i} [y_k^i - y_k^i(\beta^m)] \right\}$$

and

$$1 - \sum_{i \neq \bar{i}} y_k^i(\beta^m) \geq 0.$$

The vector x_{k+1} is then given by

$$x_{k+1}^i = y_{k+1}^i \quad \forall i \neq \bar{i}, \quad x_{k+1}^{\bar{i}} = 1 - \sum_{i \neq \bar{i}} y_{k+1}^i.$$

Similarly as for Proposition 3, it is easily shown that this algorithm converges superlinearly to the unique (global) minimum of problem (97). The algorithm can be extended trivially to the case where, in addition to the nonnegativity constraints, there is a single equality or inequality constraint, as well as to the case where the constraint set consists of a Cartesian product of simplices. Similar algorithms can be written in

explicit form for problems with a large number of nonnegativity (or upper and lower bound constraints) and a small number of additional equality or inequality constraints. Newton-like algorithms of this type are particularly effective when the problem has special structure that facilitates the solution of the linear system of equations involved in implementing the basic iteration (cf. (98)).

4. Application in discrete-time optimal control—computational results. The algorithms of the paper are particularly well suited for discrete-time optimal control problems involving a discrete time system of the form

$$(99) \quad x_{i+1} = f_i(x_i, u_i), \quad i = 0, \dots, N-1,$$

a cost functional of the form

$$(100) \quad G(x_N) + \sum_{i=0}^{N-1} g_i(x_i, u_i)$$

and simple constraints on the control vectors of the form

$$\underline{b}_i \leq u_i \leq \bar{b}_i, \quad i = 0, \dots, N-1.$$

We assume that the functions $f_i: R^{n+m} \rightarrow R^n$, $g_i: R^{n+m} \rightarrow R$ and $G: R^n \rightarrow R$ are twice continuously differentiable and N is a positive integer. Problems of this type are discussed for example in Varaiya [17], Polak [18], and Cannon, Cullum and Polak [20]. They are often characterized by large dimension, particularly when they arise from discretization of continuous-time optimal control and calculus of variations problems.

Each state vector x_i can be uniquely represented in terms of the control sequence $u = \{u_0, \dots, u_{N-1}\}$ via the system equation (99) in the form

$$x_i = \phi_i(u), \quad i = 1, \dots, N,$$

where ϕ_i are the appropriate functions. The problem is then equivalent to

$$(101) \quad \begin{aligned} &\text{minimize } J(u) = G[\phi_N(u)] + \sum_{i=0}^{N-1} g_i[\phi_i(u), u_i] \\ &\text{subject to } \underline{b}_i \leq u_i \leq \bar{b}_i, \quad i = 0, \dots, N-1. \end{aligned}$$

It is well known (see Mitter [21], Polak [18]) that the unconstrained Newton direction $-\nabla^2 J(u)^{-1} \nabla J(u)$ for this problem can be efficiently computed by means of the Riccati equation. An algorithm such as the one of Proposition 4 can also be similarly implemented via the Riccati equation. At each iteration k we first determine the set of indices $I_k^\#$ [cf. (81)]. We then compute the Newton direction with respect to the control vector coordinates corresponding to indices $i \notin I_k^\#$ via the Riccati equation, while we compute the (diagonally) scaled steepest descent direction for the remaining coordinates corresponding to indices $i \in I_k^\#$. The overall algorithm is thus very similar to the one used for the corresponding unconstrained problem. It is well suited for large scale linear-quadratic problems with simple control constraints for which pivoting methods are apparently very cumbersome and inefficient. Our computational example is of this type.

Consider the two-dimensional linear system

$$(102) \quad \begin{bmatrix} x_{i+1,1} \\ x_{i+1,2} \end{bmatrix} = \begin{bmatrix} 1 & s \\ -s & 1 \end{bmatrix} \begin{bmatrix} x_{i,1} \\ x_{i,2} \end{bmatrix} + \begin{bmatrix} 0 \\ s \end{bmatrix} u_i, \quad i = 0, 1, \dots, N-1.$$

The initial state $x_0 = (x_{0,1}, x_{0,2})$ is given and the control constraints are

$$(103) \quad -1 \leq u_i \leq 1, \quad i = 0, 1, \dots, N-1.$$

The problem is to minimize

$$(104) \quad J(u) = \frac{s}{2} \sum_{i=0}^{N-1} (x'_{i+1} Q x_{i+1} + R u_i^2),$$

where the matrix Q and the scalar R are given by

$$(105) \quad Q = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \quad R = 6.$$

This problem arises from discretization of the continuous-time problem of minimizing

$$(106) \quad \frac{1}{2} \int_0^T [x(t)' Q x(t) + R u(t)^2] dt$$

subject to

$$(107) \quad \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t)$$

and

$$(108) \quad -1 \leq u(t) \leq 1, \quad t \in [0, T].$$

If the interval $[0, T]$ is discretized into N intervals of length

$$(109) \quad s = \frac{T}{N}$$

and the approximation

$$(110) \quad \dot{x}(t) \approx \frac{1}{s} (x_{i+1} - x_i), \quad t \in [is, (i+1)s]$$

is used, then problem (102)–(105) is a discretized version of problem (106)–(110).

We show in Table 1 for a variety of values of N and s the number of iterations required by the method of Proposition 4 to obtain the exact solution for two initial states $x_0 = (15, 5)$ and $x_0 = (5, -10)$ and two initial control trajectories $u_i^0 \equiv 0$ and $u_i^0 \equiv 1$. In all runs we chose $\varepsilon = 0.01$, $\beta = 0.5$ and $\sigma = 10^{-4}$. All computations were performed in double precision and this was found essential for large values of N . The results demonstrate the ability of the method to identify the set of binding constraints in very few iterations. It is worth noting that while the table gives results for N only up to 10,000, an incomplete set of experiments was run with $N = 25,000$, and a very similar performance was observed for the method.

TABLE 1

x_0	N	s	# of binding constraints at solution	# of iterations	
				$u_i^0 \equiv 0$	$u_i^0 \equiv 1$
(15, 5)	100	0.002	0	1	1
		0.01	18	3	2
		0.04	78	3	3
		0.1	91	4	3
		0.5	100	5	5
	1,000	0.0002	0	1	1
		0.001	183	3	2
		0.004	770	4	3
		0.01	890	4	3
		0.05	705	23	16
	10,000	0.00002	0	1	1
		0.0001	1834	3	3
		0.0004	7693	5	4
		0.001	8861	6	4
		0.005	4261	10	18
(5, -10)	100	0.002	0	1	1
		0.01	48	2	1
		0.04	73	4	3
		0.1	87	5	4
		0.5	100	7	5
	1,000	0.0002	0	1	1
		0.001	478	3	1
		0.004	684	4	3
		0.01	765	5	4
		0.05	370	9	18
	10,000	0.00002	0	1	1
		0.0001	4772	3	2
		0.0004	6802	5	7
		0.001	7595	6	5
		0.005	2591	12	17

REFERENCES

- [1] A. A. GOLDSTEIN, *Convex programming in Hilbert space*, Bull. Amer. Math. Soc., 70 (1964), pp. 709-710.
- [2] E. S. LEVITIN AND B. T. POLJAK, *Constrained minimization problems*, U.S.S.R. Comput. Math. Math. Phys., 6 (1966), pp. 1-50.
- [3] G. P. MCCORMICK, *Anti-zig-zagging by bending*, Management Science, 15 (1969), pp. 315-319.
- [4] D. P. BERTSEKAS, *On the Goldstein-Levitin-Poljak gradient projection method*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 174-184.
- [5] J. C. DUNN, *Newton's method and the Goldstein step-length rule for constrained minimization problems*, this Journal, 6 (1980), pp. 659-674.
- [6] U. M. GARCIA-PALOMARES AND O. L. MANGASARIAN, *Superlinearly convergent quasi-Newton algorithms for nonlinearly constrained optimization problems*, Math. Prog., 11 (1976), pp. 1-13.
- [7] J. B. ROSEN, *The gradient projection method for nonlinear programming, Part I: linear constraints*, J. Soc. Ind. Appl. Math., 8 (1960), pp. 181-217.

- [8] P. E. GILL AND M. MURRAY, eds., *Numerical Methods for Constrained Optimization*, Academic Press, New York, Chs. 2, 3.
- [9] D. GOLDFARB, *Extension of Davidon's variable metric algorithm to maximization under linear inequality and equality constraints*, SIAM J. Applied Math., 17 (1969), pp. 739–764.
- [10] D. G. LUENBERGER, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1973, Ch. 11.
- [11] M. L. LENARD, *A computational study of active set strategies in nonlinear programming with linear constraints*, Math. Prog. 16 (1979), pp. 81–97.
- [12] G. P. MCCORMICK, *The variable reduction method for nonlinear programming*, Management Science, 17 (1970), pp. 146–160.
- [13] R. K. BRAYTON AND J. CULLUM, *An algorithm for minimizing a differentiable function subject to box constraints and errors*, J. Opt. Th. Appl., 29 (1979), pp. 521–558.
- [14] W. MURRAY, *Second derivative methods*, Numerical Methods for Unconstrained Optimization, W. Murray, ed., Academic Press, New York, 1972, pp. 57–71.
- [15] R. FLETCHER AND T. L. FREEMAN, *A modified Newton method for minimization*, J. Opt. Th. Appl., 23 (1977), pp. 357–372.
- [16] J. J. MORÉ AND D. C. SORESENSEN, *On the use of directions of negative curvature in a modified Newton method*, Math. Prog., 16 (1979), pp. 1–20.
- [17] P. P. VARAIYA, *Notes on Optimization*, Van Nostrand-Reinhold, New York, 1972.
- [18] E. POLAK, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1971.
- [19] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1981 (to appear).
- [20] M. D. CANNON, C. D. CULLUM AND E. POLAK, *Theory of Optimal Control and Mathematical Programming*, McGraw-Hill, New York, 1970.
- [21] S. K. MITTER, *Successive approximation methods for the solution of optimal control problems*, Automatica, 3 (1966), pp. 135–149.
- [22] J. C. DUNN, *Global and asymptotic convergence rate estimates for a class of projected gradient processes*, this Journal, 19 (1981), pp. 368–400.

THE SINGULAR STEADY STATE LINEAR REGULATOR*

VIOLET B. HAAS†

Abstract. We consider the problem of infimizing the steady state quadratic cost-functional, whose integrand is given by $\phi'D'D\phi + u'Ru$, with R positive semidefinite, subject to the constraint $\dot{\phi} = A\phi + Bu$, with fixed initial state. It is shown that there may be two separate cost infima, one under the assumption of bounded control energy and one under the additional assumption of partially bounded inputs. Each of them is a positive semidefinite quadratic function of the initial state whose coefficient matrix is one of the positive semidefinite solutions of a pair of algebraic Riccati relations, provided there are no system modes which are uncontrollable, observable and unstable. The cost infimum can be found by solving a sequence of algebraic Riccati equations in a reduced state space.

Let B_0 denote the restriction of B to kernel R . If the unobservable subspace of the pair (D, A) is the same as the supremal (A, B_0) invariant subspace contained in kernel D then the coefficient matrix of the cost infimum is a maximal solution of the pairs of algebraic Riccati relations. In case the cost infimum is a true minimum, then the optimal control is a feedback control when the order of singularity is finite.

1. Introduction. We consider the system described by the differential equation

$$(1.1) \quad \dot{\phi} = A\phi + Bu,$$

where $A: \mathcal{X} \rightarrow \mathcal{X}$, an n -dimensional linear vector space, $B: \mathcal{U} \rightarrow \mathcal{X}$, where \mathcal{U} is an m -dimensional linear vector space, u is a member of the class U of continuous functions having values in \mathcal{U} and $\phi = \phi(t; x, u)$ denotes that trajectory of (1.1) corresponding to the control function u and satisfying

$$(1.2) \quad \phi(0; x, u) = x.$$

The problem, Π , is that of finding $V(x)$ where

$$(1.3) \quad V(x) = \inf_{u \in U} J(x, u) = \inf_{u \in U} \int_0^\infty (\phi'Q\phi + u'Ru) dt.$$

Q and R are constant, symmetric nonnegative definite maps of appropriate dimension. The prime superscript denotes vector or matrix transposition, so that if x and y are vectors then $x'y$ denotes their inner product. The state vectors x and $\phi(t; x, u)$ belong to \mathcal{X} . Let U^0 denote that subclass of U for which

$$(i) \lim_{t \rightarrow \infty} Q\phi(t; x, u) = 0$$

and

$$(ii) \int_0^\infty u'(t)Ru(t) dt < \infty.$$

Let U^1 denote that subclass of U^0 whose members u satisfy

$$(iii) \lim_{t \rightarrow \infty} u(t) = 0.$$

Let $V^1(x) = \inf_{u \in U^1} J(x, u)$ and $V^0 = \inf_{u \in U^0} J(x, u)$. We shall see that sometimes $V(x) = V^0(x)$, sometimes $V(x) = V^1(x)$, and when $V^1(x)$ and $V^0(x)$ are both defined we may have $V^0(x) < V^1(x)$, as the example of § 6 illustrates.

Let \mathcal{B} denote the image of B and let $\langle A|\mathcal{B} \rangle$ denote the union of the subspaces $\mathcal{B}, A\mathcal{B}, \dots, A^{n-1}\mathcal{B}$, of \mathcal{X} . Let $Q = D'D$ and, to avoid trivia, assume $D \neq 0$. Define

$$(1.4) \quad \eta = \bigcap_{i=1}^n \text{Ker } (DA^{i-1}).$$

* Received by the editors March 5, 1980, and in final revised form May 11, 1981. This research was partially supported by the National Science Foundation under grant ECS-7918885.

† School of Electrical Engineering, Purdue University, West Lafayette, Indiana 47907.

Then $\langle A|\mathcal{B} \rangle$ and η are respectively the “subspace of controllable modes” of the pair (A, B) , and the “subspace of unobservable modes” of the pair (D, A) . The set η consists of those states which are indistinguishable from the zero state after observing $y = D\phi$ over any finite time interval. (See [13], Chapters 1 and 2 for details.) Let $\alpha(\lambda)$ denote the minimal polynomial of A and let $\alpha_+(\lambda)$ denote the factor of $\alpha(\lambda)$ associated with those zeros of α which lie in the closed right half of the complex plane. Define

$$\mathcal{X}^+ \triangleq \text{Ker } \alpha_+(A)$$

and let B_0 denote the restriction of B to the kernel of R . \mathcal{X}^+ is the subspace of unstable modes of A . Define \mathcal{V}_0^* as the largest of the subspaces \mathcal{V} of kernel D satisfying $A\mathcal{V} \subset \mathcal{V} + \mathcal{B}_0$. We write,

$$\mathcal{V}_0^* = \sup \mathcal{L}(A, B_0; \ker D)$$

(see [13]). It was shown in [2] that

$$(1.5) \quad \mathcal{X}^+ \subset \langle A|\mathcal{B} \rangle + \mathcal{V}_0^*$$

is a necessary and sufficient condition for the existence of a linear feedback control, $u = F\phi$ in the class U^0 . Thus, (1.5) guarantees that the class U^0 is nonempty and that the infimization problem has meaning. Condition (1.5) also guarantees the existence of a positive constant γ satisfying

$$0 \leq V^0(x) \leq \gamma |x|^2.$$

We shall at times wish to apply the stronger hypothesis

$$(1.6) \quad \mathcal{X}^+ \subset \langle A|\mathcal{B} \rangle + \eta.$$

It was shown in [3] that (1.6) is a necessary and sufficient condition for the existence of a linear feedback control u in the class U^1 .

Let R^* denote the Moore–Penrose pseudoinverse of R and let N be a matrix of maximal rank whose columns span the kernel of R . We shall show that if (1.5) holds then

$$(1.7a) \quad V(x) = V^0(x) = x'P_0x,$$

and if (1.6) holds then

$$(1.7b) \quad V^1(x) = x'P^0x,$$

where P_0 and P^0 are symmetric positive semidefinite solutions of the matrix inequality

$$(1.8) \quad A'P + PA + Q - PBR^*B'P \geq 0$$

and null space condition

$$(1.9) \quad PBN = 0.$$

If (1.5) holds with $\eta = \mathcal{V}_0^*$, then $V^0(x) = V^1(x)$. We shall show that P_0 and P^0 are also solutions of the Popov-type equations

$$A'P + PA + Q = K'K, \quad PB = K'R^{1/2}.$$

Furthermore, if (1.6) holds then this pair of equations has a solution (K, P_+) for which P_+ is maximal, and that in this case $P^0 = P_+$. We shall also show that if $V^i(x)$ is actually

a minimum for $i = 0$ or 1 , and if u^* is the optimal control then

$$(1.10) \quad (PA + A'P + Q - PBR^*B'P)\phi(t; x, u^*) \equiv 0,$$

$$(1.11) \quad \frac{d}{dt} V_x(\phi)' = -Q\phi - A'V_x(\phi),$$

$$(1.12) \quad \frac{d}{dt} (\phi'(t; x, u^*)P\phi(t; x, u^*)) = 0,$$

where P is either P_0 or P^0 as appropriate. If \hat{u} denotes the projection of u^* on the image of R then

$$(1.13) \quad \hat{u}(t) = -R^*B'P\phi(t; x, u^*),$$

and if u^* is singular of finite order then u^* is a linear function of ϕ . If (1.5) holds, but (1.6) does not, then $V^1(x)$ is not defined. Note that (1.6) implies (1.5). If (1.6) holds and η is a proper subspace of \mathcal{V}_0^* then we can expect that $P_0 < P^0$.

The optimal feedback control is found by invoking properties of the Moore–Penrose pseudoinverse of the matrix R . A number of other authors have treated the singular linear regulator problem on finite time intervals for time varying matrices. Their work, much of which is described in references [1] and [4], requires a series of transformations on both the control and state variables to obtain solutions. In order to legitimately use these transformations, certain differentiability and rank conditions must be hypothesized, integral equations must be solved, and the set of admissible controls must be enlarged to permit impulse functions. In a series of papers, [7]–[10], this author obtains solutions to the same problem without any coordinate transformations and without the extra assumptions attendant thereon. The solution of the singular linear regulator problem on a semi-infinite interval was proposed as a fruitful topic for research in [4, Chapt. V].

Examples show (see [5, p. 249]) that when R is singular, optimal trajectories can be expected to exist when x lies only in certain “good” portions of \mathcal{X} , while when x is in the remainder of \mathcal{X} , only an infimum of $J(x, u)$ exists with an “infimizing” control consisting of a concatenation of true optimal trajectories and impulsive arcs.

2. Infimization problems in a reduced space. Let $r = \text{rank } R$, let n be a positive integer, and define

$$R_n = R + \frac{1}{n}NN'.$$

We may suppose, without loss of generality, that coordinates in \mathcal{U} are so chosen that R is diagonal with its first r diagonal entries nonzero, and that NN' contains an $(m-r) \times (m-r)$ identity matrix in its lower right hand corner and that all other elements of NN' vanish. Then

$$R_n^{-1} = R^\# + nNN'.$$

We define a new problem, Π_n , having the same constraints as Π but for which the cost functional is

$$(2.1) \quad J_n(x, u) = \int_0^\infty (\phi'Q\phi + u'R_nu) dt,$$

and define

$$\inf_{u \in \mathcal{U}} J_n(x, u) = V_n(x).$$

It is a standard result (see [12, Thm. 3.7]) that if R is positive definite, if (A, B) is stabilizable and (D, A) is detectable, then the infimum is a true minimum, so that

$$(2.2) \quad V_n(x) = \min_{u \in U} J_n(x, u).$$

Assume (1.6) holds, let $\bar{\mathcal{X}}$ denote the space \mathcal{X} reduced mod η , and let T denote the canonical projection $T: \mathcal{X} \rightarrow \bar{\mathcal{X}}$. Since η is A -invariant, there is a unique map \bar{A} induced in $\bar{\mathcal{X}}$ by A and $\bar{A}T = TA$. Let \bar{D} denote the unique map defined by

$$(2.3) \quad D = \bar{D}T,$$

and let

$$(2.4) \quad \bar{Q} = \bar{D}'\bar{D}.$$

Let $\bar{x} = Tx$, $\bar{\phi} = T\phi$, $\bar{B} = TB$. Then

$$(2.5) \quad J(x, u) = \bar{J}(\bar{x}, u) \triangleq \int_0^\infty (\bar{\phi}'\bar{Q}\bar{\phi} + u'Ru) dt,$$

where

$$(2.6) \quad \dot{\bar{\phi}} = \bar{A}\bar{\phi} + \bar{B}u, \quad \bar{\phi}(0; \bar{x}, u) = \bar{x}.$$

It was shown in [3] that the pair (\bar{A}, \bar{B}) is stabilizable and the pair (\bar{D}, \bar{A}) is observable in $\bar{\mathcal{X}}$. Hence $\bar{D}\bar{\phi}(t; \bar{x}, u) = D\phi(t; x, u) \rightarrow 0$ as $t \rightarrow \infty$ implies $\lim_{t \rightarrow \infty} \bar{\phi}(t; \bar{x}, u) = \bar{0}$.

Let $\bar{\Pi}$ denote the problem of infimizing $\bar{J}(\bar{x}, u)$ subject to (2.6) and let $\bar{\Pi}_n$ denote the same problem with R replaced by R_n . Then for all \bar{x} in $\bar{\mathcal{X}}$, $\bar{\Pi}_n$ has a unique solution u_n , $u_n = -R_n^{-1}\bar{B}'\bar{P}_n\bar{\phi}$, where \bar{P}_n is the unique positive definite solution of the algebraic Riccati equation

$$(2.7) \quad \bar{A}'\bar{P} + \bar{P}\bar{A} + \bar{Q} - \bar{P}\bar{B}R^{-1}\bar{B}'\bar{P} = 0,$$

when $R = R_n$. The matrix $\bar{A} = \bar{B}R_n^{-1}\bar{B}'\bar{P}_n$ is stable, and

$$\bar{V}_n(\bar{x}) \triangleq \min_{u \in U} J_n(\bar{x}, u) = \bar{x}'\bar{P}_n\bar{x}.$$

Premultiplying (2.7) by T' , postmultiplying by T , substituting (2.3) and (2.4) into the result, setting $P = T'\bar{P}T$ and noting that $T'\bar{Q}T = Q$ shows that

$$(2.8) \quad P_n = T'\bar{P}_nT$$

is a not necessarily unique symmetric positive semidefinite solution of

$$(2.9) \quad A'P + PA + Q - PBR^{-1}B'P = 0,$$

when $R = R_n$; that

$$(2.10) \quad \min_{u \in U} J_n(x, u) = x'P_nx,$$

and the optimal control, given by

$$u_n = -R_n^{-1}\bar{B}'\bar{P}_n\bar{\phi} = -R_n^{-1}B'P_n\phi,$$

belongs to the class U^1 . We thus have the following result.

THEOREM 2.1. *Let $Q: \mathcal{X} \rightarrow \mathcal{X}$ be a symmetric positive semidefinite map and let $Q = D'D$. Let $R: \mathcal{U} \rightarrow \mathcal{U}$ be a positive definite map and let $A: \mathcal{X} \rightarrow \mathcal{X}$, $B: \mathcal{U} \rightarrow \mathcal{X}$. Let η be defined as in (1.4) and let $T: \mathcal{X} \rightarrow \bar{\mathcal{X}}$ be the canonical projection. Suppose (1.6) holds. Let $\bar{B} = TB$, let \bar{A} denote the map induced in $\bar{\mathcal{X}}$ by A and let \bar{P} denote the unique symmetric*

positive definite solution of (2.7). Then $P = T'\bar{P}T$ is a positive semidefinite solution of (2.9), $\min_{u \in U} J(x, u) = x'Px$, and the optimal control is given by

$$u = -R^{-1}B'P\phi.$$

Note that the minimization problem can be solved uniquely in the reduced state space, saving computer time and memory.

3. Solution of problem II.

THEOREM 3.1. Suppose $Q: \mathcal{X} \rightarrow \mathcal{X}$ and $R: \mathcal{U} \rightarrow \mathcal{U}$ are positive semidefinite maps. If (1.5) holds then

$$V^0(x) = x'P_0x,$$

and if (1.6) holds then

$$V^1(x) = x'P^0x,$$

where the symmetric positive semidefinite matrices P_0 and P^0 are solutions of (1.8) and (1.9). If $J(x, u)$ has a minimum in either class U^0 or U^1 , if u^* denotes the optimal control function, and if $\hat{u}(t)$ is the projection of $u^*(t)$ on the image of R then (1.10)–(1.13) hold with P replaced by P_0 or P^0 as appropriate.

Proof. We shall first suppose that (1.6) holds. If P_n is defined as in (2.7)–(2.8), then $0 \leq P_{n+1} \leq P_n$. The sequence $\{P_n\}$ thus converges to a limit P^0 , P^0 is symmetric and positive semidefinite and

$$(3.1) \quad V^1(x) \leq x'P^0x.$$

Setting $P = P_n$ and $R = R_n$ in (2.9) and taking limits as $n \rightarrow \infty$ we find that

$$(3.2) \quad A'P^0 + P^0A + Q - P^0BR^*B'P^0 = P^*,$$

where

$$P^* = \lim_{n \rightarrow \infty} nP_nBNN'B'P_n$$

is finite, symmetric and positive semidefinite. Since P^* is finite,

$$\lim_{n \rightarrow \infty} \sqrt{n}N'B'P_n < \infty$$

and

$$(3.3) \quad N'B'P^0 = 0.$$

This proves (1.8) and (1.9).

Now suppose that

$$V^1(x) = x'P^0x - \varepsilon$$

for some positive number ε . There exists u_0 in U^1 such that

$$J(x, u_0) < V^1(x) + \frac{\varepsilon}{4}.$$

We have

$$x'P_nx = V_n(x) \leq J_n(x, u_0) = J(x, u_0) + \frac{1}{n} \int_0^\infty u_0'NN'u_0 dt.$$

Now let n_0 be so large that

$$\int_0^\infty u'_0(t)NN'u_0(t) dt \leq n_0 \frac{\varepsilon}{4}$$

and let $n \geq n_0$. Then

$$x'P_nx \leq J(x, u_0) + \frac{\varepsilon}{4} < V^1(x) + \frac{\varepsilon}{2}.$$

Letting $n \rightarrow \infty$ we find

$$x'P^0x < V^1(x) + \frac{\varepsilon}{2},$$

which contradicts (3.1). Hence (1.7b) must hold.

Now suppose u^* minimizes $J(x, u)$ in U^1 so that

$$(3.4) \quad \min_{u \in U^1} J(x, u) = J(x, u^*) = x'P^0x = V(x),$$

and define $\phi^*(t) \triangleq \phi(t; x, u^*)$. By standard dynamic programming arguments,

$$(3.5) \quad \min_{u \in U^1} [V_x^1(\phi^*)(A\phi^* + Bu) + \frac{1}{2}(\phi^{*'}Q\phi^* + u'Ru)] = 0.$$

Then u^* must satisfy

$$(3.6) \quad B'V_x^1(\phi^*) + Ru^* = 0.$$

Premultiplying (3.6) by R^* and noting that R^*R is a projection map from \mathcal{U} to the image of R , we obtain

$$(3.7) \quad \hat{u} = -R^*B'P^0\phi^*,$$

and this proves (1.13). Substitution of (3.7) and (3.3) into (3.5) yields

$$(3.8) \quad \phi^{*'}(P^0A + A'P^0 + Q - P^0BR^*B'P^0)\phi^* = 0.$$

Since the map in parentheses is symmetric and positive semidefinite then (1.10) holds.

Now define $\lambda = P^0\phi^*$. Using (3.3) and (3.7) we obtain

$$(3.9) \quad \dot{\lambda} = P^0(A - BR^*B'P^0)\phi^*.$$

Substitution of (3.8) into (3.9) yields

$$(3.10) \quad \dot{\lambda} = -A'\lambda - Q\phi^*,$$

and this proves (1.11). Equation (1.12) can now be derived in the usual way. The proof for the case when (1.5) holds is deferred to § 5.

4. The optimal feedback controls. We shall define the Hamiltonian H in the standard way as

$$(4.1) \quad H \triangleq H(t, \phi, u, \lambda) = \frac{1}{2}(\phi'Q\phi + u'Ru) + \lambda'(A\phi + Bu),$$

where $\lambda = P\phi$ and P is a symmetric nonnegative solution of (1.8)–(1.9). From (1.9) and (4.1) it follows that

$$(4.2) \quad N'H'_u = N'B'\lambda = 0,$$

and that

$$N'\frac{d^j}{dt^j}H'_u = 0, \quad j = 1, 2, \dots$$

If u_i denotes the i th component of the vector u and if, along an optimal trajectory in an appropriate coordinate system in \mathcal{U} ,

$$\frac{\partial}{\partial u_i} \frac{d^j}{dt^j} H_{u_i} \equiv 0, \quad j = 0, 1, \dots, p_i - 1,$$

while

$$\frac{\partial}{\partial u_i} \frac{d^{p_i}}{dt^{p_i}} H_{u_i} \neq 0,$$

then it is known that p_i must be even, say $p_i = 2k_i$ (see [11]), and the control function u_i is said to be singular of order k_i . If $k_i = 0$ then u_i is regular. Thus, $u_1^*, u_2^*, \dots, u_r^*$ are all regular. We can completely determine all remaining optimal controls when k_i is finite for all $i = r+1, \dots, m$. Differentiating (4.2) twice, employing (4.1), (1.1) and (3.10) and denoting the projection of u^* on the kernel of R by \tilde{u} , we obtain, after dropping the asterisk from ϕ^* ,

$$(4.3) \quad N'B'QB\tilde{u} = -N'B'[Q(A - BR^*B'P^0) - A'Q - A'^2P^0]\phi.$$

Letting $\tilde{u} = Nz$ and substituting into (4.3) we find

$$(4.4) \quad \hat{z} = -(N'B'QBN)^*N'B'[QA - A'Q - QBR^*B'P^0 - A'^2P^0]\phi,$$

where \hat{z} is the projection of z on the image of $N'B'QBN$. If the controls u_{r+1}, \dots, u_m are all singular of order 1 then the matrix

$$L = (R, B'QB)$$

must have maximal rank, and the converse is true as well. In this case (4.3) may be solved together with (3.6) to determine u^* as a linear function of ϕ . If some control components are singular of order two or more, then the matrix L does not have full rank and further differentiations of (4.2) are required. Since $\lambda = P\phi$, a finite number of differentiations determines the optimal feedback control law when all controls have finite order of singularity.

If $u = F\phi$ is the optimal feedback control law then we must have

$$\int_0^\infty \phi'(t)Q\phi(t) dt < \infty, \quad \int_0^\infty u'(t)Ru(t) dt < \infty.$$

Since $\phi'(t)Q\phi(t) = \bar{\phi}'(t)\bar{Q}\bar{\phi}(t)$, $\bar{Q}\bar{\phi}(t) \rightarrow 0$ then we must have $\lim_{t \rightarrow \infty} \bar{\phi}(t) = 0$. Hence the optimal feedback control stabilizes the reduced system and belongs to U^1 .

We collect our results in the following theorem.

THEOREM 4.1. *Let $Q: \mathcal{X} \rightarrow \mathcal{X}$ and $R: \mathcal{U} \rightarrow \mathcal{U}$ be positive semidefinite maps and assume (1.6). If for $x \in \mathcal{X}$, $J(x, u)$ has a minimum in U^1 and if all control variables have finite order of singularity then the optimal control is a linear feedback control, $u = F\phi$. In the reduced space $\bar{\mathcal{X}} = \mathcal{X}/\eta$ the optimal control is also a linear feedback control, $u = \bar{F}\bar{\phi}$, where $F = \bar{F}T$, and the matrix $\bar{A} + \bar{B}\bar{F}$ is stable.*

5. Proof of Theorem 3.1 when (1.5) holds. We shall now suppose that (1.5) holds, define $\bar{\mathcal{X}}$ as \mathcal{X} reduced mod \mathcal{V}_0^* , and define T as the canonical projection of \mathcal{X} onto $\bar{\mathcal{X}}$. It was shown in [13] that if $A\mathcal{V} \subset \mathcal{V} + \mathcal{B}_0$ then there exists a map F such that $(A + B_0F)\mathcal{V} \subset \mathcal{V}$, or \mathcal{V} is $(A + B_0F)$ -invariant. Define \mathcal{F} as the set of all maps F for which $(A + B_0F)\mathcal{V}_0^* \subset \mathcal{V}_0^*$, and let $F_0 \in \mathcal{F}$. Then \mathcal{V}_0^* is the unobservable subspace of the pair $(D, A + B_0F_0)$. Let $A_0 \triangleq A + B_0F_0$. Then \mathcal{V}_0^* is A_0 -invariant. If \bar{A}_0 is the map induced in $\bar{\mathcal{X}}$ by A_0 then the pair (\bar{A}_0, TB) is stabilizable, and if \bar{D} is the unique map defined by $D = \bar{D}T$, then (\bar{D}, \bar{A}_0) is observable.

Now consider a new infimization problem with cost functional

$$(5.1) \quad J_0(x, v) = \int_0^\infty (\phi' Q \phi + v' R v) dt,$$

subject to the constraints

$$(5.2) \quad \dot{\phi} = A_0 \phi + Bv, \quad \phi(0; x, v) = x.$$

The admissible controls v must satisfy

$$\lim_{t \rightarrow \infty} D\phi(t; x, v) = 0 \quad \text{and} \quad \int_0^\infty v' R v dt < \infty.$$

Since v is arbitrary, subject only to these conditions, then the new infimization problem is equivalent to the old infimization problem over U^0 (i.e., we may suppose that $u = NF_0 \phi + v$). We shall define $J_n(x, v)$ to be the cost functional $J_0(x, v)$ with R replaced by R_n and define $\bar{J}_n(\bar{x}, v)$ accordingly.

Let $\bar{\phi} = T\phi$, $\bar{B} = TB$, $\bar{x} = Tx$ and consider the new infimization problem in the reduced space to minimize

$$(5.3) \quad \bar{J}_0(\bar{x}, v) = \int_0^\infty \bar{\phi}' \bar{D}' \bar{D} \bar{\phi} + v' R_n v dt,$$

subject to the constraints

$$(5.4) \quad \dot{\bar{\phi}} = \bar{A}_0 \bar{\phi} + \bar{B}v, \quad \bar{\phi}(0; x, v) = \bar{x}.$$

We shall henceforth allow only feedback controls of the form $v = \bar{F}\bar{\phi}$ into the competition and modify the definition of U^0 accordingly. Since $\bar{J}_0(\bar{x}, v) = J_0(x, v)$ for all v , so that

$$\inf_{u \in U} J(x, u) = \inf_{v \in U} J_0(x, v) \triangleq V^0(x) = \inf_{v \in U} \bar{J}_0(\bar{x}, v),$$

we shall direct our attention to the investigation of

$$\inf_{v \in U} \bar{J}_0(\bar{x}, v).$$

As before, there exists a positive definite symmetric matrix \bar{P}_n satisfying

$$\min_{v \in U} \bar{J}_n(\bar{x}, v) = \min_{v \in U} J_n(x, v) = \bar{x}' \bar{P}_n \bar{x} = x' P_n x,$$

where \bar{P}_n is the unique positive definite symmetric solution of the steady state Riccati equation,

$$(5.5) \quad \bar{A}_0' \bar{P}_n + \bar{P}_n \bar{A}_0 + \bar{D}' \bar{D} = \bar{P}_n \bar{B} R_n^{-1} \bar{B}' \bar{P}_n,$$

the optimal control is given by

$$v = -R_n^{-1} \bar{B}' \bar{P}_n \bar{\phi},$$

and $P_n = T' \bar{P}_n T$. Exactly as before, the sequence $\{P_n\}$ converges to a limit P_0 , where P_0 satisfies

$$(5.6) \quad A_0' P_0 + P_0 A_0 + D' D - P_0 B R^{-1} B' P_0 \geq 0,$$

and

$$(5.7) \quad N' B' P_0 = 0.$$

Since $B_0 = BN$ then (5.7) implies that $A'_0 P_0 = A' P_0$ and $P_0 A_0 = P_0 A$. Thus P_0 is a solution of (1.8)–(1.9), and exactly as before,

$$(5.8) \quad \inf_{v \in U} J(x, v) = V^0(x) = x' P_0 x.$$

By again invoking (5.7), the rest of Theorem 3.1 can be shown to hold even when (1.5) is true. Theorem 4.1 also holds if we assume (1.5) rather than (1.6).

Remarks. Note that each control u admitted to the competition has the form $u = NF_0 \phi + v$ and belongs to U^0 but not necessarily to U^1 . Note also that since $\bar{\phi}(t; \bar{x}, u) \rightarrow 0$ as $t \rightarrow \infty$ so does $v(t) = \bar{F} \bar{\phi}(t; x, v)$.

If η is a proper subspace of \mathcal{V}_0^* and (1.6) holds then it may happen that $V^0(x) < V^1(x)$, or $x' P_0 x < x' P^0 x$, as the example of the next section demonstrates.

6. An example. We consider the example of [3] where

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad D = (1, 0), \quad R = 0.$$

(A, B) is controllable, (D, A) is observable, \mathcal{V}_0^* is the span of $\text{col}(0, 1)$, and η is a proper subspace of \mathcal{V}_0^* . If $F \in \mathcal{F}$ then $F = (f, -1)$ for arbitrary f . Using $u = Fx + v$ we find that in the reduced space the differential equation becomes $\dot{\bar{\phi}} = f\bar{\phi} + v$. The algebraic Riccati equation for problem $\bar{\Pi}_n$ with this differential equation constraint is

$$n\bar{P}_n^2 - 2f\bar{P}_n - 1 = 0,$$

and for all real f its positive solution converges to zero as n increases. For the original system, we see from (1.9) that all entries of P_0 must be equal to, say, p and from (1.8) we see that $0 \leq p \leq 2$. Note that the maximal solution is $p = 2$. For any initial point on the singular line $2x_2 + x_1 = 0$ there is an optimal control law:

$$u = -x_1 - x_2 \quad \text{in the class } U^1 \quad \text{and } V^1(x) = 2(x_1 + x_2)^2.$$

However, for arbitrary initial point and arbitrary $F = (f, -1)$ in \mathcal{F} , there is a sequence $\{u_n\}$ of controls,

$$u_n = -x_2 + fx_1 + v_n,$$

where $v_n = -(f+n)x_1$, such that

$$\lim_{n \rightarrow \infty} J(x, u_n) = \lim_{n \rightarrow \infty} J_0(x, v_n) = 0.$$

Hence $V^0(x) = 0$. Each u_n for $n > 1$ belongs to the class U^0 but not to the class U^1 . It should be noted here that the optimal singular control law is unstable (in the sense that any disturbance will cause both x_2 and u to become unbounded), and therefore offers no practical advantage over a control such as $u = fx_1 - x_2$, which yields a lower cost for appropriate f , but whose magnitude increases indefinitely as $t \rightarrow \infty$.

7. The Popov equations. Here we consider the equations

$$(7.1) \quad \bar{P}\bar{A} + \bar{A}'\bar{P} + \bar{Q} = \bar{K}'\bar{K},$$

$$(7.2) \quad \bar{P}\bar{B} = \bar{K}'R^{1/2}$$

in the reduced state space. It was shown in [6] that if instead of (1.5) we suppose that the pair (\bar{A}, \bar{B}) is controllable then (7.1)–(7.2) has a solution (\bar{P}_+, \bar{K}_+) for which \bar{P}_+ is

maximal, and a solution (\bar{P}_-, \bar{K}_-) for which \bar{P}_- is minimal. Furthermore,

$$(7.3) \quad \inf_{u \in U^1} \int_0^\infty (\bar{\phi}' \bar{Q} \bar{\phi} + u' R u) dt = \bar{x}' \bar{P}_+ \bar{x}$$

and

$$(7.4) \quad \inf_{u \in U^1} \int_{-\infty}^0 (\bar{\phi}' \bar{Q} \bar{\phi} + u' R u) dt = -\bar{x}' \bar{P}_- \bar{x}.$$

We may draw the same conclusion under the weaker hypothesis (1.6). First note that in this case U^1 is not empty. We have seen that under hypothesis (1.6) there is at least one solution (\bar{P}_0, \bar{K}_0) of (7.1)–(7.2), where

$$\bar{P}_0 = \lim_{n \rightarrow \infty} \bar{P}_n, \bar{K}_0 = R^{1/2} R^* \bar{B}' \bar{P}_0 + \lim_{n \rightarrow \infty} \sqrt{n} N B' \bar{P}_n.$$

Let (\bar{P}, \bar{K}) be any solution of (7.1)–(7.2). Then

$$\int_0^T (\bar{\phi}'_0 \bar{Q} \bar{\phi}_0 + u'_0 R u_0) dt = \bar{x}' \bar{P} \bar{x} + \int_0^T |\bar{K} \bar{\phi}_0 + R^{1/2} u_0|^2 dt - \bar{\phi}_0(T)' \bar{P} \bar{\phi}_0(T).$$

Taking limits as $T \rightarrow \infty$ we see that

$$(7.5) \quad \int_0^\infty (\bar{\phi}'_0 \bar{Q} \bar{\phi}_0 + u'_0 R u_0) dt \geq \bar{x}' \bar{P} \bar{x}.$$

Inequality (7.5) holds for all solutions $(\bar{\phi}_0, u_0)$ of (2.6) with $u_0 \in U^1$, since (\bar{D}, \bar{A}) is observable. Thus,

$$\bar{x}' \bar{P}^0 \bar{x} \geq \bar{x}' \bar{P} \bar{x}$$

and \bar{P}^0 is maximal.

The analogous conclusion cannot be drawn for the solution $P^0 = T' \bar{P}^0 T$ in the space \mathcal{X} . Since (D, A) is not necessarily observable, a feedback control $u_0 = F \phi_0$ does not necessarily stabilize the system in \mathcal{X} . Given any solution (\bar{P}, \bar{K}) of (7.1)–(7.2) there is a corresponding solution $P = T' \bar{P} T$, $K = \bar{K} T$ of

$$(7.6) \quad A' P + P A + Q = K' K,$$

$$(7.7) \quad P B = K' R^{1/2}.$$

The converse, however, is not true. There may be solutions (P, K) of (7.6)–(7.7) which do not correspond to any solutions of (7.1)–(7.2).

REFERENCES

- [1] D. J. BELL AND D. H. JACOBSON, *Singular Optimal Control Problems*, Academic Press, New York, 1975.
- [2] S. P. BHATTACHARYYA, *Regulation in linear systems with partially bounded inputs*, IEEE Trans. Aut. Contr., AC-18 (1973), pp. 684–685.
- [3] ———, *Output regulation with bounded energy*, *ibid.*, pp. 381–383.
- [4] D. J. CLEMENTS AND B. D. O. ANDERSON, *Singular Optimal Control: The Linear-Quadratic Problem*, Springer-Verlag, New York, 1978.
- [5] A. E. BRYSON AND Y.-C. HO, *Applied Optimal Control*, Blaisdell, Waltham, MA, 1969.
- [6] W. A. COPPEL, *Linear-quadratic optimal control*, Proc. Royal Society of Edinburgh, 73A (1974–75), pp. 271–289.
- [7] V. B. HAAS, *On normality and conjugate point criteria for singular extremals*, this Journal, 13 (1975), pp. 1172–1182.

- [8] ———, *On the singular Bolza problem*, Adv. in Math., 24 (1977), pp. 189–205.
- [9] ———, *The Clebsch and Jacobi conditions for singular extremals*, Internat. J. Control, 27 (1978), pp. 557–570.
- [10] ———, *Positive definiteness of a quadratic functional*, IEEE Trans. Aut. Contr., AC-24 (1979), pp. 970–974.
- [11] A. J. KRENER, *The high order maximum principle and its applications to singular extremals*, this Journal, 15 (1977), pp. 256–293.
- [12] H. KWAKERNAAK AND R. SIVAN, *Linear Optimal Control Systems*, Wiley-Interscience, New York, 1972.
- [13] W. M. WONHAM, *Linear Multivariate Control*, Springer-Verlag, New York, 1979.

ON THE COMPUTATION OF THE REACHABLE/OBSERVABLE CANONICAL FORM*

R. E. KALMAN†

Abstract. This note is concerned with questions related to the definition of “canonical form” for linear systems. Some misunderstandings in a previous note by Boley in this journal [SIAM J. Control Optim., 18 (1980), pp. 624–626], are clarified and some new research problems are suggested.

In a recent note to this Journal, Boley [1980] claims that the procedure given in § 7 of my paper, Kalman [1963], for the computation of the reachability¹/observability canonical forms of a linear system $\Sigma = (F, G, H)$ is erroneous.

Although Boley’s claim is not quite legally correct (see below), he has a point: *the meaning of a “canonical form” in my paper is not sufficiently clear.* This is actually a serious omission in my paper. A small modification gives a satisfactory computing procedure but there are also deeper issues revolving around “genericity”.

My paper shows quite correctly that the state space of every linear system may be decomposed as a direct sum

$$(1) \quad X = X_A \oplus X_B \oplus X_C \oplus X_D,$$

so as to give a corresponding block decomposition of the defining matrices as

$$(2) \quad F_{\text{can}} = \begin{bmatrix} F_{AA} & F_{AB} & F_{AC} & F_{AD} \\ 0 & F_{BB} & 0 & F_{BD} \\ 0 & 0 & F_{CC} & F_{CD} \\ 0 & 0 & 0 & F_{DD} \end{bmatrix}, \quad G_{\text{can}} = \begin{bmatrix} G_A \\ G_B \\ 0 \\ 0 \end{bmatrix}, \quad H_{\text{can}} = [0 \quad H_B \quad 0 \quad H_D],$$

In this decomposition, the subspaces involved are constructed in such a way that they have the following interpretation:

- $$(3) \quad \begin{aligned} X_A &:= \text{states which are reachable but unobservable,} \\ X_B &:= \text{states which are both reachable and observable,} \\ X_C &:= \text{states which are neither reachable nor observable,} \\ X_D &:= \text{states which are observable but not reachable.} \end{aligned}$$

It is pointed out in Kalman [1962] that only the space X_A may be defined in a natural (coordinate-free) fashion in the style of abstract linear algebra; the other spaces cannot be defined uniquely and are characterized only by the direct sum property (1) and the system-theoretic properties (3). As mentioned in Kalman [1962, line following (8.1)] the matrix decomposition (2) is “canonical” in the sense that it does not depend on the

* Received by the editors February 24, 1981.

† Mathematische Systemtheorie, ETH-Zentrum, Hauptgebäude, CH-8092, Zurich, Switzerland. This research was supported in part by the U.S. Air Force under grant AFOSR 76-3034(D) and the U.S. Army Research Office under grant DAAG29-77-G-0225 through the Center for Mathematical System Theory, University of Florida, Gainesville, Florida 32611.

¹ The term “controllability” in Kalman [1963] was superseded two years later for algebraic reasons by the term “reachability”. See Kalman, Falb and Arbib [1969]. For continuous-time systems, the class which is discussed in Kalman [1963], the two concepts are equivalent. If we wish to apply the decomposition theorem to discrete-time systems, however, we must replace “controllability” by “reachability”. I use here the modern terminology.

nonuniqueness of the definitions of X_B , X_C , X_D in (3). These results are expressed in Kalman [1963, Thm. 5].

Note that the claimed decompositions (1), (3) of the state space and the corresponding special matrix form (2) are a universal result, that is, they are valid *without any assumption* on the triple $\Sigma = (F, G, H)$.

Boley's claim that "an example has been exhibited which fails to have a property essential to the commonly used procedure for computing the joint controllability-observability canonical form of a linear dynamic system" is not precisely correct. (He is, however, quite right in pointing out that the claims

$$(4) \quad n_C < n_c \quad \text{and} \quad n_D > n_d$$

of Kalman [1963, p. 174, second line after (7.9)] are false. However, these claims were intended merely as motivation for the computational procedure and are not essential for it.) In fact, the matrix (5) given by Boley satisfies (2) if we take $\dim X_A = \dim X_B = 1$, $\dim X_C = 0$ and $\dim X_D = 2$. So the procedure given in Kalman [1963, § 7] *does* reduce any triple to the form (2).

The misunderstanding arises from the fact that Boley takes it for granted that I have claimed in Kalman [1963] also the converse of the decomposition theorem cited above, namely that

$$(5) \quad \text{If a triple has the form (2) then the corresponding direct-sum decomposition of the state space necessarily has the properties listed in (3).}$$

Such a claim is *not* made in Kalman [1963], *but something should have been said about it*.

The wording of my § 7 is at least misleading in that it seems to imply that once we get (F, G, H) into the form (2) we have determined a decomposition of the state space which automatically possesses properties (3). Boley's counterexample shows that this is not so, (5) is false, and (2) does not imply (3). In other words, (2) is *not canonical in the strong sense* that it is equivalent to the decomposition (1) + (3).

As I have argued in recent publications, the word "canonical" should be invoked with much greater care in the current system-theoretic literature. Now I see that the injunction also applies to me, retroactively.

In view of these facts, nothing else can be done but fix up § 7 of my paper so that the computing procedure simultaneously meets requirements (2) *and* (3).

Fortunately, this is easily done. Apply the first half of step (b) but then use the second half (the computation of $(\bar{F}^{22}, 0, \bar{H}^2)$ in the notation of my paper) only to start the corrected version of step (c). This goes as follows, still using the notation of my paper:

(c*) Consider the triple Σ_* with state space X_* given by

$$(6) \quad F_* = \begin{bmatrix} F^{BB} & \bar{F}^{B2} \\ 0 & \bar{F}^{22} \end{bmatrix}, \quad G_* = \begin{bmatrix} G^B \\ 0 \end{bmatrix}, \quad H_* = [H^B, \bar{H}^2],$$

where F^{BB} , \bar{F}^{B2} , G^B , \bar{F}^{22} , \bar{H}^2 and H^B are computed by the first half of step (b).

Now compute the space X_C of unobservable states for Σ_* . Since $\Sigma_B = (F^{BB}, G^B, H^B)$ is observable by the first part of step (b), it is clear that $X_C \cap X_B = \{0\}$. In view of this fact we may write $X_* = X_B \oplus X_C \oplus X_D$, this relation being the only requirement on the definition of X_D . $H^C = 0$ is obvious. The definition of observability means that X_C is an F -invariant subspace. Consequently in the matrix decomposition we must have $F^{BC} = 0$, $F^{DC} = 0$. Moreover, $F^{CB} = 0$ and $F^{DB} = 0$ because X_B is also an

F -invariant subspace. If we collect these facts, the defining matrices for Σ_* become

$$\begin{bmatrix} F^{BB} & 0 & F^{BD} \\ 0 & F^{CC} & F^{CD} \\ 0 & 0 & F^{DD} \end{bmatrix}, \quad \begin{bmatrix} G^B \\ 0 \\ 0 \end{bmatrix}, \quad [H^B \quad 0 \quad H^D].$$

Combining these results with step (4) yields the form (2). Of course, the matrices F^{AB} , F^{AC} , F^{AD} will be, in general, nonzero.

Contrary to what was published in 1963, the present procedure assures that $X_A \oplus X_C$ is the space of all unobservable states. Thus (c*) implies (3).

Finally, it is interesting to note that Boley's counterexample to (3) implying (5) is a nongeneric system. In other words, he has $X_J^{\text{obs}} \oplus X_B \supsetneq X^{\text{obs}}$; this strict inclusion is possible only when certain determinants vanish.

As a matter of fact, it is easy to prove (by computing the reachability and observability matrices for (2)) that *generically* (2) is equivalent to (1) + (3). In this case (2) is indeed a canonical form in a much stronger sense than is claimed in Kalman [1963], and therefore, the original (rather clumsy) recipe given for steps (b) and (c) of § 7 may be allowed to stand.

Evidently, a full understanding of these issues requires bringing in the machinery of algebraic geometry. This will be done elsewhere.

REFERENCES

- D. L. BOLEY [1980], *On Kalman's procedure for the computation of the controllable/observable canonical form*, this Journal, 18, pp. 624–626.
- R. E. KALMAN [1962], *Canonical structure of linear dynamical systems*, Proc. Natl. Acad. Sci. U.S., 48, pp. 596–600.
- [1963], *Mathematical description of linear dynamical systems*, SIAM J. Control, 1, pp. 152–192.
- R. E. KALMAN, P. L. FALB AND M. A. ARBIB [1969], *Topics in Mathematical System Theory*, McGraw-Hill, New York.

OPTIMAL CONTROL FOR PARTIALLY OBSERVED DIFFUSIONS*

WENDELL H. FLEMING† AND ETIENNE PARDOUX‡

Abstract. Stochastic control problems are considered in which a state process X_t and an observation process Y_t are governed by Ito-sense stochastic differential equations driven by independent Brownian motions. The control U_t enters linearly in the dynamics of X_t . A “separated” control problem is introduced, in which the state at any time t is a measure Λ_t representing an unnormalized conditional distribution for X_t given Y_s, U_s for $s \leq t$. The method depends on introducing a pathwise version of Λ_t which depends continuously on observation and control trajectories Y, U . Existence of an optimal control is obtained in a suitable class, larger than the usual class of controls admissible in the strict sense that U_t is measurable on the σ -algebra $\mathcal{F}_t(Y)$ generated by observations $Y_s, s \leq t$. The dynamics of Λ_t are studied using a method of forward and backward partial differential equations. Under a suitable nondegeneracy condition, the measure Λ_t has a density $q(t, x)$ with respect to Lebesgue measure.

1. Introduction. In this paper we are concerned with optimal control problems of the following kind. Let X_t denote the process which we wish to control, Y_t the observation process and U_t the control process, $0 \leq t \leq T$, with T fixed. The state and observation processes are governed by stochastic differential equations

$$(1.1) \quad \begin{aligned} (a) \quad & dX_t = b(X_t, Y_t, U_t) dt + \sigma(X_t, Y_t) dW_t, \\ (b) \quad & dY_t = h(X_t) dt + d\tilde{W}_t. \end{aligned}$$

We shall assume that $b(x, y, u)$ is linear in u . See condition (A₂) in § 2. X_t has values in N -dimensional \mathbb{R}^N , Y_t values in \mathbb{R}^M and U_t values in $\mathcal{U} \subset \mathbb{R}^L$. X_0 has given distribution μ and $Y_0 = 0$. In (1.1), W and \tilde{W} are independent standard Wiener processes with values in $\mathbb{R}^D, \mathbb{R}^M$, respectively. The matrix σ is thus $N \times D$.

The problem is to minimize a criterion of the form

$$(1.2) \quad J = E \left\{ \int_0^T F(X_t, U_t) dt + G(X_T) \right\}.$$

It is customary to require that U_t be measurable with respect to the σ -algebra generated by the observations $Y_s, 0 \leq s \leq t$. We call this the *strict sense* version of the problem (§ 6). For several years the question of proving a general theorem about existence of optimal controls in the strict sense has been open. We do not obtain such a result here. In fact, our results together with a counterexample of Varadhan (§ 6), strongly suggest that, if there is indeed a general existence theorem for strict sense optimal controls, then standard methods are not adequate to prove it. There is a similar difficulty in proving existence of optimal controls with complete observations with singular noise coefficient σ , if the term “complete observations” is taken in the strict sense that U_t depends on the past of the Wiener process driving the system.

Instead of allowing only strict-sense controls, we obtain existence of a minimum in a wider class of controls. Roughly speaking, this wider class is obtained as follows.

* Received by the editors August 6, 1980, and in final revised form April 2, 1981. This research was supported in part by the Air Force Office of Scientific Research under grant AF-AFOSR 76-3063C and in part by the National Science Foundation under grant MCS-79-03554, and by the Centre National de la Recherche Scientifique.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. Part of the research was done during a visit by this first author to the Centro de Investigacion y de Estudios Avanzados, IPN, Mexico.

‡ Université de Provence, 13331 Marseille, Cedex 3, France.

Let

$$(1.3) \quad Z_t = \exp \left[\int_0^t h(X_s) \cdot dY_s - \frac{1}{2} \int_0^t |h(X_s)|^2 ds \right].$$

Then W_t, Y_t are independent standard Wiener processes under a new probability measure \tilde{P} related to the original probability measure \hat{P} by $d\tilde{P}/d\hat{P} = Z_T^{-1}$. In the wider sense formulation we wish to require merely that U_t be independent of future increments $Y_r - Y_t$ for $t \leq r$ and independent of the W process, with respect to \tilde{P} . In § 2 we give a precise formulation of this idea, in which the control is defined as the joint probability distribution measure π of the processes Y, U .

A principal objective of the present paper is to study an associated problem, which we call a “separated” stochastic control problem. In the separated problem the state at time t is a measure Λ_t on \mathbb{R}^N , which is an unnormalized conditional distribution of X_t . The dynamics of the measure-valued process Λ_t obey the Zakai equation of nonlinear filtering ((5.8) below). In §§ 4 and 7 we use the separated problem to obtain results about existence of optimal controls. Bismut [3] (this issue, pp. 302–309) has obtained a similar, and somewhat stronger, existence theorem by another method which does not use a separated problem. In a sequel to this paper [10] (this issue, pp. 286–301) our results are applied to obtain a nonlinear semigroup connected with the separated stochastic control problem.

Our method depends on introducing a “pathwise” version $\Lambda_t = \Lambda_t^{Y,U}$ of the unnormalized conditional distribution measure for X_t given past values of the observation process Y and the control process U (§ 3). An important fact is the continuous dependence of Λ_t on Y, U and the initial distribution μ (Lemma 3.2). In § 5 we study the dynamics of Λ_t using a method of forward and backward partial differential equations. Similar ideas were used in [17]–[19], for the nonlinear filter problem. The forward equation (5.4) is linear parabolic (possibly degenerate) with coefficients depending parametrically on observations Y_t and controls U_t . Under suitable regularity assumptions, Λ_t has a density $q(t, x)$, related in a simple way to a solution $p(t, x)$ of the forward equation via (5.6). Without the regularity assumptions, one uses instead a weak sense version (5.4') of the forward equation. In § 7 we find that under a different set of assumptions $p(t, x)$ is a solution to (5.4), in the sense of the Hilbert space theory of linear parabolic partial differential equations.

The method of backward and forward equations was applied to the nonlinear filter problem in [17], working directly with the Zakai equation and its adjoint and allowing correlations between the Wiener processes W, \tilde{W} driving the state and observation equations. However, technical difficulties are encountered in adapting that method to the control problem. Our derivation is similar to that given in [18] and [19] in the nonlinear filtering set-up with W_t and \tilde{W}_t independent.

In [9] another “separated” control problem was considered. In that formulation the “state” for the separated problem corresponds to the (normalized) conditional distribution measure of X_t given past observations. Some of the results in [9] are proved under assumptions not satisfied when X_t is a controlled, partially observed solution to (1.1a). Hence, the results of [9] are complementary to those in the present paper.

Recently Haussmann [13] also obtained an existence theorem, using different methods and a somewhat different concept of wider sense admissible control.

In [5] Christopheit proved an existence theorem for optimal stochastic controls under partial observations. In that work, the observation process is a deterministic function of (part of) the past trajectory of the state process, and the optimal control

is sought in a class of feedback controls. Both his results and his methods of proof differ significantly from ours and those of Bismut.

2. Formulation of the problem. We make the following assumptions about the functions appearing in (1.1).

(A1) σ is a bounded, continuous $N \times D$ matrix-valued function on \mathbb{R}^{N+M} . Moreover, $\sigma(\cdot, y)$ is Lipschitz on \mathbb{R}^N with Lipschitz constant not depending on $y \in \mathbb{R}^M$.

(A2) $b(x, y, u) = b^0(x, y) + b^1(x, y)u$, where b^0, b^1 are bounded, continuous functions on \mathbb{R}^{N+M} . Moreover, $b^l(\cdot, y)$ is Lipschitz on \mathbb{R}^N with Lipschitz constant not depending on $y \in \mathbb{R}^M$, for $l = 0, 1$.

Note that in (A2), b^0 has values in \mathbb{R}^N , while b^1 has $N \times L$ matrices as values.

We write $C_b(\mathbb{R}^N)$ for the space of bounded continuous real-valued functions on \mathbb{R}^N and $C_0(\mathbb{R}^N)$ for the space of continuous functions with compact support. We write $C_b^k(\mathbb{R}^N)$, $C_0^k(\mathbb{R}^N)$ for the spaces of functions whose partial derivatives of orders $\leq k$ are in $C_b(\mathbb{R}^N)$, $C_0(\mathbb{R}^N)$, respectively. Similarly, we write $C_b^k(\mathbb{R}^N; \mathbb{R}^M)$, $C_0^k(\mathbb{R}^N; \mathbb{R}^M)$ if the functions are \mathbb{R}^M -valued.

(A3) $h \in C_b^2(\mathbb{R}^N; \mathbb{R}^M)$.

In § 7 we shall assume that σ is nonsingular $N \times N$. One could also let b, σ, h depend on t , with minor changes in the results, and proofs. This would only be a generalization in § 7, since in §§ 2–6 t can be adjoined as an additional x component.

(A4) \mathcal{U} is a convex, compact subset of \mathbb{R}^L .

Choose any $T > 0$ which will be fixed throughout the paper. We formulate the control problem on the “canonical” sample space

$$\Omega = \Omega_0 \times \Omega_1 \times \Omega_2 \times \Omega_3,$$

where Ω_0, Ω_1 are $C([0, T]; \mathbb{R}^m)$ with $m = D, N$, respectively,

$$\Omega_2 = \{Y \in C([0, T]; \mathbb{R}^M) : Y_0 = 0\}, \quad \Omega_3 = L^2([0, T]; \mathcal{U}).$$

The elements $\omega = (W, X, Y, U)$ of Ω satisfy

$$\omega(t) = (W_t(\omega), X_t(\omega), Y_t(\omega), U_t(\omega)), \quad 0 \leq t \leq T.$$

We give $\Omega_0, \Omega_1, \Omega_2$ the usual norm topology and Ω_3 the weak topology, which is metrizable and separable since \mathcal{U} is compact [2, p. 238]. Let

$$\Omega^1 = \Omega_0 \times \Omega_1, \quad \Omega^2 = \Omega_2 \times \Omega_3,$$

whose respective elements are pairs $(W, X), (Y, U)$. Let $\mathcal{F}_t(W) = \sigma\{W_s, 0 \leq s \leq t\}$, with $\mathcal{F}_t(X), \mathcal{F}_t(Y)$ defined similarly. Let

$$\mathcal{F}_t(U) = \sigma\{V_s, 0 \leq s \leq t\}, \quad V_t = \int_0^t U_s ds.$$

The elements of these σ -algebras are subsets of $\Omega_0, \dots, \Omega_3$, respectively. However, can also regard them as σ -algebras of subsets of Ω, Ω^1 or Ω^2 , with the obvious identifications. For example, $A \in \mathcal{F}_t(X)$ can be identified with $\Omega_0 \times A \times \Omega_2 \times \Omega_3$. We shall also use the σ -algebras

$$\begin{aligned} \mathcal{G}_t^1 &= \mathcal{F}_t(W) \times \mathcal{F}_t(X), & \mathcal{G}_t^2 &= \mathcal{F}_t(Y) \times \mathcal{F}_t(U), \\ \mathcal{H}_t &= \mathcal{G}_t^1 \times \mathcal{G}_t^2 = \mathcal{F}_t(W) \times \dots \times \mathcal{F}_t(U). \end{aligned}$$

We note that $\mathcal{F}_T(U)$ is the Borel σ -algebra of Ω_3 , and thus, \mathcal{G}_T^2 is the Borel σ -algebra of Ω^2 .

Remark. Intuitively, by using the indefinite integral V_t instead of U_t in defining $\mathcal{F}_t(U)$, we need not be concerned with changes in U_t on subsets of $[0, T]$ of Lebesgue measure 0. An alternative to our formulation would be to consider quadruples (W, X, Y, V) instead of (W, X, Y, U) , using the uniform norm on V . By (A_2) the control enters linearly in b . Hence, one can write, in the integrated form of (1.1a),

$$\int_0^t b^1(X_s, Y_s) U_s ds = \int_0^t b^1(X_s, Y_s) dV_s,$$

the right side being a Riemann-Stieltjes integral. This device was used in [11], but we use here U_t instead.

Distribution of (W, X) conditioned on (Y, U) . Let $Y = Y_\cdot, U = U_\cdot$ be given sample paths for the observation and control processes; thus $(Y, U) \in \Omega^2$. Consider (1.1a) with initial data $W_0 = 0, X_0 = x$. Assumptions $(A1), (A2)$ imply the Ito conditions. There is a solution to (1.1a) which is pathwise unique, and hence also unique in probability law. Let $\bar{P}_x^{Y, U}$ denote the distribution measure of (W, X) given (Y, U) . Then $\bar{P}_x^{Y, U}$ lies in the space of probability measures on \mathcal{G}_T^1 . By convergence of a sequence of probability measures \bar{P}_n to \bar{P} we mean weak convergence, namely $\int_{\Omega^1} g(W, X) d\bar{P}_n \rightarrow \int_{\Omega^1} g(W, X) d\bar{P}$ for all $g \in C_b(\Omega^1)$.

LEMMA 2.1. $\bar{P}_x^{Y, U}$ depends continuously on x, Y, U .

This lemma is essentially known (cf. Stroock-Varadhan [20]). However, for completeness we outline a proof in the Appendix.

Following the motivation described in § 1, the formal definition of admissible control is as follows.

DEFINITION. An *admissible control* π is a probability measure on $(\Omega^2, \mathcal{G}_T^2)$ such that Y is a $\pi, \{\mathcal{G}_t^2\}$ Wiener process.

Remark. The projection $(Y, U) \rightarrow Y$ maps π onto Wiener measure. The definition of admissible control requires, in addition, that $\int_0^t U_s ds$ be independent of $Y_r - Y_t$ for $t \leq r \leq T$.

Let \mathfrak{A} denote the set of all admissible controls π . Given a distribution μ for X_0 , each $\pi \in \mathfrak{A}$ determines a joint distribution measure P_π of (W, X, Y, Z) as follows. Define $\bar{P}^{Y, U} = \bar{P}_\mu^{Y, U}$ by

$$\bar{P}^{Y, U}(A) = \int_{\mathbb{R}^N} \bar{P}_x^{Y, U}(A) d\mu(x), \quad A \in \mathcal{G}_T^1.$$

We then define \bar{P}_π on \mathcal{H}_T by

$$(2.1) \quad \bar{P}_\pi(dW, dX, dY, dU) = \bar{P}^{Y, U}(dW, dX) \pi(dY, dU).$$

The projection of \bar{P}_π under $(W, X, Y, U) \rightarrow (Y, U)$ is π . The family of probability measures $\bar{P}^{Y, U}$ gives a regular conditional distribution for (W, X) . If $g(W, X)$ is \mathcal{G}_t^1 -measurable, then $\bar{E}^{Y, U} g(W, X)$ is \mathcal{G}_t^2 -measurable, where we write $\bar{E}^{Y, U}, \bar{E}_\pi$ for expectations with respect to $\bar{P}^{Y, U}, \bar{P}_\pi$. We then have for any \mathcal{H}_t -measurable ψ with $\bar{E}_\pi |\psi| < \infty$

$$(2.2) \quad \bar{E}_\pi(\psi | \mathcal{G}_t^2) = \bar{E}^{Y, U}(\psi), \quad \pi - \text{a.s.}$$

We define P_π by

$$(2.3) \quad \frac{dP_\pi}{d\bar{P}_\pi} = Z_T,$$

with Z_T as in (1.3). Since $h(x)$ is bounded, $P_\pi(\Omega) = \bar{E}_\pi(Z_T) = 1$.

For each (Y, U) , W is a $\bar{P}^{Y,U}$ -standard Wiener process, and X satisfies the stochastic differential equation (1.1a) $\bar{P}^{Y,U}$ -a.s. With respect to \bar{P}_π , W and Y are independent standard Wiener processes.

LEMMA 2.2. *Let $\tilde{W}_t = Y_t - \int_0^t h(X_s) ds$. Then \tilde{W}, W are independent standard Wiener processes under P_π and the stochastic differential equations (1.1a), (1.1b) hold P_π -a.s..*

Proof. Since the pair (\tilde{W}) is a \bar{P}_π -standard Wiener process, of dimension $N + M$, the Cameron–Martin–Girsanov formula and (2.3) imply that (\tilde{W}) is a P_π -standard Wiener process. Since (1.1a) holds \bar{P}_π -a.s. and $P_\pi \ll \bar{P}_\pi$, (1.1a) holds P_π -a.s.; while (1.1b) holds by definition of \tilde{W} .

We have defined as admissible control a probability measure π belonging to the class \mathfrak{A} . Convergence of sequences π_n of admissible controls is taken in the sense of weak convergence of probability measures. \mathfrak{A} is a metric space under (for instance) the Prokhorov metric [2]. Moreover, \mathfrak{A} is a convex set.

LEMMA 2.3. *\mathfrak{A} is compact under weak sequential convergence.*

Proof. Since every measure $\pi \in \mathfrak{A}$ projects onto Wiener measure under $(Y, U) \rightarrow Y$ and the second component U lies in the compact (weak topology) space $L^2([0, T]; \mathcal{U})$, tightness of \mathfrak{A} follows by standard arguments. Hence [2, p. 37], it remains only to show that \mathfrak{A} is closed. Suppose that $\pi_n \rightarrow \pi$, $\pi_n \in \mathfrak{A}$. We must show that Y is a π , $\{\mathcal{G}_t^2\}$ Wiener process. Since π_n projects onto Wiener measure for each n , so does π . We need only verify that $Y_r - Y_t$ is independent of \mathcal{G}_t^2 for $t \leq r$. For this it suffices that for any \mathcal{G}_t^2 -measurable $\phi \in C_b(\Omega^2)$ and $f \in C_b(\mathbb{R}^M)$,

$$\int_{\Omega^2} \phi f(Y_r - Y_t) d\pi = \int_{\Omega^2} \phi d\pi \int_{\Omega^2} f(Y_r - Y_t) d\pi.$$

But this holds for each π_n , and we pass to the limit. This proves Lemma 2.3.

In § 6 we shall consider the subclass \mathfrak{A}^s of strict-sense controls.

Existence of optimal controls. In (1.2) we take $E = E_\pi$ and write $J = J(\pi)$. By Lemma 2.3, $J(\pi)$ has a minimum on \mathfrak{A} provided $J(\pi)$ is lower semicontinuous on \mathfrak{A} , $J(\pi) \geq 0$ and there exists $\pi \in \mathfrak{A}$ such that $J(\pi) < \infty$. We prove lower semicontinuity of $J(\pi)$ in two cases: (1) $F = 0$ and G satisfies condition (A5) in § 4 (see Theorem 4.1); (2) F, G satisfy (A5) and also (A1'), (A6) in § 7 (see Theorem 7.2).

Bismut [3] has obtained similar, somewhat stronger results by another method, which involves neither the separated problem nor the techniques from the Hilbert space theory of parabolic partial differential equations used in § 7. Theorem 7.2 was obtained earlier, and we believe that the method we use to prove it has independent interest.

Using Bismut's method, one can establish lower semicontinuity for a more general criterion of the form

$$J(\pi) = E_\pi \Phi(X, Y, U),$$

if Φ is lower semicontinuous on $\Omega_1 \times \Omega_2 \times \Omega_3$ and $\Phi \geq 0$.

As already pointed out, controls which are admissible in the customary sense correspond to what we call strict sense admissible controls. The existence theorems mentioned above assert that there is a minimum of $J(\pi)$ on the compact space \mathfrak{A} but not necessarily on the subset $\mathfrak{A}^s \subset \mathfrak{A}$ of strict sense admissible controls. However, we show in § 6 that the infimum of $J(\pi)$ on \mathfrak{A}^s is the same as the minimum on \mathfrak{A} , at least when $J(\pi)$ has the form (1.2) and a slightly stronger condition (A5') holds.

It is not known whether $J(\pi)$ has a minimum on \mathfrak{A}^s . Even if this should turn out to be true, it is not clear how to prove it by some version of the usual minimizing sequences technique. The difficulty is to avoid the possibility that the limit of a convergent minimizing sequence in \mathfrak{A}^s is in $\mathfrak{A} - \mathfrak{A}^s$.

3. The unnormalized conditional distribution. We wish to introduce an unnormalized conditional distribution of X_t given controls and observations up to t . Let us take a version of the \tilde{P}_π -martingale Z such that Z_t is \mathcal{H}_t -measurable for $0 \leq t \leq T$. Consider any $f \in C_b(\mathbb{R}^N)$. By (2.2) with $\psi = f(X_t)Z_t$,

$$(3.1) \quad \tilde{E}_\pi(f(X_t)Z_t | \mathcal{G}_t^2) = \bar{E}^{Y,U}(f(X_t)Z_t), \quad 0 \leq t \leq T, \quad \pi - \text{a.s.}$$

Let us rewrite (3.1) in such a way that it is defined for *all* Y, U , not just π -a.s., and depends continuously on (Y, U) . See Lemma 3.2 below. Since $h \in C_b^2(\mathbb{R}^N; \mathbb{R}^M)$, we can integrate $\int_0^t h(X_s) \cdot dY_s$ by parts:

$$\int_0^t h(X_s) \cdot dY_s = h(X_t) \cdot Y_t - \int_0^t Y_s \cdot L_s h(X_s) ds - \int_0^t Y_s \cdot \nabla h(X_s) \sigma(X_s, Y_s) dW_s,$$

where $Y_s \cdot \nabla h$ is the gradient in x of $Y_s \cdot h$ and

$$(3.2) \quad L_s = \frac{1}{2} \sum_{i,j=1}^N a_{ij}(x, Y_s) \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^N b_i(x, Y_s, U_s) \frac{\partial}{\partial x_i},$$

with $a = \sigma\sigma'$. For fixed Y, U , $\partial/\partial s + L_s$ is the backward operator corresponding to (1.1a). Let

$$(3.3) \quad e(s, x) = \frac{1}{2}(a Y_s \cdot \nabla h, Y_s \cdot \nabla h) - Y_s \cdot L_s h - \frac{1}{2}|h|^2,$$

where in (3.3) $(a\xi, \xi) = |\xi\sigma'|^2$ denotes the dot product in \mathbb{R}^N of $a\xi$ with ξ , and \cdot denotes the dot product in \mathbb{R}^M . From (1.3)

$$Z_t = \check{Z}_t \exp(Y_t \cdot h(X_t)) \exp \int_0^t e(s, X_s) ds,$$

where

$$\check{Z}_t = \exp \left[- \int_0^t Y_s \cdot \nabla h(X_s) \sigma(X_s, Y_s) dW_s - \frac{1}{2} \int_0^t (a(X_s, Y_s) Y_s \cdot \nabla h(X_s), Y_s \cdot \nabla h(X_s)) ds \right].$$

For fixed (Y, U) , let us define another probability measure $\check{P}^{Y,U}$ on $(\Omega^1, \mathcal{G}_T^1)$ by

$$(3.4) \quad \frac{d\check{P}^{Y,U}}{d\bar{P}^{Y,U}} = \check{Z}_T.$$

This corresponds to a change in drift coefficient in (1.1a) from b to $\check{b} = b - a Y_s \cdot \nabla h$, and changes L_s in (3.2) to the operator

$$(3.5) \quad \check{L}_s = L_s - (a Y_s \cdot \nabla h, \nabla).$$

(For another characterization of \check{L}_s see (5.12), also Davis [7] for the corresponding formula in nonlinear filtering.) From (3.1) we then have

$$\tilde{E}_\pi(f(X_t)Z_t | \mathcal{G}_t^2) = \check{E}^{Y,U} \left(f(X_t) \exp(Y_t \cdot h(X_t)) \exp \int_0^t e(s, X_s) ds \right),$$

where the right side is now defined for *all* $(Y, U) \in \Omega^2$, not merely π -a.s. For fixed (Y, U) , the right side is a bounded linear functional on $C_b(\mathbb{R}^N)$. Hence, for every $(Y, U) \in \Omega^2$ and $0 \leq t \leq T$ there exists a measure $\Lambda_t^{Y,U}$ on the Borel σ -algebra $\mathcal{B}(\mathbb{R}^N)$ such that

$$(3.6) \quad \langle f, \Lambda_t^{Y,U} \rangle = \check{E}^{Y,U} \left(f(X_t) \exp(Y_t \cdot h(X_t)) \exp \int_0^t e(s, X_s) ds \right),$$

for all $f \in C_b(\mathbb{R}^N)$, where for any measure ν with $\nu(\mathbb{R}^N) < \infty$

$$\langle f, \nu \rangle = \int_{\mathbb{R}^N} f(x) d\nu(x).$$

DEFINITION. $\Lambda_t^{Y,U}$ is the *unnormalized conditional distribution measure*.

The unnormalized conditional distribution measure satisfies, for all $f \in C_b(\mathbb{R}^N)$,

$$(3.7) \quad \langle f, \Lambda_t^{Y,U} \rangle = \check{E}_\pi(f(X_t) Z_t | \mathcal{G}_t^2).$$

As is well known, the (normalized) conditional distribution of X_t satisfies, for all $f \in C_b(\mathbb{R}^N)$,

$$E_\pi(f(X_t) | \mathcal{G}_t^2) = \frac{\langle f, \Lambda_t^{Y,U} \rangle}{\langle 1, \Lambda_t^{Y,U} \rangle},$$

where E_π denotes expectation with respect to the measure P_π defined by (2.3). For fixed t , let

$$(3.8) \quad v_0^{Y,U}(x) = \check{E}_x^{Y,U} \left(f(X_t) \exp(Y_t \cdot h(X_t)) \exp \int_0^t e(s, X_s) ds \right),$$

where $\check{E}_x^{Y,U}$ denotes expectation with respect to the probability measure $\check{P}_x^{Y,U}$ in (3.4) for initial state $X_0 = x$. For initial distribution μ for X_0 ,

$$\check{E}^{Y,U}(g(X)) = \int_{\mathbb{R}^N} \check{E}_x^{Y,U}(g(X)) d\mu(x).$$

Therefore, by (3.6), (3.8),

$$(3.9) \quad \langle f, \Lambda_t^{Y,U} \rangle = \langle v_0^{Y,U}, \mu \rangle,$$

for all $(Y, U) \in \Omega^2$ and $f \in C_b(\mathbb{R}^N)$.

In § 5 we shall see that (3.9) has a natural interpretation in terms of solutions to forward and backward partial differential equations.

Remark. We shall later wish to consider $\Lambda_t^{Y,U}$ corresponding to any $\mu \geq 0$ with $\mu(\mathbb{R}^N) < \infty$, not merely for probability measures μ on $\mathcal{B}(\mathbb{R}^N)$. Given Y, U and μ , the right side of (3.9) is a bounded, nonnegative linear functional of f , by (3.8). This gives an alternate way to define the measure $\Lambda_t^{Y,U}$, without the restriction $\mu(\mathbb{R}^N) = 1$, in such a way that (3.9) holds.

LEMMA 3.1. (a) $v_0^{Y,U}(x)$ is a continuous function of (x, Y, U) .

(b) Given $f \in C_0(\mathbb{R}^N)$ and $a > 0$, there exist $c, k > 0$ (depending on f, a and bounds for $|b|, |\sigma|, |\nabla h|$), such that $|v_0^{Y,U}(x)| \leq c \exp(-k|x|^2)$ for all $x \in \mathbb{R}^N$ and Y, U such that $\|Y\| \leq a$.

Proof. By Lemma A.1 (Appendix), $\check{P}_x^{Y,U}$ depends continuously on x, Y, U . Let $x_n \rightarrow x, Y_n \rightarrow Y, U_n \rightarrow U$, and let (for fixed t)

$$\Psi_n(X) = f(X_t) \exp(Y_{nt} \cdot h(X_t)) \exp \int_0^t e_n(s, X_s) ds,$$

$$\Psi(X) = f(X_t) \exp(Y_t \cdot h(X_t)) \exp \int_0^t e(s, X_s) ds,$$

where e_n is defined by (3.3) with Y, U replaced by Y_n, U_n . For any compact $\Gamma \subset C([0, T]; \mathbb{R}^N)$, $\int_0^t e_n ds \rightarrow \int_0^t e ds$ as $n \rightarrow \infty$, uniformly on Γ . This is proved by the same reasoning used in the proof of Lemma A.1. Then $\Psi_n \rightarrow \Psi$ uniformly on Γ . From this we conclude that $v_0^{Y_n, U_n}(x_n) \rightarrow v_0^{Y, U}(x)$ as $n \rightarrow \infty$, which proves (a).

To prove (b), U_s is bounded by (A4). For $\|Y\| \leq a, Y_t \cdot h(X_t)$ and $e(s, X_s)$ are bounded. Hence, for some c_1 ,

$$|v_0^{Y, U}(x)| \leq c_1 \check{P}_x^{Y, U}(\check{X}_t \in \text{spt } f).$$

However, $\check{P}_x^{Y, U}$ - a.s.

$$dX_t = \check{b}(t, X_t) dt + \sigma d\check{W}_t,$$

$$\check{b} = b - a Y_s \cdot \nabla h, \quad \check{W}_t = W_t + \int_0^t Y_s \cdot \nabla h(X_s) \sigma(X_s, Y_s) ds,$$

and \check{W}_t is a $\check{P}_x^{Y, U}$ -Wiener process. For $\|Y\| \leq a, \check{b}$ is bounded, and σ is bounded by (A1). By standard estimates

$$|X_t - x| \leq c_2 \|\zeta\| + c'_2 t, \quad \zeta_t = \int_0^t \sigma d\check{W}_s,$$

$$\check{P}_x^{Y, U}(X_t \in \text{spt } f) \leq \check{P}_x^{Y, U}(\|\zeta\| > k_1 |x| - k_2),$$

for some $k_1 > 0$ and k_2 ($\|\cdot\|$ is as usual the sup norm). Using the fact that σ is bounded and an exponential martingale inequality,

$$\check{P}_x^{Y, U}(\|\zeta\| > k_1 |x| - k_2) \leq c_3 \exp(-k|x|^2),$$

for some $c_3, k > 0$. See, for example [20, p. 87]. This proves (b).

For $r > 0$, let

$$\mathcal{M}_r = \{\mu \geq 0 \text{ on } \mathcal{B}(\mathbb{R}^N): \|\mu\| \leq r\},$$

where $\|\mu\| = \mu(\mathbb{R}^N)$. We use vague convergence for sequences of measures: $\nu_n \rightarrow \nu$ means that $\langle f, \nu_n \rangle \rightarrow \langle f, \nu \rangle$ for all $f \in C_0(\mathbb{R}^N)$. The next lemma asserts continuous dependence of $\Lambda_t^{Y, U}$ on μ, Y, U for fixed t , provided we restrict μ to \mathcal{M}_r (see the remark preceding Lemma 3.1). We recall that \mathcal{M}_r with the vague topology is metrizable and compact.

LEMMA 3.2. $\Lambda_t^{Y, U}$ is a continuous function of μ, Y, U , on $\mathcal{M}_r \times \Omega^2$.

Proof. Let $\mu_n \rightarrow \mu, (Y_n, U_n) \rightarrow (Y, U)$. Given $f \in C_0(\mathbb{R}^N)$, let

$$v_n(x) = v_0^{Y_n, U_n}(x), \quad v(x) = v_0^{Y, U}(x).$$

By (3.9) it suffices to show that

$$(*) \quad \lim_{n \rightarrow \infty} \langle v_n, \mu_n \rangle = \langle v, \mu \rangle.$$

By Lemma 3.1a, $x_n \rightarrow x$ implies $v_n(x_n) \rightarrow v(x)$. From this fact and continuity of v , $v_n \rightarrow v$ uniformly on compact subsets of \mathbb{R}^N . Since $Y_n \rightarrow Y$, $\|Y_n\| \leq a$ for some a . By Lemma 3.1b, $v_n(x) \rightarrow 0$ as $|x| \rightarrow \infty$, uniformly with respect to n . Since $\|\mu_n\| \leq r$, this implies (*) and, hence, Lemma 3.2.

4. The “separated” control problem. As in (1.2), let

$$(4.1) \quad J(\pi) = E_\pi \left\{ \int_0^T F(X_t, U_t) dt + G(X_T) \right\},$$

with E_π the expectation with respect to the probability measure P_π in (2.3). The minimum problem is: given a distribution measure μ for X_0 , find a control $\pi^* \in \mathfrak{U}$ such that $J(\pi^*) \leq J(\pi)$, for all $\pi \in \mathfrak{U}$. We assume that

(A5) F, G are continuous and $F \geq 0, G \geq 0$. There exists $\pi \in \mathfrak{U}$ such that $J(\pi) < \infty$.

We sometimes impose the stronger condition:

(A5') F, G are continuous, $F \geq 0, G \geq 0$, and for some positive $C, m, l > m$

$$|F(x, u)| \leq C(1 + |x|^m), \quad |G(x)| \leq C(1 + |x|^l), \quad \langle x|^l, \mu \rangle < \infty.$$

Since X_t satisfies the stochastic differential equation (1.1a) with bounded coefficients b, σ , $J(\pi) < \infty$, for all $\pi \in \mathfrak{U}$, provided that (A5') holds. See [12, p. 48].

From (2.3) and the fact that X_t, U_t are \mathcal{H}_t -measurable,

$$J(\pi) = \mathring{E}_\pi \left\{ \int_0^T Z_t F(X_t, U_t) dt + Z_T G(X_T) \right\}.$$

Upon taking conditional expectations and using (3.7),

$$(4.2) \quad \begin{aligned} J(\pi) &= \mathring{E}_\pi \left\{ \int_0^T \mathring{E}_\pi(Z_t F(X_t, U_t) | \mathcal{G}_t^2) dt + \mathring{E}_\pi(Z_T G(X_T) | \mathcal{G}_T^2) \right\}, \\ J(\pi) &= \int_{\Omega^2} \left\{ \int_0^T \langle F(\cdot, U_t), \Lambda_t^{Y, U} \rangle dt + \langle G, \Lambda_T^{Y, U} \rangle \right\} d\pi(Y, U). \end{aligned}$$

In the separated problem we regard the unnormalized conditional distribution measure $\Lambda_t = \Lambda_t^{Y, U}$ as the “state,” and (4.2) as the criterion to be minimized. Initially, $\Lambda_0 = \mu$. The dynamics of the measure-valued process Λ_t will be described in § 5.

In our formulation the separated control problem is completely equivalent to the problem originally formulated in § 2. An optimal control π^* for either problem is also optimal for the other.

In the case $F = 0$ we can now prove the existence of an optimal π^* . In § 7 we shall prove another existence theorem, with $F \neq 0$, using methods of partial differential equations. One cannot, in general, reduce $F \neq 0$ to $F = 0$ by adding a new state variable since linearity would then no longer hold in (A2), § 2.

THEOREM 4.1. *Let $F = 0$. There exists $\pi^* \in \mathfrak{U}$ such that $J(\pi^*) \leq J(\pi)$, for all $\pi \in \mathfrak{U}$.*

Proof. By Lemma 2.3, \mathfrak{U} is compact. It suffices to show that

$$J(\pi) = \int_{\Omega^2} \langle G, \Lambda_T^{Y, U} \rangle d\pi(Y, U)$$

is lower semicontinuous on \mathfrak{U} . For $\rho = C_0(\mathbb{R}^N)$, $H \in C_b(\mathbb{R}^1)$, $0 \leq \rho \leq 1$, $H \geq 0$, let

$$\tilde{J}(\pi) = \int_{\Omega^2} H[\langle \rho G, \Lambda_T^{Y, U} \rangle] d\pi(Y, U).$$

By Lemma 3.2, with μ fixed, the integrand is a bounded continuous function on Ω^2 . Hence \tilde{J} is continuous on \mathfrak{A} . Let $\rho = \rho_n$, $H = H_n$ be increasing sequences such that $\rho_n(x) \rightarrow 1$, $H_n(z) \rightarrow z$ as $z \rightarrow \infty$. Then $J(\pi)$ is the limit of the corresponding increasing sequence $\tilde{J}_n(\pi)$, which implies that $J(\pi)$ is lower semicontinuous on \mathfrak{A} .

5. Dynamics of Λ_μ . We begin by imposing rather stringent regularity conditions on the coefficients in (1.1) and by assuming that the initial distribution μ has a density $p_0 \in C_0^\infty(\mathbb{R}^N)$. Then Λ_t turns out to have a density $q(t, x)$ which obeys the Zakai stochastic partial differential equation, as in case of nonlinear filtering. However, it is more convenient to consider instead $p = q \exp(-Y_t \cdot h)$, which obeys the partial differential equation (5.4). Later in the section we drop the regularity assumptions, and obtain the same equation in a weak form.

The regular case. We fix $(Y, U) \in \Omega^2$, and for the present assume that U is continuous on $[0, T]$. We also assume for the present that σ, b^0, b^1, h are of class $C_b^\infty(\mathbb{R}^N)$ for Y fixed. Given $t > 0$ and $f \in C_0(\mathbb{R}^N)$, consider the following “backward” partial differential equation

$$(5.1) \quad \frac{dv}{ds} + \check{L}_s v + e(s)v = 0, \quad 0 \leq s \leq t, \quad v(t) = f \exp(Y_t \cdot h),$$

where we have written $v(s), e(s)$ for $v(s, \cdot), e(s, \cdot)$ and \check{L}_s is defined by (3.5). The Cauchy problem (5.1) has the probabilistic solution

$$(5.2) \quad v(s, x) = \check{E}_{sx}^{Y, U} \left[f(X_t) \exp(Y_t \cdot h(X_t)) \exp \int_s^t e(\theta, X_\theta) d\theta \right],$$

where $\check{P}_{sx}^{Y, U}$ is the distribution measure of (\check{W}_s, X_t) satisfying $dX_t = \check{b} dt + \sigma d\check{W}_t$, $s \leq t \leq T$, with $X_s = x$ (in particular, $\check{P}_{0x}^{Y, U} = \check{P}_x^{Y, U}$) already defined in § 3. By (3.8)

$$(5.3) \quad v(0) = v_0^{Y, U}.$$

Under our regularity conditions, $v(s) \in C^\infty(\mathbb{R}^N)$ for $0 \leq s \leq t$. This follows from the smooth dependence on the initial state x of pathwise solutions \bar{X} to $d\bar{X}_t = \bar{b} dt + \sigma d\bar{W}_t$, $\bar{X}_s = x$, with \bar{W}_t a fixed Wiener process on some $(\bar{\Omega}, \{\bar{\mathcal{F}}_t\}, \bar{P})$. By essentially the same proof as [12, p. 74], dv/ds is continuous and (5.1) holds. Moreover, each partial derivative of any order of v in the variables x_1, \dots, x_n tends to 0 exponentially as $|x| \rightarrow \infty$. For instance, by replacing X by \bar{X} and $\check{E}_{sx}^{Y, U}$ by $\bar{E} = E_{\bar{P}}$ in (5.2), and differentiating with respect to x_i , we get an estimate

$$|v_{x_i}(s, x)| \leq C \max_{s \leq \tau \leq t} \bar{E}(\chi_f |\xi_i(\tau)|),$$

with $\xi_i = \partial \bar{X} / \partial x_i$ and χ_f the indicator function of the event $\bar{X}_t \in \text{spt } f$. By [12, p. 61], $\bar{E}|\xi_i(\tau)|^p$ is bounded (independent of τ and x) for each $p > 0$. By taking $p = 2$ and using Cauchy-Schwarz, we get

$$|v_{x_i}(s, x)| \leq C_1 [\bar{P}(\bar{X}_t \in \text{spt } f)]^{1/2}.$$

Since $\bar{P}(\bar{X}_t \in B) = \check{P}_{sx}^{Y, U}(X_t \in B)$, the proof of Lemma 3.1b then shows that $v_{x_i}(s, x) \rightarrow 0$ exponentially as $|x| \rightarrow \infty$. Similarly, higher order derivatives of v tend to 0 exponentially as $|x| \rightarrow \infty$, using the fact that partial derivatives of \bar{X} of all orders with respect to x_1, \dots, x_n have bounded expectations [12, p. 61].

Let us also consider the following initial value problem for the equation adjoint to (5.1):

$$(5.4) \quad \frac{dp}{dt} = \check{L}_t^* p + e(t)p, \quad t \geq 0, \quad p(0) = p_0,$$

where $p_0 \in C_0^\infty(\mathbb{R}^N)$. The time reversal $s = T - t$ changes (5.4) into a problem of the same form as (5.1), but with \check{L}_s replaced by another degenerate parabolic operator L'_s and $e(s)$ by another $e'(s)$. Therefore, (5.4) has a unique solution with $p(t) \in C^\infty(\mathbb{R}^N)$ and with all partial derivatives of any order in x_1, \dots, x_N tending to 0 exponentially as $|x| \rightarrow \infty$.

Let us write (\cdot, \cdot) for scalar product in $L^2(\mathbb{R}^N)$. Integrations by parts imply $(v(t), p(t)) = \text{constant}$. In particular,

$$(v(t), p(t)) = (v_0^{Y,U}, p_0).$$

If p_0 is the density of μ , then we have from (3.9), since $v(t) = f \exp(Y_t \cdot h)$,

$$(5.5) \quad \int_{\mathbb{R}^N} p(t, x) \exp(Y_t \cdot h(x)) f(x) dx = \langle f, \Lambda_t^{Y,U} \rangle.$$

Let

$$(5.6) \quad q(t, x) = p(t, x) \exp(Y_t \cdot h(x))$$

(of course, $q = q^{Y,U}$ depends on the observation and control trajectories). Then (5.5) implies that $q(t)$ is the density of the unnormalized conditional distribution measure $\Lambda_t = \Lambda_t^{Y,U}$, under the above regularity assumptions. The partial differential equation (5.4) determines the dynamics of $p(t)$, hence, also of $q(t)$.

Equation (5.4) is a linear partial differential equation in which the processes Y, U enter parametrically. In contrast, the Zakai equation for $q(t)$, see (5.8) below, is a stochastic partial differential equation driven by the Y process. The technique of replacing the Zakai equation by (5.4) is analogous to the technique of Doss [8] and Sussmann [21] for reducing certain finite dimensional Ito-sense stochastic differential equations to ordinary differential equations depending parametrically on a Wiener process. The same idea has been used in nonlinear filtering by Liptser–Shiryaev [15], Clark [6] and others. See Davis [7].

The general case. Let us return to the assumptions (A1)–(A3) on σ, b^0, b^1, h . We consider fixed $(Y, U) \in \Omega^2$ and any distribution μ for X_0 . Let us rewrite (5.4) in a weak form. Define the measure $\tilde{\Lambda}_t$ by

$$(5.7) \quad \langle g, \tilde{\Lambda}_t \rangle = \langle g \exp(-Y_t \cdot h), \Lambda_t \rangle, \quad g \in C_b(\mathbb{R}^N).$$

In the regular case, $\tilde{\Lambda}_t$ has density $p(t)$. By multiplying (5.4) by $g \in C_0^\infty(\mathbb{R}^N)$ and integrating by parts, we get

$$(5.4') \quad \frac{d}{dt} \langle g, \tilde{\Lambda}_t \rangle = \langle \check{L}_t g, \tilde{\Lambda}_t \rangle + \langle e(t)g, \tilde{\Lambda}_t \rangle, \quad g \in C_0^\infty(\mathbb{R}^N).$$

This is the weak form of (5.4). The initial data are now $\tilde{\Lambda}_0 = \mu$.

THEOREM 5.1. *Equation (5.4') holds, for any $(Y, U) \in \Omega^2$, any $g \in C_b^2(\mathbb{R}^N)$ and initial distribution μ for X_0 .*

Proof. For $g \in C_0^\infty(\mathbb{R}^N)$, (5.4') holds in the regular case. For fixed Y , take $\sigma_n, b_n^0, b_n^1, h_n$ of class $C_b^\infty(\mathbb{R}^N)$, uniformly bounded and such that if Ψ_n denotes any of

$\sigma_n, b_n^0, b_n^1, h_n, h_{nx}, h_{nx_i}, h_{nx_i x_j}$, then Ψ_n is uniformly bounded and tends uniformly on compact sets as $n \rightarrow \infty$ to $\sigma, b^0, \dots, h_{x_i x_j}$. In addition, the partial derivatives $\sigma_{nx_i}, b_{nx_i}^0, b_{nx_i}^1$ are uniformly bounded. Moreover, let U_n tend to U almost everywhere on $[0, T]$, $U_{n_i} \in \mathcal{U}$, and μ_n tend weakly to μ , where U_n is continuous on $[0, T]$ and μ_n has density $p_{n0} \in C_0^\infty(\mathbb{R}^N)$. Let $\check{P}_{nx} = \check{P}_{nx}^{Y, U_n}$, where the subscript n means that σ, b^l, h are replaced by $\sigma_n, b_n^l, h_n, l = 0, 1$. Lemma A.1 implies that $\check{P}_{nx_n} \rightarrow \check{P}_x^{Y, U}$ if $x_n \rightarrow x$ as $n \rightarrow \infty$. Let $f \in C_0(\mathbb{R}^N)$. The same proof as for Lemma 3.1(a) implies that

$$v_{0n}(x) = \check{E}_{xn} \left(f(X_t) \exp(Y_t \cdot h_n(X_t)) \exp \int_0^t e_n(s, X_s) ds \right)$$

tends uniformly on any compact set to $v_0^{Y, U}(x)$. Let Λ_{nt} be the corresponding unnormalized conditional distribution, with $\Lambda_{n0} = \mu_n$. By (3.9),

$$\langle f, \Lambda_{nt} \rangle = \langle v_{0n}, \mu_n \rangle, \quad \langle f, \Lambda_t \rangle = \langle v_0, \mu \rangle,$$

where $v_0 = v_0^{Y, U}$. Then $\langle v_{0n}, \mu_n \rangle \rightarrow \langle v_0, \mu \rangle$. Since this is true for every $f \in C_0(\mathbb{R}^N)$, $\Lambda_{nt} \rightarrow \Lambda_t$ vaguely as $n \rightarrow \infty$. Also $\|\Lambda_{nt}\|$ is bounded. From (5.7), $\tilde{\Lambda}_{nt} \rightarrow \tilde{\Lambda}_t$ vaguely with $\|\tilde{\Lambda}_{nt}\|$ bounded. We rewrite (5.4') in the regular case in integrated form:

$$\langle g, \tilde{\Lambda}_n \rangle = \langle g, \mu_n \rangle + \int_0^t \langle \check{L}_{sn} g, \tilde{\Lambda}_{sn} \rangle ds + \int_0^t \langle e_n(s)g, \tilde{\Lambda}_{sn} \rangle ds.$$

For each $g \in C_0^\infty(\mathbb{R}^N)$, $\check{L}_{sn} g, e_n(s)g$ are uniformly bounded and tend to $\check{L}_s g, e(s)g$ uniformly on \mathbb{R}^N , for almost all $s \in [0, T]$. By passing to the limit, we get (5.4'), when $g \in C_0^\infty(\mathbb{R}^N)$. Finally, we approximate $g \in C_b^2(\mathbb{R}^N)$ by $g_n \in C_0^\infty(\mathbb{R}^N)$ such that $g_n, \check{L}_s g_n$ are uniformly bounded and tend to $g, \check{L}_s g$ as $n \rightarrow \infty$, uniformly on compact subsets of \mathbb{R}^N . By passing to the limit in (5.4'), we get Theorem 5.1.

We do not have a uniqueness result for (5.4'), in contrast with the nondegenerate case to be considered in § 7. Moreover, in § 7 we will be able to use results from the theory of parabolic PDE concerning the continuous dependence of solutions on the coefficients to get a stronger existence theorem for an optimal stochastic control.

The Zakai equation. The unnormalized conditional distribution Λ_t satisfies the following (Zakai) equation, written in a weak form. Recall that Y is a $\check{P}_\pi, \{\mathcal{G}_t^2\}$ -Brownian motion for every admissible control π .

THEOREM 5.2. For every $f \in C_b^2(\mathbb{R}^N)$

$$(5.8) \quad d\langle f, \Lambda_t \rangle = \langle L_t f, \Lambda_t \rangle dt + \langle hf, \Lambda_t \rangle \cdot dY_t.$$

Proof. Let $\psi(t, x) = f(x) \exp(Y_t \cdot h(x))$. Then

$$\langle f, \Lambda_t \rangle = \langle \psi(t), \tilde{\Lambda}_t \rangle,$$

where, as before, we set $\psi(t) = \psi(t, \cdot)$. For fixed x , the Ito differential rule implies that

$$d\psi = \frac{1}{2} \psi |h|^2 dt + \psi h \cdot dY.$$

Given $t > 0$, we partition $[0, t]$ into m subintervals $[t_{j-1}, t_j]$ of length $m^{-1}t$. Then

$$\begin{aligned} \langle f, \Lambda_t \rangle - \langle f, \Lambda_0 \rangle &= \sum_{j=1}^m \langle \psi(t_j), \tilde{\Lambda}_{t_j} - \tilde{\Lambda}_{t_{j-1}} \rangle + \sum_{j=1}^m \langle \psi(t_j) - \psi(t_{j-1}), \tilde{\Lambda}_{t_{j-1}} \rangle \\ &= \sum_{j=1}^m \int_{t_{j-1}}^{t_j} [\langle \check{L}_s \psi(t_j) + e(s) \psi(t_j), \tilde{\Lambda}_s \rangle] ds \\ &\quad + \sum_{j=1}^m \int_{t_{j-1}}^{t_j} \frac{1}{2} \langle \psi(s) |h|^2, \tilde{\Lambda}_{t_{j-1}} \rangle ds + \sum_{j=1}^m \int_{t_{j-1}}^{t_j} \langle \psi(s) h, \tilde{\Lambda}_{t_{j-1}} \rangle \cdot dY_s. \end{aligned}$$

(To justify the exchange of stochastic and Lebesgue integrals see the Note below.)
For fixed Y, U ,

$$\begin{aligned} |\langle \psi(s)h, \tilde{\Lambda}_s \rangle - \langle \psi(s)h, \tilde{\Lambda}_{t_{j-1}} \rangle| &= \left| \int_{t_{j-1}}^s [\langle L_\theta(\psi(s)h), \tilde{\Lambda}_\theta \rangle + \langle e(\theta)\psi(s)h, \tilde{\Lambda}_\theta \rangle d\theta] \right| \\ &\leq c(s - t_{j-1}) \leq cm^{-1}T, \end{aligned}$$

where c depends on Y, U . Hence, as $m \rightarrow \infty$, the last term tends in \tilde{P}_π -probability to $\int_0^t \langle \psi(s)h, \tilde{\Lambda}_s \rangle \cdot dY_s$. By using a similar estimate for the integrand in the middle term and elementary estimates for the first term, we get

$$\langle f, \Lambda_t \rangle - \langle f, \Lambda_0 \rangle = \int_0^t \langle \check{L}_s \psi(s) + e(s)\psi(s) + \frac{1}{2}\psi(s)|h|^2, \tilde{\Lambda}_s \rangle ds + \int_0^t \langle \psi(s)h, \tilde{\Lambda}_s \rangle \cdot dY_s.$$

A straightforward calculation, using (3.2), (3.3), (3.5), gives

$$\exp(Y_s \cdot h)L_s f = \check{L}_s \psi(s) + e(s)\psi(s) + \frac{1}{2}\psi(s)|h|^2.$$

Moreover, from (5.7),

$$\begin{aligned} \langle \exp(Y_s \cdot h)L_s f, \tilde{\Lambda}_s \rangle &= \langle L_s f, \Lambda_s \rangle, \\ \langle \psi(s)h, \tilde{\Lambda}_s \rangle &= \langle \exp(Y_s \cdot h)hf, \tilde{\Lambda}_s \rangle = \langle hf, \Lambda_s \rangle. \end{aligned}$$

Therefore

$$\langle f, \Lambda_t \rangle - \langle f, \Lambda_0 \rangle = \int_0^t \langle L_s f, \Lambda_s \rangle ds + \int_0^t \langle hf, \Lambda_s \rangle \cdot dY_s.$$

This is the integrated form of (5.8) and proves Theorem 5.2.

Note. In the proof we have used

$$(5.9) \quad \int_{t_{j-1}}^{t_j} \langle \psi(s)h, \tilde{\Lambda} \rangle \cdot dY_s = \langle \zeta, \tilde{\Lambda} \rangle,$$

where for brevity we now write $\tilde{\Lambda}_{t_{j-1}} = \tilde{\Lambda}$ and where (pointwise on \mathbb{R}^N)

$$(*) \quad \zeta = \int_{t_{j-1}}^{t_j} \psi(s)h \cdot dY_s = \psi(t_j) - \psi(t_{j-1}) - \frac{1}{2} \int_{t_{j-1}}^{t_j} \psi(s)|h|^2 ds.$$

The functions $\zeta, \psi(s)h$ are bounded and uniformly continuous on \mathbb{R}^N . The bounds and moduli of continuity depend on f and $\|Y\|$ but not on s . For $n = 1, 2, \dots$, partition $B_n = \{|x| \leq n\}$ into Borel sets $A_1^n, \dots, A_{m_n}^n$ of diameter $< n^{-1}$ and choose $x_i^n \in A_i^n$. Then

$$(5.10) \quad \int_{t_{j-1}}^{t_j} \left[\sum_i \psi(s, x_i^n)h(x_i^n)\tilde{\Lambda}(A_i^n) \right] \cdot dY_s = \sum_i \zeta(x_i^n)\tilde{\Lambda}(A_i^n).$$

For each (Y, U) , the right side tends to $\langle \zeta, \tilde{\Lambda} \rangle$ as $n \rightarrow \infty$ using $(*)$ and dominated convergence. The sum in brackets on the left side tends to $\langle \psi(s)h, \tilde{\Lambda} \rangle$ uniformly with respect to s . Hence, the stochastic integral converges in probability to the left side of (5.9) as $n \rightarrow \infty$ [12, p. 11, IV]. This proves (5.9).

THEOREM 5.3. For $K = 1, 2, \dots$, and any $j \geq 0$,

$$(5.11) \quad \tilde{E}_\pi(\langle (1+|x|^2)^{j/2}, \Lambda_t \rangle^K) \leq C\langle (1+|x|^2)^{j/2}, \mu \rangle^K,$$

where C depends on K, j and T (but not on $\pi \in \mathfrak{A}$).

Proof. For $0 < \alpha < 1$, let

$$f_\alpha(x) = (1+|x|^2)^{j/2} \exp[-\alpha(1+|x|^2)^{1/2}].$$

An easy calculation shows that $f_\alpha \in C_b^2(\mathbb{R}^N)$ and $|L_s f_\alpha| \leq C_1 f_\alpha$ for suitable C_1 depending on j . The Zakai equation (5.8) and Ito differential rule imply

$$\begin{aligned} d\langle f_\alpha, \Lambda_t \rangle^K &= [K\langle f_\alpha, \Lambda_t \rangle^{K-1} \langle L_t f_\alpha, \Lambda_t \rangle + K(K-1)\langle f_\alpha, \Lambda_t \rangle^{K-2} |\langle h f_\alpha, \Lambda_t \rangle|^2] dt \\ &\quad + K\langle f_\alpha, \Lambda_t \rangle^{K-1} \langle h f_\alpha, \Lambda_t \rangle \cdot dY_t. \end{aligned}$$

For $a > 0$, let $\tau_a = \inf \{t: \|\Lambda_t\| \geq a\}$. From (3.6) with $f = 1$, $\|\Lambda_t\| = \langle 1, \Lambda_t \rangle$ is continuous in t and $\{\mathcal{G}_t^2\}$ -adapted. Hence, τ_a is a stopping time. Let χ_a be the indicator function of the set $\{s \leq \tau_a\}$. Then

$$\begin{aligned} \mathring{E}_\pi \langle f_\alpha, \Lambda_{t \wedge \tau_a} \rangle^K &= \langle f_\alpha, \mu \rangle^K + K \mathring{E}_\pi \int_0^t \chi_a \langle f_\alpha, \Lambda_s \rangle^{K-1} \langle L_s f_\alpha, \Lambda_s \rangle ds \\ &\quad + K(K-1) \mathring{E}_\pi \int_0^t \chi_a \langle f_\alpha, \Lambda_s \rangle^{K-2} |\langle h f_\alpha, \Lambda_s \rangle|^2 ds. \end{aligned}$$

We have, since $f_\alpha > 0$ and $|L_s f_\alpha| \leq C_1 f_\alpha$,

$$|\langle L_s f_\alpha, \Lambda_s \rangle| \leq C_1 \langle f_\alpha, \Lambda_s \rangle, \quad |\langle h f_\alpha, \Lambda_s \rangle| \leq \|h\| \langle f_\alpha, \Lambda_s \rangle,$$

$$\mathring{E}_\pi \langle f_\alpha, \Lambda_{t \wedge \tau_a} \rangle^K \leq \langle f_\alpha, \mu \rangle^K + (KC_1 + K(K-1)\|h\|^2) \mathring{E}_\pi \int_0^t \chi_a \langle f_\alpha, \Lambda_s \rangle^K ds.$$

However, $\chi_a \langle f_\alpha, \Lambda_s \rangle \leq \langle f_\alpha, \Lambda_{s \wedge \tau_a} \rangle$. Gronwall's inequality then implies

$$\begin{aligned} \mathring{E}_\pi \langle f_\alpha, \Lambda_{t \wedge \tau_a} \rangle^K &\leq C \langle f_\alpha, \mu \rangle^K, \\ C &= \exp [(KC_1 + K(K-1)\|h\|^2)t]. \end{aligned}$$

We let $a \rightarrow \infty$ and then $\alpha \rightarrow 0$ to obtain (5.11).

Remark. The operator $\check{L}_s + e$ in the backward partial differential equation (5.1) can be rewritten as

$$(5.12) \quad (\check{L}_s + e)f = \exp(Y_s \cdot h)(L_s - \tfrac{1}{2}|h|^2) \exp(-Y_s \cdot h)f.$$

Let $S_{s,t} = S_{s,t}^{Y,U}$ denote the two-parameter linear semigroup on $C_b(\mathbb{R}^N)$ generated by $\check{L}_s + e$. Then

$$\begin{aligned} v_0^{Y,U}(x) &= S_{0,t}^{Y,U}[\exp(Y_t \cdot h)f], \\ \langle f, \Lambda_t \rangle &= \langle S_{0,t}^{Y,U}[\exp(Y_t \cdot h)f], \mu \rangle, \end{aligned}$$

as in (3.9). See Davis [7] for the corresponding pathwise nonlinear filtering formulas, and Davis [7], Mitter [16], for interesting interpretations in terms of multiplicative functionals of gauge, Feynman-Kac and Girsanov types.

6. Strict-sense admissible controls. We recall the notations of § 2.

DEFINITION. We say that $\pi \in \mathfrak{A}$ is a *strict-sense admissible control* if there exists $\underline{u}: \Omega_2 \rightarrow \Omega_3$ such that \underline{u} is $(\mathcal{F}_T(Y), \mathcal{F}_T(U))$ measurable, and for every \mathcal{G}_T^2 -measurable $\psi \geq 0$,

$$\int_{\Omega^2} \psi(Y, U) d\pi = \int_{\Omega_2} \psi(Y, \underline{u}(Y)) dw,$$

where w is Wiener measure on $(\Omega_2, \mathcal{F}_T(Y))$.

For any $\pi \in \mathfrak{A}$

$$\pi(dY, dU) = \pi^Y(dU)w(dY),$$

where π^Y is a regular conditional distribution for π . Strict-sense admissible controls are those such that $\pi^Y = \delta_{\underline{u}(Y)}$, w -a.s., where δ_u = Dirac measure on $(\Omega_3, \mathcal{F}_T(U))$

concentrated at u . By admitting, in § 2, controls $\pi \in \mathfrak{A}$ which are not strict-sense, we are in effect allowing the choice of U_t to depend on auxiliary randomizations. Let

$$\mathfrak{A}^s = \{\text{strict-sense admissible } \pi\}.$$

Corresponding to $\pi \in \mathfrak{A}^s$ there is a causal functional γ such that $U_t = \gamma(t, Y)$ Lebesgue \times almost everywhere [22]. Causal is in the sense that $Y_s = Y'_s$ for $0 \leq s \leq t$ implies $g(t, Y) = \gamma(t, Y')$ for $Y, Y' \in C[(0, T]; \mathbb{R}^M)$. (We do not use this result in this paper.) It can be shown that \mathfrak{A}^s is dense in \mathfrak{A} . We shall not prove this here. However, we shall show that the infimum of $J(\pi)$ on \mathfrak{A}^s is the same as on \mathfrak{A} (Theorem 6.1). For this purpose we consider approximations by piecewise constant controls.

For $m = 1, 2, \dots$, let us partition $[0, T]$ into m equal subintervals $[t_{j-1}, t_j]$, $t_j = j\Delta$, $\Delta = m^{-1}T$. Let

$$\Omega_{3m} = \{U \in \Omega_3: U_t \text{ constant on } [t_{j-1}, t_j], j = 1, \dots, m\}.$$

On Ω_{3m} weak and strong convergence of a sequence are both equivalent to pointwise convergence on each subinterval $[t_{j-1}, t_j]$. Define $\Phi_m: \Omega_3 \rightarrow \Omega_{3m}$ by $\Phi_m(U) = U_m$, where

$$U_{mt} = \begin{cases} 0, & 0 \leq t \leq \Delta, \\ \Delta^{-1} \int_{t_{j-1}}^{t_j} U_s ds, & t_j \leq t \leq t_{j+1}. \end{cases}$$

As $m \rightarrow \infty$, $\Phi_m(U) \rightarrow U$ in L^2 -norm, for every $U \in \Omega_3$. Let $\Omega_m^2 = \Omega_2 \times \Omega_{3m}$ and

$$\mathfrak{A}_m = \{\pi \in \mathfrak{A}: \pi(\Omega_m^2) = 1\}.$$

If $\pi \in \mathfrak{A}_m$, $t \in [t_j, t_{j+1})$, then U_t is independent of the increments $Y_r - Y_s$ for $t_j \leq s \leq r$ under π .

We call $\psi(Y, U)$ strongly continuous on $\Omega^2 = \Omega_2 \times \Omega_3$ if ψ is continuous when Ω_3 has the L^2 -norm topology rather than the weak topology. We also denote by Φ_m the mapping from $\Omega^2 \rightarrow \Omega_m^2$, such that $(Y, U) \rightarrow (Y, \Phi_m(U))$.

LEMMA 6.1. *Let ψ be bounded and strongly continuous on Ω^2 . Let $\pi_m = \Phi_m \pi$. Then*

$$\lim_{m \rightarrow \infty} \int_{\Omega_m^2} \psi(Y, U) d\pi_m = \int_{\Omega^2} \psi(Y, U) d\pi.$$

Proof. By definition

$$\int_{\Omega_m^2} \psi(Y, U) d\pi_m = \int_{\Omega^2} \psi(Y, \Phi_m U) d\pi.$$

Since $\Phi_m(U) \rightarrow U$ strongly, the lemma follows from the dominated convergence theorem.

In particular, we may take in Lemma 6.1 any ψ bounded and continuous on Ω^2 , where Ω_3 has the weak topology. Thus:

COROLLARY 6.1. *As $m \rightarrow \infty$, $\Phi_m \pi \rightarrow \pi$, for every $\pi \in \mathfrak{A}$.*

Let

$$\mathfrak{A}_m^s = \mathfrak{A}_m \cap \mathfrak{A}^s.$$

LEMMA 6.2. *Let ψ be bounded on Ω^2 and continuous on any compact subset of Ω_m^2 . Then*

$$\inf_{\mathfrak{A}_m} \int_{\Omega^2} \psi d\pi = \inf_{\mathfrak{A}_m^s} \int_{\Omega^2} \psi d\pi.$$

We leave the proof of this lemma, which depends on standard but tedious arguments, to the Appendix.

In addition to (A1)–(A4) in § 2, we assume (A5') in § 4. We use the “separated” formula (4.2) for $J(\pi)$.

THEOREM 6.1. $\inf_{\mathfrak{A}} J(\pi) = \inf_{\mathfrak{A}^s} J(\pi)$.

Proof. Since $\mathfrak{A}^s \subset \mathfrak{A}$, we have \leq . Let $\rho \in C_0(\mathbb{R}^N)$ with $0 \leq \rho \leq 1$, $H \in C_b(\mathbb{R}^1)$, and

$$\psi(Y, U) = \int_0^T H[\langle \rho F(\cdot, U_t), \Lambda_t^{Y,U} \rangle] dt + H[\langle \rho G, \Lambda_T^{Y,U} \rangle],$$

$$\tilde{J}(\pi) = \int_{\Omega^2} \psi d\pi.$$

By Lemma 3.2, ψ satisfies the hypotheses of Lemma 6.1 and hence also Lemma 6.2. Hence, for every $\varepsilon > 0$ and $\pi \in \mathfrak{A}$ there exist m and $\pi_1 \in \mathfrak{A}_m^s$ such that

$$\tilde{J}(\pi_1) < \tilde{J}(\pi) + \varepsilon.$$

Therefore,

$$\inf_{\mathfrak{A}^s} \tilde{J}(\pi) = \inf_{\mathfrak{A}} \tilde{J}(\pi).$$

Now take ρ_n such that $\rho_n(x) = 1$ for $|x| \leq n$, $H_n(z) = \min(z, n)$, and the corresponding $\tilde{J}_n(\pi)$. To complete the proof it suffices to show that $\tilde{J}_n(\pi) \rightarrow J(\pi)$ uniformly on \mathfrak{A} as $n \rightarrow \infty$. For brevity, we write $\Lambda_t = \Lambda_t^{Y,U}$. We have from (A5'), § 4,

$$\begin{aligned} (*) \quad 0 &\leq \tilde{E}_\pi[F(\cdot, U_t), \Lambda_t] - H_n[\rho_n F(\cdot, U_t), \Lambda_t], \\ &\leq C \left[\int_{\Omega^2} \langle (1 - \rho_n)(1 + |x|^m), \Lambda_t \rangle d\pi + \int_{B_n} \langle 1 + |x|^m, \Lambda_t \rangle d\pi \right], \end{aligned}$$

where $B_n = \{\langle (1 + |x|^m), \Lambda_t \rangle > C^{-1}n\}$. Let $l > m$ as in (A5') and $p = m^{-1}l$. From Hölder's inequality

$$\begin{aligned} \langle (1 - \rho_n)(1 + |x|^m), \Lambda_t \rangle &\leq \left(\int_{|x| \geq n} 1 d\Lambda_t \right)^{1/p'} \left(\int_{\mathbb{R}^N} (1 + |x|^m)^p d\Lambda_t \right)^{1/p} \\ &\leq c_1 n^{-l/p'} \left(\int_{|x| \geq n} |x|^l d\Lambda_t \right)^{1/p'} \left(\int_{\mathbb{R}^N} (1 + |x|^l) d\Lambda_t \right)^{1/p} \\ &\leq c_1 n^{-l/p'} \left(\int_{\mathbb{R}^N} (1 + |x|^l) d\Lambda_t \right)^{1/p + 1/p'}. \end{aligned}$$

Since $p^{-1} + (p')^{-1} = 1$ and $(p')^{-1}l = l - m$,

$$\tilde{E}_\pi \langle (1 - \rho_n)(1 + |x|^m), \Lambda_t \rangle \leq c_1 n^{-(l-m)} \tilde{E}_\pi \langle 1 + |x|^l, \Lambda_t \rangle.$$

By Theorem 5.3 the expectation on the right side is finite. By Cauchy–Schwarz,

$$\int_{B_n} \langle 1 + |x|^m, \Lambda_t \rangle d\pi \leq \pi(B_n)^{1/2} [\tilde{E}_\pi \langle 1 + |x|^m, \Lambda_t \rangle^2]^{1/2}.$$

Moreover,

$$\pi(B_n) \leq C n^{-1} \tilde{E}_\pi \langle 1 + |x|^m, \Lambda_t \rangle.$$

By using Theorem 5.3 again, the right side of (*) is bounded above by $c_2 n^{-\beta}$, where $\beta = \min(\frac{1}{2}, l - m)$. A similar estimate holds if $F(\cdot, U_t)$ is replaced by $G(\cdot)$. We then

have, for all $\pi \in \mathfrak{A}$,

$$0 \leq J(\pi) - \tilde{J}_n(\pi) \leq c_2(T+1)n^{-\beta},$$

as required. This proves Theorem 6.1.

Extreme points of \mathfrak{A} . Under the hypotheses of the existence Theorem 4.1 or of Theorem 7.2 below, $J(\pi)$ is linear and lower semicontinuous on the compact, convex set \mathfrak{A} . Hence, $J(\pi)$ has a minimum at some extreme point of \mathfrak{A} . Let

$$\mathfrak{A}^e = \{\text{extreme points of } \mathfrak{A}\}.$$

It can be shown that $\mathfrak{A}^s \subset \mathfrak{A}^e$. However, the following counter-example, due essentially to Varadhan, shows that $\mathfrak{A}^s \neq \mathfrak{A}^e$.

An example of Cirelson [4] provides a bounded causal drift coefficient $\alpha(t, \eta)$, such that the stochastic differential equation

$$d\eta_t = \alpha(t, \eta) dt + dY,$$

with Y a Wiener process, $\eta_0 = Y_0 = 0$, has no strong solution. However, the Carmon-Martin-Girsanov formula gives a weak solution, uniquely determining the joint distribution measure π' of (Y, η) on $C([0, T]; \mathbb{R}^2)$. Let $\mathcal{U} = [-1, 1]$ and $U_t = \phi^{-1}(\eta_t)$, where $\phi(u) = (1-u^2)^{-1}u$, $-1 < u < 1$. Let $\Phi(Y, U) = (Y, \eta)$, $\eta_t = \phi(U_t)$. Then $\pi = \Phi^{-1}\pi'$ is in \mathfrak{A} , but not in \mathfrak{A}^s since no strong solution exists. In fact, $\pi \in \mathfrak{A}^e$. To see this, suppose that $\pi = \lambda\pi_1 + (1-\lambda)\pi_2$, $0 < \lambda < 1$, $\pi_i \in \mathfrak{A}$ for $i = 1, 2$. Let $\pi'_i = \Phi\pi_i$. Then, for $i = 1, 2$,

$$\pi'_i\left(\eta_t = \int_0^t \alpha(s, \eta) ds + Y_t, 0 \leq t \leq T\right) = 1.$$

Hence, $\pi'_i = \pi'$, $\pi_i = \pi$, $i = 1, 2$, which implies $\pi \in \mathfrak{A}^e$.

The following characterization [23] of \mathfrak{A}^e was pointed out to the authors by J.-M. Bismut: $\pi \in \mathfrak{A}^e$ if and only if every bounded $\pi, \{\mathcal{G}_t^2\}$ -martingale M_t has the form

$$M_t = c + \int_0^t f_s dY_s,$$

with c a constant and f_t some integrable predictable process.

7. The nondegenerate case. Let us now assume, instead of (A1) in § 2:

(A1') σ is a bounded, continuous $N \times N$ matrix-valued function on \mathbb{R}^{N+M} with bounded inverse. Moreover, $\partial\sigma/\partial x_i \in C_b(\mathbb{R}^{N+M})$ for $i = 1, \dots, N$.

We also assume:

(A6) The distribution μ of X_0 has a density $p_0 \in L^2(\mathbb{R}^N)$.

Let us show that, for fixed $(Y, U) \in \Omega^2$, the forward equation (5.4) is still correct, if suitably interpreted in the L^2 theory of parabolic partial differential equations.

Consider the Sobolev space

$$H^1 = \left\{ v \in L^2(\mathbb{R}^N) : \frac{\partial v}{\partial x_i} \in L^2(\mathbb{R}^N), i = 1, \dots, N \right\},$$

and $H^{-1} = (H^1)'$. Let \hat{L}_t be the bounded linear operator from H^1 to H^{-1} , such that for all $p, v \in H^1$

$$\begin{aligned} \langle \hat{L}_t p, v \rangle = & -\frac{1}{2} \sum_{i,j=1}^N \int_{\mathbb{R}^N} a_{ij} \frac{\partial p}{\partial x_i} \frac{\partial v}{\partial x_j} dx + \sum_{i=1}^N \int_{\mathbb{R}^N} \hat{a}_i p \frac{\partial v}{\partial x_i} dx \\ & + \frac{1}{2} Y_s \cdot \sum_{i,j=1}^N \int_{\mathbb{R}^N} a_{ij} \frac{\partial h}{\partial x_j} \frac{\partial p}{\partial x_i} v dx, \end{aligned}$$

where $\langle \cdot, \cdot \rangle$ here denotes pairing of H^1 and H^{-1} and

$$\hat{a}_i(s, x) = b_i(x, Y_s, U_s) - \frac{1}{2} \sum_{j=1}^N \frac{\partial a_{ij}}{\partial x_j}(x, Y_s) - \frac{1}{2} Y_s \cdot \sum_{j=1}^N a_{ij}(x, Y_s) \frac{\partial h}{\partial x_j}.$$

In the "regular case," integrations by parts show that (5.4) is equivalent to

$$(7.1) \quad \frac{dp}{dt} = \hat{L}_t p + \hat{e} p, \quad t \geq 0, \quad p(0) = p_0,$$

where

$$\begin{aligned} \hat{e}(s, x) &= \frac{1}{2} (a Y_s \cdot \nabla h, Y_s \cdot \nabla h) - \hat{b} \cdot (Y_s \cdot \nabla h) - \frac{1}{2} |h|^2, \\ \hat{b}_i(s, x) &= b_i(x, Y_s, U_s) - \frac{1}{2} \sum_{j=1}^N \frac{\partial a_{ij}}{\partial x_j}(x, Y_s). \end{aligned}$$

The initial value problem has, for fixed Y, U , a unique solution [1]

$$p \in L^2([0, T]; H^1) \cap C([0, T]; L^2(\mathbb{R}^N)).$$

THEOREM 7.1. $q(t) = p(t) \exp(Y_t \cdot h)$ is the density of the unnormalized conditional distribution Λ_t .

Proof. From (5.6), this is true in the regular case. Following the proof of Theorem 5.1, we make approximations $\sigma_n, b_n^0, b_n^1, h_n$, with the properties described there as well as $\partial \sigma_n / \partial x_i \rightarrow \partial \sigma / \partial x_i$ uniformly on compact sets and with $a_n = \sigma_n \sigma_n^* \cong \alpha I$ ($\alpha > 0$) for all n . U_n is continuous and tends to U strongly in $L^2([0, T]; \mathcal{U})$, while μ_n has density $p_{n0} \in C_0^\infty(\mathbb{R}^N)$ tending to p_0 strongly in $L^2(\mathbb{R}^N)$. The density $p_n(t)$ of the corresponding Λ_{nt} satisfies

$$\frac{dp_n}{dt} = \hat{L}_{nt} p_n + \hat{e}_n p_n, \quad p_n(0) = p_{n0},$$

where \hat{L}_{nt}, \hat{e}_n are obtained by replacing σ, \dots, U above by σ_n, \dots, U_n .

Rewrite $A_n p_n = \hat{L}_n p_n + \hat{e}_n p_n$, and $A p = \hat{L} p + \hat{e} p$. Then:

$$\frac{d}{dt}(p - p_n) = A_n(p - p_n) + g_n, \quad p(0) - p_n(0) = p_0 - p_{n0},$$

where $g_n = (A - A_n)p$.

It follows from the above hypotheses that there exists c , independent of n , such that for all $v \in H^1$:

$$\langle -A_n v, v \rangle + c \|v\|_{L^2(\mathbb{R}^N)}^2 \geq \frac{\alpha}{2} \|v\|_{H^1}^2.$$

Consequently, by standard PDE arguments (see [1]) there exist c' and c'' such that

$$\sup_{0 \leq t \leq T} |p(t) - p_n(t)|_{L^2(\mathbb{R}^N)}^2 \leq c' |p_0 - p_{n0}|_{L^2(\mathbb{R}^N)}^2 + c'' \|g_n\|_{L^2(0, T; H^{-1})}^2.$$

One easily checks that $g_n \rightarrow 0$ in $L^2(0, T; H^{-1})$.

Finally, $p_n(t) \rightarrow p(t)$ in $L^2(\mathbb{R}^N)$. Then

$$\lim_{n \rightarrow \infty} \langle f, \Lambda_{nt} \rangle = \lim_{n \rightarrow \infty} \int_{\mathbb{R}^N} f \exp(Y_t \cdot h) p_n dx = \int_{\mathbb{R}^N} f \exp(Y_t \cdot h) p dx,$$

for any $f \in C_0(\mathbb{R}^N)$. However, the proof of Theorem 5.1 showed that $\langle f, \Lambda_{nt} \rangle \rightarrow \langle f, \Lambda_t \rangle$.

Thus, $q = p \exp(Y_t \cdot h)$ is the density of Λ_t , which is Theorem 7.1.

Let us write $p = p^{Y,U}$ to emphasize the dependence on Y, U of the solution to the initial value problem (7.1). From (4.2) and Theorem 7.1 we can rewrite the criterion to be minimized as

$$(7.2) \quad J(\pi) = \int_{\Omega^2} \left[\int_0^T \int_{\mathbb{R}^N} F(x, U_t) p^{Y,U}(t, x) \exp(Y_t \cdot h(x)) dx dt + \int_{\mathbb{R}^N} G(x) p^{Y,U}(T, x) \exp(Y_T \cdot h(x)) dx \right] d\pi(Y, U).$$

Let us suppose:

(A5'') Condition (A5) in § 4 holds, and $F(x, \cdot)$ is convex on \mathcal{U} for all $x \in \mathbb{R}^N$.

THEOREM 7.2. *There exists $\pi^* \in \mathfrak{A}$ such that $J(\pi^*) \leq J(\pi)$ for all $\pi \in \mathfrak{A}$.*

Let us first prove two lemmas.

LEMMA 7.1. *For every $\rho \in C_0(\mathbb{R}^N)$, $\rho \geq 0$, and $(Y, U) \in \Omega^2$, the function $\psi(V)$ defined by*

$$\psi(V) = \int_0^T \int_{\mathbb{R}^N} \rho(x) F(x, V_t) p^{Y,U}(t, x) \exp[Y_t \cdot h(x)] dx dt$$

is lower semicontinuous on Ω_3 .

Proof. Since $\psi(V)$ is convex from (A5''), it suffices to show that it is continuous on $L^2(0, T; \mathcal{U})$ endowed with the strong topology.

Let $V^n \rightarrow V$ in $L^2(0, T; \mathcal{U})$ strongly. Let $V^{n'}$ be a subsequence such that $V_t^{n'}$ converges for almost all t . Then $\psi(V^n) \rightarrow \psi(V)$. Consequently, any convergent subsequence of $\{\psi(V^n)\}$ has $\psi(V)$ as its limit. But ψ is uniformly bounded on $L^2(0, T; \mathcal{U})$. It follows that $\psi(V^n) \rightarrow \psi(V)$.

LEMMA 7.2. *Let $(Y_n, U_n) \rightarrow (Y, U)$ in Ω^2 . Denote $p^n = p^{Y_n, U_n}$, $p = p^{Y, U}$. Then for every D bounded open subset of \mathbb{R}^N with smooth boundary,*

(a) $p^n(T) \rightarrow p(T)$ in $L^2(D)$ weakly.

(b) $p^n \rightarrow p$ in $L^2((0, T) \times D)$ strongly.

Proof. Equation (7.1) can be rewritten in the form:

$$(7.3) \quad \frac{dp}{dt} + A_0 p + U_t A_1 p = 0, \quad p(0) = 0,$$

where, for all $p, v \in H^1$,

$$\begin{aligned} \langle A_0 p, v \rangle &= \frac{1}{2} \sum_{i,j=1}^N \int_{\mathbb{R}^N} a_{ij} \frac{\partial p}{\partial x_i} \frac{\partial v}{\partial x_j} dx + \sum_{i=1}^N \alpha_i p \frac{\partial v}{\partial x_i} dx - \sum_{i=1}^N \int_{\mathbb{R}^N} \gamma_i \frac{\partial p}{\partial x_i} v dx + \int_{\mathbb{R}^N} \delta p v dx, \\ \langle A_1 p, v \rangle &= - \sum_{i=1}^N \int_{\mathbb{R}^N} \beta_i p \frac{\partial v}{\partial x_i} dx + \int_{\mathbb{R}^N} \theta p v dx, \end{aligned}$$

with

$$\begin{aligned} \alpha_i(t, x) &= -b_i^0(x, Y_t) + \frac{1}{2} \sum_{j=1}^N \frac{\partial a_{ij}}{\partial x_j}(x, Y_t) + \frac{1}{2} Y_t \cdot \sum_{j=1}^N a_{ij}(x, Y_t) \frac{\partial h(x)}{\partial x_j}, \\ \beta_i(t, x) &= b_i^1(x, Y_t), \\ \gamma_i(t, x) &= \frac{1}{2} Y_t \cdot \sum_{j=1}^N a_{ij}(x, Y_t) \frac{\partial h}{\partial x_j}(x), \end{aligned}$$

$$\delta(t, x) = -\frac{1}{2}(a Y_t \cdot \nabla h, Y_t \cdot \nabla h) + Y_t \cdot \sum_{i=1}^N \frac{\partial h}{\partial x_i}(x) \left[b_i^0(x, Y_t) - \frac{1}{2} \sum_{j=1}^N \frac{\partial a_{ij}}{\partial x_j}(x, Y_t) \right] - \frac{1}{2}|h|^2(x),$$

$$\theta(t, x) = Y_t \cdot \sum_{i=1}^N b_i^1(x, Y_t) \frac{\partial h}{\partial x_i}(x),$$

all these coefficients being continuous and bounded functions of (t, Y_t) .

It follows from standard arguments, after multiplication of (7.3) by p and making use of (A2'), that there exists a constant K (depending only on $\sup_t |Y_t|$) such that:

$$(7.4) \quad |p(t)|_{L^2(\mathbb{R}^N)}^2 + \frac{\alpha}{2} \int_0^t \|\nabla p(s)\|_{(L^2(\mathbb{R}^N))^N}^2 ds \leq |p_0|^2 + K \int_0^t |p(s)|_{L^2(\mathbb{R}^N)}^2 ds.$$

Let now $(Y_n, U_n) \rightarrow (Y, U)$ in Ω^2 . Then $\sup_t Y_t^n$ is uniformly bounded, and it follows from (7.4) and (7.3) that $(p^n, dp^n/dt, p^n(T))$ remains in a bounded subset of $L^2(0, T; H^1) \times L^2(0, T; H^{-1}) \times L^2(\mathbb{R}^N)$. We can then extract a subsequence, still denoted p^n , such that: $(p^n, dp^n/dt, p^n(T)) \rightarrow (\bar{p}, \bar{q}, \bar{r})$ weakly, and it is easy to check that $\bar{q} = d\bar{p}/dt$, $\bar{r} = \bar{p}(T)$. Then $\bar{p} \in L^2(0, T; H^1)$. If we still denote by p^n and \bar{p} the restriction of p^n and \bar{p} to $[0, T] \times D$, we have:

- (i) $p^n \rightarrow \bar{p}$ in $L^2(0, T; H^1(D))$ weakly,
- (ii) $dp^n/dt \rightarrow d\bar{p}/dt$ in $L^2(0, T; H^{-1}(D))$ weakly,
- (iii) $p^n(T) \rightarrow \bar{p}(T)$ in $L^2(D)$ weakly,

where D is open, bounded and with smooth boundary.

Since D is bounded, the injection from $H^1(D)$ into $L^2(D)$ is compact, and it follows from (i) and (ii) by a compactness Lemma [14] that

- (iv) $p^n \rightarrow \bar{p}$ in $L^2([0, T] \times D)$ strongly.

It remains to show that $\bar{p} = p$. Choose any $\phi \in C_0^\infty(\mathbb{R}^1)$, and $v \in C_0^\infty(D)$. Multiply (7.3)ⁿ by ϕv , and integrate by parts:

$$\begin{aligned} \phi(T)(p^n(T), v) + \int_0^T \phi(t) \langle p^n, (A_0^n)^* v \rangle dt + \int_0^T U_t^n \phi(t) \langle p^n, (A_1^n)^* v \rangle dt \\ = \phi(0)(p_0, v) + \int_0^T \frac{d\phi}{dt}(p^n, v) dt, \end{aligned}$$

where (\cdot, \cdot) denotes the inner product in $L^2(\mathbb{R}^N)$.

Now, $(A_0^n)^* v \rightarrow A_0^* v$ in $L^2(0, T; H^{-1}(D))$ strongly and $(A_1^n)^* v \rightarrow A_1^* v$ in $L^2([0, T] \times D)$ strongly. It follows from (i), (iii) and (iv) that we can take the limit in the above equality. Since D, v and ϕ are arbitrary, $\bar{p} = p$, the unique solution of (7.3).

Proof of Theorem 7.2. As in Theorem 4.1, it suffices to show that $J(\pi)$ is lower semicontinuous on \mathfrak{A} , and this will be true if for all $\rho \in C_0(\mathbb{R}^N)$, $\rho \geq 0$ and $H \in C_b(\mathbb{R}^1)$ monotone, the following functional is lower semicontinuous on \mathfrak{A} :

$$\begin{aligned} \tilde{J}(\pi) = \int_{\Omega^2} H \left(\int_0^T \int_{\mathbb{R}^N} \rho(x) F(x, U_t) \exp(Y_t \cdot h(x)) p^{Y, U}(t, x) dx dt \right. \\ \left. + \int_{\mathbb{R}^N} \rho(x) G(x) \exp(Y_T \cdot h(x)) p^{Y, U}(T, x) dx \right) d\pi(Y, U). \end{aligned}$$

A sufficient condition for J to be l.s.c. (lower semicontinuous) on \mathfrak{A} is that the integrand be l.s.c. on Ω^2 .

Since H is continuous and monotone, it suffices to show that the following functional is l.s.c. on Ω^2 :

$$\begin{aligned}\mathbb{H}(Y, U) &= \int_0^T \int_{\mathbb{R}^N} \rho(x) F(x, U_t) \exp(Y_t \cdot h(x)) p^{Y, U}(t, x) dx dt \\ &\quad + \int_{\mathbb{R}^N} \rho(x) G(x) \exp(Y_T \cdot h(x)) p^{Y, U}(T, x) dx.\end{aligned}$$

Now let (Y^n, U^n) be a sequence such that $(Y^n, U^n) \rightarrow (Y, U)$ in Ω^2 , and consider (with the notation of Lemma 7.2):

$$\begin{aligned}\mathbb{H}(Y^n, U^n) - \mathbb{H}(Y, U) &= \int_0^T \int_{\mathbb{R}^N} \rho[F(U^n) - F(U)] \exp(Y_t \cdot h) p dx dt \\ &\quad + \int_0^T \int_{\mathbb{R}^N} \rho F(U^n) [\exp(Y_t^n \cdot h) p^n - \exp(Y_t \cdot h) p] dx dt \\ &\quad + \int_{\mathbb{R}^N} \rho G[\exp(Y_T^n \cdot h) p^n(T) - \exp(Y_T \cdot h) p(T)] dx.\end{aligned}$$

When $n \rightarrow \infty$, it follows from Lemma 7.1 that \liminf of the first term in the right-hand side is ≥ 0 . The two other terms tend to zero from Lemma 7.2. Then $\liminf_{n \rightarrow \infty} \mathbb{H}(Y^n, U^n) \geq \mathbb{H}(Y, U)$.

Appendix. In this Appendix we prove two results used in the paper. The first result concerns the continuous dependence on the coefficients and initial state of solutions to martingale problems associated with stochastic differential equations of the form

$$\begin{aligned}\text{(A.1)} \quad dX_t &= (\beta^0(t, X_t) + \beta^1(t, X_t) U_t) dt + \gamma(t, X_t) dW_t, \quad 0 < t < T, \\ X_0 &= x.\end{aligned}$$

Let us write for brevity $X'_t = (W_t, X_t)$, and consider the “canonical” sample space $\Omega^1, \{\mathcal{G}_t^1\}$ in the notation of § 2. For $f \in C_0^2(\mathbb{R}^{D+N})$, let

$$\begin{aligned}\text{(A.2)} \quad M_f(t) &= f(X'_t) - f(X'_0) - \int_0^t L'_s f(X'_s) ds, \\ \text{(A.3)} \quad L'_t f &= \frac{1}{2} \Delta_w f + \frac{1}{2} \sum_{i,j=1}^N \alpha_{ij}(t, x) f_{x_i x_j} + \sum_{i=1}^N \sum_{k=1}^D \gamma_{ik}(t, x) f_{x_i w_k} \\ &\quad + (\beta^0(t, x) + \beta^1(t, x) U_t) \cdot \nabla_x f,\end{aligned}$$

where $\Delta_w f(w, x)$ is the Laplacian with respect to w , ∇_x the gradient in x , and $\alpha = \gamma\gamma'$. The martingale problem is to find a probability measure P_x on $\{\mathcal{G}_T^1\}$ such that $P_x(X'_0 = (0, x)) = 1$ and $M_f(t)$ is a $P_x, \{\mathcal{G}_t^1\}$ -martingale for every $f \in C_0^2(\mathbb{R}^{D+N})$. See [20, Chapt. 6].

Let us call a function β of class \mathcal{L}_K if β is Borel measurable on $[0, T] \times \mathbb{R}^N$, $|\beta(t, x)| \leq K$, and $\beta(t, \cdot)$ is Lipschitz with constant K . If β^0, β^1, γ are of class \mathcal{L}_K and $U \in L^2([0, T]; \mathcal{U})$, then the Ito conditions hold in (A.1). This implies existence and pathwise uniqueness of solutions to (A.1), and consequently existence and uniqueness of the solution P_x to the martingale problem.

We write P_{nx} for the solution to the martingale problem if β^l, γ, U are replaced by β_n^l, γ_n, U_n , $n = 1, 2, \dots, l = 0, 1$.

LEMMA A.1. Assume that β_n^l, γ_n are of class \mathcal{L}_K and tend to β^l, γ as $n \rightarrow \infty$, uniformly on compact subsets of $[0, T] \times \mathbb{R}^N$, $l = 0, 1$. Moreover, assume that $U_n \rightarrow U$ weakly in $L^2([0, T]; \mathcal{U})$, $x_n \rightarrow x$ as $n \rightarrow \infty$. Then $P_{n x_n} \rightarrow P_x$.

Proof. The sequence $P_n = P_{n x_n}$ of probability measures is tight [20, Chapt. 6.1]. Hence, any subsequence has a further subsequence tending to a limit P_0 . It suffices to show that P_0 is a solution to the martingale problem. Uniqueness then implies $P_0 = P_x$. Clearly, $P_0(X'_0 = (0, x)) = 1$. Let us write M_{nf}, L'_{nt} in (A.2), (A.3) when β^l, γ, U are replaced by β_n^l, γ_n, U_n . Let us show that for fixed $f \in C_0^2(\mathbb{R}^{D+N})$ and compact $\Gamma \subset \Omega^1$,

$$(A.4) \quad \lim_{n \rightarrow \infty} \int_0^t L'_{ns} f(X'_s) ds = \int_0^t L'_s f(X'_s) ds$$

uniformly on $[0, T] \times \Gamma$. Since $M_{nf}(t)$ is a $P_n, \{\mathcal{G}_t^1\}$ martingale and $P_n \rightarrow P_0$ (n in a subsequence), (A.4) will imply that $M_f(t)$ is a $P_0, \{\mathcal{G}_t^1\}$ martingale. Now

$$L'_{ns} f - L'_s f = (\beta_n^1 - \beta^1) U_{ns} \cdot \nabla_x f + \beta^1 (U_{ns} - U_s) \nabla_x f + \theta_n(s, x),$$

where $\theta_n \rightarrow 0$ uniformly on compact subsets of $[0, T] \times \mathbb{R}^N$. Since U_{ns} is bounded (see (A₄), §2) and $\beta_n^1 \rightarrow \beta^1$ uniformly on compact sets,

$$\begin{aligned} \lim_{n \rightarrow \infty} \int_0^t [\beta_n^1(s, X_s) - \beta^1(s, X_s)] U_{ns} \cdot \nabla_x f(X'_s) ds &= 0, \\ \lim_{n \rightarrow \infty} \int_0^t \theta_n(s, X_s) ds &= 0 \end{aligned}$$

uniformly for $0 \leq t \leq T, X' \in \Gamma$. To obtain (A.4) it remains to show that

$$(A.5) \quad \lim_{n \rightarrow \infty} \int_0^t \beta^1(s, X_s) (U_{ns} - U_s) \cdot \nabla f(X'_s) ds = 0$$

uniformly on $[0, T] \times \Gamma$. Now $\beta^1(s, \cdot)$ and ∇f are bounded and Lipschitz, with some constant K . Moreover, functions $X' \in \Gamma$ are uniformly bounded and equicontinuous. Therefore, given $\varepsilon > 0$, the integral in (A.5) can be approximated to within ε , uniformly with respect to $X' \in \Gamma$ and $n = 1, 2, \dots$, by a finite sum

$$(A.6) \quad \sum_{i=1}^m \int_{t_{i-1}}^{t_i} \beta^1(s, x_i) (U_{ns} - U_s) \cdot \nabla f(x'_i) ds,$$

where $0 = t_0 < t_1 < \dots < t_m = T$ and x'_1, \dots, x'_m are suitably chosen. (The t_i, x'_i depend on ε and X .) Since $U_n \rightarrow U$ weakly, (A.6) tends to 0 as $n \rightarrow \infty$. This proves Lemma A.1.

Note. In this paper we appeal to Lemma A.1 three times. In Lemma 2.1 we take

$$\begin{aligned} \gamma_n(t, x) &= \sigma(x, Y_{nt}), & \gamma(t, x) &= \sigma(t, Y_t), \\ \beta_n^l(t, x) &= b^l(x, Y_{nt}), & \beta^l(t, x) &= b^l(x, Y_t), \end{aligned}$$

where $\|Y_n - Y\| \rightarrow 0$ as $n \rightarrow \infty$ (sup norm). In that case, $P_x = P_x^{Y, U}$. In the proof of Lemma 3.1 we use instead of b^l the modified drift coefficient \check{b}^l corresponding to the change of probability measures (3.4). Then $P_x = \check{P}_x^{Y, U}$ and L_s is replaced by \check{L}_s in (3.5). Finally, we use Lemma A.1 in the proof of Theorem 5.1 as indicated there.

In § 6 we postponed the proof of Lemma 6.2.

Proof of Lemma 6.2. Since $\mathfrak{A}_m^s \subset \mathfrak{A}_m$, it suffices to show that, for every $\pi \in \mathfrak{A}_m$ and $\varepsilon > 0$, there exists $\pi_1 \in \mathfrak{A}_m^s$ such that

$$(\#) \quad \int_{\Omega^2} \psi d\pi_1 < \int_{\Omega^2} \psi d\pi + \varepsilon.$$

Let us fix Δ and consider different $T = m\Delta$, $m = 1, 2, \dots$. We prove $(\#)$ by induction on m . For $m = 1$, each admissible control π , with U_t constant on $[0, \Delta]$ π -almost surely, corresponds to a product measure $\pi = w \times \alpha$, where w is Wiener measure on $\Omega_{21} = C([0, \Delta]; \mathbb{R}^M)$ and α is a probability measure on \mathcal{U} . Let u^* minimize $\int_{\Omega_{21}} \psi(Y, u) dw(Y)$ on \mathcal{U} . The control π_1 such that $U_t = u^*$, $0 \leq t < \Delta$ with probability 1 satisfies

$$\int_{\Omega_1^2} \psi d\pi_1 = \int_{\Omega_{21}} \psi(Y, u^*) dw(Y),$$

where $\Omega_1^2 = \Omega_{21} \times \Omega_3$. We then have

$$\int_{\Omega_1^2} \psi d\pi = \int_{\mathcal{U}} \int_{\Omega_{21}} \psi(Y, u) dw(Y) d\alpha(u) \geq \int_{\Omega_1^2} \psi d\pi_1$$

as required.

Now suppose that $(\#)$ has been proved when m is replaced by $m-1$ (i.e., T by $T-\Delta$). Let (\bar{Y}, \bar{U}) denote the restriction of (Y, U) to $[0, T-\Delta]$, and $\bar{\Omega}^2$ the space of such (\bar{Y}, \bar{U}) . Let $\bar{\pi}$ be the measure on $\bar{\Omega}^2$ induced from π by restriction. We write similarly $\bar{\Omega}_m^2, \bar{\Omega}_2, \bar{w}$ when T is replaced by $T-\Delta$. Let $U_t = U_m$ on $[T-\Delta, T]$, where $U_m \in \mathcal{U}$. Let $Y_{mt} = Y_t - Y_{t-T+\Delta}$ on $[T-\Delta, T]$, and w_m Wiener measure on $C_m = C([T-\Delta, T]; \mathbb{R}^M)$. We can identify Y with (\bar{Y}, Y_m) and piecewise constant U with (\bar{U}, U_m) . Let

$$\gamma(\bar{Y}, \bar{U}, u) = \int_{C_m} \psi(\bar{Y}, Y_m, \bar{U}, u) dw_m(Y_m),$$

$$\zeta(\bar{Y}, \bar{U}) = \min_{\mathcal{U}} \gamma(\bar{Y}, \bar{U}, u).$$

Since ψ is bounded and continuous on any compact subset of Ω_m^2 , ζ is bounded and continuous on any compact subset of $\bar{\Omega}_m^2$.

Consider any $\pi \in \mathfrak{A}_m$, with corresponding $\bar{\pi}$ determined by restriction. By induction (with ψ replaced by ζ) there exists $\bar{\pi}_1$ which is strict sense admissible, such that U_t is constant on $[t_{j-1}, t_j]$, $j = 1, \dots, m-1$, $\bar{\pi}_1$ -almost surely and

$$\int_{\bar{\Omega}^2} \zeta d\bar{\pi}_1 < \int_{\bar{\Omega}^2} \zeta d\bar{\pi} + \frac{\varepsilon}{3}.$$

We define $\phi: \bar{\Omega}^2 \rightarrow \mathcal{U}$ as follows. Let $K \subset \bar{\Omega}^2$ be compact with

$$\bar{\pi}(\bar{\Omega}^2 - K) + \bar{\pi}_1(\bar{\Omega}^2 - K) < \varepsilon(3\|\psi\|)^{-1},$$

where $\|\cdot\| = \sup$ norm. Choose a partition $K = K_1 \cup \dots \cup K_n$ with each $K_i \in \mathcal{G}_{T-\Delta}^2$ and $(\bar{Y}_i, \bar{U}_i) \in K_i$ such that

$$\zeta(\bar{Y}_i, \bar{U}_i) < \zeta(\bar{Y}, \bar{U}) + \frac{\varepsilon}{6}, \quad \gamma(\bar{Y}, \bar{U}, u) \leq \gamma(\bar{Y}_i, \bar{U}_i, u) + \frac{\varepsilon}{6}$$

for all $(\bar{Y}, \bar{U}) \in K_i$, $u \in \mathcal{U}$. Let $u_i \in \mathcal{U}$ minimize $\gamma(\bar{Y}_i, \bar{U}_i, u)$ on \mathcal{U} , $i = 1, \dots, n$. Let

$u_0 \in \mathcal{U}$ be arbitrary; and take

$$\phi(\bar{Y}, \bar{U}) = \begin{cases} u_i, & (\bar{Y}, \bar{U}) \in K_i, \\ u_0, & (\bar{Y}, \bar{U}) \in \bar{\Omega}^2 - K. \end{cases}$$

The control $\pi_1 \in \mathfrak{A}_m^s$ is defined by taking $U_m = \phi(\bar{Y}, \bar{U})\pi_1$ -almost surely, and $\bar{\pi}_1$ the restriction of π_1 . Then

$$\begin{aligned} \int_{\Omega^2} \psi d\pi_1 &= \int_{\bar{\Omega}^2} \gamma(\bar{Y}, \bar{U}, \phi(\bar{Y}, \bar{U})) d\pi_1 \\ &\cong \sum_{i=1}^n \int_{K_i} \zeta(\bar{Y}_i, \bar{U}_i) d\bar{\pi}_1 + \int_{\bar{\Omega}^2 - K} \gamma(\bar{Y}, \bar{U}, u_0) d\bar{\pi}_1 + \frac{\varepsilon}{6} \\ &\cong \int_{\bar{\Omega}^2} \zeta d\bar{\pi}_1 + \frac{2\varepsilon}{3} \cong \int_{\bar{\Omega}^2} \zeta d\bar{\pi} + \varepsilon. \end{aligned}$$

On the other hand,

$$\begin{aligned} \int_{\Omega^2} \psi d\pi &= \int_{\bar{\Omega}^2} \int_{\mathcal{U}} \gamma(\bar{Y}, \bar{U}, u) d\pi^{\bar{Y}, \bar{U}}(u) d\bar{\pi}(\bar{Y}, \bar{U}) \\ &\cong \int_{\bar{\Omega}^2} \zeta(\bar{Y}, \bar{U}) d\bar{\pi}(\bar{Y}, \bar{U}). \end{aligned}$$

This gives (#), and hence Lemma 6.2.

Acknowledgment. The authors wish to thank V. E. Benes, J. M. Bismut and S. R. S. Varadhan for helpful comments in connection with § 6.

REFERENCES

- [1] A. BENSOUSSAN AND J. L. LIONS, *Application des inéquations variationnelles en contrôle stochastique*, Dunod, Paris, 1978.
- [2] P. BILLINGSLEY, *Convergence of Probability Measures*, Paris, John Wiley, New York, 1968.
- [3] J.-M. BISMUT, *Partially observed diffusions and their control*, this Journal, this issue, pp. 302–309.
- [4] B. S. TSIREL'SON, *An example of a stochastic differential equation having no strong solution*, Theor. Prob. Appl., 20 (1975), pp. 416–418.
- [5] N. CHRISTOPEIT, *Existence of optimal stochastic controls under partial observation*, Z. Warsch. Verw. Gebiete, 51 (1980), pp. 201–213.
- [6] J. M. C. CLARK, *The design of robust approximations to the stochastic differential equations of nonlinear filtering*, in Communications Systems and Random Process Theory, J. Skwirzynski, ed., Sijthoff and Noordhoff, Groningen, 1978.
- [7] M. H. A. DAVIS, *Pathwise nonlinear filtering*, Stochastic Systems, M. Hazelwinkel, ed., NATO Advanced Study Institute Series, Reidel, Dordrecht.
- [8] H. DOSS, *Liens entre équations différentielles stochastiques et ordinaires*, Ann. Inst. H. Poincaré, 13 (1977), pp. 99–125.
- [9] W. H. FLEMING, *Measure-valued processes in the control of partially-observable stochastic systems*, Applied Math. Optim., 6 (1980), pp. 271–285.
- [10] ———, *Nonlinear semigroup for controlled partially observed diffusions*, this Journal, this issue, pp. 286–301.
- [11] W. H. FLEMING AND M. NISIO, *On the existence of optimal stochastic controls*, J. Math. Mech., 15 (1966), pp. 777–794.
- [12] I. I. GIKHMAN AND A. V. SKOROKHOD, *Stochastic Differential Equations*, Springer-Verlag, New York, 1972.
- [13] U. G. HAUSSMANN, *Existence of partially observable stochastic optimal controls*, Proc. 3rd Working Conference on Stochastic Differential Systems, Visegrad, Hungary, Lecture Notes in Control and Information Science, Springer, New York, to appear.

- [14] J. L. LIONS, *Quelques méthodes de résolution des problèmes aux limites nonlineaire*, Dunod, Paris, 1969.
- [15] R. S. LIPTSER AND A. N. SHIRYAEV, *Statistics of Random Processes I*, Springer-Verlag, New York, 1977 (transl. from Russian).
- [16] S. K. MITTER, *On the analogy between mathematical problems of nonlinear filtering and quantum physics*, LIDS Rep. P-1006, Massachusetts Institute of Technology, Cambridge, Recherche de Automatica, to appear.
- [17] E. PARDOUX, *Stochastic partial differential equations and filtering of diffusion processes*, Stochastics.
- [18] ———, *Nonlinear filtering, prediction and smoothing*, Stochastic Systems, M. Hazewinkel, ed., NATO Advanced Study Institute Series, Reidel, Dordrecht.
- [19] ———, *Equations du filtrage nonlinéaire, de la prédiction et du lissage*, Publ. Math. Appli. Marseille-Toulon, No. 80-6, Université de Provence, Stochastics, submitted.
- [20] D. W. STROOCK AND S. R. S. VARADHAN, *Multidimensional Diffusion Processes*, Springer-Verlag, New York, 1979.
- [21] H. J. SUSSMANN, *On the gap between deterministic and stochastic ordinary differential equations*, Ann. Prob., 6 (1978), pp. 19–41.
- [22] M. P. YERSHOV, *Nonanticipating solutions of stochastic equations*, Proc. 3rd Japan-USSR Symposium on Probability Theory, Lecture Notes in Mathematics 550, Springer-Verlag, New York, 1976.
- [23] M. YOR, *Sur l'étude des martingales continues extrémales*, Stochastics, 2 (1979), pp. 191–196.

NONLINEAR SEMIGROUP FOR CONTROLLED PARTIALLY OBSERVED DIFFUSIONS*

WENDELL H. FLEMING†

Abstract. In this paper a "separated" control problem associated with controlled, partially observed diffusion processes is considered. The state in the separated problem is an unnormalized conditional distribution measure. The corresponding Nisio nonlinear semigroup associated with the separated problem is found.

1. Introduction. In this paper we are concerned with stochastic control problems of the following kind. Let X_t denote the state of a process being controlled, Y_t the observation process and U_t the control process, $t \geq 0$. The state and observation processes are governed by stochastic differential equations

$$(1.1) \quad \begin{aligned} (a) \quad & dX_t = b(X_t, U_t) dt + \sigma(X_t) dW_t, \\ (b) \quad & dY_t = h(X_t) dt + d\tilde{W}_t. \end{aligned}$$

We shall assume that $b(x, u)$ is linear in u . See condition (A2) in § 3. X_t has values in N -dimensional \mathbb{R}^N , Y_t values in \mathbb{R}^M and U_t values in $\mathcal{U} \subset \mathbb{R}^L$. X_0 has given distribution μ , and $Y_0 = 0$. In (1.1), W and \tilde{W} are independent standard Wiener processes, with values in \mathbb{R}^D , \mathbb{R}^M respectively. The problem is to find an admissible control minimizing some criterion J .

For instance, we may take $J = EG(X_{t_1})$ for some fixed time $t_1 > 0$. In case of completely observed, controlled diffusions (with $Y_t = X_t$ rather than Y_t as in (1.1b)), the problem can be treated using dynamic programming. Let $V(x, t_1)$ denote the minimum of J , for initial data $X_0 = x$. Under suitable assumptions, $V(x, t)$ has continuous partial derivatives $\partial V / \partial t$, $\partial V / \partial x_i$, $\partial^2 V / \partial x_i \partial x_j$, $i, j = 1, \dots, N$, $x = (x_1, \dots, x_N)$. Among these assumptions is the condition that the symmetric matrix $a = \sigma\sigma'$ has a bounded inverse a^{-1} . The function V then satisfies the dynamic programming equation [4, Chapt. VI.6]

$$(1.2) \quad \frac{\partial V}{\partial t} = LV,$$

$$(1.3) \quad LV = \min_{u \in \mathcal{U}} \left[\frac{1}{2} \sum_{i,j=1}^N a_{ij}(x) \frac{\partial^2 V}{\partial x_i \partial x_j} + \sum_{i=1}^N b_i(x, u) \frac{\partial V}{\partial x_i} \right].$$

The assumptions that $a(x)$ has a bounded inverse can sometimes be weakened, by considering generalized solutions to the dynamic programming equation [4, p. 177] and more recently [6]–[8].

In [9] Nisio introduced another treatment which is valid under much less restrictive conditions. Let $\mathcal{S}_t G(x) = V(x, t)$. Then Nisio showed that \mathcal{S}_t is a nonlinear semigroup on the space $C_b(\mathbb{R}^N)$ of continuous bounded functions f on \mathbb{R}^N . Moreover, the operator L in (1.3) agrees with the generator of the semigroup \mathcal{S}_t on the space $C_b^2(\mathbb{R}^N)$ of those

* Received by the editors October 27, 1980, and in final form May 11, 1981. This research was supported in part by the U.S. Air Force Office of Scientific Research under grant AF-AFOSR 76-3063C, in part by the National Science Foundation under grant MCS-79-03554. In addition, part of the research was done during a visit by the author to the Centro de Investigacion y de Estudios Avanzados, IPN, Mexico.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912.

f such that $f, f_{x_i}, f_{x_i x_j}$ are in $C_b(\mathbb{R}^N)$ for $i, j = 1, \dots, N$. For another treatment of this nonlinear semigroup see [1, Chapt. IV.5.1].

In this paper we find a nonlinear semigroup \mathcal{T}_t associated with the partially observed control problem. In this case one should regard as the true “state” the conditional distribution of X_t given past data or some quantity equivalent to the conditional distribution. For technical reasons, it is more convenient to consider an unnormalized conditional distribution Λ_t for X_t . We have $\Lambda_t \in \mathcal{M}$, where \mathcal{M} is the space of finite measures on \mathbb{R}^N . The problem we consider is to control the measure-valued process Λ_t such that a criterion of the form $J = E\phi(\Lambda_t)$ is minimized. The dynamics of the Λ_t -process are governed by the Zakai equation, written in a weak form as (3.1) below.

If one writes $V(\mu, t_1)$ for the minimum of J , given initial data $\Lambda_0 = \mu$, then $V(\mu, t)$ formally satisfies a dynamic programming equation of the form

$$(1.4) \quad \frac{\partial V}{\partial t} = \mathcal{L}V,$$

where $\mathcal{L}V = \min_{u \in \mathcal{U}} \mathcal{L}^u V$ and \mathcal{L}^u is the generator of the linear semigroup \mathcal{T}_t^u associated for a constant control u with the process Λ_t (Λ_t is Markov for constant u). Equation (1.4) is called Mortensen’s equation. However, (1.4) has been treated rigorously only in very special cases.

Following Nisio, we write $V(\mu, t) = \mathcal{T}_t \phi(\mu)$. The purpose of this paper is to show that \mathcal{T}_t is a nonlinear semigroup, on a space $C(\mathcal{M})$, with $\mathcal{T}_t \phi$ continuous in t , and to describe the generator \mathcal{L} on a dense subspace of $C(\mathcal{M})$. We rely heavily on results from [3] (this issue, pp. 261–285). In particular, it was shown in [3] that Λ_t can be defined pathwise in such a way that Λ_t depends continuously on observation and control trajectories (Y, U) and on $\mu = \Lambda_0$. This and other results from [3] needed in this paper are summarized as 3.1–3.4 below.

For the case of a controlled Markov chain X_t subject to observations Y_t of the form (1.1b), a corresponding nonlinear semigroup was constructed by Davis [2].

2. The spaces $C_K(\mathcal{M}), C(\mathcal{M})$. Let $C_b(\mathbb{R}^N)$ denote the space of bounded, continuous f on \mathbb{R}^N and $C_0(\mathbb{R}^N)$ the space of continuous f with compact support. Let $C_b^k(\mathbb{R}^N), C_0^k(\mathbb{R}^N)$ be the spaces of f such that f together with all partial derivatives of orders $\leq k$ are in $C_b(\mathbb{R}^N), C_0(\mathbb{R}^N)$ respectively. Similarly, for \mathbb{R}^m -valued functions on \mathbb{R}^N we write $C_b^k(\mathbb{R}^N; \mathbb{R}^m), C_0^k(\mathbb{R}^N; \mathbb{R}^m)$.

Let $\mathcal{B}(\mathbb{R}^N)$ denote the Borel σ -algebra of \mathbb{R}^N , and

$$(2.1) \quad \mathcal{M} = \{\text{measures } \mu \geq 0 \text{ on } \mathcal{B}(\mathbb{R}^N); \mu(\mathbb{R}^N) < \infty\}.$$

We write

$$\langle f, \mu \rangle = \int_{\mathbb{R}^N} f(x) d\mu(x)$$

for the scalar product and

$$\|\mu\| = \langle 1, \mu \rangle = \mu(\mathbb{R}^N).$$

By convergence of sequences in \mathcal{M} we mean w^* -convergence: $\mu_n \rightarrow \mu$ if and only if $\langle f, \mu_n \rangle \rightarrow \langle f, \mu \rangle$ as $n \rightarrow \infty$ for every $f \in C_b(\mathbb{R}^N)$ such that $f(x) \rightarrow 0$ as $|x| \rightarrow \infty$.

We denote real-valued functions on \mathcal{M} by ϕ, ψ, \dots . For $K = 0, 1, 2, \dots$, let

$$(2.2) \quad \|\phi\|_K = \sup_{\mu \in \mathcal{M}} \frac{|\phi(\mu)|}{1 + \|\mu\|^K}.$$

By ϕ continuous on \mathcal{M} we mean, of course, continuity of ϕ under w^* -sequential convergence. Let

$$(2.3) \quad C_K(\mathcal{M}) = \{\phi \text{ continuous on } \mathcal{M} : \|\phi\|_K < \infty\}.$$

Then $\|\cdot\|_K$ is a norm on $C_K(\mathcal{M})$. Let

$$(2.4) \quad C(\mathcal{M}) = \bigcup_{K=0}^{\infty} C_K(\mathcal{M}).$$

For $r < \infty$, let

$$(2.5) \quad \mathcal{M}_r = \{\mu \in \mathcal{M} : \|\mu\| \leq r\}.$$

We give $C(\mathcal{M})$ the following metric:

$$(2.6) \quad d(\phi, \psi) = \sum_{l=1}^{\infty} 2^{-l} (\sup_{\mathcal{M}_l} |\phi(\mu) - \psi(\mu)| \wedge 1).$$

Thus d -convergence of ϕ_n to ϕ is equivalent to convergence of $\phi_n(\mu)$ to $\phi(\mu)$ uniformly on \mathcal{M}_r for every $r < \infty$. For each K , $\|\cdot\|_K$ is a lower semicontinuous function under d -convergence. Moreover, from (2.2), $\phi_n, \phi \in C_K(\mathcal{M})$ and $\|\phi_n - \phi\|_K \rightarrow 0$ imply $d(\phi_n, \phi) \rightarrow 0$ as $n \rightarrow \infty$.

Let

$$\tilde{\mathcal{M}} = \{\mu \geq 0 \text{ on } \mathcal{B}(\mathbb{R}^N) : \mu(B) < \infty \text{ for every compact } B\},$$

with the vague topology: $\mu_n \rightarrow \mu$ vaguely means $\langle f, \mu_n \rangle \rightarrow \langle f, \mu \rangle$ as $n \rightarrow \infty$ for every $f \in C_0(\mathbb{R}^N)$. $\tilde{\mathcal{M}}$ is a Polish space. In fact, one can choose a metric $\delta(\mu, \nu)$ for $\tilde{\mathcal{M}}$ of the form

$$(2.7) \quad \delta(\mu, \nu) = \sum_{m=1}^{\infty} 2^{-m} (|\langle f_m, \mu \rangle - \langle f_m, \nu \rangle| \wedge 1)$$

for a suitably chosen sequence $f_m \in C_0(\mathbb{R}^N)$.

For each $r < \infty$, \mathcal{M}_r is a compact subset of $\tilde{\mathcal{M}}$. For sequences in \mathcal{M}_r , vague convergence is equivalent to w^* -convergence. Moreover $\mu_n, \mu \in \mathcal{M}$ and $\mu_n \rightarrow \mu$ w^* imply $\|\mu_n\| \leq r$ for some r . Thus, we have:

LEMMA 2.1. ϕ is continuous on \mathcal{M} , under w^* -sequential convergence if and only if $\phi|_{\mathcal{M}_r}$ is vaguely continuous for every $r < \infty$.

This furnishes an alternate characterization of $C(\mathcal{M})$, in terms of the vague topology rather than in terms of w^* -sequential convergence.

A measure $\mu \in \mathcal{M}$ can be approximated by measures $\rho\mu$ with compact support, as follows. Let $\rho \in C_0(\mathbb{R}^N)$, $0 \leq \rho \leq 1$, and define $\rho\mu$ by $\langle f, \rho\mu \rangle = \langle \rho f, \mu \rangle$ for all $f \in C_b(\mathbb{R}^N)$. Define ϕ^ρ by

$$(2.8) \quad \phi^\rho(\mu) = \phi(\rho\mu), \quad \mu \in \mathcal{M}.$$

Then $\phi \in C_K(\mathcal{M})$ implies $\phi^\rho \in C_K(\mathcal{M})$ and $\|\phi^\rho\|_K \leq \|\phi\|_K$. We write $\mu|_B$ for the restriction of μ to a compact set B : $(\mu|_B)(A) = \mu(A \cap B)$ for all $A \in \mathcal{B}(\mathbb{R}^N)$. Let

$$(2.9) \quad C_K^0(\mathcal{M}) = \{\psi \in C_K(\mathcal{M}) : \text{there exists } B \text{ compact such that } \psi(\mu) = \psi(\mu|_B) \text{ for all } \mu \in \mathcal{M}\}.$$

In particular, $\phi^\rho \in C_K^0(\mathcal{M})$ if $\phi \in C_K(\mathcal{M})$, and ϕ^ρ is defined by (2.8).

LEMMA 2.2. For every $\phi \in C_K(\mathcal{M})$, there exists a sequence $\phi_n \in C_K^0(\mathcal{M})$ such that $\|\phi_n\|_K \leq \|\phi\|_K$ and $d(\phi_n, \phi) \rightarrow 0$ as $n \rightarrow \infty$.

Proof. Let $\rho_n \in C_0(\mathbb{R}^N)$, with $0 \leq \rho_n \leq 1$, $\rho_n(x) = 1$ for $|x| \leq n$ and $\rho_n(x) = 0$ for $|x| \geq n+1$. Let $\phi_n = \phi^{\rho_n}$. Then $\|\phi_n\|_K \leq \|\phi\|_K$. Since $\phi_n(\mu) = \phi(\rho_n \mu)$, it suffices to show that $\phi(\rho_n \mu) - \phi(\mu)$ tends to 0 uniformly on \mathcal{M}_r for every $r < \infty$. Let

$$\eta_n = \max_{\mathcal{M}_r} |\phi(\rho_n \mu) - \phi(\mu)| = |\phi(\rho_n \mu_n) - \phi(\mu_n)|$$

for some $\mu_n \in \mathcal{M}_r$ (recall that \mathcal{M}_r is compact). We have $\rho_n \mu_n \in \mathcal{M}_r$. For each $f \in C_0(\mathbb{R}^N)$, $\langle f, \rho_n \mu_n \rangle = \langle f, \mu_n \rangle$ for all large enough n . Consider any subsequence such that μ_n tends to a limit μ . Then $\rho_n \mu_n$ also tends to μ for n in this subsequence. Since $\phi|_{\mathcal{M}_r}$ is continuous, both $\phi(\rho_n \mu_n)$ and $\phi(\mu_n)$ tend to $\phi(\mu)$. If $\limsup_{n \rightarrow \infty} \eta_n > 0$, we could find some such subsequence for which $|\phi(\rho_n \mu_n) - \phi(\mu_n)|$ tends to a positive limit, a contradiction. This proves Lemma 2.2.

LEMMA 2.3. *Let $\psi \in C_K^0(\mathcal{M})$, and B compact such that $\psi(\mu) = \psi(\mu|B)$ for all $\mu \in \mathcal{M}$. Then there exists a sequence $\psi_n \in C_K^0(\mathcal{M})$ such that $\|\psi_n\|_K \leq \|\psi\|_K$, $d(\psi_n, \psi) \rightarrow 0$ as $n \rightarrow \infty$ and $\psi_n(\mu) = 0$ whenever $\mu(B) \geq n$.*

Proof. Choose $\rho \in C_0(\mathbb{R}^N)$ with $0 \leq \rho \leq 1$, $\rho(x) = 1$ for all $x \in B$. Let $g_n \in C_0(\mathbb{R}^1)$, with $0 \leq g_n \leq 1$, $g_n(s) = 1$ if $s \leq n-1$, $g_n(s) = 0$ if $s \geq n$. Let

$$\psi_n(\mu) = g_n(\langle \rho, \mu \rangle) \psi(\mu).$$

Since $|\psi_n(\mu)| \leq |\psi(\mu)|$, $\|\psi_n\|_K \leq \|\psi\|_K$. For $\mu \in \mathcal{M}_r$, $\langle \rho, \mu \rangle \leq r$. Hence, $\psi_n(\mu) = \psi(\mu)$ if $n \geq r+1$, which implies $\psi_n \rightarrow \psi$ uniformly on \mathcal{M}_r . Thus, $d(\psi_n, \psi) \rightarrow 0$ as $n \rightarrow \infty$. Finally, $\mu(B) \geq n$ implies $\langle \rho, \mu \rangle \geq n$, and hence, $\psi_n(\mu) = 0$. This proves Lemma 2.3.

The set \mathcal{D} of “test functions.” In § 5 we shall define a “generator” for the nonlinear semigroup on the following set of functions ϕ , depending on finitely many scalar products:

$$(2.10) \quad \begin{aligned} \mathcal{D} = \{ \phi : \phi(\mu) = F(\langle f_1, \mu \rangle, \dots, \langle f_J, \mu \rangle), \\ F \in C_b^\infty(\mathbb{R}^J), f_1, \dots, f_J \in C_0^\infty(\mathbb{R}^N), J = 1, 2, \dots \}. \end{aligned}$$

(More properly, we should say “pregenerator,” instead of “generator,” since we do not characterize the domain of the nonlinear semigroup.) In § 4 we shall weaken slightly the conditions on F, f_1, \dots, f_J , to obtain certain sets \mathcal{D}_m containing \mathcal{D} .

LEMMA 2.4. *For every $\phi \in C_K(\mathcal{M})$ there exists a sequence $\psi_n \in \mathcal{D}$ such that $\|\psi_n\|_K \leq \|\phi\|_K + n^{-1}$ and $d(\psi_n, \phi) \rightarrow 0$ as $n \rightarrow \infty$.*

Proof. By Lemmas 2.2 and 2.3 it suffices to suppose that, in addition, there exist compact B and $a > 0$ such that $\phi(\mu) = \phi(\mu|B)$ for all μ and $\phi(\mu) = 0$ if $\mu(B) \geq a$. Following a similar construction in [5, § 3], given $\varepsilon > 0$, we take $g_1, \dots, g_J, x_1, \dots, x_J$ with the following properties:

$$\begin{aligned} g_j \in C_0^\infty(\mathbb{R}^N), \quad g_j \geq 0, \quad \text{diam}(\text{spt } g_j) < \varepsilon, \\ \sum_{j=1}^J g_j(x) \leq 1, \quad x \in \mathbb{R}^N, \quad \sum_{j=1}^J g_j(x) = 1, \quad x \in B, \\ x_j \in B \cap \text{spt } g_j. \end{aligned}$$

Let

$$\mathbb{R}_+^J = \{z \in \mathbb{R}^J : z_j \geq 0 \text{ for } j = 1, \dots, J\},$$

$$\tilde{F}(z) = \phi\left(\sum_{j=1}^J z_j \delta_{x_j}\right),$$

where δ_x denotes the Dirac measure at x . Then $\tilde{F} \in C_0(\mathbb{R}_+^J)$. In fact, $\tilde{F}(z) = 0$ whenever

$$\sum_{j=1}^J z_j \delta_{x_j}(B) = \sum_{j=1}^J z_j \geq a.$$

By regularizing, there exists $F \in C_0^\infty(\mathbb{R}_+^J)$ such that $|F(z) - \tilde{F}(z)| \leq \varepsilon$ for all $z \in \mathbb{R}_+^J$. Then, taking $z_j = \langle g_j, \mu \rangle$,

$$\psi(\mu) = F(\langle g_1, \mu \rangle, \dots, \langle g_J, \mu \rangle)$$

is in \mathcal{D} . For all μ ,

$$|\psi(\mu)| \leq \left| \phi \left(\sum_{j=1}^J \langle g_j, \mu \rangle \delta_{x_j} \right) \right| + \varepsilon \leq \|\phi\|_K \left(1 + \left\| \sum_{j=1}^J \langle g_j, \mu \rangle \delta_{x_j} \right\|_K^K \right) + \varepsilon \leq \|\phi\|_K (1 + \|\mu\|_K^K) + \varepsilon.$$

Therefore, $\|\psi\|_K \leq \|\phi\|_K + \varepsilon$.

We take $\varepsilon = \varepsilon_n = n^{-1}$, and corresponding $g_{jn}, x_{jn}, j = 1, \dots, J_n$. The corresponding ψ_n obtained from the construction above has the properties required in Lemma 2.4. To show that $d(\psi_n, \phi) \rightarrow 0$, it suffices to show that $\psi_n(\mu) \rightarrow \phi(\mu)$ uniformly on \mathcal{M}_r for any $r > 0$ as $n \rightarrow \infty$. Now

$$|\psi_n(\mu) - \phi[G_n(\mu)]| \leq \varepsilon_n, \quad G_n(\mu) = \sum_{j=1}^{J_n} \langle g_{jn}, \mu \rangle \delta_{x_{jn}}.$$

On \mathcal{M}_r both vague and w^* -convergence of a sequence are equivalent to convergence in the metric δ in (2.7). For each m , $|\langle f_m, G_n(\mu) \rangle - \langle f_m, \rho_n \mu \rangle| \rightarrow 0$ as $n \rightarrow \infty$ uniformly for $\mu \in \mathcal{M}_r$, where $\rho_n = \sum_j g_{jn}$. Therefore, $\delta(G_n(\mu), \rho_n \mu) \rightarrow 0$ uniformly on \mathcal{M}_r as $n \rightarrow \infty$. Since \mathcal{M}_r is compact and ϕ continuous on \mathcal{M}_r , ϕ is uniformly continuous on \mathcal{M}_r . Thus, $|\phi[G_n(\mu)] - \phi(\rho_n \mu)| \rightarrow 0$ uniformly on \mathcal{M}_r . Since $\rho_n(x) = 1$ on B , $\phi(\rho_n \mu) = \phi(\mu) = \phi(\mu|B)$. This proves that $\psi_n(\mu) \rightarrow \phi(\mu)$ uniformly on \mathcal{M}_r , as required.

3. The control problem for Λ_r . We begin with a summary of assumptions and notation, together with a review of concepts from [3]. We make the same assumptions as in [3] about the coefficients in (1.1):

(A1) σ is a bounded, Lipschitz $N \times D$ matrix-valued function on \mathbb{R}^N .

(A2) $b(x, u) = b^0(x) + b^1(x)u$, where b^0, b^1 are bounded, Lipschitz functions on \mathbb{R}^N .

Note that b^0 has values in \mathbb{R}^N , and b^1 has $N \times L$ matrices as values. In § 5 we shall impose additional smoothness conditions on σ, b^0, b^1 .

(A3) $h \in C_b^2(\mathbb{R}^N; \mathbb{R}^M)$.

(A4) \mathcal{U} is a convex, compact subset of \mathbb{R}^L .

We use Y to denote an \mathbb{R}^M -valued function, and U a \mathcal{U} -valued function, of time $t \geq 0$. Let Y_t, U_t denote their respective values at time t . Let

$$\Omega = \{(Y, U): Y_0 = 0, Y \in C([0, \infty); \mathbb{R}^M), U \in L^2([0, T]; \mathcal{U}) \text{ for each } T < \infty\}.$$

Let Ω_T denote the set of restrictions to $[0, T]$ of functions $(Y, U) \in \Omega$. As in [3], we give Ω_T a metric in which convergence of a sequence (Y_n, U_n) means uniform convergence on $[0, T]$ of Y_n and weak convergence of U_n in $L^2([0, T]; \mathcal{U})$. We give Ω a metric in which convergence of (Y_n, U_n) is equivalent to convergence of (Y_n, U_n) restricted to $[0, T]$ for every $T < \infty$. Let

$$\mathcal{F}_t(Y) = \sigma\{Y_s, 0 \leq s \leq t\},$$

$$\mathcal{F}_t(U) = \sigma\{V_s, 0 \leq s \leq t\}, \quad V_t = \int_0^t U_\theta d\theta,$$

$$\mathcal{G}_t = \mathcal{F}_t(Y) \times \mathcal{F}_t(U).$$

These are σ -algebras of subsets of Ω . However, if $t \leq T$, they can also be regarded as σ -algebras of subsets of Ω_T . In [3], Ω_T was denoted by Ω^2 and \mathcal{G}_t by \mathcal{G}_t^2 .

Let \mathcal{G}_∞ be the least σ -algebra containing \mathcal{G}_t for all $t \geq 0$.

DEFINITION. An *admissible control* on $[0, T]$ is a probability measure π_T on $(\Omega_T, \mathcal{G}_T)$, such that Y is a $\pi_T, \{\mathcal{G}_t\}$ -Wiener process for $0 \leq t \leq T$.

An *admissible control* is a probability measure π on $(\Omega, \mathcal{G}_\infty)$ such that Y is a $\pi, \{\mathcal{G}_t\}$ -Wiener process for $t \geq 0$.

The definition of admissible control on $[0, T]$ is exactly as in [3]. If π is an admissible control, then its restriction π_T to \mathcal{G}_T is admissible on $[0, T]$.

Let \mathcal{A}_T denote the set of all admissible controls π_T on $[0, T]$. Then \mathcal{A}_T is compact under weak sequential convergence of probability measures [3, Lemma 2.3]. Let \mathcal{A} denote the set of all admissible controls with the weak sequential convergence topology. Then \mathcal{A} is a compact metric space under (for instance) the Prokhorov metric. Moreover, $\pi_n \rightarrow \pi$ if and only if the restrictions $\pi_{n,T}$ tend to π_T as $n \rightarrow \infty$ for each T finite.

The *unnormalized conditional distribution measure* Λ_t . For every $\mu \in \mathcal{M}$, $(Y, U) \in \Omega$ and $t \geq 0$, we define $\Lambda_t = \Lambda_{t\mu}^{YU}$ by formula [3, (3.9)]. (In [3] we wrote Λ_t^{YU} , but now we wish to emphasize its dependence on the initial value $\mu = \Lambda_0$.) From its definition, $\Lambda_t \in \mathcal{M}$ and Λ_t is \mathcal{G}_t -measurable as a function of $(Y, U) \in \Omega$. In [3, § 3] we interpreted Λ_t as an unnormalized conditional distribution of X_t in (1.1a) with respect to the σ -algebra \mathcal{G}_t generated by the observation and control past up to t . The normalized conditional distribution of X_t is $\|\Lambda_t\|^{-1} \Lambda_t$. The intuitive reason for conditioning on \mathcal{G}_t , rather than on $\mathcal{F}_t(Y)$, is that U_t is not necessarily $\mathcal{F}_t(Y)$ -measurable π -almost surely, when $\pi \in \mathcal{A}$. For the smaller class of strict-sense admissible controls [3, § 6], one can condition on $\mathcal{F}_t(Y)$ instead of \mathcal{G}_t .

We shall need the following properties of Λ_t , proved in [3].

Property 3.1. For each $t \geq 0$, $r < \infty$, $\Lambda_{t\mu}^{YU}$ is continuous on $\mathcal{M}_r \times \Omega$. See [3, Lemma 3.2].

Property 3.2. For each finite T, r, a there exists $\rho = \rho(T, r, a)$ such that $0 \leq t \leq T$, $\|\mu\| \leq r$, $\|Y\|_T \leq a$ imply $\|\Lambda_{t\mu}^{YU}\| \leq \rho$. Here $\|Y\|_T = \max_{0 \leq t \leq T} |Y(t)|$. See [3, (3.6)]; since Λ_t depends linearly on $\mu = \Lambda_0$, it suffices to consider $\|\mu\| = 1$.

Property 3.3. The Zakai equation holds:

$$(3.1) \quad d\langle f, \Lambda_t \rangle = \langle L^U f, \Lambda_t \rangle dt + \langle hf, \Lambda_t \rangle \cdot dY_t, \text{ all } f \in C_b^2(\mathbb{R}^N).$$

See [3, Thm. 5.2]. Here, for constant control $u \in \mathcal{U}$, L^u is the generator of the diffusion process in \mathbb{R}^N corresponding to (1.1a):

$$(3.2) \quad L^u f = \frac{1}{2} \sum_{i,j=1}^N a_{ij}(x) f_{x_i x_j} + (b^0(x) + b^1(x)u) \cdot \nabla f,$$

with $a = \sigma\sigma'$.

Property 3.4. For every $T < \infty, K = 1, 2, \dots$, there exists γ_{KT} such that

$$E_\pi \|\Lambda_t\|^K \leq \gamma_{KT} \|\mu\|^K, \quad 0 \leq t \leq T$$

for all $\pi \in \mathcal{A}$. See [3, Thm. 5.3] with $j = 0$.

For $t \geq 0, \mu \in \mathcal{M}, \pi \in \mathcal{A}, \phi \in C(\mathcal{M})$ let

$$(3.3) \quad J(t, \mu, \pi, \phi) = E_\pi \phi(\Lambda_{t\mu}^{YU}).$$

Since $\phi \in C_K(\mathcal{M})$ for some K , the expectation exists by Properties 3.1 and 3.4.

LEMMA 3.5. *Let $\|\phi_n\|_K \leq C$ and $d(\phi_n, \phi) \rightarrow 0$ as $n \rightarrow \infty$. Then*

$$J(t, \mu, \pi, \phi) = \lim_{n \rightarrow \infty} J(t, \mu, \pi, \phi_n)$$

uniformly on $[0, T] \times \mathcal{M}_r \times \mathcal{A}$, for any finite T, r .

Proof. Consider $\Gamma \subset \Omega$, and let $\Gamma_T \subset \Omega_T$ denote the set of restrictions to $[0, T]$ of $(Y, U) \in \Gamma$. Then

$$(*) \quad |E_\pi \phi_n(\Lambda_t) - E_\pi \phi(\Lambda_t)| \leq \int_\Gamma |\phi_n(\Lambda_t) - \phi(\Lambda_t)| d\pi + \int_{\Gamma'} |\phi_n(\Lambda_t) - \phi(\Lambda_t)| d\pi,$$

with $\Gamma' = \Omega - \Gamma$. If Γ_T is a compact subset of Ω_T , then $\|Y\|_T$ is bounded on Γ . By 3.2, $0 \leq t \leq T$, $(Y, U) \in \Gamma$, $\mu \in \mathcal{M}_r$ imply $\Lambda_t \in \mathcal{M}_\rho$ for some ρ . Since $d(\phi_n, \phi) \rightarrow 0$, $\phi_n \rightarrow \phi$ uniformly on \mathcal{M}_ρ . Therefore, the first term on the right side of $(*)$ tends to 0 as $n \rightarrow \infty$, uniformly with respect to $(t, \mu, \pi) \in [0, T] \times \mathcal{M}_r \times \mathcal{A}$.

It remains to show that, given $\varepsilon > 0$, Γ can be chosen such that the last term in $(*)$ is less than ε , uniformly on $[0, T] \times \mathcal{M}_r \times \mathcal{A}$. Now

$$|\phi_n(\Lambda_t) - \phi(\Lambda_t)| \leq (\|\phi_n\|_K + \|\phi\|_K)(1 + \|\Lambda_t\|^K) \leq 2C(1 + \|\Lambda_t\|^K).$$

By Cauchy-Schwarz and Property 3.4,

$$\int_{\Gamma'} \|\Lambda_t\|^K d\pi \leq \pi(\Gamma')^{1/2} \left(\int_{\Gamma'} \|\Lambda_t\|^{2K} d\pi \right)^{1/2} \leq \pi(\Gamma')^{1/2} \gamma_{2K, T}^{1/2} \|\mu\|^K.$$

Under $(Y, U) \rightarrow Y$, π projects onto Wiener measure w . Let $A \subset C([0, T]; \mathbb{R}^N)$ be compact with $Y_0 = 0$ for all $Y \in A$ and

$$w[C([0, T]; \mathbb{R}^N) - A] < \varepsilon^2 [2C(1 + \gamma_{2K, T}^{1/2} r^K)]^{-2}.$$

We choose Γ such that $\Gamma_T = A \times L^2([0, T]; \mathcal{U})$. Since $L^2([0, T]; \mathcal{U})$ is compact (weak topology), Γ_T is compact. We have

$$\int_{\Gamma'} |\phi_n(\Lambda_t) - \phi(\Lambda_t)| d\pi \leq 2C(\pi(\Gamma') + \pi(\Gamma')^{1/2} \gamma_{2K, T}^{1/2} r^K) < \varepsilon,$$

as required. This proves Lemma 3.5.

LEMMA 3.6. *For each $t \geq 0$, $\phi \in C(\mathcal{M})$, $r < \infty$, $J(t, \mu, \pi, \phi)$ is continuous on $\mathcal{M}_r \times \mathcal{A}$.*

Proof. Let $g(\mu, Y, U) = \phi(\Lambda_{t, \mu}^{Y, U})$. By 3.1, 3.2, g is continuous on $\mathcal{M}_r \times \Omega$ (recall that $(Y_n, U_n) \rightarrow (Y, U)$ implies $\|Y_n - Y\|_t \rightarrow 0$, and hence, $\|Y_n\|_t \leq a$ for some a). Moreover, $g(\mu, \cdot, \cdot)$ is \mathcal{G}_t -measurable.

Suppose first that $\phi(\mu)$ is bounded on \mathcal{M} . Let $\mu_n \rightarrow \mu$, $\pi_n \rightarrow \pi$ with $\mu_n \in \mathcal{M}_r$. By definition of weak convergence,

$$\lim_{n \rightarrow \infty} \int_\Omega g(\mu, Y, U) d\pi_n = \int_\Omega g(\mu, Y, U) d\pi.$$

Moreover, $|g(\mu_n, Y, U) - g(\mu, Y, U)| \rightarrow 0$ as $n \rightarrow \infty$, uniformly on any $\Gamma \subset \Omega$ such that the set Γ_t of restrictions to $[0, t]$ of $(Y, U) \in \Gamma$ is compact. As in the proof of Lemma 3.5, we can choose Γ such that $\pi_n(\Omega - \Gamma)$ is arbitrarily small, uniformly with respect to n . This proves Lemma 3.6 in case $\phi(\mu)$ is bounded on \mathcal{M} .

Now take any $\phi \in C_K(\mathcal{M})$. By Lemmas 2.2 and 2.3, there exist $\varphi_n \in C_K(\mathcal{M})$ such that $|\phi_n(\mu)|$ is bounded on \mathcal{M} for each n , $\|\phi_n\|_K$ is bounded, and $d(\phi_n, \phi) \rightarrow 0$ as $n \rightarrow \infty$. Lemma 3.6 now follows from Lemma 3.5.

The control problem. Given t, μ, ϕ , we consider the problem of minimizing $J(t, \mu, \pi, \phi) = E_\pi \phi(\Lambda_t)$ on the space \mathcal{A} of admissible controls π . We can regard the Zakai equation (3.1) as governing the dynamics of the “state” process Λ_t for this control problem. Since Λ_t is an unnormalized conditional distribution measure for X_t in the partially observed control system (1.1), we call the problem of minimizing $E_\pi \phi(\Lambda_t)$ a “separated” optimal control problem.

Following Nisio [9], let

$$(3.4) \quad \mathcal{T}_t \phi(\mu) = \min_{\pi \in \mathcal{A}} J(t, \mu, \pi, \phi).$$

This minimum is attained by Lemma 3.6. Since $\phi(\Lambda_t)$ is \mathcal{G}_t -measurable, the minimum is the same taken in the class \mathcal{A}_t of admissible controls on $[0, t]$:

$$(3.5) \quad \mathcal{T}_t \phi(\mu) = \min_{\pi_t \in \mathcal{A}_t} J(t, \mu, \pi_t, \phi).$$

For the special case $\phi(\mu) = \langle G, \mu \rangle$, $J = E_\pi \langle G, \Lambda_t \rangle$, which is of the form considered in the existence theorem [3, Thm. 4.1]. However, if ϕ has this special linear form, $\mathcal{T}_t \phi(\mu)$ is not linear in μ . Hence, we define \mathcal{T}_t on the bigger space $C(\mathcal{M})$ and not merely on the space of ϕ of the form $\phi(\mu) = \langle G, \mu \rangle$.

THEOREM 3.1. $\phi \in C_K(\mathcal{M})$ implies $\mathcal{T}_t \phi \in C_K(\mathcal{M})$.

Proof. By Lemma 3.6 and the fact that \mathcal{M}_r and \mathcal{A} are compact, $\|\mu_n\| \leq r$ and $\mu_n \rightarrow \mu$ imply $\mathcal{T}_t \phi(\mu_n) \rightarrow \mathcal{T}_t \phi(\mu)$. Since any w^* -convergent sequence μ_n has $\|\mu_n\|$ bounded, $\mathcal{T}_t \phi$ is continuous on \mathcal{M} . From 3.4,

$$|J(t, \mu, \pi, \phi)| \leq \|\phi\|_K E \int_{\Omega} (1 + \|\Lambda_t\|^K) d\pi \leq \|\phi\|_K (1 + \gamma_{Kt} \|\mu\|^K) \leq (1 + \gamma_{Kt}) \|\phi\|_K (1 + \|\mu\|^K).$$

Thus, $\|\mathcal{T}_t \phi\|_K \leq (1 + \gamma_{Kt}) \|\phi\|_K$, which proves Theorem 3.1.

COROLLARY. \mathcal{T}_t maps $C(\mathcal{M})$ into $C(\mathcal{M})$.

In the next section we establish the semigroup property of \mathcal{T}_t .

4. The semigroup property. The purpose of this section is to prove the following two theorems.

THEOREM 4.1. For every $\phi \in C(\mathcal{M})$, $s, t \geq 0$,

$$\mathcal{T}_{s+t} \phi = \mathcal{T}_s \mathcal{T}_t \phi.$$

Theorems 3.1 and 4.1 imply that \mathcal{T}_t is a (nonlinear) semigroup on $C(\mathcal{M})$. Let $C_b(\mathcal{M})$ denote the space of bounded continuous functions on \mathcal{M} (it is the same as $C_K(\mathcal{M})$ when $K = 0$.) From (3.3), (3.4) $\|\mathcal{T}_t \phi - \mathcal{T}_t \psi\|_0 \leq \|\phi - \psi\|_0$. Hence, when restricted to $C_b(\mathcal{M})$, \mathcal{T}_t is a contracting semigroup on $C_b(\mathcal{M})$.

THEOREM 4.2. For every $\phi \in C(\mathcal{M})$, $\mathcal{T}_t \phi$ is a continuous function of $t \in [0, \infty)$ in the d -metric on $C(\mathcal{M})$.

The proof of Theorem 4.1 will be based on a series of three lemmas. We begin by temporarily imposing rather stringent conditions on the coefficients in (1.1), and on Y, U, μ . We say that the coefficients are regular if σ, b^0, b^1, g are of class $C_b^\infty(\mathbb{R}^N; \mathbb{R}^l)$ for the appropriate $l = ND, N, NL, M$, respectively. Let us denote by $C_e^{1,2}$ the class of functions q on $[0, \infty) \times \mathbb{R}^N$ with the following properties:

- (i) q and the partial derivatives $q_b, q_{x_i}, q_{x_i x_j}$ are continuous, $i, j = 1, \dots, N$.

(ii) For each $T > 0$, there exist $C, k > 0$ (depending perhaps on T) such that

$$|r(x, t)| \leq C \exp(-k|x|^2), \quad 0 \leq t \leq T,$$

where r denotes any of the functions $q, q_{x_i}, q_{x_i x_j}$.

For brevity, we write $q(t) = q(t, \cdot)$.

LEMMA 4.1. Assume that the coefficients in (1.1) are regular, and that $Y \in C^1([0, \infty); \mathbb{R}^N)$, $U \in C([0, \infty); \mathcal{U})$. Then:

(a) If μ has a density $p_0 \in C_0^\infty(\mathbb{R}^N)$, then $\Lambda_t (= \Lambda_{t\mu}^{YU})$ has a density $q \in C_e^{1,2}$, satisfying the partial differential equation

$$(4.1) \quad \frac{dq}{dt} = (L^{U_t})^* q + hq \cdot \dot{Y}_t - \frac{1}{2}|h|^2 q, \quad t \geq 0, \quad q(0) = p_0.$$

(b) If $q \in C_e^{1,2}$ is a solution of (4.1) with $q(0)$ the density of μ , then $q(t)$ is the density of Λ_t for all $t \geq 0$.

Here $(L^u)^*$ denotes the formal adjoint of the operator L^u in (3.2), and $\dot{Y}_t = dY/dt$. Note that part (a) of the lemma, but not part (b), requires that $q(0)$ has compact support.

Proof of Lemma 4.1. To prove (a), we recall from [3, § 5] that

$$(4.2) \quad p(t) = q(t) \exp(-Y_t \cdot h)$$

is a solution in $C_e^{1,2}$ to the partial differential equation

$$(4.3) \quad \begin{aligned} \frac{dp}{dt} &= \check{L}_t^* p + e(t)p, \quad \text{where} \\ \check{L}_t &= L_t - (aY_t \cdot \nabla h, \nabla), \quad L_t = L^{U_t}, \\ e(t) &= \frac{1}{2}(aY_t \cdot \nabla h, Y_t \cdot \nabla h) - Y_t \cdot L_t h - \frac{1}{2}|h|^2, \end{aligned}$$

where $(a\xi, \eta) = \sum_{i,j=1}^N a_{ij}\xi_i\eta_j$ and \cdot denotes the \cdot product in \mathbb{R}^M . The operators L_t^*, \check{L}_t^* are related by

$$(4.4) \quad (\check{L}_t^* p) \exp(Y_t \cdot h) = L_t^* q - eq - \frac{1}{2}|h|^2 q.$$

Equation (4.4) follows upon multiplying both sides of (4.3) by $f \in C_0(\mathbb{R}^N)$, integrating by parts, and using the relation

$$\exp(Y_t \cdot h)L_t f = \check{L}_t[f \exp(Y_t \cdot h)] + e(t)f \exp(Y_t \cdot h) + \frac{1}{2}|h|^2 f \exp(Y_t \cdot h).$$

Then equation (4.1) follows from (4.3), (4.4) and the product rule applied to $(d/dt)[p \exp(Y_t \cdot h)]$.

To prove (b), if $q \in C_e^{1,2}$ satisfies (4.1), then the above calculation shows that $p(t)$ defined by (4.2) is a solution in $C_e^{1,2}$ to (4.3). It follows from [3, (5.5)] that $q(t)$ is the density of Λ_t . (In the derivation of [3, (5.5)] it was stated that $q(0) \in C_0(\mathbb{R}^N)$. However, the proof there is based on integrations by parts, and is the same if $q \in C_e^{1,2}$). This proves Lemma 4.1.

For $s \geq 0$, let us introduce the notation

$$Y_\tau^s = Y_{s+\tau} - Y_s, \quad U_\tau^s = U_{s+\tau}, \quad \tau \geq 0.$$

In particular, $Y_0^s = 0$; and $(Y, U) \in \Omega$ implies $(Y^s, U^s) \in \Omega$.

LEMMA 4.2. For every $(Y, U) \in \Omega$, $\mu \in \mathcal{M}$, $s, t \geq 0$,

$$(4.5) \quad \Lambda_{s+t, \mu}^{YU} = \Lambda_{t, \Lambda_s}^{Y^s U^s}, \quad \text{where } \Lambda_s = \Lambda_{s\mu}^{YU}.$$

Proof. Step 1. First assume the conditions of Lemma 4.1 on b^l, σ, h, Y, U , and that $\mu = \Lambda_0$ has a density $p_0 \in C_0^\infty(\mathbb{R}^N)$. By Lemma 4.1(a), Λ_τ has a density $q(\tau) \in C_e^{1,2}$ satisfying (4.1) for $\tau \geq 0$. Let $q^s(\tau) = q(s + \tau)$. Then q^s is a solution in $C_e^{1,2}$ of (4.1), with (Y, U) replaced by (Y^s, U^s) ; note that $\dot{Y}_{s+\tau} = \dot{Y}_\tau$ and $q^s(0) = q(s)$. By Lemma 4.1(b), $q^s(t)$ is the density of $\Lambda_{t\Lambda_s}^{Y^s U^s}$. This proves (4.5) under these conditions.

Step 2. Again assume regular coefficients $b^l, \sigma, h, l = 0, 1$. Let $(Y, U) \in \Omega, \mu \in \mathcal{M}$. Let $(Y_n, U_n) \rightarrow (Y, U), \mu_n \rightarrow \mu$, where Y_n, U_n, μ_n satisfy the conditions in Step 1 for each n . Write $\Lambda_s^n = \Lambda_{s\mu_n}^{Y_n U_n}$. By Property 3.1, as $n \rightarrow \infty$,

$$\Lambda_{s+t, \mu_n}^{Y_n U_n} \rightarrow \Lambda_{s+t, \mu}^{YU}, \quad \Lambda_s^n \rightarrow \Lambda_s, \quad \Lambda_{t\Lambda_s^n}^{Y_n^s U_n^s} \rightarrow \Lambda_{t\Lambda_s}^{Y^s U^s}.$$

At the last step we used the fact that $(Y_n^s, U_n^s) \rightarrow (Y^s, U^s)$. This implies (4.5).

Step 3. Fix $\mu \in \mathcal{M}, (Y, U) \in \Omega$. Let σ_n, b_n^l, h_n be regular for each n , uniformly bounded together with their first order partial derivatives and tending uniformly to σ, b^l, h as $n \rightarrow \infty, l = 0, 1$. Write $\Lambda_{t\mu}^n = \Lambda_{t\mu}^{nYU}$ to indicate that the coefficients depend on n . The proof of [3, Thm. 5.1] shows the following: $\nu_n \rightarrow \nu, \nu_n \in \mathcal{M}_r$ implies $\Lambda_{\tau\nu_n}^n \rightarrow \Lambda_{\tau\nu}$ for any $\tau \geq 0$. We then have as $n \rightarrow \infty$

$$\Lambda_{s+t, \mu}^n \rightarrow \Lambda_{s+t, \mu}, \quad \Lambda_{s\mu}^n \rightarrow \Lambda_{s\mu}.$$

Similarly, if we write $\Lambda_{s\mu}^n = \Lambda_s^n$, then

$$\Lambda_{t\Lambda_s^n}^{nY^s U^s} \rightarrow \Lambda_{t\Lambda_s}^{Y^s U^s}.$$

This implies (4.5), and hence, Lemma 4.2.

As in §3, let π_s denote the restriction to \mathcal{G}_s of $\pi \in \mathcal{A}$. Let π_s^{YU} be a regular conditional distribution for (Y^s, U^s) , given \mathcal{G}_s .

LEMMA 4.3. *If $\pi \in \mathcal{A}$, then:*

(a) $\pi_s^{YU} \in \mathcal{A}, \pi_s$ -almost surely.

(b) $J(s+t, \mu, \pi, \phi) = \int_{\Omega} J(s, \Lambda_{s\mu}^{YU}, \pi_s^{YU}, \phi) d\pi_s$

for any $\phi \in C(\mathcal{M})$.

Proof. To prove (a), it suffices to verify that, for any \mathcal{G}_s -measurable $\Phi \in C_b(\Omega)$, \mathcal{G}_t -measurable $\Psi \in C_b(\Omega), F \in C_b(\mathbb{R}^M)$ and $r > t$

$$E_\pi[\Psi(Y, U)\Phi(Y^s, U^s)F(Y_r^s - Y_t^s)] = E_\pi[\Psi(Y, U)\Phi(Y^s, U^s)]E_\pi F(Y_r^s - Y_t^s).$$

But this follows from independence under π of the random variables $\Psi(Y, U)\Phi(Y^s, U^s)$ and $F(Y_r^s - Y_t^s)$.

Part (b) is immediate from (3.3), Lemma 4.2 and properties of conditional expectations.

Proof of Theorem 4.1. For every $\pi \in \mathcal{A}$, Lemma 4.3, the definition (3.4) of $\mathcal{T}_t\phi$ and (3.5) imply

$$\begin{aligned} J(s+t, \mu, \pi, \phi) &= \int_{\Omega} J(s, \Lambda_{s\mu}^{YU}, \pi_s^{YU}, \phi) d\pi_s \\ &\geq \int_{\Omega} \mathcal{T}_t\phi(\Lambda_{s\mu}^{YU}) d\pi_s = E_{\pi_s}\mathcal{T}_t\phi(\Lambda_{s\mu}^{YU}) \geq \mathcal{T}_s\mathcal{T}_t\phi(\mu). \end{aligned}$$

Since this is true for every $\pi \in \mathcal{A}$,

$$\mathcal{T}_{s+t}\phi(\mu) \geq \mathcal{T}_s\mathcal{T}_t\phi(\mu).$$

To prove the opposite inequality, we make the following construction. Let $\rho > 0$, $\delta > 0$ to be chosen later. Let $A_0 = \mathcal{M} - \mathcal{M}_\rho$ and A_1, \dots, A_m disjoint Borel subsets of \mathcal{M}_ρ , such that

$$\mathcal{M}_\rho = A_1 \cup \dots \cup A_m,$$

and for $\nu, \nu' \in A_i$, $i = 1, \dots, m$, $\pi \in \mathcal{A}$,

$$|J(t, \nu, \pi, \phi) - J(t, \nu', \pi, \phi)| < \delta.$$

This is possible by Lemma 3.6. Choose $\mu_i \in A_i$ and $\pi_i \in \mathcal{A}$ such that

$$J(t, \mu_i, \pi_i, \phi) < \mathcal{T}_t \phi(\mu_i) + \delta.$$

For all $\nu \in A_i$,

$$(*) \quad J(t, \nu, \pi_i, \phi) < \mathcal{T}_t \phi(\nu) + 3\delta.$$

Let $\pi_0 \in \mathcal{A}$ be arbitrary. Let

$$\pi_s^{YU} = \pi_i \quad \text{if } \Lambda_{s\mu}^{YU} \in A_i.$$

Given $\pi_s \in \mathcal{A}_s$, this defines $\pi \in \mathcal{A}$ such that π_s^{YU} is a regular conditional distribution for Y^s, U^s given \mathcal{G}_s and $\pi|_{\mathcal{G}_s} = \pi_s$. By Lemma 4.3 and (*), with $\nu = \Lambda_{s\mu}^{YU} = \Lambda_s$,

$$\begin{aligned} J(s+t, \mu, \pi, \phi) &= \int_{\Omega} J(s, \Lambda_s, \pi_s^{YU}, \phi) d\pi_s \\ &\leq \int_{\mathcal{M}_\rho} \mathcal{T}_t \phi(\Lambda_s) d\pi_s + \int_{A_0} J(s, \Lambda_s, \pi_0, \phi) d\pi_s + 3\delta. \end{aligned}$$

Since $\mathcal{T}_{s+t} \phi(\mu) \leq J(s+t, \mu, \pi, \phi)$, we have

$$\mathcal{T}_{s+t} \phi(\mu) \leq E_{\pi_s} \mathcal{T}_t \phi(\Lambda_s) + \int_{A_0} J(s, \Lambda_s, \pi_0, \phi) d\pi_s + \int_{A_0} |\mathcal{T}_t \phi(\Lambda_s)| d\pi_s + 3\delta.$$

Now $\phi \in C_K(\mathcal{M})$ for some K . We have, for some C_1 ,

$$|J(s, \Lambda_s, \pi_0, \phi)| \leq C_1(1 + \|\Lambda_s\|^K), \quad |\mathcal{T}_t \phi(\Lambda_s)| \leq C_1(1 + \|\Lambda_s\|^K),$$

while by Property 3.4 and the fact that $A_0 = \{\nu: \|\nu\| > \rho\}$

$$\int_{A_0} (1 + \|\Lambda_s\|^K) d\pi_s \leq \rho^{-K} E_{\pi_s} [\|\Lambda_s\|^K + \|\Lambda_s\|^{2K}] \leq C_2 \rho^{-K} (1 + r^{2K})$$

for $\mu \in \mathcal{M}_r$. Therefore, given $\varepsilon > 0$, we can choose ρ large enough and δ small enough that

$$\mathcal{T}_{s+t} \phi(\mu) \leq E_{\pi_s} \mathcal{T}_t \phi(\Lambda_s) + \varepsilon,$$

for all $\mu \in \mathcal{M}_r$ and $\pi_s \in \mathcal{A}_s$. Upon taking the inf over π_s (recall (3.5)), we have

$$\mathcal{T}_{s+t} \phi(\mu) \leq \mathcal{T}_s \mathcal{T}_t \phi(\mu) + \varepsilon.$$

Since ε is arbitrary, we obtain Theorem 4.1.

In preparation for the proof of Theorem 4.2, and for § 5, let us introduce the following family of operators \mathcal{L}^u , for constant controls $u \in \mathcal{U}$. Let

$$\begin{aligned} \tilde{\mathcal{D}} &= \{\phi: \phi(\mu) = F(\langle f_1, \mu \rangle, \dots, \langle f_J, \mu \rangle), \\ &\quad F \in C^2(\mathbb{R}^J), f_1, \dots, f_J \in C_0^2(\mathbb{R}^J), J = 1, 2, \dots\}, \end{aligned}$$

and for each integer $m \geq 0$, let

$$(4.6) \quad \mathcal{D}_m = \{\phi \in \tilde{\mathcal{D}} : |F_{z_j}(z)| \leq C(1 + |z|^{m+1}), |F_{z_j z_k}(z)| \leq C(1 + |z|^m), j, k = 1, \dots, J\}.$$

We have the inclusions $\mathcal{D} \subset \mathcal{D}_m \subset C_{m+2}(\mathcal{M})$.

For $\phi \in \mathcal{D}_m$ and $u \in \mathcal{U}$, let

$$(4.7) \quad \mathcal{L}^u \phi(\mu) = \sum_{j=1}^J F_{z_j}(\cdot \cdot \cdot) \langle L^u f_j, \mu \rangle + \sum_{j,k=1}^J F_{z_j z_k}(\cdot \cdot \cdot) \langle h f_j, \mu \rangle \cdot \langle h f_k, \mu \rangle$$

where $\cdot \cdot \cdot$ denotes that the partial derivatives $F_{z_j}, F_{z_j z_k}$ are evaluated at the vector $z = (\langle f_1, \mu \rangle, \dots, \langle f_J, \mu \rangle)$. It might seem that $\mathcal{L}^u \phi$ depends not just on ϕ , but also on F, f_1, \dots, f_J . However, it follows from (4.13) below that this difficulty does not occur.

LEMMA 4.4. *Let $\phi \in \mathcal{D}_m$. Then there exists c such that:*

- (a) $\mathcal{L}^u \phi \in C_{m+2}(\mathcal{M})$, $\|\mathcal{L}^u \phi\|_{m+2} \leq c$ for all $u \in \mathcal{U}$.
- (b) *The mapping $(u, \mu) \rightarrow \mathcal{L}^u \phi(\mu)$ is continuous from $\mathcal{U} \times \mathcal{M}_r$ into \mathbb{R}^1 for every $r < \infty$.*

This follows at once from (4.7).

Let us next apply the Ito differential rule to $\phi(\Lambda_t)$; for $\phi \in \mathcal{D}_m$,

$$\phi(\Lambda_t) = F(\langle f_1, \Lambda_t \rangle, \dots, \langle f_J, \Lambda_t \rangle).$$

We get, using the Zakai equation (3.1),

$$(4.8) \quad d\phi(\Lambda_t) = \mathcal{L}^{U_t} \phi(\Lambda_t) dt + \sum_{j=1}^J F_{z_j}(\cdot \cdot \cdot) \langle h f_j, \Lambda_t \rangle \cdot dY_t,$$

where $\cdot \cdot \cdot$ denotes $(\langle f_1, \Lambda_t \rangle, \dots, \langle f_J, \Lambda_t \rangle)$. Since $|F_{z_j}| \leq C(1 + |z|^{m+1})$, the components of $F_{z_j}(\langle f_1, \mu \rangle, \dots, \langle f_J, \mu \rangle) \langle h f_j, \mu \rangle$ are in $C_{m+2}(\mathcal{M})$. From property 3.4, the integrals on $[0, t]$ of the last term in (4.8) is a square integrable $\pi, \{\mathcal{G}_t\}$ martingale for any $\pi \in \mathcal{A}$. By taking $E_\pi \int_0^t$ in (4.8) and using Lemma 4.4(a) we get

$$(4.9) \quad E_\pi \phi(\Lambda_t) = \phi(\mu) + E_\pi \int_0^t \mathcal{L}^{U_\theta} \phi(\Lambda_\theta) d\theta,$$

for any $\phi \in \mathcal{D}_m$, $\pi \in \mathcal{A}$ and any initial data $\mu = \Lambda_0$.

LEMMA 4.5. *Let $\psi \in \mathcal{D}_m$, $0 \leq s \leq t \leq T$. Then there exists α (depending on ψ, m and T) such that*

$$\|\mathcal{T}_t \psi - \mathcal{T}_s \psi\|_{m+2} \leq \alpha(t-s).$$

Proof. Consider any $\pi \in \mathcal{A}$. By (4.9)

$$|E_\pi \phi(\Lambda_t) - E_\pi \phi(\Lambda_s)| \leq E \int_s^t |\mathcal{L}^{U_\theta} \phi(\Lambda_\theta)| d\theta \leq \max_{u \in \mathcal{U}} \|\mathcal{L}^u \phi\|_{m+2} \int_s^t (1 + E\|\Lambda_\theta\|^{m+2}) d\theta.$$

By Lemma 4.4(a) and Property 3.4,

$$|E_\pi \phi(\Lambda_t) - E_\pi \phi(\Lambda_s)| \leq c(1 + \gamma_{m+2,T})(t-s)(1 + \|\mu\|^{m+2}).$$

Since this holds for all $\pi \in \mathcal{A}$, we get Lemma 4.5 with $\alpha = c(1 + \gamma_{m+2,T})$.

Proof of Theorem 4.2. For some K , $\phi \in C_K(\mathcal{M})$. By Lemma 2.4, there exists $\psi_n \in \mathcal{D}$, $n = 1, 2, \dots$, such that $\|\psi_n\|_K$ is bounded and $d(\psi_n, \phi) \rightarrow 0$. Fix $T > 0$. For $0 \leq s \leq T$, we write

$$\mathcal{T}_t \phi - \mathcal{T}_s \phi = [\mathcal{T}_t \phi - \mathcal{T}_t \psi_n] + [\mathcal{T}_t \psi_n - \mathcal{T}_s \psi_n] + [\mathcal{T}_s \psi_n - \mathcal{T}_s \phi].$$

Lemma 3.5 implies that the first and third terms on the right side tend to 0 as $n \rightarrow \infty$, uniformly for $0 \leq s < t \leq T$ and $\mu \in \mathcal{M}_r$.

Lemma 4.5 with $m = 0$ implies, for $\mu \in \mathcal{M}_r$,

$$|\mathcal{T}_t \psi_n(\mu) - \mathcal{T}_s \psi_n(\mu)| \leq \alpha_n(1+r^2)(t-s),$$

where α_n is some constant. Let

$$\eta(\varepsilon, r) = \sup \{|\mathcal{T}_t \phi(\mu) - \mathcal{T}_s \phi(\mu)| : \mu \in \mathcal{M}_r, 0 \leq s < t \leq T, t-s < \varepsilon\}.$$

For each r , $\eta(\varepsilon, r) \rightarrow 0$ as $\varepsilon \rightarrow 0$. This implies $d(\mathcal{T}_t \phi, \mathcal{T}_s \phi) \rightarrow 0$ as $t-s \rightarrow 0$, as required. This proves Theorem 4.2.

Constant controls. In particular, let us consider a constant control u . In our formulation, this corresponds to taking $\pi = \pi^u = w \times \delta_u$, where w is Wiener measure on $C([0, \infty); \mathbb{R}^m)$ and δ_u is the Dirac measure on $L_{\text{loc}}^2([0, \infty); \mathcal{U})$ concentrated on the constant trajectory $U_t \equiv u$. We can then write $E(=E_w)$, instead of E_{π^u} , and obtain from (4.9)

$$(4.10) \quad E\phi(\Lambda_t) = \phi(\mu) + E \int_0^t \mathcal{L}^u \phi(\Lambda_\theta) d\theta, \quad \phi \in \mathcal{D}_m.$$

For constant u , we may regard Λ_t as defined on the sample space $C([0, \infty); \mathbb{R}^M)$ of Y -trajectories, endowed with the family $\{\mathcal{F}_t(Y)\}$ of σ -algebras and with Wiener measure w . It follows from Lemma 4.2 that $\Lambda_t = \Lambda_{t,\mu}^u$ is a Markov process (u fixed), with which is associated the linear semigroup \mathcal{T}_t^u on $C(\mathcal{M})$:

$$(4.11) \quad \mathcal{T}_t^u \phi(\mu) = E\phi(\Lambda_t),$$

where $E = E_w$.

From (4.10) we have, for $\phi \in \mathcal{D}_m$,

$$(4.12) \quad t^{-1}[\mathcal{T}_t^u \phi(\mu) - \phi(\mu) - t\mathcal{L}^u \phi(\mu)] = t^{-1} \int_0^t [\mathcal{T}_\theta^u(\mathcal{L}^u \phi)(\mu) - \mathcal{L}^u \phi(\mu)] d\theta.$$

Since $\mathcal{L}^u \phi \in C(\mathcal{M})$, the same proof as for Theorem 4.2 shows that $\mathcal{T}_\theta^u(\mathcal{L}^u \phi) \rightarrow \mathcal{L}^u \phi$ as $\theta \rightarrow 0^+$, uniformly on \mathcal{M}_r for each $r < \infty$ (alternatively we could apply Theorem 4.2 with the control space \mathcal{U} replaced by a new one-element control space $\{u\}$.) Hence the left side of (4.12) tends to 0 and $t \rightarrow 0^+$ uniformly on \mathcal{M}_r , which implies

$$(4.13) \quad \mathcal{L}^u \phi = \text{d-lim}_{t \rightarrow 0^+} t^{-1}[\mathcal{T}_t^u \phi - \phi], \quad \phi \in \mathcal{D}_m.$$

This shows that for each $m = 0, 1, 2, \dots$, \mathcal{D}_m is contained in the domain of the generator of the linear semigroup \mathcal{T}_t^u and that \mathcal{L}^u agrees on \mathcal{D}_m with the generator.

5. The generator of the semigroup \mathcal{T}_t^u . We define the operator \mathcal{L} on the dense subset \mathcal{D} of $C(\mathcal{M})$ by

$$(5.1) \quad \mathcal{L}\phi(\mu) = \min_{u \in \mathcal{U}} \mathcal{L}^u \phi(\mu), \quad \phi \in \mathcal{D}.$$

Lemma 4.4 implies that $\mathcal{L}\phi \in C_2(\mathcal{M})$ for every $\phi \in \mathcal{D}$.

We need slightly stronger hypotheses on σ, b^0, b^1 than (A1), (A2) in § 3:

(A1') Condition (A1) holds and, in addition, $a \in C_b^2(\mathbb{R}^N; \mathbb{R}^{N^2})$, where $a = \sigma\sigma'$.

(A2') $b(x, u) = b^0(x) + b^1(x)u$, where $b^0 \in C_b^2(\mathbb{R}^N; \mathbb{R}^N)$ and $b^1 \in C_b^2(\mathbb{R}^N; \mathbb{R}^{N^L})$.

When (A1'), (A2'), (A3) hold, $f \in C_0^\infty(\mathbb{R}^N)$ implies $L^u f \in C_0^\infty(\mathbb{R}^N)$ and $hf \in C_0^2(\mathbb{R}^N; \mathbb{R}^M)$. From (4.7), $\phi \in \mathcal{D}$ implies $\mathcal{L}^u \phi \in \mathcal{D}_2$.

THEOREM 5.1. *For every $\phi \in \mathcal{D}$*

$$(5.2) \quad \mathcal{L}\phi = \mathbf{d}\text{-}\lim_{t \rightarrow 0^+} t^{-1}(\mathcal{T}_t\phi - \phi).$$

This theorem justifies our calling \mathcal{L} the pregenerator of the nonlinear semigroup \mathcal{T}_t . Our proof of Theorem 5.1 follows the same general line of reasoning as Nisio [9].

The proof of Theorem 5.1 depends on the following estimates for the semigroups \mathcal{T}_t^u , for any constant control $u \in \mathcal{U}$. By the same calculation used in the proof of Theorem 3.1,

$$(5.3) \quad \|\mathcal{T}_t^u\phi\|_K \leq (1 + \gamma_{Kt})\|\phi\|_K, \quad \phi \in C_K(\mathcal{M}).$$

For $\phi \in \mathcal{D}_m$, 3.4 and (4.10) imply

$$(5.4) \quad \|\mathcal{T}_t^u\phi - \phi\|_{m+2} \leq \|\mathcal{L}^u\phi\|_{m+2}(1 + \gamma_{m+2,t})t.$$

Lemma 4.4(a) gives a bound for $\|\mathcal{L}^u\phi\|_{m+2}$.

Now consider $\phi \in \mathcal{D}$,

$$\phi(\mu) = F(\langle f_1, \mu \rangle, \dots, \langle f_J, \mu \rangle),$$

with $F \in C_b^\infty(\mathbb{R}^J)$, $f_j \in C_0^\infty(\mathbb{R}^N)$. Then

$$\begin{aligned} \mathcal{L}^u\phi &= \sum_{j=1}^J \phi_j + \sum_{j,k=1}^J \phi_{jk}, \\ \phi_j(\mu) &= F_{z_j}(\langle f_1, \mu \rangle, \dots, \langle f_J, \mu \rangle) \langle L^u f_j, \mu \rangle, \\ \phi_{jk}(\mu) &= F_{z_j z_k}(\langle f_1, \mu \rangle, \dots, \langle f_J, \mu \rangle) \langle h f_j, \mu \rangle \cdot \langle h f_k, \mu \rangle, \\ \mathcal{T}_t^u\phi - \phi - t\mathcal{L}^u\phi &= \int_0^t [\mathcal{T}_\theta^u(\mathcal{L}^u\phi) - \mathcal{L}^u\phi] d\theta \\ &= \sum_j \int_0^t [\mathcal{T}_\theta^u\phi_j - \phi_j] d\theta + \sum_{j,k} \int_0^t [\mathcal{T}_\theta^u\phi_{jk} - \phi_{jk}] d\theta. \end{aligned}$$

Since $\phi_j, \phi_{jk} \in \mathcal{D}_2$, we can apply (5.4) to ϕ_j, ϕ_{jk} to get, for $0 \leq t \leq 1$,

$$(5.5) \quad \|\mathcal{T}_t^u\phi - \phi - t\mathcal{L}^u\phi\|_4 \leq \beta t^2,$$

where the constant β depends on ϕ but not on $u \in \mathcal{U}$.

LEMMA 5.1. *For $\phi \in \mathcal{D}$,*

$$\mathcal{T}_t\phi - \phi \geq \int_0^t \mathcal{T}_\theta(\mathcal{L}\phi) d\theta.$$

Proof. By (4.9), for any $\pi \in \mathcal{A}$,

$$\begin{aligned} E_\pi\phi(\Lambda_t) - \phi(\mu) &= E_\pi \int_0^t \mathcal{L}^{U_\theta}\phi(\Lambda_\theta) d\theta \geq E_\pi \int_0^t \mathcal{L}\phi(\Lambda_\theta) d\theta = \int_0^t E_\pi \mathcal{L}\phi(\Lambda_\theta) d\theta \\ &\geq \int_0^t \mathcal{T}_\theta(\mathcal{L}\phi)(\mu) d\theta. \end{aligned}$$

The minimum over \mathcal{A} of the left side is $T_t\phi(\mu) - \phi(\mu)$. This proves Lemma 5.1.

Proof of Theorem 5.1. Observe that $\mathcal{T}_t\phi \leq \mathcal{T}_t^u\phi$ for all $u \in \mathcal{U}$ (constant controls are suboptimal). Then, for $\phi \in \mathcal{D}$, $0 < t \leq 1$,

$$t^{-1}[\mathcal{T}_t\phi - \phi - t\mathcal{L}^u\phi] \leq t^{-1}[\mathcal{T}_t^u\phi - \phi - t\mathcal{L}^u\phi].$$

In particular, given μ , we take u such that $\mathcal{L}^u\phi(\mu) = \mathcal{L}\phi(\mu)$ [recall (5.1)]. By (5.5), when $0 < t \leq 1$,

$$t^{-1}[\mathcal{T}_t\phi(\mu) - \phi(\mu) - t\mathcal{L}\phi(\mu)] \leq \beta t(1 + \|\mu\|^4).$$

Therefore, uniformly for $\mu \in \mathcal{M}_r$,

$$\limsup_{t \rightarrow 0^+} t^{-1}[\mathcal{T}_t\phi(\mu) - \phi(\mu)] \leq \mathcal{L}\phi(\mu).$$

On the other hand, by Lemma 5.1,

$$\liminf_{t \rightarrow 0^+} t^{-1}[\mathcal{T}_t\phi(\mu) - \phi(\mu)] \geq \liminf_{t \rightarrow 0} t^{-1} \int_0^t \mathcal{T}_\theta(\mathcal{L}\phi)(\mu) d\theta.$$

Since $\mathcal{L}\phi \in C(\mathcal{M})$, Theorem 4.2 implies that $\mathcal{T}_\theta(\mathcal{L}\phi)(\mu) \rightarrow \mathcal{L}\phi(\mu)$ as $\theta \rightarrow 0^+$, uniformly on \mathcal{M}_r . Hence,

$$\lim_{t \rightarrow 0^+} t^{-1}[\mathcal{T}_t\phi(\mu) - \phi(\mu)] = \mathcal{L}\phi(\mu)$$

uniformly on \mathcal{M}_r for each r . This proves Theorem 5.1.

Remark. The nonlinear semigroup \mathcal{T}_t can be obtained from the family of linear semigroups \mathcal{T}_t^u , $u \in \mathcal{U}$, by the following procedure used in [9]. For $\Delta > 0$, let $\mathcal{J}_\Delta\phi(\mu) = \min_{u \in \mathcal{U}} \mathcal{T}_\Delta^u\phi(\mu)$. For $n = 1, 2, \dots$ and dyadic rational $t = m2^{-n}$ ($m = 1, 2, \dots$), let

$$\mathcal{T}_t^n\phi = \mathcal{J}_{\Delta_n}^m\phi, \quad \Delta_n = 2^{-n}, \quad \phi \in C(\mathcal{M}).$$

It is easy to show that, for dyadic rational $t = m2^{-n}$,

$$\mathcal{T}_t^n\phi \geq \mathcal{T}_t^{n+1}\phi \geq \dots \geq \mathcal{T}_t\phi.$$

By considering controls piecewise constant in time, one can show that $\mathcal{T}_t^n\phi \rightarrow \mathcal{T}_t\phi$ as $n \rightarrow \infty$, if t is dyadic rational. Choose n large enough such that $t = m2^{-n}$. Let $\tau_k = k2^{-n}$ and

$$\mathcal{A}_{nt} = \{\pi \in \mathcal{A}_t: \pi[U_\tau = U_{\tau_k} \text{ for } \tau \in [\tau_k, \tau_{k+1}), k = 0, 1, \dots, m-1] = 1\}$$

By induction on m (for fixed n) and a construction like that in the proof of Theorem 4.1, it can be shown that

$$\mathcal{T}_t^n\phi(\mu) = \min_{\pi \in \mathcal{A}_{nt}} J(t, \mu, \pi, \phi).$$

By [3, Cor. 6.1], every $\pi \in \mathcal{A}_t$ is the limit of π_{nt} as $n \rightarrow \infty$, with $\pi_{nt} \in \mathcal{A}_{nt}$. Lemma 3.6 then implies that $\mathcal{T}_t^n\phi(\mu) \rightarrow \mathcal{T}_t\phi(\mu)$ as $n \rightarrow \infty$.

REFERENCES

- [1] A. BENSOUSSAN AND J. L. LIONS, *Applications des inéquations variationnelles en contrôle stochastique*, Dunod, Paris, 1978.
- [2] M. H. A. DAVIS, *Nonlinear semigroups in the control of partially-observable stochastic systems*, in Measure Theory and Applications to Stochastic Analysis, Lecture Notes in Mathematics, 695, Springer-Verlag, New York, 1978.
- [3] W. H. FLEMING AND E. PARDOUX, *Existence of optimal controls for partially observed diffusions*, this Journal, this issue, pp. 261–285.
- [4] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

- [5] W. H. FLEMING AND M. VIOT, *Some measure-valued Markov processes in population genetics theory*, Indiana Univ. Math. J., 28 (1979), pp. 817–844.
- [6] N. V. KRYLOV, *On passing to the limit in degenerate Bellman equations* I, II, Math. Sb., 106 (1978), pp. 214–233; 107 (1978), pp. 56–68. English tr. Math USSR Sb. 34 (1978); 35 (1979).
- [7] P. L. LIONS, *Control of diffusion processes in R^N* , C.R. Acad. Sci., Paris, 288 Ser. A (1979), p. 339.
- [8] ———, *Degenerate Hamilton–Jacobi–Bellman equations with Dirichlet boundary conditions*, C.R. Acad. Sci., Paris, 289 Ser. A (1979), p. 329.
- [9] M. NISIO, *On a nonlinear semigroup attached to optimal stochastic control*, Publ. RIMS Kyoto Univ. 13 (1976), pp. 513–537.

PARTIALLY OBSERVED DIFFUSIONS AND THEIR CONTROL*

JEAN-MICHEL BISMUT†

Abstract. This paper is an extension of a previous article of Fleming and Pardoux [SIAM J. Control Optim., 20 (1982), pp. 261–265] on the control of partially observed diffusions, when the control enters linearly into the controlled stochastic differential equation.

Introduction. The purpose of this paper is to give another approach to the problems of optimal stochastic control with incomplete observation considered by Fleming and Pardoux in [3] (this issue, pp. 261–265) and to give a theorem about the existence of optimal controls in the sense of [3].

Consider the system of stochastic differential equations

$$\begin{aligned} dx &= (f(t, x, y) + g(t, x, y)u_t) dt + X_1(t, x, y) dw^1 + \cdots + X_m(t, x, y) dw^m, \\ x(0) &= z, \\ dy &= C(t, x) dt + d\eta, \\ y(0) &= 0 \end{aligned} \tag{0.1}$$

where $w = (w^1, \dots, w^m)$, $\eta = (\eta^1, \dots, \eta^d)$ are independent Brownian motions, z, w, η are independent and u is a control which is assumed to be adapted to y at least in a generalized sense made precise by Fleming and Pardoux in [3]. The criterion to be minimized is of the form

$$E \left[\int_0^T L(t, x_t, u_t) dt + G(x_T) \right]. \tag{0.2}$$

In [3] the problem is solved in two cases:

- (a) when $L = 0$, with compactness methods on the probability laws in (0.1);
- (b) when L is not 0, when the probability law of z is smooth, with techniques which involve results on the partial differential equation of filtering in (0.1).

In this paper we will show how it is possible to work systematically with compactness properties of the probability measures associated to system (0.1); this is done by using a very general result of Stroock and Varadhan [6] (which is, in fact, partly used in [3]). In particular, any reference to the filtering equation is bypassed.

In § 1 the space of control laws is defined as in [3]. In § 2 the problem of optimal stochastic control is solved.

We essentially follow Fleming and Pardoux [3], with some technical improvements in their method. In particular, the key integration by parts argument is taken from [3]. Some care is given in giving a rigorous foundation to stochastic integral manipulations, which require in fact nontrivial results on the definition of stochastic integrals.

Contrary to the situation in our previous work [1], where the structure of the filtering equation was essential since the Girsanov transformation could be directly applied on the filtering equation itself, the relation between the filtering equation and the optimal control is not absolutely clear in this case.

1. The space of controls. $L_\infty(R^+; R^k)$ denotes the vector space of measurable essentially bounded functions defined on $(R^+, \mathcal{B}(R^+), dt)$ with values in R^k .

* Received by the editors April 13, 1981.

† Département de Mathématiques, Université Paris-Sud, 91405 Orsay, France.

$L_\infty(R^+; R^k)$ is the strong dual of $L_1(R^+; R^k)$, which is the set of integrable functions defined on $(R^+, \mathcal{B}(R^+), dt)$ with values in R^k .

U is a compact convex set in R^k .

To each $u \in L_\infty(R^+; R^k)$, we associate $h^u \in \mathcal{C}(R^+; R^k)$ defined by

$$(1.1) \quad h^u(t) = \int_0^t u_s ds.$$

It is clear that the mapping $u \rightarrow h^u$ is injective.

$L_\infty(R^+; U)$ is the set of $u \in L_\infty(R^+; R^k)$ such that $u_t \in U$ a.e. We endow $L_\infty(R^+; U)$ with the weak-star topology of $L_\infty(R^+; R^k)$. $L_\infty(R^+; U)$ is then a compact metrizable space. Moreover, if u_n is a sequence in $L_\infty(R^+; U)$, the convergence of the sequence u_n is equivalent to the uniform convergence on compact sets of h^{u_n} .

$\mathcal{C}^U(R^+; R^k)$ denotes the image of $L_\infty(R^+; U)$ by the mapping h . $\mathcal{C}^U(R^+; R^k)$ is a compact subspace of $\mathcal{C}(R^+; R^k)$.

In the sequel we will identify $\mathcal{C}(R^+; R^k) \times \mathcal{C}(R^+; R^d)$ and $\mathcal{C}(R^+; R^k \times R^d)$ without explicit mention.

DEFINITION 1.1. On the space $\mathcal{C}(R^+; R^k \times R^d)$, whose standard element is written (h, y) , with $h \in \mathcal{C}(R^+; R^k)$ and $y \in \mathcal{C}(R^+; R^d)$, for $t \geq 0$, the following σ -fields are defined:

$$(2.2) \quad F_t^h = \mathcal{B}(h_s; s \leq t), \quad F_t^y = \mathcal{B}(y_s; s \leq t), \quad F_t^{h,y} = \mathcal{B}(h_s, y_s; s \leq t).$$

DEFINITION 1.2. P^d denotes the Brownian measure on $\mathcal{C}(R^+; R^d)$, with $P^d(y_0 = 0) = 1$.

As in [3], we now define the set of admissible controls as a set of probability measures.

DEFINITION 1.3. Π denotes the set of probability measures P on $\mathcal{C}(R^+; R^k \times R^d)$ such that

- (a) $P(h \in \mathcal{C}^U(R^+; R^k)) = 1$;
- (b) y_t is a martingale with respect to the family of σ -fields $\{F_t^{h,y}\}_{t \geq 0}$;
- (c) The law of y under P is P^d .

The following result is elementary:

PROPOSITION 1.4. Π is a nonempty compact convex set of probability measures on $\mathcal{C}(R^+; R^k \times R^d)$.

Proof. The tightness of the probability measures in Π is trivial due to (c) in Definition 1.3 and the compactness of $\mathcal{C}^U(R^+; R^k)$. Π is trivially convex and closed. \square

As mentioned in [3], Π is the closure of the set Π^s of the probability measures P in Π such that, if $\{F_t^{y^{**}}\}_{t \geq 0}$ is the filtration on $\mathcal{C}(R^+; R^k \times R^d)$ obtained by completing $\{F_t^y\}_{t \geq 0}$ with the negligible sets in $F_t^{h,y}$ (for the measure P) and taking the right-continuous regularization of this new family of σ -fields as in [2], then h is adapted to $\{F_t^{y^{**}}\}_{t \geq 0}$. The set Π is then the closure of the set Π^s of natural control laws, i.e., the laws for which u is a nonanticipating function of y .

2. Optimal stochastic control. We first define the various functions appearing in system (0.1).

$a(t, x, y)$ is a function defined on $R^+ \times R^n \times R^d$ with values in the set of symmetric (n, n) matrices which are nonnegative.

$f(t, x, y)$ is a function defined on $R^+ \times R^n \times R^d$ with values in R^n .

$g(t, x, y)$ is a function defined on $R^+ \times R^n \times R^d$ with values in the set of (n, k) matrices.

We now make two types of assumptions on a, f, g which will be used in the sequel.

H1. (a) There exist functions $X_1(t, x, y), \dots, X_m(t, x, y)$, defined on $R^+ \times R^n \times R^d$ with values in R^n , which are bounded, measurable in t , uniformly Lipschitz in x and continuous in y such that

$$(2.1) \quad a^{ij}(t, x, y) = \sum_{k=1}^m (X_k^i X_k^j)(t, x, y).$$

(b) $f(t, x, y), g(t, x, y)$ are bounded, measurable in the variable t , uniformly Lipschitz in x and continuous in y .

A second set of assumptions is as follows:

H2. (a) $a(t, x, y)$ is bounded, continuous, and has positive definite values.

(b) $f(t, x, y)$ and $g(t, x, y)$ are bounded, measurable in (t) and continuous in (x, y) .

From now on we assume that H1 or H2 are satisfied by a, f, g (of course, they can be simultaneously satisfied).

(a) *The probability measures $P^{z, h, y}$.* We first define the solution of the first stochastic differential equation for a given trajectory of (u, y) .

Note that in the sequel, when $h \in \mathcal{C}^U(R^+; R^d)$, u_t is the element of $L_\infty(R^+; U)$ which is the a.e. derivative of h_t .

DEFINITION 2.1. For $(z, h, y) \in R^n \times \mathcal{C}^U(R^+; R^k) \times \mathcal{C}(R^+; R^d)$, $P^{z, h, y}$ is the unique probability measure on $\mathcal{C}(R^+; R^n)$ which is the solution of the martingale problem associated to $(x, f(t, x, y_t) + g(t, x, y_t)u_t, a(t, x, y_t))$ in the sense of Stroock and Varadhan [6], i.e., $P^{z, h, y}(x_0 = z) = 1$, and moreover, if $L_t^{u, y}$ is the differential operator

$$(2.2) \quad L_t^{u, y} = (f(t, x, y_t) + g(t, x, y_t)u_t)^k \frac{\partial}{\partial x^k} + \frac{1}{2} a^{ij}(t, x, y_t) \frac{\partial^2}{\partial x^i \partial x^j},$$

then, if B is a C^∞ function defined on R^n with values in R which has compact support,

$$(2.3) \quad B(x_t) - \int_0^t (L_s^{u, y} B)(x_s) ds$$

is a martingale with respect to the filtration $\{F_t^x\}_{t \geq 0}$ of $\mathcal{C}(R^+; R^n)$.

Note that by [6, Thms 6.3.4 and 7.2.1], under H1 or under H2, $P^{z, h, y}$ exists and is unique.

We now have the key result of [3]–[6].

THEOREM 2.2. $P^{z, h, y}$ depends continuously on $(z, h, y) \in R^n \times \mathcal{C}^U(R^+; R^k) \times \mathcal{C}(R^+; R^d)$.

Proof. Note that all the martingale problems which we consider are well-posed in the sense of [6]. Theorem 2.2 is then an immediate consequence of [3, Lemma A1] and [6, Thm. 11.3.4]. \square

(b) *The probability measures Q^P .* Ω is the probability space $\mathcal{C}(R^+; R^n \times R^k \times R^d)$. $\{G_t\}_{t \geq 0}$ is the filtration on Ω defined by

$$(2.4) \quad G_t = \mathcal{B}(x_s, h_s, y_s; s \leq t).$$

Note that, since there is no risk of confusion, we will still use the notation F_t^h, F_t^y, \dots to describe the filtration generated on Ω by the processes h, y, \dots .

μ is a fixed probability measure on R^n .

DEFINITION 2.3. Let P be a probability measure on $\mathcal{C}(R^+; R^k \times R^d)$, which belongs to Π . Q^P is the probability measure on Ω which is given by

$$(2.5) \quad dQ^P(x, h, y) = \int_{R^n} d\mu(z) dP(h, y) dP^{z, h, y}(x).$$

We now show some "obvious" properties of the probability measures Q^P . We choose a given P in Π . Under Q^P , y is a Brownian motion, and the law of x_0 is exactly μ . Since the conditional law of x on $\mathcal{C}(R^+; R^n)$, given by $x_0 = z, h, y$, is equal to $P^{z,h,y}$, if $H(t, x)$ is a C^∞ function defined on $R^+ \times R^n$ which has compact support, then

$$(2.6) \quad H(t, x_t) - \int_0^t (H_s(s, x_s) + L_s^{u,y} H(s, x_s)) ds$$

is a conditional martingale with respect to $\{F_t^x\}_{t \geq 0}$, and is then a martingale on Ω with respect to the filtration $\{G_t\}_{t \geq 0}$.

Moreover, it is clear that $P^{z,h,y}$, when restricted to F_t^x , only depends on the trajectory of (h, y) up to time t . It is then not hard to prove that under Q^P , y is still a martingale with respect to the filtration $\{G_t\}_{t \geq 0}$ and is, in fact, a Brownian martingale. Finally, since (2.6), given (h, y) , is a conditional martingale, it is easy to prove that (2.6) and y^1, \dots, y^d are orthogonal martingales in the sense of [4], i.e., the product of (2.6) and of any of the y^i is still a martingale.

This shows, in particular, that for the measure Q^P , in the sense of [6], (x_t, y_t) are Ito processes on Ω endowed with the filtration $\{G_t\}_{t \geq 0}$, whose characteristics in the sense of [6] are

$$(2.7) \quad \begin{pmatrix} a(t, x_t, y_t) & 0 \\ 0 & I \end{pmatrix}, \quad \begin{pmatrix} f(t, x_t, y_t) + g(t, x_t, y_t)u_t \\ 0 \end{pmatrix}.$$

(c) *A few remarks on stochastic integrals.* $C(t, x)$ is a function defined on $R^+ \times R^n$ with values in R^d , which is bounded, continuous, once differentiable in the variable t , twice differentiable in the variable x and whose differentials are bounded and continuous.

P is a fixed probability measure in Π . Since for the probability measure Q^P on Ω , (x, y) are Ito processes whose characteristics are given by (2.7), we may use the standard Ito calculus to obtain

$$(2.8) \quad \int_0^T \langle C(s, x_s), dy_s \rangle = \langle C(T, x_T), y_T \rangle - \int_0^T \left\langle y_s, C_x(s, x_s) dx_s \right. \\ \left. + \left(\frac{1}{2} a^{ij}(s, x_s, y_s) C_{x^i x^j}(s, x_s) + C_s(s, x_s) \right) ds \right\rangle.$$

Now, by the results given in [5]–[7], we know that, if $T \in R^+$, there exists a universally measurable function $J_T(x, y)$ on $\mathcal{C}(R^+; R^n \times R^d)$ such that, if Q is any probability law on $\mathcal{C}(R^+; R^n \times R^d)$ for which x is a semi-martingale, $J_T(x, y)$ is Q a.e. equal to the stochastic integral $\int_0^T \langle y_s, C_x(s, x_s) dx_s \rangle$.

Now under Q^P , x is a semi-martingale on Ω with respect to $\{G_t^+\}_{t \geq 0}$. Using the result of Stricker [2, § VII–60], that x is a $\{F_t^{y,x^+}\}_{t \geq 0}$ semi-martingale, and the result of Meyer [2, VIII–13] on the invariance of the stochastic integral when restricting the filtration, we find that, Q^P a.s.,

$$(2.9) \quad \int_0^T \langle C(s, x_s), dy_s \rangle = \langle C(T, x_T), y_T \rangle - J_T(x, y) \\ - \int_0^T \left\langle \frac{1}{2} a^{ij}(s, x_s, y_s) C_{x^i x^j}(s, x_s) + C_s(s, x_s), y_s \right\rangle ds.$$

(d) *The probability measure Q'^P .* We are now able to define completely the probability law of the system described in (0.1). In fact, starting from measure Q^P , we will do a Girsanov transformation [6, § 4] to change the drift of y .

DEFINITION 2.4. Let $T \in R^+$. $Z_T(x, y)$ is the universally measurable random variable defined on $\mathcal{C}(R^+; R^n \times R^d)$ by

$$(2.10) \quad Z_T(x, y) = \exp \left\{ \langle C(T, x_T), y_T \rangle - J_T(x, y) - \frac{1}{2} \int_0^T (y_s^l a^{ij}(s, x_s, y_s) C_{x^l x^j}^l(s, x_s) + |C(s, x_s)|^2) ds - \int_0^T y_s^l C_s^l(s, x_s) ds \right\}.$$

We finally define the law of (h, y, x) for a given P .

DEFINITION 2.5. Let $P \in \Pi$. Q'^P is the probability measure on $\mathcal{C}(R^+; R^n \times R^k \times R^d)$ given by

$$(2.11) \quad dQ'^P(x, h, y) = Z_T(x, y) dQ^P(x, h, y).$$

Of course, due to (2.9), Z_T is for Q^P an exponential martingale (in T), and since C is bounded, Q'^P is a probability measure. Moreover, under Q'^P , (x_t, y_t) are Ito processes on $(\Omega, \{G_t\}_{t \geq 0})$ whose characteristics are

$$(2.12) \quad \begin{pmatrix} a(t, x_t, y_t) & 0 \\ 0 & I \end{pmatrix}, \quad \begin{pmatrix} f(t, x_t, y_t) + g(t, x_t, y_t) u_t \\ 1_{t \leq T} C(t, x_t) \end{pmatrix};$$

this is by a well-known result on the Girsanov transformation [6, § 4].

Now, using (2.5) we have

$$(2.13) \quad dQ'^P(x, h, y) = \int_{R^n} d\mu(z) dP(h, y) Z_T(x, y) dP^{z, h, y}(x).$$

By Fubini's theorem, to integrate with respect to Q'^P we first fix (z, h, y) , integrate with respect to $Z_T(x, y) dP^{z, h, y}(x)$ and then integrate with respect to (z, h, y) .

We now claim

PROPOSITION 2.6. For any (z, h, y) in $R^n \times \mathcal{C}^U(R^+; R^k) \times \mathcal{C}(R^+; R^d)$, when the space $\mathcal{C}(R^+; R^n)$ is endowed with the probability measure $P^{z, h, y}$, the random variable $J_T(x, y)$ is $P^{z, h, y}$ a.s. equal to the stochastic integral

$$(2.14) \quad J_T(x, y) = \int_0^T \langle y_s, C_x(s, x_s) dx_s \rangle.$$

Proof. Let S be the probability measure on $\mathcal{C}(R^+; R^d)$ which is the Dirac measure at y . Then, under the probability measure $dS(\bar{y}) dP^{z, h, \bar{y}}(x)$ on $\mathcal{C}(R^+; R^n \times R^d)$, x_t is a semi-martingale with respect to the filtration $\{F_t^{x, y}\}_{t \geq 0}$. We then know that $J_T(x, y)$ is a version of the stochastic integral on the r.h.s. of (2.14). \square

(e) *The probability measure $R^{z, h, y}$.* As in [3], we introduce the probability measures $R^{z, h, y}$ as a technical tool.

Let $\sigma(t, x, y)$ be the positive square root of $a(t, x, y)$. Using [6, Thm. 4.5.2], we know that for any $(z, h, y) \in R^n \times \mathcal{C}^U(R^+; R^k) \times \mathcal{C}(R^+; R^d)$, there exists a Brownian motion w (on a possibly enlarged probability space) such that, for the probability measure $P^{z, h, y}$, the semi-martingale x_t is such that

$$(2.15) \quad dx = (f(t, x_t, y_t) + g(t, x_t, y_t) u_t) dt + \sigma(t, x_t, y_t) dw.$$

We now proceed as in [3]. Let Z_T^{iy} be the random variable defined on $(\mathcal{C}(R^+; R^n), P^{z,h,y})$ by

$$(2.16) \quad Z_T^{iy}(x) = \exp \left\{ - \int_0^T \langle y_s, C_x(s, x_s) \sigma(s, x_s, y_s) dw \rangle - \frac{1}{2} \int_0^T \langle a(s, x_s, y_s) C_x^*(s, x_s) y_s, C_x^*(s, x_s) y_s \rangle ds \right\}.$$

Now, since C_x is bounded and y is continuous, by [6, Thm. 6.4.3], we know that $Z_T^{iy} dP^{z,h,y}$ is a probability measure on $\mathcal{C}(R^+; R^n)$, which we call $R^{z,h,y}$. We now have the elementary

PROPOSITION 2.7. *The measure $R^{z,h,y}$ is the probability measure on $\mathcal{C}(R^+; R^n)$ which is the unique solution of the martingale problem associated to $(x, f(t, x, y_t) + g(t, x, y_t)u_t - a(t, x, y_t)C_x^*(t, x)y_t, a(t, x, y_t))$. $R^{z,h,y}$ depends continuously on $(z, h, y) \in R^n \times \mathcal{C}^U(R^+; R^k) \times \mathcal{C}(R^+; R^d)$.*

Proof. Since $P^{z,h,y}$ is the solution of the martingale problem defined in Definition 2.1, by a standard use of the Girsanov transformation [6, § 4], the first result is easily proved. Note that we are again under the conditions of uniqueness of the solution of the martingale problem. Finally, the continuous dependence of $R^{z,h,y}$ is a consequence of [3, Lemma A1] and [6, Thm. 11.3.4]. \square

(f) *Continuous dependence of Q'^P on P .* We now prove that Q'^P depends continuously on $P \in \Pi$.

Let $V_T(x, h, y)$ be the continuous function on $\mathcal{C}(R^+; R^n) \times \mathcal{C}^U(R^+; R^k) \times \mathcal{C}(R^+; R^d)$

$$V_T(x, h, y)$$

$$= \exp \left\{ \langle C(T, x_T), y_T \rangle + \int_0^T \left(-\frac{1}{2} a^{ij}(s, x_s, y_s) C_{x^i x^j}^l(s, x_s) y_s^l - y_s^l C_s^{l i}(s, x_s) \right. \right. \\ \left. \left. - \frac{1}{2} |C(s, x_s)|^2 - \langle y_s, C_x(s, x_s)(f(s, x_s, y_s) + g(s, x_s, y_s)u_s) \rangle \right. \right. \\ \left. \left. + \frac{1}{2} \langle a(s, x_s, y_s) C_x^*(s, x_s) y_s, C_x^*(s, x_s) y_s \rangle \right) ds \right\}.$$

We now have the essential result.

THEOREM 2.8. *For $T \in R^+$, the restriction of Q'^P to G_T depends continuously on $P \in \Pi$.*

Proof. Using (2.13), Proposition 2.6 and (2.16), it is clear that for $P \in \pi$

$$(2.17) \quad dQ'^P(x, h, y) = \int_{R^n} d\mu(z) dP(h, y) V_T(x, h, y) dR^{z,h,y}(x).$$

Let $W(x, h, y)$ be a bounded continuous function on $\mathcal{C}([0, T]; R^n \times R^k \times R^d)$. We must prove that

$$(2.18) \quad P \in \Pi \rightarrow \int W(x, h, y) dQ'^P(x, h, y)$$

is continuous. Let $r_k(v)$ be a bounded continuous function on R with values in $[0, 1]$,

which is equal to 1 on $\{|v| \leq k\}$ and to 0 for $|v| \geq k+1$. We claim that

$$(2.19) \quad P \in \Pi \rightarrow \int W(x, h, y) r_k(\sup_{s \leq T} |y_s|) dQ'^P(x, h, y)$$

is continuous. To prove (2.19), it suffices to show that the mapping

$$(2.20) \quad (z, h, y) \in R^n \times \mathcal{C}^U(R^+; R^k) \times \mathcal{C}(R^+; R^d) \\ \rightarrow \int r_k(\sup_{s \leq T} |y_s|) W(x, h, y) V_T(x, h, y) dR^{z, h, y}(x)$$

is bounded and continuous. Now, when $\sup_{s \leq T} |y_s|$ is uniformly bounded, $V_T(x, h; y)$ is also uniformly bounded. The function $r_k(\sup_{s \leq T} |y_s|) W(x, h, y) V_T(x, h, y)$ being bounded and continuous, the continuity of (2.20) follows easily from the continuity of $R^{z, h, y}$ and Prokhorov's theorem.

To prove the continuity of (2.18), it suffices to show that when $k \rightarrow +\infty$, (2.19) converges uniformly to (2.18) on Π . By (2.12) we know that, under Q'^P , $y_t - \int_0^t C(s, x_s) ds$ is a Brownian martingale. If M is a positive constant such that, for any (s, x) , $|C(s, x)| \leq M$, we have

$$(2.21) \quad \left| \int W(x, y, h) (1 - h_k(\sup_{s \leq T} |y_s|)) dQ'^P \right| \leq C Q'^P(\sup_{s \leq T} |y_s| \geq k) \\ \leq C P^d(\sup_{s \leq T} |y_s| \geq k - MT).$$

Since the r.h.s. of (2.21) tends to 0 when $k \rightarrow +\infty$, the l.h.s. of (2.21) tends to 0 uniformly in $P \in \Pi$. The theorem is proved. \square

(g) *Optimal stochastic control.* We will now solve an optimal stochastic control problem. Let $L(t, x, y, u)$ be a positive continuous function defined on $R^+ \times R^n \times R^d \times U$, which is convex in the variable u . We now have the following elementary result:

PROPOSITION 2.9. *The mapping $(x, h, y) \in \mathcal{C}(R^+; R^n) \times \mathcal{C}^U(R^+; R^k) \times \mathcal{C}(R^+; R^d) \rightarrow \int_0^T L(t, x_t, y_t, u_t) dt$ is l.s.c.*

Proof. Let (x^n, y^n, u^n) be a sequence converging to (x, y, u) . By uniform continuity it is clear that the sequence of continuous functions $(t, v) \in [0, T] \times U \rightarrow L(t, x_t^n, y_t^n, v)$ converges uniformly and boundedly to the function $(t, v) \rightarrow L(t, x_t, y_t, v)$. Now using the convexity of L in the variable u , it is easy to see that

$$(2.22) \quad \int_0^T L(t, x_t, y_t, u_t) dt \leq \liminf \int_0^T L(t, x_t, y_t, u_t^n) dt.$$

The proposition follows. \square

G is a lower bounded continuous function defined on $R^n \times R^d$ with values in R .

We now have the existence result of an optimal control law:

THEOREM 2.10. *The functional*

$$(2.23) \quad P \in \Pi \rightarrow \left[\int_0^T L(t, x_t, y_t, u_t) dt + G(x_T, y_T) \right] dQ'^P(x, h, y)$$

has a minimum on Π .

Proof. From Theorem 2.8 and Proposition 2.9, (2.23) is l.s.c. on Π . By Proposition 1.4 Π is compact. The theorem is proved. \square

Remark. The result of Theorem 2.10 is slightly stronger than existence theorems proved in [3]. The method in [3] involved a “separated” control problem, which is avoided here. See [3, § 4, § 7].

The proof of Theorem 2.10 gives the following slightly more general result:

THEOREM 2.11. *Let $(x, h, y) \rightarrow \Phi(x, h, y)$ be l.s.c. on $\mathcal{C}(R^+; R^n) \times \mathcal{C}^U(R^+; R^k) \times \mathcal{C}(R^+; R^d)$ with $\Phi \geq 0$. Then $P \in \Pi \rightarrow \int \Phi(x, h, y) dQ^P(x, h, y)$ has a minimum on Π .*

Acknowledgment. The author is much indebted to Professor W. Fleming for very helpful discussions and encouragement.

REFERENCES

- [1] J. M. BISMUT, *Un problème de contrôle stochastique avec observation partielle*, Z. Wahrsch. Verw. Gebiete, 49 (1979), pp. 63–95.
- [2] C. DELLACHERIE AND P. A. MEYER, *Probabilités et Potentiels*, 2 éd., Chap. I–IV, Hermann, Paris 1975, Chap. V–VIII, Hermann, Paris 1980.
- [3] W. H. FLEMING AND E. PARDOUX, *Optimal control for partially observed diffusions*, this Journal, this issue, pp. 261–285.
- [4] P. A. MEYER, *Cours sur les intégrales stochastiques*, Séminaire de Probabilités X, Lecture Notes in Mathematics 511, Springer, Berlin–Heidelberg–New York, 1976, pp. 245–400.
- [5] C. DOLEANS-DADE, *Intégrales stochastiques par rapport à une famille de probabilités*, Séminaire de Probabilités V, Lecture Notes in Mathematics 191, Springer, Berlin–Heidelberg–New York, 1971, pp. 141–146.
- [6] D. W. STROOCK AND S. R. S. VARADHAN, *Multidimensional Diffusion Processes*, Grundlehren der Mathematischen Wissenschaften 233, Springer, Berlin–Heidelberg–New York, 1979.
- [7] P. A. MEYER, *Sur un théorème de C. Stricker*, Séminaire de Probabilités 11, Lecture Notes in Mathematics 581, Springer, Berlin–Heidelberg–New York, 1977, pp. 482–489.

NONLINEAR PERTURBATIONS OF CONTROL-SEMILINEAR CONTROL SYSTEMS*

KEVIN A. GRASSE†

Abstract. We consider the class of nonlinear, autonomous control systems on a differentiable manifold that depend semilinearly on the control variable. This class includes the much-studied class of control-linear systems. For such control-semilinear systems, we prove that global controllability is preserved under nonlinear perturbations satisfying a mild boundedness condition. Our techniques enable us to obtain controllability results for piecewise-constant and smooth controls, as well as for measurable controls.

1. Introduction. Our purpose in this paper is to study the effect of “reasonable” perturbations on global controllability properties of a certain class of nonlinear control systems. In a previous paper [4] the author gave a sufficient condition for the property of controllability to a compact set on a fixed time interval to be preserved under small perturbations. This condition was formulated for arbitrary nonlinear, nonautonomous control systems. By restricting the class of control systems under consideration, one can naturally expect to obtain stronger results. We shall do this here by focusing our attention on nonlinear, autonomous control systems that are “semilinear” in the control variable (precise definitions will follow). Our results will show that, for such systems, *global* controllability is preserved under nonlinear perturbations that satisfy a rather mild boundedness condition.

The problems considered here are motivated by and related to the work of Brunovsky and Lobry [1]. In fact, the connection between our work and theirs is sufficiently close to merit a precise delineation of how the results presented here complement and/or improve the results of [1]. Consequently, we begin by giving a brief summary of some of the work of Brunovsky and Lobry.

Let M be a connected C^∞ manifold of dimension n and let $V^\infty(M)$ denote the Lie algebra of all C^∞ vector fields on M under the usual bracket multiplication. A finite subset $\{\xi_1, \dots, \xi_m\}$ of $V^\infty(M)$ induces a control-linear control vector field (or control system) on M ,

$$(1) \quad \dot{x} = \sum_{i=1}^m u_i(t) \xi_i(x),$$

where the u_i are bounded measurable functions of \mathbb{R} into \mathbb{R} . Let L denote the subalgebra of $V^\infty(M)$ generated by the vector fields ξ_1, \dots, ξ_m , and let

$$L(x) = \{\rho(x) | \rho \in L\}$$

for each x in M . The set $L(x)$ is clearly a vector subspace of $T_x M$, the tangent space to M at x . As is well known, the condition $\dim L(x) = n$ for every x in M is sufficient for the global controllability of (1) from every x in M [5], [6]. Furthermore, this condition is also necessary for global controllability in the real-analytic case [11], although it is not necessary in the C^∞ case. In [1], Brunovsky and Lobry prove, or at least develop the techniques to prove, the following results (see also [7], [8]).

* Received by the editors November 25, 1980.

† Department of Mathematics, University of Oklahoma, Norman, Oklahoma 73019.

THEOREM. *If $\dim L(x) = n$ for every x in M and if ρ is any C^∞ vector field on M , then the control vector field*

$$\dot{x} = \rho(x) + \sum_{i=1}^m u_i(t) \xi_i(x)$$

is globally controllable from every x in M .

THEOREM. *If $\dim L(x) = n$ for every x in M and if $g : M \times \mathbb{R}^m \rightarrow TM$ is a C^∞ control vector field on M that is globally bounded on $M \times \mathbb{R}^m$ (e.g., with respect to a Riemannian metric on M), then the control vector field*

$$\dot{x} = g(x, u(t)) + \sum_{i=1}^m u_i(t) \xi_i(x)$$

is globally controllable from every x in M , where $u(t) = (u_1(t), \dots, u_m(t))$.

We seek to improve these results in several ways. First, we will work with C^1 mappings, as opposed to C^∞ mappings. It is important to note that the bracket multiplication does not yield a Lie-algebra structure on the set of all C^1 vector fields on M because the bracket of two C^1 vector fields is only of class C^0 in general. Hence, we must reformulate the hypothesis in the above results on the dimension of the Lie algebra L . In our treatment we will simply replace this hypothesis by the assumption that (1) is globally controllable from every x in M . Since the hypothesis on the dimension of the Lie algebra implies global controllability, but not conversely, this will, in effect, weaken the assumptions in the above results.

Second, instead of considering the control-linear system (1), we will prove the above results for the more general *control-semilinear* system,

$$(2) \quad \dot{x} = \sum_{i=1}^m u_i(t) \xi_i(x, v(t)),$$

where the u_i are as above, v is a bounded measurable mappint of \mathbb{R} into \mathbb{R}^p and $\xi_i : M \times \mathbb{R}^p \rightarrow TM$ is a C^1 control vector field on M with control space \mathbb{R}^p for each $i = 1, \dots, m$. Thus, the control present in the system is comprised of two components $(v(t), u(t))$, and we require the system to be linear in the second component. This enables us to handle at least some degree of nonlinearity in the control variable.

While these remarks indicate some improvement of certain aspects of the results in [1], we must stress that our work cannot be regarded as a complete generalization of [1], because there the authors consider controllability by a rather restricted set of controls. Our only efforts in this direction will be to show that the perturbed system is controllable by piecewise-constant controls and, in the control-linear case, by C^∞ controls.

We will employ two major tools in this paper: a perturbation theorem proved previously by the author [4, Thm. 4.9] and the notion of *normal reachability* introduced by H. Sussmann [13]. The author's theorem assumes the existence of sufficiently many so-called normal values, and there is a close connection between normal values and normal reachability in control-semilinear systems, as we will show.

The development of the results in this paper will be carried out within the notational framework of [4]. Consequently, we will assume a familiarity with the contents of this reference and simply provide a brief resumé of some essential definitions here. It should be noted that the results of [4] applied to nonautonomous systems, so the notation was designed to keep track of the initial time. Here, we will consider autonomous systems exclusively, where the initial time is usually taken to be

zero. Thus, in using the notation of [4] we will be constantly reminding ourselves that the initial time is zero. While this may seem slightly redundant, we felt it preferable to developing a new notational scheme for the autonomous case.

2. Preliminaries. Let M denote a finite-dimensional, second-countable, connected, Hausdorff differentiable manifold of class C^k with $k \geq 2$ and set $n = \dim M$. These assumptions imply, in particular, that M is a paracompact and metrizable topological space. Let TM denote the tangent bundle and $\pi : TM \rightarrow M$ the canonical projection. We recall that TM is a differentiable manifold of class C^{k-1} and π is a C^{k-1} submersion. It will be convenient to fix, once and for all, a metric d on M compatible with the manifold topology and a *Finsler structure* $\omega : TM \rightarrow \mathbb{R}$ [4, Def. 4.6].

For $l \in \mathbb{N}$, we let L_∞^l denote the Banach space of (equivalence classes of) essentially bounded, measurable mappings of \mathbb{R} into \mathbb{R}^l . The essential-supremum norm is denoted by $\|\cdot\|_\infty$, and elements of L_∞^l are referred to as *controls*. We will also have occasion to make use of the vector subspaces P_∞^l and D_∞^l of L_∞^l defined by

$$P_\infty^l = \{u \in L_\infty^l \mid u \text{ is piecewise constant on every compact subinterval of } \mathbb{R}\}$$

and

$$D_\infty^l = \{u \in L_\infty^l \mid u \text{ is of class } C^\infty \text{ on } \mathbb{R}\}.$$

We make the obvious identification $L_\infty^{p+m} = L_\infty^p \times L_\infty^m$ for $p, m \in \mathbb{N}$.

A C^1 (autonomous) *control vector field* on M with *control space* \mathbb{R}^l is a C^1 mapping $\xi : M \times \mathbb{R}^l \rightarrow TM$ such that $(\pi \circ \xi)(x, w) = x$ for every (x, w) in $M \times \mathbb{R}^l$. For (s, x, u) in $\mathbb{R} \times M \times L_\infty^l$, we let $J(s, x, u)$ denote the maximal subinterval of \mathbb{R} containing s on which a solution to the initial-value problem

$$\dot{\sigma}(t) = \xi(\sigma(t), u(t)), \quad \sigma(s) = x,$$

can be defined. This maximal solution is denoted by

$$\mu_{(s,x,u)} : J(s, x, u) \rightarrow M$$

and is called the *maximal response of ξ with initial condition (s, x) corresponding to the control u* . Note that the interval $J(s, x, u)$ is necessarily open in \mathbb{R} .

DEFINITION 2.1. If $\xi : M \times \mathbb{R}^l \rightarrow TM$ is a C^1 control vector field on M , then we let $\mathcal{D}(\xi)$ denote the subset of $\mathbb{R} \times \mathbb{R} \times M \times L_\infty^l$ given by

$$\mathcal{D}(\xi) = \{(t, s, x, u) \in \mathbb{R} \times \mathbb{R} \times M \times L_\infty^l \mid t \in J(s, x, u)\},$$

and we define a mapping $\mu : \mathcal{D}(\xi) \rightarrow M$ by

$$\mu(t, s, x, u) = \mu_{(s,x,u)}(t).$$

The mapping μ is called the *global flow* of ξ .

As was pointed out in [4], the set $\mathcal{D}(\xi)$ is open in $\mathbb{R} \times \mathbb{R} \times M \times L_\infty^l$, and the mapping μ is continuous.

The following two propositions contain some elementary properties of the global flow μ of a C^1 control vector field $\xi : M \times \mathbb{R}^l \rightarrow TM$. The easy proofs are omitted. We preface these results with one piece of notation. If $r \in \mathbb{R}$ and $u \in L_\infty^l$, then $u_r \in L_\infty^l$ denotes the control defined by $u_r(t) = u(t+r)$.

PROPOSITION 2.2. Let $(s, x, u) \in \mathbb{R} \times \mathbb{R} \times M \times L_\infty^l$, and let $r \in \mathbb{R}$. Then

$$J(s+r, x, u_{-r}) = J(s, x, u) + r$$

and

$$\mu(t+r, s+r, x, u_{-r}) = \mu(t, s, x, u)$$

for every t in $J(s, x, u)$. In particular, if u is a constant control, then

$$\mu(t+r, s+r, x, u) = \mu(t, s, x, u)$$

for every t in $J(s, x, u)$.

PROPOSITION 2.3. Let $s_0 < s_1 < s_2$ be real numbers, let $w_1, w_2 \in \mathbb{R}^l$, and define controls u_1, u_2 in P_∞^l by

$$u_1(t) = \begin{cases} w_1, & s_0 \leq t \leq s_1, \\ 0 & \text{otherwise,} \end{cases}$$

$$u_2(t) = \begin{cases} w_2, & s_1 \leq t \leq s_2, \\ 0 & \text{otherwise.} \end{cases}$$

If $(s_1, s_0, x, u_1) \in \mathcal{D}(\xi)$ and $(s_2, s_1, \mu(s_1, s_0, x, u_1), u_2) \in \mathcal{D}(\xi)$, then $(s_2, s_0, x, u_1 + u_2) \in \mathcal{D}(\xi)$ and

$$\mu(s_2, s_0, x, u_1 + u_2) = \mu(s_2, s_1, \mu(s_1, s_0, x, u_1), u_2).$$

There is an obvious analogue of Proposition 2.3 in the case where we have real numbers $s_0 < s_1 < \dots < s_q$ and points w_1, \dots, w_q in \mathbb{R}^l .

Notation. Let $\xi: M \times \mathbb{R}^l \rightarrow TM$ be a C^1 control vector field with global flow $\mu: \mathcal{D}(\xi) \rightarrow M$. For each (t, s, x) in $\mathbb{R} \times \mathbb{R} \times M$, we let $\mathcal{D}_{(t,s,x)}(\xi)$ denote the subset of L_∞^l defined by

$$\mathcal{D}_{(t,s,x)}(\xi) = \{u \in L_\infty^l \mid (t, s, x, u) \in \mathcal{D}(\xi)\},$$

and if $\mathcal{D}_{(t,s,x)}(\xi)$ is nonempty, we let $\mu_{(t,s,x)}: \mathcal{D}_{(t,s,x)}(\xi) \rightarrow M$ denote the mapping defined by

$$\mu_{(t,s,x)}(u) = \mu(t, s, x, u).$$

We remark that $\mathcal{D}_{(t,s,x)}(\xi)$ is open in L_∞^l , and $\mu_{(t,s,x)}$ is of class C^1 (see [4]).

DEFINITION 2.4. Let $\xi: M \times \mathbb{R}^l \rightarrow TM$ be a C^1 control vector field on M with global flow $\mu: \mathcal{D}(\xi) \rightarrow M$, let $(t, t_0, x_0) \in \mathbb{R} \times \mathbb{R} \times M$ with $t \geq t_0$, and let \mathcal{U} be a subset of L_∞^l . The attainable set of ξ from (t_0, x_0) at time t with controls in \mathcal{U} is defined by

$$\mathcal{A}_\xi(t_0, x_0; t \mid \mathcal{U}) = \mu_{(t,t_0,x_0)}(\mathcal{D}_{(t,t_0,x_0)}(\xi) \cap \mathcal{U}).$$

The attainable set of ξ from (t_0, x_0) with controls in \mathcal{U} with unspecified endtime is defined by

$$\mathcal{A}_\xi(t_0, x_0 \mid \mathcal{U}) = \bigcup_{t \geq t_0} \mathcal{A}_\xi(t_0, x_0; t \mid \mathcal{U}).$$

Remark 2.5. A subset $\mathcal{U} \subseteq L_\infty^l$ is said to be invariant under a shift in parameter if, for every $r \in \mathbb{R}$ and $u \in \mathcal{U}$, we have $u_r \in \mathcal{U}$ (recall that $u_r(t) = u(t+r)$). Clearly P_∞^l , D_∞^l and L_∞^l are all invariant under a shift in parameter. If \mathcal{U} is invariant under a shift in parameter, then Proposition 2.2 immediately yields the relations

$$\mathcal{A}_\xi(t_0, x_0; t \mid \mathcal{U}) = \mathcal{A}_\xi(0; x_0; t - t_0 \mid \mathcal{U})$$

and

$$\mathcal{A}_\xi(t_0, x_0 \mid \mathcal{U}) = \mathcal{A}_\xi(0, x_0 \mid \mathcal{U}).$$

Hence, for such families of control \mathcal{U} , there is no loss of generality in taking the initial time to be zero.

Notation. We set

$$\mathcal{A}_\xi(x_0; t \mid \mathcal{U}) = \mathcal{A}_\xi(0, x_0; t \mid \mathcal{U})$$

for $t \geq 0$ and

$$\mathcal{A}_\xi(x_0|\mathcal{U}) = \mathcal{A}_\xi(0, x_0|\mathcal{U}).$$

To further simplify the notation when using the families of controls L_∞^l and P_∞^l , we set

$$\begin{aligned} \mathcal{A}_\xi(x_0; t) &= \mathcal{A}_\xi(x_0; t|L_\infty^l), & A_\xi(x_0; t) &= \mathcal{A}_\xi(x_0; t|P_\infty^l), \\ \mathcal{A}_\xi(x_0) &= \mathcal{A}_\xi(x_0|L_\infty^l), & A_\xi(x_0) &= \mathcal{A}_\xi(x_0|P_\infty^l). \end{aligned}$$

In other words, if we do not specify the initial time, then it is zero; a script \mathcal{A} with no family of controls specified means we are using the full set of controls L_∞^l ; a nonscript A means we are using the set of piecewise-constant controls P_∞^l .

Remark 2.6. Since $P_\infty^l \subseteq L_\infty^l$, it is clear that we have the inclusions

$$A_\xi(x_0; t) \subseteq \mathcal{A}_\xi(x_0; t), \quad A_\xi(x_0) \subseteq \mathcal{A}_\xi(x_0).$$

It is also true that these inclusions are dense, although this particular fact will not be of essential importance here. In the next section we will give a sufficient condition for having equality of the sets $A_\xi(x_0)$ and $\mathcal{A}_\xi(x_0)$.

DEFINITION 2.7. Let $\xi: M \times \mathbb{R}^l \rightarrow TM$ be a C^1 control vector field on M , and let \mathcal{U} be a subset of L_∞^l . We say that ξ is *completely controllable by controls in \mathcal{U}* if $\mathcal{A}(x|\mathcal{U}) = M$ for every x in M .

We now introduce a control vector field of a more specialized form.

DEFINITION 2.8. Let m be a positive integer, and let p be a nonnegative integer. A C^1 control vector field

$$\xi: M \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow TM$$

is said to be *control semilinear* if there exist C^1 control vector fields $\xi_i: M \times \mathbb{R}^p \rightarrow TM$ for $i = 1, \dots, m$ such that

$$\xi(x, z, w) = \sum_{i=1}^m w_i \xi_i(x, z)$$

for all (x, z, w) in $M \times \mathbb{R}^p \times \mathbb{R}^m$, where $w = (w_1, \dots, w_m)$.

Remark 2.9. An important special case of this definition occurs when the ξ_i are (uncontrolled) vector fields $\xi_i: M \rightarrow TM$. Then

$$\xi(x, w) = \sum_{i=1}^m w_i \xi_i(x)$$

is referred to as a *control-linear* control vector field.

For $\alpha > 0$ and $u \in L_\infty^l$, we let $u^\alpha \in L_\infty^l$ denote the control defined by $u^\alpha(t) = u(t/\alpha)$ for each t in \mathbb{R} . The proofs of the following proposition and corollary are routine.

PROPOSITION 2.10. Let $\xi: M \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow TM$ be a C^1 control-semilinear control vector field on M with global flow $\mu: \mathcal{D}(\xi) \rightarrow M$. Let $(r, 0, x, v, u)$ be an element of $\mathcal{D}(\xi)$ with $r > 0$. Then for every $\alpha > 0$ we have

$$(\alpha r, 0, x, v^\alpha, (1/\alpha)u^\alpha) \in \mathcal{D}(\xi)$$

and

$$\mu(\alpha r, 0, x, v^\alpha, (1/\alpha)u^\alpha) = \mu(r, 0, x, v, u).$$

In particular, if v and u are constant controls, then

$$\mu(\alpha r, 0, x, v, (1/\alpha)u) = \mu(r, 0, x, v, u).$$

COROLLARY 2.11. *Let $\xi: M \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow TM$ be a C^1 control-semilinear control vector field on M with global flow $\mu: \mathcal{D}(\xi) \rightarrow M$. Let \mathcal{U} be a subset of $L_\infty^{p+m} = L_\infty^p \times L_\infty^m$ such that $(v, u) \in \mathcal{U}$, $\alpha > 0$, and $r \in \mathbb{R}$ imply $(v^\alpha, ru^\alpha) \in \mathcal{U}$. Then, for each x in M and $t > 0$, we have*

$$\mathcal{A}_\xi(x; t | \mathcal{U}) = \mathcal{A}_\xi(x | \mathcal{U}).$$

The condition on the set of controls \mathcal{U} in Corollary 2.11 is somewhat technical, but is certainly satisfied if \mathcal{U} is L_∞^{p+m} , P_∞^{p+m} or D_∞^{p+m} , which are the three cases of interest.

3. Families of vector fields. Recall that a C^1 vector field on M is a C^1 mapping $X: M \rightarrow TM$ such that $\pi \circ X$ is the identity mapping on M . For each x in M , we let $t \rightarrow X_t(x)$ denote the maximal integral curve of X passing through x at time $t = 0$. The mapping $(t, x) \rightarrow X_t(x)$, which is called the *global flow* of X , is defined on an open subset of $\mathbb{R} \times M$ and is of class C^1 .

DEFINITION 3.1. Let S be family of C^1 vector fields on M . For x in M and $t \geq 0$, we define the *attainable set of S from x at time t* by

$$A_S(x; t) = \{X_{t_k}^k \circ \cdots \circ X_{t_1}^1(x) | k \in \mathbb{N}, X^1, \dots, X^k \in S, \\ \text{each } t_i \text{ is nonnegative, } \sum_{i=1}^k t_i = t, \text{ and the expression is defined}\}.$$

We define the *attainable set of S from x with unspecified endtime* by

$$A_S(x) = \bigcup_{t \geq 0} A_S(x; t).$$

DEFINITION 3.2. A family S of C^1 vector fields on M is called *symmetric* if $X \in S$ implies $-X \in S$.

An elementary consequence of symmetry is that we can, in effect, move both forward and backward in time along the integral curves of the vector fields in S . This follows because moving backward along an integral curve of $X \in S$ is equivalent to moving forward along an integral curve of $-X$. A deeper consequence of symmetry is contained in the following theorem, which was proved independently and simultaneously by Peter Stefan and Héctor Sussmann.

THEOREM 3.3 [9], [12]. *Let S be a symmetric family of C^1 vector fields on the manifold M . Then the family of attainable sets $\{A_S(x) | x \in M\}$ forms a partition of M and each attainable set has the structure of a C^1 immersed submanifold of M . Moreover, for each x in M we have*

$$\{X(x) | X \in S\} \subseteq T_x A_S(x).$$

We remark that the submanifolds formed by the attainable sets need not all be of the same dimension.

As is well known, every C^1 control vector field $\xi: M \times \mathbb{R}^l \rightarrow TM$ induces a family of C^1 vector fields on M . For each w in \mathbb{R}^l , we define a C^1 vector field X^w by $X^w(x) = \xi(x, w)$, and we set

$$S(\xi) = \{X^w | w \in \mathbb{R}^l\}.$$

We call $S(\xi)$ the *family of vector fields associated to the control vector field ξ* . The global flows X^w and ξ obviously satisfy the relation

$$X_t^w(x) = \mu(t, 0, x, w)$$

for every t in the interval $J(0, x, w)$, where we identify $w \in \mathbb{R}^l$ with the constant control in L_∞^l taking the value w . Hence, there is a one-to-one correspondence between integral

curves of $S(\xi)$ and responses of ξ corresponding to piecewise-constant controls. From this we obtain the relations

$$A_{S(\xi)}(x; t) = A_\xi(x; t), \quad A_{S(\xi)}(x) = A_\xi(x)$$

for every x in M and $t \geq 0$.

DEFINITION 3.4. A C^1 control vector field $\xi : M \times \mathbb{R}^l \rightarrow TM$ is said to be *symmetric* if, for every w in \mathbb{R}^l , there exists a w' in \mathbb{R}^l such that $\xi(x, w) = -\xi(x, w')$ for every x in M .

It is clear that if ξ is a symmetric control vector field on M , then the associated family of vector fields $S(\xi)$ is symmetric. We also note that a control-semilinear control vector field is obviously symmetric.

DEFINITION 3.5. Let M be a C^k manifold ($k \geq 2$), and let $N \subseteq M$ be a C^1 immersed submanifold of M . We denote the inclusion map by $i : N \rightarrow M$. A C^1 control vector field $\xi : M \times \mathbb{R}^l \rightarrow TM$ is *tangent to N* if, for every x in N and w in \mathbb{R}^l , we have

$$\xi(x, w) \in \text{image } di_x.$$

In this case, ξ induces a C^0 control vector field $\tilde{\xi} : N \times \mathbb{R}^l \rightarrow TN$ on N , which is defined by the relation $\tilde{\xi} \circ (i \times 1_{\mathbb{R}^l}) = di \circ \xi$.

PROPOSITION 3.6. Let $\xi : M \times \mathbb{R}^l \rightarrow TM$ be a C^1 control vector field on the C^k manifold M ($k \geq 2$) with global flow $\mu : \mathcal{D}(\xi) \rightarrow M$ and suppose that ξ is tangent to a C^1 immersed submanifold $N \subseteq M$. If $(\bar{t}, 0, \bar{x}, \bar{u}) \in \mathcal{D}(\xi)$ is such that

$$\mu(\bar{t}, 0, \bar{x}, \bar{u}) \in N,$$

then there exists an $\varepsilon > 0$ such that

$$|t - \bar{t}| < \varepsilon \Rightarrow \mu(t, 0, \bar{x}, \bar{u}) \in N.$$

Proof. Let $\tilde{\xi} : N \times \mathbb{R}^l \rightarrow TN$ denote the induced C^0 control vector field on N , and set $\bar{y} = \mu(\bar{t}, 0, \bar{x}, \bar{u})$. Let $p = \dim N$, let (φ, U) be a chart of N with $\bar{y} \in U$, and let $\xi_U : \varphi(U) \times \mathbb{R}^l \rightarrow \mathbb{R}^p$ denote the local representative of $\tilde{\xi}$ with respect to (φ, U) [4; Def. 2.4]. The mapping $(t, y) \mapsto \xi_U(y, \bar{u}(t))$ is clearly measurable in t , continuous in y , and locally bounded in both variables. By the Carathéodory existence theorem for ordinary differential equations [3], there exists an $\varepsilon > 0$ and an absolutely continuous mapping $\alpha : (\bar{t} - \varepsilon, \bar{t} + \varepsilon) \rightarrow \varphi(U)$ such that $\alpha(\bar{t}) = \varphi(\bar{y})$ and

$$\dot{\alpha}(t) = \xi_U(\alpha(t), \bar{u}(t)) \quad \text{a.e. for } t \in (\bar{t} - \varepsilon, \bar{t} + \varepsilon).$$

It is important to note that we cannot infer uniqueness of such a solution from the Carathéodory theorem alone. However, a quick check of the definitions shows that the mapping

$$i \circ \varphi^{-1} \circ \alpha : (\bar{t} - \varepsilon, \bar{t} + \varepsilon) \rightarrow M$$

is a response of the control vector field ξ with initial condition (\bar{t}, \bar{y}) corresponding to the control \bar{u} . Since ξ is of class C^1 , its responses for a prescribed set of initial conditions and control are unique, so that

$$(3) \quad (i \circ \varphi^{-1} \circ \alpha)(t) = \mu(t, \bar{t}, \bar{y}, \bar{u})$$

for every t in $(\bar{t} - \varepsilon, \bar{t} + \varepsilon)$. Using the transitivity of the flow, we obtain, for every t in $(\bar{t} - \varepsilon, \bar{t} + \varepsilon)$,

$$\mu(t, \bar{t}, \bar{y}, \bar{u}) = \mu(t, \bar{t}, \mu(\bar{t}, 0, \bar{x}, \bar{u}), \bar{u}) = \mu(t, 0, \bar{x}, \bar{u}),$$

which is obviously an element of N by (3). \square

The following theorem shows that the condition of symmetry on a control vector field is sufficient to ensure the equality of the attainable sets via measurable and piecewise-constant controls.

THEOREM 3.7. *Let $\xi : M \times \mathbb{R}^l \rightarrow TM$ be a C^1 symmetric control vector field on M and let $S(\xi)$ denote the associated family of C^1 vector fields. Then, for every x in M , we have*

$$\mathcal{A}_\xi(x) = A_\xi(x) = A_{S(\xi)}(x).$$

Proof. Since ξ is symmetric, $S(\xi)$ is symmetric, and Theorem 3.3 implies that $\{A_{S(\xi)}(x) | x \in M\}$ is a partition of M by C^1 immersed submanifolds. Choose a point x_α from each attainable set $A_{S(\xi)}(x)$, so that for an appropriate index set I the family $\{A_{S(\xi)}(x_\alpha) | \alpha \in I\}$ represents all of the attainable sets of $S(\xi)$ and each attainable set is listed only once.

Fix a point \bar{x} in M , and let $\beta \in I$ be such that $A_{S(\xi)}(\bar{x}) = A_{S(\xi)}(x_\beta)$. We have already commented that the equality $A_\xi(\bar{x}) = A_{S(\xi)}(\bar{x})$ and the inclusion $A_\xi(\bar{x}) \subseteq \mathcal{A}_\xi(\bar{x})$ hold. To prove the theorem, it suffices to prove the inclusion $\mathcal{A}_\xi(\bar{x}) \subseteq A_{S(\xi)}(\bar{x})$.

If $y \in \mathcal{A}_\xi(\bar{x})$, then there exist u in L_∞^l and $t \geq 0$ such that $y = \mu(t, 0, \bar{x}, u)$, where $\mu : \mathcal{D}(\xi) \rightarrow M$ is the global flow of ξ . For each α in I , define a subset J_α of the closed interval $[0, t]$ by

$$J_\alpha = \{s \in [0, t] | \mu(s, 0, \bar{x}, u) \in A_{S(\xi)}(x_\alpha)\}.$$

Observe that the family $\{J_\alpha | \alpha \in I\}$ is a partition of $[0, t]$, since $\{A_{S(\xi)}(x_\alpha) | \alpha \in I\}$ is a partition of M . Furthermore, each J_α is open relative to $[0, t]$ by Proposition 3.6. Because $J_\beta \neq \emptyset$ and $[0, t]$ is connected, we must have $J_\beta = [0, t]$, and hence

$$y = \mu(t, 0, \bar{x}, u) \in A_{S(\xi)}(x_\beta) = A_{S(\xi)}(\bar{x}).$$

Since $y \in \mathcal{A}_\xi(\bar{x})$ was arbitrary, we conclude that $\mathcal{A}_\xi(\bar{x}) \subseteq A_{S(\xi)}(\bar{x})$, which completes the proof. \square

We remark that results similar to Theorem 3.7 have been obtained by G. Stefani and P. Zecca [10]. They work in the context of multivalued vector fields and, consequently, their methods of proof differ somewhat from the above.

COROLLARY 3.8. *Let $\xi : M \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow TM$ be a C^1 control-semilinear control vector field on M . Then, for every x in M and $t > 0$, we have*

$$\mathcal{A}_\xi(x; t) = \mathcal{A}_\xi(x) = A_\xi(x) = A_\xi(x; t).$$

Proof. The first and last equalities follow from Corollary 2.11. The second equality follows from the theorem, because a control-semilinear control vector field is symmetric. \square

The proof of the next theorem is similar to that of Theorem 3.7.

THEOREM 3.9. *Let $\xi : M \times \mathbb{R}^l \rightarrow TM$ be a C^1 symmetric control vector field on M and let $\eta : M \times \mathbb{R}^l \rightarrow TM$ be a C^1 (possibly nonsymmetric) control vector field on M such that η is tangent to the submanifold $A_\xi(x)$ for every x in M . Then for every x in M we have*

$$\mathcal{A}_{\xi+\eta}(x) \subseteq \mathcal{A}_\xi(x) = A_\xi(x).$$

It is important to note that if η is not tangent to *all* of the submanifolds $A_\xi(x)$, then a response of $\xi + \eta$ can escape one of the submanifolds $A_\xi(\bar{x})$, even if η is tangent to $A_\xi(\bar{x})$. We omit the easy examples.

4. Normal values and normal reachability. In this section we begin by reviewing the definition of a normal value of a differentiable mapping and refining, to some extent, our previous perturbation theorem [4, Thm. 4.9]. This refinement will enable us to deal

with perturbations of systems that are controllable by piecewise-constant or C^∞ controls, as well as by measurable controls. Our perturbation theorem requires the existence of sufficiently many normal values. Using Sussmann's notion of normal reachability, we will show that if ξ is control semilinear with global flow μ , then every point in the interior of the attainable set $\mathcal{A}_\xi(x_0; t)$ ($x_0 \in M$, $t > 0$) is a normal value of the mapping

$$\mu_{(t,0,x_0)} : \mathcal{D}_{(t,0,x_0)}(\xi) \rightarrow M$$

(this notation was introduced just prior to Def. 2.4).

DEFINITION 4.1 [4]. Let X and Y be Banach manifolds of class C^1 and let $h : X \rightarrow Y$ be a C^1 mapping. A point y_0 in $h(X)$ is called a *normal value* of h if there exists at least one x_0 in $h^{-1}(y_0)$ such that the differential dh_{x_0} is a split-surjective linear mapping.

Remark 4.2. In our applications X will be an open subset of a Banach space and Y will be the finite-dimensional manifold M . Recall that surjectivity and split-surjectivity of the differential are equivalent in this situation [4; Rem. 3.6].

THEOREM 4.3. Let E be a Banach space and let F be a vector subspace of E . Let U be an open subset of E , and let $h : U \rightarrow M$ be a C^1 mapping. Suppose that $C \subseteq h(U)$ is a compact set with the following property: for every y in C there exists a closed subspace G_y of E such that $G_y \subseteq F$, and y is a normal value of $h|_{U \cap G_y}$. Then there exist a compact subset K of $U \cap F$ and an $\varepsilon > 0$ such that, if $\tilde{h} : K \rightarrow M$ is any continuous mapping satisfying $d(\tilde{h}(x), h(x)) \leq \varepsilon$ for every x in K , then $C \subseteq \tilde{h}(K)$.

Proof. The proof is almost identical to the proof of [4, Thm. 3.7]. In fact, only one modification is necessary. When applying the inverse-mapping theorem to obtain a local right inverse of h at $y \in C$, we apply the inverse-mapping theorem to the mapping $h|_{U \cap G_y}$, and thus obtain a local right inverse of h taking values in $G_y \subseteq F$. \square

Remark 4.4. In our applications, E will be L_∞ and F will be P_∞^l , D_∞^l or L_∞^l . The subspace G_y will always be finite dimensional and hence closed in L_∞^l .

Notation. If $\eta : M \times \mathbb{R}^l \rightarrow TM$ is a C^1 control vector field on M and $K \subseteq M \times \mathbb{R}^l$ is a compact subset, then we set

$$\|\eta\|_K = \max \{ \|\eta(x, w)\|_\omega \mid (x, w) \in K \},$$

where $\omega : TM \rightarrow \mathbb{R}$ is the given Finsler structure on TM .

THEOREM 4.5. Let $\xi : M \times \mathbb{R}^l \rightarrow TM$ be a C^1 control vector field on M with global flow $\mu : \mathcal{D}(\xi) \rightarrow M$ and let \mathcal{U} be a vector subspace of L_∞^l . For x_0 in M and $\bar{t} > 0$, suppose that C is a compact subset of the attainable set $\mathcal{A}_\xi(x_0; \bar{t}|\mathcal{U})$ having the following property: for every y in C there exists a closed subspace \mathcal{G}_y of L_∞^l such that $\mathcal{G}_y \subseteq \mathcal{U}$ and y is a normal value of the mapping

$$\mu_{(\bar{t},0,x_0)}|_{\mathcal{D}_{(\bar{t},0,x_0)}(\xi)} \cap \mathcal{G}_y.$$

Then there exist a $\delta > 0$ and a compact subset K of $M \times \mathbb{R}^l$ such that, if $\eta : M \times \mathbb{R}^l \rightarrow TM$ is any C^1 control vector field on M satisfying $\|\eta\|_K \leq \delta$, then $C \subseteq \mathcal{A}_{\xi+\eta}(x_0; \bar{t}|\mathcal{U})$.

Proof. The theorem follows from Theorem 4.3 in much the same way that [4, Thm. 4.9] follows from [4, Thm. 3.7]. We omit the details. \square

We next introduce Sussmann's concept of normal reachability in families of vector fields. As we will see, there is a close connection between normal values and normal reachability for control-semilinear control vector fields.

DEFINITION 4.6 [13]. Let S be a family of C^1 vector fields on M . We say that a point y in M is *normally k -reachable* from a point x in M , $0 \leq k \leq n$, if there exist a positive integer q , vector fields X^1, \dots, X^q in S , and positive real numbers s_1, \dots, s_q

such that $X_{s_q}^q \circ \cdots \circ X_{s_1}^1(x)$ is defined and equals y and the C^1 mapping

$$(t_1, \dots, t_q) \mapsto X_{t_q}^q \circ \cdots \circ X_{t_1}^1(x),$$

which is defined in a neighborhood of (s_1, \dots, s_q) in \mathbb{R}^q , has rank k at (s_1, \dots, s_q) .

The following fundamental result is an immediate consequence of the work of Sussmann [12]. It is also contained implicitly in the work of Stefan [9].

THEOREM 4.7. *Let S be a symmetric family of C^1 vector fields on M , let $x \in M$, and let k be the dimension of the submanifold $A_S(x)$. Then every point y in $A_S(x)$ is normally k -reachable from x .*

THEOREM 4.8. *Let $\xi : M \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow TM$ be a C^1 control-semilinear control vector field on M with global flow $\mu : \mathcal{D}(\xi) \rightarrow M$, and fix x_0 in M . Then the following three statements are equivalent:*

- (i) $\dim \mathcal{A}_\xi(x_0) = n$.
- (ii) Every point y in $\mathcal{A}_\xi(x_0)$ is normally n -reachable from x_0 via the associated family of vector fields $S(\xi)$.
- (iii) For every y in $\mathcal{A}_\xi(x_0)$ and $\bar{t} > 0$, y is a normal value of the mapping

$$\mu_{(\bar{t}, 0, x_0)} : \mathcal{D}_{(\bar{t}, 0, x_0)}(\xi) \rightarrow M.$$

Proof. If $S(\xi)$ denotes the family of C^1 vector fields on M associated to ξ , then $S(\xi)$ is symmetric, $A_{S(\xi)}(x) = \mathcal{A}_\xi(x)$ for every x in M and each attainable set has the structure of a C^1 immersed submanifold of M . These observations are contained in Theorems 3.3 and 3.7 and will be useful to us here.

The implication (i) \Rightarrow (ii) follows from Theorem 4.7.

Assume that statement (iii) holds. Let y be a point in $\mathcal{A}_\xi(x_0)$, and let $\bar{t} > 0$ be a positive real number. By assumption, there exists a control u in $\mathcal{D}_{(\bar{t}, 0, x_0)}(\xi)$ such that $\mu_{(\bar{t}, 0, x_0)}(u) = y$, and the differential

$$d(\mu_{(\bar{t}, 0, x_0)})_u : L_\infty^{p+m} \rightarrow T_y M$$

is a surjective linear mapping. The surjective-mapping theorem implies that the interior of

$$\text{image}(\mu_{(\bar{t}, 0, x_0)}) = \mathcal{A}_\xi(x_0; \bar{t})$$

relative to M is nonempty. Therefore $\mathcal{A}_\xi(x_0)$ is a submanifold of M having nonempty interior. Since $\mathcal{A}_\xi(x_0)$ is a connected submanifold of the paracompact manifold M , it follows that $\mathcal{A}_\xi(x_0)$ is a second countable in its submanifold topology [2]. We conclude that $\dim \mathcal{A}_\xi(x_0) = n$, and this proves the implication (iii) \Rightarrow (i).

It remains to prove the implication (ii) \Rightarrow (iii). In the following argument we will identify the point (z, w) in $\mathbb{R}^p \times \mathbb{R}^m$ with the constant control in L_∞^{p+m} taking the value (z, w) . Let y be a point in $\mathcal{A}_\xi(x_0) = A_{S(\xi)}(x_0)$. By assumption, y is normally n -reachable from x_0 so there exists a positive integer q , constant controls

$$(z_1, w_1), \dots, (z_q, w_q) \in \mathbb{R}^p \times \mathbb{R}^m$$

and positive real numbers s_1, \dots, s_q such that the C^1 mapping

$$(4) \quad \begin{aligned} & (t_1, \dots, t_q) \\ & \mapsto \mu(t_q, 0, \mu(t_{q-1}, 0, \mu(\dots, \mu(t_1, 0, x_0, z_1, w_1), \dots), z_{q-1}, w_{q-1}), z_q, w_q) \end{aligned}$$

is defined on an open neighborhood of (s_1, \dots, s_q) in \mathbb{R}^q , takes the value y at (s_1, \dots, s_q) and is of rank n at (s_1, \dots, s_q) . Using the elementary properties of the global flow μ listed earlier, we will derive an alternative expression for the mapping (4).

Let U be an open neighborhood of (s_1, \dots, s_q) in \mathbb{R}^q such that $(t_1, \dots, t_q) \in U$ implies $t_i > 0$ for each $i = 1, \dots, q$ and (t_1, \dots, t_q) is in the domain of the mapping (4). Denote the restriction of the mapping (4) to the set U by $f: U \rightarrow M$. Then for each (t_1, \dots, t_q) in U we have

$$\begin{aligned} f(t_1, \dots, t_q) &= \mu(t_q, 0, \mu(t_{q-1}, 0, \mu(\dots, \mu(t_1, 0, x_0, z_1, w_1), \dots), z_{q-1}, w_{q-1}), z_q, w_q) \\ &= \mu\left(\frac{t_q}{s_q} s_q, 0, \mu\left(\frac{t_{q-1}}{s_{q-1}} s_{q-1}, 0, \mu\left(\dots, \mu\left(\frac{t_1}{s_1} s_1, 0, x_0, z_1, w_1\right), \dots\right), z_{q-1}, w_{q-1}\right), z_q, w_q\right) \\ &= \mu\left(s_q, 0, \mu\left(s_{q-1}, 0, \mu\left(\dots, \mu\left(s_1, 0, x_0, z_1, \frac{t_1}{s_1} w_1\right), \dots\right), z_{q-1}, \frac{t_{q-1}}{s_{q-1}} w_{q-1}\right), z_q, \frac{t_q}{s_q} w_q\right), \end{aligned}$$

where the last equality is a consequence of Proposition 2.10.

Set $s_0 = 0$ and define controls $v_1, \dots, v_q \in L_\infty^p$, $u_1, \dots, u_q \in L_\infty^m$ by

$$v_i(t) = \begin{cases} z_i & s_0 + \dots + s_{i-1} \leq t \leq s_0 + \dots + s_{i-1} + s_i \\ 0 & \text{otherwise,} \end{cases}$$

and

$$u_i(t) = \begin{cases} (1/s_i) w_i & s_0 + \dots + s_{i-1} \leq t \leq s_0 + \dots + s_{i-1} + s_i \\ 0 & \text{otherwise.} \end{cases}$$

Using this notation and Propositions 2.2 and 2.3, we obtain

$$\begin{aligned} f(t_1, \dots, t_q) &= \mu(s_q + \dots + s_1, s_{q-1} + \dots + s_1, \mu(s_{q-1} + \dots + s_1, s_{q-2} + \dots + s_1, \\ &\quad \mu(\dots, \mu(s_1, 0, x_0, v_1, t_1 u_1), \dots), v_{q-1}, t_{q-1} u_{q-1}), v_q, t_q u_q) \\ &= \mu(s_q + \dots + s_1, 0, x_0, v_1 + \dots + v_q, t_1 u_1 + \dots + t_q u_q). \end{aligned}$$

Define a mapping $\rho: \mathbb{R}^q \rightarrow L_\infty^{p+m}$ by

$$\rho(t_1, \dots, t_q) = (v_1 + \dots + v_q, t_1 u_1 + \dots + t_q u_q).$$

Since ρ is an affine mapping, it is clearly of class C^∞ . If we set $\bar{s} = s_1 + \dots + s_q$, then our above computation shows that the mapping $f: U \rightarrow M$ admits a factorization

$$f = \mu_{(\bar{s}, 0, x_0)} \circ (\rho|_U).$$

Letting $v = v_1 + \dots + v_q$ and $u = s_1 u_1 + \dots + s_q u_q$, we see that v and u are piecewise-constant controls and, by the chain rule,

$$df_{(s_1, \dots, s_q)} = d(\mu_{(\bar{s}, 0, x_0)})_{(v, u)} \circ D\rho_{(s_1, \dots, s_q)}.$$

Since the differential $df_{(s_1, \dots, s_q)}$ is surjective by assumption, the differential $d(\mu_{(\bar{s}, 0, x_0)})_{(v, u)}$ is also surjective. We conclude that $y = \mu_{(\bar{s}, 0, x_0)}(v, u)$ is a normal value of the mapping $\mu_{(\bar{s}, 0, x_0)}$.

Finally, if $\bar{t} > 0$ is an arbitrary positive real number, then for $\alpha = \bar{t}/\bar{s}$ we have

$$\begin{aligned} y &= \mu(s_q, 0, \mu(s_{q-1}, 0, \mu(\dots, \mu(s_1, 0, x_0, z_1, w_1), \dots), z_{q-1}, w_{q-1}), z_q, w_q) \\ &= \mu(\alpha s_q, 0, \mu(\alpha s_{q-1}, 0, \mu(\dots, \mu(\alpha s_1, 0, x_0, z_1, \alpha^{-1} w_1), \dots), \\ &\quad z_{q-1}, \alpha^{-1} w_{q-1}), z_q, \alpha^{-1} w_q). \end{aligned}$$

Hence, the mapping

$$g(t_1, \dots, t_q) \\ = \mu(t_q, 0, \mu(t_{q-1}, 0, \mu(\dots, \mu(t_1, 0, x_0, z_1, \alpha^{-1}w_1), \dots), z_{q-1}, \alpha^{-1}w_{q-1}), z_q \alpha^{-1}w_q)$$

is defined in a neighborhood of $(\alpha s_1, \dots, \alpha s_q)$ and satisfies

$$g(t_1, \dots, t_q) = f(t_1/\alpha, \dots, t_q/\alpha).$$

In particular, we have

$$y = g(\alpha s_1, \dots, \alpha s_q)$$

and

$$dg_{(\alpha s_1, \dots, \alpha s_q)} = (1/\alpha)df_{(s_1, \dots, s_q)}.$$

This shows that g has rank n at $(\alpha s_1, \dots, \alpha s_q)$. Since

$$\alpha s_q + \dots + \alpha s_q = \alpha \bar{s} = \bar{t},$$

the preceding argument shows that y is a normal value of the mapping $\mu_{(\bar{t}, 0, x_0)}$. This completes the proof of the implication (ii) \Rightarrow (iii) and the theorem. \square

Imbedded in the above proof is the following technical result, which will be useful in verifying the hypothesis of Theorem 4.5 for control-semilinear control vector fields.

COROLLARY 4.9. *Let $\dim \mathcal{A}_\xi(x_0) = n$. Then, for every y in $\mathcal{A}_\xi(x_0)$ and every $\bar{t} > 0$, there exists a finite-dimensional subspace \mathcal{G}_y of P_∞^{p+m} such that y is a normal value of the mapping*

$$\mu_{(t, 0, x_0)}|_{\mathcal{D}_{(t, 0, x_0)}(\xi)} \cap \mathcal{G}_y.$$

Proof. The proof of the theorem shows that $\mu_{(\bar{t}, 0, x_0)}$ factors through a subspace of P_∞^{p+m} spanned by a finite number of piecewise-constant controls. \square

5. Applications to the perturbation problem. We will now apply the preceding results to study the effect of nonlinear perturbations on certain controllability properties of control-semilinear control vector fields. Our first theorem is the analog of [4, Thm. 4.9] for control-semilinear control vector fields. In this case the formulation is somewhat more satisfying than that of [4, Thm. 4.9] because we do not require any assumption concerning the existence of normal values.

THEOREM 5.1. *Let $\xi: M \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow TM$ be a C^1 control-semilinear control vector field on M , and let $x_0 \in M$ be such that $\dim \mathcal{A}_\xi(x_0) = n$. Let C be a compact subset of $\mathcal{A}_\xi(x_0)$, and let $\bar{t} > 0$ be given. Then there exist a $\delta > 0$ and a compact subset K of $M \times \mathbb{R}^p \times \mathbb{R}^m$ such that if $\eta: M \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow TM$ is any C^1 control vector field on M satisfying $\|\eta\|_K \leq \delta$, then $C \subseteq A_{\xi+\eta}(x_0; \bar{t})$.*

Proof. By Corollary 3.8, we have $C \subseteq \mathcal{A}_\xi(x_0) = A_\xi(x_0; \bar{t})$. The result is now an immediate consequence of Corollary 4.9 and Theorem 4.5 with $\mathcal{U} = P_\infty^{p+m}$. \square

Our next objective is to develop results of a character similar to Theorem 5.1 for global controllability (as opposed to controllability to a compact set) and globally bounded perturbations (as opposed to perturbations that are bounded, but possibly very small, on a compact set).

PROPOSITION 5.2. *Let $\xi: M \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow TM$ be a C^1 control-semilinear control vector field on M and let $\mathcal{U} \subseteq L_\infty^{p+m}$ be such that $\alpha > 0$ and $(v, u) \in \mathcal{U}$ imply $(v^\alpha, (1/\alpha)u^\alpha) \in \mathcal{U}$. Let $\eta: M \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow TM$ be an arbitrary C^1 control vector field on M , and let $\varepsilon > 0$ be given. Define a C^1 control vector field $\eta_\varepsilon: M \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow TM$ by*

$$\eta_\varepsilon(x, z, w) = \varepsilon \eta(x, z, (1/\varepsilon)w)$$

for (x, z, w) in $M \times \mathbb{R}^p \times \mathbb{R}^m$. Then, for each x_0 in M and $\bar{t} > 0$, we have

$$\mathcal{A}_{\xi+\eta}(x_0; \bar{t}|\mathcal{U}) = \mathcal{A}_{\xi+\eta_\varepsilon}(x_0; \bar{t}/\varepsilon|\mathcal{U}).$$

Proof. Let ρ and ρ_ε denote the global flows of $\xi + \eta$ and $\xi + \eta_\varepsilon$, respectively. If $y \in \mathcal{A}_{\xi+\eta}(x_0; \bar{t}|\mathcal{U})$, then there exists a control (v, u) in \mathcal{U} such that

$$\rho(\bar{t}, 0, x_0, v, u) = y.$$

Define a mapping $\beta : [0, \bar{t}/\varepsilon] \rightarrow M$ by

$$\beta(t) = \rho(\varepsilon t, 0, x_0, v, u).$$

Then $\beta(0) = x_0$, $\beta(\bar{t}/\varepsilon) = y$ and, for almost every t in $[0, \bar{t}/\varepsilon]$, we have

$$\begin{aligned} \dot{\beta}(t) &= \varepsilon D_1 \rho(\varepsilon t, 0, x_0, v, u) \\ &= \varepsilon \left[\sum_{i=1}^m u_i(\varepsilon t) \xi_i(\beta(t), v(\varepsilon t)) + \eta(\beta(t), v(\varepsilon t), u(\varepsilon t)) \right] \\ &= \sum_{i=1}^m \varepsilon u_i^{1/\varepsilon}(t) \xi_i(\beta(t), v^{1/\varepsilon}(t)) + \eta_\varepsilon(\beta(t), v^{1/\varepsilon}(t), \varepsilon u^{1/\varepsilon}(t)). \end{aligned}$$

Consequently, for t in $[0, \bar{t}/\varepsilon]$, we have

$$\beta(t) = \rho_\varepsilon(t, 0, x_0, v^{1/\varepsilon}, \varepsilon u^{1/\varepsilon})$$

and, in particular,

$$y = \beta(\bar{t}/\varepsilon) = \rho_\varepsilon(\bar{t}/\varepsilon, 0, x_0, v^{1/\varepsilon}, \varepsilon u^{1/\varepsilon}) \in \mathcal{A}_{\xi+\eta_\varepsilon}(x_0; \bar{t}/\varepsilon|\mathcal{U}).$$

Since $y \in \mathcal{A}_{\xi+\eta}(x_0; \bar{t}|\mathcal{U})$ was arbitrary, we conclude that

$$\mathcal{A}_{\xi+\eta}(x_0; \bar{t}|\mathcal{U}) \subseteq \mathcal{A}_{\xi+\eta_\varepsilon}(x_0; \bar{t}/\varepsilon|\mathcal{U}).$$

The reverse inclusion follows in a similar manner or by symmetry. \square

COROLLARY 5.3. $\mathcal{A}_{\xi+\eta}(x_0|\mathcal{U}) = \mathcal{A}_{\xi+\eta_\varepsilon}(x_0|\mathcal{U})$.

DEFINITION 5.4. Let $\eta : M \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow TM$ be a C^1 control vector field on M . We say that η satisfies the *modified boundedness condition in \mathbb{R}^m* if, for every pair of compact sets $K_1 \subseteq M$ and $K_2 \subseteq \mathbb{R}^p$, there exists a constant $B > 0$ such that

$$(x, z, w) \in K_1 \times K_2 \times \mathbb{R}^m \Rightarrow \|\eta(x, z, w)\|_\omega \leq B.$$

Remark 5.5. Observe that if either

- (a) the Finsler norm of η is globally bounded on $M \times \mathbb{R}^p \times \mathbb{R}^m$, or
- (b) η does not depend on $w \in \mathbb{R}^m$,

then η satisfies the modified boundedness condition in \mathbb{R}^m . For our applications, these are the two cases of interest.

THEOREM 5.6. Let $\xi : M \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow TM$ be a C^1 control-semilinear control vector field on M , and let $x_0 \in M$ be such that $\dim \mathcal{A}_\varepsilon(x_0) = n$. Suppose that $\eta : M \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow TM$ is an arbitrary C^1 control vector field on M which satisfies the modified boundedness condition in \mathbb{R}^m . If C is a compact subset of $\mathcal{A}_\xi(x_0)$, then there exists $\bar{s} > 0$ such that $C \subseteq \mathcal{A}_{\xi+\eta}(x_0; \bar{s})$.

Proof. Fix $\bar{t} > 0$. By Theorem 5.1, there exist a $\delta > 0$ and a compact subset K of $M \times \mathbb{R}^p \times \mathbb{R}^m$ such that if $\zeta : M \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow TM$ is a C^1 control vector field on M satisfying $\|\zeta\|_K \leq \delta$, then $C \subseteq \mathcal{A}_{\xi+\zeta}(x_0; \bar{t})$. Let

$$\lambda_1 : M \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow M, \quad \lambda_2 : M \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow \mathbb{R}^p$$

denote the projections on the indicated factors. Because η satisfies the modified

boundedness condition in \mathbb{R}^m , there exists $B > 0$ such that

$$(x, z, w) \in \lambda_1(K) \times \lambda_2(K) \times \mathbb{R}^m \Rightarrow \|\eta(x, z, w)\|_\omega \leq B.$$

Choose $\varepsilon > 0$ so that $\varepsilon B \leq \delta$, and define a C^1 control vector field $\eta_\varepsilon : M \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow TM$ by

$$\eta_\varepsilon(x, z, w) = \varepsilon \eta(x, z, (1/\varepsilon)w)$$

for (x, z, w) in $M \times \mathbb{R}^p \times \mathbb{R}^m$. Then for (x, z, w) in K we have

$$\|\eta_\varepsilon(x, z, w)\|_\omega = \|\varepsilon \eta(x, z, (1/\varepsilon)w)\|_\omega \leq \varepsilon B \leq \delta,$$

and this implies that $C \subseteq A_{\xi+\eta_\varepsilon}(x_0; \bar{t})$. We can now apply Proposition 5.2 to conclude that

$$C \subseteq A_{\xi+\eta_\varepsilon}(x_0; \bar{t}) = A_{\xi+\eta}(x_0; \varepsilon \bar{t}),$$

so the conclusion of the theorem holds with $\bar{s} = \varepsilon \bar{t}$. \square

COROLLARY 5.7. *If ξ , η and x_0 are as above, then $\mathcal{A}_\xi(x_0) \subseteq A_{\xi+\eta}(x_0)$. In particular, if ξ is completely controllable by measurable (or, equivalently, piecewise-constant) controls, then $\xi + \eta$ is completely controllable by piecewise-constant controls.*

Proof. The theorem implies that $C \subseteq A_{\xi+\eta}(x_0)$ for every compact subset C of $\mathcal{A}_\xi(x_0)$, whence the inclusion $\mathcal{A}_\xi(x_0) \subseteq A_{\xi+\eta}(x_0)$ follows. The second statement follows from this and the definition of complete controllability.

For emphasis, we give two restatements of Corollary 5.7, corresponding to the two special instances of the modified boundedness condition listed in Remark 5.5.

COROLLARY 5.8. *If $\xi : M \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow TM$ is a C^1 control-semilinear control vector field on M that is completely controllable by measurable controls and if $\eta : M \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow TM$ is a C^1 control vector field on M whose Finsler norm is globally bounded on $M \times \mathbb{R}^p \times \mathbb{R}^m$, then $\xi + \eta$ is completely controllable by piecewise-constant controls.*

COROLLARY 5.9. *If $\xi : M \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow TM$ is a C^1 control-semilinear control vector field on M that is completely controllable by measurable controls and if $\eta : M \times \mathbb{R}^p \rightarrow TM$ is an arbitrary C^1 control vector field on M (with control space \mathbb{R}^p), then $\xi + \eta$ is completely controllable by piecewise-constant controls.*

In the event that ξ is not completely controllable, or the submanifold $\mathcal{A}_\xi(x_0)$ is not of maximal dimension, it is still possible to obtain results for perturbation control vector fields η that are tangent to $\mathcal{A}_\xi(x_0)$. The method of proof seems apparent; we simply restrict our attention to the submanifold $\mathcal{A}_\xi(x_0)$ and apply the preceding results. However, one technical difficulty arises. If ξ is of class C^1 , then the submanifold $\mathcal{A}_\xi(x_0)$ is of class C^1 , so the restriction of ξ to $\mathcal{A}_\xi(x_0)$ is only of class C^0 , in general. Since the preceding results were formulated for manifolds of class at least C^2 and control vector fields of class at least C^1 , they cannot be applied directly to the restriction of ξ to $\mathcal{A}_\xi(x_0)$. We will circumvent this problem by assuming in the following two results that M is of class at least C^3 and the control vector fields on M are of class at least C^2 . These results are actually true for C^2 manifolds and C^1 control vector fields, but the required arguments appear to be rather technical and uninteresting.

THEOREM 5.10. *Let M be of class C^k , $k \geq 3$, and let $\xi : M \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow TM$ be a C^2 control-semilinear control vector field on M . Let $\eta : M \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow TM$ be a C^2 control vector field on M that is tangent to the submanifold $\mathcal{A}_\xi(x)$ for every x in M and satisfies the modified boundedness condition in \mathbb{R}^m . Then*

$$\mathcal{A}_\xi(x) = A_{\xi+\eta}(x) = \mathcal{A}_{\xi+\eta}(x)$$

for every x in M .

Proof. By Theorem 3.9, for every x in M we have the inclusions

$$(5) \quad A_{\xi+\eta}(x) \subseteq \mathcal{A}_{\xi+\eta}(x) \subseteq \mathcal{A}_{\xi}(x).$$

Each of the submanifolds $\mathcal{A}_{\xi}(x)$ is of class C^2 , since ξ is of class C^2 , and the vector fields ξ, η induce C^1 control vector fields on these submanifolds by restriction. Because the restriction of ξ to the submanifold $\mathcal{A}_{\xi}(x)$ is obviously globally controllable (with respect to $\mathcal{A}_{\xi}(x)$), we can apply Corollary 5.7 to conclude that

$$\mathcal{A}_{\xi}(x) \subseteq A_{\xi+\eta}(x)$$

for every x in M . Combining this inclusion with the inclusions (5), we obtain the result. \square

A similar technique can be used to prove the following theorem.

THEOREM 5.11. *Let M be of class C^k , $k \geq 3$, and let $\xi: M \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow TM$ be a C^2 control-semilinear control vector field on M . Let $\eta: M \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow TM$ be a C^2 control vector field on M that is tangent to one of the submanifolds $\mathcal{A}_{\xi}(x_0)$, and suppose that the restriction $\tilde{\eta}: \mathcal{A}_{\xi}(x_0) \times \mathbb{R}^p \times \mathbb{R}^m \rightarrow T\mathcal{A}_{\xi}(x_0)$ of η to $\mathcal{A}_{\xi}(x_0)$ satisfies the modified boundedness condition in \mathbb{R}^m . Then we have $\mathcal{A}_{\xi}(x_0) \subseteq A_{\xi+\eta}(x_0)$.*

6. Controllability by C^∞ controls. In this section we will specialize to C^1 control vector fields $\xi: M \times \mathbb{R}^m \rightarrow TM$ that are linear in the control variable; i.e.,

$$\xi(x, w) = \sum_{i=1}^m w_i \xi_i(x),$$

where ξ_1, \dots, ξ_m are C^1 (uncontrolled) vector fields on M . Our aim is to examine situations in which nonlinear perturbations of such control vector fields are completely controllable by C^∞ controls. We begin with a technical lemma.

LEMMA 6.1. *Let $\xi: M \times \mathbb{R}^m \rightarrow TM$ be a C^1 control-linear control vector field on M with global flow $\mu: \mathcal{D}(\xi) \rightarrow M$ and fix $x_0 \in M$, $\bar{t} > 0$. If \mathcal{G} is a finite-dimensional subspace of P_∞^m , then there exists a linear mapping $\Gamma: \mathcal{G} \rightarrow D_\infty^m$ such that*

$$\mu_{(\bar{t}, 0, x_0)}(u) = \mu_{(\bar{t}, 0, x_0)}(\Gamma(u))$$

for every u in $\mathcal{D}_{(\bar{t}, 0, x_0)}(\xi) \cap \mathcal{G}$.

Proof. Let $0 = s_0 < s_1 < \dots < s_q = \bar{t}$ be a partition of $[0, \bar{t}]$ such that, for every u in \mathcal{G} , u is constant on the open interval (s_{i-1}, s_i) for each $i = 1, \dots, q$. Such a partition can be obtained as follows: take a finite number of controls in \mathcal{G} which span \mathcal{G} , determine for each one of these controls a partition on whose open subintervals the control is constant, and form the common refinement of the partitions so obtained.

Let $\sigma: \mathbb{R} \rightarrow \mathbb{R}$ be a C^∞ function having the following properties:

- (a) $\sigma|_{[s_0, s_q]}$ is a one-to-one, increasing mapping of $[s_0, s_q]$ onto $[s_0, s_q]$;
- (b) $\sigma(s_i) = s_i$ for every $0 \leq i \leq q$;
- (c) $\sigma^{(k)}(s_i) = 0$ for every $0 \leq i \leq q$ and $k \geq 1$;
- (d) $\sigma((-\infty, s_0]) = s_0$, $\sigma([s_q, \infty)) = s_q$.

For each u in \mathcal{G} , let $\Gamma(u)$ be defined by

$$\Gamma(u)(t) = \dot{\sigma}(t)u(\sigma(t)).$$

Since $u \circ \sigma$ is constant on the intervals $(-\infty, s_0)$, (s_q, ∞) , and (s_{i-1}, s_i) for $i = 1, \dots, q$ and since $\sigma^{(k)}(s_i) = 0$ for every $i = 0, \dots, q$ and $k \geq 1$, it is easy to see that $\Gamma(u): \mathbb{R} \rightarrow \mathbb{R}^m$ is C^∞ . Hence, we obtain a mapping $\Gamma: \mathcal{G} \rightarrow D_\infty^m$ which is clearly linear.

Let $u \in \mathcal{D}_{(\bar{i}, 0, x_0)}(\xi) \cap \mathcal{G}$ and define $\varphi : [0, \bar{i}] \rightarrow M$ by $\varphi(t) = \mu(t, 0, x_0, u)$. Then $\varphi \circ \sigma$ is a mapping of $[0, \bar{i}]$ into M and satisfies $(\varphi \circ \sigma)(0) = x_0$ and

$$(\dot{\varphi} \circ \dot{\sigma})(t) = \dot{\sigma}(t) \dot{\varphi}(\sigma(t)) = \dot{\sigma}(t) \sum_{i=1}^m u_i(\sigma(t)) \xi_i(\varphi(\sigma(t))) = \xi((\varphi \circ \sigma)(t), \Gamma(u)(t))$$

for almost every t in $[0, \bar{i}]$. We conclude that

$$(\varphi \circ \sigma)(t) = \mu(t, 0, x_0, \Gamma(u))$$

for every t in $[0, \bar{i}]$. Therefore, we have

$$\mu_{(\bar{i}, 0, x_0)}(u) = \varphi(\bar{i}) = \varphi(\sigma(\bar{i})) = \mu_{(\bar{i}, 0, x_0)}(\Gamma(u)),$$

which completes the proof. \square

COROLLARY 6.2. *If $\xi : M \times \mathbb{R}^m \rightarrow TM$ is a C^1 control-linear control vector field on M , then for every x in M and $t > 0$ the six attainable sets*

$$\mathcal{A}_\xi(x; t), \quad \mathcal{A}_\xi(x), \quad A_\xi(x; t), \quad A_\xi(x), \quad \mathcal{A}_\xi(x; t|D_\infty^m), \quad \mathcal{A}_\xi(x|D_\infty^m)$$

are all equal.

Proof. The equality of the first four sets follows from Corollary 3.8 and the equality of the last two follows from Corollary 2.11. The inclusion

$$\mathcal{A}_\xi(x; t|D_\infty^m) \subseteq \mathcal{A}_\xi(x; t) = A_\xi(x; t)$$

is obvious, since $D_\infty^m \subseteq L_\infty^m$, and the inclusion

$$A_\xi(x; t) \subseteq \mathcal{A}_\xi(x; t|D_\infty^m)$$

follows from the lemma. \square

LEMMA 6.3. *Let $\xi : M \times \mathbb{R}^m \rightarrow TM$ be a C^1 control-linear vector field on M and let $x_0 \in M$ be such that $\dim \mathcal{A}_\xi(x_0) = n$. Then, for every y in $\mathcal{A}_\xi(x_0)$ and every $\bar{i} > 0$, there exists a finite-dimensional subspace \mathcal{F}_y of D_∞^m such that y is a normal value of the mapping*

$$\mu_{(\bar{i}, 0, x_0)}|_{\mathcal{D}_{(\bar{i}, 0, x_0)}(\xi) \cap \mathcal{F}_y}.$$

Proof. By Corollary 4.9 there exists a finite-dimensional subspace \mathcal{G}_y of P_∞^m such that y is a normal value of the mapping

$$\mu_{(\bar{i}, 0, x_0)}|_{\mathcal{D}_{(\bar{i}, 0, x_0)}(\xi) \cap \mathcal{G}_y}.$$

Lemma 6.1 yields a linear mapping $\Gamma : \mathcal{G}_y \rightarrow D_\infty^m$ such that

$$\mu_{(\bar{i}, 0, x_0)}(u) = (\mu_{(\bar{i}, 0, x_0)} \circ \Gamma)(u)$$

for every u in $\mathcal{D}_{(\bar{i}, 0, x_0)}(\xi) \cap \mathcal{G}_y$. If we set $\mathcal{F}_y = \Gamma(\mathcal{G}_y)$, then it is clear that y is a normal value of the mapping $\mu_{(\bar{i}, 0, x_0)}$ restricted to the set $\mathcal{D}_{(\bar{i}, 0, x_0)}(\xi) \cap \mathcal{F}_y$. \square

The following theorem is the analog of Theorem 5.1 for C^∞ controls.

THEOREM 6.4. *Let $\xi : M \times \mathbb{R}^m \rightarrow TM$ be a C^1 control-linear control vector field on M , and let $x_0 \in M$ be such that $\dim \mathcal{A}_\xi(x_0) = n$. Let C be a compact subset of $\mathcal{A}_\xi(x_0)$ and let $\bar{i} > 0$ be given. Then there exist a $\delta > 0$ and a compact subset K of $M \times \mathbb{R}^m$ such that, if $\eta : M \times \mathbb{R}^m \rightarrow TM$ is any C^1 control vector field on M satisfying $\|\eta\|_K \leq \delta$, then $C \subseteq \mathcal{A}_{\xi+\eta}(x_0; \bar{i}|D_\infty^m)$.*

Proof. By Corollary 6.2, we have $C \subseteq \mathcal{A}_\xi(x_0) = \mathcal{A}_\xi(x_0; \bar{i}|D_\infty^m)$. The result is now an immediate consequence of Lemma 6.3 and Theorem 4.5 with $\mathcal{U} = D_\infty^m$. \square

THEOREM 6.5. *Let $\xi : M \times \mathbb{R}^m \rightarrow TM$ be a C^1 control-linear control vector field on M and let $x_0 \in M$ be such that $\dim \mathcal{A}_\xi(x_0) = n$. Suppose that $\eta : M \times \mathbb{R}^m \rightarrow TM$ is an arbitrary C^1 control vector field on M which satisfies the modified boundedness condition in \mathbb{R}^m . If C is a compact subset of $\mathcal{A}_\xi(x_0)$, then there exists $\bar{s} > 0$ such that*

$$C \subseteq \mathcal{A}_{\xi+\eta}(x_0; \bar{s}|D_\infty^m).$$

Proof. The proof is very similar to that of Theorem 5.6. We omit the details. \square

COROLLARY 6.6. *If ξ , η and x_0 are as above, then we have*

$$\mathcal{A}_\xi(x_0) \subseteq \mathcal{A}_{\xi+\eta}(x_0|D_\infty^m).$$

In particular, if ξ is completely controllable by measurable (or, equivalently, piecewise-constant) controls, then $\xi + \eta$ is completely controllable by C^∞ controls.

In conclusion, we note that results analogous to Corollaries 5.8, 5.9 and Theorems 5.10, 5.11 are valid for C^∞ controls (as opposed to piecewise-constant controls) in the control-linear case (i.e., the case $p = 0$). The formulation of the precise statements and proofs of these results is a routine matter.

REFERENCES

- [1] P. BRUNOVSKY AND C. LOBRY, *Contrôlabilité bang bang, contrôlabilité différentiable, et perturbation des systèmes non linéaires*, Ann. Mat. Pura Appl., Ser. 4, 105 (1975), pp. 93–119.
- [2] C. CHEVALLEY, *Theory of Lie Groups*, Princeton University Press, Princeton, NJ, 1946.
- [3] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
- [4] K. A. GRASSE, *Perturbations of nonlinear controllable systems*, this Journal, 19 (1981), pp. 203–220.
- [5] H. HERMES, *On local and global controllability*, SIAM J. Control, 12 (1974), pp. 252–261.
- [6] C. LOBRY, *Quelques aspects qualitatifs de la théorie de la commande*, Thèse, l'Université de Grenoble, 1972.
- [7] C. LOBRY, *Controllability of non-linear control dynamical systems*, in Control Theory and Topics in Functional Analysis (3 vols.), International Atomic Energy Agency, Vienna, 1976, pp. 361–383.
- [8] C. LOBRY, *Bases mathématiques de la théorie de systèmes asservis non linéaires*, l'U.E.R. de Mathématiques et Informatique 7505, l'Université de Bordeaux, 1976.
- [9] P. STEFAN, *Accessible sets, orbits, and foliations with singularities*, Proc. London Math. Soc., 29 (1974), pp. 699–713.
- [10] G. STEFANI AND P. ZECCA, *Multivalued differential equations on manifolds with applications to control theory*, Illinois J. Math., to appear.
- [11] H. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.
- [12] H. SUSSMANN, *Orbits of families of vector fields and integrability of distributions*, Trans. Amer. Math. Soc., 180 (1973), pp. 171–188.
- [13] H. SUSSMANN, *Some properties of vector field systems that are not altered by small perturbations*, J. Differential Equations, 20 (1976), pp. 293–315.

STABILITY, EXTENDED SPACES AND NUMERICAL RANGES*

D. J. ALLWRIGHT[†] AND A. I. MEES[†]

Abstract. This paper looks at a generalized version of the ideas of passivity and positivity used in studying stability of nonlinear feedback systems. We generalize these ideas to normed spaces that may not be Hilbert spaces by using the concept of the numerical range and extending this to nonlinear operators. The stability theorem derived by this method includes the usual results for passive and positive operators but also allows one to obtain different results by applying it to non-Hilbert spaces. Before coming to the stability theorem, it is necessary to examine more carefully the way in which extended spaces are constructed. A certain technical property, needed in the transition between extended and original spaces in stability proofs, holds automatically in Hilbert spaces: it need not hold in general but we show how it can always be achieved by an initial enlargement of the original normed space.

1. Introduction. This paper presents an approach to stability which generalizes the positivity and passivity approaches. The idea is to take a standard nonlinear feedback system with input u and output y , and find conditions under which $\|y\|/\|u\|$ is bounded for all inputs u of interest. It is customary and sensible to try to give the conditions a graphical interpretation which makes questions of robustness and the like easy to answer, and we shall demonstrate that the usual interpretations can be given to our results.

For example, Freedman, Falb and Zames [5] established a stability criterion for systems whose transfer function matrix is normal, i.e., commutes with its adjoint. We indicate how to obtain this result and a generalization that allows nonnormal matrices whose departure from normality is suitably bounded.

We shall use the notion of an extended space, as introduced by Zames [15], [16]. In most cases, the extended spaces are constructed through the use of projections P^t that truncate a signal to zero after time t . However, these projections are inappropriate for certain spaces so we give a treatment of extended spaces that only depends on the P^t obeying certain axioms. These axioms cover the commonly occurring cases and are more general than the resolution space setting introduced into control theory by Krein [6], [7]. In the course of this treatment of extended spaces we have to clarify a certain concept that we call "fullness". The formal definition is in § 2.1 below, and the idea is that we say a space E of signals is full if every element of the extended space that was not already in E has $\sup_t \|P^t x\| = \infty$. When E is a Hilbert space, this will follow from the basic axioms on the P^t , but in general it does not. However, we show that it can always be achieved, by an initial enlargement of E if necessary. In most examples it will be clear what the correct enlargement is.

In stability studies in the context of extended spaces it has been common since the work of Zames [16] to work in terms of *relations* instead of operators: whereas an operator is a rule that associates at most one element of a space E_2 with each element of E_1 , a relation is simply a subset of $E_1 \times E_2$. Thus if $T \subseteq E_1 \times E_2$ is a relation and $x \in E_1$, there may be one or many $y \in E_2$ such that $(x, y) \in T$ (in which case we say that x lies in the *domain* of T) or there may be none at all. For an operator, we write $y = Tx$ in place of $(x, y) \in T$. Every operator is a relation, but examples of relations that are not operators include hysteresis elements and more general nonlinear dynamical systems for which a given input may give rise to many outputs depending on the initial conditions. Dealing with relations also takes account of the fact that the

* Received by the editors July 2, 1980, and in revised form June 17, 1981.

[†] Department of Pure Mathematics and Mathematical Statistics, Cambridge University, Cambridge CB2 1SB, England.

output of a system may be subject to small fluctuations that are not exactly determined by the input.

We can define the usual operations for relations as well as for operators. Thus if $T \subseteq E_1 \times E_2$, $T' \subseteq E_1 \times E_2$ and $T_0 \subseteq E_2 \times E_3$, define

$$\begin{aligned}
 (1.1) \quad & T + T' = \{(x_1, x_2 + x'_2); (x_1, x_2) \in T, (x_1, x'_2) \in T'\}, \\
 & \lambda T = \{(x_1, \lambda x_2); (x_1, x_2) \in T\}, \\
 & T^{-1} = \{(x_2, x_1); (x_1, x_2) \in T\}, \\
 & T_0 T = \{(x_1, x_3); \text{for some } x_2, (x_1, x_2) \in T, (x_2, x_3) \in T_0\}.
 \end{aligned}$$

We shall see that other concepts carry over too: causality, for instance, is very important in our development. For this reason, we look carefully at truncations, extended spaces and full spaces, building on the ideas of [4] and [11].

2. Extended spaces: definitions and basic properties. In defining causality, one usually uses some “forgetting” operators P^t , such that if x, x' are two signals, then $P^t x = P^t x'$ if and only if x and x' cannot be distinguished before time t . Often the signals lie in a space $L^p((0, \infty); \mathbb{R}^n)$ and the function $P^t x$ is defined to be x on $[0, t]$ and 0 thereafter. However, P^t has to be defined differently on some other spaces of interest. For instance if $E = C_b([0, \infty); \mathbb{R}^n)$, the continuous bounded functions from $[0, \infty)$ into \mathbb{R}^n with $\|x\| = \sup |x(t)|$, then truncation to zero after time t will generally produce a discontinuous function. The natural choice for P^t in this space is to make $P^t x$ agree with x up to time t and then be constantly equal to $x(t)$. If we wanted to work in the subspace of E for which $x(t) \rightarrow 0$ as $t \rightarrow \infty$, we would have to make some further modification. Clearly we need to develop the theory in a way that depends only on certain axioms holding for the P^t which are sufficiently general to cover all commonly occurring cases.

2.1 An axiomatic approach to extended spaces. For our definitions, the signals x, x' are assumed to lie in a (real or complex) normed linear space E , and we let $\mathcal{B}(E)$ denote the normed algebra of continuous linear operators on E . We let \mathcal{T} denote $[0, \infty)$, the set of times of interest, and in what follows t and s will always denote elements of \mathcal{T} .

DEFINITION 2.1. A *normed signal space* is a normed linear space E endowed with a map $P: [0, \infty) \rightarrow \mathcal{B}(E)$ enjoying the following properties:

- (1) for all s and all $t \geq s$, $P^s P^t = P^s$;
- (2) if $x, y \in E$ and $P^t x = P^t y$ for all t , then $x = y$;
- (3) for each t , $\|P^t\| \leq 1$;
- (4) for each $x \in E$, $\sup_t \|P^t x\| = \|x\|$.

If also the following holds, (E, P) is said to be *full*:

- (5) if $(x_t, t \in \mathcal{T})$ is a family of elements of E satisfying
 - (i) for all s and all $t \geq s$, $P^s x_t = x_s$, and
 - (ii) $\sup_t \|x_t\| < \infty$,
 then there is some $x \in E$ such that for all t , $x_t = P^t x$.

Generalization 1. One might wish to take \mathcal{T} as some other set, such as \mathbb{Z}_+ for discrete time systems. This will not affect the subsequent theory, provided that \mathcal{T} is partially ordered and contains a monotonic cofinal sequence (t_n) (i.e., a sequence such that $t_{n+1} \geq t_n$ and for any $t \in \mathcal{T}$, some $t_n \geq t$). Any subset of \mathbb{R} has this property.

Generalization 2. To define a signal space when E does not have all the structure of a normed linear space, one just retains as much of the above definition as makes

sense. Thus if E is an arbitrary set, the P' can be any maps from E to E satisfying axioms (1) and (2). For a topological signal space, we would require E to have a topology and the P' to be continuous as well. For a linear signal space, we would require E to be a linear space and the P' to be linear maps.

The extended space. Some descriptive terms are useful. An element of $P'E$ will be called a *t-initial segment* or generally an *initial segment*. The initial segments of $x \in E$ are the elements $P^t x$. Thus (2) says that an element of E is uniquely determined by its initial segments. If $x_t = P^t x$ then (1) says that $(x_t, t \in \mathcal{T})$ is a *compatible family* in the sense that for all s and all $t \geq s$, $P^s x_t = x_s$. Now we just define the extended space E_e to be the set of all compatible families of initial segments. If $x = (x_t, t \in \mathcal{T}) \in E_e$, we define $P_e^t: E_e \rightarrow E$ by $P_e^t x = x_t$. There is a natural map $J: E \rightarrow E_e$ taking an element of E to its own initial segments and by (2), J is an injection. Further $P_e^t Jx = P^t x$ for $x \in E$.

Now we want to see how much of the structure of E goes over to E_e . If E is a topological signal space, then E_e inherits a subspace topology from the product topology on $E^{\mathcal{T}}$. With this topology J and the P_e^t are continuous, but J is not necessarily a homeomorphism. If E is a linear signal space, then the natural definitions make E_e a linear space, and P_e^t and J are then linear maps. In this case we may identify x with Jx so that P_e^t extends P^t . Henceforth we make this identification so $E \subseteq E_e$ and we denote P_e^t by P^t without risk of confusion. Then if $x \in E_e$ and $t \geq s$, $P^s P^t x = P^s x_t = x_s = P^s x$ and so (1) still holds for the extended P^t and (2) holds trivially.

The normed case. First observe that each P^t is a projection, (i.e., $P^t P^t = P^t$) so $\|P^t\|$ is 0 or 1. Also, if $t \geq s$, $\|P^s x\| \leq \|P^t x\|$ so $\|P^s\| \leq \|P^t\|$. The extra condition (5) for a full signal space is that if $x \in E_e$ and $\sup_t \|P^t x\| < \infty$ then $x \in E$. This is vital in stability proofs; typically, one shows that the output has norm-bounded initial segments, and needs to deduce that it lies in the original (unextended) space. Not all normed signal spaces are full.

However, Feintuch and Sacks [4] showed that a reflexive signal space is automatically full. (Although their proof was in a more specialized setting, one easily checks that it remains valid under our weaker assumptions.) We shall now show that if E is not full there is a unique way of making it up to a full signal space.

FILLING-UP THEOREM 2.2. *Let (E, P) be a normed signal space. Then there is a unique full normed signal space (\hat{E}, P) with $E \subseteq \hat{E} \subseteq E_e$. If E is a Banach space, so is \hat{E} .*

Proof. The only possible choice for \hat{E} is

$$\{x \in E_e: \sup_t \|P^t x\| < \infty\}.$$

For if x lies in this set and \hat{E} is full, $x \in \hat{E}$, while if $x \in \hat{E}$, the condition must certainly hold. Thus \hat{E} is as stated and axiom (5) automatically holds for it. Then \hat{E} is of course a linear subspace of E_e , and the norm on it must be $\|x\| = \sup_t \|P^t x\|$ to satisfy axiom (4). We note that this is a norm extending the original norm on E , so it is safely denoted by the same symbol. Now (3) holds too, so only the completeness assertion remains. Let $(x^{(n)})$ be a Cauchy sequence in \hat{E} , and suppose E is complete. Then $P^t x^{(n)}$ is Cauchy in E and so converges to $x_t \in E$. The $(x_t, t \in \mathcal{T})$ form a compatible family and so define an element $x \in E_e$, for which $\sup_t \|P^t x\| \leq \sup_n \|x^{(n)}\| < \infty$, so $x \in \hat{E}$. Then

$$\|x^{(n)} - x\| = \sup_t \|P^t x^{(n)} - x_t\| \leq \sup \{\|x^{(m)} - x^{(n)}\|: m \geq n\} \rightarrow 0,$$

as $n \rightarrow \infty$. Thus \hat{E} is complete as required. \square

Having proved Theorem 2.2, we see that stability theorems for full signal spaces are not too restrictive. When E is not full, \hat{E} will usually be rather obvious, as in Example 2 below. Nevertheless, the problem of proving stability theorems in nonfull spaces remains.

For the particular case of Hilbert spaces, they are reflexive and therefore already full. Furthermore the axioms then imply that each P^t is an orthogonal projection, and that $P^t x \rightarrow x$ as $t \rightarrow \infty$. Thus the only way that a Hilbert signal space fails to be a Hilbert resolution space is if $P^t x$ is not required to be continuous in t .

Example 1. If $E = L^p((0, \infty); \mathbb{R}^n)$ for $1 \leq p \leq \infty$, and $P^t x$ is defined by truncating x to zero after time t , then E is already full (although if $p = 1$ or $p = \infty$, E is not reflexive, and when $p = \infty$, $P^t x \not\rightarrow x$ as $t \rightarrow \infty$). The extended space E_e is the set of functions whose restrictions to $[0, t]$ lie in $L^p(0, t)$ for each t .

Example 2. If we take our time set as \mathbb{Z}_+ , and E as the space c of convergent sequences with the sup norm, and if $P^n x$ is defined by truncating the sequence to zero after the n th term, then we have a Banach signal space that is not full. It fills up to give l^∞ , and the extended space consists of all sequences.

Example 3. Let E be the space of all functions from $[0, \infty)$ to \mathbb{R}^n that are square integrable and absolutely continuous with square integrable derivative. If $|\cdot|$ denotes the Euclidean norm on \mathbb{R}^n , let

$$\|x\|^2 = \int_0^\infty |x(t)|^2 + |\dot{x}(t)|^2 dt.$$

This makes E a Hilbert space. To define projections on this it is natural to require $P^t x$ to depend only on the restriction of x to $[0, t]$ and to agree with x on this interval. Subject to this, $P^t x$ must have the minimum norm possible (to make $\|P^t\| = 1$) and one easily shows that this forces

$$P^t x(s) = \begin{cases} x(s) & \text{for } s \leq t, \\ x(t) e^{t-s} & \text{for } s \geq t. \end{cases}$$

This makes (E, P) a Hilbert signal space, and so automatically full.

Alternative definition of E_e . We could define a metric on E by

$$d(x, y) = \min(1, \inf \{t^{-1} : P^t x = P^t y\}).$$

(The 1 is there only to make d finite. The metric tells us for how long the signals are equal to one another.) Then E_e can be regarded as the metric completion of (E, d) . The extension of the metric will still be given by the above formula. This viewpoint allows us to define the extension S_e of any subset S of E as its closure in (E_e, d) . Explicitly,

$$S_e = \{x \in E_e : P^t x \in P^t S \text{ for all } t\}.$$

If S is a linear subspace, so is S_e .

2.2 Extension of causal relations. If E_1, E_2 are normed signal spaces, then $E_1 \times E_2$ is naturally a normed signal space if we define $\|(x_1, x_2)\|$ as $\max(\|x_1\|, \|x_2\|)$ or $(\|x_1\|^p + \|x_2\|^p)^{1/p}$ for some $p \geq 1$. Thus every relation $T \subseteq E_1 \times E_2$ has an extension $T_e \subseteq (E_1 \times E_2)_e = E_{1e} \times E_{2e}$. We want to examine the connections between properties of T and properties of T_e . First we need to define causality.

DEFINITION 2.3. $T \subseteq E_1 \times E_2$ is *causal* if whenever $(x_1, x_2) \in T$ and $t \in \mathcal{T}$, there is some x'_2 with $(P^t x_1, x'_2) \in T$ and $P^t x'_2 = P^t x_2$.

Remark. It is not entirely clear what the best definition of causality for relations is, but the above is adequate for our purposes. It implies that the domain of T is P' invariant, and reduces to the familiar definition for operators that if $x, x' \in \text{dom}(T)$ and $P'x = P'x'$ then $P'Tx = P'Tx'$.

PROPOSITION 2.4. *If $T \subseteq E_1 \times E_2$ is causal, so is T_e . If T is a causal operator, so is T_e . If T is a causal linear operator, so is T_e .*

Proof. Let $(x_1, x_2) \in T_e$. Then there exists $(x'_1, x'_2) \in T$ with $P'x'_1 = P'x_1$ and $P'x'_2 = P'x_2$. By causality of T there exists $(P'x'_1, x''_2) \in T$ with $P'x''_2 = P'x'_2$, and so $(P'x_1, x''_2) \in T_e$ and we have causality of T_e . The rest is trivial.

2.3 Numerical properties of relations. We recall the definition of gain, since our approach to numerical ranges is analogous. The natural way to define the gain of a relation $T \subseteq E_1 \times E_2$ on normed spaces is

$$(2.1) \quad g(T) = \inf \{M \geq 0: \text{if } (x_1, x_2) \in T, \|x_2\| \leq M\|x_1\|\}.$$

This obviously satisfies (if T, T', T_0 are as in (1.1))

$$(2.2) \quad g(T + T') \leq g(T) + g(T'), \quad g(\lambda T) = |\lambda|g(T), \quad g(T_0 T) \leq g(T_0)g(T).$$

However, $g(T)$ may be infinite because of the behavior of T near $x_1 = 0$. One is more often interested in the large scale properties of T and so uses its gain in the large,

$$g_l(T) = \inf \{M \geq 0, \exists \beta: \forall (x_1, x_2) \in T, \|x_2\| \leq M\|x_1\| + \beta\}.$$

This also satisfies (2.2) and

$$g_l(T) \leq g(T).$$

When T is a relation on extended spaces, one defines its extended gain $g_e(T)$ and extended gain in the large $g_{el}(T)$ as follows:

$$g_e(T) = \inf \{M \geq 0: \forall (x_1, x_2) \in T \text{ and } \forall t, \|P'x_2\| \leq M\|P'x_1\| + t\},$$

$$g_{el}(T) = \inf \{M \geq 0: \exists \beta: \forall (x_1, x_2) \in T \text{ and } \forall t, \|P'x_2\| \leq M\|P'x_1\| + \beta + t\}.$$

Each of these satisfies (2.2) and $g_{el}(T) \leq g_e(T)$ and we have the following proposition, whose proof is easy.

PROPOSITION 2.5. *If $T \subseteq E_1 \times E_2$ is causal, then*

$$g_e(T_e) \leq g(T) \quad \text{and} \quad g_{el}(T_e) \leq g_l(T).$$

Now we are ready to define the numerical range of a relation. For brevity we write $f\|x$ (pronounced “ f supports x ”) whenever $f \in E^*$, $x \in E$ and $\|f\| \cdot \|x\| = f(x) = 1$. Then the numerical range $V(T)$ of a relation $T \subseteq E \times E$ is defined by

$$V(T) = \{f(y): (x, y) \in T, f\|x\}.$$

Observe that this is a subset of \mathbb{R} or \mathbb{C} according as E is a real or complex normed space. It satisfies

$$(2.3) \quad V(T + T') \subseteq V(T) + V(T'), \quad V(\lambda T) = \lambda V(T).$$

The numerical radius is

$$(2.4) \quad v(T) = \sup \{|z|: z \in V(T)\}$$

and it clearly satisfies

$$(2.5) \quad v(T) \leq g(T).$$

The large scale numerical range has to be defined rather more carefully. Let $\tilde{\mathbb{C}}$ denote \mathbb{C} compactified by adjoining a circle C_∞ of points at ∞ . Then the numerical range in the large is

$$V_l(T) = \{p \in \tilde{\mathbb{C}} : \exists (x_n, y_n) \in T, f_n \|x_n, f_n(y_n) \rightarrow p \text{ and } \|x_n\| \rightarrow \infty\}.$$

Then (2.3) hold for the extended numerical range provided that we make the following conventions about sums of sets that may contain points at ∞ :

- (i) if $z \in \mathbb{C}$ and $p \in C_\infty$, $z + p = p$;
- (ii) if $p, q \in C_\infty$ and are not opposite each other, $p + q$ is the shorter closed arc of C_∞ joining them;
- (iii) if p, q are opposite points of C_∞ , $p + q = \tilde{\mathbb{C}}$.

Also, it is easy to see that

$$V_l(T) \subseteq \overline{V(T)},$$

where the closure on the right is taken in $\tilde{\mathbb{C}}$, and also that the analogue in the large of (2.5) holds: in particular, if $g_l(T)$ is finite, $V_l(T)$ does not contain any points at ∞ . For relations on extended spaces, we have the following definitions:

$$V_e(T) = \{f(P^t y) : (x, y) \in T, f \|P^t x\},$$

$$V_{el}(T) = \{p \in \tilde{\mathbb{C}} : \exists t_n \in \mathcal{T}, (x_n, y_n) \in T, f_n \|P^{t_n} x_n, f_n(P^{t_n} y_n) \rightarrow p, \|P^{t_n} x_n\| \rightarrow \infty\}.$$

Each of these satisfies (2.3). Moreover,

$$V_{el}(T) \subseteq \overline{V_e(T)},$$

where the closure is in $\tilde{\mathbb{C}}$, the analogues of (2.5) hold and we have the following proposition.

PROPOSITION 2.6. *If $T \subseteq E \times E$ is causal, then*

$$V_e(T_e) \subseteq V(T) \quad \text{and} \quad V_{el}(T_e) \subseteq V_l(T).$$

Before proceeding, we briefly explain the relation of the numerical range to the more usual concepts in system theory of positivity and passivity. In a real Hilbert space, given x there is a unique $f \|x$, given by $f(y) = \langle y, x \rangle / \|x\|^2$. Thus a positive operator on a real Hilbert space is one satisfying $V(T) \subseteq [0, \infty)$, and it is strictly positive if $\inf V(T) > 0$. On a real Hilbert space, a passive operator has $V_{el}(T) \subseteq [0, \infty]$ and a strictly passive one has $\inf V_{el}(T) > 0$.

3. Numerical ranges and stability. Our stability theorem relates to the system shown in Fig. 1, namely,

$$(3.1) \quad (e, y) \in T_1, \quad (y, u - e) \in T_2,$$

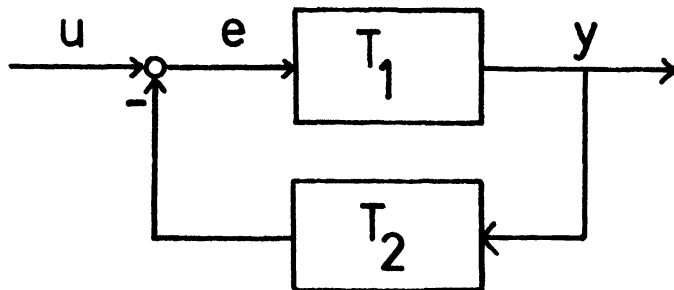


FIG. 1.

where T_1, T_2 are causal relations on a full normed signal space E . The whole system is thus specified by the relation

$$(3.2) \quad (u, y) \in (T_1^{-1} + T_2)^{-1}.$$

We suppose that $u \in E$ and that when T_1, T_2 are extended according to Proposition 2.4, there is at least one solution with $e \in E_e, y \in E_e$. The theorem may be regarded as a generalization to non-Hilbert spaces of the classical results of Zames [15], [16].

THEOREM 3.1. *In the situation above, let δ be the distance of the origin from the set $V(T_1^{-1}) + V(T_2)$. If $\delta > 0$ then in fact $y \in E$ and*

$$(3.3) \quad g((T_1^{-1} + T_2)^{-1}) \leq \delta^{-1}.$$

Let δ_l be the distance of the origin from the set $V_l(T_1^{-1}) + V_l(T_2)$. If $\delta_l > 0$ then $y \in E$ and

$$(3.4) \quad g_l((T_1^{-1} + T_2)^{-1}) \leq \delta_l^{-1}.$$

Proof. We start with the easier (3.3). By hypothesis, we have some $(e, y) \in T_{1e}$ and $(y, u - e) \in T_{2e}$. We claim that for any t , $\|P^t y\| \leq \delta^{-1} \|P^t u\|$, and since $u \in E$ and E is full, this is enough to prove that $y \in E$ and $\|y\| \leq \delta^{-1} \|u\|$, as required. For if $P^t y = 0$, the claim is trivial. Otherwise, let $f \|P^t y$, and then by definition of V_e ,

$$f(P^t e) \in V_e(T_{1e}^{-1}) \quad \text{and} \quad f(P^t(u - e)) \in V_e(T_{2e}).$$

Thus by Proposition 2.6 and the hypothesis of the Theorem 3.1 $|f(P^t u)| \geq \delta$. So

$$\|P^t y\| \leq \|P^t y\| \delta^{-1} \|f\| \|P^t u\| = \delta^{-1} \|P^t u\|,$$

as required. For (3.4) let $0 < \eta < \delta_l$, $|z| \leq \eta$, and then by Proposition 2.6,

$$z \notin V_{el}(T_{1e}^{-1}) + V_{el}(T_{2e}).$$

So there is some $\varepsilon > 0$ and some finite β such that if $f \|P^t y, (y, e) \in T_{1e}^{-1}, (y, u - e) \in T_{2e}$ and $f(P^t u) \in N_\varepsilon(z)$ then $\|P^t y\| \leq \beta$. Now, by compactness, finitely many of the discs $N_\varepsilon(z)$ cover $\{z : |z| \leq \eta\}$, so taking the maximum of the corresponding β , we have that

$$\text{if } |f(P^t u)| \leq \eta \quad \text{then} \quad \|P^t y\| \leq \beta.$$

But, as above, if $|f(P^t u)| \geq \eta$, $\|P^t y\| \leq \eta^{-1} \|P^t u\|$. Thus in either case, $\|P^t y\| \leq \eta^{-1} \|P^t u\| + \beta$, so when $u \in E$, $y \in E$ by fullness, and $\|y\| \leq \eta^{-1} \|u\| + \beta$. This holds for any $\eta < \delta_l$, thus proving (3.4). \square

Theorem 3.1 is related to Safonov's general stability theorem [13], though it is a bit more specialized: Safonov allows more general sets than $V(T_1^{-1})$ and $V(T_2)$. The advantage of sticking to numerical ranges is that they are relatively well studied and there is a lot of useful machinery available, so they lead very directly to practical stability tests.

If E is a real Hilbert space, then it is clear from our remarks at the end of § 2.3 that Theorem 3.1 includes the standard results on passive and positive operators [16]. The results of this section show that in fact E may be any normed space, and also help in the final interpretation. Practicalities of numerical range and norm calculation will probably dictate that in most cases the theorems are applied in much the same way as standard passivity results, although in principle they can do better because the numerical range will usually be a strict subset of the region defined by the norm.

To make Theorem 3.1 useful, we shall have to show how to estimate numerical ranges. The key result is the following proposition, which we have already pointed out in passing.

PROPOSITION 3.2. *If $T = C + R$, then $V(T) \subseteq V(C) + V(R)$.*

Proposition 3.2 is useful when we know the numerical range of C and can put some bounds on that of R , which brings us to the question of when we can find a numerical range exactly. Apart from trivial cases where explicit calculation is possible, it appears that the only time when $V(T)$ (or rather $\overline{V(T)}$, which is good enough for Theorem 3.1) can be found is when T is linear and normal. (On a Hilbert space, a normal operator is one that commutes with its adjoint: in general the definition is rather more subtle [2].)

THEOREM 3.3. *If T is a continuous normal linear operator on E , with spectrum $\sigma(T)$, then $\overline{V(T)} = \mathbb{F} \cap \overline{\text{co}} \sigma(T)$, where $\overline{\text{co}}$ is the closure of the convex hull and \mathbb{F} ($=\mathbb{R}$ or \mathbb{C}) is the ground field of E .*

Proof. See [2].

Thus if $T = C + R$ and C is normal, we can now write

$$V(T) \subseteq \overline{\text{co}} \sigma(C) + \Delta \|R\|,$$

where Δ is the unit disk in \mathbb{C} or the interval $[-1, 1]$ in \mathbb{R} . We recall the definition of sector-boundedness for relations, another notion introduced by Zames [15], [16].

DEFINITION 3.4. If E is a real normed space and $T \subseteq E \times E$ satisfies $(x, y) \in T \Rightarrow \|y - Kx\| \leq \mu \|x\|$, then T is said to be *sector-bounded* with *sector bounds* $\alpha = K - \mu$ and $\beta = K + \mu$ and we write $T \in \text{sector} [\alpha, \beta]$.

Note that if E is a Hilbert space this reduces to the usual definition $\langle Tx - \alpha x, Tx - \beta x \rangle \leq 0$.

PROPOSITION 3.5. $T \in \text{sector} [\alpha, \beta]$ implies $V(T) \subseteq [\alpha, \beta]$.

Proof. Let $C = (\alpha + \beta)/2$ in the above.

Notice that if we are given sector bounds we have limits on $V(T)$ at once, while even if the relation is not sector-bounded it may sometimes be possible to bound $V(T)$ by some other method. One of the most helpful facts is that if E is a Hilbert space then $T \in \text{sector} [\alpha, \beta]$ and $\alpha\beta > 0$ implies $V(T^{-1}) \subseteq [\beta^{-1}, \alpha^{-1}]$.

Another case where we can estimate $V(T)$ rather easily is when $E = L^2(H)$ where H is a real Hilbert space and T is a linear transfer operator, with convolution kernel γ and transfer function G . Then, writing η for the Fourier transform of x and with $f\|x$, we have

$$f(Tx) = \int_0^\infty \langle \gamma^* x(t), x(t) \rangle \frac{dt}{\|x\|^2} = \int_{-\infty}^\infty \langle G(i\omega)\eta(\omega), \eta(\omega) \rangle \frac{d\omega}{\|\eta\|^2}.$$

This clearly lies in the convex hull of the points $V(G(i\omega))$, so we have shown that

$$V(T) \subseteq \mathbb{R} \cap \overline{\text{co}} \bigcup_{\omega \in \mathbb{R}} V(G(i\omega)).$$

Now $G(i\omega)$ is a matrix in any basis for H so $V(G(i\omega))$ should be relatively easy to find. In particular, if H is finite dimensional there are many numerical methods available [1], [8], [9]. In the simplest case of all, when $H = \mathbb{R}$, we can see that $f(Tx)$ only depends on $\text{Re } G(i\omega)$, so

$$V(T) \subseteq \overline{\text{co}} \bigcup_{\omega \in \mathbb{R}} \text{Re } (G(i\omega)) = [\inf \text{Re } (G(i\omega)), \sup \text{Re } (G(i\omega))].$$

The large scale numerical range $V_l(t)$ can be estimated by methods similar to those we have outlined for $V(T)$. The basic idea is stated in the following proposition.

PROPOSITION 3.6. *If $T, T' \subseteq E \times E$, $\phi: \mathbb{R} \rightarrow \mathbb{R}$ has $\phi(u)/u \rightarrow 0$ as $u \rightarrow \infty$, and $\forall (x, y) \in T, \exists (x, y') \in T'$ with $\|y - y'\| \leq \phi(\|x\|)$, then*

$$V_l(T) \subseteq V_l(T').$$

The proof is trivial, and to apply the result, one will typically take T' such that $V(T')$ can be bounded by the methods outlined previously. For instance, let T be an ordinary linear dynamical system, possibly starting from nonzero initial conditions. Then T' could be taken as the same system starting from zero initial conditions, which is a convolution operator, for which we have already shown how to estimate $V(T')$.

4. Examples

Example 4.1. In our standard system, let $E = L^2(\mathbb{R})$, let $T_1 \in \text{sector } [0, \beta]$ and let T_2 be a linear transfer operator with transfer function G . Let us assume T_1 and T_2 are causal so we know how to calculate their numerical ranges. We have

$$V(T_1^{-1}) + V(T_2) \subseteq [\beta^{-1}, \infty) + [\inf \operatorname{Re} G(i\omega), \sup \operatorname{Re} G(i\omega)]$$

so the system is stable if $\inf \operatorname{Re} G(i\omega) > -\beta^{-1}$. In graphical terms, the Nyquist locus of G must be bounded away from, and lie to the right of, a straight line through $-\beta^{-1}$ which is parallel to the imaginary axis. This is obviously a form of circle criterion and can be poleshifted to give the more usual form.

Example 4.2. If T_2 has an $n \times n$ matrix transfer function G , the only difference from the single loop case is that we have to find a way to get a decent estimate of $V(G(i\omega))$. The simplest possibility is to write $G = C + R$, where C is the matrix of diagonal elements of G . Then Proposition 3.2 gives us

$$V(G(i\omega)) \subseteq \operatorname{co} \{G(i\omega)_{jj} : 1 \leq j \leq n\} + \Delta \|R(i\omega)\|.$$

That is, at each frequency ω we plot the diagonal elements of $G(i\omega)$ as points in \mathbb{C} , surrounded by disks of radius $\|R(i\omega)\|$. There is no need to take the convex hull, because for stability we only have to check that the convex hull lies in a convex set, namely the half-plane $\operatorname{Re} z > -\beta^{-1}$. This is so if each of the disks lies in that half-plane. As ω varies, the discs sweep out bands, and the system is stable if the bands lie in the half-plane.

One can be more sophisticated: for example, MacFarlane's characteristic locus stability criterion [10] requires one to plot eigenvalue loci, and these can be banded. Wilkinson [14] shows that every complex matrix is equivalent under unitary transformation to an upper triangular matrix:

$$U^*GU = \tilde{G}, \quad \tilde{G}_{kj} = 0 \quad \text{if } k > j,$$

where U has columns obtained from the generalized eigenvectors of G by Schmidt orthonormalization. Now we may apply Proposition 3.2 to \tilde{G} instead of G . If G happens to be normal, \tilde{G} is diagonal and the bands have zero width. In that case, we obtain a result analogous to the generalized circle criterion of Freedman, Falb and Zames [5]. (They took the set of times to be a locally compact Abelian group, which we have taken as \mathbb{R} .)

A further refinement is to write $R = H + iJ$, with H and J Hermitian and therefore normal. We now have

$$V(R) \subseteq V(H) + iV(J)$$

and $V(H)$ and $V(J)$ are the intervals bounded by the minimum and maximum eigenvalues of H and J . The disk in the above has now been replaced by a rectangle contained in it, making the bands somewhat narrower [12].

Example 4.3. The infinite dimensional version of the Freedman, Falb and Zames [5] criterion mentioned above can be obtained from Theorems 3.1 and 3.3. If we add Proposition 3.2 we have an obvious generalization in which a suitably "smudged out"

version of the spectrum of the linear part has to avoid a disk on $[-\alpha^{-1}, -\beta^{-1}]$ as diameter. The result we obtain will be akin to that of Cook [3], though he works in terms of diagonal elements of a specific representation of the linear element as an infinite matrix instead of in terms of the spectrum. Our result also extends to the case where E is not a Hilbert space as long as we can ensure that $V(T_1^{-1}) \subseteq [\beta^{-1}, \alpha^{-1}]$. This may require a condition different from $T_1 \subseteq \text{sector} [\alpha, \beta]$.

REFERENCES

- [1] C. S. BALLANTINE, *Numerical range of a matrix; some effective criteria*, Lin. Alg. Appl., 19 (1978), pp. 117–188.
- [2] F. F. BONSALL AND J. DUNCAN, *Numerical Ranges of Operators on Normed Spaces and of Elements of Normed Algebras*, University Press, Cambridge, 1971.
- [3] P. A. COOK, *Circle criteria for stability in Hilbert space*, this Journal, 13 (1975), pp. 593–610.
- [4] A. FEINTUCH AND R. SAEKS, *Extended spaces and the resolution topology*, Internat. J. Control, 33 (1981), pp. 347–354.
- [5] M. I. FREEDMAN, P. L. FALB AND G. ZAMES, *A Hilbert space stability theory over locally compact Abelian groups*, SIAM J. Control, 7 (1969), pp. 479–495.
- [6] I. C. GOHBERG AND M. G. KREIN, *Volterra operators in Hilbert Space*, AMS Transl. of Math. Monographs 24, American Mathematical Society, Providence, RI, 1970.
- [7] ———, *Factorization of operators in Hilbert space*, Acta Sci. Math. Szeged, 25 (1964); English transl. in Amer. Math. Soc. Transl. (2) 51 (1966), pp. 155–188.
- [8] C. R. JOHNSON, *A Gershgorin inclusion set for the field of values of a finite matrix*, Proc. Amer. Math. Soc., 41 (1973), pp. 57–60.
- [9] ———, *Numerical determination of the field of values of a general complex matrix*, SIAM J. Numer. Anal., 15 (1978), pp. 595–602.
- [10] A. G. J. MACFARLANE AND I. POSTLETHWAITE, *The generalized Nyquist stability criterion and multivariable root loci*, Internat. J. Control, 25 (1977), pp. 81–127.
- [11] A. I. MEES, *Dynamics of Feedback Systems*, Wiley, Chichester, 1980.
- [12] A. I. MEES AND D. P. ATHERTON, *Domains containing the field of values of a matrix*, Lin. Alg. Appl., 26 (1979), pp. 289–296.
- [13] M. G. SAFONOV, *On stability theory*, Proc. IEEE Conference on Decision and Control, San Diego, 1979.
- [14] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon, London, 1965.
- [15] G. ZAMES, *On the stability of nonlinear, time-varying feedback systems*, Proc. 1964 Nat. Electronics Conf., vol. 20, pp. 725–730.
- [16] ———, *On the input-output stability of time-varying nonlinear feedback systems: Part I: Concepts of loop gain, conicity and positivity*, IEEE Trans. Automat. Control, AC-11 (1966), pp. 228–238; *Part II: Circles in the frequency plane and sector nonlinearities*, Ibid., AC-11 (1966), pp. 466–476.

OBSERVABILITY AND CONTROLLABILITY FOR SMOOTH NONLINEAR SYSTEMS*

A. J. VAN DER SCHAFT†

Abstract. The definition of a smooth nonlinear system as proposed recently by Willems, is elaborated as a natural generalization of the more common definitions of a smooth nonlinear input–output system. Minimality for such systems can be defined in a very direct geometric way, and already implies a usual notion of observability, namely, local weak observability. As an application of this theory, it is shown that observable nonlinear Hamiltonian systems are necessarily controllable, and vice versa.

1. Introduction. In the last decade there has been important work on a differential geometric approach to nonlinear input state–output systems, which in local coordinates have the form

$$(1.1) \quad \dot{x} = g(x, u), \quad y = h(x),$$

where x is the *state* of the system, u is the *input* and y the *output* (for a survey see Brockett [3]). Most of the attention has been directed to the formulation in this context of fundamental system theoretic concepts like controllability, observability, minimality and realization theory. Some basic papers are, for instance, Hermann–Krener [6], Sussmann [12], and recently Jakubczyk [9].

In spite of some very natural formulations and elegant results which have been achieved, there are certain disadvantages in the whole approach, from which we summarize the following points.

a) Normally the equations

$$(1.2) \quad \dot{x} = g(x, u)$$

are interpreted as a family of vector fields on a manifold parametrized by u ; i.e., for every fixed \bar{u} , $g(\cdot, \bar{u})$ is a globally defined vector field. As noted already by Brockett [4], Takens [15] and Willems [17] there are some serious objections to this setting. In fact, the last author proposes another framework by looking at (1.2) as a coordinatization of

$$\begin{array}{ccc} B & \xrightarrow{\quad g \quad} & TX \\ & \searrow & \swarrow \\ & X & \end{array}$$

where B is a *fiber bundle* above the state space manifold X and the fibers of B are the *state dependent* input spaces, while TX is as usual the tangent bundle of X (the possible velocities at every point of X).

b) The usual definition of *observability* for this kind of system (cf. [6]) has some drawbacks. In fact, observability is defined as *distinguishability*; i.e., for every x_1 and x_2 (elements of X) there exists a *certain* input function (in principle dependent on x_1 and x_2) such that the output function of the system starting from x_1 under the influence of this input function is different from the output function of the system starting from x_2 under the influence of this same input function. Of course, from a practical point of view this notion of observability is not very useful, and also is not in accord with the usual definition of observability or reconstructibility for general systems (cf. [10]).

* Received by the editors December 3, 1980, and in final form June 5, 1981.

† Mathematics Institute, P.O. Box 800, 9700 AV Groningen, the Netherlands.

Hence, despite the work of Sussmann [13] on *universal* inputs, i.e., input functions which distinguish between every two states x_1 and x_2 , this approach remains unsatisfactory.

c) In the class of nonlinear systems (1.1) *memoryless* systems

$$(1.3) \quad y = h(u)$$

are not included! Of course, one could extend the system (1.1) to the form

$$(1.4) \quad \dot{x} = g(x, u), \quad y = h(x, u),$$

but this gives, if one wants to regard observability as distinguishability, the following rather complicated notion of observability. As can be seen from [2], distinguishability of (1.4) with $y \in \mathbb{R}^p$, $u \in \mathbb{R}^m$ and $x \in \mathbb{R}^n$ is equivalent to distinguishability of

$$(1.5) \quad \dot{x} = g(x, u), \quad \bar{y} = \bar{h}(x)$$

where $\bar{h} : \mathbb{R}^n \rightarrow (\mathbb{R}^p)^{\mathbb{R}^m}$ is defined by $\bar{h}(x)(u) = h(x, u)$.

Checking the Lie algebra conditions for distinguishability as described in [6] for the system (1.5) is not very easy!

d) As noted by Willems [17], in a description of a physical system ("physical" interpreted in a broad sense) it is often not clear how to distinguish a priori between inputs and outputs. Especially in the case of a nonlinear system, it could be possible that a separation of what we shall call *external variables* in input variables and output variables should be interpreted only *locally*. An example is the (nearly) ideal diode given by the I - V characteristic in Fig. 1. For $I < 0$ it is natural to regard I as the input and V as the output, while for $V > 0$ it is natural to see V as the input and I

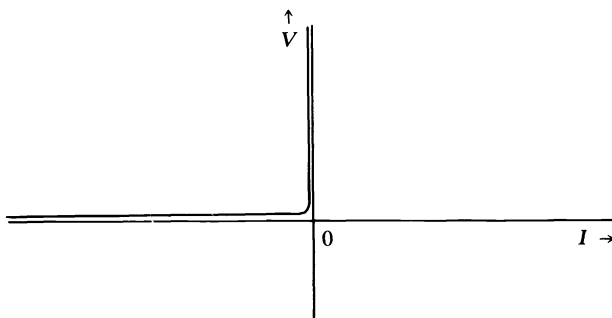


FIG. 1

as the output. Around 0 an input-output description should be given in the scattering variables $(I - V, I + V)$. Moreover, in the case of nonlinear systems it can happen that a *global* separation of the external variables in inputs and outputs is simply not possible! This results in a definition of a system which is a generalization of the usual input-output framework. It appears that various notions like the definitions of autonomous (i.e., without inputs), memoryless, time-reversible, Hamiltonian and gradient systems are very natural in this framework (see [16], [17]).

The organization of this paper is as follows. In § 2 we give the definition of a nonlinear system as proposed in [17], and give some connections with the more usual input-output settings. In § 3 we define *minimality* of such a system and derive *local* conditions from this global definition. It is very surprising that this results in the same

kind of conditions as given in recent papers on nonlinear disturbance decoupling; see [7], [8] and especially the setting proposed by Nijmeijer [11]. These local conditions imply local weak observability for systems which locally can be represented in an input–output form without a feedthrough term. Finally, in § 4 the definition of minimality is tested in the case of Hamiltonian systems as defined in [16], and we can derive the theorem that an “observable” full Hamiltonian system is necessarily “controllable”, and vice versa. Surprisingly, it appears that this need not hold for gradient systems!

2. Definition of a smooth nonlinear system. As proposed in [17] and argued in [16], [17], smooth (say C^∞) systems can be represented in the commutative diagram

$$(2.1) \quad \begin{array}{ccc} B & \xrightarrow{f} & TX \times W \\ & \searrow \pi \quad \swarrow \pi_X & \\ & X & \end{array}$$

where (all spaces are smooth manifolds) B is a fiber bundle above X with projection π , TX is the tangent bundle of X , π_X the natural projection of TX on X and f is a smooth map. W is the space of external variables (think of the inputs *and* the outputs). X is the state space and the fiber $\pi^{-1}(x)$ in B above $x \in X$ represents the space of inputs (to be seen initially as *dummy* variables), which is state dependent (think of forces acting at different points of a curved surface).

This definition formalizes the idea that at every point $x \in X$ we have a set of possible velocities (elements of TX) and possible values of the external variables (elements of W), namely the space

$$f(\pi^{-1}(x)) \subset T_x X \times W.$$

We denote the system (2.1) by $\Sigma(X, W, B, f)$. It is easily seen that in local coordinates x for X , v for the fibers of B , w for W , and with f factored in $f = (g, h)$, the system is given by

$$(2.2) \quad \dot{x} = g(x, v), \quad w = h(x, v).$$

Of course one should ask oneself how this kind of system formulation is connected with the usual input–output setting. In fact, by adding more and more assumptions successively to the very general formulation (2.1) we shall distinguish among three important situations, of which the last is equivalent to the “usual” interpretation of system (1.1).

(i) Suppose the map h restricted to the fibers of B is an *immersive* map into W (this is equivalent to asking that the matrix $\partial h / \partial v$ be injective). Then:

LEMMA 2.1. *Let h restricted to the fibers of B be an immersion into W . Let (\bar{x}, \bar{v}) and \bar{w} be points in B and W respectively such that $h(\bar{x}, \bar{v}) = \bar{w}$. Then locally around (\bar{x}, \bar{v}) and \bar{w} there are coordinates (x, v) for B (such that v are coordinates for the fibers of B), coordinates (w_1, w_2) for W and a map \tilde{h} such that h has the form*

$$(2.3) \quad (x, v) \mapsto (w_1, w_2) = (\tilde{h}(x, v), v).$$

Proof. The lemma follows from the implicit function theorem.

Hence *locally* we can interpret a part of the external variables, i.e., w_1 , as the outputs, and a complementary part, i.e., w_2 , as the inputs! When we denote w_1 by y

and w_2 by u , then system (2.2) has the form (of course only locally)

$$(2.4) \quad \dot{x} = y(x, u), \quad y = \tilde{h}(x, u).$$

(ii) Now we not only assume that $\partial h / \partial v$ is injective, which results in a *local* input-output parametrization (2.4), but we also assume that the output set denoted by Y is *globally* defined. Moreover, we assume that W is a fiber bundle above Y , which we will call $p : W \rightarrow Y$, and that h is a bundle morphism (i.e., maps fibers of B into fibers of W). Then:

LEMMA 2.2. *Let $h : B \rightarrow W$ be a bundle morphism, which is a diffeomorphism restricted to the fibers. Let $\bar{x} \in X$ and $\bar{y} \in Y$ be such that $h(\pi^{-1}(\bar{x})) = p^{-1}(\bar{y})$. Take coordinates x around \bar{x} for X and coordinates y around \bar{y} for Y . Let (\bar{x}, \bar{v}) be a point in the fiber above \bar{x} and let (\bar{y}, \bar{u}) be a point in the fiber above \bar{y} such that $h(\bar{x}, \bar{v}) = (\bar{y}, \bar{u})$. Then there are local coordinates v around \bar{v} for the fibers of B , coordinates u around \bar{u} for the fibers of W and a map $\tilde{h} : X \rightarrow Y$ such that h has the form*

$$(2.5) \quad (x, v) \xrightarrow{h} (y, u) = (\tilde{h}(x), v).$$

Proof. Choose a locally trivializing chart $(0, \varphi)$ of W around \bar{y} . Then $\varphi : p^{-1}(0) \rightarrow 0 \times U$, with U the standard fiber of W . Take local coordinates u around $\bar{u} \in U$. Then (y, u) forms a coordinate system for W around (\bar{y}, \bar{u}) . Because h is a bundle morphism, h has the form

$$(x, \tilde{v}) \xrightarrow{h} (y, u) = (\tilde{h}(x), h'(x, \tilde{v})).$$

where (x, \tilde{v}) is a coordinate system for B around (\bar{x}, \bar{v}) . Now adapt this last coordinate system by defining

$$v = (h')^{-1}(x, u) \quad \text{with } x \text{ fixed.}$$

Because h restricted to the fibers is a diffeomorphism, v is well defined and (x, v) forms a coordinate system for B in which h has the form

$$(x, v) \xrightarrow{h} (y, u) = (\tilde{h}(x), u). \quad \square$$

Hence under the conditions of Lemma 2.2 our system is locally (around $\bar{x} \in X$ and $\bar{y} \in Y$) described by

$$(2.6) \quad \dot{x} = g(x, u), \quad y = \tilde{h}(x).$$

This input-output formulation is essentially the same as the one proposed by Brockett [4] and Takens [15], who take the input spaces as the fibers of a bundle above a globally defined output space Y . In fact, this situation should be regarded as the normal setting for nonlinear control systems.

(iii) Take the same assumptions as in (ii) and assume moreover that W is a *trivial* bundle, i.e., $W = Y \times U$, and that B is a trivial bundle, i.e., $B = X \times V$. Because h is a diffeomorphism on the fibers, we can identify U and V . In this case the output set Y and the input set U are *globally* defined, and the system is described by

$$(2.7) \quad \dot{x} = g(x, u), \quad y = \tilde{h}(x),$$

where for each fixed \bar{u} , $g(\cdot, \bar{u})$ is a globally defined vector field on X . This is the "usual" interpretation of (1.1).

Remarks on (i).

1. When h restricted to the fibers of B is *not* an immersion we have a situation where we could speak of “hidden inputs”. In fact, in this case there are variables in the fibers of B which can affect the internal state behavior via the equation $\dot{x} = g(x, v)$ but which cannot be directly identified with some of the external variables.

2. The splitting of the external variables into inputs and outputs as described in Lemma 2.1 is of course by no means unique! This fact has interesting implications, even in the linear case, which we shall not pursue further here.

Remarks on (ii).

1. From Lemma 2.2 it is clear that the coordinatization of the fibers of the bundle W uniquely determines, via h , the coordinatization of the fibers of B . It should be remarked that a coordinatization of the fibers of W is locally equivalent to the existence of an (integrable) *connection* on the bundle W , and that one coordinatization is linked with another by what is essentially an output feedback transformation, i.e., a bundle isomorphism from W into itself. Again we will not comment further on this point.

2. A beautiful example of this kind of system is the Lagrangian system (see Takens [15]). Here the output space is equal to the configuration space Q of a mechanical system. The state space X is the configuration space with the velocity space, so $X = TQ$. The space W is equal to T^*Q (the cotangent bundle of Q), with the fibers of T^*Q representing the external forces. When we denote the natural projection of TQ on Q by ρ , then B is just ρ^*T^*Q (the pullback bundle via ρ). Now given a function $L : TQ \rightarrow \mathbb{R}$ (called the Lagrangian) we can construct a symplectic form $d(\partial L / \partial \dot{q}) \wedge dq$ (with (q, \dot{q}) coordinates for TQ) on TQ which uniquely determines a map $g : B \rightarrow TTQ$ (cf. [15]). Finally, in coordinates the system is given by

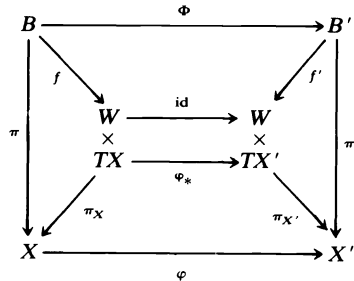
$$(2.8) \quad \ddot{q} = F(q, \dot{q}) + \sum_j u_j Z_j(q, \dot{q}), \quad y = q,$$

with the vector fields $F(q, \dot{q})$ and $Z_j(q, \dot{q})$ satisfying certain conditions. Moreover the vector fields Z_j commute, i.e., $[Z_i, Z_j] = 0$ for all i, j , a fact which has a very interesting interpretation (cf. [5], [15]).

Remark on (iii). Most cases where B can be taken as trivial are generated by a space X such that TX is a trivial bundle. For instance, when X is a Lie group TX is automatically trivial.

3. Minimality and observability

3.1. Minimality. We want to give a definition of minimality for a general (smooth) nonlinear system



DEFINITION 3.1 (see [16]). Let $\Sigma(X, W, B, f)$ and $\Sigma'(X', W, B', f')$ be two smooth systems. Then we say $\Sigma' \leq \Sigma$ if there exist surjective submersions $\varphi : X \rightarrow X'$, $\Phi : B \rightarrow B'$

such that the diagram

$$(3.1) \quad \begin{array}{ccc} B & \xrightarrow{f} & TX \times W \\ & \searrow & \swarrow \\ & X & \end{array}$$

commutes.

Σ is called *equivalent* to Σ' (denoted $\Sigma \sim \Sigma'$) if φ and Φ are diffeomorphisms.

We call Σ *minimal* if $\Sigma' \leq \Sigma \Rightarrow \Sigma' \sim \Sigma$.

Remark 1. This definition formalizes the idea that we call Σ' *less complicated* than Σ ($\Sigma' \leq \Sigma$) if Σ' consists of a set of trajectories in the state space, smaller than the set of trajectories of Σ , but which generates the same *external behavior*. (The external behavior Σ_e of $\Sigma(X, W, B, f)$ consists of the possible functions $w: \mathbb{R} \rightarrow W$ generated by $\Sigma(X, W, B, f)$. Hence, when we define

$$\Sigma := \{(x, w): \mathbb{R} \rightarrow X \times W \mid x \text{ absolutely continuous and } (\dot{x}(t), w(t)) \in f(\pi^{-1}(x(t))) \text{ a.e.}\},$$

then Σ_e is just the projection of Σ on $W^{\mathbb{R}}$.)

Remark 2. Notice that we only formalize the *regular* case by asking that Φ and φ be surjective as well as submersive. In fact we could, for instance, allow that at isolated points φ or Φ are not submersive. However, we will at this time not go into this problem, and we will treat only the regular case as described in Definition 3.1.

Remark 3. Notice that $\Sigma_1 \leq \Sigma_2$ and $\Sigma_2 \leq \Sigma_1$ need not imply $\Sigma_1 \sim \Sigma_2$. This fact leads to very interesting problems which we will not pursue further at this time.

Of course, Definition 3.1 is an elegant but rather abstract definition of minimality. From a differential geometric point of view it is very natural to see what these conditions of commutativity mean *locally*. In fact, we will see in Theorem 3.7 that locally these conditions of commutativity do have a very direct interpretation. But first we have to state some preparatory lemmas and theorems.

Let us look at (3.1). Because Φ is a submersion it induces an involutive distribution D on B given by

$$D := \{Z \in TB \mid \Phi_* Z = 0\}$$

(the foliation generated by D is of the form $\Phi^{-1}(c)$ with c constant). In the same way φ induces an involutive distribution E on X . Now the information in the diagram (3.1) is contained in three subdiagrams (we assume $f = (g, h)$ and $f' = (g', h')$):

$$\begin{array}{lll} \text{I} & \begin{array}{ccc} B & \xrightarrow{\Phi} & B' \\ h \downarrow & & \downarrow h' \\ W & \xleftarrow{\text{id}} & W \end{array} & \\ \text{II} & \begin{array}{ccc} B & \xrightarrow{\Phi} & B' \\ \pi \downarrow & & \downarrow \pi' \\ X & \xrightarrow{\varphi} & X' \end{array} & \\ \text{III} & \begin{array}{ccc} B & \xrightarrow{\Phi} & B' \\ g \downarrow & & \downarrow g' \\ TX & \xrightarrow{\varphi_*} & TX' \end{array} & \end{array}$$

LEMMA 3.2. *Locally the diagrams I, II, III are equivalent, respectively, to*

$$(3.2) \quad \begin{array}{ll} \text{I':} & D \subset \ker dh, \\ \text{II':} & \pi_* D = E, \\ \text{III':} & g_* D \subset TE = T\pi_*(D). \end{array}$$

Proof. I' and II' are trivial. For III' observe that, when φ induces a distribution E on X , then φ_* induces the distribution TE on TX . \square

Now we want to relate conditions I', II', III' with the theory of nonlinear disturbance decoupling, and especially with the formulation of it given in [11]. Consider in local coordinates the system

$$\dot{x} = f(x) + \sum_{i=1}^m u_i g_i(x) \quad \text{on a manifold } X.$$

We can interpret this as an affine distribution on X (for each $x \in X$, we give an affine subspace of $T_x X$). We call this affine distribution Δ . Now define

$$\Delta_0 := \Delta - \Delta := \{Y - Z \mid Y, Z \in \Delta\}.$$

It is easily seen that Δ_0 is a distribution on X , given in local coordinates by $\text{span}\{g_1(x), \dots, g_m(x)\}$ (the directions in which we can steer). Define $A(\Delta_0) := \{D \mid D \text{ an involutive distribution such that } D + \Delta_0 \text{ is involutive}\}$. Then in [11] it is proved that

THEOREM 3.3. *Let $D \in A(\Delta_0)$. Then the condition*

$$(3.3) \quad [\Delta, D] \subseteq D + \Delta_0$$

(we call such a $D \in A(\Delta_0)$ $\Delta \pmod{\Delta_0}$ invariant) is equivalent to the two conditions

- a) *there exists a vector field $F \in \Delta$ such that $[F, D] \subset D$;*
- b) *there exist vector fields $B_i \in \Delta_0$ such that $\text{span}\{B_i\} = \Delta_0$ and $[B_i, D] \subset D$.*

With the aid of this theorem the disturbance decoupling problem is readily solved.

The key to connecting our situation with this theory is given by the concept of the *extended system*, which is of interest in itself.

DEFINITION 3.4 (*extended system*). Let

$$\begin{array}{ccc} B & \xrightarrow{f} & TX \times W \\ & \searrow \pi & \swarrow \pi_x \\ & X & \end{array}$$

Then we define the *extended system* of $\Sigma(X, W, B, f)$ as follows: We define Δ_0 as the vertical tangent space of B , i.e.,

$$\Delta_0 := \{Z \in TB \mid \pi_* Z = 0\}.$$

Note that Δ_0 is automatically involutive.

Now take a point $(\bar{x}, \bar{v}) \in B$. Then $g(\bar{x}, \bar{v})$ is an element of $T_{\bar{x}} X$. Now define

$$\Delta(\bar{x}, \bar{v}) := \{Z \in T_{(\bar{x}, \bar{v})} B \mid \pi_* Z = g(\bar{x}, \bar{v})\}.$$

So $\Delta(\bar{x}, \bar{v})$ consists of the possible lifts of $g(\bar{x}, \bar{v})$ in (\bar{x}, \bar{v}) . Then it is easy to see that Δ is an affine distribution on B , and that $\Delta - \Delta = \Delta_0$. We call the affine system (Δ, Δ_0) on B constructed in this way, together with the output function $h: B \rightarrow W$, the *extended system* $\Sigma^e(X, W, B, f)$.

We have the following:

LEMMA 3.5.

a) Let D be an involutive distribution on B such that $D \cap \Delta_0$ has constant dimension. Then $\pi_* D$ is a well-defined and involutive distribution on X if and only if $D + \Delta_0$ is an involutive distribution.

b) Let D be an involutive distribution on B and let $D \cap \Delta_0$ have constant dimension. Then the following two conditions are equivalent:

- i) $\pi_* D$ is a well-defined and involutive distribution on X , and $g_* D \subset T\pi_* D$.
- ii) $[\Delta, D] \subset D + \Delta_0$.

Proof. a) Let $D + \Delta_0$ be involutive. Because D and Δ_0 are involutive this is equivalent to $[D, \Delta_0] \subset D + \Delta_0$. Applying Theorem 3.3 to this case gives a basis $\{Z_1, \dots, Z_k\}$ of D such that $[Z_i, \Delta_0] \subset \Delta_0$. In coordinates (x, u) for B , this last expression is equivalent to $Z_i(x, u) = (Z_{ix}(x), Z_{iu}(x, u))$, where Z_{ix} and Z_{iu} are the components of Z_i in the x - and u -directions, respectively. Hence $\pi_* D = \text{span} \{Z_{1x}, \dots, Z_{kx}\}$ and is easily seen to be involutive. The converse statement is trivial.

b) Assume i); then there exist coordinates (x, u) for B such that $D = \{\partial/\partial x_1, \dots, \partial/\partial x_k\}$ (the integral manifolds of D are contained in the sections $u = \text{constant}$). Then $g_* D \subset T\pi_* D$ is equivalent to

$$\left(\frac{\partial g}{\partial x_i} \right)_{j^e \text{ comp}} = 0$$

with $i = 1, \dots, k$ and $j = k+1, \dots, n$ (n is the dimension of X). From these expressions $[\Delta, D] \subset D + \Delta_0$ readily follows. The converse statement is based on the same argument. \square

Now we are prepared to state the main theorem of this section. First we have to give another definition.

DEFINITION 3.6 (*local minimality*). Let $\Sigma(X, W, B, f)$ be a smooth system. Let $\bar{x} \in X$. Then $\Sigma(X, W, B, f)$ is called *locally minimal* (around \bar{x}) if when D and E are distributions (around \bar{x}) which satisfy conditions I', II', III' of Lemma 3.2, then D and E must be the zero distributions.

It is readily seen from Definition 3.1 that minimality of $\Sigma(X, W, B, f)$ locally implies local minimality (locally every involutive distribution can be factored out).

Combining Lemma 3.2, Definition 3.4 and Lemma 3.5 we can state:

THEOREM 3.7. $\Sigma(X, W, B, f = (g, h))$ is locally minimal if and only if the extended system $\Sigma^e(X, W, B, f = (g, h))$ satisfies the condition that there exist no nonzero involutive distribution D on B such that

$$(3.4) \quad \begin{aligned} (i) \quad & [\Delta, D] \subset D + \Delta_0, \\ (ii) \quad & D \subset \ker dh. \end{aligned}$$

Remark 1. It is very surprising that the condition of minimality locally comes down to a condition on the extended system, which is in some sense an infinitesimal version of the original system.

Remark 2. Actually there is a conceptual algorithm to check local minimality (cf. [11]). Define

$$\Delta^{-1}(\Delta_0 + D) := \{\text{vector fields } Z \text{ on } B \mid [\Delta, Z] \subseteq \Delta_0 + D\}.$$

Then we can define the sequence $\{D^\mu\}$, $\mu = 0, 1, 2, \dots$ as follows:

$$\begin{aligned} D^0 &= \ker dh, \\ D^\mu &= D^{\mu-1} \cap \Delta^{-1}(\Delta_0 + D^{\mu-1}), \quad \mu = 1, 2, \dots \end{aligned}$$

Then $\{D^\mu\}$, $\mu = 0, 1, 2, \dots$, is a decreasing sequence of involutive distributions, and for some $k \leq \dim(\ker dh)$ $D^k = D^\mu$ for all $\mu \geq k$. Then D^k is the *maximal* involutive distribution which satisfies

$$(i) \quad [\Delta, D^k] \subset D^k + \Delta_0,$$

$$(ii) \quad D^k \subset \ker dh.$$

From Theorem 3.7 it follows that $\Sigma(X, W, B, f)$ is locally minimal if and only if $D^k = 0$! Notice that the maximum numbers of steps needed in this algorithm is equal to the dimension of $\ker dh$, and hence at least smaller than $\dim B$.

3.2. Observability. It is natural to suppose that our definition of minimality has something to do with controllability and observability. However, because the definition of a nonlinear system (2.1) also includes autonomous systems, (i.e., no inputs), minimality cannot be expected to imply, in general, some kind of controllability. In fact an autonomous linear system

$$\dot{x} = Ax, \quad y = Cx$$

is easily seen to be minimal if and only if (A, C) is observable (cf. [17]). Moreover, it seems natural to define a notion of *observability* only in the case that the system (2.1) has at least a local input–output representation; i.e., we make the standing assumption that $(\partial h / \partial v)$ is injective (see Lemma 2.1). Therefore, *locally* we have as our system

$$(3.5) \quad (= (2.5)) \quad \dot{x} = g(x, u), \quad y = \tilde{h}(x, u)$$

for every possible input–output coordinatization (y, u) of W (see Remark (i) 2 in § 2). For such an input–output system local minimality implies the following notion of observability, which we will call *local distinguishability*.

PROPOSITION 3.8. *Choose a local input–output parametrization as in (3.5). Then local minimality implies that the only involutive distribution E on X which satisfies*

$$i) \quad [g(\cdot, u), E] \subset E \quad \text{for all } u \quad (E \text{ is invariant under } g(\cdot, u)),$$

ii) $E \subset \ker d_x \tilde{h}(\cdot, u)$ for all u ($d_x \tilde{h}$ means differentiation with respect to x) is the zero distribution.

Proof. Let E be a distribution on X which satisfies i) and ii). Then we can lift E in a trivial way to a distribution D on B by requiring that the integral manifolds of D be contained in the sections $u = \text{constant}$. Then one can see that D satisfies $[\Delta, D] \subset D + \Delta_0$ and $D \subset \ker dh$. Hence $D = 0$ and $E = 0$. \square

Remark. It is easily seen that, under the condition $(\partial h / \partial v)$ injective, local minimality is in fact equivalent to the condition in Proposition 3.8. This is because $(\partial h / \partial v)$ injective implies that there cannot be a distribution D on B such that $D \cap \Delta_0 \neq 0$ and $D \subset \ker dh$. So from Lemma 3.5 a) it follows that the only involutive distributions D with $D \subset \ker dh$ and $D + \Delta_0$ involutive are of the form E_{lift} , with E an involutive distribution on X .

Actually, for nonlinear systems which can be represented in the input–output form without a feedthrough term (2.6), we can state the following:

COROLLARY. *Suppose there exists an input–output coordinatization*

$$(3.6) \quad (= (2.6)) \quad \dot{x} = g(x, u), \quad y = \tilde{h}(x),$$

Then local minimality implies local weak observability (cf. [6], [12]).

Proof. As can be seen from Proposition 3.8, local minimality in this more restricted case implies that the only involutive distribution E on X which satisfies

- i) $[g(\cdot, u), E] \subset E$ for all u ,
- ii) $E \subset \ker d\tilde{h}$

is the zero distribution. It can be readily seen [cf. 8] that the biggest distribution which satisfies i) and ii) is given by the null space of the codistribution P generated by elements of the form

$$L_{g(\cdot, u^1)} L_{g(\cdot, u^2)} \cdots L_{g(\cdot, u^k)} d\tilde{h}, \quad \text{with } u^i \text{ arbitrary.}$$

Because this distribution has to be zero, the codistribution P equals T_x^*X , in every $x \in X$. This is, apart from singularities (which we don't want to consider), equivalent to local weak observability as defined in [6].

Moreover, let (3.6) be locally weakly observable. Then all feedback transformations $u \mapsto v = \alpha(x, u)$ which leave the form (3.6) invariant (i.e., y is only the function x) are exactly the output feedback transformations $u \mapsto v = \alpha(y, u)$. It can be easily seen in local coordinates that after such output feedback is applied the modified system is still locally weakly observable. \square

In Proposition 3.8 and its corollary we have shown that local minimality implies a notion of observability which generalizes the usual notion of local weak observability. Now we will define a much stronger notion. Let us denote the (defined only locally) vector field $\dot{x} = g(x, \bar{u})$ for fixed \bar{u} by $g^{\bar{u}}$ and the function $\tilde{h}(x, \bar{u})$ by $h^{\bar{u}}$ (with g and \tilde{h} as in (3.5)).

DEFINITION 3.9. Let $\Sigma(X, W, B, f) = (g, h)$ be a smooth nonlinear system. It is called *strongly observable* if for every possible input-output coordinatization (3.5) the autonomous system

$$(3.7) \quad \dot{x} = g^{\bar{u}}(x), \quad y = h^{\bar{u}}(x)$$

with \bar{u} constant is *locally weakly observable* (for a definition see [6] or [12]), for all \bar{u} .

Remark. Let $\Sigma(X, W, B, f = (g, h))$ be strongly observable. Take one input-output coordinatization (y, u) . The system has the form (in these coordinates)

$$\dot{x} = g(x, u), \quad y = \tilde{h}(x, u).$$

Because the system is strongly observable, every *constant* input-function (constant in *this* coordinatization) distinguishes between two nearby states. However, in every other input-output coordinatization every constant (i.e., in *this* coordinatization) input function also distinguishes. This implies that in the first coordinatization every C^∞ input function distinguishes. Because the C^∞ input functions are dense in a reasonable set of input functions, every input function in this coordinatization distinguishes.

PROPOSITION 3.10. Consider the Pfaffian system constructed as follows:

$$P = dh^{\bar{u}} + L_{g^{\bar{u}}} dh^{\bar{u}} + L_{g^{\bar{u}}}(L_{g^{\bar{u}}} dh^{\bar{u}}) + \cdots + L_{g^{\bar{u}}}^{(n_{\bar{u}}-1)} dh^{\bar{u}},$$

with n the dimension of X and $L_{g^{\bar{u}}}$ the Lie derivative with respect to $g^{\bar{u}}$. As is well known from [6], the condition that the Pfaffian system P as defined above satisfies the condition $P_x = T_x^*X$ for all $x \in X$ (the so called observability rank condition) implies that the system

$$\dot{x} = g^{\bar{u}}(x), \quad y = h^{\bar{u}}(x)$$

is locally weakly observable. Hence, when the observability rank condition is satisfied for all \bar{u} , the system is strongly observable.

We will call the Pfaffian system P the *observability codistribution*.

Remark 1. As is known from [6], local weak observability of the system

$$\dot{x} = g^{\bar{u}}(x), \quad y = h^{\bar{u}}(x)$$

implies that the observability rank condition (i.e., $\dim P_x = T_x^*X$) is satisfied almost everywhere (in fact, in the analytic case everywhere). Because we don't want to go into singularity problems, for us local weak observability and the observability rank condition are the same.

Remark 2. It is easily seen that when for one input-output coordinatization the observability rank condition for all u is satisfied, then for *every* input-output coordinatization the observability rank condition for all u is satisfied. This follows from the fact that the observability rank condition is an open condition.

3.3. Controllability. The aim of this section is to define a kind of controllability which is “dual” to the definition of local distinguishability (Proposition 3.8) and which we shall use in the following section. The notion of controllability we shall use is the so-called “strong accessibility”, introduced in [14].

DEFINITION 3.11. Let $\dot{x} = g(x, u)$ be a nonlinear system in local coordinates. Define $R(T, x_0)$ as the set of points reachable from x_0 in exactly time T ; in other words,

$$R(T, x_0) := \{x_1 \in X \mid \exists \text{ state trajectory } x(t) \text{ generated by } g \\ \text{such that } x(0) = x_0 \text{ and } x(T) = x_1\}.$$

We call the system *strongly accessible* if for all $x_0 \in X$, and for all $T > 0$ the set $R(T, x_0)$ has a nonempty interior.

For systems of the form (in local coordinates)

$$(3.8) \quad \dot{x} = f(x) + \sum_{i=1}^m u_i g_i(x)$$

(i.e., affine systems) we can define A as the smallest Lie algebra which contains $\{g_1, \dots, g_m\}$ and which is invariant under f (i.e., $[f, A] \subset A$). It is known (cf. [14]) that $A_x = T_x X$ for every $x \in X$ implies that the system (3.8) is strongly accessible. In fact, when the system is analytic, strong accessibility and the rank condition $A_x = T_x X$ for every $x \in X$, are equivalent. We will call A the *controllability distribution* and the rank condition the controllability rank condition. Now it is clear that for affine systems (3.8) this kind of controllability is an elegant “dual” of local weak observability.

We know that the extended system (see Definition 3.4) is an affine system. Hence for this system we can apply the rank condition described above. This makes sense because the strong accessibility of $\Sigma(X, W, B, f)$ is very much related to the strong accessibility of $\Sigma^e(X, W, B, f)$, as can be seen from the following two propositions.

PROPOSITION 3.12. *If $\Sigma^e(X, W, B, f = (g, h))$ is strongly accessible, then $\Sigma(X, W, B, f = (g, h))$ is strongly accessible.*

Proof. In local coordinates the dynamics of Σ^e and Σ are given by

$$\begin{array}{ll} \text{I} & \dot{x} = g(x, u) \quad (\Sigma), \\ & \dot{x} = g(x, v) \quad (\Sigma^e), \\ \text{II} & \dot{v} = u. \end{array}$$

Now it is trivial that when for Σ^e we can steer to a point x_1 then we also can for Σ (even with an input that is smoother). \square

The converse is harder:

PROPOSITION 3.13. *Let $\Sigma(X, W, B, f = (g, h))$ be strongly accessible. Assume moreover that the fibers of B are connected. Then also $\Sigma^e(X, W, B, f = (g, h))$ is strongly accessible.*

Proof (sketch, see also [4]). Take the same representation of Σ and Σ^e as in the proof of Proposition 3.12. Let $x_0 \in X$ and let x_1 be in the (nonempty) interior of $R_\Sigma(x_0, T)$ (the reachable set of system Σ). Then it is possible that x_1 is reachable from x_0 by an input function $v(t)$ which cannot be generated by the differential equation $\dot{v} = u$ (with u for instance L^2). However, we know that the set of the v generated in this way is dense in, for instance L^2 . (For this we certainly need that the fibers of B are connected.) Because we only have to prove that the interior of a set is nonempty, this makes no difference. Now it is obvious from the equations

$$\dot{x} = g(x, v), \quad \dot{v} = u$$

that if we can reach an open set in the x -part of the (extended) state, then this is surely possible in the whole (x, v) -state. \square

4. Hamiltonian and gradient systems

4.1. Hamiltonian systems. A linear input-output system

$$\dot{x} = Ax + Bu, \quad y = Cx + Du$$

is called *Hamiltonian* if

$$\begin{aligned} A^T J + JA &= 0, \\ B^T J &= C, \quad \text{where } J \text{ is the symplectic form } \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix}, \\ D &= D^T \quad (\text{see [16]}), \end{aligned}$$

and is called a *gradient system* if

$$\begin{aligned} TA &= A^T T, \\ TB &= C^T, \quad \text{where } T \text{ is a nonsingular symmetric matrix,} \\ D &= D^T \quad (\text{see [18]}). \end{aligned}$$

It can be easily checked that for both kind of systems observability implies controllability, and vice versa.

We want to see whether we can derive a similar result for nonlinear Hamiltonian and gradient systems as defined in [16]. We start with the Hamiltonian case.

1. Let

$$\begin{array}{ccc} B & \xrightarrow{f} & TM \times W \\ \pi \searrow & & \swarrow \pi_M \\ & M & \end{array}$$

be a smooth system.

Now take M a symplectic manifold with symplectic form ω (see [1]). Because M is a symplectic manifold we can also define in a canonical way a symplectic form, denoted by $\hat{\omega}$, on TM (see [16]). Darboux's theorem tells us that we can find coordinates

(q_i, p_i) for M and $(q_i, p_i, \dot{q}_i, \dot{p}_i)$ for TM such that

$$\omega = \sum_{i=1}^n dq_i \wedge dp_i \quad \text{and} \quad \dot{\omega} = \sum_{i=1}^n d\dot{q}_i \wedge dp_i + dq_i \wedge d\dot{p}_i.$$

Now also take W a symplectic manifold with symplectic form ω^e .

Finally, we can make $TM \times W$ into a symplectic manifold by defining the symplectic form

$$\Omega := \pi_1^* \dot{\omega} - \pi_2^* \omega^e \quad (\text{the minus sign is a matter of convention})$$

with π_1 and π_2 the natural projections of $TM \times W$ on TM and W respectively.

DEFINITION 4.1 (see [16, Def. 4.1]). $\Sigma(M, W, B, f)$ with M and W as above is called *full Hamiltonian* if $f(B)$ is a Lagrangian submanifold of $(TM \times W, \Omega)$

PROPOSITION 4.2 (see [16, Prop. 4.2]). *Let $\Sigma(M, W, B, f)$ be full Hamiltonian. Then there exist coordinates for TM as above, coordinates $\{y_1, \dots, y_m, u_1, \dots, u_m\}$ for W and a function $H(q_1, \dots, q_n, p_1, \dots, p_n, u_1, \dots, u_m)$ such that the system is locally described by*

$$(4.1) \quad \begin{aligned} \dot{q}_i &= \frac{\partial H}{\partial p_i}, & i &= 1, \dots, n, \\ \dot{p}_i &= -\frac{\partial H}{\partial q_i}, \\ y_j &= c_j \frac{\partial H}{\partial u_j}, & j &= 1, \dots, m, \quad \text{with } c_j = \pm 1, \\ \omega^e &= \sum_{j=1}^m c_j dy_j \wedge du_j. \end{aligned}$$

Remark. We see that in this case the freedom in the input–output parametrization is restricted to the so called *canonical* coordinates for ω^e , i.e., only coordinates (y, u) such that $\omega^e = \sum c_j dy_j \wedge du_j$.

From Proposition 4.2 the following proposition easily follows:

PROPOSITION 4.3.

a) *Let $\Sigma(M, W, B, f)$ be locally minimal. Then f must be an immersion.*

b) *Let $\Sigma(M, W, B, f = (g, h))$ be full Hamiltonian and assume f is an immersion. Then h restricted to the fibers of B must be an immersion.*

Proof. a) From the definition of $\Sigma(M, W, B, f)$ it follows that we only have to prove that f restricted to the fibers is an immersion. Now suppose that f restricted to the fibers is not an immersion. Then the distribution $\ker d_v f \subset \Delta_0$ with d_v the derivative in the direction of the fiber is not equal to zero and satisfies (trivially) the conditions of Lemma 3.2., i.e., a contradiction.

b) Take a local input–output coordinatization (y, u) as in Proposition 4.2. Then the whole system is parametrized by the “input variables” u_1, \dots, u_m . Therefore the image of h restricted to the fibers has to be of dimension m and the image of g restricted to the fibers has to be of dimension at most m . Because f restricted to the fibers is an immersion, it is clear that the dimension of the fibers of B must be m and so h restricted to the fibers is an immersion. \square

Now we can state the main theorem of this section.

THEOREM 4.4. *Let $\Sigma(M, W, B, f = (g, h))$ be a full Hamiltonian system. Suppose f is an immersion. Then, for every input–output coordinatization of $\Sigma(M, W, B, f)$ as in*

Proposition 4.2,

$$(4.2) \quad \begin{aligned} \dot{q}_i &= \frac{\partial H}{\partial p_i}, \\ \dot{p}_i &= -\frac{\partial H}{\partial q_i} \end{aligned} \quad i = 1, \dots, n, \quad y_j = c_j \frac{\partial H}{\partial u_j}, \quad j = 1, \dots, m,$$

the following is true (see Proposition 3.8):

(4.2) is strongly accessible \Leftrightarrow (4.2) is locally distinguishable.

COROLLARY. *If $\Sigma(M, W, B, f)$ is locally minimal, then it follows from Proposition 4.3 that f is an immersion. Moreover, it follows from Proposition 3.8 that (4.2) is locally distinguishable. Therefore, by Theorem 4.4, the system (4.2) is also strongly accessible.*

Proof. Let us denote by X_H the vector field

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \quad \dot{p}_i = -\frac{\partial H}{\partial q_i}.$$

As is proved in Proposition 3.8, local distinguishability of (4.2), or equivalently local minimality, comes down to the following. Let \bar{O} be the vector space of functions spanned by $\{u_1, \dots, u_m, \partial H/\partial u_1, \dots, \partial H/\partial u_m\}$ (for simplicity take $c_j = 1$). Now add to \bar{O} all the functions generated by taking Lie derivatives of functions in \bar{O} with respect to the vector fields (on B) X_H and $\partial/\partial u_1, \dots, \partial/\partial u_m$. We denote the vector space spanned by all these functions by O . We shall give O the following notation:

$$O = \left\{ u_1, \dots, u_m, \frac{\partial H}{\partial u_1}, \dots, \frac{\partial H}{\partial u_m}, + \text{invariance under } X_H \text{ and } \frac{\partial}{\partial u_i} \right\}.$$

Then local distinguishability of (4.2) is equivalent to

$$dO(x, u) = T_{(x, u)}^* B \quad \text{for every } (x, u) \in B.$$

We can rewrite $\partial H/\partial u_i$ as $L_{\partial/\partial u_i} H$. Also it is easy to prove that $d_x(L_{\partial/\partial u_i} H) = L_{\partial/\partial u_i}(d_x H)$, $i = 1, \dots, m$ (d_x denotes differentiation with respect to x). Therefore:

$$\begin{aligned} dO &= \{du_1, \dots, du_m, d(L_{\partial/\partial u_1} H), \dots, d(L_{\partial/\partial u_m} H) + \text{invariance under } X_H \text{ and } \partial/\partial u_i\} \\ &= \{du_1, \dots, du_m, d_x(L_{\partial/\partial u_1} H), \dots, d_x(L_{\partial/\partial u_m} H) + \text{invariance under } X_H \text{ and } \partial/\partial u_i\} \\ &= \{du_1, \dots, du_m, L_{\partial/\partial u_1} d_x H, \dots, L_{\partial/\partial u_m} d_x H + \text{invariance under } X_H \text{ and } \partial/\partial u_i\} \end{aligned}$$

Now we turn to strong accessibility. As proved in Propositions 3.12 and 3.13, strong accessibility of (4.2) is equivalent to strong accessibility of the extended system of (4.2). Therefore when we define the vector space of vector fields A by the vector fields adding all the Lie derivatives of the vector fields $\partial/\partial u_1, \dots, \partial/\partial u_m$ with respect to X_H and $\partial/\partial u_i$, $i = 1, \dots, m$, i.e.,

$$A = \left\{ \frac{\partial}{\partial u_1}, \dots, \frac{\partial}{\partial u_m}, + \text{invariance under } X_H \text{ and } \frac{\partial}{\partial u_i} \right\},$$

then we have to show that

$$A(x, u) = T_{(x, u)} B \quad \text{for every } (x, u) \in B.$$

(A is the controllability distribution of the extended system.) It immediately follows that

$$A = \left\{ \frac{\partial}{\partial u_1}, \dots, \frac{\partial}{\partial u_m}, L_{\partial/\partial u_1} X_H, \dots, L_{\partial/\partial u_m} X_H, + \text{invariance under } X_H \text{ and } \frac{\partial}{\partial u_i} \right\}$$

Now we will show that the map $\alpha : TM \rightarrow T^*M$, defined by $\alpha(Y) = \omega(Y, -)$, with $Y \in TM$, together with the map which sends $\partial/\partial u_i$ to du_i , $i = 1, \dots, m$, is an isomorphism between A and dO , and therefore $A(x, u) = T_{(x, u)}B$ if and only if $dO(x, u) = T_{(x, u)}^*B$. The following observations are sufficient:

- i) $\alpha(X_H) = d_x H$.
- ii) $\alpha(L_{\partial/\partial u_i} X_H) = L_{\partial/\partial u_i} \alpha(X_H) = L_{\partial/\partial u_i} (d_x H)$.
- iii) Because $L_{X_H} \omega = 0$ and also $L_{\partial/\partial u_i} \omega = 0$, and because Lie brackets of Hamiltonian vector fields (Hamiltonian with respect to the degenerate form ω on B) are again Hamiltonian, A is generated by Hamiltonian vector fields.
- iv) Take an arbitrary Hamiltonian vector field X_G in A . Then:

$$\alpha(L_{X_H} X_G) = L_{X_H} \alpha(X_G) \quad \text{because } L_{X_H} \omega = 0,$$

$$\alpha(L_{\partial/\partial u_i} X_G) = L_{\partial/\partial u_i} \alpha(X_G) \quad \text{because } L_{\partial/\partial u_i} \omega = 0.$$

This easily gives the induction argument that A is mapped onto dO . \square

Remark 1. It is also possible to derive a duality result for strong observability (see Definition 3.9). The notion of dual controllability appears to be stronger than that of strong accessibility. However we will leave this for the moment.

Remark 2. Of course duality between strong accessibility and local distinguishability is closely related to the existence of a Lie algebra morphism between a Lie algebra of Hamiltonian vector fields equipped with the Lie bracket and a Lie algebra of Hamilton functions provided with the Poisson bracket (cf. [1]). We will explore this relationship in a future paper [19].

Remark 3. Consider the expression $\{\bar{u} \partial H / \partial u, H^{\bar{u}}\}$ with $\{\cdot, \cdot\}$ the Poisson bracket on M and $H^{\bar{u}}$ a function on M defined by $H^{\bar{u}}(q, p) := H(q, p, \bar{u})$. This expression equals

$$\sum_{j=1}^m \bar{u}_j \{h_j^{\bar{u}}, H^{\bar{u}}\} = \sum_{j=1}^m \bar{u}_j \cdot \frac{d}{dt} h_j^{\bar{u}},$$

with $\bar{u} = (\bar{u}_1, \dots, \bar{u}_m)$ and $h_j^{\bar{u}}(q, p) := j\text{th component of } (\partial H / \partial u)(q, p, \bar{u})$, and has a direct interpretation in the sense that when we interpret u as the external force and y as the position (see [16]) the expressions equal the instantaneous external work.

4.2. Gradient systems. Following [16] a system $\Sigma(M, W, B, f)$ is called a *full gradient system* if

- (i) M is a Riemannian manifold with (possibly indefinite) metric $\langle \cdot, \cdot \rangle$;
- (ii) W is a symplectic manifold with symplectic form ω^e ;
- (iii) $\langle \cdot, \cdot \rangle$ induces a bundle isomorphism α between TM and T^*M by setting $\alpha(X) := \langle X, - \rangle$, for $X \in TM$.

Because T^*M has a canonical 2-form Ω , TM has the symplectic form $\alpha^* \Omega$. Then $TM \times W$ is also a symplectic manifold with symplectic form $\pi_1^* \alpha^* \Omega - \pi_2^* \omega^e$ (π_1 and π_2 are the natural projections of $TM \times W$ on TM and W respectively).

Now we ask that $f(B)$ be a *Lagrangian submanifold* of $(TM \times W, \pi_1^* \alpha^* \Omega - \pi_2^* \omega^e)$.

PROPOSITION 4.6. *Parallel to Proposition 4.2 we can prove that locally this definition reduces to the existence of coordinates $\{x_1, \dots, x_n\}$ for M and $\{y_1, \dots, y_m\}$*

$u_1, \dots, u_m\}$ for W and the existence of a potential function $V(x, u)$ such that

$$(4.3) \quad \begin{aligned} G(x)\dot{x} &= \frac{\partial V}{\partial x}(x, u), \\ y_j &= c_j \frac{\partial V}{\partial u_j}(x, u) \quad \text{with } c_j = \pm 1 \end{aligned}$$

and $\omega^e = \sum c_j dy_j \wedge du_j$ (so (y, u) are canonical coordinates). $G(x)$ represents the Riemannian metric $\langle \cdot, \cdot \rangle$.

Now one could suppose, guided by the similarity in definition of Hamiltonian and gradient systems, and also by the linear situation as sketched before, that Theorem 4.3 should have an analogue in the gradient case. However it is easy to construct an example of a *nonlinear* gradient system which is strongly observable but not strongly accessible.

COUNTEREXAMPLE. Take $V(x_1, x_2, x_3, u) = e^{x_3}x_1x_2 + x_3u + u^2$, and as Riemannian metric the Euclidean metric on \mathbb{R}^3 . This generates a gradient system

$$\begin{aligned} \dot{x}_1 &= e^{x_3}x_2 := g_1(x), & \dot{x}_2 &= e^{x_3}x_1 := g_2(x) & \dot{x}_3 &= e^{x_3}x_1x_2 + u := g_3(x) + u, \\ y &= x_3 + u := h(x) + u, \end{aligned}$$

which is locally weakly observable because

$$\begin{aligned} \text{i)} \quad dh &= dx_3, \\ \text{ii)} \quad L_g dh &= d(e^{x_3}x_1x_2) = e^{x_3}x_2 dx_1 + e^{x_3}x_1 dx_2 + e^{x_3}x_1x_2 dx_3. \end{aligned}$$

Because $dh = dx_3$ and e^{x_3} is merely a factor > 0 we only have to consider $x_1 dx_2 + x_2 dx_1$.

$$\begin{aligned} \text{iii)} \quad L_g(x_1 dx_2 + x_2 dx_1) &= d(x_1 e^{x_3}x_1 + x_2 e^{x_3}x_2) \\ &= e^{x_3}(x_1 dx_1 + x_2 dx_2) + (\text{factors in } dx_3). \end{aligned}$$

Now because $x_1 dx_2 + x_2 dx_1$ and $x_1 dx_1 + x_2 dx_2$ form a basis of $T^*\mathbb{R}^2$ in almost every $(x_1, x_2) \in \mathbb{R}^2$ the observability codistribution has full dimension, and so the system is locally weakly observable (even strongly observable, as can be readily seen). But the system is not strongly accessible, because

$$\left[g, \frac{\partial}{\partial x_3} \right] = e^{x_3}x_2 \frac{\partial}{\partial x_1} + e^{x_3}x_1 \frac{\partial}{\partial x_2} + e^{x_3}x_1x_2 \frac{\partial}{\partial x_3} = g.$$

Therefore the controllability distribution has dimension at most two.

5. Conclusion. We have shown that the definition of a smooth nonlinear system in § 2 can be readily interpreted as a generalization of more usual input–output formulations. Further we can define a natural notion of minimality for such systems which implies the usual definition of observability for nonlinear systems. It would be interesting to look for a natural *realization theory* in this context. The definition of minimality suggests a more local theory than the realization theory of nonlinear input–output systems as developed in [9]. This aspect (see also [17]) is presently under investigation. We also expect to find a natural interpretation of the definition of strong observability in such a realization theory. The results of § 4 indicate that, contrary to the linear case, nonlinear gradient systems may be, at least from a system theoretic point of view, more complex than nonlinear Hamiltonian systems.

Acknowledgments. Discussions with H. Nijmeijer and J. C. Willems were very stimulating. I also would like to thank D. Bruin for pointing out to me an error in an earlier version of the manuscript.

REFERENCES

- [1] V. I. ARNOLD, *Mathematical Methods of Classical Mechanics*, Springer, New York 1978 (translation of the 1974 Russian edition).
- [2] J. BASTO GONCALVES, *Equivalence of gradient systems*, Control Theory Centre Report 84. University of Warwick, England.
- [3] R. W. BROCKETT, *Nonlinear systems and differential geometry*, Proc. IEEE, 64 (1976), pp. 61–72.
- [4] ———, *Global descriptions of nonlinear control problems*, Vector bundles and nonlinear control theory, Notes for a CBMS conference, to appear.
- [5] M. I. FREEDMAN AND J. C. WILLEMS, *Smooth representation of systems with differentiated inputs*, IEEE Trans. Autom. Control, AC-23 (1978), pp. 16–20.
- [6] R. HERMANN AND A. J. KRENER, *Nonlinear controllability and observability*, IEEE Trans. Autom. Control, AC-22 (1977), pp. 728–740.
- [7] R. W. HIRSCHORN, *(A, B)-invariant distributions and disturbance decoupling of nonlinear systems*, this Journal, 19 (1981), pp. 1–19.
- [8] A. ISIDORI, A. J. KRENER, C. GORI GIORGI AND S. MONACO, *Nonlinear decoupling via feedback: a differential geometric approach*, IEEE Trans. Autom. Control, submitted.
- [9] B. JAKUBCZYK, *Existence and uniqueness of realizations of nonlinear systems*, this Journal, 18 (1980), pp. 455–471.
- [10] R. E. KALMAN, P. L. FALB AND M. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.
- [11] H. NIJMEIJER, *Controlled invariance for affine nonlinear control systems*, submitted.
- [12] H. J. SUSSMANN, *Existence and uniqueness of minimal realizations of nonlinear systems*, Math. Systems Theory, 10 (1977), pp. 263–284.
- [13] ———, *Single-input observability of continuous-time systems*, Math. Systems Theory, 12 (1979), pp. 371–393.
- [14] H. J. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.
- [15] F. TAKENS, *Variational and conservative systems*, Rapport ZW-7603, Mathematics Institute, Groningen, the Netherlands, 1976.
- [16] A. J. VAN DER SCHAFT, *Hamiltonian dynamics with external forces and observations*, submitted.
- [17] J. C. WILLEMS, *System theoretic models for the analysis of physical systems*, Ricerche di Automatica, Special Issue on Systems Theory and Physics, to appear.
- [18] ———, *Dissipative dynamical systems, part II: Linear systems with quadratic supply rates*, Arch. Rat. Mech. Anal. 45 (1972), pp. 352–393.
- [19] A. J. VAN DER SCHAFT, *Controllability and observability for affine nonlinear Hamiltonian systems*, in preparation.

ON HAMILTONIAN FLOWS AND SYMPLECTIC TRANSFORMATIONS*

FRANK H. CLARKE†

Abstract. Any linear symplectic transformation is realized by the flow on a convex energy surface.

The basic object with which we are concerned in this article is a “convex energy surface” S , by which we mean a set S which is the boundary of a compact convex set in \mathbb{R}^{2n} containing the origin in its interior. We shall make no further assumptions on S , and in particular none about smoothness, but for the purposes of this introduction let us suppose for the moment that a C^1 function $H: \mathbb{R}^{2n} \rightarrow \mathbb{R}$ exists such that $S = H^{-1}(1)$ and such that $\nabla H \neq 0$ on S . We use (as is customary) the symbol J for the $2n \times 2n$ matrix

$$\begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix},$$

where I is the $n \times n$ identity matrix. The Hamiltonian system of differential equations

$$(1) \quad J\dot{z}(t) = \nabla H(z(t))$$

defines a flow on S , in view of the well-known fact that solutions of (1) evolve on level sets of H . (If z is decomposed into $(x, p) \in \mathbb{R}^n \times \mathbb{R}^n$, (1) may be written in the more familiar form $-\dot{p} = H_x, \dot{x} = H_p$.)

It is a useful observation (see Rabinowitz [9]) that the oriented curves on S (“orbits”) which result from solutions of (1) (on $(-\infty, \infty)$) do not depend on the choice of the function H representing S . This permits, in the study of orbits, the use of any convenient H . One such is the (Minkowski) gauge g of S , defined as follows:

$$(2) \quad g(z) = \min \{ \lambda \geq 0 : z \in \lambda S \}.$$

Clearly $g^{-1}(1) = S$, and if S were assumed to be smooth, then g would be C^1 (except at the origin) with $\nabla g \neq 0$ on S . Under our hypotheses, g is not necessarily differentiable on S . However, the function g is convex (hence locally Lipschitz) and in consequence admits at every point z a subdifferential in the sense of convex analysis (see Rockafellar [10]) (or, equivalently in the presence of convexity, a “generalized gradient,” see Clarke [2]). This is a set, denoted $\partial g(z)$, which reduces to $\{\nabla g(z)\}$ if g is differentiable at z . If now we give ourselves any locally Lipschitz function $H: \mathbb{R}^{2n} \rightarrow \mathbb{R}$ such that $H^{-1}(1) = S$ and $0 \notin \partial H$ on S (g is one such), we may consider the natural extension of (1) to the nonsmooth case; that is, the “Hamiltonian inclusion”

$$(3) \quad J\dot{z}(t) \in \partial H(z(t)) \quad \text{a.e.,}$$

where a solution z of (3) is an absolutely continuous function and “almost everywhere” refers to t and Lebesgue measure. It follows just as in the smooth case (see [5]) that this procedure defines orbits on S , independently of the particular H . This nonsmooth setting is the one within which the tools we use were developed, so it seems to be the natural one to invoke, but the reader who is uncomfortable with strange derivatives may

* Received by the editors October 28, 1980, and in revised form May 5, 1981. This paper was written while the author was visiting the Electrical Engineering and Computer Science Department, University of California, Berkeley, and typed using National Science Foundation grant ECS-79-13148. This research was supported by a Canada Council Killam Research Fellowship and by the Natural Sciences and Engineering Research Council of Canada under grant A9082.

† Department of Mathematics, University of British Columbia, Vancouver, British Columbia, Canada V6T 1W4.

simply assume that S is a C^1 -manifold. Even in that case, however, the proof of the theorem requires considering the generalized gradient of a nonsmooth function.

A $2n \times 2n$ matrix M is said to be *symplectic* if

$$(4) \quad M^*JM = J$$

(where $*$ denotes transpose). Symplectic transformations are of great importance in classical mechanics, where “the motion of a mechanical system corresponds to the continuous evolution or unfolding of a canonical (i.e., symplectic) transformation” (H. Goldstein [7, § 8.6]). The following, our main result, can be viewed as a kind of converse: any (linear) symplectic transformation is realized on some orbit of any convex energy surface.

THEOREM. *Let S be a convex energy surface and let M be a symplectic matrix. Then, for some s in S , there is an orbit on S joining s to Ms .*

Remark. As a referee has pointed out, implicit in the theorem is the “antipodality” result that any convex energy surface S contains a point s whose image under M also lies in S . (This fact can also be deduced from known properties of symplectic matrices.) In certain cases of small perturbations of orbits, our results are related to work of J. Moser [8].

The proof, which hinges upon a direct variational principle, achieves the desired result by producing a solution z on S of (3) on an interval $[0, T]$ ($T > 0$), where H is the gauge g of S , having the property that $z(T) = Mz(0)$. We shall need the *polar* Σ of S . This is defined as the unique convex set Σ whose “support function” is g ; i.e., Σ is the convex set satisfying, for every s in R^{2n} ,

$$\max \{ \langle s, \sigma \rangle : \sigma \in \Sigma \} = g(s).$$

It follows (see [5]) that Σ is a convex compact set containing the origin in its interior. We denote by Y the set of all absolutely continuous functions $y : [0, 1] \rightarrow R^{2n}$ satisfying $J\dot{y}(t) \in \Sigma$ almost everywhere, and we let $A : Y \rightarrow R^{2n}$ be defined by

$$A(y) = y(1) - My(0).$$

Finally, let E_1 be the eigenspace of M for eigenvalue 1, and let π and p be the projections onto E_1 and E_1^\perp respectively.

LEMMA 1. *There exist positive constants c and k such that, for every y in Y ,*

$$|p(y(0))| \leq c + k|A(y)|.$$

Proof. There is a constant k such that, for every v in E_1^\perp , $k|(I - M)v| \geq |v|$, and a constant b such that $|y(1) - y(0)| \leq b$ for all y in Y (since Σ is bounded). Let any y in Y be given. Since $(I - M)$ acts as a projection on E_1^\perp , we have

$$(I - M)p(y(0)) = (I - M)y(0) = y(0) - y(1) + A(y),$$

and consequently

$$|p(y(0))| \leq k(b + |A(y)|).$$

The lemma follows by taking $c = bk$.

We denote by f the functional on Y defined by

$$f(y) = -\frac{1}{2} \int_0^1 \langle J\dot{y}, y \rangle dt.$$

We now define a parametrized family $P(v)$ of problems: to minimize $f(y)$ over the elements y of Y which satisfy $|y(0)| \leq c + k$ and $A(y) = v$. We denote by $\alpha(v)$ the minimum in problem $P(v)$.

LEMMA 2. *The minimum $\alpha(v)$ is finite and attained for all v near 0; the function $\alpha(\cdot)$ is Lipschitz in a neighborhood of 0.*

Proof. Since Σ contains a neighborhood of 0, it follows that the feasible set for $P(v)$ is nonempty for small v . The uniform bounds on $y(0)$ and \dot{y} imply that f is bounded on this feasible set. If y_i is a minimizing sequence for $P(v)$, we may assume (by extracting subsequences) that \dot{y}_i converges weakly in L^1 to a function u satisfying $Ju \in \Sigma$ almost everywhere, and that $y_i(0)$ converges to $y(0)$, where y is an element of Y such that $\dot{y} = u$ almost everywhere. It follows that y_i converges strongly to y in L^1 and that $f(y_i)$ converges to $f(y)$; y is the required solution (a more detailed version of essentially this argument appears in [6]). That α is Lipschitz near 0 is a direct application of [1, Prop. 4.1].

LEMMA 3. *There is a nonconstant solution \hat{y} of $P(0)$ such that $|\hat{y}(0)| \leq c$.*

That there is a solution \hat{y} follows from the preceding lemma. If \hat{y} is constant, then $\alpha(0) = 0$. But this cannot be, since it is easy to produce y with zero endpoints such that $f(y) < 0$, and for sufficiently small ε we will have εy feasible for $P(0)$ with $f(\varepsilon y) < 0$. Now let us observe that whenever $A(y) = 0$ and u is any element of E_1 , we have $A(y - u) = 0$ and

$$\begin{aligned} f(y - u) &= f(y) + \frac{1}{2} \int \langle J\dot{y}, u \rangle dt = f(y) + \frac{1}{2} \langle J(M - I)y(0), u \rangle \\ &= f(y) + \frac{1}{2} \langle y(0), (I - M^*)Ju \rangle = f(y), \end{aligned}$$

in view of the fact that, under (4), $Mu = u$ is equivalent to $M^*Ju = Ju$. Thus by "subtracting off" we may choose the solution \hat{y} to satisfy $\pi(\hat{y}(0)) = 0$ and the required result now follows from Lemma 1.

It follows from the definition of α that for any y in Y satisfying $|y(0)| \leq c + k$, we have

$$(5) \quad f(y) \geq \alpha(A(y)),$$

and we have equality for any y solving $P(v)$ for any v , in particular for \hat{y} . We may interpret (5) as follows: the function \hat{y} provides a strong local solution (in the sense of the calculus of variations, as in [3]) for the problem of minimizing

$$-\frac{1}{2} \int_0^1 \langle J\dot{y}, y \rangle dt - \alpha(y(1) - My(0)),$$

subject to $J\dot{y} \in \Sigma$ (the constraint $|y(0)| \leq c + k$ is inactive near \hat{y} as a consequence of Lemma 3).

We now apply the necessary conditions of [3] to this "differential inclusion problem." These assert the existence of an absolutely continuous function q satisfying

$$(6) \quad (-\dot{q}, \dot{\hat{y}}) \in \partial H(\hat{y}, q) \quad \text{a.e.},$$

$$(7) \quad (q(0), -q(1)) \in \partial l(\hat{y}(0), \hat{y}(1)),$$

where l is the function

$$l(y, y') = -\alpha(y' - My)$$

and where H is the Hamiltonian of the problem:

$$\begin{aligned} H(y, q) &= \max \{ \langle q, w \rangle + \tfrac{1}{2} \langle Jw, y \rangle : Jw \in \Sigma \} \\ &= \max \left\{ \left\langle Jq + \left(\frac{y}{2} \right), Jw \right\rangle : Jw \in \Sigma \right\} = g \left(Jq + \left(\frac{y}{2} \right) \right). \end{aligned}$$

Relation (6) thus reduces to

$$J\dot{y} = -2\dot{q} \in \partial g \left(Jq + \left(\frac{\hat{y}}{2} \right) \right) \quad \text{a.e.}$$

It ensues that the function Jq is a translate of $\hat{y}/2$, so that the function $\hat{z} = Jq + (\hat{y}/2)$ is nonconstant (Lemma 3). We know by [11, Thm. 3] that $g(\hat{z}(t)) = h$ for some constant h ; by the preceding observation, h is strictly positive. Note the relation

$$(8) \quad J\dot{\hat{z}} \in \partial g(\hat{z}) \quad \text{a.e.}$$

We now turn to (7), which can be rewritten $q(1) = r$, $q(0) = M^*r$, where r is an element of $\partial\alpha(0)$. Armed with this we calculate

$$M\dot{\hat{z}}(0) = MJq(0) + \frac{M\hat{y}(0)}{2} = MJM^*q(1) + \frac{\hat{y}(1)}{2} = Jq(1) + \frac{\hat{y}(1)}{2} = \dot{\hat{z}}(1).$$

We now effect a transformation which places \hat{z} on S (rather than hS). Let $z(t) = \hat{z}(th)/h$ for $0 \leq t \leq 1/h = T$. Then $g(z(t)) = 1$, since g is positively homogeneous of degree 1, and we have (in view of (8))

$$J\dot{z} \in \partial g(z) \quad \text{a.e.,}$$

since $\dot{z} = \dot{\hat{z}}$ and ∂g is positively homogeneous of degree 0. Of course, z continues to satisfy $z(T) = Mz(0)$, and so we have arrived at the z alluded to at the beginning of the proof. \square

We shall say an orbit is *closed* if it consists of a closed curve on S , i.e., if it corresponds to a periodic solution of (3): $z(jT) = z(0)$ for every integer j . This is actually equivalent to the orbit being a closed set (an observation for which we thank R. Conley), so that no ambiguity lies in the term "closed orbit."

COROLLARY 1. *Any convex energy surface S admits at least one closed orbit.*

Of course we derive this by taking $M = I$ in the theorem. This global existence result is due to P. Rabinowitz [9] and A. Weinstein [12]. A proof of this corollary, simpler in this special case but similar in spirit, is given in [4], where the use of the polar or dual variational principle was inaugurated. We refer to [6] for an extension of this device to the prescribed period (as opposed to prescribed energy) case.

The following consequence of the theorem is reminiscent of Noether's theorem in the calculus of variations, inasmuch as an invariance of the system implies an invariance of (in this case, some) extremals:

COROLLARY 2. *Let the symplectic matrix M satisfy $MS = S$. Then there is an orbit C such that $MC = C$. If M is of finite order, there is a closed orbit with this property.*

Proof. First we apply the theorem and, as mentioned earlier, get a function $z(t)$ on $[0, T]$ satisfying (3) for $H = g$ and also $z(T) = Mz(0)$. We now observe that $z(\cdot)$ can be extended to $[T, 2T]$ as a solution of (3) as follows: let $z(t) = Mz(t - T)$ for $T < t \leq 2T$, and note: z is continuous at T , and

$$\begin{aligned} J\dot{z}(t) &= JM\dot{z}(t - T) = -JMJ \partial g(z(t - T)) \\ &= -JMJM^* \partial g(Mz(t - T)) = \partial g(z(t)) \quad \text{as required.} \end{aligned}$$

(We have used the identity $M^*\partial g(Mz) = \partial g(z)$, which follows by differentiating $g(Mz) = g(z)$.) We extend z indefinitely forward this way, and also in reverse time, to get an orbit C such that $MC = C$. If M has finite order (i.e., $M^K = I$) then of course we cycle in finitely many steps, and C is closed.

We now proceed to show that the hypothesis that M is symplectic cannot be deleted from the theorem. We do this by studying the case in which S is the sphere, for which a convenient Hamiltonian is $H(x, p) = \Sigma \{(x^i)^2 + (p^i)^2\}/2$ (summation 1 to n). The Hamiltonian system (2) is easily solved, and we find

$$x^i(t) = x_0^i \cos t = x_1^i \sin t, \quad p^i(t) = x_1^i \cos t - x_0^i \sin t.$$

It follows that $Mz(0) = z(t)$ is equivalent to

$$(9) \quad Mu = \{(\cos t)I - (\sin t)J\}u,$$

where I is the $2n \times 2n$ identity matrix and u is the vector (x_0, x_1) . For (x, p) to be nonzero, we must have a nontrivial solution u of (9), for some $t > 0$. It is not difficult to find (nonsymplectic) matrices M for which this is not possible, demonstrating that the theorem is false for general M . As a consequence of the theorem and the example, we deduce:

COROLLARY 3. *If M is symplectic, then there exist numbers α, β such that $\alpha^2 + \beta^2 = 1$ and*

$$\det(M - \alpha I - \beta J) = 0.$$

REFERENCES

- [1] J. P. AUBIN AND F. H. CLARKE, *Shadow prices and duality for a class of optimal control problems*, this Journal, 17 (1979), pp. 567–586.
- [2] F. H. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.
- [3] ———, *Extremal arcs and extended Hamiltonian systems*, Trans. Amer. Math. Soc., 231 (1977), pp. 349–367.
- [4] ———, *Solution périodique des équations hamiltoniennes*, Comptes Rendus Acad. Sci. Paris, 287 (1978), pp. 951–952.
- [5] ———, *Periodic solutions to Hamiltonian inclusions*, J. Differential Equations, 40 (1981), pp. 1–6.
- [6] F. H. CLARKE AND I. EKELAND, *Hamiltonian trajectories having prescribed minimal period*, Comm. Pure Appl. Math., 33 (1980), pp. 103–116.
- [7] H. GOLDSTEIN, *Classical Mechanics*, Addison-Wesley, Reading, MA, 1950.
- [8] J. MOSER, *A fixed-point theorem in symplectic geometry*, Acta Math., 141 (1978), pp. 17–34.
- [9] P. H. RABINOWITZ, *Periodic solutions of Hamiltonian systems*, Comm. Pure Appl. Math., 31 (1978), pp. 157–184.
- [10] R. T. ROCKAFELLAR, *Generalized Hamiltonian equations for convex problems of Lagrange*, Pacific J. Math., 33 (1970), pp. 411–427.
- [11] ———, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [12] A. WEINSTEIN, *Periodic orbits for convex Hamiltonian systems*, Ann. Math., 108 (1978), pp. 507–518.

ITERATIVE PROCEDURES FOR CONSTRAINED AND UNILATERAL OPTIMIZATION PROBLEMS*

J. WARGA†

Abstract. We study an iterative procedure for determining an “extremal” of a (nonconvex) optimization problem with unilateral constraints patterned on the unilateral problems of optimal control. We also investigate auxiliary procedures for determining $\inf \{ \alpha | (\alpha, 0) \in A \}$, where A is a closed convex subset of a Hilbert space $\mathbb{R} \times H$, and for finding a point $s(q)$ in A nearest some given point q .

1. Introduction. Let T be a compact metric space, K a compact convex subset of a normed vector space, and

$$\phi = (\phi^0, \phi^1) = (\phi^0, \phi^{1,1}, \dots, \phi^{1,m}) : K \rightarrow \mathbb{R} \times C(T, \mathbb{R}^m)$$

a continuous function with continuous directional derivatives

$$(x, y) \rightarrow D\phi(x; y - x) = \lim_{\omega \rightarrow 0^+} \omega^{-1} [\phi(x + \omega(y - x)) - \phi(x)] : K \times K \rightarrow \mathbb{R} \times C(T, \mathbb{R}^m),$$

where \mathbb{R}^m is provided with the Euclidean norm and $C(T, \mathbb{R}^m)$ is the Banach space of continuous functions from T to \mathbb{R}^m with the sup norm $|\cdot|_{\text{sup}}$. We wish to consider Problem P_{III} of minimizing ϕ^0 on the set $\{x \in K | \phi^1(x) \leq 0\}$, where the inequality $\phi^1(x) \leq 0$ means $\phi^{1,j}(x)(t) \leq 0$ ($j = 1, \dots, m, t \in T$). This problem is patterned on the unilateral problems of optimal control (otherwise known as problems with state inequality constraints), where K represents the space of relaxed controls and ϕ is defined by certain functional-integral equations (such as those discussed in [9, Ch. V–VII]) and, in particular, by ordinary differential equations.

Necessary conditions for minimum for Problem P_{III} (see e.g. [9, Thm. V. 2.3, p. 303]) involve unknown Lagrange coefficient measures and, except in simple cases, are rather difficult to apply in computing prospective minima. We shall therefore consider a direct method in the form of the iterative Procedure III (of § 3) which formally resembles the finite-dimensional modified method of centers, due to Huard and Polak [7, Algorithm 4.2.27, p. 155] and bears certain features of the feasible directions algorithm of Polak and Mayne [8] (for optimal control problems with endpoint inequalities). However, when applied to optimal control (in § 4), Procedure III differs from [8] not only because it admits state inequality constraints for infinitely many values of time but also because it admits time-varying and nonconvex restrictions on the control variables, requires no differentiability with respect to the latter, and yields a “strong” extremal satisfying Pontryagin’s maximum principle. In its operational aspects, Procedure III differs from [8] in using the linear structure of relaxed controls and not of ordinary controls and in selecting its preferred direction by keeping account not only of the ε -active constraints but of all the constraints and their differing proximities to the “active condition.”

Procedure III shares with many iterative methods the property of requiring nested iterations. In the present case, we must apply the iterative Procedure II at each step of Procedure III and the iterative Procedure I at each step of Procedure II. Specifically, we attempt to solve our basic Problem P_{III} of finding $\min \{ \phi^0(x) | x \in K, \phi^1(x) \leq 0 \}$ by assuming known a “feasible” point $x_0 \in K$ satisfying $\phi^1(x_0) \leq 0$ and constructing

* Received by the editors November 4, 1980, and in revised form June 10, 1981. This work was supported in part by the National Science Foundation under grant MCS-7903394.

† Department of Mathematics, Northeastern University, Boston, Massachusetts 02115.

iteratively a sequence (x_i) in K as follows: Given x_i , we apply Procedure II to find the solution (α_i, y_i) of Problem $\hat{P}(x_i)$, defined by

$$\hat{P}(x): \min \{-\alpha | D\phi^0(x; y - x) \leq -\alpha, \phi^1(x) + D\phi^1(x; y - x) \leq -\alpha, \alpha \in \mathbb{R}, y \in K\},$$

and then set $x_{i+1} = x_i + \theta_i(y_i - x_i)$, where θ_i minimizes $\phi^0(x_i + \theta(y_i - x_i))$ on the set

$$\{\theta \in [0, 1] | \phi^1(x_i + \theta(y_i - x_i)) \leq 0\}.$$

In turn, Procedure II solves each Problem $\hat{P}(x_i)$ by an iteration involving a repeated use of Procedure I to find $\min \{|b - a| | a \in A\}$ for appropriately defined points b and convex sets A .

In an attempt to “finitize” these methods (and somewhat in the spirit of [8]), we also introduce related Procedures I_1 , II_1 and III_1 in which all the internal iterations are terminated after a finite number of steps and the j th step of every iteration may be affected by an error whose preassigned bound η_j must decrease to 0 as $j \rightarrow \infty$.

In the special case when each of the functions

$$x \rightarrow \phi^0(x), \quad x \rightarrow \phi^{1,j}(x)(t) \quad (j = 1, \dots, m, \quad t \in T)$$

is convex, Problem P_{III} is equivalent to the problem of minimizing α subject to $(\alpha, 0) \in A$, where

$$A = \{(\phi^0(x), \phi^1(x) + v^1(x)) | x \in K, 0 \leq v^1 \leq c, v^1 \in C(T, \mathbb{R}^m)\}$$

is a closed, bounded and convex subset of $\mathbb{R} \times C(T, \mathbb{R}^m)$ and $c \geq |\phi^1|_{\sup}$. In the general case, where the above convexity assumptions do not hold, every step of Procedure III involves a similar problem, namely $\hat{P}(x_i)$, but with $A \subset \mathbb{R} \times \mathbb{R} \times C(T, \mathbb{R}^m)$ defined in terms of ϕ and $D\phi$. We solve this problem by embedding $C(T, \mathbb{R}^m)$ in a Hilbert space $L^2(\mu, \mathbb{R}^m)$ and applying Procedure II to the corresponding problem P_{II} in a Hilbert space. Procedure II is a generalization of the “tangent plane” algorithm of Wolfe¹ and Morrison [4], [5]. It requires at every step the application of Procedure I for finding the point $s(b)$ in A nearest a given point b . Procedure I is a rather straightforward generalization of the finite-dimensional methods of Gilbert [3], Barr [1] and Pecsvaradi and Narendra [6] and of the method of [10] (applicable to compact convex sets in a Hilbert space).

It is not surprising that, while Procedures I, I_1 , II, and II_1 always yield the solution of the corresponding convex optimization problems P_I and P_{II} , Procedure III yields an “extremal” point for the nonconvex Problem P_{III} but not necessarily a minimizing point. In the special case of unilateral problems of optimal control, an “extremal” point turns out to be an extremal control, that is, a control satisfying the customary generalization of Pontryagin’s maximum principle and transversality conditions [9, Thm. VI. 2.3, p. 357].

Theorem 2.1 provides specific error estimates for every step of Procedures I and I_1 . Theorem 2.3 demonstrates that, for “normal” problems, Procedure II converges superlinearly. We have little hope of learning much about the rapidity of convergence of Procedure III in the general case. In fact, the steepest descent method for scalar C^1 functions over compact convex subsets of \mathbb{R}^n is a special case of Procedure III, and little can be said in general about its rapidity of convergence. It would be interesting to conduct numerical experimentation with Procedure III applied to unilateral optimal control problems but we have not had the means to do so.

¹ The “tangent plane” algorithm is credited by Morrison [5] to the referee of his paper who, according to Kowalik [4], was P. Wolfe.

We discuss Procedures I, I₁, II and II₁ in § 2, Procedures III and III₁ in § 3, and an application to a unilateral problem of optimal control in § 4. The proofs appear in §§ 5, 6 and 7.

2. The convex optimization problem. We shall denote by $x \cdot y$ the inner product in a (real) Hilbert space and by $|\cdot|$ the norm in a (real) normed vector space.

Procedure I. Let B be a closed, bounded and convex subset of a Hilbert space \mathcal{Y} and $q \in \mathcal{Y}$. Let $x_0 \in B$ and x_1, x_2, \dots be constructed as follows: given x_i , select $y_i \in B$ such that

$$(x_i - q) \cdot y_i \leq (x_i - q) \cdot y \quad (y \in B),$$

choose as B_i any closed convex subset of B containing x_i and y_i , and choose as x_{i+1} the point in B_i that minimizes the distance to q .

Procedure I is clearly a special case of

Procedure I₁. Let B be a closed, bounded and convex subset of a Hilbert space \mathcal{Y} and $q \in \mathcal{Y}$. Let $x_0 \in B$, $\eta_0 \geq \eta_1 \geq \eta_2 \dots$, $\lim_i \eta_i = 0$, and let x_1, x_2, \dots be constructed as follows: given x_i , select $y_i \in B$ such that

$$2(x_i - q) \cdot y_i \leq 2(x_i - q) \cdot y + \eta_i \quad (y \in B),$$

choose as B_i any convex subset of B containing x_i and y_i , and choose as x_{i+1} any point in B_i such that $|x_{i+1} - q| \leq |x_i - q|$ and

$$|x_{i+1} - q|^2 \leq |x_i - q|^2 + \eta_i \quad (x \in B_i).$$

THEOREM 2.1. Let x_0, x_1, \dots be constructed by Procedure I₁, and let $s_B(q)$ denote the (unique) point in B nearest q . Then $\lim_i x_i = s_B(q)$ and

$$0 \leq |x_i - q|^2 - |s_B(q) - q|^2 \leq -2(x_i - q) \cdot (y_i - x_i) + \eta_i \rightarrow 0.$$

If $\mathcal{Y} = \mathbb{R}^n$ and B_i is the segment joining x_i and y_i , then Procedure I coincides with a method of Gilbert [3] and is related to the Frank-Wolfe procedure [2]. If $\mathcal{Y} = \mathbb{R}^n$ and B_i is chosen as a polyhedron containing certain x_k or y_k for $k \leq i$, then Procedure I coincides with the methods of Barr [1] and of Pecsvaradi and Narendra [6].

Now let H be a (real) Hilbert space, $\mathcal{H} = \mathbb{R} \times H$ the corresponding Hilbert space with the inner product

$$(\alpha, x) \cdot (\beta, y) = \alpha\beta + x \cdot y,$$

and A a closed convex subset of \mathcal{H} . We write $s(q) = (s^0(q), s^1(q))$ for $s_A(q)$ and set

$$C = \{\beta | (\beta, 0) \in A\}, \quad \gamma = \inf C, \quad g = (\gamma, 0) \in \mathcal{H},$$

with the usual convention that $\inf \phi = \infty$. If $p \in \mathcal{H}$, $p \neq 0$, $\bar{a} \in A$ and

$$p \cdot \bar{a} \leq p \cdot a \quad (a \in A),$$

we say that p is an *inward normal* to A at \bar{a} . The problem of finding γ is referred to as Problem P_{II}.

Procedure II that we define below resembles the Newton-Raphson method and exhibits similar convergence properties. Geometrically, it consists in iteratively replacing a point $b_i = (\beta_i, 0) \in \mathbb{R} \times H$ by a point $b_{i+1} = (\beta_{i+1}, 0)$ obtained by intersecting the axis $\mathbb{R} \times \{0\}$ with the hyperplane through $s(b_i)$ normal to $s(b_i) - b_i$.

Procedure II. Let $\beta_0 \in \mathbb{R}$, $\beta_0 \leq \gamma$. For $i = 0, 1, 2, \dots$, let $b_i = (\beta_i, 0) \in \mathcal{H}$, and let

$$\beta_{i+1} = s^0(b_i) + (s^0(b_i) - \beta_i)^{-1} |s^1(b_i)|^2 \quad \text{if } s^0(b_i) - \beta_i > 0.$$

If $s^0(b_i) - \beta_i \leq 0$, we set $\beta_j = \beta_i$ ($j > i$) and terminate the iteration.

THEOREM 2.2. Let β_1, β_2, \dots be determined by Procedure II. Then

$$\beta_i < \infty, \quad \beta_i \leq \beta_{i+1} \leq \gamma \quad (i = 0, 1, \dots).$$

If the iteration terminates at the k th step with $s(b_k) = b_k$ then $\beta_k = \gamma$. If the iteration terminates at the k th step with $s(b_k) \neq b_k$ then $\gamma = \infty$, i.e., Problem P_{II} has no solution. If the iteration continues indefinitely then (β_i) is a (strictly) increasing sequence converging to γ in the extended real number system. Furthermore,

$$\lim_i b_i = \lim_i s(b_i) = g = (\gamma, 0) \quad \text{if } \gamma < \infty.$$

THEOREM 2.3. Assume that $\gamma < \infty$ and the iteration of Procedure II continues indefinitely. Let

$$\lambda_i = s(b_i) - b_i, \quad \rho_i = (\lambda_i^0)^{-1} \lambda_i, \quad e_i = \gamma - \beta_i.$$

Then

- (1) $(|\rho_i|)$ is a (strictly) increasing sequence,
- (2) either $\lim_i |\rho_i| = \infty$ or there exist $\bar{\rho} = (1, \bar{\rho}^1) \in \mathcal{H}$ and $J \subset (1, 2, \dots)$ such that $\lim_{i \in J} \rho_i = \bar{\rho}$ weakly and $\bar{\rho}$ is an inward normal to A at $g = (\gamma, 0)$, and
- (3) if there exists an inward normal $\bar{\rho} = (1, \bar{\rho}^1)$ to A at g then

$$\lim \frac{e_{i+1}}{e_i} = 0.$$

Remark. Theorem 2.3 shows that, for "normal" problems (to which statement 3) is applicable), Procedure II converges superlinearly. If H is finite-dimensional and $\gamma < \infty$ then the boundary point $g = (\gamma, 0)$ of A admits an inward normal $\bar{\rho} = (\bar{\rho}^0, \bar{\rho}^1)$ with $\bar{\rho}^0 \geq 0$. Then either $\bar{\rho}^0 > 0$, in which case we may assume that $\bar{\rho}^0 = 1$ and statement 3) of Theorem 2.3 is applicable, or $\bar{\rho}^0 = 0$ and then $\bar{\rho}^1$ is an inward normal at 0 to the projection

$$P_A = \{x \in H \mid (\alpha, x) \in A\}$$

of A on H . However, if H is infinite-dimensional then the boundary point $g = (\gamma, 0)$ may admit no inward normal to A . For example, if μ is the Lebesgue measure on $[0, 1]$, $H = L^2(\mu)$ and A is the collection of all points in $\mathbb{R} \times L^2(\mu)$ of the form $(\alpha, \alpha h - f)$, where $\alpha \in [-1, 1]$, $h(t) = \sqrt{t}$, $|f(0)| \leq 1$ and f has a Lipschitz constant ≤ 1 , then $C = \{0\}$ and $g = (0, 0)$ but A admits no inward normal at $(0, 0)$ (because the points τf ($\tau > 0$) form a dense subset of $L^2(\mu)$). In fact, the arguments used in the proof of Theorem 2.3 can also demonstrate the following characterization of boundary points without inward normals:

PROPOSITION. Let A be a closed convex subset of a Hilbert space \mathcal{Y} and $\bar{x} \in \partial A$. Then A has no supporting hyperplane at \bar{x} if and only if

$$\lim_{x \rightarrow \bar{x}, x \notin A} |s(x) - x|^{-1} (s(x) - x) \cdot u = 0 \quad \text{for all } u \in \mathcal{Y}.$$

Procedure II_1 below will truncate the (infinite) iteration of Procedure I_1 after a finite number of steps according to a specific test. Procedure I_1 will be applied to the case where $B = A$ and q is some point $b \in \mathbb{R} \times \{0\}$. Thus, starting with some point a_0 ,

Procedure I_1 will generate a sequence $(f_i(b, a_0))$ in A converging to $s(b)$, and we shall refer in Procedure II_1 to such sequences $(f_i(b, a_0))$ directly. We shall also use the notation

$$\Gamma_n(p) = \{a \in A \mid p \cdot a \leq p \cdot x + \eta_n(x \in A)\}.$$

In defining Procedure II we assumed that we had the means of finding $s(b)$ for any $b \in \mathbb{R} \times \{0\} \subset \mathcal{H}$, and the existence of $s(b)$ is guaranteed if A is closed and convex. However, in Procedure II_1 we will also require the computation of elements of $\Gamma_n(p)$ and the existence of such elements is guaranteed only if A is bounded. Thus Procedure II_1 is applicable with the latter assumption (which is always satisfied in the context of Procedures III and III_1).

Procedure II_1 . Let $0 < M < 1$, $\eta_0 > \eta_1 > \eta_2 \cdots$ and $\lim_j \eta_j = 0$. Let $x_0 \in A$, $\beta_0 < \gamma$ and $n(0) = 0$. We iteratively construct x_j , β_j , λ_j , $n(j)$ ($j = 0, 1, 2, \dots$), as follows. We set

$$b_j = (\beta_j, 0) \in \mathcal{H}, \quad \lambda_j = (\lambda_j^0, \lambda_j^1) = x_j - b_j.$$

If $\lambda_j^0 > 0$, we set $\rho_j = (\lambda_j^0)^{-1} \lambda_j$ and determine some $y_j \in \Gamma_{n(j)}(\rho_j)$.

(1) If $\lambda_j^0 > 0$ and $\rho_j \cdot y_j - 2\eta_{n(j)} \geq \beta_j + M|\lambda_j|$, we set

$$\beta_{j+1} = \rho_j \cdot y_j - 2\eta_{n(j)}, \quad x_{j+1} = x_j, \quad n(j+1) = n(j) + 1;$$

(2) otherwise, we determine $f_i(b_j, x_j)$ for $i = 1, 2, \dots$. If, for some $i_0 = i_0(j)$, case (1) applies with x_j , $n(j)$ replaced by $f_{i_0}(b_j, x_j)$, $n(j) + i_0(j)$, respectively, we terminate this internal iteration and set

$$\beta_{j+1} = \beta_j, \quad x_{j+1} = f_{i_0}(b_j, x_j), \quad n(j+1) = n(j) + i_0(j).$$

We shall say that Procedure II_1 stops at N if case 2) applies for $j = N$ and $i_0(N)$ does not exist.

THEOREM 2.4. Let A be convex, closed and bounded, and let Procedure II_1 be applied to Problem P_{II} . Then $C \neq \emptyset$ if and only if Procedure II_1 does not stop at any j and the sequence (β_j) is nondecreasing and bounded. In that case

$$\lim_j \beta_j = \gamma, \quad \lim_j x_j = g = (\gamma, 0).$$

If $C = \emptyset$ then either Procedure II_1 does not stop at any j and then $\lim_j \beta_j = \gamma = \infty$ or Procedure II_1 stops at some N and then $f_i^0(b_N, x_N) - \beta_N \leq 0$ for all sufficiently large i .

Remark 2.5. Theorem 2.4 shows that if $C \neq \emptyset$ then Procedure II_1 yields a sequence $(x_j) = ((x_j^0, x_j^1))$ converging to $g = (\gamma, 0)$. It follows then that, for any $\eta > 0$, there exists j_0 such that

$$|x_j^0 - \gamma| \leq \eta, \quad |x_j^1| \leq \eta \quad (j \geq j_0),$$

a property that we shall refer to directly in Procedure III_1 .

3. The nonconvex unilateral optimization problem. We next consider Problem P_{III} described in § 1. Consistently with our use of the notation $\phi^1(x) \leq 0$, we write $h + \alpha$ for $(h^1 + \alpha, \dots, h^l + \alpha)$ when $h = (h^1, \dots, h^l): T \rightarrow \mathbb{R}^l$.

Procedure III , which we shall apply in an attempt to solve Problem P_{III} , will require the solution of auxiliary problems $\hat{P}(x)$ of the following form. Set $\hat{\phi} = (0, \phi^1)$. For a given point $x \in K$, determine $(\hat{\alpha}, \hat{y}) \in \mathbb{R} \times K$ that minimizes $-\alpha$ on the set

$$S(x) = \{(\alpha, y) \mid \hat{\phi}(x) + D\phi(x; y - x) + \alpha \leq 0\}.$$

We shall verify that each Problem $\hat{P}(x)$ is a special case of Problem P_{II} .

Let μ be a positive Radon measure on T such that $\mu(G) > 0$ for every open $G \subset T$. If T is a subset of some \mathbb{R}^k , it is convenient to choose μ to be the k -dimensional Lebesgue measure on T . Otherwise, we may construct μ by selecting a dense denumerable subset $\{t_1, t_2, \dots\}$ of T and setting

$$\mu(E) = \sum_{t_i \in E} 2^{-i} \quad \text{for every Borel subset } E \text{ of } T.$$

Let $|\cdot|_2$ denote the usual norm of the Hilbert space $L^2(\mu, \mathbb{R}^m)$. It is easy to see that any compact subset P of $(C(T, \mathbb{R}^m), |\cdot|_{\sup})$ is also a compact subset of $(L^2(\mu, \mathbb{R}^m), |\cdot|_2)$ and that, in P , $|\cdot|_{\sup}$ -convergence and $|\cdot|_2$ -convergence are equivalent. Furthermore (by Lemma 7.1 of § 7), it follows from the continuity of $(x, y) \rightarrow D\phi(x; y - x): K \times K \rightarrow \mathbb{R} \times C(T, \mathbb{R}^m)$ that $y \rightarrow D\phi(x; y - x)$ is linear under convex combinations, i.e.,

$$D\phi(x; \theta y_1 + (1 - \theta)y_2 - x) = \theta D\phi(x; y_1 - x) + (1 - \theta)D\phi(x; y_2 - x)$$

for $0 \leq \theta \leq 1$ and $x, y_1, y_2 \in K$. Finally, since $D\phi$ is bounded with respect to $|\cdot|_{\sup}$, for each $x \in K$ there exists $c_x \in \mathbb{R}$ such that

$$\hat{\phi}(x) + D\phi(x; y - x) \geq -c_x \quad (y \in K).$$

We now set, for a given $x \in K$,

$$\begin{aligned} A = A(x) = \{(-\alpha, \hat{\phi}(x) + D\phi(x; y - x) + \alpha + v) | y \in K, 0 \leq \alpha \leq c_x, \\ v \in L^2(\mu, \mathbb{R}^m), 0 \leq v(t) \leq c_x \quad \mu\text{-a.e.}\} \end{aligned}$$

and observe that A is a closed, bounded and convex subset of the Hilbert space $\mathbb{R} \times \mathbb{R} \times L^2(\mu, \mathbb{R}^m)$. It is also clear that

$$\min \{-\alpha | (\alpha, y) \in S(x)\} = \min \{-\alpha | (-\alpha, 0) \in A(x)\}.$$

Thus each Problem $\hat{P}(x)$ is a special case of Problem P_{II} .

Procedure III. Let $x_0 \in K$ and $\phi^1(x_0) \leq 0$. Given $x_i \in K$ ($i = 0, 1, \dots$), with $\phi^1(x_i) \leq 0$, we determine x_{i+1} as follows: we solve Problem $\hat{P}(x_i)$ which must have a solution (α_i, y_i) because $(0, x_i) \in S(x_i)$ and $y \rightarrow D\phi(x_i; y - x_i)$ is continuous. If $\alpha_i = 0$, we set $x_j = x_i$ ($j > i$) and terminate the iteration. If $\alpha_i > 0$, we determine θ_i that minimizes $\phi^0(x_i + \theta(y_i - x_i))$ on the set

$$\{\theta \in [0, 1] | \phi^1(x_i + \theta(y_i - x_i)) \leq 0\},$$

and let $x_{i+1} = x_i + \theta_i(y_i - x_i)$.

Procedure III is a special case of

Procedure III₁. Let $x_0 \in K$, $\phi^1(x_0) \leq 0$, $\eta_0 \geq \eta_1 \geq \dots$ and $\lim_j \eta_j = 0$. Given $x_i \in K$ ($i = 0, 1, 2, \dots$), we determine x_{i+1} as follows. We determine (α_i, y_i) that is an approximate solution of Problem $\hat{P}(x_i)$ in the sense that

$$\hat{\phi}(x_i) + D\phi(x_i; y_i - x_i) + \alpha_i \leq \eta_i$$

and

$$|-\alpha_i - \inf \{-\alpha | (\alpha, y) \in S(x_i)\}| \leq \eta_i.$$

(We have observed in Remark 2.5 that such an (α_i, y_i) is produced by Procedure II₁ in a finite number of steps.) We set

$$\mathcal{T}_i = \{\theta \in [0, 1] | \phi^0(x_i + \theta(y_i - x_i)) \leq \phi^0(x_i), \phi^1(x_i + \theta(y_i - x_i)) \leq 0\},$$

determine some $\theta_i \in \mathcal{T}_i$ such that

$$\phi^0(x_i + \theta_i(y_i - x_i)) \leq \inf_{\theta \in \mathcal{T}_i} \phi^0(x_i + \theta(y_i - x_i)) + \eta_i,$$

and let x_{i+1} be any point in K such that

$$\phi^0(x_{i+1}) \leq \phi^0(x_i + \theta_i(y_i - x_i)), \quad \phi^1(x_{i+1}) \leq 0.$$

THEOREM 3.1. *Let x_1, x_2, \dots be constructed by Procedure III₁. Then $\phi^1(x_i) \leq 0$ and $(\phi^0(x_i))$ is a nonincreasing sequence. If x_∞ is the limit of any convergent subsequence of (x_i) in the compact set K then $\phi^1(x_\infty) \leq 0$ and there exist $\lambda^0 \geq 0$ and a positive Radon measure $\lambda^1 = (\lambda^{1,1}, \dots, \lambda^{1,m})$ on T with values in \mathbb{R}^m such that*

- (a) $\lambda = (\lambda^0, \lambda^1) \neq 0, \lambda^{1,j}(\{t \in T | \phi^{1,j}(t) < 0\}) = 0$ ($j = 1, \dots, m$),
- (b) $\lambda^0 D\phi^0(x_\infty; y - x_\infty) + \int D\phi^1(x_\infty; y - x_\infty)(t) \cdot \lambda^1(dt) \geq 0$ ($y \in K$),

where

$$\int h(t) \cdot \lambda^1(dt) = \sum_{j=1}^m \int h^j(t) \lambda^{1,j}(dt) \quad \text{for } h = (h^1, \dots, h^m).$$

We refer to a point $\bar{x} \in K$ as extremal for Problem P_{III} if $\phi^1(\bar{x}) \leq 0$ and there exist λ^0 and λ^1 satisfying relations (a) and (b) of Theorem 3.1 with \bar{x} replacing x_∞ .

4. Unilateral problems of optimal control. A large class of optimal control problems with unilateral and inequality restrictions are special cases of Problem P_{III}. Such problems may be defined by functional-integral equations of a rather general type (e.g., such as studied in [9, Ch. V–VII]) but, to avoid some technicalities and to simplify the exposition, we shall restrict ourselves to problems defined by ordinary differential equations.

Let $T = [t_0, t_1] \subset \mathbb{R}$, R be a compact metric space, R^* a Lebesgue measurable mapping of T into the class of closed subsets of R with the Hausdorff metric, and \mathcal{S}^* the corresponding set of relaxed controls [9, Ch. IV], that is, functions $t \rightarrow \sigma(t)$ whose values are Radon probability measures on R such that $\sigma(t)(R^*(t)) = 1$ almost everywhere in T and the function

$$t \rightarrow \int c(r) \sigma(t)(dr) : T \rightarrow \mathbb{R}$$

is Lebesgue measurable for each continuous $c : R \rightarrow \mathbb{R}$. We recall [9, Thm. IV. 3.11, p. 287] that \mathcal{S}^* is a compact and convex subset of a normed vector space (namely $L^1(T, C(R))^*$ with a “weak” norm whose topology restricted to \mathcal{S}^* is the weak star topology).

We assume given an open set $V \subset \mathbb{R}^n$, $v_0 \in V$, and functions

$$f : T \times V \times R \rightarrow \mathbb{R}^n, \quad h^0 : V \rightarrow \mathbb{R}, \quad h^1 : V \rightarrow \mathbb{R}^{m_1}, \quad h^2 : T \times V \rightarrow \mathbb{R}^{m_2}$$

such that each $f(\cdot, v, r)$ is Lebesgue measurable, $f(t, \cdot, \cdot)$, h^0 , h^1 and h^2 are continuous, and $f(t, \cdot, r)$, h^0 , h^1 and $h^2(t, \cdot)$ have derivatives f_v , h_v^0 , h_v^1 and h_v^2 (with respect to $v \in V$) such that $f_v(t, \cdot, \cdot)$, h_v^0 , h_v^1 and h_v^2 are continuous. For any Radon measure s on R , we write

$$f(t, v, s) = \int f(t, v, r) s(dr),$$

and assume that there exist a compact $D \subset V$ and an integrable $\psi : T \rightarrow \mathbb{R}$ such that

$$|f(t, v, r)| \leq \psi(t), \quad |f_v(t, v, r)| \leq \psi(t) \quad (t \in T, \quad v \in D, \quad r \in R)$$

and such that the differential equation

$$y(t) = v_0 + \int_{t_0}^t f(\tau, y(\tau), \sigma(\tau)) d\tau \quad (t \in T)$$

has a unique solution $y(\sigma)$, with $y(\sigma)(t) \in D$ for all $\sigma \in \mathcal{S}^*$ and $t \in T$.

We now consider the optimal control problem of finding $\bar{\sigma} \in \mathcal{S}^*$ that minimizes $h^0(y(\sigma)(t_1))$ subject to the restrictions

$$h^1(y(\sigma)(t_1)) \leq 0, \quad h^2(t, y(\sigma)(t)) \leq 0 \quad (t \in T).$$

This is equivalent to solving Problem P_{III} in which

$$K = \mathcal{S}^*, \quad m = m_1 + m_2, \quad \phi^0(\sigma) = h^0(y(\sigma)(t_1)),$$

$$\phi^1(\sigma)(t) = (h^1(y(\sigma)(t_1)), \quad h^2(t, y(\sigma)(t))).$$

In fact, we can verify (see e.g. [9, Thm. VI. 1.1, p. 348]) that $\phi = (\phi^0, \phi^1)$, as defined above, as well as the function

$$(\sigma, \nu) \rightarrow D\phi(\sigma; \nu - \sigma) : \mathcal{S}^* \times \mathcal{S}^* \rightarrow \mathbb{R} \times C(T, \mathbb{R}^m),$$

are continuous, and that

$$D\phi^0(\sigma; \nu - \sigma) = h_v^0(y(\sigma)(t_1))\eta(t_1),$$

$$D\phi^1(\sigma; \nu - \sigma)(t) = (h_v^1(y(\sigma)(t_1))\eta(t_1), h_v^2(t, y(\sigma)(t))\eta(t)),$$

where $\eta : T \rightarrow \mathbb{R}^n$ is the solution of

$$\eta(t) = \int_{t_0}^t [f_v(\tau, y(\sigma)(\tau), \sigma(\tau))\eta(\tau) + f(\tau, y(\sigma)(\tau), \nu(\tau) - \sigma(\tau))] d\tau \quad (t \in T).$$

It follows that

$$\eta(t) = Z(\sigma)(t)^{-1} \int_{t_0}^t Z(\sigma)(\tau) f(\tau, y(\sigma)(\tau), \nu(\tau) - \sigma(\tau)) d\tau \quad (t \in T),$$

where $Z(\sigma) : T \rightarrow \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ is the matrix-valued solution of

$$Z(\sigma)(t) = I + \int_t^{t_1} Z(\sigma)(\tau) f_v(\tau, y(\sigma)(\tau), \sigma(\tau)) d\tau \quad (t \in T)$$

and I denotes the unit $n \times n$ matrix.

We shall next show how Procedures I, II and III (respectively, I₁, II₁ and III₁) can be applied to this unilateral optimal control problem. We assume known some $\sigma_0 \in \mathcal{S}^*$ such that $h^1(y(\sigma_0)(t_1)) \leq 0$ and $h^2(t, y(\sigma_0)(t)) \leq 0$ ($t \in T$). For any σ_i satisfying similar inequalities, we consider Problem $\hat{P}(\sigma_i)$ which has a solution (α_i, ν_i) . Each Problem $\hat{P}(\sigma_i)$ can be solved using Procedure II or II₁ if we can find

(a) a number $c = c_{\sigma_i}$ such that

$$\hat{\phi}(\sigma_i) + D\phi(\sigma_i; \nu - \sigma_i) \geq -c \quad (\nu \in \mathcal{S}^*),$$

(b) a number $\beta_0 \leq -\alpha_i$, and

(c) a method for determining for each $p \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{m_1} \times L^2(\mu, \mathbb{R}^{m_2})$, a point $a_p \in A(\sigma_i)$ such that $p \cdot a_p \leq p \cdot a$ ($a \in A(\sigma_i)$), where $A(\sigma_i)$ corresponds to Problem $\hat{P}(\sigma_i)$.

In fact, we can find c_{σ_i} of condition (a) by computing $y(\sigma_i)$, $Z(\sigma_i)$, $\hat{\phi}(\sigma_i)$ and setting

$$c_{\sigma_i} = -\inf \{h^{1,l}(y(\sigma_i)(t_1)), h^{2,l}(y(\sigma_i)(t)) | j = 1, \dots, m_1, l = 1, \dots, m_2, t \in T\} \\ -\inf \{d_0, d_1(j), e(l, t) | j = 1, \dots, m_1, l = 1, \dots, m_2, t \in T\},$$

where

$$d_0 = \int_{t_0}^{t_1} \min_{r \in R^*(\tau)} h_v^0(y(\sigma_i)(t_1)) Z(\sigma_i)(\tau) f(\tau, y(\sigma_i)(\tau), \delta_r - \sigma_i(\tau)) d\tau,$$

$$d_1(j) = \int_{t_0}^{t_1} \min_{r \in R^*(\tau)} h_v^{1,j}(y(\sigma_i)(t_1)) Z(\sigma_i)(\tau) f(\tau, y(\sigma_i)(\tau), \delta_r - \sigma_i(\tau)) d\tau,$$

$$e(l, t) = \int_{t_0}^t \min_{r \in R^*(\tau)} h_v^{2,l}(y(\sigma_i)(t)) Z(\sigma_i)(t)^{-1} Z(\sigma_i)(\tau) f(\tau, y(\sigma_i)(\tau), \delta_r - \sigma_i(\tau)) d\tau,$$

and δ_r denotes the Dirac measure at r . The number β_0 of condition (b) can be chosen as $-c$ and, for a given $p = (p_0^0, p_1^0, p_0^1, p_1^1) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{m_1} \times L^2(\mu, \mathbb{R}^{m_2})$, we can determine a_p by finding $\alpha_i \in [0, c_{\sigma_i}]$, $v_i = (v_i^1, v_i^2) \in \mathbb{R}^{m_1+1} \times L^2(\mu, \mathbb{R}^{m_2})$ with $v_i \in [0, c_{\sigma_i}]$, and $v_i \in \mathcal{S}^{\#}$ that minimize

$$(p_1^0, p_0^1, p_1^1) \cdot D\phi(\sigma_i; \nu - \sigma_i) + (-p_0^0 + 1 \cdot (p_1^0, p_0^1, p_1^1))\alpha + (p_0^0, p_1^0) \cdot v^1 + p_1^1 \cdot v^2,$$

where 1 denotes a function with all values and components equal to 1. This clearly yields values of α_i and v_i equal to 0 or c_{σ_i} depending on the sign of their respective coefficients. An easy computation also shows that a minimizing $v_i \in \mathcal{S}^{\#}$ is obtained by setting $v_i(\tau) = \delta_{\rho(\tau)}$ ($\tau \in T$), where, for each $\tau \in T$, $r = \rho(\tau)$ minimizes

$$k(\tau)^T f(\tau, y(\sigma_i)(\tau), \delta_r - \sigma_i(\tau)),$$

and

$$k(\tau)^T = \left[p_1^0 h_v^0(y(\sigma_i)(t_1)) + p_0^1 h_v^1(y(\sigma_i)(t_1)) \right. \\ \left. + \int_{\tau}^{t_1} p_1^1(t)^T h_v^2(y(\sigma_i)(t)) Z(\sigma_i)(t)^{-1} dt \right] Z(\sigma_i)(\tau).$$

The corresponding choice of a_p is

$$a_p = (-\alpha_i, \hat{\phi}(\sigma_i) + D\phi(\sigma_i; \nu_i - \sigma_i) + \alpha_i + v_i).$$

At this point we ought to make a remark concerning certain computational aspects of our procedures as applied to optimal control. These procedures involve relaxed controls which are measure-valued functions. However, it is well known (see, e.g., [9, Ch. VI]) that for each relaxed control σ there exists at least one Gamkrelidze-type control $\tilde{\sigma}$, represented by a function

$$t \rightarrow (\alpha_0(t), \dots, \alpha_n(t), \rho_0(t), \dots, \rho_N(t)): T \rightarrow \mathbb{R}^{n+1} \times \mathbb{R}^{n+1},$$

such that $y(\tilde{\sigma}) = y(\sigma)$; and the procedures that we apply retain their convergence properties if we replace arbitrary relaxed controls by equivalent Gamkrelidze-type controls. Thus the storage of relaxed controls in a computer memory is equivalent to the storage of ordinary control functions.

Finally, we observe that a point $\tilde{\sigma}$ that is extremal for Problem P_{III} as defined in this section yields a point $(y(\tilde{\sigma}), \tilde{\sigma})$ that is extremal in the sense of [9, Def. V. 2.0, p. 298] (see also [9, Thm. VI. 2.3(1), p. 358]). Thus Procedure III, when applied to unilateral problem of the optimal control of ordinary differential equations, yields an extremal control, i.e., one satisfying Pontryagin's principle and transversality conditions.

5. Proof of Theorem 2.1.

LEMMA 5.1. Let B be a bounded convex subset of a real normed vector space and $f: B \rightarrow \mathbb{R}$ a continuous function bounded below with directional derivatives $Df(x; y - x)$ such that, for each $x \in B$, $y \rightarrow Df(x; y - x)$ is bounded below and the family of functions $\{x \rightarrow Df(x; y - x) | y \in B\}$ is equiuniformly continuous. For $i = 0, 1, 2, \dots$, let

$$\eta_i \geq 0, \quad \lim_i \eta_i = 0, \quad x_i, y_i \in B,$$

$$Df(x_i; y_i - x_i) \leq \inf_{y \in B} Df(x_i; y - x_i) + \eta_i,$$

$$f(x_{i+1}) \leq \inf_{\theta \in [0,1]} f(x_i + \theta(y_i - x_i)) + \eta_i, \quad f(x_{i+1}) \leq f(x_i).$$

Then $\lim_i Df(x_i; y_i - x_i) = 0$. Furthermore, if f is convex then

$$\inf f(B) \leq f(x_i) \leq \inf f(B) - Df(x_i; y_i - x_i) + \eta_i \rightarrow \inf f(B).$$

Proof. We have

$$Df(x_i; y_i - x_i) \leq Df(x_i; x_i - x_i) + \eta_i = \eta_i \rightarrow 0.$$

Now assume, for purposes of contradiction, that there exist $J \subset (0, 1, 2, \dots)$ and $\varepsilon > 0$ such that

$$Df(x_i; y_i - x_i) \leq -\varepsilon \quad (i \in J).$$

Because of the equiuniform continuity assumption, there exists $\bar{\theta}$ such that $0 < \bar{\theta} < 1$ and

$$(1 - \theta)Df(x_i + \theta(y_i - x_i); y_i - x_i) = Df(x_i + \theta(y_i - x_i); y_i - [x_i + \theta(y_i - x_i)]) \leq -\varepsilon/2 \quad (i \in J, \quad 0 \leq \theta \leq \bar{\theta}).$$

Now set $\psi_i(\theta) = f(x_i + \theta(y_i - x_i))$ ($\theta \in [0, 1]$), and let ψ_i^+ denote the right derivative of ψ_i . Then

$$\psi_i^+(\theta) = Df(x_i + \theta(y_i - x_i); y_i - x_i) \leq -\varepsilon/2 \quad (i \in J, \quad 0 \leq \theta < \bar{\theta}).$$

It follows that $\psi_i(\bar{\theta}) \leq \psi_i(0) - \bar{\theta}\varepsilon/2$ ($i \in J$), and therefore

$$f(x_{i+1}) \leq f(x_i) - \bar{\theta}\varepsilon/2 + \eta_i \quad (i \in J).$$

Since $(f(x_i))$ is nonincreasing, this implies $\lim_i f(x_i) = -\infty$ and contradicts the assumption that f is bounded below. Thus $\lim_i Df(x_i; y_i - x_i) = 0$.

Now assume that f is convex. Then

$$Df(x; y - x) \leq f(y) - f(x) \quad (x, y \in B);$$

hence

$$Df(x_i; y_i - x_i) \leq \inf_{y \in K} Df(x_i; y - x_i) + \eta_i \leq \inf f(B) - f(x_i) + \eta_i,$$

thus proving the last assertion of the lemma. Q.E.D.

Proof of Theorem 2.1. Let $f(x) - |x - q|^2$ ($x \in B$). Then f is convex, $Df(x; y - x) = 2(x - q) \cdot (y - x)$, and all the assumptions of Lemma 5.1 are satisfied. It follows that

$$|s_B(q) - q|^2 \leq |x_i - q|^2 \leq |s_B(q) - q|^2 - 2(x_i - q) \cdot (y_i - x_i) + \eta_i \rightarrow |s_B(q) - q|^2$$

and therefore $\lim_i x_i = s_B(q)$ (because B is a convex subset of a Hilbert space). Q.E.D.

6. Proof of Theorems 2.2, 2.3 and 2.4.

LEMMA 6.1. *If β_1, β_2, \dots are determined by Procedure II and if $\beta_i < \gamma < \infty$ then*

$$\beta_i < s^0(b_i) \leq \beta_{i+1} \leq \gamma \quad \text{and} \quad \beta_{i+1} - \beta_i = [s^0(b_i) - \beta_i]^{-1} |s(b_i) - b_i|^2.$$

Proof. If $\beta_i < \gamma$ then $b_i \notin A$ and therefore $\nu = s(b_i) - b_i \neq 0$. Since ν is an inward normal to A at $s(b_i)$, we have

$$\nu \cdot s(b_i) \leq \nu \cdot g = \nu^0 \gamma;$$

hence

$$0 < |\nu|^2 = \nu \cdot [s(b_i) - b_i] \leq \nu^0 (\gamma - \beta_i)$$

and

$$(1) \quad \nu^0 = s^0(b_i) - \beta_i > 0, \quad \nu^0 \beta_{i+1} = \nu \cdot s(b_i) \leq \nu^0 \gamma.$$

Furthermore,

$$(2) \quad \beta_{i+1} - \beta_i = (\nu^0)^{-1} |\nu|^2 > 0,$$

and

$$(3) \quad \beta_{i+1} - s^0(b_i) = (\nu^0)^{-1} |\nu^1|^2 \geq 0.$$

Our conclusion follows from inequalities (1)–(3). Q.E.D.

LEMMA 6.2. *Let*

$$p_0 = (\pi_0, 0), \quad c \in A, \quad \lambda = s(p_0) - p_0, \quad \rho = (\lambda^0)^{-1} \lambda, \quad \nu = (1, \nu^1),$$

and assume that

$$\nu \cdot c \leq \nu \cdot x \quad (x \in A), \quad \lambda^0 > 0, \quad \rho \cdot s(p_0) < \nu \cdot c.$$

Then

$$|\rho| < |\nu| \quad \text{and} \quad |\rho|^{-1} [\rho \cdot s(p_0) - \pi_0] \geq |\nu|^{-1} [\nu \cdot c - \pi_0].$$

Proof. Let $\pi_1 = \rho \cdot s(p_0)$ and $\pi_2 = \nu \cdot c$. We have

$$\rho \cdot [s(p_0) - p_0] = (\lambda^0)^{-1} |s(p_0) - p_0|^2 > 0;$$

hence

$$\pi_1 = \rho \cdot s(p_0) > \rho \cdot p_0 = \pi_0.$$

Thus $\pi_0 < \pi_1 < \pi_2$.

We observe that the hyperplane $L = \{x | \nu \cdot x = \pi_2\}$ separates the point p_0 from A . Thus, denoting by $d[x, S]$ the distance from a point x to a set S , we have

$$|\lambda| = d[p_0, A] \geq d[p_0, L] = |\nu|^{-1} (\pi_2 - \pi_0).$$

Furthermore,

$$|\lambda|^2 = \lambda \cdot [s(p_0) - p_0] = \lambda^0 [\rho \cdot s(p_0) - \pi_0] = \lambda^0 (\pi_1 - \pi_0)$$

and therefore

$$|\lambda| = \frac{\lambda_0}{|\lambda|} (\pi_1 - \pi_0) = \frac{\pi_1 - \pi_0}{|\rho|} \geq \frac{\pi_2 - \pi_0}{|\nu|}.$$

Since $\pi_0 < \pi_1 < \pi_2$, it follows that $|\rho| < |\nu|$. Q.E.D.

Proof of Theorem 2.2. Step 1. We first assume that $C \neq \emptyset$. If $\beta_0 = \gamma$ then $s(b_0) = b_0$ and the iteration terminates. If $\beta_0 < \gamma$ but the iteration terminates for $i = k$ then, by Lemma 6.1,

$$\beta_i < s^0(b_i) \leq \beta_{i+1} \leq \gamma \quad (i < k)$$

and

$$\beta_k = \gamma; \quad \text{hence } s(b_k) - b_k = 0.$$

It remains therefore to consider the case when the iteration continues indefinitely. Then, by Lemma 6.1, (β_i) is an increasing sequence with an upper bound γ and therefore (β_i) has a limit $\bar{\beta} \leq \gamma$. Thus, again by Lemma 6.1,

$$\lim_i [s^0(b_i) - \beta_i]^{-1} |s(b_i) - b_i|^2 = \lim_i (\beta_{i+1} - \beta_i) = 0,$$

$$\lim_i (s^0(b_i) - \beta_i) = 0.$$

Therefore, setting $\bar{b} = (\bar{\beta}, 0)$, we obtain

$$|s(\bar{b}) - \bar{b}| = \lim_i |s(b_i) - b_i| = 0,$$

and thus $\bar{b} \in A$, hence $\bar{\beta} = \gamma$.

We have shown above that if $C \neq \emptyset$ then either the iteration terminates for some $i = k$ with $s(b_k) - b_k = 0$ or the iteration continues indefinitely. Thus, if the iteration stops at $i = k$ with $s(b_k) - b_k \neq 0$ then $C = \emptyset$. It remains, therefore, to consider the case when $C = \emptyset$ and the iteration continues indefinitely. Then

$$\beta_{i+1} - \beta_i = (s^0(b_i) - \beta_i)^{-1} |s(b_i) - b_i|^2 \geq |s(b_i) - b_i|.$$

It is now easy to verify that this relation implies $\lim_i \beta_i = \gamma = \infty$. Indeed, if $\beta_0 \leq \beta_i \leq \bar{\beta}$ for some $\bar{\beta} \in \mathbb{R}$ and all i , then the compact set $[\beta_0, \bar{\beta}] \times \{0\}$, which has no points in common with A , must be at a positive distance d from A . Thus $\beta_{i+1} - \beta_i \geq d$ for all i , hence $\lim_i \beta_i = \infty$, contradicting the assumption that $\beta_i \leq \bar{\beta}$. Q.E.D.

Proof of Theorem 2.3. If we set, for any $i \in \{0, 1, 2, \dots\}$,

$$\pi_0 = \beta_i, \quad c = s(b_{i+1}), \quad \nu = \rho_{i+1}$$

then we have

$$\nu \cdot c \leq \nu \cdot x \quad (x \in A)$$

and, by Lemma 6.1,

$$\lambda_i^0 > 0, \quad \rho_i \cdot s(b_i) = \beta_{i+1} < \beta_{i+2} = \nu \cdot c.$$

It follows then from Lemma 6.2 that

$$|\rho_i| < |\rho_{i+1}|,$$

thus proving statement (1) and showing that either $\lim_i |\rho_i| = \infty$ or the sequence $(|\rho_i|)$ is bounded. Assume that the latter is the case. Then there exist a sequence $J \subset (0, 1, 2, \dots)$ and $\bar{\rho} \in \mathcal{H}$ such that

$$\lim_{i \in J} \rho_i = \bar{\rho} \text{ weakly.}$$

Since $\rho_i = (1, \rho_i^1)$, it follows that $\bar{\rho} = (1, \bar{\rho}^1)$. Furthermore, we have

$$\rho_i \cdot s(b_i) \leq \rho_i \cdot x \quad (i = 1, 2, \dots, x \in A)$$

and

$$\lim_i s(b_i) = g = (\gamma, 0).$$

It follows that, for every $x \in A$,

$$\begin{aligned} \bar{\rho} \cdot g &= \lim_{i \in J} \rho_i \cdot g = \lim_{i \in J} (\rho_i \cdot s(b_i) + \rho_i \cdot [g - s(b_i)]) \\ &\leq \lim_{i \in J} \rho_i \cdot x + \lim_{i \in J} |\rho_i| |g - s(b_i)| = \bar{\rho} \cdot x. \end{aligned}$$

Thus $\bar{\rho} = (1, \bar{\rho}^1)$ is an inward normal to A at g , proving statement (2).

Now assume that $\bar{p} = (1, \bar{p}^1)$ is an inward normal to A at g . Then it follows from Lemma 6.2 (with $\pi_0 = \beta_i$, $c = g$, $\nu = \bar{p}$) that

$$|\rho_i| < |\bar{p}| \quad (i = 0, 1, 2, \dots).$$

Thus the sequence $(|\rho_i|)$ is bounded and, as we have shown, this sequence is increasing and there exist $\bar{\rho}$ and J as described above. Since $\lim_{i \in J} \rho_i = \bar{\rho}$ weakly, we have

$$|\bar{\rho}| \leq \liminf_{i \in J} |\rho_i| = \lim_i |\rho_i| \leq |\bar{p}|.$$

Since the above inequalities hold for any inward normal $\bar{p} = (1, \bar{p}^1)$ to A at g , they must hold for $\bar{p} = \bar{\rho}$, implying that

$$\lim_i |\bar{\rho}_i| = |\bar{\rho}|.$$

We next consider the error term e_i . If we replace, in Lemma 6.2, p_0 , ρ , ν and c by, respectively, b_i , ρ_i , $\bar{\rho}$ and g then Lemma 6.2 yields

$$|\rho_i|^{-1}(\beta_{i+1} - \beta_i) = |\rho_i|^{-1}(\rho_i \cdot s(b_i) - \beta_i) \geq |\bar{\rho}|^{-1}(\bar{\rho} \cdot g - \beta_i) = |\bar{\rho}|^{-1}(\gamma - \beta_i).$$

Thus $e_i - e_{i+1} \geq |\bar{\rho}|^{-1} |\rho_i| e_i$; hence $e_{i+1} \leq (1 - |\bar{\rho}_i|/|\bar{\rho}|)e_i$ ($i = 0, 1, 2, \dots$). Since $\lim_i (1 - |\bar{\rho}_i|/|\bar{\rho}|) = 0$, this shows that

$$\lim_i \frac{e_{i+1}}{e_i} = 0. \quad \text{Q.E.D.}$$

LEMMA 6.3. *If Procedure II₁ does not stop at $j = 0, 1, \dots, N$ then*

$$x_j \in A, \quad \beta_{j-1} \leq \beta_j < \gamma \quad (j = 1, 2, \dots, N).$$

Proof. It is clear from the construction that $\beta_{j-1} \leq \beta_j$ and $x_j \in A$. We shall next prove by induction that $\beta_j < \gamma$. Indeed, this is trivially true if $\gamma = \infty$. If $\gamma < \infty$, we assume that $\beta_{j-1} < \gamma$. Then, if case (2) applies for $j-1$, we have $\beta_j = \beta_{j-1} < \gamma$. If case (1) applies for $j-1$ then

$$\beta_j = \rho_{j-1} \cdot y_{j-1} - 2\eta_{n(j-1)} \leq \inf_{y \in A} \rho_{j-1} \cdot y - \eta_{n(j-1)} < \rho_{j-1} \cdot (\gamma, 0) = \gamma. \quad \text{Q.E.D.}$$

LEMMA 6.4. *Let $C \neq \emptyset$. Then Procedure II₁ does not stop at any N .*

Proof. Assume that case (2) applies at $j = N$. For $i = 1, 2, \dots$, let

$$\begin{aligned} \beta &= \beta_N, \quad b = b_N, \quad c = s(b), \quad x_{N,i} = f_i(b, x_N), \\ \lambda_{N,i} &= x_{N,i} - b, \quad \rho_{N,i} = (\lambda_{N,i}^0)^{-1} \lambda_{N,i} \quad \text{if } \lambda_{N,i}^0 > 0. \end{aligned}$$

Then, by Lemma 6.3, $\beta < \gamma$ and therefore $b \notin A$. Thus $\lim_i \lambda_{N,i} = c - b \neq 0$. Now $g = (\gamma, 0) \in A$ and $(c - b) \cdot c \leq (c - b) \cdot x$ ($x \in A$), and therefore

$$(c^0 - \beta)(\gamma - \beta) = (c - b) \cdot (g - b) \geq |c - b|^2 > 0;$$

hence

$$(1) \quad \lim_i \lambda_{N,i}^0 = c^0 - \beta > 0.$$

It follows that $\lambda_{N,i}^0 > 0$ for all large i , say $i \geq i_1$.

Now, let $y_{N,i} \in \Gamma_{n(j)+i}(\rho_{N,i})$ ($i \geq i_1$). Then

$$\inf_{y \in A} \rho_{N,i} \cdot y \leq \rho_{N,i} \cdot y_{N,i} \leq \inf_{y \in A} \rho_{N,i} \cdot y + \eta_{n(j)+i};$$

hence

$$\begin{aligned} \lim_i \rho_{N,i} \cdot y_{N,i} - \beta &= \lim_i \inf_{y \in A} \rho_{N,i} \cdot y - \beta = (c^0 - \beta)^{-1}(c - b) \cdot c - \beta \\ &= (c^0 - \beta)^{-1}|c - b|^2 \geq |c - b| > M|c - b| = M \lim_i |\lambda_{N,i}|. \end{aligned}$$

We now conclude that there exists an integer i_0 such that

$$\rho_{N,i_0} \cdot y_{N,i_0} - \beta \geq M|\lambda_{N,i_0}| + 2\eta_{n(N)+i_0}. \quad \text{Q.E.D.}$$

Proof of Theorem 2.4. First assume that $C \neq \emptyset$. Then, by Lemmas 6.3 and 6.4, Procedure II₁ does not stop at N and

$$x_j \in A, \quad \beta_j \leq \beta_{j+1} < \gamma < \infty \quad (j = 0, 1, 2, \dots).$$

Thus (β_j) converges to some $\bar{\beta} \leq \gamma$. Furthermore, case (1) applies for infinitely many values of j , say $j_1 < j_2 < \dots$. Then

$$M|x_{j_i} - b_{j_i}| \leq \beta_{j_i+1} - \beta_{j_i} \rightarrow 0$$

and therefore

$$(\bar{\beta}, 0) = \lim_j b_j = \lim_i b_{j_i} = \lim_i x_{j_i} \in A.$$

Since $\bar{\beta} \leq \gamma$, this shows that $\bar{\beta} = \gamma$. Furthermore, since the sequence (x_j) is made up of the elements of (x_{j_i}) , with some of the latter repeated finitely often, we have

$$\lim_j x_j = g = (\gamma, 0).$$

This proves the “only if” part of the first statement of the theorem.

Now assume that Procedure II₁ does not stop at any N . Then, by Lemma 6.3, the sequence (β_j) is nondecreasing and converges in the extended real number system to some $\bar{\beta} \leq \gamma$. Thus $\bar{\beta} = \infty$ implies $C = \emptyset$. If $\bar{\beta} < \infty$ and j_1, j_2, \dots are the values of j when case (1) applies then our previous argument shows that $(\bar{\beta}, 0) \in A$ and therefore $\bar{\beta} = \gamma < \infty$. This proves the “if” part of the first statement.

It remains to consider the case when Procedure II₁ stops at some N which, as we have seen, implies $C = \emptyset$. Then

$$\lambda_{N,i} = f_i(b_N, x_N) - b_N \xrightarrow{i} s(b_N) - b_N \neq 0.$$

We cannot have $\lambda_{N,i}^0 > 0$ for infinitely many i , say $i = i_1, i_2, \dots$, because then the same argument as in Lemma 6.4 shows that

$$\rho_{N,i_k} \cdot y_{N,i_k} - \beta > M|\lambda_{N,i_k}|$$

for sufficiently large values of k , implying that Procedure II₁ does not stop at N . Thus $\lambda_{N,i}^0 \leq 0$ for all sufficiently large i . Q.E.D.

7. Proofs related to Procedure III.

LEMMA 7.1. *Let ϕ be as described in § 1. Then*

$$D\phi(x; \theta y_1 + (1-\theta)y_2 - x) = \theta D\phi(x; y_1 - x) + (1-\theta)D\phi(x; y_2 - x) \\ (x, y_1, y_2 \in K, \quad 0 \leq \theta \leq 1).$$

Proof. Let $\psi: K \rightarrow \mathbb{R}$ and

$$(x, y) \rightarrow D\psi(x; y - x): K \times K \rightarrow \mathbb{R}$$

be continuous, and let $\bar{x}, y_1, y_2 \in K$. Set

$$h(\alpha, \beta) = \psi(\bar{x} + \alpha(y_1 - \bar{x}) + \beta(y_2 - \bar{x})) \quad (\alpha, \beta \geq 0, \quad \alpha + \beta \leq 1).$$

Then, for $\alpha + \beta < 1$,

$$h_\alpha^+(\alpha, \beta) = \lim_{\omega \rightarrow 0^+} \omega^{-1} [h(\alpha + \omega, \beta) - h(\alpha)] = D\psi(\bar{x} + \alpha(y_1 - \bar{x}) + \beta(y_2 - \bar{x}); y_1 - \bar{x})$$

and thus $(\alpha, \beta) \rightarrow h_\alpha^+(\alpha, \beta)$ is continuous. Thus $\alpha \rightarrow h(\alpha, \beta)$ is locally Lipschitz continuous and therefore, for each $\beta < 1$,

$$h_\alpha^+(\alpha, \beta) = h_\alpha(\alpha, \beta) \quad \text{for almost all } \alpha,$$

where h_α denotes the partial derivative. It follows that

$$h(\alpha + \varepsilon, \beta) = h(\alpha, \beta) + \int_\alpha^{\alpha + \varepsilon} h_\alpha^+(t, \beta) dt$$

for small ε ; hence $h_\alpha(t, \beta) = h_\alpha^+(t, \beta)$ for all t and thus h_α exists for all α, β with $\alpha, \beta > 0, \alpha + \beta < 1$, and is continuous. A similar conclusion holds for h_β , showing that h is C^1 on the open set $\alpha, \beta > 0, \alpha + \beta < 1$. Thus, for all such α, β ,

$$\begin{aligned} D\psi(\bar{x} + \alpha(y_1 - \bar{x}) + \beta(y_2 - \bar{x}); \theta(y_1 - \bar{x}) + (1-\theta)(y_2 - \bar{x})) \\ = \lim_{\omega \rightarrow 0^+} \omega^{-1} [h(\alpha + \theta\omega, \beta + (1-\theta)\omega) - h(\alpha, \beta)] \\ = \theta h_\alpha(\alpha, \beta) + (1-\theta)h_\beta(\alpha, \beta) \\ = \theta D\psi(\bar{x} + \alpha(y_1 - \bar{x}) + \beta(y_2 - \bar{x}); y_1 - \bar{x}) \\ + (1-\theta)D\psi(\bar{x} + \alpha(y_1 - \bar{x}) + \beta(y_2 - \bar{x}); y_2 - \bar{x}). \end{aligned}$$

It follows, by letting $\alpha, \beta \rightarrow 0$, that

$$D\psi(\bar{x}; \theta y_1 + (1-\theta)y_2 - \bar{x}) = \theta D\psi(\bar{x}; y_1 - \bar{x}) + (1-\theta)D\psi(\bar{x}; y_2 - \bar{x}).$$

Now let l be a continuous linear functional on $\mathbb{R} \times C(T, \mathbb{R}^m)$, and let $\psi = l \circ \phi$. Then our preceding argument shows that

$$l(D\phi(\bar{x}; \theta y_1 + (1-\theta)y_2 - x) - \theta D\phi(\bar{x}; y_1 - \bar{x}) - (1-\theta)D\phi(\bar{x}; y_2 - \bar{x})) = 0$$

for all such continuous linear functionals, whence our conclusions follows. Q.E.D.

Proof of Theorem 3.1. Step 1. Since $(0, x_i) \in S(x_i)$, we have $\alpha_i \geq -\eta_i$. We shall prove that $\lim_i \alpha_i = 0$. Indeed, assume the contrary, and let $J \subset (1, 2, \dots)$ and $\varepsilon_0 > 0$ be such that

$$\alpha_i \geq 2\varepsilon_0, \quad \eta_i \leq \varepsilon_0 \quad (i \in J);$$

hence

$$(1) \quad \hat{\phi}(x_i) + D\phi(x_i; y_i - x_i) \leq -\varepsilon_0 \quad (i \in J).$$

Since K is compact, K is bounded and the function $(x, y) \rightarrow D\phi(x; y - x)$ is uniformly continuous. Therefore there exists $\bar{\theta}$ such that $0 < \bar{\theta} \leq \frac{1}{2}$ and

$$(2) \quad \hat{\phi}(x_i + \theta(y_i - x_i)) + D\phi(x_i + \theta(y_i - x_i); y_i - [x_i + \theta(y_i - x_i)]) \leq -\varepsilon_0/2 \\ (\theta \in [0, \bar{\theta}], \quad i \in J).$$

Let $\psi_i(\theta) = (\psi_i^0(\theta), \psi_i^1(\theta)) = \phi(x_i + \theta(y_i - x_i))$, and let $\psi_i^+(\theta)$ denote the right derivative of ψ_i at θ . Then

$$D\phi(x_i + \theta(y_i - x_i); y_i - [x_i + \theta(y_i - x_i)]) = (1 - \theta)\psi_i^+(\theta)$$

and, by (2),

$$(3) \quad (\psi_i^0)^+(\theta) \leq -\varepsilon_0/2,$$

$$(4) \quad \psi_i^1(\theta) + (1 - \theta)(\psi_i^1)^+(\theta) \leq -\varepsilon_0/2 \quad (i \in J, \quad \theta \in [0, \bar{\theta}]).$$

Since, for each $t \in T$, each component of $\psi_i(\theta)(t)$ is a continuous function of θ , relation (4) implies that

$$\psi_i^1(\bar{\theta}) = \phi^1(x_i + \bar{\theta}(y_i - x_i)) \leq 0 \quad (i \in J),$$

while relation (3) yields

$$\phi^0(x_i + \bar{\theta}(y_i - x_i)) = \psi_i^0(\bar{\theta}) \leq \psi_i^0(0) - \varepsilon_0\bar{\theta}/2 \\ = \phi^0(x_i) - \varepsilon_0\bar{\theta}/2 \quad (i \in J).$$

The above two relations show that

$$\phi^0(x_{i+1}) \leq \phi^0(x_i + \bar{\theta}(y_i - x_i)) + \eta_i \leq \phi^0(x_i) - \varepsilon_0\bar{\theta}/2 + \eta_i \quad (i \in J).$$

Since $(\phi^0(x_i))$ is nonincreasing, this implies that $\lim_i \phi^0(x_i) = -\infty$ which cannot be because ϕ^0 is continuous on the compact set K .

Step 2. We have thus shown that $\lim_i \alpha_i = 0$. Now let x_∞ be the limit of $(x_i)_{i \in J}$ for some $J \subset (1, 2, \dots)$. For each $i \in J$, set

$$F_i = \hat{\phi}(x_i) + D\phi(x_i; K - x_i), \\ Z_i = \{z = (z^0, z^1) \in \mathbb{R} \times C(T, \mathbb{R}^m) \mid z < -\alpha_i - \eta_i\}.$$

Then F_i and Z_i are nonempty disjoint convex subsets of $\mathbb{R} \times C(T, \mathbb{R}^m)$ and Z_i has a nonempty interior. It follows that there exists $\lambda_i = (\lambda_i^0, \lambda_i^1) \in \mathbb{R} \times C(T, \mathbb{R}^m)^*$ such that $|\lambda_i| = 1$ and

$$(5) \quad \lambda_i f \geq \lambda_i z \quad (f \in F_i, \quad z \in Z_i).$$

Let

$$s = 1 + \sup \{\alpha_i + \eta_i \mid i = 0, 1, 2, \dots\}, \quad z_0^0 = -s, \quad z_0^1 = -s, \quad z_0 = (z_0^0, z_0^1).$$

Since all the sets Z_i contain the same open ball $B(z_0, 1)$ of center z_0 and radius 1, it follows from (5) that

$$(6) \quad \lambda_i(\hat{\phi}(x_i) + D\phi(x_i; y - x_i)) \geq \lambda_i z_0 + |\lambda_i| = \lambda_i z_0 + 1 \quad (y \in K, \quad i \in J).$$

We may choose $J_1 \subset J$ such that $(\lambda_i)_{i \in J_1}$ converges in the weak star topology to some λ . For $y = x_i$, relation (6) yields

$$\lambda \hat{\phi}(x_\infty) = \lim_{i \in J_1} \lambda_i \hat{\phi}(x_i) \geq \lim_{i \in J_1} \lambda_i z_0 + 1 = \lambda z_0 + 1$$

and this shows that $\lambda \neq 0$. It follows now from (5) that, for every $z \in \mathbb{R} \times C(T, \mathbb{R}^m)$ with $z < 0$, we have

$$(7) \quad \lambda(\hat{\phi}(x_\infty) + D\phi(x_\infty; y - x_\infty)) \geq \lambda z \quad (y \in K).$$

Now λ^1 can be identified with a Radon measure $(\lambda^{1,1}, \dots, \lambda^{1,m})$ on T with values in \mathbb{R}^m . Thus (7) yields, for $y = x_\infty$,

$$\int \phi^1(x_\infty)(t) \cdot \lambda^1(dt) \geq \lambda^0 z^0 + \int z^1(t) \cdot \lambda^1(dt) \quad (z < 0),$$

whence we conclude that λ^0 and λ^1 are nonnegative and $\lambda^{1,j}(\{t \in T | \phi^{1,j}(x_\infty)(t) < 0\}) = 0$. Finally, relation (b) follows from (7) by letting $z \rightarrow 0$. Q.E.D.

REFERENCES

- [1] R. O. BARR, *An efficient computational procedure for a generalized quadratic programming problem*, this Journal, 7 (1969), pp. 415–429.
- [2] M. FRANK AND P. WOLFE, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1956), pp. 95–110.
- [3] E. G. GILBERT, *An iterative procedure for computing the minimum of a quadratic form on a convex set*, this Journal, 4 (1966), pp. 61–80.
- [4] J. KOWALIK, *A new method for constrained optimization problems*, Operations Res., 17 (1969), pp. 973–983.
- [5] D. D. MORRISON, *Optimization by least squares*, SIAM J. Numer. Anal., 5 (1968), pp. 83–86.
- [6] T. PECSVARADI AND K. NARENDRA, *A new iterative procedure for the minimization of a quadratic form on a convex set*, this Journal, 8 (1970), pp. 396–402.
- [7] E. POLAK, *Computational Methods in Optimization*, Academic Press, New York, 1971.
- [8] E. POLAK AND D. Q. MAYNE, *A feasible directions algorithm for optimal control problems with control and terminal inequality constraints*, IEEE Trans. Autom. Control, AC22 (1977), pp. 741–751.
- [9] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
- [10] ———, *Steepest descent with relaxed controls*, this Journal, 15 (1977), pp. 674–682.

CONTROLABILITE DES SYSTEMES BILINEAIRES*

J. P. GAUTHIER† ET G. BORNARD‡

Abstract. This paper deals with bilinear systems. A complementary result to the well-known Jurdjevic-Kupka criterion for controllability is obtained.

Resumé. Ce travail traite de la controlabilité des systèmes bilinéaires. On présente une amélioration du critère de Jurdjevic-Kupka dans le cas de $SL(n, R)$.

Sous des hypothèses ouvertes, le résultat obtenu est une condition nécessaire et suffisante de controlabilité dans le cas d'un espace d'état non compact, ce qui est nouveau (si on exclut le cas classique des systèmes linéaires) pour les systèmes comportant une dérive.

1. Introduction. Les systèmes que l'on considère sont de la forme:

$$(1) \quad \dot{x} = Ax + uBx$$

ou encore:

$$(2) \quad \dot{x} = uAx + vBx$$

où u et v sont dans R et x est dans $GL^+(n, R)$, $SL(n, R)$ ou $R^n - \{0\}$ selon le cas. $GL^+(n, R)$ est le groupe des automorphismes de R^n à déterminant positif; (GL^+ est connexe). $SL(n, R)$ est la restriction du groupe précédent au groupe des automorphismes à déterminant égal à 1.

Dans les deux premiers cas, Ax et Bx doivent être compris comme des champs de vecteurs invariants à droite sur les groupes considérés, avec:

$$Ax = A(x) = A \circ x$$

où \circ représente la composition des applications linéaires.

Dans le cas de $R^n - \{0\}$, la transitivité de l'action naturelle de ces groupes sur $R^n - \{0\}$ fait que la controlabilité passe de $GL^+(n, R)$ ou $SL(n, R)$ à $R^n - \{0\}$. Nous verrons que, dans le cas des résultats qu'on présente ici, la controlabilité est en fait *équivalente* sur ces 3 espaces.

On notera $gl(n, R)$ et $sl(n, R)$ pour les algèbres de Lie de ces deux groupes (respectivement: $M(n)$ l'ensemble des matrices carrées d'ordre n à coefficients réels, et l'ensemble des matrices carrées réelles d'ordre n de trace nulle).

R est le centre de $gl(n, R)$, et $sl(n, R)$ est une algèbre de Lie simple. On a:

$$gl(n, R) = sl(n, R) \oplus R$$

En conséquence, avec l'aide de résultats classiques de controlabilité (Jurdjevic-Kupka [1, prop. 19, p. 75] pour les systèmes de la forme (1), Boothby [7, Lemme p. 305], pour les systèmes de la forme (2)), on peut voir que la controlabilité sur $SL(n, R)$ (partie simple de $GL^+(n, R)$) est *équivalente* à la controlabilité sur $GL^+(n, R)$ si et

* Received by the editors August 15, 1980, and in revised form March 31, 1981.

† Centre National de la Recherche Scientifique, Ecole Nationale Polytechnique Alger, 10 av. Pasteur, El Harrach, Alger, Algerie. This work was completed while the author was at the Laboratoire d'Automatique de Grenoble.

‡ Centre National de la Recherche Scientifique—Laboratoire d'Automatique de Grenoble, Ecole Nationale Supérieure d'Ingenieurs Electriciens de Grenoble, BP. 46, 38402 St. Martin d'Heres, France.

seulement si:

- (3) —cas (1): $\text{tr}(B) \neq 0$,
 —cas (2): $\text{tr}(A)$ ou $\text{tr}(B) \neq 0$.

En fait, dans [1] et [7], il est seulement dit que la controlabilité sur $SL(n, R)$ implique la controlabilité sur $GL^+(n, R)$ dans le cas (3). Pour vérifier la réciproque, il suffit de voir que, si (3) est faux, alors $\det(x)$ dans (1) ou (2) n'a pas toute la liberté voulue.

En ce qui concerne les groupes et compte tenu de ce qui vient d'être dit, nous pouvons donc restreindre l'étude à $SL(n, R)$ (c'est à dire supposer $\text{tr}(A) = 0$ et $\text{tr}(B) = 0$). Pour $R^n - \{0\}$, nous préciserons plus loin.

Dans tout ce qui suit on supposera que B a des valeurs réelles, ce qui n'est pas indispensable, mais simplifie la démarche.

Nous ferons, sur A et B , les hypothèses suivantes: Soit Ad la représentation adjointe sur $SL(n, R)$; on supposera que B est un élément "*fortement régulier*"; c'est à dire que $\text{Ker}(\text{Ad}(B))$ est une sous-algèbre de Cartan (isomorphe à l'ensemble des matrices diagonales de trace nulle), et chacun des sous-espaces propres de $\text{Ad}(B)$ associés à une valeur propre non nulle est de dimension 1 (cas réel). Cette hypothèse est *fondamentale* pour nos objectifs.

Il est facile de vérifier que l'ensemble de ces hypothèses peut se résumer à:

$B = \text{diag}(b_1, b_2, \dots, b_n)$, matrice diagonale de trace nulle avec:

- H1. $b_1 < b_2 < \dots < b_n$,
 $b_i - b_j \neq b_k - b_m$ pour $(i, j) \neq (k, m)$.

Dans le cas des systèmes de la forme (1), nous ferons aussi l'hypothèse suivante, introduite par Jurdjevic-Kupka dans [1]:

- H2. $a_{1n} \cdot a_{n1} > 0$.

Nous pouvons maintenant donner le résultat suivant, qui a été le point de départ de ce travail:

THÉORÈME 1 (Jurdjevic-Kupka) [1]. *Supposons: H1; H2;*

- H3. $a_{ij} \neq 0$ pour $|i - j| = 1$.

Alors, le système (1) est controlable sur $SL(n, R)$.

Il y a par ailleurs le résultat suivant (que nous donnons pour des raisons qui apparaîtront clairement immédiatement après):

THÉORÈME 2. *Supposons: H1; H3.*

Alors, le système (2) est controlable sur $SL(n, R)$.

Notons que, dans ce cas, ceci veut simplement dire que l'algèbre de Lie engendrée par A et B est exactement $sl(n, R)$.

Notre contribution principale est constituée par les résultats suivants:

THÉORÈME 3 (Résultat principal). *Supposons: H1 et H2 dans le cas (1); H1 dans le cas (2). Alors, une condition nécessaire et suffisante de controlabilité sur $SL(n, R)$ est:*

- H4. *A est une matrice permutation-irréductible.*

COROLLAIRE 1. *H4, sous les mêmes hypothèses, est aussi une condition nécessaire et suffisante de controlabilité sur $R^n - \{0\}$.*

Le paragraphe 2 définit et donne des caractérisations classiques des matrices "permutation-irréductibles" (et fournit un algorithme de calcul). Le paragraphe 3 donne la preuve de notre résultat principal. Le paragraphe 4 fait quelques commentaires et donne un résultat complémentaire.

2. Matrices “permutation-irréductibles”.

DÉFINITION 1. Une matrice carrée A ($n \times n$), réelle (ou complexe) est dite *permutation-réductible* s’il existe une matrice de *permutation* P telle que:

$$P^{-1}AP = \begin{bmatrix} A_1 & A_2 \\ 0 & A_3 \end{bmatrix}$$

ou A_3 est une matrice carrée $r \times r$ avec $0 < r < n$.

DÉFINITION 2. Une matrice carrée A ($n \times n$), est dite *permutation-irréductible* si elle n’est pas permutation-réductible.

Remarque. L’hypothèse H4 signifie qu’il n’existe pas de sous-ensemble de $SL(n, R)$ qui soit simultanément A -invariant et B -invariant.

Nous noterons maintenant P -réductible et P -irréductible. Nous allons caractériser ces matrices. La preuve des résultats énoncés ici peut être trouvée dans [2] ou [3] par exemple.

Une matrice carrée A , ($n \times n$), étant donnée, on définit le *graphe* de A de la façon suivante: considérons un ensemble de n points ordonnés de 1 à n . Ce sont les noeuds du graphe. On mène un *arc orienté* du noeud i au noeud j si et seulement si $a_{ij} \neq 0$.

THÉORÈME 4. Une matrice A est P -irréductible si et seulement si son graphe est fortement connexe (i.e., si, pour tout couple (i, j) de noeuds, il y a dans le graphe de A un chemin orienté allant de i à j).

Donnons maintenant un algorithme qui permet de tester si une matrice A est irréductible (ce résultat est utile pour l’utilisation pratique du théorème 3).

On note par \tilde{A} la matrice suivante, associée à A :

$$\tilde{a}_{ii} = 0, \quad \tilde{a}_{ij} = \begin{cases} 1 & \text{si } a_{ij} \neq 0 \\ 0 & \text{si } a_{ij} = 0 \end{cases} \quad \text{si } i \neq j.$$

THÉORÈME 5. Soit A une matrice carrée $n \times n$. Alors, la suite \tilde{A}^p des puissances booléennes de \tilde{A} se stabilise pour $p \leq n - 1$, et A est P -irréductible si et seulement si $(\tilde{A}^{n-1})_{ij} = 1$ pour tout (i, j) .

Par “puissances booléennes” on veut dire que le “produit” de matrices est considéré sur l’anneau de Boole ordinaire, les matrices considérées n’ayant que 0 ou 1 pour termes.

3. Preuve du résultat principal. Notons (comme dans [1]) par $LS(T)$, l’ensemble Lie-saturé d’un sous-ensemble T de $SL(n, R)$; en particulier, on considérera:

$$T = \{A, B, -B\} \quad \text{dans le cas (1)}$$

et

$$T = \{A, -A, B, -B\} \quad \text{dans le cas (2)}.$$

$LS(T)$ est le plus grand sous-ensemble de $SL(n, R)$ qui est équivalent à T du point de vue de l’accessibilité. La définition exacte de $LS(T)$ est la suivante ([1]):

$$LS(T) = L(T) \cap \text{Sat}(T)$$

où:

- $L(T)$ est la sous-algèbre de Lie de $sl(n, R)$ engendrée par les éléments de T ;
- $\text{Sat}(T)$ est l’ensemble de tous les éléments X de $sl(n, R)$ tels que

$$\{\exp(tX) | t \in R^+\} \subset Cl(S^+(T));$$

- Cl représente la fermeture dans la topologie naturelle;
- $S^+(T)$ est le sous-semigroupe de $SL(n, R)$ engendré par l'union $\bigcup_{X \in T} \{\exp(tX) | t \in R^+\}$ de tous les semigroupes engendrés par les éléments de T .

Remarquons que, dans le cas (2), $LS(T) = L(T)$.

Dans [1], on peut trouver les propriétés suivantes, pour $LS(T)$:

- (4) $LS(T)$ est un cône, convexe, fermé.

Si X et $\pm Y$ sont dans $LS(T)$, alors,

- (5)
$$e^{\text{Ad}(vY)} X = \exp(vY) X \exp(-vY)$$

est aussi dans $LS(T)$ pour tout v dans R .

- (6) Si $\pm X$ et $\pm Y$ sont dans $LS(T)$, alors, $\pm[X, Y]$ est aussi dans $LS(T)$.

Une condition nécessaire et suffisante de controlabilité de (1)

- (7) ou (2) sur $SL(n, R)$ est que $LS(T) = sl(n, R)$.

Nous emploierons, dans la suite, les notations suivantes, pour certaines matrices carrées $n \times n$. E_{ij} est la matrice telle que

$$(E_{ij})_{km} = \delta_{ik} \cdot \delta_{jm}, \text{ où } \delta_{ij} = \begin{cases} 1 & \text{si } i = j, \\ 0 & \text{sinon.} \end{cases}$$

De plus, A étant donné, ${}^i A^i$ et ${}_j A_i$ sont définis comme suit:

- toute colonne de ${}^i A^i$ est zéro, à l'exception de $j^{\text{ième}}$ qui est A^i , la $i^{\text{ième}}$ colonne de A ,
- toute ligne de ${}_j A_i$ est zéro, à l'exception de la $j^{\text{ième}}$ qui est A_i , la $i^{\text{ième}}$ ligne de A ,

et on utilisera souvent les identités suivantes:

$$E_{ij}A = {}_i A_j, \quad AE_{ij} = {}^i A^j, \quad [E_{ij}, A] = {}_i A_j - {}^i A^j.$$

Nous pouvons maintenant énoncer les lemmes nécessaires:

LEMME 1. Supposons que $\pm E_{ij}$ et E_{jm} ($i \neq j, j \neq m, i \neq m$) soient dans $LS(T)$. Alors, $\pm E_{im}$ est aussi dans $LS(T)$.

Preuve. D'après (5), on a

$$H(v) = \exp(-E_{ij}v) E_{jm} \exp(E_{ij}v) \in LS(T)$$

pour tout v dans R . Mais on a

$$H(v) = E_{jm} - vE_{im} = |v| \left(\frac{E_{jm}}{|v|} \pm E_{im} \right).$$

Du fait de (4), $(E_{jm}/|v| \pm E_{im})$ est donc dans $LS(T)$, et puisque $LS(T)$ est fermé (encore (4)), en faisant $|v| \rightarrow \infty$, on obtient le résultat.

LEMME 2. Supposons que A et $\pm E_{ij}$ soient dans $LS(T)$ avec $i \neq j$. Alors $\pm({}^i A^i - {}_i A_j)$ est aussi dans $LS(T)$.

Preuve. Même preuve que pour le lemme 1, avec

$$H(v) = \exp(-E_{ij}v) A \exp(E_{ij}v).$$

LEMME 3. Supposons que A , $\pm B$, et $\pm E_{ij}$ soient dans $LS(T)$ avec $i \neq j$. Soit c_{km} ($k \neq m$) un terme différent de zéro, (non diagonal), se trouvant dans ${}^iA^i$ ou dans ${}_iA_j$.

Alors, $\pm c_{km}$ est aussi dans $LS(T)$.

Preuve. D'après le lemme 2, $\pm({}^iA^i - {}_iA_j)$ est dans $LS(T)$. Alors, pour tout v dans R , on sait que

$$H(v) = \pm \exp(vB)({}^iA^i - {}_iA_j) \exp(-vB)$$

est dans $LS(T)$. Mais

$$H(v)_{km} = \pm({}^iA^i - {}_iA_j)_{km} e^{(b_k - b_m)v}.$$

Soit (k_0, m_0) tel que $(b_{k_0} - b_{m_0})$ est le maximum (minimum) des $(b_k - b_m)$ tels que ${}_{km} = ({}^iA^i - {}_iA_j)_{km} \neq 0$.

Considérons alors $K(v)$:

$$K(v) = \pm \frac{\exp(vB)({}^iA^i - {}_iA_j) \exp(-vB)}{e^{(b_{k_0} - b_{m_0})v}}.$$

Chacun des termes de $K(v)$ est soit nul, soit de la forme:

$$\pm c_{km} \frac{e^{(b_k - b_m)v}}{e^{(b_{k_0} - b_{m_0})v}}.$$

Clairement, quand $v \rightarrow +\infty$, $(-\infty)$, tous ces termes tendent vers 0 sauf un. D'autre part, encore d'après (4), $\lim_{v \rightarrow \infty} K(v) = \pm c_{k_0 m_0} E_{k_0 m_0}$ est dans $LS(T)$. Par conséquent, $\pm E_{k_0 m_0}$ est dans $LS(T)$.

Maintenant considérons $\pm\{{}^iA^i - {}_iA_j) - c_{k_0 m_0} E_{k_0 m_0}\}$. On peut itérer le raisonnement qui a été fait. Mais $(b_{k_0} - b_{m_0})$ aura diminué, et, successivement, on obtiendra dans $LS(T)$ tous les termes voulus; ce qui prouve le résultat.

LEMME 4. A et $\pm B$ étant dans $LS(T)$, supposons que A soit P -irréductible, et que $\pm E_{ij}$ et $\pm E_{ji}$, pour un couple (i, j) donné, soient dans $LS(T)$. Alors, $LS(T) = sl(n, R)$.

Preuve. A étant irréductible, pour tout $m \neq j$, il existe un chemin orienté dans le graphe de A , du noeud j au noeud m . Ce chemin peut être choisi sans boucle, c'est-à-dire que, si $a_{ji_1}, a_{i_1 i_2}, \dots, a_{i_p m}$ sont les termes $\neq 0$ dans A correspondant à ce chemin, il n'y a pas de répétition d'indice, et en particulier:

$$i_k \neq j, \quad i_k \neq m \quad \text{pour tout } k.$$

Mais, nous avons besoin de plus, et on aimerait que ce chemin ne passe pas par i . Deux cas se présentent alors:

- ou bien ce chemin ne passe pas par i ,
- ou bien, pour un certain k , $i_k = i$.

Alors, choisissons le plus grand k dans le chemin tel que $i_k = i$ et coupons ce chemin. On obtient $a_{ii_{k+1}}, a_{i_{k+1} i_{k+2}}, \dots, a_{i_p m}$, tel que, pour tout h ,

$$i_{k+h} \neq i, \quad i_{k+h} \neq m, \quad i_{k+h} \neq j.$$

Et alors, dans tous les cas, on a obtenu: soit un chemin sans boucle de j à m , ne passant pas par i , soit un chemin sans boucle de i à m ne passant pas par j .

Si on est dans la seconde situation, on se ramène à la première par une permutation des indices i et j qui laisse inchangées les hypothèses du lemme.

Supposons donc que ce chemin soit de j à m , et soit $a_{ji_1}, a_{i_1 i_2}, \dots, a_{i_p m}$, la séquence correspondante.

On sait que $\pm E_{ij} \in LS(T)$. Le terme $\pm(a_{ji_1}E_{ii_1})_{ii_1}$ est un terme $\neq 0$ non diagonal dans $\pm_i A_j$ (c'est pourquoi il a fallu considérer plus haut l'éventualité où $i_k = i$). En appliquant le lemme 3, on obtient que $\pm E_{ii_1} \in LS(T)$.

En itérant le processus, on voit que $\pm E_{ii_2}, \dots, \pm E_{im}$ sont dans $LS(T)$.

Maintenant, considérons $\pm E_{ji} \in LS(T)$, et $E_{im} \in LS(T)$. D'après le lemme 1, ceci implique que $\pm E_{jm}$ est dans $LS(T)$.

Prenant maintenant un chemin de m à j , et faisant la même démonstration en sens inverse, on obtient aisément que $\pm E_{mi} \pm E_{mj}$ sont dans $LS(T)$ pour tout m , $m \neq i$, $m \neq j$.

Soit $k \neq i, j, m$. En appliquant le même raisonnement que précédemment on obtient que $\pm E_{km}$ et $\pm E_{mk}$ sont dans $LS(T)$.

Par conséquent, $\pm E_{km} \in LS(T)$ pour tout couple (k, m) , $k \neq m$. Mais l'ensemble des E_{km} , $k \neq m$, qui est l'ensemble des vecteurs racines de $sl(n, R)$, engendre $sl(n, R)$ en tant qu'algèbre de Lie. On a donc $LS(T) = sl(n, R)$, de qui prouve le résultat.

Preuve du théorème 3 (résultat principal).

Nécessité. Supposons que A soit une matrice P -réductible. Il y a alors une matrice de permutation P telle que

$$(8) \quad P^{-1}AP = \begin{vmatrix} c_1 & c_2 \\ 0 & c_3 \end{vmatrix},$$

c_3 étant une matrice carrée $a \times r$, $0 < r < n$.

Or une matrice de permutation laisse trivialement B sous forme diagonale, ce qui implique que A et B sont tous deux conjugués à des matrices de la forme (8).

Mais l'ensemble des matrices de la forme (8) constitue une sous-algèbre de Lie de $sl(n, R)$; par conséquent $L(T) \neq sl(n, R)$, donc, $LS(T) = L(T) \cap \text{Sat}(T) \neq sl(n, R)$, ce qui contredit (7).

Suffisance.

Cas des systèmes de la forme (2). A étant irréductible, B fortement régulier, on montre aisément que $\pm E_{ij} \in LS(T)$ pour tout couple (i, j) tel que $a_{ij} \neq 0$, $i \neq j$ (la preuve est du même type que celle du lemme 3).

Considérons maintenant un couple (i, j) , et le chemin de i à j :

$$a_{ii_1} \neq 0, \quad a_{i_1 i_2} \neq 0, \quad \dots, \quad a_{i_p j} \neq 0,$$

auquel correspond la suite de matrices

$$\pm E_{ii_1}, \quad \pm E_{i_1 i_2}, \quad \dots, \quad \pm E_{i_p j}.$$

En appliquant le lemme 1 et en itérant, on voit que $\pm E_{ij} \pm E_{ij} \in LS(T)$ pour tout couple (i, j) , $i \neq j$. Donc, $LS(T) = sl(n, R)$.

Cas des systèmes de la forme (1). On montre d'abord comme dans [1], qu'à cause de l'hypothèse H2 ($a_{1n} \cdot a_{n1} < 0$), $\pm E_{1n}$ et $\pm E_{n1}$ sont dans $LS(T)$.

Appliquons (5) avec $X = A$, $Y = B$, divisons par $e^{(b_1 - b_n)v}$ et faisons successivement tendre v vers $+\infty$ et $-\infty$. (On peut le faire parce que $LS(T)$ est un cône fermé).

Pour simplifier, supposons $a_{1n} > 0$, $a_{n1} < 0$. Alors, on obtient $E_{1n} \in LS(T)$, $-E_{n1} \in LS(T)$. Donc, $E_{1n} - E_{n1} \in LS(T)$; mais les trajectoires du champ $X = E_{1n} - E_{n1}$ sont périodiques. Par conséquent:

$$\{\exp(tX), t < 0\} \subset S^+(T),$$

et donc

$$-X = E_{n1} - E_{1n} \in \text{Sat}(T).$$

Mais comme $-X \in L(T)$, on a donc, $-X \in LS(T)$.

Maintenant, avec (4) encore,

$$E_{n1} - E_{1n} + E_{1n} = E_{n1} \quad \text{et} \quad E_{n1} - E_{1n} - E_{n1} = -E_{1n}$$

sont dans $LS(T)$. On a ainsi $\pm E_{1n}$ et $\pm E_{n1}$ dans $LS(T)$.

Appliquons maintenant le lemme 4, on en déduit que $LS(T) = sl(n, R)$, et avec (7), le système est controlable.

Preuve du corollaire 1. La condition est suffisante, c'est clair. Pour la nécessité, examinons la forme (8): On voit que sur $R^n - \{0\}$, il y a un sous espace invariant, celui engendré par les $n - r$ premières composantes.

4. Commentaires. Examinons ce que signifie exactement le théorème 3 et son corollaire.

Appelons, comme dans [10] par "*condition du rang*", le fait que le rang de l'algèbre de Lie engendrée par le système soit, en tout point de l'espace d'état, égal à la dimension de l'espace d'état. Ceci n'est en fait qu'*exceptionnellement* une condition nécessaire et *suffisante* de controlabilité: Cas des systèmes linéaires sur R^n , cas d'un système symétrique ($X \in T, -X \in T$), cas où la "dérive" du système est périodique ou "Poisson stable", cas d'un groupe compact, cas du produit semi-direct d'un groupe compact K par R^n , K agissant irréductiblement sur R^n [11].

Le théorème 3 et son corollaire montrent que:

THÉORÈME 6. *Supposons H1, sur $SL(n, R)$ ou sur $R^n - \{0\}$, indifféremment. Alors, la "condition du rang" est équivalente à l'hypothèse H4 (P-irréductibilité de A).*

Mais, de plus:

THÉORÈME 7. *Supposons H1 et H2, sur $SL(n, R)$ ou sur $R^n - \{0\}$, indifféremment. Alors, la "condition du rang" est une condition nécessaire et suffisante de controlabilité pour le système (1) comportant une dérive (non compacte).*

Ceci nous semble très important et doit être compris comme la contribution majeure de ce travail.

Les théorèmes 1 et 2 sont obtenus immédiatement, comme corollaires du théorème 3, dès lors que l'on remarque que la matrice de Jacobi:

$$J = \begin{vmatrix} 0 & 1 & & 0 \\ & & & 1 \\ 1 & & & \\ & & & 1 \\ 0 & & & 0 \end{vmatrix}$$

est irréductible (ou que les termes $\neq 0$ dans J correspondent justement aux vecteurs racines associés aux racines primitives de $sl(n, R)$, relativement à la sous-algèbre de Cartan définie par B ; c'est le point de vue émis dans [1]).

Maintenant, on peut se poser la question suivante: La P-irréductibilité des matrices joue-t-elle un rôle dans le cas de la controlabilité sur les groupes généraux?

La réponse semble être oui, comme le montrent les considérations suivantes:

Soit g une algèbre de Lie qui est la forme réelle d'une algèbre de Lie complexe simple g^c , et G le groupe de Lie correspondant à g .

Soit B un élément régulier de g . ($\text{Ker}(\text{Ad}(B))$ est une sous-algèbre de Cartan, c'est-à-dire une sous-algèbre maximale commutative égale à son propre normalisateur dans g^c .)

De tels éléments forment un ouvert dense sur g .

Soit ϕ une représentation irréductible de g sur un espace vectoriel complexe V .

(Dans le cas d'un groupe simple, on peut prendre $V = g^c$ et $\phi = \text{Ad}$.) $\tilde{\phi}$, défini par $\tilde{\phi}(x + iy) = \phi(x) + i\phi(y)$, définit une représentation irréductible de g^c sur g^c .

B est inclus dans la sous-algèbre de Cartan qu'il définit ($\text{Ad}(B)(B) = 0$), et tous les éléments X dans cette sous-algèbre de Cartan sont tels que les matrices $\phi(X)$ sont diagonalisables (simultanément). Par conséquent, on peut supposer que $\phi(B) = \tilde{\phi}(B)$ est diagonale.

THÉORÈME 7. *Soit B tel que ci-dessus, et $A \in g$; alors, une condition nécessaire de controlabilité du système de champs de vecteurs invariants à droite sur G : $T = \{A, \pm B\}$, est que $\phi(A)$ soit P -irréductible.*

Preuve. Pour que le système soit controlable, il faut que $L(T) = g$. Par conséquent, il faut que $L(T)$ engendre aussi g^c , les éléments de $L(T)$ étant pris en tant qu'éléments d'une algèbre de Lie complexe.

Supposons que $\phi(A)$ soit P -réductible. Il existe alors une représentation irréductible de g sur V telle que $\phi(A)$ et $\phi(B)$ soient tous deux sous la forme (8); mais il y a un sous-espace $\phi(g^c)$ invariant sur V , et donc la représentation ϕ n'est pas irréductible.

En particulier, $\text{Ad}(A)$ doit être P -irréductible.

Remerciements. Nous remercions vivement le Professeur Y. Kupka pour l'attention qu'il a portée à ce travail et pour les conseils précieux qu'il nous a prodigués.

BIBLIOGRAPHIE

- [1] V. JURDJEVIC ET Y. KUPKA, *Accessibility on semi-simple Lie groups and their homogeneous spaces*. Nov. 1977, à paraître dans "Annales de l'Institut Fourier".
- [2] P. VARGA, *Matrix Iterative Analysis*. Prentice-Hall, Englewood Cliffs, NJ, 1962.
- [3] F. ROBERT, *Matrices nonnégatives et normes vectorielles*. Cours de 3e cycle, INPG ENSIMAG, 1973.
- [4] H. SAMELSON, *Notes on Lie Algebras*. Van Nostrand-Reinhold Mathematical studies, Van Nostrand, New York, 1969.
- [5] L. AUSLANDER ET R. E. MACKENZIE, *Introduction to Differentiable Manifolds*. McGraw-Hill, New York, 1963.
- [6] W. M. BOOTHBY, *An Introduction to Differentiable Manifolds and Riemannian Geometry*. Academic Press, New York, 1975.
- [7] ———, *A transitivity problem from control theory*, J. Differential Equations, 17 (1975), pp. 296–307.
- [8] W. M. BOOTHBY AND E. N. WILSON, *Determination of the transitivity of bilinear systems*, SIAM J. Control. Optim., 17 (1979), pp. 212–221.
- [9] N. NAIMARK ET A. STERN, *Théorie des représentations des groupes*, Mir, Moscow, 1979.
- [10] R. HERMANN ET A. J. KRENER, *Nonlinear controllability and observability*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 728–740.
- [11] B. BONNARD, V. JURDJEVIC, Y. KUPKA ET G. SALLET, *Controlabilité sur le produit semi direct d'un groupe compact par un E.V. réel*. Trans. Amer. Math. Soc., to appear.
- [12] C. LOBRY, *Controlabilité des systèmes non-linéaires*, SIAM J. Control, 8 (1970), pp. 573–605.
- [13] ———, *Quelques aspects qualitatifs de la théorie de la commande*, Thèse d'état, Grenoble, 1972.
- [14] H. SUSSMANN ET V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.

ON THE EXISTENCE OF OPTIMAL CONTROLS FOR PARTIALLY OBSERVED DIFFUSIONS*

U. G. HAUSSMANN†

Abstract. The problem of optimally controlling a partially observed diffusion process is shown to have a solution in two cases: when the set of admissible controls is compact, or when the set of admissible controls is the set of randomized controls.

1. Introduction. Consider the following control problem. X_t is the process to be controlled, Y_t is the observation, and $u(t)$, the control, is to depend only on $\{Y_s: 0 \leq s \leq t\}$, for $0 \leq t \leq T < \infty$, T nonrandom. Assume $X_t \in \mathbb{R}^d$, $Y_t \in \mathbb{R}^m$, $u(t) \in U \subset \mathbb{R}^p$ the set of control points, satisfy

$$(1.1) \quad dX_t = f(t, X_t, Y_t, u(t, Y \cdot)) dt + \sigma(t, X_t, Y_t, u(t, Y \cdot)) dw_t,$$

$$(1.2) \quad dY_t = h(t, X_t, Y_t) dt + \bar{\sigma}(t, Y_t) d\bar{w}_t, \quad Y_0 = 0.$$

For convenience only assume that X_0 is fixed. Here w , \bar{w} are independent \mathbb{R}^d and \mathbb{R}^m valued Wiener processes on (Ω, \mathcal{F}, P) . The problem is to minimize

$$(1.3) \quad J[u] = E \left\{ \int_0^T f_0(t, X_t, Y_t, u(t, Y \cdot)) dt + g(X_T, Y_T) \right\}.$$

The question of existence of such optimal controls has been open for about 20 years. Recently Christopheit [3] and Kohlmann [8] solved problems somewhat more general than (1.1)–(1.3) assuming uniform equicontinuity in x of the admissible controls as well as compactness (as functions of t) in $L_1([0, T])$ of the set of admissible controls. Practically speaking, their hypotheses imply compactness of the set of admissible controls as a subset in $L_1([0, T] \times \mathcal{C}^m[0, T])$ ($\mathcal{C}^m[0, T]$ is the space of continuous functions: $[0, T] \rightarrow \mathbb{R}^m$). Assuming this compactness (without necessarily the equicontinuity in x of u) an optimal control is obtained in the obvious direct way if f , σ are Lipschitz in x . We state and prove this simple result precisely in the next section, adding some examples.

The hypothesis of compactness of the admissible controls is very annoying; it is too strong. Recently Fleming and Pardoux [5] showed, using results from nonlinear filtering theory as well as PDE's and weak convergence, that a *randomized* optimal control does exist. They assume that f is affine in u and U is convex compact. In the completely observable case, i.e., $u(t)$ is a function of $\{X_s: s \leq t\}$, it is known that the set of "states" is sequentially compact, so that $X(u^n) \rightarrow X$ if $\{u^n\}$ is a minimizing sequence. Then the convexity hypothesis " $f(t, x, y, U)$ convex for each (t, x, y) " can be used to show that X is generated by an admissible control which must then be optimal [1], [9]. If we strengthen this hypothesis to

$$(C) \quad f(t, \cdot, y, U) \text{ is convex for each } (t, y),$$

i.e., if for u_1, u_2 in U , $0 < \lambda < 1$, there is a point u in U , depending on t, y, u_1, u_2, λ , such that for all x

$$\lambda f(t, x, y, u_1) + (1 - \lambda) f(t, x, y, u_2) = f(t, x, y, u),$$

* Received by the editors March 3, 1981, and in final form June 9, 1981. This research was supported by the Natural Sciences and Engineering Research Council of Canada under grant A8051.

† Department of Mathematics, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Y4.

then we prove that there exists a randomized optimal control for the problem (1.1), (1.2), (1.3). Thus our result is the same as that in [5], except that we have relaxed the hypotheses on the system parameters. Originally it was hoped that our formulation would identify cases where a nonrandomized optimal control exists, but this was not achieved. Our method of proof consists of first eliminating h by a Girsanov transformation, also used in [5], and then following the method of Kushner, [9], but in a function space setting. In contrast to [5], we do not use filtering theory or PDE's. In § 3 we state the main result that the minimum cost over the controls adapted to the observations is attained by a randomized control, relegating its proof to § 4. Some extensions, including the result in [5] are established in § 5. The appendix contains some results from measure theory.

It should be added that Kushner [20, Chap. 11], and Bismut [19] have also obtained results for similar problems.

2. The compact case. We consider the problem

$$(2.1) \quad \min \{J[u]: u \in \tilde{\mathcal{U}}\}$$

subject to

$$(2.2) \quad X_t = X_0 + \int_0^t f(s, X_s, Y_s, u(s, Y_s)) ds + \int_0^t \sigma(s, X_s, Y_s, u(s, Y_s)) dw_s,$$

$$(2.3) \quad Y_t = \int_0^t h(s, X_s, Y_s) ds + \int_0^t \bar{\sigma}(s, Y_s) d\bar{w}_s,$$

where $\tilde{\mathcal{U}}$ will be defined below. We assume:

(A₁) X_0 is fixed.

(A₂) $h, f, f_0, g, \sigma, \bar{\sigma}$ are Borel measurable and continuous in (x, u) for each (t, y) .

(A₃) $\bar{\sigma}(t, y)$ is invertible for each (t, y) , $|\bar{\sigma}(t, y)| \leq K$,

$$|\bar{\sigma}(t, y) - \bar{\sigma}(t, \tilde{y})| \leq K|y - \tilde{y}|, \quad |\bar{\sigma}(t, y)^{-1}h(t, x, y)| \leq K(1 + |y|).$$

$$(A_4) \quad |f(t, x, y, u) - f(t, \tilde{x}, y, u)| + |\sigma(t, x, y, u) - \sigma(t, \tilde{x}, y, u)| \leq K|x - \tilde{x}|,$$

$$|f(t, x, y, u)| + |\sigma(t, x, y, u)| \leq K(1 + |x|).$$

$$(A_5) \quad |f_0(t, x, y, u)| + |g(x, y)| \leq K(1 + |x|^r + |y|^r), \quad r \text{ a finite integer.}$$

We also assume that we are given a probability space with filtration, $(\Omega, \mathcal{F}, Q, \{\mathcal{F}_t\}_{[0, T]})$, carrying independent standard Wiener processes $w \in \mathbb{R}^q$, $\beta \in \mathbb{R}^m$ ($\bar{\sigma}$ is $m \times m$, σ is $d \times q$). We write $\mathcal{C}(X; Y)$ for the set of continuous functions from X to Y . $\mathcal{C}([0, T]; \mathbb{R}^m)$ is written as $\mathcal{C}^m[0, T]$. On $\mathcal{C}^m[0, T]$ we let \mathcal{G} denote the Borel algebra and \mathcal{G}_t the σ -algebra generated by the paths up to time t , i.e., generated by the sets $\{y \in \mathcal{C}^m[0, T]: y(s) \in B\}$ where s is any time in $[0, t]$ and B is any Borel set in \mathbb{R}^m . We let \mathcal{B}^m be the Borel algebra on \mathbb{R}^m .

$\tilde{\mathcal{U}}$ is assumed to be a subset of \mathcal{U}^s , the *strictly admissible controls*. We define \mathcal{U}^s to be the set of functions possessing the following two properties:

- (i) $u: ([0, T] \times \mathcal{C}^m[0, T], \mathcal{B} \times \mathcal{G}) \rightarrow (\mathcal{U}, \mathcal{B}^p)$ is measurable,
- (ii) $u(t, \cdot)$ is \mathcal{G}_t -measurable.

(Measurable means that the preimage of a measurable set is measurable.) Further conditions will be imposed on $\tilde{\mathcal{U}}$ below.

We can now define the solutions of (2.2), (2.3) considered. By (A₃) we can solve

$$(2.4) \quad dY_t = \bar{\sigma}(t, Y_t) d\beta_t$$

uniquely and independently of u . Given $u \in \mathcal{U}^s$, we substitute $Y_t(\omega)$, $u(t, Y_t(\omega))$ into (2.2) and solve it uniquely by (A₄). Moreover (A₄) also implies that $\sup_u E \sup_t |X_t|^q < \infty$ for all $q < \infty$. Finally we define ($'$ denotes transpose)

$$(2.5) \quad \tilde{Z}_t = \exp \left\{ \int_0^t [\bar{\sigma}(s, Y_s)^{-1} h(s, X_s, Y_s)]' d\beta_s - \frac{1}{2} \int_0^t |\bar{\sigma}(s, Y_s)^{-1} h(s, X_s, Y_s)|^2 ds \right\}.$$

From (A₃) it follows that P defined by

$$(2.6) \quad \frac{dP}{dQ} = \tilde{Z}_T$$

is a probability measure and by [7] if

$$(2.7) \quad d\bar{w}_t \equiv d\beta_t - \bar{\sigma}(t, Y_t)^{-1} h(t, X_t, Y_t) dt$$

then w_t , \bar{w}_t are independent standard Wiener processes on $(\Omega, \mathcal{F}, \mathcal{P}, \{\mathcal{F}_t\})$. Hence X , Y satisfy (2.2), (2.3) on (Ω, \mathcal{F}, P) , and solutions of (2.2), (2.3) are law unique. Note that we may begin directly with solutions \tilde{X} , \tilde{Y} of (2.2), (2.3) where w , \bar{w} may depend on u . Then we fix one w^0 and β^0 for all u and use these as w , β in our construction to obtain (X, Y) with the same distribution as (\tilde{X}, \tilde{Y}) , hence with the same expected cost.

Let us write E_P for expectation with respect to P . Then

$$J[u] = E_P \left\{ \int_0^T f_0 dt + g \right\} = E_Q \left\{ \left(\int_0^T f_0 dt + g \right) \tilde{Z}_T \right\}.$$

Observe that (A₃) implies that there exists $p_1 > 1$, $K_1 < \infty$, such that for all u

$$(2.8) \quad E_Q \tilde{Z}_T^{p_1} \leq K_1.$$

Let R be the product measure $\lambda \times Q \circ Y^{-1}$ where λ is Lebesgue measure.

THEOREM 2.1. Assume (A₁)–(A₅). If $\tilde{\mathcal{U}}$ is compact in $L_1([0, T] \times \mathcal{C}^m[0, T], \mathcal{B} \times \mathcal{G}, R); U$ then an optimal control exists for the problem (2.1)–(2.3).

Proof. Let $\{u^n\}$ be a minimizing sequence so that $J[u^n] \rightarrow J^* = \inf \{J[u] : u \in \tilde{\mathcal{U}}\}$. By compactness there exists $u^0 \in \tilde{\mathcal{U}}$ and a subsequence, again denoted by $\{u^n\}$, such that $u^n \rightarrow u^0$ in L_1 . By [6, Thm. 2, p. 52] and the remark following it, we have

$$E \sup_t |X_t^n - X_t^0|^2 \rightarrow 0,$$

where X^k is the solution of (2.2) corresponding to u^k , $k = 0, 1, 2, \dots$. Thus for a further subsequence, again denoted by $\{u^n\}$, we have

$$\sup_t |X_t^n - X_t^0| \rightarrow 0 \quad \text{w.p.1.}$$

Hence for each t , $f_0(t, X_t^n, Y_t, u^n(t, Y_t)) \rightarrow f_0(t, X_t^0, Y_t, u^0(t, Y_t))$ w.p.1, and $g(X_T^n, Y_T) \rightarrow g(X_T^0, Y_T)$ w.p.1. Since $\sup_n E \sup_t |X_t^n|^q < \infty$, then (2.8) and (A₅) imply

$$E_Q \left\{ \tilde{Z}_T^n \left(\int_0^T |f_0(t, X_t^n, Y_t, u^n) - f_0(t, X_t^0, Y_t, u^0)| dt + |g(X_T^n, Y_T) - g(X_T^0, Y_T)| \right) \right\} \rightarrow 0.$$

Moreover (A₃) and the continuity of h in x imply

$$E_Q \left| \int_0^T (\bar{\sigma}^{-1}[h(X^n) - h(X^0)]' d\beta \right|^2 \rightarrow 0,$$

hence $\tilde{Z}_t^n \rightarrow \tilde{Z}_t^0$ in probability. Equation (2.8) now gives

$$E_Q \left\{ (\tilde{Z}_T^n - \tilde{Z}_T^0) \left(\int_0^T f_0(t, X_t^0, Y_t, u^0) dt + g(X_T^0, Y_T) \right) \right\} \rightarrow 0,$$

so that $J[u^n] \rightarrow J[u^0]$ and the proof is complete.

Remark 2.1. According to [4, IV. 8.18] $\tilde{\mathcal{U}}$ is compact in $L_1([0, T] \times \mathcal{C}^m[0, T])$ if

(a) $\tilde{\mathcal{U}}$ is closed and bounded;

(b) given that π is the family of *finite* subalgebras of $\mathcal{B} \times \mathcal{G}$ partially ordered by inclusion (except that the elements of the subalgebra may be modified by sets of measure zero), then

$$\limsup_{\mathcal{A} \in \pi} \sup_{u \in \tilde{\mathcal{U}}} \int \int |E_R\{u(t, y) | \mathcal{A}\} - u(t, y)| dR = 0.$$

We give some examples of sets $\tilde{\mathcal{U}}$ which satisfy (i), (ii), (a), and (b).

Example 1. If U is compact, and if $\tilde{\mathcal{U}}$ is equicontinuous and adapted, e.g., for each t, y there are continuous functions $r_{ty}(\cdot), r_y(\cdot)$ such that $r_{ty}(z) \downarrow 0, r_y(z) \downarrow 0$ with z and

$$(2.9) \quad |u(t, \bar{y}) - u(t, y)| \leq r_y \left(\sup_{0 \leq s \leq t} |\bar{y}(s) - y(s)| \right),$$

$$(2.10) \quad |u(t, y) - u(s, y)| \leq r_{ty}(|t - s|),$$

then $\tilde{\mathcal{U}}$ is compact in $\mathcal{C}([0, T] \times \mathcal{C}^m[0, T]; U)$ with the topology of uniform convergence on compact subsets. Since the identity map $\mathcal{C} \rightarrow L_1$ is continuous by the bounded convergence theorem, then $\tilde{\mathcal{U}}$ is compact in L_1 .

Example 2. $\tilde{\mathcal{U}}$ is compact in $\mathcal{C}(\mathcal{C}^m[0, T]; L_1([0, T]; U))$, hence in L_1 , if U is compact, and if (2.9), (2.10) hold with $r_y(z) \downarrow 0$ with z and

$$\lim_{h \rightarrow 0} \int_0^T r_{ty}(h) dt = 0.$$

This is essentially the setting for the results of Christopheit [3] and Kohlmann [8], although they treat more general systems than (2.2), (2.3).

Example 3 (Finite sampled observations). Of the three examples, this is probably the only one of any importance. If $\{0 = t_0, t_1, \dots, t_M = T\}$ is a finite subset of $[0, T]$, if $\{D_1, D_2, \dots, D_N\}$ is a finite, disjoint Borel partition of \mathbb{R}^m , and if \mathcal{A} is the finite algebra generated by

$$E_{ij} = \{(t, y): t \in (t_{i-1}, t_i], y(t_{i-1}) \in D_j\},$$

then $\tilde{\mathcal{U}} \equiv \{u: [0, T] \times \mathcal{C}^m[0, T] \rightarrow U, u \text{ is } \mathcal{A}\text{-measurable}\}$ is compact for U compact. Note $\tilde{\mathcal{U}} \simeq U^{MN}$.

In fact the general compactness criterion in L_1 requires that the admissible controls can be approximated in L_1 by "finite" controls uniformly in $\tilde{\mathcal{U}}$.

Let us add that X_0 need not be constant. It can be an \mathcal{F}_0 measurable random variable, such that $E_Q |X_0|^{\bar{q}} < \infty$ for some $\bar{q} > rp_1/(p_1 - 1)$ with p_1 as in (2.8). In the proof it is sufficient to use this \bar{q} . In case $|\bar{\sigma}^{-1}h|$ is bounded, then any $p_1 < \infty$ will do and all we need is $\bar{q} > r$.

3. The convex case. We turn now to a more general problem. Let us state some assumptions. (σ is now independent of u , but only for convenience, cf. Remark 4.4.)

(H₁) U is a compact metric space, $U \neq \emptyset$.

(H₂) f, f_0 are Borel measurable, continuous in (x, y) uniformly in u , a.e. t , and continuous in u locally uniformly in x for each (t, y) , and

$$|f(t, x, y, u)| \leq K(1 + |x| + |y|), \quad |f_0(t, x, y, u)| \leq K(1 + |x|^r + |y|^r).$$

$\sigma(t, x, y)$ is bounded, Borel-measurable, and continuous in (x, y) , a.e. t . Note that "locally uniformly in x " means that for any compact set C the property holds uniformly in x for x in C .

(H₃) $\bar{\sigma}(t, y)$ is Borel-measurable, invertible, bounded and locally Lipschitz in y .

(H₄) h is Borel-measurable, continuous in (x, y) for almost all t and

$$|\bar{\sigma}(t, y)^{-1}h(t, x, y)| \leq K(1 + |x| + |y|).$$

(H₅) g is continuous and

$$|g(x, y)| \leq K(1 + |x|^r + |y|^r).$$

(H₆) X_0 is fixed in \mathbb{R}^d .

The traditional problem is to establish the existence of a control which minimizes $J[u]$ over the strictly admissible controls \mathcal{U}^s , where $u: [0, T] \times \mathcal{C}^m[0, T] \rightarrow U$ is *strictly admissible* if it is $(\mathcal{B} \times \mathcal{G})$ -measurable and $\{\mathcal{G}_t\}$ -adapted, and if there is a probability space with filtration, $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}_{0 \leq t \leq T})$ carrying independent standard Wiener processes w, \bar{w} , and $\{\mathcal{F}_t\}$ -adapted processes X, Y , such that (1.1) and (1.2) hold. With the assumption of Lipschitz continuity of f, σ as in the last section, the two definitions of \mathcal{U}^s coincide.

Unfortunately we cannot prove the existence of an optimal strictly admissible control. We can only show that $J^* \equiv \inf \{J[u]: u \in \mathcal{U}^s\}$ is attained by a *randomized* control.

\mathcal{U} , the set of *admissible* controls, is the set of randomized controls, where $u \in \mathcal{U}$ if u is an adapted stochastic process on some probability space with filtration $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}_{0 \leq t \leq T})$ carrying independent Wiener processes w, \bar{w} , a \mathcal{H} -valued random variable ζ (\mathcal{H} is a metric space) and adapted stochastic processes X, Y such that

$$(3.1) \quad dX_t = f(t, X_t, Y_t, u_t) dt + \sigma(t, X_t, Y_t) dw_t,$$

$$(3.2) \quad dY_t = h(t, X_t, Y_t) dt + \bar{\sigma}(t, Y_t) d\bar{w}_t,$$

and such that

- (i) $u(t, \omega)$ is $\mathcal{B} \times \mathcal{F}^{Y_\zeta}$ -measurable;
- (ii) $\{u_r: r \leq t\}$ and $\{Y_s: s > t\}$ are conditionally independent given \mathcal{F}_t^Y , with respect to a new measure Q to be introduced in the next section (Q is obtained from P by a Girsanov transformation to remove the drift in (3.2));
- (iii) (u, Y) and w are independent with respect to Q ;
- (iv) Y and ζ are independent with respect to Q ;
- (v) (X, w, \bar{w}) and ζ are conditionally independent given Y, u .

Our notation here is that \mathcal{F}^{Y_ζ} denotes the subalgebra of \mathcal{F} generated by ζ and $\{Y_s: 0 \leq s \leq T\}$, and \mathcal{F}_t^Y denotes the subalgebra generated by $\{Y_s: 0 \leq s \leq t\}$ and the null sets of \mathcal{F} for any stochastic process Y and random variable ζ .

We make some remarks now.

(a) The growth conditions imply that $\sup_u \sup_t E|Y_t| < \infty$ for all k , and $\sup_u \sup_t E|X_t|^k < \infty$ if $E|X_0|^k < \infty$ again true for all k by H₆. Thus $J[u] < \infty$.

(b) ζ represents the randomization. If ζ is in fact constant, then (i) and (ii) imply that u is $\{\mathcal{F}_t^Y\}$ adapted, i.e., u is strictly admissible, so that $\mathcal{U}^s \in \mathcal{U}$.

(c) In general (i), (ii) do *not* imply that u is $\{\mathcal{F}_t^Y\}$ -adapted, only that for all measurable sets A , the process $(t, \omega) \rightarrow Q(u_t \in A | \mathcal{F}_t^Y)(\omega)$ is $\{\mathcal{F}_t^Y\}$ -adapted. Here $Q(u_t \in A | \mathcal{F}_t^Y)$ denotes the conditional probability under Q that $u_t \in A$ given Y . Hence, although the control is conditionally independent of the future observations and of the Brownian motion (at least with respect to Q), we cannot claim that there is a feedback law expressing u_t as a function of $\{Y_s: s \leq t\}$ or even of $\{(\zeta, Y_s): s \leq t\}$. Note that if $\bar{\sigma}(t, y) = \bar{\sigma}(t)$, then (ii) is equivalent to: $\{(u_s, Y_s): s \leq t\}$ and $\{Y_s - Y_t: s > t\}$ are independent with respect to Q .

(d) If μ, p represent the distributions of Y, ζ respectively under Q , and if by (i) u is written as $u(Y, \zeta) \equiv u_Y(\zeta)$, then the joint distribution of (Y, u) is given by

$$Q_{Yu}(A \times B) \equiv \int_A p[u_Y^{-1}(B)] \mu(dy),$$

since Y, ζ are independent. This measure is called a randomized control in [5]; hence our randomized controls are randomized controls in the sense of [5].

Let us write $l(t, x, y, u)'$ for $(f(t, x, y, u)', f_0(t, x, y, u))$.

THEOREM 3.1. Assume (H_1) – (H_6) . If l satisfies condition (C), then $J^* \equiv \inf \{J[u]: u \in \mathcal{U}^s\} = J[u^*]$ for some $u^* \in \mathcal{U}$.

We remark that l satisfies (C) if

$$l(t, x, y, u) = \sum_{i=1}^N \phi_i(t, x, y) \psi_i(t, y, u)$$

and $\Psi(t, y, U)$ is convex for each (t, y) , where $\Psi = (\psi_1, \psi_2, \dots, \psi_N)$, a composed matrix.

Finally X_0 can be allowed to be random; i.e., one specifies the distribution D on \mathbb{R}^d and then demands of the admissible controls that $D = P \circ X_0^{-1}$. (H_6) must then be replaced by: for $\mu > p_1 r / (p_1 - 1)$, $\mu \geq 2$,

$$\int_{\mathbb{R}^d} |x|^\mu D(dx) < \infty,$$

where p_1 is related to the densities $\exp Z_T$ as in § 2. Again if $|\bar{\sigma}^{-1}h|$ is bounded then we only need $\mu > r$, $\mu \geq 2$.

4. The proof. We shall break the proof of Theorem 3.1 into several lemmas. Note that K with or without subscripts always stands for a constant. Let $\{u^n\}$ be a minimizing sequence in \mathcal{U}^s and write $\alpha(t, y)$ for $\bar{\sigma}(t, y)\bar{\sigma}(t, y)'$. On each $(\Omega_n, \mathcal{F}_n, P_n, \{\mathcal{F}_t^n\})$ define

$$(4.1) \quad Z_t^n = \int_0^t h(s, X_s^n, Y_s^n)' \alpha(s, Y_s^n)^{-1} dY_s^n - \frac{1}{2} \int_0^t |\bar{\sigma}(s, Y_s^n)^{-1} h(s, X_s^n, Y_s^n)|^2 ds,$$

and define Q_n by

$$\frac{dQ_n}{dP_n} = e^{-Z_7^n}.$$

Hence $(\Omega_n, \mathcal{F}_n, Q_n, \{\mathcal{F}_t^n\})$ is a probability space carrying the independent Brownian motions w^n and β^n with

$$\beta_t^n = \bar{w}_t^n + \int_0^t \bar{\sigma}(s, Y_s^n)^{-1} h(s, X_s^n, Y_s^n) ds,$$

and we have $u_t^n \equiv u^n(t, Y^n)$.

$$(4.2) \quad dX_t^n = f(t, X_t^n, Y_t^n, u_t^n) dt + \sigma(t, X_t^n, Y_t^n) dw_t^n,$$

$$(4.3) \quad dY_t^n = \bar{\sigma}(t, Y_t^n) d\beta_t^n.$$

Q_n is the measure Q referred to in (iii) of the definition of randomized control.

Let us denote E_{Q_n} by E_Q^n . Then

$$E_Q^n \left\{ e^{Z_T^n} \left[\int_0^T f_0(t, X_t^n, Y_t^n, u_t^n) dt + g(X_T^n, Y_T^n) \right] \right\} \rightarrow J^*$$

as $n \rightarrow \infty$. We define

$$(4.4) \quad \begin{aligned} \bar{F}_t^n &= \int_0^t f(s, X_s^n, Y_s^n, u_s^n) ds : \Omega \rightarrow \mathcal{C}^d[0, T], \\ F_t^n &= \int_0^t f(s, \cdot, Y_s^n, u_s^n) ds : \Omega \rightarrow \mathcal{C}([0, T]; \mathcal{C}^d(\mathbb{R}^d)), \\ B_t^n &= \int_0^t \sigma(s, X_s^n, Y_s^n) dw_s^n : \Omega \rightarrow \mathcal{C}^d[0, T], \\ \bar{B}_t^n &= \int_0^t \bar{\sigma}(s, Y_s^n) d\beta_s^n : \Omega \rightarrow \mathcal{C}^m[0, T], \\ H_t^n &= \int_0^t (\bar{\sigma}(s, Y_s^n)^{-1} h(s, X_s^n, Y_s^n))' d\beta_s^n : \Omega \rightarrow \mathcal{C}[0, T], \\ G_t^n &= \int_0^t f_0(s, \cdot, Y_s^n, u_s^n) ds : \Omega \rightarrow \mathcal{C}([0, T]; \mathcal{C}(\mathbb{R}^d)), \\ \bar{G}_t^n &= \int_0^t f_0(s, X_s^n, Y_s^n, u_s^n) ds : \Omega \rightarrow \mathcal{C}[0, T], \\ \Phi_t^n &= (w_t^n, X_t^n, Y_t^n, \bar{F}_t^n, \bar{G}_t^n, H_t^n, B_t^n, \bar{B}_t^n, F_t^n, G_t^n). \end{aligned}$$

Observe that w.p.1

$$(4.5) \quad J[u^n] = E_Q^n \{ e^{Z_T^n} [\bar{G}_T^n + g(X_T^n, Y_T^n)] \},$$

$$(4.6) \quad X_t^n = X_0 + \bar{F}_t^n + B_t^n,$$

$$(4.7) \quad Y_t^n = \bar{B}_t^n,$$

$$(4.8) \quad Z_t^n = H_t^n - \frac{1}{2} \int_0^t |\bar{\sigma}(s, Y_s^n)^{-1} h(s, X_s^n, Y_s^n)|^2 ds.$$

Let \tilde{Q}_n be the measure induced by Φ^n on $\mathcal{C}^{q+3d+2m+2}[0, T] \times \mathcal{C}([0, T]; \mathcal{C}^{d+1}(\mathbb{R}^d))$.

LEMMA 4.1. *The sequence of measures $\{\tilde{Q}_n\}$ is tight.*

Proof. The growth conditions imply that for any k there is a constant C_k such that $E_Q^n \sup_t \{|Y_t^n|^k + |X_t^n|^k\} \leq C_k$ for all n , so that (H₂) and (H₄) imply

$$(4.9) \quad E_Q^n \left(\int_t^{t+\Delta} |f| + |\sigma|^2 + |\bar{\sigma}^{-1} h|^2 + |\bar{\sigma}|^2 ds \right)^2 \leq K \Delta^2.$$

From [2, Thm. 12.3] it follows that $\{(w_t^n, X_t^n, Y_t^n, \bar{F}_t^n, \bar{G}_t^n, H_t^n, B_t^n, \bar{B}_t^n)\}$ is tight (cf. [9]).

Recall that $l = (f', f_0)'$, and define $L_t^n = (F_t^n, G_t^n)'$. It remains to show that $\{L_t^n\}$ is tight; i.e., given $\varepsilon > 0$ we must construct $C_\varepsilon \subset \mathcal{C}([0, T]; \mathcal{C}^{d+1}(\mathbb{R}^d))$, C_ε compact, such that for all n , $Q_n \circ (L^n)^{-1}(C_\varepsilon) \geq 1 - \varepsilon$, or $\Pr \{L^n \in C_\varepsilon\} \geq 1 - \varepsilon$. Since $\{Y^n\}$ is tight there

is a compact set $C_0 \subset \mathcal{C}^m[0, T]$ such that $\Pr\{Y^n \in C_0\} \geq 1 - \varepsilon$; hence for some k_0 , $\Pr\{\sup_t |Y_t^n| > k_0\} < \varepsilon$ for all n . Let $A_n = \{\omega : \sup_t |Y_t^n| \leq k_0\}$. Then $Q_n(A_n) \geq 1 - \varepsilon$ for all n . According to Ascoli's theorem [13, p. 179], $\bigcup_{n=1}^\infty \bigcup_{\omega \in A_n} \{L_t^n(\omega)\} \equiv D$ is conditionally compact if $\{L_t^n(\omega)\}$ is equicontinuous, and for each t , $\bigcup_n \bigcup_{\omega \in A_n} \{L_t^n(\omega)\}$ is conditionally compact in $\mathcal{C}^{d+1}(\mathbb{R}^d)$ using the (metric) topology of uniform convergence on compact subsets. The metric is given by

$$\rho(\phi, \psi) = \sum_{k=1}^{\infty} 2^{-k} \sup_{|x| \leq k} \frac{|\phi(x) - \psi(x)|}{1 + |\phi(x) - \psi(x)|}.$$

1. *Equicontinuity.* Given $\eta > 0$ fix r such that $2^{-r} \leq \eta/2$. Then

$$\sup_{|t-s| < \delta} \rho(L_t^n, L_s^n) \leq \frac{\eta}{2} + \sup_{|t-s| < \delta} \sup_{|x| \leq r} \int_s^t |l(\tau, x, Y_\tau^n, u_\tau^n)| d\tau \leq \eta,$$

if δ is sufficiently small, since l is locally bounded and $\sup_t |Y_t^n| \leq k_0$.

2. *Conditional compactness.* Fix t . We must show that for any $N < \infty$, $\bigcup_n \bigcup_{\omega} \{L_t^n(\omega)\}$ is conditionally compact in $\mathcal{C}^{d+1}(\{x : |x| \leq N\})$. But

$$\sup \{|L_t^n(\omega)(0)| : \omega \in A_n, n = 1, 2, \dots\} < \infty,$$

and

$$\sup_{|x-\bar{x}| < \delta} \left| \int_0^t l(s, x, Y_s^n, u_s^n) - l(s, \bar{x}, Y_s^n, u_s^n) ds \right| < \eta$$

for δ sufficiently small by Lemma A.4 of the appendix. Hence again by the Ascoli theorem, $\bigcup_n \bigcup_{\omega} \{L_t^n(\omega)\}$ is conditionally compact.

If C_ε is the closure of D then it is compact. Moreover

$$\begin{aligned} Q_n \circ (L^n)^{-1}(C_\varepsilon) &\geq Q_n \circ (L^n)^{-1} \left(\bigcup_{i=1}^\infty \{L^i(\omega) : \omega \in A_i\} \right) \\ &\geq Q_n \circ (L^n)^{-1} \{L^n(\omega) : \omega \in A_n\} \\ &\geq Q_n(A_n) \geq 1 - \varepsilon. \end{aligned}$$

This establishes the lemma.

It now follows from Prokhorov's theorem [2] that there is a subsequence of $\{\tilde{Q}_n\}$, i.e., of $\{Q_n \circ (\Phi^n)^{-1}\}$, again denoted by $\{\tilde{Q}_n\}$, which converges weakly to \tilde{Q}_0 . Since our space is complete, separable, metric, we can apply Skorokhod's lemma [12, p. 10] to conclude that there are new random variables $\hat{\Phi}^n$ all defined on the same space (Ω, \mathcal{F}, Q) such that $\sup_t \rho[\hat{\Phi}_t^n, \hat{\Phi}_t^0] \rightarrow 0$ w.p.1., and Φ^n and $\hat{\Phi}^n$ have the same distribution. In fact (4.4)–(4.8) also hold for the $\hat{\cdot}$ processes for some Brownian motion $(\hat{w}^n, \hat{\beta}^n)$, as explained in [9, p. 355]. Let us now drop the $\hat{\cdot}$ notation. If we use (4.8) with $n = 0$ to define Z^0 then (4.5)–(4.8) also hold for Φ^0 .

We must now find the required ζ .

Let V be a continuous stochastic process assuming values in \mathcal{V} , a metric space with Borel sets \mathcal{B}_V . Recall that $\mathcal{F}^Y = (Y^0)^{-1}(\mathcal{G})$.

LEMMA 4.2. *There exist a (Lusin) metric space \mathcal{H} and probability space $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{Q})$ carrying processes $(\hat{Y}, \hat{L}, \hat{V}, \hat{\zeta})$ such that*

- (i) $\hat{\Omega} = \Omega \times \mathcal{H}$, $(\hat{Y}, \hat{L}, \hat{V}, \hat{\zeta})(t, \omega, z) = (Y^0(t, \omega), L^0(t, \omega), V(t, \omega), z)$;
- (ii) $(\hat{Y}, \hat{L}, \hat{V})$ have the same distribution as (Y^0, L^0, V) ;
- (iii) $\hat{V}, \hat{\zeta}$ are conditionally independent given (\hat{Y}, \hat{L}) ;

(iv) \hat{Y} is independent of ζ ;

(v) there is an $\mathcal{F}^Y \times \mathcal{B}_{\mathcal{H}}$ measurable function \tilde{L} such that $\hat{L} = \tilde{L}$ a.e. \hat{Q} .

Proof. Write \mathcal{C}_1 for $\mathcal{C}^m[0, T]$, \mathcal{C}_2 for $\mathcal{C}([0, T]; \mathcal{C}^{d+1}(\mathbb{R}^d))$, $\tilde{\mathcal{C}}$ for $\mathcal{C}_1 \times \mathcal{C}_2$, \mathcal{G}_t , \mathcal{G}_t^2 , $\tilde{\mathcal{G}}_t$ for the corresponding canonical Borel filtrations, and \mathcal{G}^2 , $\tilde{\mathcal{G}}$ for \mathcal{G}_T^2 , $\tilde{\mathcal{G}}_T$ (we already defined \mathcal{G}_t with the admissible controls). Define a measure \tilde{P} on $(\tilde{\mathcal{C}}, \tilde{\mathcal{G}})$ by

$$\tilde{P}(A) = Q(\{\omega : (Y^0(\omega), L^0(\omega)) \in A\}).$$

Set $\mathcal{F}^{YL} = (Y^0, L^0)^{-1}(\tilde{\mathcal{G}})$, i.e., the σ -algebra generated by (Y^0, L^0) , so that the regular conditional probability $Q(\cdot | \mathcal{F}^{YL})$ can be regarded as a measurable function on $\tilde{\mathcal{C}}$. Hence for $A \in \mathcal{F}$ (abusing notation)

$$Q(A) = \int_{\Omega} Q(A | \mathcal{F}^{YL}) dQ = \int_{\tilde{\mathcal{C}}} Q(A | \mathcal{F}^{YL}) d\tilde{P},$$

and if $\tilde{A} \in \tilde{\mathcal{G}}$

$$Q(A \cap (Y^0, L^0)^{-1}\tilde{A}) = \int_{\tilde{A}} Q(A | \mathcal{F}^{YL}) d\tilde{P}.$$

We shall now apply a result of [14]. Let μ , a measure on $(\mathcal{C}_1, \mathcal{G})$, be the distribution of Y^0 , and define

$$\tilde{\Omega} = \{(Y^0(\omega), L^0(\omega)) : \omega \in \Omega, (Y^n(\omega), L^n(\omega)) \rightarrow (Y^0(\omega), L^0(\omega))\}.$$

Then $\tilde{\Omega} \subset \tilde{\mathcal{C}}$ is a separable metric space and $\tilde{P}(\tilde{\Omega}) = 1$. Let us define $\gamma : \tilde{\Omega} \rightarrow \mathcal{C}_1$ by $\gamma(y, l) = y$.

Clearly γ is continuous and $\mu(A) = \tilde{P} \circ \gamma^{-1}(A)$. Also by Lemma A.1 (or as in the proof of the previous lemma) if $C \subset \mathcal{C}_1$ is compact, then $\gamma^{-1}(C)$ is conditionally compact.

In [14] it is assumed that $\tilde{\Omega}$ is complete but this is not actually required for the proof. It follows [14, pp. 260–264] that there exists a metric space \mathcal{H} (which is a Borel subset of a compact metric space, hence Lusin) supporting a Borel measure p and probability measures $P_z(\cdot)$ such that

$$\tilde{P}(\cdot) = \int_{\mathcal{H}} P_z(\cdot) p(dz).$$

P_z can be extended to $\tilde{\mathcal{G}}$ in the obvious manner. We have, moreover, for each z , $\mu = P_z \circ \gamma^{-1}$, and $\tilde{\mathcal{G}} = \gamma^{-1} \mathcal{G} \pmod{P_z}$, i.e., if $A \in \tilde{\mathcal{G}}$ then there is $B \in \gamma^{-1} \mathcal{G}$ such that $P_z(A \Delta B) = 0$. It follows that $\tilde{\mathcal{G}} = \mathcal{G} \times \{\phi, \mathcal{C}_2\} \pmod{P_z}$. Finally we note also that $z \rightarrow P_z(A)$ is Borel measurable for each A in $\tilde{\mathcal{G}}$; cf. the proof in [14].

Define \hat{P} on $\tilde{\mathcal{G}} \times \mathcal{B}_{\mathcal{H}}$ by

$$\hat{P}(A \times B) = \int_B P_z(A) dp.$$

Now define \hat{Q} on $(\Omega \times \mathcal{H}, \mathcal{F} \times \mathcal{B}_{\mathcal{H}})$ by

$$\hat{Q}(A \times B) = \int_{\tilde{\mathcal{C}} \times B} Q(A | \mathcal{F}^{YL}) d\tilde{P}.$$

\hat{Q} is well defined by Fubini's theorem [15, II.14]. Observe that, if $A \in \mathcal{F}^{YL}$ so $A = (Y^0, L^0)^{-1}\tilde{A}$, $\tilde{A} \in \tilde{\mathcal{G}}$, $B \in \mathcal{B}_{\mathcal{H}}$, then

$$\hat{Q}(A \times B) = \hat{P}(\tilde{A} \times B).$$

Define $(\hat{Y}, \hat{L}, \hat{V}, \zeta)$ from (i).

To establish (ii) notice that

$$\begin{aligned}
 \Pr \{((\hat{Y}, \hat{L}), \hat{V}) \in A \times B\} &= \int_{\tilde{\mathcal{C}} \times \mathcal{H}} Q((Y^0, L^0)^{-1}(A) \cap V^{-1}(B) | \mathcal{F}^{YL}) d\hat{P} \\
 &= \int_{\tilde{\mathcal{C}} \times \mathcal{H}} 1_A Q(V^{-1}(B) | \mathcal{F}^{YL}) d\hat{P} \\
 &= \int_{\mathcal{H}} \int_{\tilde{\mathcal{C}}} 1_A Q(V^{-1}(B) | \mathcal{F}^{YL}) dP_z dp \\
 &= \int_{\tilde{\mathcal{C}}} 1_A Q(V^{-1}(B) | \mathcal{F}^{YL}) d\tilde{P} \\
 &= \int_A Q(V^{-1}(B) | \mathcal{F}^{YL}) d\tilde{P} \\
 &= Q([(Y^0, L^0), V]^{-1}(A \times B)) \\
 &= \Pr \{((Y^0, L^0), V) \in A \times B\},
 \end{aligned}$$

where we have used [15, II, 14/16]. Thus (ii) holds.

Next consider $\hat{Q}(\hat{V}^{-1}(\cdot) | \mathcal{F}^{YL} \times \mathcal{B}_{\mathcal{H}})$. For $A \in \mathcal{F}^{YL}$, $A = (Y^0, L^0)^{-1}\tilde{A}$, $B \in \mathcal{B}_{\mathcal{H}}$, $C \in \mathcal{B}_{\mathcal{V}}$ we have

$$\begin{aligned}
 \int_{A \times B} \hat{Q}(\hat{V}^{-1}(C) | \mathcal{F}^{YL} \times \mathcal{B}_{\mathcal{H}}) d\hat{Q} &= \hat{Q}((A \times B) \cap \hat{V}^{-1}(C)) \\
 &= \int_{\tilde{A} \times B} Q(V^{-1}(C) | \mathcal{F}^{YL}) d\tilde{P} \\
 &= \int_{A \times B} Q(V^{-1}(C) | \mathcal{F}^{YL})(Y^0(\omega), L^0(\omega)) d\hat{Q}.
 \end{aligned}$$

From this it follows that

$$\hat{Q}(\hat{V}^{-1}(\cdot) | \mathcal{F}^{YL} \times \mathcal{B}_{\mathcal{H}}) = Q(V^{-1}(\cdot) | \mathcal{F}^{YL}) = \hat{Q}(\hat{V}^{-1}(\cdot) | \mathcal{F}^{YL} \times \{\phi, \mathcal{H}\});$$

hence (iii) is established.

Since $\mu = P_z \circ \gamma^{-1}$ then $\Pr \{\hat{Y} \in A, \zeta \in B\} = \mu(A)p(B)$ and (iv) follows.

Finally for any A in $\tilde{\mathcal{G}}$ and any function ψ such that $\psi: \tilde{\mathcal{C}} \times \mathcal{H} \rightarrow \mathbb{R}$ is $\mathcal{G} \times \{\phi, \mathcal{C}_2\} \times \mathcal{B}_{\mathcal{H}}$ measurable we have

$$\begin{aligned}
 E_{\hat{P}}[\hat{P}[A \times \mathcal{H} | \mathcal{G} \times \{\phi, \mathcal{C}_2\} \times \mathcal{B}_{\mathcal{H}}] \psi] &= E_{\hat{P}}[\psi 1_{A \times \mathcal{H}}] \\
 &= E_p E_{P_z}[\psi 1_{A \times \mathcal{H}}] \\
 &= E_p E_{P_z}[\psi E_{P_z}[1_A | \mathcal{G} \times \{\phi, \mathcal{C}_2\}]] \\
 &= E_{\hat{P}}[P_z[A | \mathcal{G} \times \{\phi, \mathcal{C}_2\}] \psi].
 \end{aligned}$$

Hence $\hat{P}(A \times \mathcal{H} | \mathcal{G} \times \{\phi, \mathcal{C}_2\} \times \mathcal{B}_{\mathcal{H}})(\cdot, \cdot, z) = P_z(A | \mathcal{G} \times \{\phi, \mathcal{C}_2\})(\cdot, \cdot)$ a.e. \hat{P} .

Since $\tilde{\mathcal{G}} = \mathcal{G} \times \{\phi, \mathcal{C}_2\} \pmod{P_z}$, then

$$P_z[A | \mathcal{G} \times \{\phi, \mathcal{C}_2\}] = 1_A \quad \text{a.e. } (P_z).$$

Thus (cf. [14] also)

$$1_{A \times \mathcal{H}} = \hat{P}(A \times \mathcal{H} | \mathcal{G} \times \{\phi, \mathcal{C}_2\} \times \mathcal{B}_{\mathcal{H}}) \quad \text{a.e. } \hat{P},$$

hence $\tilde{\mathcal{G}} \times \{\phi, \mathcal{H}\} \subset \overline{\mathcal{G} \times \{\phi, \mathcal{C}_2\} \times \mathcal{B}_{\mathcal{H}}}$, the completion of $\mathcal{G} \times \{\phi, \mathcal{C}_2\} \times \mathcal{B}_{\mathcal{H}}$ in $(\tilde{\mathcal{C}} \times \mathcal{H}, \tilde{\mathcal{G}} \times \mathcal{B}_{\mathcal{H}}, \hat{P})$, i.e., augmented by subsets of null sets of $\tilde{\mathcal{G}} \times \mathcal{B}_{\mathcal{H}}$. Now (v) follows by Lemma A.2.

We can now consider all variables to be defined on $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{Q})$ by setting $V = (w, X, \bar{F}, \bar{G}, H, B, \bar{B}, \Phi^1, Z^1, \Phi^2, Z^2, \dots)$. Now, replacing \hat{L} by \tilde{L} and dropping the $\hat{\cdot}$, we still have (4.4)–(4.8) as before. (The superscript 0 has been dropped from Φ^0 .)

Let us now show that the limit L (i.e., \tilde{L}) can be written as an integral. Note, since we have dropped $\hat{\cdot}$ then ω refers to the pair (old ω, z).

LEMMA 4.3. *There exists a $\mathcal{B} \times \mathcal{F}^{Y_\zeta}$ -measurable function $\bar{l}: [0, T] \times \Omega \rightarrow \mathcal{C}^{d+1}(\mathbb{R}^d)$ such that w.p.1*

$$(4.10) \quad L_t(\omega) = \int_0^t \bar{l}(s, \omega) ds \quad \text{all } t.$$

Proof. $\mathcal{B} \times \mathcal{F}^{Y_\zeta}$ is generated by the semialgebra of sets $(s, t] \times A$, $A \in \mathcal{F}^{Y_\zeta}$, plus $\{0\} \times A$, $A \in \mathcal{F}^{Y_\zeta}$. Define

$$\mu_N((s, t] \times A) = \int_A [L_t(\omega) - L_s(\omega)] 1_N(\omega) dQ,$$

where 1_N is the characteristic function of

$$C_N = \{\omega : \sup_t |Y_t(\omega)| \leq N\}.$$

Thus there exists n_ω such that if $\omega \in C_N$ and $n > n_\omega$ then $\sup_t |Y_t^n(\omega)| \leq N+1$. Let $\bar{Q} = \lambda \times Q$, $\bar{\mathcal{F}}$ be the completion under \bar{Q} of $\mathcal{B} \times \mathcal{F}^{Y_\zeta}$. For $M \leq \infty$, consider L as an element of $\mathcal{C}([0, T]; \mathcal{C}^{d+1}(S_M))$ with $S_M = \{x : |x| \leq M\}$. Observe that

$$(4.11) \quad \begin{aligned} \|\mu_N((s, t] \times A)\|_M &= \left\| \int_A \lim_n [L_t^n(\omega) - L_s^n(\omega)] 1_N dQ \right\|_M \\ &= \left\| \int_A \lim_n \int_s^t l(\tau, \cdot, Y_\tau^n, u_\tau^n) d\tau 1_N dQ \right\|_M \\ &\leq K_M^N \bar{Q}((s, t] \times A), \end{aligned}$$

since l is locally bounded. It follows that μ_N can be extended to a vector measure

$$\bar{\mathcal{F}} \rightarrow \mathcal{C}^{d+1}(S_M), \quad \mu_N \ll \bar{Q}, \quad \mu_N \text{ of bounded variation.}$$

We shall now apply a theorem of Sion.

First we must establish that for each $B' \in \bar{\mathcal{F}}$ with $\bar{Q}(B') > 0$ there is a subset $B \in \bar{\mathcal{F}}$, $\bar{Q}(B) > 0$, such that

$$\left\{ \frac{\mu_N(A)}{\bar{Q}(A)} : A \in \bar{\mathcal{F}}, A \subset B, \bar{Q}(A) > 0 \right\}$$

is conditionally compact. We use the Ascoli theorem again. Since

$$\phi_m(t) \equiv \sup \left\{ |l(t, x, y, u) - l(t, \bar{x}, y, u)| : |x|, |\bar{x}| \leq M, |y| \leq N, u \in U, |x - \bar{x}| < \frac{1}{m} \right\}$$

converges to 0 a.e., and since l is locally integrable, then $\phi_m \rightarrow 0$ almost uniformly. Let R be a set such that $\lambda([0, T] \setminus R) < \bar{Q}(B')/2$ and such that $\phi_m \rightarrow 0$ uniformly on R so that l is continuous in x on S_M uniformly in $(t, y, u) \in R \times \{y : |y| \leq N\} \times U$. Define $\tilde{l} = l$ for $t \in R$, $\tilde{l} = 0$ for $t \notin R$, so that \tilde{l} is continuous in $x \in S_M$ uniformly in $(t, y, u) \in$

$[0, T] \times \{y: |y| \leq N\} \times U$. With $B = B' \cap (R \times \Omega)$ we have $\bar{Q}(B) > \bar{Q}(B')/2 > 0$ and $l(t, x, Y_t^n, u_t^n) = \tilde{l}(t, x, Y_t^n, u_t^n)$ for all $n, x \in S_M, (t, \omega) \in B$. Define $\tilde{\mu}_N$ from \tilde{l} in the same way as μ_N was defined from l . For $A \subset B, \bar{Q}(A) > 0, \eta > 0$, there exists a set $O = \bigcup_{i=1}^{\infty} O_i \supset A, O_i$ disjoint, $O_i = (s_i, t_i) \times B_i$ such that $\bar{Q}(O \setminus A) \leq \eta \bar{Q}(A)$, hence $\bar{Q}(O) \leq (1 + \eta) \bar{Q}(A)$. Now for $x, \bar{x} \in S_M$, we have, using (4.11),

$$\begin{aligned} \frac{|\mu_N(A)(x) - \mu_N(A)(\bar{x})|}{\bar{Q}(A)} &= \frac{|\tilde{\mu}_N(A)(x) - \tilde{\mu}_N(A)(\bar{x})|}{\bar{Q}(A)} \\ &\leq \frac{|\tilde{\mu}_N(O \setminus A)(x) - \tilde{\mu}_N(O \setminus A)(\bar{x})|}{\bar{Q}(A)} + \frac{|\tilde{\mu}_N(O)(x) - \tilde{\mu}_N(O)(\bar{x})|}{\bar{Q}(A)} \\ &\leq 2K_M^N \frac{\bar{Q}(O \setminus A)}{\bar{Q}(A)} + \frac{\bar{Q}(O)}{\bar{Q}(A)} \frac{|\tilde{\mu}_N(O)(x) - \tilde{\mu}_N(O)(\bar{x})|}{\bar{Q}(O)} \\ &\leq 2K_M^N \eta + \frac{(1 + \eta)(*)}{\bar{Q}(O)} \end{aligned}$$

where

$$(*) = \left| \sum_i \int_{B_i} 1_N \lim_n \int_{s_i}^{t_i} \tilde{l}(t, x, Y_t^n, u_t^n) - \tilde{l}(t, \bar{x}, Y_t^n, u_t^n) dt dQ \right| \leq \eta \bar{Q}(O)$$

for $|x - \bar{x}| < \delta$ by the continuity of \tilde{l} in x , uniformly in (t, y, u) . Thus $\{\mu_N(A)/\bar{Q}(A): A \subset B\}$ is equicontinuous. Since this set is also bounded by K_M^N , it is conditionally compact. Hence by [11, Thms. 4.2, p. 110, 2.8, p. 48 and 6.2, p. 76] there is a Bochner integrable function $\bar{l}_N(t, \omega): [0, T] \times \Omega \rightarrow \mathcal{C}^{d+1}(S_M)$ such that

$$\mu_N(A) = \int_A \bar{l}_N d\bar{Q}.$$

It follows that $1_N L_t = \int_0^t \bar{l}_N(s, \omega) ds$ if $\omega \notin D_N$, a null set. Let $D = \bigcup_{N=1}^{\infty} D_N$. Then for $\omega \in C_N \setminus D, k \geq 0$

$$L_t(\omega) = \int_0^t \bar{l}_{N+k}(s, \omega) ds.$$

Define $\bar{l}^M(t, \omega) = \bar{l}_1(t, \omega)1_1(\omega) + \sum_{N=2}^{\infty} \bar{l}_N(t, \omega)[1_N(\omega) - 1_{N-1}(\omega)]$, so that

$$L_t(\omega) = \int_0^t \bar{l}^M(s, \omega) ds \quad \text{w.p.1.}$$

Next observe that $\sup_{|x| \leq M} |\bar{l}^{M+1}(t, \omega)(x) - \bar{l}^M(t, \omega)(x)| = 0$ a.e.; thus if we define

$$\hat{l}(t, \omega)(x) = \sum_{M=0}^{\infty} \bar{l}^{M+1}(t, \omega)(x) 1_{\{M \leq |x| < M+1\}}(x)$$

then $\hat{l}(t, \omega) \in \mathcal{C}^{d+1}(\mathbb{R}^d)$ a.e. and \hat{l} is $\bar{\mathcal{F}}$ -measurable. Hence there is a $\mathcal{B} \times \mathcal{F}^{Y^c}$ -measurable function \bar{l} such that $\bar{l} = \hat{l}$ a.e. Moreover (4.10) holds. This establishes the lemma.

COROLLARY. If $\bar{L}_t = (\bar{F}_t', \bar{G}_t)'$ then

$$\sup_t \left| \bar{L}_t - \int_0^t \bar{l}(s, \omega)(X_s) ds \right| = 0 \quad \text{w.p.1.}$$

Proof. Fix $\varepsilon > 0$, $\eta > 0$. Since $(X^n, Y^n) \rightarrow (X, Y)$ then there exist $n_0, k < \infty$ such that for all $n > n_0$

$$\Pr \{ \|X^n\| + \|Y^n\| \geq k \} \leq \eta.$$

By Lemma A.4 it follows that there exists ε' depending only on ε and k such that

$$\sup_t \left| \int_0^t l(s, x(s), Y_s^n, u_s^n) - l(s, \bar{x}(s), Y_s^n, u_s^n) ds \right| < \varepsilon,$$

whenever $\|Y^n\|, \|x\|, \|\bar{x}\| \leq k$ and $\|x - \bar{x}\| < 3\varepsilon'$. But now there exist n_1 and δ depending on ε' and η such that for $n > n_1$

$$\Pr \{ \|X^n - X\| \geq \varepsilon' \} \leq \eta, \quad \Pr \left\{ \sup_{|s-t| < \delta} |X(s) - X(t)| \geq \varepsilon' \right\} \leq \eta.$$

If we let $\pi = \{0 = t_0 < t_1 < \dots < t_k = T\}$ be a partition of $[0, T]$ with mesh less than δ , and if for any $x \in \mathcal{C}^d[0, T]$ we set

$$\pi x(t) = x(t_i), \quad t_i \leq t < t_{i+1},$$

then it follows that for $n > n_1$

$$\Pr \{ \|\pi X^n - X^n\| \geq 3\varepsilon' \} \leq 2\eta,$$

and thus

$$\Pr \left\{ \sup_t \left| \int_0^t l(s, \pi X^n(s), Y_s^n, u_s^n) ds - \bar{L}_t^n \right| \geq \varepsilon \right\} \leq 3\eta$$

provided $n > n_2(\varepsilon, \eta) = \max(n_0, n_1)$, and π is a finite partition with mesh less than $\delta(\varepsilon, \eta)$. But,

$$\int_0^t l(s, \pi X^n(s), Y_s^n, u_s^n) ds = \sum_i [L_{t \wedge t_{i+1}}^n(X_{t_i}^n) - L_{t \wedge t_i}^n(X_{t_i}^n)].$$

If we take the limit as $n \rightarrow \infty$, recalling that L^n converges uniformly on the compact set $|x| \leq k$, we have, using the previous lemma,

$$\Pr \left\{ \sup_t \left| \sum_i \int_{t \wedge t_i}^{t \wedge t_{i+1}} \bar{l}(s, \omega)(X_{t_i}) ds - \bar{L}_t \right| \geq \varepsilon \right\} \leq 3\eta.$$

Now let the mesh $\delta \rightarrow 0$ to obtain

$$\Pr \left\{ \sup_t \left| \bar{L}_t - \int_0^t \bar{l}(s, \omega)(X_s) ds \right| \geq \varepsilon \right\} \leq 3\eta.$$

Since ε, η are arbitrary the result follows.

We know now that

$$(4.12) \quad X_t = X_0 + \int_0^t \bar{f}(s, \omega)(X_s) ds + B(t).$$

Let $\mathcal{F}_t^0 \equiv \mathcal{F}_t^{wXYL}$. Then $(w, X, Y, B, \bar{B}, F, G)$ is $\{\mathcal{F}_t^0\}$ -adapted. The next step is to show that

$$\bar{f}(s, \omega)(X_s) = f(s, X_s, Y_s, u_s), \quad \bar{f}_0(s, \omega)(X_s) = f_0(s, X_s, Y_s, u_s).$$

LEMMA 4.4. *There is an $\{\mathcal{F}_t^0\}$ -adapted $\mathcal{B} \times \mathcal{F}^{Y^c}$ -measurable process $u^*: [0, T] \times \Omega \rightarrow U$ such that $\bar{l}(t, \omega) = l(t, \cdot, Y_t, u_t^*)$ a.e. (t, ω) .*

Proof. We shall use the McShane–Warfield implicit function theorem. By Lemma 4.3 \bar{l} is $\mathcal{B} \times \mathcal{F}^{Y_t}$ -measurable. As L is $\{\mathcal{F}_t^0\}$ -adapted, so is \bar{l} . Since $X^n \rightarrow X$ w.p.1., then w.p.1. there exists a compact set $C(\omega) \subset \mathbb{R}^d$ such that $X_t^n(\omega) \in C(\omega)$ for all n, t . The convergence $L^n \rightarrow L$ w.p.1. implies w.p.1.

$$(4.13) \quad \sup_t \sup_{x \in C(\omega)} \left| \int_0^t [l(s, x, Y_s^n, u_s^n) - \bar{l}(s, \omega)(x)] ds \right| \rightarrow 0$$

as $n \rightarrow \infty$. It follows from the local boundedness of l and from (4.13) that $l(\cdot, \cdot, Y_t^n, u_t^n) \rightarrow \bar{l}(\cdot, \omega)(\cdot)$ weakly in $L_1([0, T] \times C(\omega); \mathbb{R}^{d+1})$, so a sequence of convex combinations converges strongly.

Since $\sup_t \{Y^n(t, \omega) : n = 1, 2, \dots, 0 \leq t \leq T\}$ is bounded for each ω , since $\sup_t |Y^n - Y| \rightarrow 0$, and since l is locally bounded and continuous in y uniformly in u , a.e. t, x , then w.p.1

$$\lim_{n \rightarrow \infty} \int_0^T \int_{C(\omega)} |l(t, x, Y_t^n, u_t^n) - l(t, x, Y_t, u_t^n)| dx dt = 0.$$

The above two facts imply that a sequence of convex combinations of the $l(\cdot, \cdot, Y_t, u_t^n)$ converges to $\bar{l}(\cdot, \omega)$ in $L_1([0, T]; L_1(C(\omega); \mathbb{R}^{d+1}))$. For a subsequence this convergence is a.e. t . Since l satisfies the condition C , then each convex combination lies in $l(t, \cdot, Y_t, U)$. Moreover, since for each (t, ω) the function $l : U \rightarrow L_1(C(\omega); \mathbb{R}^{d+1})$ is continuous, then w.p.1, a.e. t , $\bar{l}(t, \omega) \in l(t, \cdot, Y_t(\omega), U)$ as a set in $L_1(C(\omega); \mathbb{R}^{d+1})$, i.e., there exists u in U such that $\bar{l}(t, \omega)(x) = l(t, x, Y_t(\omega), u)$ for almost all x . But the continuity in x of \bar{l} and l implies the result for all x , i.e., w.p.1., a.e. t , $\bar{l}(t, \omega) \in l(t, \cdot, Y_t(\omega), U)$ as a set in $\mathcal{C}^{d+1}(\mathbb{R}^d)$. Since $\Omega = [0, 1] \times \mathcal{X}$ (the probability space in the Skorokhod imbedding is $[0, 1]$) then the set

$$A = \{(t, \omega, l(t, \cdot, Y_t(\omega), u)) : (t, \omega, u) \in [0, T] \times \Omega \times U\}$$

is Borel-measurable by [15, Chapt. III, Thm. 21b]. Hence each (t, ω) section $A_{t\omega} = l(t, \cdot, Y_t(\omega), U)$ is measurable. It follows that

$$\{(t, \omega) : \bar{l}(t, \omega) \in l(t, \cdot, Y_t(\omega), U)\}$$

is measurable, and hence by Fubini's theorem

$$\bar{l}(t, \omega) \in l(t, \cdot, Y_t(\omega), U)$$

a.e. (t, ω) . Hence if we modify \bar{l} on a null set in $\mathcal{B} \times \mathcal{F}$, calling the new function \bar{l}_0 , then $\bar{l}_0(t, \omega) \in l(t, \cdot, Y_t, U)$ for all (t, ω) .

If we write \mathcal{P} for all $\mathcal{B} \times \mathcal{F}^{Y_t}$ -measurable sets such that each t section is \mathcal{F}_t^0 -measurable, and if $\bar{\mathcal{P}}$ is the completion of \mathcal{P} in $([0, T] \times \Omega, \mathcal{B} \times \mathcal{F}, \lambda \times Q)$, then \bar{l}_0 is $\bar{\mathcal{P}}$ -measurable.

The McShane–Warfield implicit function theorem, as extended by Benes [10], [1], guarantees the existence of a function $\bar{u} : [0, T] \times \Omega \rightarrow U$, $\bar{\mathcal{P}}$ -measurable, such that $\bar{l}_0(t, \omega) = l(t, \cdot, Y_t, \bar{u}(t, \omega))$. According to Lemma A.3, there is a \mathcal{P} -measurable u^* such that $u^* = \bar{u}$ a.e.

COROLLARY. u_t^* and $\{Y_s : s > t\}$ are conditionally independent given \mathcal{F}_t^Y ; moreover (u^*, Y) and w are independent.

PROOF. The last assertion is trivial since (u^n, Y^n) , w^n , i.e., (L^n, Y^n) , w^n are independent. To prove the other result we first show that Y is an $\{\mathcal{F}_t^0\}$ martingale. Clearly Y is an $\{\mathcal{F}_t^0\}$ adapted process with continuous paths. For any real, continuous,

bounded function ϕ , $t_i \leq t$, $i = 1, \dots, k$, $0 \leq t \leq t+s \leq T$,

$$\begin{aligned} 0 &= E_Q \phi(w_{t_1}^n, \dots, w_{t_k}^n, X_{t_1}^n, \dots, X_{t_k}^n, Y_{t_1}^n, \dots, Y_{t_k}^n, L_{t_1}^n, \dots, L_{t_k}^n)(Y_{t+s}^n - Y_t^n) \\ &\rightarrow E_Q \phi(w_{t_1}, \dots, L_{t_k})(Y_{t+s} - Y_t), \end{aligned}$$

since $\Phi^n \rightarrow \Phi$ w.p.1, Y^n is an \mathcal{F}_t^n -martingale, $E|Y_{t+s}^n - Y_t^n|^4 \leq Ks^2$, and w^n, X^n, Y^n, u^n are \mathcal{F}_t^n -adapted. Thus

$$E_Q\{Y_{t+s} - Y_t | \mathcal{F}_t^0\} = 0.$$

Similarly

$$\begin{aligned} &E_Q \phi(w_{t_1}^n, \dots, L_{t_k}^n)(Y_{t+s}^n - Y_t^n)(Y_{t+s}^n - Y_t^n)' \\ &= E_Q \phi(w_{t_1}^n, \dots, L_{t_k}^n) \int_t^{t+s} \bar{\sigma}(\tau, Y_\tau^n) \bar{\sigma}(\tau, Y_\tau^n)' d\tau \\ &\rightarrow E_Q \phi(w_{t_1}, \dots, L_{t_k}) \int_t^{t+s} \bar{\sigma}(\tau, Y_\tau) \bar{\sigma}(\tau, Y_\tau)' d\tau \end{aligned}$$

and

$$\rightarrow E_Q \phi(w_{t_1}, \dots, L_{t_k})(Y_{t+s} - Y_t)(Y_{t+s} - Y_t)'.$$

Hence the increasing process of the martingale Y is $\int_0^t \bar{\sigma} \bar{\sigma}' ds$, and a representation theorem of Doob yields an $\{\mathcal{F}_t^0\}$ -Brownian motion β_t such that (cf. [17, (14.47)])

$$Y_t = \int_0^t \bar{\sigma}(s, Y_s) d\beta_s.$$

Since u^* is $\{\mathcal{F}_t^0\}$ -adapted, the result now follows.

LEMMA 4.5. *There is a process β such that (β_s^w) is a standard Wiener process on $(\Omega, \mathcal{F}, Q, \{\mathcal{F}_t^0\})$ and such that*

$$\begin{pmatrix} B_t \\ \bar{B}_t \\ H_t \end{pmatrix} = \int_0^t \begin{pmatrix} \sigma(s, X_s, Y_s) & 0 \\ 0 & \bar{\sigma}(s, Y_s) \\ 0 & h(s, X_s, Y_s)'(\bar{\sigma}(s, Y_s)^{-1})' \end{pmatrix} d \begin{pmatrix} w_s \\ \beta_s \end{pmatrix}.$$

Proof. As in the proof of the previous corollary, it can be shown that

$$M_t = \begin{pmatrix} w_t \\ B_t \\ \bar{B}_t \\ H_t \end{pmatrix}$$

is a sample continuous $\mathcal{F}_t^0 \vee \mathcal{F}_t^H$ -martingale with increasing process $\int_0^t \Gamma_s ds$ where

$$\Gamma \equiv \begin{pmatrix} I & \sigma' & 0 & 0 \\ \sigma & \sigma\sigma' & 0 & 0 \\ 0 & 0 & \bar{\sigma}\bar{\sigma}' & h \\ 0 & 0 & h' & \bar{h}'\bar{h} \end{pmatrix} = \begin{pmatrix} I & 0 \\ \sigma & 0 \\ 0 & \bar{\sigma} \\ 0 & \bar{h}' \end{pmatrix} \begin{pmatrix} I & \sigma' & 0 & 0 \\ 0 & 0 & \bar{\sigma}' & \bar{h} \end{pmatrix},$$

where $\bar{h} = \bar{\sigma}^{-1}h$. Then [17, (14.47)] implies that there is a Brownian motion $(\bar{\beta})$ such that

$$M_t = \int_0^t \begin{pmatrix} I & 0 \\ \sigma & 0 \\ 0 & \bar{\sigma} \\ 0 & \bar{h}' \end{pmatrix} d\begin{pmatrix} \bar{\beta} \\ \beta \end{pmatrix}.$$

We observe that $w = \bar{\beta}$, $\int \bar{\sigma}^{-1} dY = \beta$ so that $(\bar{\beta}) = (\beta)$ is $\{\mathcal{F}_t^0\}$ adapted and the result follows.

We can now conclude that

$$(4.14) \quad X_t = X_0 + \int_0^t f(s, X_s, Y_s, u_s^*) ds + \int_0^t \sigma(s, X_s, Y_s) dw_s,$$

$$(4.15) \quad Y_t = \int_0^t \bar{\sigma}(s, Y_s) d\beta_s,$$

$$\begin{aligned} Z_t &\equiv \int_0^t (\bar{\sigma}(s, Y_s)^{-1} h(s, X_s, Y_s))' d\beta_s - \frac{1}{2} \int_0^t |\bar{\sigma}(s, Y_s)^{-1} h(s, X_s, Y_s)|^2 ds \\ &= \lim_n Z_t^n \end{aligned}$$

and w.p.1

$$\begin{aligned} e^{Z_T^n} \left[\int_0^T f_0(t, X_t^n, Y_t^n, u^n(t, Y_t^n)) dt + g(X_T^n, Y_T^n) \right] \\ \rightarrow e^{Z_T} \left[\int_0^T f_0(t, X_t, Y_t, u_t^*) dt + g(X_T, Y_T) \right]. \end{aligned}$$

(H₃) and (H₄) imply that $\{e^{Z_T^n}\}$ are bounded in L_p for some $p > 1$, (H₂), (H₅) imply the same for the term in the square bracket above for all $p < \infty$, so that $J[u^n] \rightarrow J[u^*]$. But also $J[u^n] \rightarrow J^*$. Thus u^* is optimal. Note that we can now use Z_T to define P by

$$\frac{dP}{dQ} = \exp Z_T,$$

so that by Girsanov's theorem, (1.1), (1.2) are satisfied on (Ω, \mathcal{F}, P) with $\bar{w}_t = \beta_t - \int_0^t \bar{\sigma}^{-1} h ds$.

Finally we verify that u^* satisfies part (v) of the definition of randomized control. Since $\bar{\mathcal{F}}^{Yu^*} = \bar{\mathcal{F}}^{YL}$ ($\bar{\cdot}$ denotes completion in (Ω, \mathcal{F}, P)), then by Lemma 4.2 (iii), it follows that for any bounded, measureable functions g of (X, Y, w, β) , and h of ζ , we have

$$\begin{aligned} E_P\{g(X, Y, w, \beta)h(\zeta)|\bar{\mathcal{F}}^{Yu^*}\} &= \frac{E_Q\{e^Z gh|\bar{\mathcal{F}}^{Yu^*}\}}{E_Q\{e^Z|\bar{\mathcal{F}}^{Yu^*}\}} \\ &= E_Q\{h|\bar{\mathcal{F}}^{Yu^*}\} \frac{E_Q\{e^Z g|\bar{\mathcal{F}}^{Yu^*}\}}{E_Q\{e^Z|\bar{\mathcal{F}}^{Yu^*}\}} \\ &= E_Q\{h|\bar{\mathcal{F}}^{Yu^*}\} E_P\{g|\bar{\mathcal{F}}^{Yu^*}\}, \end{aligned}$$

since Z is (X, Y, w, β) -measurable and uniformly integrable. Setting $g = 1$ we find

$$E_P\{h(\zeta)|\bar{\mathcal{F}}^{Yu^*}\} = E_Q\{h(\zeta)|\bar{\mathcal{F}}^{Yu^*}\}.$$

Upon substituting this in the above, we find that the conditional independence of (X, Y, w, β) and ζ is preserved under the Girsanov transformation. Since \bar{w} is a function of (X, Y, w, β) the result also holds for (X, Y, w, \bar{w}) .

Remark 4.1. We could allow $f, f_0, h, \sigma, \bar{\sigma}$ to depend, at time t , not just on y_t but rather on $\{y_s, s \leq t\}$, e.g., $f: [0, T] \times \mathbb{R}^d \times \mathcal{C}^m[0, T] \times U \rightarrow \mathbb{R}^d$ such that $f(t, x, \cdot, u)$ is \mathcal{G}_t -measurable. Now in (H_2) , (H_4) , (H_5) we replace $|y|$ by $\sup_{s < t} |y_s|$.

Remark 4.2. From property (iii) of admissible controls and from the conditional independence of w and ζ , it follows that (w, β) and ζ are independent since $\mathcal{F}^\beta = \mathcal{F}^Y$. Thus under Q_z , the conditional probability given $\zeta = z$, we have

$$dX_t = f(t, X_t, Y_t, u_t^*) dt + \sigma(t, X_t, Y_t) dW_t \quad (\text{Brownian motion}),$$

$$dY_t = \bar{\sigma}(t, Y_t) dW_t \quad (\text{Brownian motion}).$$

Now $u^* = u^*(t, Y_t)$ since $\zeta = z$ is fixed. If we knew that for fixed z , u^* were strictly admissible, then one of these z would give an optimal control in \mathcal{U}^s . Unfortunately, although Y is a martingale under Q_z , we don't know that it is a martingale relative to $\{\mathcal{F}_t^{Y^L}\}$, so that we cannot use the proof given above that u^* is independent of future Y increments under Q_z . Hence u^* is *not* necessarily strictly admissible.

The difficulty is the same as the one which implies that the stochastic differential equation

$$(4.16) \quad d\xi = \phi(t, \xi) dt + dw$$

may have a weak but no strong solution.

In the terminology of Ershov, the Girsanov transformation gives a nonanticipative solution, i.e., a solution such that for each t $\{w_s: 0 \leq s \leq T\}$ and $\{\xi_s: 0 \leq s \leq t\}$ are conditionally independent given \mathcal{F}_t^w . However, as the Tsirel'son example shows, the solution may not be extreme, i.e., $\mathcal{F}_t^\xi \neq \mathcal{F}_t^w$. Note that Ershov has shown that a solution is nonanticipative and extreme if and only if it is causal, i.e., $\mathcal{F}_t^\xi = \mathcal{F}_t^w$. On the other hand, he has also shown [14], that an extreme solution exists, but of course not necessarily nonanticipative. In our context u plays the role of ξ and Y that of w . Under Q , u^* is nonanticipative; cf. property (ii) of admissible controls. Moreover, Q is an average of the Q_z 's, under each of which u^* is an extreme "solution", i.e., $\mathcal{F}^{Yu} = \mathcal{F}^Y \pmod{Q_z}$, but not necessarily a causal solution, i.e., $\mathcal{F}_t^{Yu} \neq \mathcal{F}_t^Y \pmod{Q_z}$, i.e., not necessarily nonanticipative.

Remark 4.3. In (H_2) , (H_4) the conditions

$$|f| + |\bar{\sigma}^{-1}h| \leq K(1 + |x| + |y|), \quad |\sigma| \leq K$$

can be replaced by

$$|\sigma| + |f| \leq K(1 + |x| + |y|^r), \quad |\bar{\sigma}^{-1}h| \leq K(1 + |y|).$$

We still have

$$\sup_u E \sup_t (|X_t|^k + |Y_t|^k) < \infty, \quad E \exp(\pm Z_T^n) = 1, \quad \{\exp Z_T^n\} \text{ uniformly integrable.}$$

Remark 4.4. σ can be allowed to depend on u . In the proofs we simply add the process

$$E_t^n = \int_0^t \sigma(s, \cdot, Y_s^n, u_s^n) \sigma(s, \cdot, Y_s^n, u_s^n)' ds$$

to Φ^n and to L^n . Now

$$l = (f, f_0 e_1, \sigma \sigma')$$

is required to satisfy C , where $e'_1 = (1, 0, \dots, 0)$.

5. Extensions. Let us now indicate how a slightly different problem can be treated, (cf. [16], [5]).

THEOREM 5.1. *Assume (H_1) – (H_6) . If f is affine in u for each (t, x, y) , if U is convex, and if $f_0(t, x, y, \cdot)$ is convex on U for each (t, x, y) , then there exists $u^0 \in \mathcal{U}$ such that*

$$J[u^0] = \inf \{J[u] : u \in \mathcal{U}\}.$$

Proof. Let $\{u^n\} \subset \mathcal{U}$ be a minimizing sequence. Since U is convex, compact, then $L_2([0, T]; U)$ with the weak topology is a complete, separable, compact metric space, so we can add u^n to Φ^n . We still have tightness and, after the Skorokhod embedding, (4.4)–(4.8) hold for the $\hat{\cdot}$ process.

We now replace L_t^0 by u^0 and \mathcal{C}^2 by $L_2([0, T]; U)$ in Lemma 4.2, so u^0 can be taken as \mathcal{F}^{Y^c} -measurable. Moreover since $u^n \rightarrow u^0$ weakly in L_2 and f is affine, then $f(t, X^n, Y^n, u^n) \rightarrow f(t, X^0, Y^0, u^0)$. This replaces Lemmas 4.3 and 4.4. The corollary to Lemma 4.4 as well as Lemma 4.5 still go through.

Finally, since $u^n \rightarrow u^0$ weakly in $L_2[0, T]$, w.p.1, then $u^n \rightarrow u^0$ weakly in $L_2([0, T] \times \Omega)$ and by [4, v.7.43] there exists a sequence $\{v^n\}$, $v^n \in \text{co}\{u^n, u^{n+1}, u^{n+2}, \dots\}$, such that $v^n \rightarrow u^0$ strongly in $L_2([0, T] \times \Omega)$. By the convexity of f_0 we have that

$$f_0(t, x, y, v^n) \leq \sum_{k \geq n} \alpha_k^n f_0(t, x, y, u^k),$$

where $v^n = \sum_{k \geq n} \alpha_k^n u^k$. But $(X^n, Y^n, Z^n) \rightarrow (X^0, Y^0, Z^0)$ w.p.1, hence in probability, so (H_2) implies that for any $\varepsilon > 0$ there exists n_0 and a set A with $Q(A) < \varepsilon$ such that for $n > n_0$, $\omega \notin A$,

$$\begin{aligned} e^{Z_T^n} \left[\int_0^T f_0(t, X_t^n, Y_t^n, v_t^n) dt + g(X_T^n, Y_T^n) \right] \\ \leq e^{Z_T^n} \sum_{k \geq n} \alpha_k^n \left[\int_0^T f_0(t, X_t^n, Y_t^n, u_t^k) dt + g(X_T^n, Y_T^n) \right] \\ \leq \sum_{k \geq n} \alpha_k^n e^{Z_T^k} \left[\int_0^T f_0(t, X_t^k, Y_t^k, u_t^k) dt + g(X_T^k, Y_T^k) \right] + \varepsilon. \end{aligned}$$

Since $J[u_n] \geq J[u_k]$ for $n \leq k$ it follows upon taking expectations and limits $n \rightarrow \infty$, $\varepsilon \rightarrow 0$, that $J[u^0] \leq \inf_{\mathcal{U}} J[u]$. The proof is complete.

We remark that Theorem 3.1 states that $J^* = \inf \{J[u] : u \in \mathcal{U}^s\}$ is attained by a randomized control $u \in \mathcal{U}$, whereas for the affine-convex case Theorem 5.1 states that $\inf \{J[u] : u \in \mathcal{U}\}$ is attained. In fact if U is convex in addition to the assumptions of Theorem 3.1, then we can add $u^n \in L_2([0, T]; U)$ as an extra component to Φ^n , and the proof of Theorem 3.1 continues to hold. We obtain the following theorem.

THEOREM 5.2. *Assume (H_1) – (H_6) and assume U is convex. If l satisfies condition C then $\inf \{J[u] : u \in \mathcal{U}\}$ is attained.*

The point is that we need $\{u^n\}$ to be tight so that, after the Skorokhod imbedding, (4.6) holds for the $\hat{\cdot}$ process. It follows that $u^n \rightarrow u^0$ weakly. In the linear theory this

u^0 is optimal since $\int f(u^n) \rightarrow \int f(u^0)$, but in the nonlinear theory we ignore u^0 and construct the optimal u^* via measurable selection, cf. Lemma 4.4.

If we have $\inf \{J[u]: u \in \mathcal{U}\} = \inf \{J[u]: u \in \mathcal{U}^s\}$ then of course Theorem 3.1 implies that $\inf \{J[u]: u \in \mathcal{U}\}$ is attained without requiring U to be convex. Following [5] we have the following proposition.

PROPOSITION 5.3. *Assume (H₁)–(H₆) and assume that f, σ are Lipschitz continuous in x uniformly in (y, u) . Then*

$$\inf \{J[u]: u \in \mathcal{U}\} = \inf \{J[u]: u \in \mathcal{U}^s\}.$$

Proof. For $u \in \mathcal{U}$, $\varepsilon > 0$, we can define $u_m \in \mathcal{U}$ such that $u_m(t)$ is constant on $[t_i, t_{i+1})$, $t_i = \tilde{t} + iT/m$, $i = 0, \pm 1, \pm 2, \dots$ and such that $\|u_m - u\|_2 < \varepsilon$. Here \tilde{t} depends on u but not m , and u_m is constructed as in [18, pp. 94–95].

Now suppose that $V_n \rightarrow V \in \mathcal{U}$ strongly in L_2 . Then using the Lipschitz condition we can define x_n such that

$$\begin{aligned} dx_n &= f(t, x_n, y, V_n) dt + \sigma(t, x_n, y) dw, \\ dx &= f(t, x, y, V) dt + \sigma(t, x, y) dw, \\ dy &= \bar{\sigma}(t, y) d\beta. \end{aligned}$$

For a subsequence, denoted again by x_n , we have $x_n \rightarrow x$ w.p.1, uniformly in t , cf. [6, Chap. 2., Thm. 2]. Thus $J[V_n] \rightarrow J[V]$, i.e., J is continuous in the norm topology.

Since $\mathcal{U}^s \subset \mathcal{U}$, the proof follows if we establish that for any $\varepsilon > 0$, u_m , there exists $v_m \in \mathcal{U}^s$ such that $J[v_m] \leq J[u_m] + \varepsilon$.

Let $\mathcal{U}_m(\tilde{t})$ be those elements in \mathcal{U} which have sample paths constant on $[t_i, t_{i+1})$, $i = 0, \pm 1, \pm 2, \dots$ so $u_m \in \mathcal{U}_m(\tilde{t})$, and let $\mathcal{U}_m^s = \mathcal{U}_m(\tilde{t}) \cap \mathcal{U}^s$. It can be shown, by adapting the proof by induction in [5, Appendix] that for any bounded continuous function $\psi: \mathcal{C}^m[0, T] \times U^m \rightarrow \mathbb{R}$, $\varepsilon > 0$, and $u \in \mathcal{U}_m(\tilde{t}) \cong U^m$ there exists $V \in \mathcal{U}_m^s$ such that

$$E\psi(Y, V) \leq E\psi(Y, u) + \varepsilon.$$

Use is made here of the Lipschitz continuity of $\bar{\sigma}$ and of property (ii) of the definition of admissible control.

If

$$\phi(Y, u) = E \left\{ e^{Z_T} \left[\int_0^T f_0(t, X_t, Y_t, u) dt + g(X_T, Y_T) \right] \middle| \mathcal{F}^{Yu} \right\}$$

then $J[u] = E\phi(Y, u)$. As above, if $(Y_n, u_n) \rightarrow (Y, u)$, then $x_n \rightarrow x$ and so ϕ is continuous. Finally, since there exists $p > 1$, $K < \infty$ such that $E|\phi(Y, u)|^p \leq K$ for all u , then for $\varepsilon > 0$ there exists M such that

$$\sup_{u \in \mathcal{U}} E|\psi(Y, u) - \phi(Y, u)| < \frac{\varepsilon}{3},$$

where

$$\psi(Y, u) = \begin{cases} \phi(Y, u) & \text{if } |\phi(Y, u)| \leq M, \\ M \frac{\phi(Y, u)}{|\phi(Y, u)|} & \text{otherwise.} \end{cases}$$

Now there exists V_m in \mathcal{U}_m^s such that

$$J[V_m] - J[u_m] = J[V_M] - E\psi(Y, V_m) + E\psi(Y, V_m) - E\psi(Y, u_m) + E\psi(Y, u_m) - J[u_m] \\ \leq \varepsilon,$$

so the proof is complete.

We remark that if σ is invertible then probably the Lipschitz continuity of f , σ is not required if x_m is defined using the Girsanov approach.

Appendix.

LEMMA A.1. If $\tilde{\Omega} = \{(Y^0(\omega), L^0(\omega)) : \omega \in \Omega, (Y^n(\omega), L^n(\omega)) \rightarrow (Y^0(\omega), L^0(\omega))\}$,

$$\gamma(y, \lambda) = y,$$

$K \subset \mathcal{C}_1$ is compact, then $\gamma^{-1}(K)$ is conditionally compact.

Proof. For $y \in K$, there exists $k_0 < \infty$ such that $\|y\| = \sup_t |y_t| \leq k_0 - 1$. Let $A_n = \{\omega : \|Y^n(\omega)\| \leq k_0\}$. Then $(Y^0)^{-1}(K) \subset \varliminf A_n$. Set $K_0 = \{L^0(\omega) : (Y^0(\omega), L^0(\omega)) = \lim (Y^n(\omega), L^n(\omega)), \omega \in \varliminf A_n\}$. Then $\gamma^{-1}(K) = \{(y, \lambda) \in \tilde{\Omega} : y \in K\} \subset K \times K_0$, so that the result follows if K_0 is conditionally compact.

We shall now show that the elements of K_0 are equicontinuous. Fix $\eta > 0$. If $\lambda \in K_0$, then $\lambda = L^0(\omega)$ for some ω and

$$\rho(\lambda_t, \lambda_s) \leq \rho(L_t^0, L_t^n) + \rho(L_t^n, L_s^n) + \rho(L_s^n, L_s^0) \leq \frac{\eta}{2} + \rho(L_t^n, L_s^n)$$

if $n > n_0(\eta, \omega)$ since $\|L^n(\omega) - L^0(\omega)\| \rightarrow 0$ for each ω . Thus for any $\delta > 0$

$$\sup_{|t-s| < \delta} \rho(\lambda_t, \lambda_s) \leq \frac{\eta}{2} + \lim_n \sup_{|t-s| < \delta} \rho(L_t^n, L_s^n).$$

But as in part 1 of the proof of Lemma 4.1,

$$\sup_{|t-s| < \delta} \rho(L_t^n, L_s^n) < \frac{\eta}{4}$$

if $\omega \in A_n$, i.e., $\|Y^n\| \leq k_0$, and $\delta = \delta(\eta, k_0)$. Thus for $\omega \in \bigcap_{n \geq m} A_n$ (some $m < \infty$)

$$\sup_{|t-s| < \delta} \rho(\lambda_t, \lambda_s) \leq \frac{3\eta}{4} < \eta$$

if $\delta = \delta(\eta, k_0)$. This establishes the equicontinuity.

Now we shall prove the conditional compactness of $\{\lambda_t : \lambda \in K_0\}$ for fixed t . For each $N < \infty$, consider $\{\lambda_t : \lambda \in K_0\} \subset \mathcal{C}^{d+1}(\{x : |x| \leq N\})$. For $n > n_1(\omega)$

$$\sup_{|x| \leq N} |\lambda_t(x) - L_t^n(\omega, x)| < 1,$$

and by (H₂)

$$\|L_t^n(\omega)\| = \sup_{|x| \leq N} \left| \int_0^t l(s, x, Y_s^n(\omega), u_s^n) ds \right| \leq Kt(1 + N^r + k_0^r)$$

if $\omega \in A_n$. Hence for $\omega \in \bigcap_{n \geq m} A_n$

$$\|\lambda_t\| \leq 1 + Kt(1 + N^r + k_0^r),$$

and so $\{\lambda_t: \lambda \in K_0\}$ is bounded. Moreover for fixed η

$$\begin{aligned} |\lambda_t(x) - \lambda_t(\bar{x})| &\leq |L_t^0(\omega, x) - L_t^n(\omega, x)| + |L_t^n(\omega, x) - L_t^n(\omega, \bar{x})| + |L_t^n(\omega, \bar{x}) - L_t^0(\omega, \bar{x})| \\ &\leq \frac{\eta}{2} + |L_t^n(\omega, \bar{x}) - L_t^n(\omega, \bar{x})| \end{aligned}$$

if $n > n_0(\eta, \omega)$. By Lemma 4 it follows that there exists $\delta(N, k_0, t)$ such that $|L_t^n(\omega, x) - L_t^n(\omega, \bar{x})| < \eta/4$ if $|x - \bar{x}| < \delta(N, k_0, t)$ and $\omega \in A_n$.

Hence for $\omega \in \bigcap_{n \geq m} A_n$

$$|\lambda_t(x) - \lambda_t(\bar{x})| \leq \frac{3\eta}{4} < \eta$$

if $|x - \bar{x}| < \delta(N, k_0, t)$. But again $\lim A_n = \bigcup_m \bigcap_{n \geq m} A_n$ so the equicontinuity follows. Thus by Ascoli's theorem $\{\lambda_t: \lambda \in K_0\}$ is conditionally compact, hence the lemma is established.

LEMMA A.2. If $1_{A \times \mathcal{H}} = \hat{P}(A \times \mathcal{H} | \mathcal{G} \times \{\phi, \mathcal{C}_2\} \times \mathcal{B}_{\mathcal{H}})$ a.e. \hat{P} for A in $\tilde{\mathcal{G}}$, then there exists $\tilde{L}, \mathcal{F}^Y \times \mathcal{B}_{\mathcal{H}}$ -measurable, such that $\hat{L} = \tilde{L}$ a.e. (\hat{Q}) .

Proof. If we set $\psi(\omega, z) = (\hat{Y}(\omega, z), \hat{L}(\omega, z), \zeta(\omega, z))$ then for A in $\mathcal{F}^{\hat{Y}\hat{L}}$ there is \tilde{A} in \mathcal{G} such that $A = \psi^{-1}(\tilde{A} \times \mathcal{H})$. Also for B in $\mathcal{F}^{\hat{Y}\hat{L}\zeta}$, $\hat{Q}(B) = \hat{P}(\psi(B))$. Thus for A in $\mathcal{F}^{\hat{Y}\hat{L}}$, the hypothesis gives

$$1_A = \hat{Q}(A | \mathcal{F}^{\hat{Y}\zeta}) \quad \text{a.e. } \hat{Q}.$$

Since $\hat{L}: (\hat{\Omega}, \mathcal{F}^{\hat{Y}\hat{L}}) \rightarrow (\mathcal{C}_2, \mathcal{G}_2)$ is measurable, there exist step functions h^k such that $h^k \rightarrow \hat{L}$ a.e., [11, II, Thms. 2.6, 2.7]. If

$$h^k = \sum_{i=1}^{\infty} h_i^k 1_{A_i^k}$$

and $B_i^k = \{(\omega, z): \hat{Q}(A_i^k | \mathcal{F}^{\hat{Y}\zeta}) = 1\} \setminus \bigcup_{j < i} B_j^k$ then $\{A_i^k\}_{i=1}^{\infty}$ are disjoint subsets of $\mathcal{F}^{\hat{Y}\hat{L}}$, $\{B_i^k\}_{i=1}^{\infty}$ are disjoint subsets of $\mathcal{F}^{\hat{Y}\zeta}$, and

$$A_i^k \triangle B_i^k \subset N_i^k \in \mathcal{F}^{\hat{Y}\hat{L}\zeta}$$

$\hat{Q}(N_i^k) = 0$. Define

$$\bar{h}^k = \sum_{i=1}^{\infty} h_i^k 1_{B_i^k}.$$

Then \bar{h}^k is $\mathcal{F}^{\hat{Y}\zeta}$ -measurable and $\bar{h}^k = h^k$ a.e. It follows that $\bar{h}^k \rightarrow \hat{L}$ a.e. Let \hat{A} be the convergence set of $\{\bar{h}^k\}$. Then $\hat{A} \in \mathcal{F}^{\hat{Y}\zeta}$ and it follows that $\hat{Q}(\hat{A}) = 1$. Define

$$\tilde{h}^k(\omega, z) = \begin{cases} \bar{h}^k(\omega, z) & \text{if } (\omega, z) \in \hat{A}, \\ 0 & \text{otherwise.} \end{cases}$$

Then \tilde{h}^k converges pointwise to a limit, called \tilde{L} , which is $\mathcal{F}^{\hat{Y}\zeta}$ -measurable by [11, II, Thm. 3.2]. Moreover

$$\tilde{L} = \lim_k \tilde{h}^k \stackrel{\text{a.e.}}{=} \lim_k \bar{h}^k \stackrel{\text{a.e.}}{=} \lim_k h^k \stackrel{\text{a.e.}}{=} \hat{L}.$$

LEMMA A.3. If \bar{u} is $\bar{\mathcal{P}}$ -measurable then there exists u^* which is \mathcal{P} -measurable such that

$$\bar{u} = u^* \quad \text{a.e.}$$

Proof. Since $A \in \bar{\mathcal{P}}$ only if there is $B \in \mathcal{P}$, $N \in \mathcal{B} \times \mathcal{F}$, such that $(\lambda \times Q)(N) = 0$, $A \triangle B \subset N$, then the result follows by the second half of the proof of Lemma A.2.

LEMMA A.4. If $u: [0, T] \rightarrow U$ -measurable, y in $\mathcal{C}^m[0, T] \cap \{\|y\| \leq k_0\}$ then the mapping $x(\cdot) \rightarrow \int_0^\cdot l(s, x(s), y(s), u(s)) ds$ of $\mathcal{C}^{d+1}[0, T] \rightarrow \mathcal{C}^{d+1}[0, T]$ is uniformly continuous on $\{x: \|x\| \leq N\}$, uniformly in y, u .

Proof. Fix $\varepsilon > 0$. Let

$$B_m = \left\{ s: \sup \left\{ |l(s, x, y, u) - l(s, \bar{x}, y, u)|: |x - \bar{x}| < \frac{1}{m}, \right. \right. \\ \left. \left. |x|, |\bar{x}| \leq N, |y| \leq k_0, u \in U \right\} < \frac{\varepsilon}{2T} \right\}.$$

Since l is uniformly continuous in (x, y, u) on compact sets, a.e. t , then

$$\lim_{m \rightarrow \infty} B_m = [0, T] \setminus (\text{null set}).$$

If λ denotes Lebesgue measure, it follows that

$$T = \lambda(\lim B_m) \leq \lim \lambda(B_m) \leq T.$$

By (H_2) , l is bounded on $[0, T] \times \{x: \|x\| \leq N\} \times \{y: \|y\| \leq k_0\} \times U$ by say $\|l\|$. Choose m so large that

$$\lambda(B_m) \geq T - \frac{\varepsilon}{(4\|l\|)}.$$

Now if $\|x - \bar{x}\| < 1/m$, then

$$\begin{aligned} & \sup_t \left| \int_0^t l(s, x(s), y(s), u(s)) - l(s, \bar{x}(s), y(s), u(s)) ds \right| \\ & \leq \int_0^T |l(s, x(s), y(s), u(s)) - l(s, \bar{x}(s), y(s), u(s))| ds \\ & \leq \int_{B_m} + \int_{[0, T] \setminus B_m} |l(s, x, y, u) - l(s, \bar{x}, y, u)| ds \\ & < \int_{B_m} \frac{\varepsilon}{2T} ds + \int_{[0, T] \setminus B_m} 2\|l\| ds \\ & < \varepsilon. \end{aligned}$$

REFERENCES

- [1] V. E. BENEŠ, *Existence of optimal stochastic controls*, this Journal, 9 (1971), pp. 446–472.
- [2] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.
- [3] N. CHRISTOPEIT, *Existence of optimal stochastic controls under partial observation*, Z. Wahrsch. Verw. Gebiete, 51 (1980), pp. 201–213.
- [4] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part I*, Interscience, New York, 1966.
- [5] W. H. FLEMING AND E. PARDOUX, *Existence of optimal controls for partially observed diffusions*, this Journal, 20 (1982), pp. 261–283.
- [6] I. I. GIKHMAN AND A. W. SKOROKHOD, *Stochastic Differential Equations*, Springer-Verlag, New York, 1972.
- [7] I. V. GIRSANOV, *On transforming a certain class of stochastic processes by absolutely continuous substitution of measures*, Theory Prob. Appl., 5 (1960), pp. 285–301.

- [8] M. KOHLMANN, *Existence of optimal controls for a partially observed semimartingale control problem*, preprint, Bonn University.
- [9] H. J. KUSHNER, *Existence results for optimal stochastic controls*, J. Optim. Theory Appl., 15 (1975), pp. 347–359.
- [10] E. J. MCSHANE AND R. B. WARFIELD, *On Filippov's implicit function lemma*, Proc. Amer. Math. Soc., 18 (1967), pp. 41–47.
- [11] M. SION, *A Theory of Semigroup Valued Measures*, Lecture Notes in Mathematics 35, Springer-Verlag, New York, 1973.
- [12] A. V. SKOROKHOD, *Studies in the Theory of Random Processes*, Addison-Wesley, Reading, MA, 1965.
- [13] H. L. ROYDEN, *Real Analysis*, Macmillan, New York, 1968.
- [14] M. P. ERSHOV, *The Choquet theorem and stochastic equations*, Analysis Matematica, 1 (1975), pp. 259–271.
- [15] C. DELLACHÉRIE AND P.-A. MEYER, *Probabilités et potentiel*, Hermann, Paris, 1975.
- [16] U. G. HAUSSMANN, *Existence of partially observable stochastic optimal controls*, Proc. of IFIP Conference on Stochastic Differential Systems, Hungary 1980, Lecture Notes in Mathematics, Springer-Verlag, New York.
- [17] J. JACOD, *Calcul stochastique et problèmes de martingales*, Lecture Notes in Mathematics 714, Springer-Verlag, New York.
- [18] R. S. LIPTSER AND A. N. SHIRYAYEV, *Statistics of Random Processes*, vol. I, Springer-Verlag, New York, 1977.
- [19] J. M. BISMUT, *Un problème de contrôle stochastique avec observation partielle*, Z. Wahrsch. Verw. Geb., 49 (1979), pp. 63–95.
- [20] H. J. KUSHNER, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Academic Press, New York, 1977.

INTERNAL AND EXTERNAL STABILITY OF LINEAR TIME-VARYING SYSTEMS*

BRIAN D. O. ANDERSON†

Abstract. Linear, finite-dimensional, time-varying systems are studied. State variable representations of systems with a bounded-input, bounded-output stability property which are uniformly stabilizable and detectable are shown to have their associated homogeneous state-variable systems exponentially stable.

1. Introduction. This paper connects input–output and Lyapunov stability results for finite-dimensional linear systems.

To put the result in perspective, recall that for a finite-dimensional time-invariant linear system in state-variable form: first, if the homogeneous state variable equation is asymptotically stable, the system is bounded-input, bounded-output stable; second, if the system is bounded-input, bounded-output stable and its state variable realization is controllable and observable, then the homogeneous state-variable equation is asymptotically stable. Of course, in the time-invariant case, asymptotic stability is equivalent to exponential stability.

Our aim here is to generalize these results, first by considering time-varying systems, and, second, by relaxing controllability and observability to a form of stabilizability and detectability.

As background, we note first [1], [2] which consider connections between input–output and Lyapunov stability for time-varying systems when the input and output vectors have the same dimension as, and are linked in a uniformly nonsingular way to, the state vector. Reference [3] discusses results for time-varying systems with state-variable realizations in phase-variable form, while [4], [5], [6] provide a major generalization by showing for time-varying systems that input–output stability together with uniform controllability and observability imply exponential stability of the associated homogeneous state variable equation. (In time-varying systems, it is possible to have asymptotic, but nonexponential, stability.)

The advance on [4]–[6] provided in this paper is the weakening of the uniform controllability and observability conditions to uniform stabilizability and detectability. For one of the first uses of these concepts for time-varying systems, see [7] and for a general treatment, see [8]. The result of this paper is of course trivial in the time-invariant case. But in the time-varying case, one cannot, via a Lyapunov basis change, break the system up into a uniformly controllable and observable part which is input–output equivalent to the original system, and an exponentially stable part—hence the nontrivial nature of the problem.

Background information is reviewed in § 2. In § 3, the connection between bounded-input, bounded-output behavior and bounded-input, bounded-state behavior is examined. The main result is proved in § 4. We elect to present the results in discrete time, thus simplifying application of the results of [8], which has results expressed in discrete time. Doubtless with minor adjustment, the results are valid for continuous time. Interestingly, the proof is not just a minor adjustment of that applying in the controllable/observable situation.

* Received by the editors November 11, 1980, and in revised form June 8, 1981. This work was supported by the Australian Research Grants Committee.

† Department of Electrical and Computer Engineering, University of Newcastle, New South Wales, 2308, Australia.

2. Detectability, stabilizability and BIBO behavior. Consider the linear finite-dimensional system

$$(2.1a) \quad x_{k+1} = F_k x_k + G_k u_k,$$

$$(2.1b) \quad y_k = H'_k x_k,$$

where $x_k \in R^n$ is the state, $u_k \in R^m$ is the input, $y_k \in R^p$ is the output, and F_k, G_k, H_k are matrices of appropriate dimension. The state transition matrix is denoted $\phi_{k,l}$ for $k \geq l$, where $\phi_{k,k} = I$, $\phi_{k+1,k} = F_k$ and $\phi_{k,l} = \phi_{k,k-1} \phi_{k-1,l}$.

Standing Assumption. The sequences F_k, G_k and H_k are bounded.

We now define the concepts of uniform detectability and stabilizability [7], [8].

DEFINITION 2.1. The pair $[F_k, H_k]$ (regarded as a sequence indexed by k) is *uniformly detectable* if there exist integers $s \geq t \geq 0$ and constants d, b with $0 \leq d < 1$, $0 < b < \infty$ such that, whenever

$$(2.2) \quad \|\phi_{k+t,k} \xi\| \geq d \|\xi\|$$

for some ξ and k , then

$$(2.3) \quad \xi' M_{k+s,k} \xi \geq b \xi' \xi,$$

where

$$(2.4) \quad M_{k+s,k} = \sum_{i=k}^{k+s} \phi'_{i,k} H_i H'_i \phi_{i,k}.$$

(The idea is that when a zero-input trajectory starting at $x_k = \xi$ fails to converge much towards the origin (see (2.2)), then x_k should be observable to a minimum level (see (2.3)).)

DEFINITION 2.2. The pair $[F_k, G_k]$ (regarded as a sequence indexed by k) is *uniformly stabilizable* if there exist integers $s \geq t \geq 0$ and constants d, b with $0 \leq d < 1$, $0 < b < \infty$, such that, whenever

$$(2.5) \quad \|\phi_{k+1,k+1-t} \xi\| \geq d \|\xi\|$$

for some ξ, k , then

$$(2.6) \quad \xi' Y_{k,k-s} \xi \geq b \xi' \xi,$$

where

$$(2.7) \quad Y_{k,k-s} = \sum_{i=k-s}^k \phi_{k+1,i+1} G_i G'_i \phi'_{k+1,i+1}.$$

We shall need several results established in [7]. The most important are stated in the following lemma.

LEMMA 2.1. *With the definitions*

$$(2.8) \quad \hat{F}_k = F'_{-k}, \quad \hat{H}_k = G_{-k},$$

$[\hat{F}_k, \hat{H}_k]$ is uniformly detectable if and only if $[F_k, G_k]$ is uniformly stabilizable, and $x_{k+1} = F_k x_k$ is exponentially stable if and only if $\hat{x}_{k+1} = \hat{F}_k \hat{x}_k$ is exponentially stable.

In the statement of the lemma, the definition of exponential stability of $x_{k+1} = F_k x_k$ is standard: the transition matrix must satisfy $\|\phi_{k,l}\| \leq \alpha \beta^{k-l}$ for some $\alpha \in [1, \infty)$, and $\beta \in [0, 1)$ and all $k \geq l$. The key idea in proving the lemma is to show that $\phi_{k,l} = \hat{\phi}'_{-l+1, -k+1}$; application of Definitions 2.1 and 2.2 and the definition of exponential stability then yields the result.

We also need to define bounded-input, bounded-output stability. This is done in a standard way [4], [5], [9].

DEFINITION 2.3. The linear finite-dimensional system (2.1) has the bounded-input, bounded-output (BIBO) (l^p) property for $1 \leq p \leq \infty$ if, with W_{kl} the impulse response matrix of the system, the input bound

$$(2.9) \quad \sum_{k=-\infty}^{+\infty} [\|u_k\|^p]^{1/p} \leq \alpha$$

implies the output bound

$$(2.10) \quad \sum_{k=-\infty}^{+\infty} [\|y_k\|^p]^{1/p} \leq \beta \alpha$$

for some constant β , independent of $\{u_k\}$ and α , with $\{y_k\}$ related to $\{u_k\}$ by

$$(2.11) \quad y_k = \sum_{l < k} W_{kl} u_l.$$

We shall use the following readily established fact. The result in question is virtually established in [9, see pp. 113–114]; the key idea is to note first that $\|W_{kl}\|$ is exponentially bounded in the same manner as $\phi_{k,l}$.

LEMMA 2.2. *If the homogeneous equation $x_{k+1} = F_k x_k$ associated with (2.1) is exponentially stable, then (2.1) is BIBO (l^p) for all $p \leq [1, \infty]$.*

Our task in the next two sections is to prove a converse, given uniform detectability and stabilizability.

3. Connection between BIBO and BIBS behavior. Prior to obtaining the main result in the next section, we establish here that, given a uniform detectability condition, BIBO behavior implies bounded-input bounded-state (BIBS) behavior. The result is intuitively reasonable, and is known when detectability is replaced by observability [4], [5]. We use two preliminary lemmas.

LEMMA 3.1. *Let T_k be an orthogonal matrix and suppose that state-variable equations for $\bar{x}_k = T_k x_k$ are constructed from (2.1). Then, in obvious notation,*

$$(3.1) \quad \bar{M}_{k+s,k} = T_k M_{k+s,k} T_k'.$$

The proof is obvious. It is clear from this lemma that there is no loss of generality in establishing that BIBO behavior implies BIBS behavior for a system (2.1) for which, for each k , $M_{k+s,k}$ is diagonal, with diagonal entries ordered in decreasing magnitude. Suppose this is done. Let us also write, for a uniformly detectable system,

$$(3.2) \quad M_{k+s,k} = M_{k+s,k}^1 \dot{+} M_{k+s,k}^2$$

with $\dot{+}$ denoting direct sum, where the diagonal entries of $M_{k+s,k}^1$ are all greater than or equal to b , and those of $M_{k+s,k}^2$ are all less than b . Note that the dimension of $M_{k+s,k}^1$ may not be constant with k . Then we can state:

LEMMA 3.2. *Suppose (2.1) is uniformly detectable, and let W_{kl} denote the impulse response of (2.1). Let k_0 be an arbitrary but fixed integer, suppose notation is as defined in § 2, and let $M_{k+s,k}^1$ be as defined above. Let the input sequence $\{u_k\}$ and x_{k_0-t+1} be*

arbitrary. Define a sequence $x_k^* \in R^n$ for $k \geq k_0 - t + 1$ by

$$(3.3) \quad Hx_{k_0}^* = x_{k_0-1}^* = \cdots = x_{k_0-t+1}^*$$

$$(3.4) \quad x_{k+t}^* = \phi_{k+t,k} \left\{ \begin{bmatrix} (M_{k+s,k}^1)^{-1} & 0 \\ 0 & 0 \end{bmatrix} \sum_{i=k}^{k+s} \phi_{i,k}' H_i \left[y_i - \sum_{l=k}^{i-1} W_{il} u_l \right] \right. \\ \left. + \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} x_k^* \right\} + \sum_{l=k}^{k+t-1} \phi_{k+t-1,l} G_l u_l.$$

Then for $k \geq k_0 - t + 1$

$$(3.5) \quad x_{k+t}^* - x_{k+t} = \phi_{k+t,k} \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} (x_k^* - x_k).$$

The partitioning of the matrix $\begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}$ in (3.4) and (3.5) is like that of $M_{k+s,k}$ in (3.2).
Proof. Observe that

$$y_i - \sum_{l=k}^{i-1} W_{il} u_l = H_i' \phi_{ik} x_k.$$

Recalling the definition of $M_{k+s,k}$ in (2.4), we see that (3.4) implies

$$\begin{aligned} x_{k+t}^* &= \phi_{k+t,k} \left\{ \begin{bmatrix} (M_{k+s,k}^1)^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} M_{k+s,k}^1 & 0 \\ 0 & M_{k+s,k}^2 \end{bmatrix} x_k + \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} x_k^* \right\} + \sum_{l=k}^{k+t-1} \phi_{k+t-1,l} G_l u_l \\ &= \phi_{k+t,k} \left\{ \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} x_k + \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} x_k^* \right\} + \sum_{l=k}^{k+t-1} \phi_{k+t-1,l} G_l u_l \\ &= \phi_{k+t,k} x_k + \sum_{l=k}^{k+t-1} \phi_{k+t-1,l} G_l u_l + \phi_{k+t,k} \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} (x_k^* - x_k) \\ &= x_{k+t} + \phi_{k+t,k} \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} (x_k^* - x_k). \end{aligned}$$

Equation (3.5) is immediate.

Equations (3.3) and (3.4) in effect are defining a type of state estimator, with x_k^* supposed to estimate x_k . In case observability is present, one has $x_{k+t}^* = x_{k+t}$. If it is lacking, (3.5) holds, and as argued below, see (3.7), this ensures that the observation error asymptotically is zero.

Now we can establish the main result of this section, which relates BIBO and BIBS stability.

PROPOSITION 3.1. *Consider the system (2.1) and suppose it is uniformly detectable. Then for any one $p \in [1, \infty]$, the system is BIBS (l^p) if and only if it is BIBO (l^p).*

Proof. Because H_k is bounded, BIBS obviously implies BIBO. So we must prove BIBO implies BIBS.

Assume that $M_{k+s,k}$ is diagonal with $M_{k+s,k}^1$ as described above Lemma 3.2. Equation (3.4) can be reorganized as

$$(3.6) \quad x_{k+t}^* = \phi_{k+t,k} \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} x_k^* + w_k,$$

where w_k is a finite length moving average of $\{u_k\}$ and $\{y_k\}$ terms, with bounded weights. The BIBO hypothesis implies that if $[\sum_{k \geq k_0} \|u_k\|^p]^{1/p} \leq \alpha$, then $[\sum_{k \geq k_0} \|w_k\|^p]^{1/p} \leq \eta \alpha$ for some $\eta > 0$, independent of α , $\{u_k\}$ and k_0 . Also, recalling the special form of $M_{k+s,k}$ in (3.2), we see from the detectability definitions that any

vector of the form $[0, \xi_2']'$ has the property

$$\left\| \phi_{k+t,k} \begin{bmatrix} 0 \\ \xi_2 \end{bmatrix} \right\| < d \|\xi_2\|$$

or for all ξ_1, ξ_2 of appropriate dimension

$$\left\| \phi_{k+t,k} \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} \right\| < d \left\| \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} \right\|,$$

so that

$$(3.7) \quad \left\| \phi_{k+t,k} \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} \right\| < d < 1.$$

This means that (3.6) is BIBO (l^p) by Lemma 2.2. Using the $\{w_k\}$ bound, we see that x_k^* is such that $[\sum_{k \geq k_0} \|x_k^*\|^p]^{1/p} < \delta\alpha$ for some δ , independent of α , $\{u_k\}$ and k_0 . The boundary conditions on x_k^* in Lemma 3.2 and on x_k in the BIBO Definition 2.3 imply $x_k^* - x_k = 0$ for $k = k_0 - t + 1, \dots, k_0$ and so by (3.5), $x_k^* = x_k$ for all k . The result is then immediate.

4. Connection between BIBO behavior and exponential stability. In the previous section, we have argued that BIBO behavior and uniform detectability imply BIBS behavior. Here, we shall argue that BIBS behavior and uniform stabilizability imply exponential stability of the homogeneous state equation of the system. We shall use two preliminary lemmas. The main idea of the proof of the main result is to combine the use of duality established in the lemmas with the BIBO/BIBS connection.

LEMMA 4.1. *Let W_{kl} be the impulse response of the system (2.1) and \hat{W}_{kl} the impulse response of*

$$(4.1a) \quad \hat{x}_{k+1} = F'_{-k} \hat{x}_k + H_{-k} \hat{u}_k,$$

$$(4.1b) \quad \hat{y}_k = G'_{-k} \hat{x}_k.$$

Then $\hat{W}_{kl} = W'_{-l, -k}$.

The proof is trivial by direct calculation, and is omitted.

LEMMA 4.2. *The system (2.1) is BIBO (l^p) for p satisfying $1 \leq p \leq \infty$ if and only if the system (4.1) is BIBO (l^q) for q satisfying $p^{-1} + q^{-1} = 1$.*

Proof. Using the Hölder inequality, the duality of l^p and l^q , and simple manipulation, we obtain the following set of equivalences.

System (2.1) is BIBO (l^p)

$$\Leftrightarrow \forall \{u_k\} \in l^p, \{\hat{u}_k\} \in l^q, \left\| \sum_{k=-\infty}^{+\infty} \hat{u}'_k \left(\sum_{l=-\infty}^{k-1} W_{kl} u_l \right) \right\| \leq c [\sum \|u_k\|^p]^{1/p} [\sum \|\hat{u}_k\|^q]^{1/q}$$

for some constant c , independent of $\{u_k\}, \{\hat{u}_k\}$

$$\Leftrightarrow \sum_{l=-\infty}^{+\infty} \left\| \sum_{k=l+1}^{\infty} \hat{u}'_k W_{kl} \right\| u_l \leq c [\sum \|u_k\|^p]^{1/p} [\sum \|\hat{u}_k\|^q]^{1/q}$$

$$\Leftrightarrow \left\| \sum_{m=-\infty}^{+\infty} u'_{-m} \left(\sum_{n=-\infty}^{m-1} W'_{-n, -m} \hat{u}_{-n} \right) \right\| \leq c [\sum \|u_{-m}\|^p]^{1/p} [\sum \|\hat{u}_{-n}\|^q]^{1/q}$$

(on transposing and setting $m = -l, n = -k$)

$$\Leftrightarrow \left\| \sum_{n=-\infty}^{\infty} u'_{-m} \left(\sum_{n=-\infty}^{m-1} \hat{W}_{mn} \hat{u}_{-n} \right) \right\| \leq c [\sum \|u_{-m}\|^p]^{1/p} [\sum \|\hat{u}_{-n}\|^q]^{1/q}$$

$$\Leftrightarrow (4.1) \text{ is BIBO } (l^q).$$

Now we can state the main result.

THEOREM 4.1. *Suppose for the system (2.1), the pairs $[F_k, G_k]$ and $[F_k, H_k]$ are uniformly stabilizable and detectable. Then if (2.1) is BIBO (l^p) for any one $p \in [1, \infty]$, $x_{k+1} = F_k x_k$ is exponentially stable.*

Proof. If (2.1) is BIBO (l^p), then (2.1) is BIBS (l^p) by Proposition 3.1. Then (2.1a) in conjunction with $y_k = x_k$ is BIBO (l^p) and so by Lemma 4.2, the following system is BIBO (l^q):

$$(4.2a) \quad \hat{x}_{k+1} = F'_{-k} \hat{x}_k + \hat{u}_q,$$

$$(4.2b) \quad \hat{y}_k = G'_{-k} \hat{x}_k.$$

By Proposition 3.1, (4.2a) is BIBS (l^q), so (4.2a) together with $\hat{y}_k = \hat{x}_k$ is BIBO (l^q). By Lemma 4.2 again,

$$(4.3a) \quad x_{k+1} = F_k x_k + u_k,$$

$$(4.3b) \quad y_k = x_k$$

is BIBO (l^p), and in particular BIBS (l^p). By a standard result (see [1], [2]), the associated homogeneous equation is exponentially stable.

The following is an immediate consequence of Lemma 2.2 and Theorem 4.1.

COROLLARY 4.1. *Suppose that for the system (2.1), the pairs $[F_k, G_k]$ $[F_k, H_k]$ are uniformly stabilizable and detectable. Then (2.1) is BIBO (l^p) for any one $p \in [1, \infty]$ if and only if it is BIBO (l^p) for all $p \in [1, \infty]$.*

REFERENCES

- [1] O. PERRON, *Die Stabilitätsfrage bei Differential Gleichungen*, Math. Z., 32 (1930), pp. 703–728.
- [2] R. E. KALMAN AND J. E. BERTRAM, *Control system analysis and design by the second method of Lyapunov*, Trans. ASME J. Basic Engng., 82, Ser. D (1960), pp. 371–400.
- [3] B. D. O. ANDERSON, *Stability properties of linear systems in phase-variable form*, Proc. IEE, 115 (1968), pp. 340–341.
- [4] L. M. SILVERMAN AND B. D. O. ANDERSON, *Controllability, observability and stability of linear systems*, SIAM J. Control, 6 (1968), pp. 121–130.
- [5] B. D. O. ANDERSON AND J. B. MOORE, *New results in linear system stability*, SIAM J. Control, 7 (1969), pp. 398–414.
- [6] B. D. O. ANDERSON, *External and internal stability of linear systems—a new connection*, Proc. Joint Automatic Control Conference, Washington Univ., August 1971, pp. 53–58; also, IEEE Trans. Automat. Control, AC-17, (1972), pp. 107–111.
- [7] W. W. HAGER AND L. L. HOROWITZ, *Convergence and stability properties of the discrete Riccati equation and the associated control and filtering problems*, this Journal, 14 (1976), pp. 295–312.
- [8] B. D. O. ANDERSON AND J. B. MOORE, *Detectability and stabilizability of time-varying discrete-time linear systems*, this Journal, 19 (1981), pp. 20–32.
- [9] C. A. DESOER AND M. VIDYASAGAR, *Feedback Systems: Input-Output Properties*, Academic Press, New York, 1975.

FINITE ELEMENT APPROXIMATION OF PARABOLIC TIME OPTIMAL CONTROL PROBLEMS*

GREG KNOWLES†

Abstract. The numerical approximation of a parabolic time optimal control problem via piecewise linear splines, is considered. At each stage of the approximation a bang-bang approximate control is selected by solving for its switching times as the solution of a constrained nonlinear least squares optimization problem. The well-posedness of the approximation scheme is shown and the rate of convergence to the exact solution investigated. Numerical results for some one- and two-dimensional parabolic control problems are given.

1. Introduction. In this paper we consider the numerical solution of the following parabolic time optimal control problem in a bounded n -dimensional domain Ω with C^∞ boundary Γ :

$$\begin{aligned} (1) \quad & \frac{\partial u}{\partial t} = Au := \sum_{i,j=1}^n \frac{\partial}{\partial x_i} \left(a_{ij} \frac{\partial u}{\partial x_j} \right) && \text{in } \Omega, \quad t > 0, \\ (2) \quad & u(\mathbf{x}, 0) = u_0(\mathbf{x}), && \mathbf{x} \in \Omega, \\ (3) \quad & \frac{\partial u}{\partial \nu_A}(\mathbf{x}, t) + bu(\mathbf{x}, t) = \sum_{i=1}^m g_i(\mathbf{x})f_i(t), && \mathbf{x} \in \Gamma, \quad t > 0. \end{aligned}$$

The operator A is assumed symmetric, and uniformly elliptic, that is there exists an $\alpha > 0$ such that

$$(4) \quad \sum_{i,j=1}^n a_{ij}(\mathbf{x}) \xi_i \xi_j \geq \alpha |\xi|^2$$

for all $\mathbf{x} \in \Omega$, and $\xi \in R^n$ ($|\cdot|$ denotes the Euclidean norm). The coefficients of A , (a_{ij}) will be assumed C^∞ in $\bar{\Omega}$, and $(a_{ij}) = (a_{ji})$. In (3) $\partial/\partial \nu_A$ represents the conormal derivative, b is a nonnegative constant, g_1, \dots, g_m are given fixed functions, and $\mathbf{f} = (f_i)$ will be taken as the control function and allowed to vary inside the set of admissible controls

$$U = \{\mathbf{f} = (f_i): f_i \text{ is Lebesgue measurable and } |f_i| \leq 1, i = 1, \dots, m\}.$$

The solution of (1)–(3) for such an \mathbf{f} (when it exists) will be denoted by $u(\mathbf{x}, t; \mathbf{f})$ or just $u(t; \mathbf{f})$ when the explicit dependence on \mathbf{x} is not required. In fact, under the minimal assumptions, that $g_i \in L^2(\Gamma)$, $i = 1, \dots, m$, $u_0 \in L^2(\Omega)$, and $\{a_{ij}\} \in C^\infty(\bar{\Omega})$, then (1)–(3) has a unique weak solution $u \in H^{3/2, 3/4}(\Omega \times [0, T])$ for each $\mathbf{f} \in U$, any $T > 0$. (For the definition and properties of the Sobolev spaces $H^r(\Omega)$, $H^r(\Gamma)$, $H^{r,s}(\Omega \times [0, T])$ r, s real, we refer to [10].)

The control problem whose numerical solution will be considered here is: given a desired final state $u_1 \in L^2(\Omega)$, and a terminal error $E > 0$, find an admissible control \mathbf{f} which minimizes the time of transfer from u_0 to $B(u_1, E) = \{v \in L^2(\Omega): \|u_1 - v\|_{L^2(\Omega)} \leq E\}$. (Of course, we suppose $u_0 \notin B(u_1, E)$.)

Although this and similar control problems have been intensively treated in the literature, e.g., [3], [5], [8], there still remains little work done on numerical methods for solving these problems on general regions Ω and for nonsymmetric or time

* Received by the editors December 3, 1980, and in final form June 16, 1981. This research was supported by the National Science Foundation under grant MCS78-25526.

† Department of Mathematics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213.

dependent operators A . Note that in the case where Ω has a special shape and the separation of variables solution of (1)–(3) can be obtained explicitly, there exist good numerical schemes [6], [9], [20]. However, to apply these ideas to general regions Ω would require numerical precomputation of the eigenfunctions and eigenvalues by, say, finite differences or finite elements. Further, in the case of time dependent, nonsymmetric or nonlinear problems, these methods cannot usually be applied. In this note we suggest a finite element algorithm for approximating the control problem (1)–(3) directly, and give estimates for the rate of convergence when the coefficients (a_{ij}) of A are symmetric and time independent. The numerical methods presented here apply with change to nonsymmetric, time dependent operators A , although we have not yet been able to carry over the corresponding error estimates.

As a final remark we note that most of the theoretical work indicates that optimal controls for parabolic problems should be bang-bang (that is, $|f_i(t)| \equiv 1$, a.e.) with at most a countable number of switches accumulating at the optimal time, t^* . Accordingly, at each stage in the approximation we shall construct an approximating control which is bang-bang with a finite number of switches, so it can be computed by searching for these switching times.

The second section of this paper discusses the approximation scheme, in the third the error estimates are derived and in the fourth and fifth numerical methods and results are presented.

2. The finite element approximation. In the sequel we denote the norm and inner product on $L^2(\Omega)$ by $\|\cdot\|$, and (\cdot, \cdot) respectively. The inner product on $L^2(\Gamma)$, will be denoted $\langle \cdot, \cdot \rangle$, the norm by $|\cdot|$, and $\|\cdot\|_{s,\Omega}$, $\|\cdot\|_{s,\Gamma}$ will denote the usual Sobolev norms on $H^s(\Omega)$, $H^s(\Gamma)$.

We consider a standard Galerkin scheme for approximating (1)–(3), namely, suppose $\phi \in H^1(\Omega)$ and u satisfies (1)–(3). Then

$$(5) \quad (u_n, \phi) = (Au, \phi).$$

So, formally integrating by parts and substituting into (3) gives

$$(6) \quad \frac{d}{dt}(u(t), \phi) + B(u(t), \phi) = \sum_{i=1} \langle g_i, \phi \rangle f_i(t) \quad \text{for } \phi \in H^1(\Omega),$$

$$(7) \quad u(0) = u_0,$$

where

$$(8) \quad B(u, \phi) = \sum_{i,j=1} \int_{\Omega} a_{ij}(\mathbf{x}) \frac{\partial u}{\partial x_j} \frac{\partial \phi}{\partial x_i} d\mathbf{x} + b \int_{\Gamma} u \phi d\Gamma$$

for $u, \phi \in H^1(\Omega)$. As remarked earlier, for $u_0 \in L^2(\Omega)$ and $\{g_i\} \in L^2(\Gamma)$, the solution of (6), (7) exists and is unique.

To approximate (6), (7) we suppose finite dimensional subspaces $\{S_h\} \subset H^1(\Omega)$ are given such that for h small positive, there exists a constant C , independent of h and v for which,

$$\inf_{w \in S_h} \{\|v - w\| + h\|v - w\|_{1,\Omega}\} \leq Ch^s \|v\|_{s,\Omega}, \quad v \in H^s(\Omega) \quad \text{for } 1 \leq s \leq 2,$$

and that the following “inverse estimate” holds;

$$(9) \quad \|w\|_{1,\Omega} \leq Ch^{-1} \|w\| \quad \text{for all } w \in S_h.$$

We also assume that the family $\{S_h\}_{h>0}$ is dense in $L^2(\Omega)$. For examples of such subspaces, which include the usual piecewise linear finite element spaces, see [13].

We may now pose the semi-discrete problem. Find $u_h(t) \in S_h$ for $t > 0$, such that

$$(10) \quad \frac{d}{dt}(u_h(t), \phi) + B(u_h(t), \phi) = \sum \langle g_i, \phi \rangle f_i(t),$$

and

$$(11) \quad (u_h(0), \phi) = (u_0, \phi) \quad \text{for all } \phi \in S_h.$$

Once a basis for the finite dimensional space S_h , $\{\phi_1, \dots, \phi_N\}$ is given ($N = N(h)$ is the dimension of S_h), we can write

$$(12) \quad u_h(t) = \sum_{i=1}^N Q_i(t) \phi_i, \quad u_h(0) = \sum_{i=1}^N Q_i^0 \phi_i$$

and (10) is equivalent to the system of ordinary differential equations

$$(13) \quad M\dot{\mathbf{Q}}(t) + (K + bF)\mathbf{Q}(t) = H\mathbf{f}(t),$$

$$(14) \quad \mathbf{Q}(0) = \mathbf{Q}^0$$

where the mass and stiffness matrices M and K are

$$M = \{(\phi_i, \phi_j)\}, \quad K = \left\{ \sum_{k,l} \int_{\Omega} a_{kl}(\mathbf{x}) \frac{\partial \phi_i}{\partial x_k} \frac{\partial \phi_j}{\partial x_l} d\mathbf{x} \right\}, \quad F = \{(\phi_i, \phi_j)\}$$

and

$$H = \{(\phi_i, \phi_j)\}.$$

For convenience we set

$$(15) \quad L = K + bF.$$

The system (10), (11) can be regarded as the state equation of the finite dimensional control problem (in the space S_h with L^2 norm) of steering $u_h(0)$ to $B_h(u_1^h, E) = \{v \in S_h : \|u_1^h - v\| \leq E\}$ in minimum time, where u_1^h is the L^2 projection of u_1 onto S_h . (Of course, this problem is solved numerically by solving the time optimal control problem for the system of ordinary differential equations (13), (14): we will discuss this further in § 4.) The optimal control \mathbf{f}_h and minimum time t_h for (10), (11) will then be taken as the approximate optimal control and minimum time for (1)–(3) at this stage of the approximation.

The questions which will be answered next then, concern the existence of \mathbf{f}_h and t_h , the rate of convergence of t_h to the minimum time for the original problem, t^* , and the rate of convergence to zero of the “end point error” of using control \mathbf{f}_h to time t_h , in other words, the convergence to zero of the distance between $u(\mathbf{f}_h, t_h)$ and the target set $B(u_1, E)$.

3. Convergence of the approximate control problem. In this section we discuss the well-posedness of the approximate problem and its convergence as $h \rightarrow 0$. Firstly we assume a controllability assumption for (1)–(3) in order to guarantee the existence of an optimal control and the minimum time for the original problem. In this section K and C will denote generic constants.

(H) Suppose that there exist a time $t > 0$ and an admissible control \mathbf{f} such that $\|u(t, \mathbf{f}) - u_1\| < E$.

Under assumption (H) it follows by standard arguments that the minimum time to hit $B(u_1, E)$, using admissible controls, exists; we denote it by t^* , and let \mathbf{f}^* be an admissible time optimal control [18].

THEOREM 1. *If (H) holds, $g_i \in L^2(\Gamma)$, $i = 1, 2, \dots, m$, $u_0 \in L^2(\Omega)$, then for any h sufficiently small an optimal control \mathbf{f}_h steering $u_h(0)$ to $B_h(u_1^h, E)$ in minimum time t_h exists. Further, \mathbf{f}_h can be chosen bang-bang, with each component having at most N switches.*

Proof. By (H) there exists an admissible control \mathbf{f} such that in some time $t > 0$

$$(16) \quad \|u(t, \mathbf{f}) - u_1\| = \sigma < E.$$

By the usual compactness arguments, we know that

$$(17) \quad \|u(t, \mathbf{f}) - u_h(t, \mathbf{f})\| \rightarrow 0 \quad \text{as } h \rightarrow 0,$$

(for fixed t, \mathbf{f}); consequently, for h sufficiently small we have

$$(18) \quad \|u(t, \mathbf{f}) - u_h(t, \mathbf{f})\| < \frac{E - \sigma}{2}.$$

Furthermore, since the spaces $\{S_h\}_{h>0}$ are dense in $L^2(\Omega)$,

$$\|u_1 - u_1^h\| < \frac{E - \sigma}{2}$$

for h sufficiently small. Hence

$$(19) \quad \|u_1^h - u_h(t, \mathbf{f})\| \leq \|u(t, \mathbf{f}) - u_h(t, \mathbf{f})\| + \|u(t, \mathbf{f}) - u_1\| + \|u_1 - u_1^h\| \leq E$$

for h small, so there exists an admissible control transferring $u_h(0)$ to $B_h(u_1^h, E)$ in finite time. Accordingly, by [7], [11] there must exist a bang-bang optimal control effecting this transfer in minimum time, t_h . Since the finite dimensional state equation ((10) or (13)) has constant coefficients, we may choose \mathbf{f}_h finite switching [7], [11]. However, the eigenvalues of the matrix $-M^{-1}L$ are real, so a result of Feld'baum [12] guarantees that each component of \mathbf{f}_h has at most N switches.

Remark. If the system (10), (13) is normal, that is, writing $H = [\mathbf{h}_1, \dots, \mathbf{h}_m]$ in columns, if $\text{rank} [\mathbf{h}_i, M^{-1}L\mathbf{h}_i, \dots, (M^{-1}L)^{N-1}\mathbf{h}_i] = N$, for every $i = 1, 2, \dots, m$, then the optimal control \mathbf{f}_h for (10), (11) ((13), (14)) is unique.

Next, we estimate the distance between $u(t_h, \mathbf{f}_h)$ and $u_h(t_h, \mathbf{f}_h)$ in the $L^2(\Omega)$ norm. This result will play a central role in the following theory.

PROPOSITION 1. *Suppose $u_0 \in H^{3/2}(\Omega)$ and $g_i \in L^2(\Gamma)$, $i = 1, 2, \dots, m$. Then for any small $\delta > 0$ there exists a constant C_δ , independent of h , for which*

$$(20) \quad \|u(t_h, \mathbf{f}_h) - u_h(t_h, \mathbf{f}_h)\| \leq C_\delta h^{3/2-\delta} \left(\|u_0\|_{3/2, \Omega} + \sum_{i=1}^m |g_i| \right)$$

for all $h > 0$ sufficiently small.

Proof. Without loss of generality take $m = 1$, and let $\delta \in (0, \frac{3}{2})$ be given. Choose $\rho \in (0, \delta)$. We show initially that $u(t, f_h) \in H^{3/2-\rho}(\Omega)$ for $t \in [0, t_h]$, and

$$(21) \quad \|u(t, f_h)\|_{3/2-\rho, \Omega} \leq K_\rho (\|u_0\|_{3/2, \Omega} + |g_1|), \quad 0 \leq t \leq t_h,$$

where K_ρ is independent of h . The proof of (21) is a simple consequence of results in [14] and [15].

Denote by G the solution operator of the elliptic boundary value problem

$$(22) \quad Au = 0 \quad \text{in } \Omega,$$

$$(23) \quad \frac{\partial u}{\partial \nu_A} + bu = w \quad \text{on } \Gamma,$$

that is, $u = Gw$. Then it is well known that for $w \in L^2(\Gamma)$, $Gw \in H^{3/2}(\Omega)$ and

$$\|Gw\|_{3/2,\Omega} \leq C|w|$$

[10]. Let \mathcal{A} be the operator defined by $-A$ on the subspace $\mathcal{D}(\mathcal{A})$ of $H^2(\Omega)$ of functions satisfying the homogeneous form of the boundary condition (23). \mathcal{A} generates a C_0 , analytic, contraction semi-group $\{S(t)\}_{t \geq 0}$ of bounded linear operators, $S(t): L^2(\Omega) \rightarrow L^2(\Omega)$. In terms of \mathcal{A} , G and $\{S(t)\}$ the solution of (1)–(3) can be written as

$$(24) \quad u(t, f_h) = S(t)u_0 + \int_0^t \mathcal{A}S(t-\tau)Gg_1f_h(\tau) d\tau, \quad t \geq 0$$

[14], [16]. Since $\{S(t)\}$ is an analytic semi-group and $Gg \in H^{3/2}(\Omega) \subset H^{3/2-\rho}(\Omega) = \mathcal{D}(\mathcal{A}^{3/4-\rho/2})$ [15 (2.5), (2.6)], we have

$$\|\mathcal{A}S(\tau)Gg_1\|_{3/2-\rho,\Omega} \leq K_\rho \|\mathcal{A}S(\tau)\mathcal{A}^{3/4-\rho/2}Gg_1\|, \quad \tau > 0.$$

Further, $\mathcal{A}^{3/4-\rho/2}(Gg_1) \in H^\rho(\Omega)$, and so by [14, Thm. 2]

$$\|\mathcal{A}S(\tau)Gg_1\|_{3/2-\rho,\Omega} \leq \frac{K_\rho}{\tau^{1-\rho}} \|\mathcal{A}^{3/4-\rho/2}Gg_1\|_{\rho,\Omega} \leq \frac{K_\rho}{\tau^{1-\rho}} \|Gg_1\|_{3/2,\Omega} \leq \frac{K_\rho}{\tau^{1-\rho}} |g_1|, \quad \tau > 0.$$

Accordingly, for $0 \leq t \leq t_h$

$$(25) \quad \left\| \int_0^t \mathcal{A}S(t-\tau)Gg_1f_h(\tau) d\tau \right\|_{3/2-\rho,\Omega} \leq \int_0^t \|\mathcal{A}S(t-\tau)Gg_1\|_{3/2-\rho,\Omega} d\tau \leq K_\rho t_h^\rho |g_1|.$$

We know from the proof of Theorem 1 that for h small $\{t_h\}$ is bounded above, and so for such h , $(25)_1 \leq K_\rho |g_1|$, where the constant K_ρ is independent of h . The full estimate (21) follows, as

$$\|S(t)u_0\|_{3/2-\rho,\Omega} \leq \|u_0\|_{3/2-\rho,\Omega} \leq \|u_0\|_{3/2,\Omega}, \quad t \geq 0.$$

The approximation result (20) can now be proven by the method of [17, (0.12)]. In fact the proof follows as in [17] except that the lower regularity of the solution here causes the estimate [17, (0.10)] to now become

$$\|\sigma(t)\| \leq Ch^{3/2-\rho} \|u(t, f_h)\|_{3/2-\rho,\Omega}$$

and consequently, the analogue of [17, (0.12)] here is

$$(26) \quad \begin{aligned} \|u(t_h, f_h) - u_h(t_h, f_h)\| &\leq Ch^{3/2-\rho} \left(1 + \log \left(1 + \frac{t_h}{h^2} \right) \right) \sup_{0 \leq t \leq t_h} \|u(t, f_h)\|_{3/2-\rho,\Omega} \\ &\leq C_\rho h^{3/2-\rho} \left(1 + \log \left(1 + \frac{t_h}{h^2} \right) \right) (\|u_0\|_{3/2,\Omega} + |g_1|), \end{aligned}$$

by virtue of (21). To derive (20) from (26) we use the fact that $\{t_h\}$ is bounded above for h small, for then $h^{3/2-\rho} (1 + \log(1 + t_h/h^2)) \leq C_\delta h^{3/2-\delta}$, by the initial choice of ρ .

LEMMA 1. *Under the assumptions of Proposition 1*

$$(27) \quad \lim_{h \rightarrow 0} t_h \geq t^*.$$

Proof. The method of proof is fairly standard; see, e.g., [9]. Let $\lim t_h = \bar{t} \geq 0$, and choose a subsequence $h_n \rightarrow 0$, $n = 1, 2, \dots$, such that $t_{h_n} \rightarrow \bar{t}$. To simplify the notation we shall denote $t_n = t_{h_n}$ and also set $f_n(\tau) = f_{h_n}(\tau)\chi_{[0, \bar{t}]}(\tau)$, $\tau \geq 0$, $n = 1, 2, \dots$. The controls $\{f_n\}$ form a bounded subset of $L^2(0, \bar{t})$ and hence have a weakly convergent subsequence, which we again denote by $\{f_n\}$, converging weakly in $L^2(0, \bar{t})$ to some \bar{f} , and $|\bar{f}| \leq 1$ a.e. We have

(2.8)

$$u(\bar{t}, \bar{f}) - u_{h_n}(t_n, f_n) = (u(\bar{t}, \bar{f}) - u(\bar{t}, f_n)) + (u(\bar{t}, f_n) - u_{h_n}(\bar{t}, f_n)) + (u_{h_n}(\bar{t}, f_n) - u_{h_n}(t_n, f_n)).$$

The first term on the right-hand side of (28) converges weakly to 0 in $L^2(\Omega)$, by the weak convergence of $f_n \rightarrow \bar{f}$ in $L^2(0, \bar{t})$ (see (24)). If $\bar{t} = 0$ the second term in (28) becomes $u_0 - u_0^{h_n}$ which converges to 0 in $L_2(\Omega)$ norm by the choice of u_0 ; if $\bar{t} > 0$ then a proof similar to that of Proposition 1 shows that this term still converges to 0 in $L^2(\Omega)$. Finally the third term in (28) converges to 0 in $L^2(\Omega)$ since $t_n \rightarrow \bar{t}$, and $|f_n| \leq 1$. The proof of this follows from the variation of parameters solution for u_h : the analogue of (24). Consequently,

$$u(\bar{t}, \bar{f}) - u_{h_n}(t_n, f_n) \xrightarrow{w} 0,$$

$n \rightarrow \infty$, in $L^2(\Omega)$, and as $u_{h_n}(t_n, f_n) \in B(u_1, E)$, and this ball is weakly closed, we must have $u(\bar{t}, \bar{f}) \in B(u_1, E)$. By the minimality of t^* , $\bar{t} \geq t^*$, as required.

Remark. To establish (27) it is enough to assume $u_0 \in L^2(\Omega)$. As a consequence the methods of [18] would allow one to extend the conclusions of Proposition 1 to initial conditions $u_0 \in L^2(\Omega)$. We won't dwell on this point here as it introduces further technicalities.

In the case $u_0 = 0$, we can show the convergence of t_h to t^* under an extra controllability assumption on u_1 . It is essentially a specialization of the well-known conditions for approximate controllability, to the particular target point u_1 [16], [15], [19]; it implies that u_1 can be reached arbitrarily closely with L^∞ controls. For simplicity we state the condition only for $m = 1$, the general statement may be found in [19, Appendix B, Lemma B1, (i), (ii)].

Let $\{\lambda_j\}$ and $\{\phi_j\}$ be the eigenvalues and corresponding eigenfunctions for $-\mathcal{A}$. Define

$$K(g_1) = \{k = 1, 2, \dots : \langle g_1, \phi_k \rangle \neq 0\}.$$

Then we will assume that

- (29) (i) $\lambda_i \neq \lambda_j$ for $i, j \in K(g_1)$,
(ii) if $\langle g_1, \phi_i \rangle = 0$, then $\langle u_1, \phi_j \rangle = 0$ for $j = 1, 2, \dots$.

THEOREM 2. If (29) holds, $u_0 = 0$, and $g_i \in L^2(\Gamma)$, $i = 1, 2, \dots, m$, then

$$\lim_{h \rightarrow 0} t_h = \overline{\lim_{h \rightarrow 0} t_h} = t^*.$$

Proof. By Lemma 1, it is enough to show that $\overline{\lim_{h \rightarrow 0} t_h} \leq t^*$. The proof of this follows along the lines of [20, Theorem 2.1].

Suppose $\rho > 0$ is given. Set $\tilde{t} = t^* + \rho$, and define an admissible control $\tilde{\mathbf{f}}$ on $(0, \tilde{t})$ by

$$\tilde{\mathbf{f}}(\tau) = \begin{cases} 0, & \tau \in (0, \tilde{t} - t^*), \\ \mathbf{f}^*(\tau - \tilde{t} + t^*), & \tau \in (\tilde{t} - t^*, \tilde{t}). \end{cases}$$

Since $u(t^*, \mathbf{f}^*) = u(\tilde{t}, \tilde{\mathbf{f}})$, $\|u(\tilde{t}, \tilde{\mathbf{f}}) - u_1\| \leq E$.

If we consider the control problem

$$\min \{ \|u(\tilde{t}, \mathbf{f}) - u_1\| : \mathbf{f} = (f_i), |f_i| \leq 1 \},$$

then this problem has an optimal solution $\hat{\mathbf{f}}$, $|\hat{f}_i| \leq 1$ [19], and

$$(30) \quad e \triangleq \|u(\tilde{t}, \hat{\mathbf{f}}) - u_1\| < E.$$

For if $e = 0$, then the result is clear; if not, assumption (29) implies that at least one component of any optimal control must be bang-bang [19, Thms. 6, and B1]. Since $\tilde{\mathbf{f}} = 0$ on $(0, \tilde{t} - t^*)$, it cannot be optimal, and so

$$e < \|u(\tilde{t}, \tilde{\mathbf{f}}) - u_1\| \leq E.$$

Using the same argument as in Theorem 1, $\|u_h(\tilde{t}, \hat{\mathbf{f}}) - u_1^h\| \leq E$ for h sufficiently small, and hence $\tilde{t} \geq t_h$ by the minimality of t_h . In other words,

$$\tilde{t} = t^* + \rho \geq \lim_{h \rightarrow 0} t_h,$$

and since $\rho > 0$ was arbitrary the result follows.

Using the proposition it is now easy to show the convergence to zero of the distance between the target set $B(u_1, E)$, and the solution of (1)–(3) obtained by using the approximate control \mathbf{f}_h to time t_h , $u(t_h, \mathbf{f}_h)$. In other words, as we would hope, the approximate control \mathbf{f}_h leads the solution of (1)–(3) increasingly close to the terminal target set, as $h \rightarrow 0$.

THEOREM 3. *If $u_0, u_1 \in H^{3/2}(\Omega)$ and $g_i \in L^2(\Gamma)$, $i = 1, 2, \dots, m$, then for arbitrarily small $\delta > 0$,*

$$(31) \quad \|u(t_h, \mathbf{f}_h) - u_1\| \leq E + O(h^{3/2-\delta}).$$

Proof. By the triangle inequality

$$\begin{aligned} \|u(t_h, \mathbf{f}_h) - u_1\| &\leq \|u_h(t_h, \mathbf{f}_h) - u_1^h\| + \|u(t_h, \mathbf{f}_h) - u_h(t_h, \mathbf{f}_h)\| + \|u_1 - u_1^h\| \\ &\leq E + K_\delta h^{3/2-\delta} + Kh^{3/2} \|u_1\|_{3/2, \Omega} \leq E + O(h^{3/2-\delta}). \end{aligned}$$

The estimate for $\|u_1 - u_1^h\|$ is a well-known consequence of approximation theory [13, Chap. 3].

The final result of this section gives an estimate for the rate of convergence of t_h to t^* in the case $u_1 = 0$ and $b > 0$.

THEOREM 4. *If $u_1 = 0$, $b > 0$, $u_0 \in H^{3/2}(\Omega)$ and $g_i \in L^2(\Gamma)$, $i = 1, 2, \dots, m$, then for arbitrarily small $\delta > 0$,*

$$(32) \quad |t_h - t^*| = O(h^{3/2-\delta}).$$

Proof. We will prove (32) in two parts,

$$(33) \quad t^* \leq t_h + O(h^{3/2-\delta}),$$

$$(34) \quad t_h \leq t^* + O(h^{3/2-\delta}).$$

Since we have already shown that $\|u(t_h, \mathbf{f}_h)\| \leq E + O(h^{3/2-\delta})$, if we can find an admissible control transferring $x_0 = u(t_h, \mathbf{f}_h)$ to $B(0, E)$ in time $t_1 = O(h^{3/2-\delta})$, then (33) will be proven; as we will then know that $B(0, E)$ is reachable in time $t_h + t_1 \leq t_h + O(h^{3/2-\delta})$, and the minimum time t^* must be less than or equal to this.

Conversely, if \mathbf{f}^* is an optimal control for the original problem, then

$$\|u_h(t^*, \mathbf{f}^*)\| \leq \|u_h(t^*, \mathbf{f}^*) - u(t^*, \mathbf{f}^*)\| + \|u_h(t^*, \mathbf{f}^*)\| \leq E + O(h^{3/2-\delta})$$

by Proposition 1. So again, if we can find an admissible control transferring $x_h^0 = u_h(t^*, \mathbf{f}^*)$ to $B_h(0, E)$ in S_h in time $t_1^h = O(h^{3/2-\delta})$, then the minimum time for the approximate problem $t_h \leq t^* + t_1^h \leq t^* + O(h^{3/2-\delta})$, giving (34).

We will show both these problems can be solved by using the admissible control $\mathbf{f} \equiv 0$. In fact, consider the state equations (8), (10) with $\mathbf{f} \equiv 0$, and initial conditions x_0, x_h^0 respectively that is,

$$(35) \quad \begin{aligned} \frac{d}{dt}(y(t), \phi) + B(y(t), \phi) &= 0 \quad \text{all } \phi \in H^1, \\ (y(0), \phi) &= (x_0, \phi). \end{aligned}$$

$$(36) \quad \begin{aligned} \frac{d}{dt}(y_h(t), \phi) + B(y_h(t), \phi) &= 0 \quad \text{all } \phi \in S_h, \\ (y_h(0), \phi) &= (x_h^0, \phi). \end{aligned}$$

Now since $b > 0$, one may show that the solutions of (35), (36) satisfy estimates of the following type (cf. [4, Problem 2.2.1]). There exists a $\rho > 0$, (independent of h) such that

$$(37) \quad \|y(t)\| \leq e^{-\rho t} \|x_0\|,$$

$$(38) \quad \|y_h(t)\| \leq e^{-\rho t} \|x_h^0\|.$$

Consequently, the solution of (35), (36) decays to zero as t increases, so if $t_1, (t_1^h)$, the first time for which $\|y(t_1)\| = E$ ($\|y_h(t_1^h)\| = E$), then by (37), (38)

$$E = \|y(t_1)\| \leq e^{-\rho t_1} \|x_0\| \leq e^{-\rho t_1} (E + K_\delta h^{3/2-\delta}).$$

Solving for t_1 , then gives, $t_1 \leq C_\delta h^{3/2-\delta}$. Similarly, $t_1^h \leq C_\delta h^{3/2-\delta}$ as required.

4. The numerical solution of the approximate control problem. In this section we discuss a numerical method for the solution of approximate control problems (13), (14) in R^N ($N = N(h)$, is the dimension of S_h). For notational simplicity we take $u_1 = 0$. Each of these problems has the following form:

Find the minimum time t_N^* to transfer $\mathbf{Q}(0) = \mathbf{Q}^0$ to $W = \{\mathbf{x} \in R^N : |\mathbf{x}| \leq \delta\}$ where

$$(39) \quad M\dot{\mathbf{Q}}(t) + L\mathbf{Q}(t) = H\mathbf{f}(t), \quad t > 0,$$

and $\mathbf{f} \in U$. The matrices M and L are $N \times N$ and H is $N \times m$. We write $H = [\mathbf{h}_1, \dots, \mathbf{h}_m]$ in columns. It has already been shown that the optimal control \mathbf{f}_N^* for this problem is bang-bang, and each component has at most N switches. Furthermore, the theoretical results for parabolic time optimal control problems indicate that the switching times of the optimal control should accumulate at the final time. Accordingly, it is to be expected, and in fact was observed in practice, that the switches of \mathbf{f}_N^* would "bunch up" at the final time t_N^* . To catch this with an evenly spaced grid on $[0, t_N^*]$ would require an excessively small mesh size, and consequently an excessive amount of computer time. So rather than solve (39) via methods based on the maximum principle and approximate the control on an evenly spaced grid, it was decided to solve for the switching times of \mathbf{f}_N^* directly. (This method will also give a solution even when the system (39) is not normal.) Namely, if the control f_j switches at $t_{j1}, \dots, t_{jN}, j = 1, \dots, m$, then once the initial value $f_j(0) = \pm 1$ is specified, the solution of (39) with initial

condition \mathbf{Q}_0 is now a function of (t_{jl}) , that is,

$$\mathbf{Q}(t) = \mathbf{Q}(t : t_{jl}), \quad t \geq 0.$$

Without loss of generality, we shall from now assume $f_j(0) = -1$ for all j .

The ordinary differential equations (39) were solved numerically by a slightly modified Crank–Nicolson difference scheme. As is well-known this method is unconditionally stable and achieves second order accuracy. If Δt denotes the time mesh size, and $\mathbf{Q}_n = \mathbf{Q}(n \Delta t)$, $\mathbf{f}_n = \mathbf{f}(n \Delta t)$, then, the scheme can be written

$$(40) \quad \left(M + \frac{\Delta t}{2}L\right)\mathbf{Q}_{n+1} = \left(M - \frac{\Delta t}{2}L\right)\mathbf{Q}_n + \frac{(\mathbf{f}_{n+1} + \mathbf{f}_n)\Delta t}{2}$$

(e.g., [13, § 7.1.]) Notice, however, that the forcing term \mathbf{f} is discontinuous, and if (40) is applied directly the truncation error in approximating \mathbf{f} will reduce the accuracy of the scheme.

To avoid this difficulty the switching times (t_{jl}) were sorted in increasing order, $0 < t_{j_0 l_0} < t_{j_1 l_1} < \dots$, say, and the equation solved firstly over $[0, t_{j_0 l_0}]$ using the final value from the previous interval as initial condition, etc. In this way the forcing term \mathbf{f} is constant over each subinterval and the accuracy is increased. The disadvantage is that the mesh sizes Δt on each subinterval may be different and so the matrices $M + (\Delta t/2)L$, $M - (\Delta t/2)L$, will have to be recomputed. However, it turned out in practice that the number of switching times was not large (never more than 6) and the above technique worked well. Notice also that although M and L may be large, they are typically sparse and banded, and can be stored in compact form. Both the matrices $M + (\Delta t/2)L$, $M - (\Delta t/2)L$, can be similarly stored, and the inversion necessary in (40) can be performed efficiently using special routines for banded, sparse matrices.

The approximate control problem can now be restated. Find (t_{jl}) such that

$$0 < t_{j1} < \dots < t_{jN} = t_N, \quad j = 1, 2, \dots, m,$$

with $\sum_{k=1}^N Q_k^2(t_N : t_{jl}) \leq \delta^2$, and t_N minimal, or, alternatively,

$$(P) \quad \text{minimize } t_N$$

subject to

$$(41) \quad \sum_{k=1}^N Q_k^2(t_N : t_{jl}) \leq \delta^2$$

and

$$(42) \quad 0 < t_{j1} < \dots < t_{jN} = t_N \quad \text{for } j = 1, 2, \dots, m.$$

(P) is a nonlinear program whose numerical solution we shall discuss next. First we give a lemma which indicates how the partial derivatives of \mathbf{Q} with respect to (t_{jl}) may be computed. Its proof can easily be derived from the variation of parameters solution of (39).

LEMMA 2. For $j = 1, 2, \dots, m$, $l = 1, 2, \dots, N-1$, we have

$$(43) \quad \frac{\partial \mathbf{Q}(t : t_{jl})}{\partial t_{jl}} = \begin{cases} 0 & \text{for } t < t_{jl}, \\ \mathbf{Z}_{jl}(t) & \text{for } t \geq t_{jl}, \end{cases}$$

where \mathbf{Z}_{jl} is the solution of the following system of ordinary differential equations:

$$(44) \quad M\dot{\mathbf{Z}}_{jl}(\tau) + L\mathbf{Z}_{jl}(\tau) = 0 \quad \text{for } \tau \in [t_{jl}, t],$$

with initial condition

$$M\mathbf{Z}_{jl}(t_{jl}) = 2(-1)^l \mathbf{h}_j.$$

Similarly, for $l = N$, we have,

$$(45) \quad \frac{\partial Q_j(t_{jN} : t_{jl})}{\partial(t_{jN})} = \mathbf{Z}_{jN}, \quad j = 1, 2, \dots, m,$$

where

$$(46) \quad M\mathbf{Z}_{jN} + L\mathbf{Q}(t_{jN} - 0) = H\mathbf{f}(t_{jN} - 0)$$

(-0) indicates left-hand limit).

To solve the program (P) numerically we first replaced the inequality constraints (42) via the change of variables,

$$(47) \quad \begin{aligned} t_{j1} &= y_{j1}^2, \\ t_{j2} &= y_{j1}^2 + y_{j2}^2, \\ &\dots\dots \\ t_{jN} &= y_{j1}^2 + y_{j2}^2 + \dots + y_{jN}^2; \end{aligned}$$

then the solution was obtained by a two-step method. We first approximated (P) by a penalty method,

$$\text{minimize} \quad \sum_{j=1}^M t_{jN} + \sigma \sum_{j=1}^N Q_j^2(t_{jN} : t_{jl})$$

for increasing values of σ until $\sum_{j=1}^N Q_j^2(t_{jN} : t_{jl})$ was sufficiently small. In terms of the new variables (y_{jl}) this becomes

$$(P') \quad \text{minimize} \quad \sum_{l=1}^N \sum_{j=1}^M y_{jl}^2 + \sigma \sum_{j=1}^N \tilde{Q}_j^2(y_{jl}),$$

where

$$\tilde{Q}_j(y_{jl}) = Q_j(y_{j1}^2 + y_{j2}^2 + y_{jN}^2 : y_{jl}).$$

The problem (P') is a nonlinear least squares minimization and was solved by the Levenberg–Marquart algorithm. Once the residuals had been reduced sufficiently (this usually only required one or two unconstrained minimizations of (P')), this approximate result from (P') was used as an initial guess for the unconstrained minimization

$$(P'') \quad \text{minimize} \quad \sum_{j=1}^N \tilde{Q}_j^2(y_{jl}).$$

The problem was considered solved when the value of (P'') became $\leq \delta$. In this way the penalty method is used to force the solution into a neighborhood of the optimum, obtaining an approximate answer, and then the locally quadratic convergence of Newton type methods can be taken advantage of to achieve as accurate an answer as required with only one more unconstrained minimization. The gradients of the functions \tilde{Q}_j , with respect to y_{jl} needed for the minimization routines, can be obtained from (43), (45) by the chain rule.

Finally we indicate the relation of δ to E . We have chosen $t_N (= t_h)$ and \mathbf{f}_h such that $\|u_h(t_N, \mathbf{f}_h)\| = E$. In terms of the coordinates $\mathbf{Q}(t)$ of $u_h(t)$ this implies that $\mathbf{Q}(t_N)M\mathbf{Q}^T(t_N) = E^2$, and hence

$$\|\mathbf{Q}(t_N)\|_{R^N}^2 \in \left[\frac{E^2}{\lambda_{\max}(M)}, \frac{E^2}{\lambda_{\min}(M)} \right],$$

where $\lambda_{\max}(M)$, $\lambda_{\min}(M)$ are the largest and smallest eigenvalues of the mass matrix M . These depend on h ; e.g., for piecewise linear splines are typically $O(h)$ [13, Chapt. 4]. In this case then we should choose $\delta = O(E/\sqrt{h})$.

5. Numerical results. We now present some numerical results to illustrate these techniques.

First, we consider a one-dimensional parabolic boundary control problem.

Example 1.

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2}, & 0 < x < 1, \quad t > 0, \\ u(x, 0) &= u_0(x), \\ \frac{\partial u}{\partial x}(0, t) &= 0, \\ \frac{\partial u}{\partial x}(1, t) &= f(t), \quad |f(t)| \leq 1. \end{aligned}$$

With the final endpoint $u_1(x) \equiv 0$, $0 \leq x \leq 1$, this problem was solved with S_h continuous piecewise linear splines, for $h = \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}$, and initial temperatures $u_0(x) = 1.0$, $u_0(x) = 0.5$, $u_0(x) = x^2$, $0 \leq x \leq 1$. The switching times for the approximate optimal controls are given in Tables 1, 2 and 3. We chose $E = 10^{-4}$.

The graphs in Figs. 1 and 2 are plots of $-\ln h$ against $-\ln \|\text{err}\|$, the $L^2(0, 1)$ norm of the endpoint errors, $\|u(t_h, \mathbf{f}_h) - u_1\|$ for Example 1 with initial temperature $u_0(x) = 1.0$, $u_0(x) = 0.5$ and $u_0(x) = x^2$, $0 \leq x \leq 1$.

TABLE 1
Example 1 with $v_0(x) = 1$

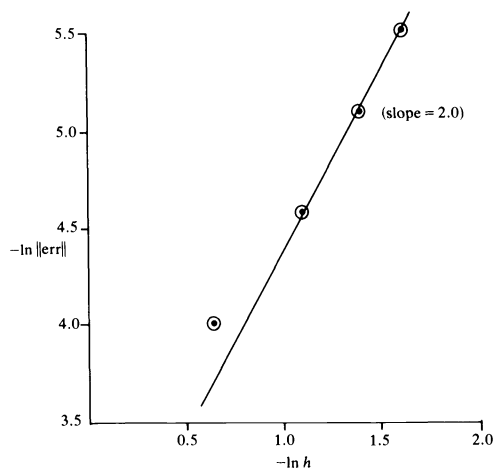
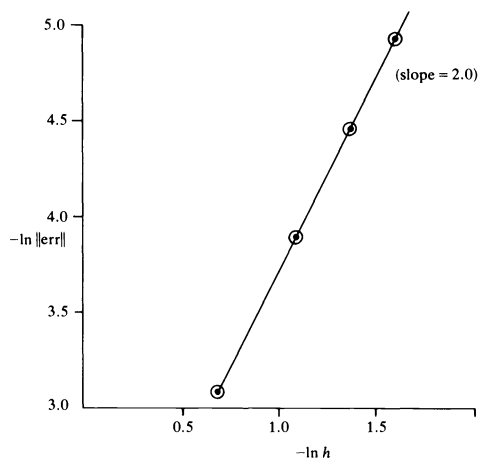
h	Switching times					
$\frac{1}{2}$	1.05992	1.13344	1.14706			
$\frac{1}{3}$	1.06558	1.14524	1.16589	1.17246		
$\frac{1}{4}$	1.06855	1.15197	1.17459	1.18599	1.18966	
$\frac{1}{5}$	1.06995	1.15552	1.17891	1.19139	1.19828	1.20055

TABLE 2
Example 1 with $v_0(x) = 0.5$

h	Switching times					
$\frac{1}{2}$	0.560196	0.633793	0.647193			
$\frac{1}{3}$	0.566185	0.646837	0.666684	0.672064		
$\frac{1}{4}$	0.569016	0.653506	0.675165	0.684476	0.687607	
$\frac{1}{5}$	0.570711	0.657389	0.679969	0.690781	0.697532	0.699872

TABLE 3
Example 1 with $v_0(x) = x^2$

h	Switching times					
$\frac{1}{2}$	0.434721	0.508173	0.521904			
$\frac{1}{3}$	0.416631	0.495836	0.515645	0.521023		
$\frac{1}{4}$	0.410994	0.493814	0.515489	0.524743	0.5279	
$\frac{1}{5}$	0.408228	0.493192	0.515969	0.526123	0.531777	0.533603

FIG. 1. Example 1 with $u_0(x) = 1.0$, $u_0(x) = 0.5$.FIG. 2. Example 1 with $u_0(x) = x^2$.

Finally, these techniques were applied to the two-dimensional parabolic boundary control problem on the square $0 \leq x, y \leq 1$.

Example 2.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \quad 0 < x, y < 1,$$

$$u(x, y, 0) = u_0(x, y),$$

$$\frac{\partial u}{\partial z}(0, y, t) = \frac{\partial u}{\partial x}(1, y, t) = \frac{\partial u}{\partial y}(x, 0, t) = 0,$$

$$\frac{\partial u}{\partial y}(x, 1, t) = f(t), \quad |f(t)| \leq 1.$$

Again $u_1(x, y) \equiv 0$, and S_h is the space of continuous piecewise linear splines over the square, $0 \leq x, y \leq 1$, with a right directed triangular grid.

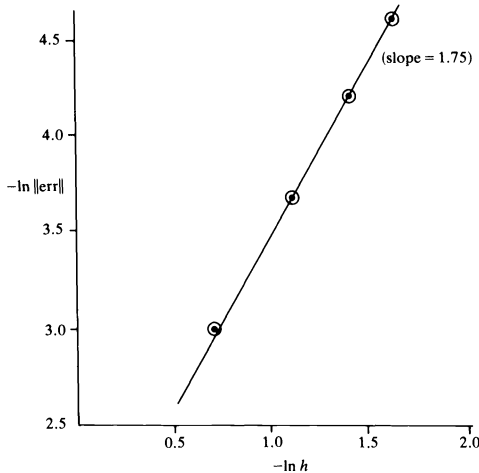
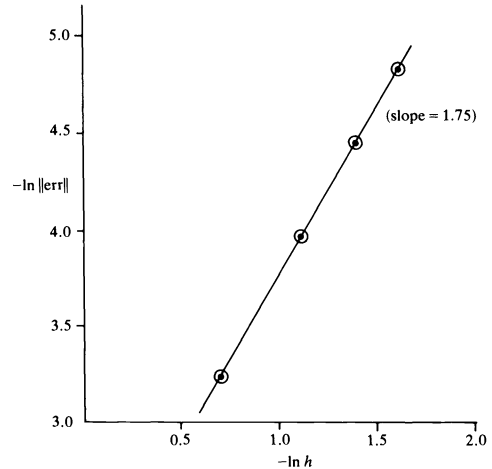
The switching times for $h = \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}$, and $u_0(x, y) = x^2 y^2$, $u_0(x, y) = xy$ are given in Tables 4 and 5 and the L^2 errors are plotted in Figs. 3 and 4.

TABLE 4
Example 2 with $v_0(x, y) = x^2 y^2$

h	Switching times					
$\frac{1}{2}$	0.218783	0.296082	0.324177	0.332284		
$\frac{1}{3}$	0.194829	0.28363	0.322531	0.33873	0.343023	
$\frac{1}{4}$	0.188473	0.283957	0.327662	0.346092	0.353744	0.355975
$\frac{1}{5}$	0.184071	0.28140	0.329418	0.351158	0.359173	0.361134

TABLE 5
Example 2 with $v_0(x, y) = xy$

h	Switching times					
$\frac{1}{2}$	0.332166	0.405045	0.417726			
$\frac{1}{3}$	0.323679	0.40205	0.421144	0.425871		
$\frac{1}{4}$	0.321137	0.402004	0.422227	0.429865	0.432316	
$\frac{1}{5}$	0.320157	0.402405	0.423293	0.431524	0.435779	0.437099

FIG 3. Example 2 with $u_0(x, y) = x^2 y^2$.FIG 4. Example 2 with $u_0(x, y) = xy$.

It appears from the numerical results presented that the rate of convergence is $O(h^2)$ for $n = 1$ and $O(h^{1.75})$ for $n = 2$, and that the general estimate (31) is not optimal in these cases. We would conjecture that it is optimal for $n \geq 2$. The higher rate of convergence for $n = 1, 2$ is probably due to the extra regularity of f_h as $h \rightarrow 0$: in these cases it would appear f_h is more regular in time than just L^∞ as $h \rightarrow 0$. Perhaps, more important than the asymptotic rate of convergence of the error, however, is the ability of the scheme presented here to capture the low frequency components of the solution on a very coarse grid. Typically, the largest values of $h = \frac{1}{2}, \frac{1}{3}$ give satisfactory results for control problems in which most of the energy is concentrated in the low frequency modes, as is often the case in practice. Furthermore, in this case the computer time required to solve such problems is minimal, typically of the order of $\frac{1}{2}$ –3 seconds CPU time on a DEC 2050. For instance, in Example 1 with constant initial temperature the L^2 error was reduced by a factor 50 on the coarsest grid ($h = \frac{1}{2}$) and by a factor of 20 for $u_0(x) = x^2$.

Finally, the most complex problem solved here, Example 2 with $h = \frac{1}{5}$ and 6 switches required ~ 40 seconds of CPU time.

Acknowledgment. I would like to acknowledge the help of George Fix and Dick MacCamy for their invaluable advice on various aspects of this paper, and to thank the University of Witwatersrand and Herriot-Watt University for their hospitality during which parts of this work were completed.

REFERENCES

- [1] S. AGMON, *Lectures on Elliptic Boundary value problems*, Van Nostrand, Princeton, NJ, 1965.
- [2] I. BABUSKA AND K. AZIZ, *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, Academic Press, New York, 1972.
- [3] Y. EGOROV, *Certain problems in the theory of optimal control*, Dokl. Akad. Nauk SSSR, 145 (1962), pp. 720–723.
- [4] A. FRIEDMAN, *Partial Differential Equations*, Holt Rinehart and Winston, New York, 1969.
- [5] ———, *Optimal control for parabolic equations*, J. Math. Anal. Appl., 18 (1967), pp. 479–491.
- [6] R. GOLDWYN, K. SRIRAM AND M. GRAHAM, *Time optimal control of a linear diffusion process*, this Journal, 5 (1967), pp. 295–308.
- [7] H. HALKIN, *A generalization of LaSalle's bang-bang principle*, this Journal, 2 (1965), pp. 199–202.
- [8] G. KNOWLES, *Time optimal control in infinite-dimensional spaces*, this Journal, 14 (1976), pp. 919–933.

- [9] ———, *Some problems in the control of distributed systems and their numerical solution*, this Journal, 15 (1977), pp. 830–840.
- [10] J. LIONS AND E. MAGENES, *Non-homogeneous Boundary Value Problems and Applications*, vols. I, II, Springer-Verlag, Heidelberg, 1972.
- [11] C. OLECH, *Extremal solutions of a control system*, J. Differential Equations 2 (1966), pp. 74–101.
- [12] L. PONTRYAGIN, V. BOLTYANSKII, R. GRAMKRELIDZE AND E. MISCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.
- [13] G. STRANG AND G. FIX, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
- [14] D. WASHBURN, *A bound on the boundary input map for parabolic equations with an application to time optimal control*, this Journal, 17 (1979), pp. 652–671.
- [15] R. TRIGGIANI, *Boundary feedback stabilization of parabolic equations*, Appl. Math. Optimization, 6 (1980), pp. 201–220.
- [16] H. O. FATTORINI, *Boundary control systems*, this Journal, 6 (1968), pp. 349–385.
- [17] A. H. SCHATZ, V. THOMÉE AND L. B. WAHLBIN, *Maximum norm stability and error estimates in parabolic finite element equations*, Comm. Pure Applied Math., 33 (1980), pp. 265–304.
- [18] J. H. BRAMBLE, A. H. SCHATZ, V. THOMÉE AND L. B. WAHLBIN, *Some convergence estimates for semi-discrete Galerkin type approximations for parabolic problems*, SIAM J. Numer. Anal., 14 (1977), pp. 218–241.
- [19] K. GLASHOFF AND N. WECK, *Boundary control of parabolic differential equations in arbitrary dimensions: Supremum norm problems*, this Journal, 14 (1976), pp. 662–681.
- [20] K. SCHITTKOWSKI, *Numerical solution of a time optimal parabolic boundary control problem*, J. Optim. Theory Appl., 27 (1979), pp. 271–290.

REGULARITY OF OPTIMAL BOUNDARY CONTROLS FOR PARABOLIC EQUATIONS, I. ANALYTICITY*

THOMAS I. SEIDMAN†

Abstract. Let (\mathbf{A}, β) be a second order elliptic operator on $\Omega \subset \mathbb{R}^d$ with a compatible boundary operator and let ϕ_* be the optimal boundary control for $\dot{u} = \mathbf{A}u + f$, $\beta u = \phi$ on $(0, T)$, $u(0) = \omega_0$ minimizing J of the form

$$J(\phi) = \int_0^T |\phi|_{\kappa}^2 + \int_0^T |u - u_T|_{\lambda}^2 + |u(T) - \omega_T|_{\mu}^2.$$

Then, if f , u_T , etc., are analytic in t on $(0, T)$ so is ϕ_* . A similar result is obtained for the infinite horizon problem.

1. Introduction. Consider, for the moment (as a model problem), the minimization of the quadratic cost criterion:

$$(1.1) \quad J = J(\phi) := \int_{\mathcal{Q}} \phi^2 + \lambda \int_{\mathcal{Q}} [u - u_T]^2 + \mu \int_{\Omega} [u(T) - \omega_T]^2$$

($0 \leq \lambda, \mu$). Here the *target trajectory* u_T and the terminal ($t = T$) *target state* ω_T are specified and u is determined as the solution of the parabolic problem

$$(1.2) \quad \dot{u} - \Delta u = f \quad \text{on } \mathcal{Q} := (0, T) \times \Omega,$$

$$(1.3) \quad u|_{\mathcal{S}} = \phi \quad \text{on } \mathcal{S} := [0, T] \times \partial\Omega,$$

$$(1.4) \quad u(0) = \omega_0 \quad \text{on } \Omega \subset \mathbb{R}^d.$$

In (1.2)–(1.4) the *initial state* ω_0 and the *source* inhomogeneity f are specified so we consider only the dependence on the *control* ϕ , i.e., $u = u(\cdot; \phi)$.

Our object in this paper is to investigate the smoothness of the minimizer ϕ_* of J under suitable assumptions. This investigation is a continuation of [12] and of certain concerns in [11] (Remark 5.4, asserting analyticity in t of ϕ_* for the case $\lambda = 0$, etc.). It was also stimulated by comparison with Lasiecka's results [3].

It is natural to view the minimization of (1.1) as minimization over ϕ in $L^2(\mathcal{S})$. As $J: L^2(\mathcal{S}) \rightarrow [0, \infty]$ need not even be lower semicontinuous, our first task is to demonstrate the existence of a unique minimizer (*optimal boundary control*) ϕ_* and to obtain some useful characterizations of it. In particular, we show that ϕ_* satisfies an equation of the form

$$(1.5) \quad (\mathbf{1} + \mathbf{T})\phi_* = \psi_0,$$

where ψ_0 is representable as w_ν (exterior normal derivative) for some solution w of another parabolic equation. The equation (1.5) and the characterization of ψ_0 can then be used to prove regularity of ϕ_* . The regularity argument proceeds as follows:

- (i) for $\varepsilon > 0$, introduce a Banach space \mathcal{Y} of $H^s(\partial\Omega)$ -valued functions, each of which is analytic in t on $(\varepsilon, T - \varepsilon)$,
- (ii) show $\mathbf{T}^2\psi_0$ is (the restriction to \mathcal{S} of) an element $\hat{\psi}_2$ of \mathcal{Y} ,

* Received by the editors April 10, 1980, and in revised form May 1, 1981. This research was partly supported by the U.S. Army Research Office under grant DAAG-29-77-G-0061 and partly by sabbatical leave from University of Maryland Baltimore County while the author was visiting at Carnegie-Mellon University, Pittsburgh, Pennsylvania.

† University of Maryland Baltimore County, Baltimore, Maryland 21228.

(iii) construct $\hat{\mathbf{T}}$ on \mathcal{Y} such that the diagram

$$(1.6) \quad \begin{array}{ccc} \mathcal{Y} & \xrightarrow{\hat{\mathbf{T}}} & \mathcal{Y} \\ \downarrow \mathbf{E} & & \downarrow \mathbf{E} \\ L^2(\mathcal{S}) & \xrightarrow{\mathbf{T}} & L^2(\mathcal{S}) \end{array}$$

is commutative (\mathbf{E} a suitable embedding) and show $\hat{\mathbf{T}}$ is compact,

(iv) conclude that -1 cannot be in the spectrum of $\hat{\mathbf{T}}$ and, finally,

(v) use this to show that

$$(1.7) \quad \phi_* = \psi_0 - \mathbf{E}\hat{\mathbf{T}}[\hat{\psi}_0 - \hat{\mathbf{T}}\hat{\psi}_0 + (\mathbf{1} + \hat{\mathbf{T}})^{-1}\hat{\psi}_2]$$

is $H^s(\partial\Omega)$ -analytic on $(\varepsilon, T - \varepsilon)$ with $\varepsilon > 0$ arbitrary.

This argument gives a very strong regularity result—analyticity in t —on the *open* interval $(0, T)$. No results are given here for the closed interval $[0, T]$ (i.e., asymptotic as $t \rightarrow 0+$, $T-$); such considerations are deferred to another paper, to be concerned with possibly time-dependent problems and the relation of regularity with convergence rates of numerical approximations. Meanwhile, note that [3] shows (for $\mu = 0$) that φ_* is in $L^2([0, T] \rightarrow H^{1/2}(\partial\Omega))$ while [13] shows, for the more regular case of Neumann control, that, e.g., φ_* is in $C([0, T] \rightarrow H^s(\partial\Omega))$ for any $s < 2$ (any $s < 3$ if $\mu = 0$), given smooth data while [4] gives φ_* in $H^{3/2-\varepsilon, 3/4-\varepsilon}(\mathcal{S})$.

Although semigroup methods are used here to obtain analyticity only for the autonomous case, the next section is presented more generally, as this involves no particular difficulty and seems to be of intrinsic interest and of possible future applicability.

2. Formulation. Somewhat more generally than (1.2)–(1.4), we consider the parabolic partial differential equation

$$(2.1) \quad \dot{u} - \mathbf{A}u = f \quad \text{on } \mathcal{Q} := (0, T) \times \Omega,$$

with the boundary condition

$$(2.2) \quad \beta u = \phi \quad \text{on } \mathcal{S} := (0, T) \times \partial\Omega$$

and the initial condition

$$(2.3) \quad u(0) = \omega_0 \quad \text{on } \Omega.$$

Here, Ω is a bounded region in \mathbb{R}^d with “sufficiently smooth” boundary $\partial\Omega$. Next \mathbf{A} is a second order, uniformly elliptic operator

$$(2.4) \quad \mathbf{A}: \omega \mapsto \left[\sum_{j,k} a_{j,k} \frac{\partial^2 \omega}{\partial x_j \partial x_k} + \sum_k b_k \frac{\partial \omega}{\partial x_k} + c\omega \right] \quad \text{in } \Omega,$$

with “sufficiently smooth” (real-valued) coefficients $\{a_{j,k}, b_k, c\}$. Finally, β is an associated *boundary operator* of the usual form

$$(2.5) \quad \beta: \omega \mapsto \alpha\omega + \beta\omega_\nu \quad \text{at } \partial\Omega,$$

where ω_ν is the *exterior conormal derivative*

$$\omega_\nu := \frac{\partial \omega}{\partial \nu} := \sum_{j,k} a_{j,k} \frac{\partial \omega}{\partial x_k} n_j \quad (\mathbf{n} := \text{unit exterior normal to } \partial\Omega),$$

with “sufficiently smooth” (real) coefficients, normalized so that $\alpha^2 + \beta^2 \equiv 1$. We

distinguish between—and admit—the three cases (*Dirichlet*, *Neumann* and *Robin* boundary conditions):

$$(2.6) \quad \begin{array}{ll} \text{case D: } \alpha \equiv 1, \beta \equiv 0, & \text{order: } \hat{m} = 0, \\ \text{case N: } \alpha \equiv 0, \beta \equiv 1, & \text{order: } \hat{m} = 1, \\ \text{case R: } \beta \text{ never } 0, & \text{order: } \hat{m} = 1. \end{array}$$

The results obtained will be quite similar in each case but the arguments are somewhat more delicate for case D. (It seems likely that the arguments used for case D could also be applied, with some considerable technical complication, to cases in which, e.g., β might vanish on *part* of $\partial\Omega$, but we exclude such considerations.)

The assumed “sufficient smoothness” of $\partial\Omega$ and the coefficients is, to begin with, presumed to imply existence of a unique solution u of (2.1)–(2.3)—in some appropriate sense to be specified later—for suitable f, ϕ, ω_0 . In later sections, taking (\mathbf{A}, β) to be autonomous we make the technical requirement that the *elliptic* problem

$$(2.7) \quad \mathbf{A}\omega = f \quad \text{in } \Omega, \quad \beta\omega = g \quad \text{at } \partial\Omega$$

has a unique solution for “arbitrary” (f, g) , i.e., without consistency conditions. (This would appear to eliminate such a standard problem as case N for the usual heat equation but actually involves no loss of generality as one can always consider $e^{-Ct}u(t)$ instead, which adds the constant C to \mathbf{A} . The resulting modification of the cost functional J to be considered will not affect the conditions which will be imposed.) A further implication of the smoothness assumptions will be, e.g., the equivalence of $|\omega|_s$ and $|\mathbf{A}\omega|_{s-2}$ (here $|\cdot|_s$ denotes the $H^s(\Omega)$ -norm, etc.) for ω such that $\beta\omega = 0$ and suitable s (compare Theorems 2.5.4 and 2.8.3 of [6]); see Remark 6.6 below.

Somewhat more generally than (1.1) we introduce seminorms $\|\cdot\|_{\kappa}, \|\cdot\|_{\lambda}, \|\cdot\|_{\mu}$ for functions on $\mathcal{S}, \mathcal{Q}, \Omega$ respectively, and set

$$(2.8) \quad J = J(\phi) := \|\phi\|_{\kappa}^2 + \|u - u_T\|_{\lambda}^2 + \|u(T) - \omega_T\|_{\mu}^2,$$

where u_T, ω_T are specified and u is determined from ϕ by (2.1)–(2.3). We will be assuming that $\|\cdot\|_{\kappa}$ is equivalent to the usual $L^2(\mathcal{S})$ -norm, that $\|\cdot\|_{\lambda}$ is dominated by the usual $L^2(\mathcal{Q})$ -norm and either that $\|\cdot\|_{\mu}$ is dominated by the usual $L^2(\Omega)$ -norm (*regular* case) or else—indicated by writing symbolically: $\mu = \infty$ —that we require *exact control* to ω_T , so that

$$\|u(T) - \omega_T\|_{\infty} := \begin{cases} 0, & u(T) = \omega_T, \\ \infty, & u(T) \neq \omega_T. \end{cases}$$

See Remark 6.1 below, however.

If we have “smooth” nonnegative functions κ, λ, μ on $\bar{\mathcal{S}}, \bar{\mathcal{Q}}, \bar{\Omega}$, respectively (with $1/\kappa$ also bounded), then we may define corresponding multiplication operators κ, λ, μ on $L^2(\mathcal{S})$, etc.; by a slight abuse of notation we also write $\kappa = \kappa(t)$ for the operator on $L^2(\partial\Omega)$ defined by multiplication by $\kappa(t, \cdot)$ (for $0 \leq t \leq T$), and similarly for $\lambda = \lambda(t)$. The situation of interest, then, is

$$(2.9) \quad \begin{aligned} \|\phi\|_{\kappa}^2 &:= \int_0^T \langle \phi(t), \kappa(t)\phi(t) \rangle_{L^2(\partial\Omega)} dt, \\ \|u\|_{\lambda}^2 &:= \int_0^T \langle u(t), \lambda(t)u(t) \rangle_{L^2(\Omega)} dt, \\ \|\omega\|_{\mu}^2 &:= \langle \omega, \mu\omega \rangle_{L^2(\Omega)} \quad \text{if } \mu \neq \infty. \quad (\text{For } \mu = \infty, \text{ as above.}) \end{aligned}$$

Note that the assumptions on the functions κ, λ, μ ensure that one has

$$(2.10) \quad \|\cdot\|_{\kappa} \text{ is equivalent to the usual } L^2(\mathcal{S})\text{-norm; } \kappa \text{ is continuously invertible.} \\ \kappa(t)^{-1/2}, \lambda(t) \text{ are positive (semi-) definite operators on } L^2(\partial\Omega), L^2(\Omega) \text{ and,} \\ \text{for suitable } \sigma, \sigma', \text{ they can be interpreted as bounded (uniformly in } t \text{ for} \\ 0 \leq t \leq T) \text{ operators on } H^{\sigma}(\partial\Omega), H^{\sigma'}(\Omega). \text{ Similarly (unless } \mu = \infty), \mu \text{ is} \\ \text{positive (semi-) definite on } L^2(\Omega).$$

We will assume (2.9)–(2.10) without necessarily taking κ, λ, μ to be multiplication operators.

It will be convenient to let u_0 be the solution of (2.1)–(2.3) with $\phi = 0$ so that

$$(2.11) \quad \dot{u}_0 - \mathbf{A}u_0 = f \quad \text{on } \mathcal{Q}, \quad \mathbf{B}u_0 = 0 \quad \text{on } \mathcal{S}, \quad u_0(0) = \omega_0$$

and to introduce the solution operator \mathbf{B} for (2.1)–(2.3) with $f = 0, \omega_0 = 0$ so $u := \mathbf{B}\phi$ means that

$$(2.12) \quad \dot{u} = \mathbf{A}u \quad \text{on } \mathcal{Q}, \quad \mathbf{B}u = \phi \quad \text{on } \mathcal{S}, \quad u(0) = 0.$$

By linearity the solution u of (2.1)–(2.3) then has the form: $u = u_0 + \mathbf{B}\phi$. Observe that in any of the cases D, N, R of (2.6) we have that

$$(2.13) \quad \mathbf{B} \text{ is continuous from } L^2(\mathcal{S}) \text{ to } L^2(\mathcal{Q}).$$

(This follows from results of [7, pp. 78 ff.]; see Remark 6.6, below.) We also introduce the (possibly unbounded) operator \mathbf{B}_T given by

$$(2.14) \quad \mathbf{B}_T\phi := [\mathbf{B}\phi](T) \quad \text{with } \mathfrak{D}(\mathbf{B}_T) := \{\phi \in L^2(\mathcal{S}): [\mathbf{B}\phi](T) \in L^2(\Omega)\}.$$

Note that $[\mathbf{B}\phi](t)$ is always a well-defined (smooth) function on Ω for each $t > 0$, including $t = T$, by interior regularity for (2.1) under mild smoothness conditions on the coefficients in (2.4). In case D, however, $[\mathbf{B}\phi](T)$ need not be well behaved near $\partial\Omega$ and need not be square-integrable on Ω —indeed, the results of [7] suggest that then $[\mathbf{B}\phi](t)$ might better be sought in $[H^{1/2}(\Omega)]^*$ than in $L^2(\Omega)$.

LEMMA 2.15. *In case D of (2.6) the operator \mathbf{B}_T is densely defined and closed. In cases N and R one has $\mathfrak{D}(\mathbf{B}_T) = L^2(\mathcal{S})$ and \mathbf{B}_T is continuous to $L^2(\Omega)$ (indeed, to $H^{1/2}(\Omega)$).*

Proof. For case D, closure follows from the continuity of $\phi \mapsto [\mathbf{B}\phi](T)|_{\mathcal{K}}$ from $L^2(\mathcal{S})$ to $C(\mathcal{K})$ for arbitrary closed $\mathcal{K} \subset \Omega$ which, in turn, follows from (2.13) and interior regularity. Density of $\mathfrak{D}(\mathbf{B}_T)$ follows, e.g., from the fact that \mathbf{B} takes $L^{\infty}(\mathcal{S})$ to $L^{\infty}(\mathcal{Q})$ by the maximum principle so $L^{\infty}(\mathcal{S}) \subset \mathfrak{D}(\mathbf{B}_T)$. For cases N and R one has $\hat{m} = 1$ and, for ϕ in $L^2(\mathcal{S})$, [7] gives $u = \mathbf{B}\phi$ in $H^{3/2, 3/4}(\mathcal{Q})$ so the trace on $\{T\} \times \Omega$ is continuous to $H^{1/2}(\Omega)$ by [7, Thm. 4.2.1]. \square

Note that the consideration of case D with $\mathbf{0} \neq \mu$ is in “violation” of the structures of [7, Remark 6.1.4, p. 160].

THEOREM 1. *Let u_0, u_T, ω_T, μ be such that J given by (2.8), (2.9) is not identically ∞ , i.e., such that*

$$(2.16) \quad \mathfrak{D}_J := \{\phi \in L^2(\mathcal{S}): J(\phi) < \infty\} \quad \text{is nonempty.}$$

Then there exists a unique minimizer ϕ_ of J . One then has*

$$(2.17) \quad (\kappa + \mathbf{B}^*\lambda\mathbf{B})\phi_* = \mathbf{B}^*z_0 + \tilde{\psi} =: \psi_0$$

with

$$(2.18) \quad z_0 = \lambda(u_T - u_0) \quad \text{and} \quad \tilde{\psi} \in \mathfrak{M} := \mathfrak{N}(\mathbf{B}_T)^{\perp}.$$

If μ is bounded on $L^2(\Omega)$, then $\tilde{\psi}$ is actually in $\mathfrak{N}(\mathbf{B}_T^*)$: the optimally controlled trajectory $u_* = u_0 + \mathbf{B}\phi_*$ is then such that

$$(2.19) \quad \omega_* := \mu[\omega_T - u_*(T)] \text{ is in } \mathfrak{D}(\mathbf{B}_T^*) \text{ and } \psi_0 = \mathbf{B}_T^* \omega_* + \mathbf{B}^* z_0.$$

Proof. First consider the case of exact control: $\mu = \infty$. Given any fixed ϕ_* in \mathfrak{D}_J , one has $(\phi_* + h)$ in \mathfrak{D}_J if and only if h is in $\mathfrak{N}(\mathbf{B}_T)$ so $[u(T) - \omega_T]$ continues to vanish. For such h and any r in \mathbb{R} ,

$$(2.20) \quad \begin{aligned} J(\phi_* + rh) &= \|\phi_* + rh\|_{\kappa}^2 + \|[u_* - u_T] + r\mathbf{B}h\|_{\lambda}^2 \\ &= J(\phi_*) + r^2[\|h\|_{\kappa}^2 + \|\mathbf{B}h\|_{\lambda}^2] + 2r[\langle \phi_*, h \rangle_{\kappa} + \langle u_* - u_T, \mathbf{B}h \rangle_{\lambda}], \end{aligned}$$

so $h \mapsto J(\phi_* + h)$ is a coercive, strictly convex, continuous function on the closed (by Lemma 2.15) subspace $\mathfrak{N}(\mathbf{B}_T)$ and hence uniquely attains its minimum. Thus, J uniquely attains its minimum on \mathfrak{D}_J and, from (2.20), this minimum will be ϕ_* (if and) only if

$$\begin{aligned} 0 &= \langle \phi_*, h \rangle_{\kappa} + \langle u_* - u_T, \mathbf{B}h \rangle_{\lambda} \\ &= \langle \kappa \phi_*, h \rangle_{L^2(\mathcal{S})} + \langle \lambda[u_0 - u_T + \mathbf{B}\phi_*], \mathbf{B}h \rangle_{L^2(\Omega)} \end{aligned}$$

for every h in $\mathfrak{N}(\mathbf{B}_T)$. Thus,

$$\psi := (\kappa + \mathbf{B}^* \lambda \mathbf{B}) \phi_* - \mathbf{B}^* \lambda [u_0 - u_T]$$

must be orthogonal to $\mathfrak{N}(\mathbf{B}_T)$ —which is just (2.17), (2.18).

Now suppose μ is bounded on $L^2(\Omega)$. Introduce the Hilbert space

$$\hat{\mathfrak{H}} := \mathfrak{H}_{\kappa} \times \mathfrak{H}_{\lambda} \times \mathfrak{H}_{\mu},$$

where \mathfrak{H}_{κ} is $L^2(\mathcal{S})$ with $\|\cdot\|_{\kappa}$, etc. (The definitions of the Hilbert spaces \mathfrak{H}_{λ} , \mathfrak{H}_{μ} may involve factoring out the null spaces of $\|\cdot\|_{\lambda}$, $\|\cdot\|_{\mu}$ —which could be nontrivial as we only assumed in (2.10) that λ , μ were semidefinite.) We define

$$\hat{\mathbf{B}}_T: L^2(\mathcal{S}) \supset \mathfrak{D}_* \rightarrow \mathfrak{H}_{\mu}: \phi \mapsto [\mathbf{B}\phi](T)$$

with, of course,

$$\mathfrak{D}_* := \{\phi \in L^2(\mathcal{S}): [\mathbf{B}\phi](T) \in \mathfrak{H}_{\mu}\}.$$

Observe that

$$\|[\mathbf{B}\phi](T)\|_{\mu} = \|\mu^{1/2}[\mathbf{B}\phi](T)\|_{L^2(\mathcal{S})},$$

so for μ bounded we have $\mathfrak{D}_* \supset \mathfrak{D}(\mathbf{B}_T)$, whence, by (2.15),

$$(2.21) \quad \mathfrak{D}_* \text{ is dense in } L^2(\mathcal{S}).$$

(The conclusions would follow if one had (2.16), (2.21), even if μ were unbounded on $L^2(\Omega)$; see Remark 6.1, below.) Again given any fixed ϕ_* in \mathfrak{D}_J , we now set

$$\Phi_* := [\phi_*, u_* - u_T, u_*(T) - \omega_T]$$

and observe that our definitions are such that Φ_* is in $\hat{\mathfrak{H}}$ and $J(\phi_*) = \|\Phi_*\|_{\hat{\mathfrak{H}}}^2$. If $\phi_* + \tilde{h} =: \tilde{\phi}$ is also in \mathfrak{D}_J we have

$$\tilde{\Phi} := [\tilde{\phi}, \tilde{u} - u_T, \tilde{u}(T) - \omega_T] \text{ in } \hat{\mathfrak{H}},$$

so that

$$\tilde{\Phi} - \Phi_* = [\tilde{h}, \tilde{u} - u_*, \tilde{u}(T) - u_*(T)] \text{ is in } \hat{\mathfrak{H}}.$$

But $\tilde{u} - u_* = \mathbf{B}\tilde{h}$ and $\tilde{u}(T) - u_*(T) = [\mathbf{B}\tilde{h}](T)$ so we have \tilde{h} in \mathfrak{D}_* . Conversely, if \tilde{h} is in \mathfrak{D}_* , then

$$\mathbf{H}h := [\tilde{h}, \mathbf{B}\tilde{h}, \hat{\mathbf{B}}_T\tilde{h}] \text{ is in } \hat{\mathfrak{H}}$$

and $\tilde{\Phi} = \Phi_* + \mathbf{H}\tilde{h}$ is also in $\hat{\mathfrak{H}}$, so that $\tilde{\phi} = \phi_* + \tilde{h}$ is in \mathfrak{D}_J . Let \mathfrak{H}_J be \mathfrak{D}_* with the norm $\|h\|_J = \|\mathbf{H}h\|_{\hat{\mathfrak{H}}}$. Then $\mathfrak{D}_J = \phi_* + \mathfrak{H}_J$ and

$$J(\phi_* + h) = \|\phi_* + \mathbf{H}h\|_{\hat{\mathfrak{H}}}^2, \quad \text{for } h \in \mathfrak{H}_J,$$

so $h \mapsto J(\phi_* + h)$ is clearly coercive, strictly convex and continuous on \mathfrak{H}_J and must therefore uniquely attain its minimum on \mathfrak{H}_J , whence J uniquely attains its minimum on $L^2(\mathcal{S})$. For r in \mathbb{R} and any h in $\mathfrak{H}_J = \mathfrak{D}_*$, one has

$$J(\phi_* + rh) = J(\phi_*) + 2r\langle \Phi_*, \mathbf{H}h \rangle_{\hat{\mathfrak{H}}} + r^2\|h\|_J^2,$$

so ϕ_* will be the minimizer of J (if and) only if $\langle \Phi_*, \mathbf{H}h \rangle_{\hat{\mathfrak{H}}} = 0$ for all such h . Thus, for every h in the dense set $\mathfrak{D}(\mathbf{B}_T) \subset \mathfrak{D}_* \subset L^2(\mathcal{S})$,

$$\begin{aligned} \langle \mu[u_*(T) - \omega_T], \mathbf{B}_T h \rangle_{L^2(\Omega)} &= \langle u_*(T) - \omega_T, \hat{\mathbf{B}}_T h \rangle_{\mu} \\ (2.22) \quad &= -\langle \phi_*, h \rangle_{\kappa} - \langle u_* - u_T, \mathbf{B}h \rangle_{\lambda} \\ &= -\langle (\kappa + \mathbf{B}^* \lambda \mathbf{B})\phi_* - \mathbf{B}^* \lambda [u_T - u_0], h \rangle_{L^2(\mathcal{S})}. \end{aligned}$$

As the final expression in (2.22) depends continuously on h topologized in $L^2(\mathcal{S})$, so does the first, which means that $\mu[u_*(T) - \omega_T]$ is in the domain of \mathbf{B}_T^* and we may set

$$\tilde{\psi} := \mathbf{B}_T^* \mu[u_*(T) - \omega_T] \in \mathfrak{M}.$$

Then (2.22) holding for h in a dense set gives (2.17), (2.19). \square

It is clear that application of this theorem will require a more detailed examination of \mathbf{B}^* , \mathbf{B}_T^* and \mathfrak{M} in terms of partial differential equations. We proceed formally, assuming that $\partial\Omega$ and the coefficients are “smooth enough” to justify the manipulations for later application and interpretation.

For fixed $t(0 < t < T)$, let u, v be “smooth enough” functions on Ω . Using the divergence theorem,

$$\begin{aligned} \int_{\Omega} [\mathbf{A}u]v &= \int_{\Omega} \sum_{j,k} a_{jk} \frac{\partial^2 u}{\partial x_j \partial x_k} v + \int_{\Omega} \sum_k b_k \frac{\partial u}{\partial x_k} v + \int_{\Omega} cuv \\ (2.23) \quad &= \int_{\Omega} u \sum_{j,k} \frac{\partial^2}{\partial x_j \partial x_k} (a_{jk}v) - \int_{\Omega} u \sum_k \frac{\partial}{\partial x_k} (b_k v) + \int_{\Omega} cuv \\ &\quad + \int_{\partial\Omega} \sum_{j,k} a_{jk} \frac{\partial u}{\partial x_j} n_k v - \int_{\partial\Omega} u \sum_{j,k} a_{jk} \frac{\partial v}{\partial x_k} n_j + \int_{\partial\Omega} \left(\sum_j \left[b_j - \sum_k \frac{\partial a_{jk}}{\partial x_k} \right] n_j \right) uv. \end{aligned}$$

Now let \mathbf{A}^* be the *formal adjoint* of \mathbf{A} , given by

$$\begin{aligned} \mathbf{A}^*: v &\mapsto \sum_{j,k} \frac{\partial^2}{\partial x_j \partial x_k} (a_{jk}v) + \sum_k \frac{\partial}{\partial x_k} (b_k v) + cv \\ (2.24) \quad &= \sum_{j,k} a_{jk}^* \frac{\partial^2 v}{\partial x_j \partial x_k} + \sum_k b_k^* \frac{\partial v}{\partial x_k} + c^* v \quad \text{in } \Omega, \end{aligned}$$

where

$$a_{jk}^* := a_{jk} (= a_{kj}), \quad b_k^* := 2 \sum_j \frac{\partial a_{jk}}{\partial x_j} - b_k, \quad c^* := \sum_{j,k} \frac{\partial^2 a_{jk}}{\partial x_j \partial x_k} + \sum_k \frac{\partial b_k}{\partial x_k} + c,$$

and let $\hat{\mathbf{B}}$ and \mathbf{B}' be the boundary operators

$$(2.25) \quad \begin{aligned} \hat{\mathbf{B}}: v &\mapsto (\alpha - \gamma\beta)v + \beta v_\nu \quad \text{at } \partial\Omega, \\ \mathbf{B}': v &\mapsto (\beta + \gamma\alpha)v - \alpha v_\nu \quad \text{at } \partial\Omega, \end{aligned}$$

with

$$\gamma := [\mathbf{b} - (\nabla \mathbf{a})] \cdot \mathbf{n} := \sum_j \left(b_j - \sum_k \frac{\partial a_{jk}}{\partial x_k} \right) n_j \quad \text{on } \partial\Omega.$$

We note that \mathbf{B}' will *only* be applied to v for which $\hat{\mathbf{B}}v = 0$; in this case one easily sees that

$$(2.26) \quad \mathbf{B}': v \mapsto \begin{cases} -v_\nu, & \text{if } \beta \equiv 0, \\ (1/\beta)v, & \text{if } \beta \neq 0, \end{cases} \quad \text{if } \hat{\mathbf{B}}v = 0.$$

Note that the order of $\hat{\mathbf{B}}$ is always the same \hat{m} as in (2.6) while the order of \mathbf{B}' (as given by (2.26) for $\hat{\mathbf{B}}v = 0$) is m' , given by

$$(2.27) \quad m' = 1 \quad \text{in case D}, \quad m' = 0 \quad \text{in cases N, R}.$$

(If \mathbf{A} were in divergence form, corresponding to having $b_k = \sum_j (\partial a_{jk} / \partial x_j)$ for each k , then one would have $\mathbf{A}^* = \mathbf{A}$ and $\gamma = 0$ so $\hat{\mathbf{B}} = \mathbf{B}$.) After manipulation of (2.4), (2.24)–(2.26), one then has from (2.23) the *fundamental identity*

$$(2.28) \quad \int_{\Omega} [\mathbf{A}u]v = \int_{\Omega} u[\mathbf{A}^*v] + \int_{\partial\Omega} (\mathbf{B}u)(\mathbf{B}'v) \quad \text{if } \hat{\mathbf{B}}v = 0.$$

Suppose $u = \mathbf{B}\phi$ so u satisfies (2.12) and let w be the solution of

$$(2.29) \quad -\dot{w} = \mathbf{A}^*w + z \quad \text{on } \mathcal{Q},$$

$$(2.30) \quad \hat{\mathbf{B}}w = 0 \quad \text{on } \mathcal{S},$$

$$(2.31) \quad w(T) = 0.$$

Then

$$\begin{aligned} \langle \mathbf{B}\phi, z \rangle_{L^2(\mathcal{Q})} &= \int_{\mathcal{Q}} uz = - \int_{\mathcal{Q}} u\dot{w} - \int_{\mathcal{Q}} u[\mathbf{A}^*w] \\ &= - \int_{\Omega} uw|_0^T + \int_{\mathcal{Q}} \dot{u}w - \int_{\mathcal{Q}} u[\mathbf{A}^*w] = 0 + \int_{\mathcal{S}} \phi(\mathbf{B}'w), \end{aligned}$$

using integration by parts over $[0, T]$, applying the conditions $u(0) = 0 = w(T)$ and, finally, using (2.28) and the equations. Thus for z in $L^2(\mathcal{Q})$ we have

$$(2.32) \quad \mathbf{B}^*z = \mathbf{B}'w \quad \text{with } w \text{ given by (2.29)–(2.31)}.$$

Next, suppose $u = \mathbf{B}\phi$ with ϕ in $\mathfrak{D}(\mathbf{B}_T)$ and, for a suitable function \hat{w} on Ω , let v be the solution of

$$(2.33) \quad -\dot{v} = \mathbf{A}^*v \quad \text{on } \mathcal{Q},$$

$$(2.34) \quad \hat{\mathbf{B}}v = 0 \quad \text{on } \mathcal{S},$$

$$(2.35) \quad v(T) = \hat{w}.$$

Then

$$\begin{aligned}\langle \mathbf{B}_T \phi, \hat{\omega} \rangle_{L^2(\Omega)} &= \int_{\Omega} u(T)v(T) = \int_{\Omega} uv|_0^T \\ &= \int_{\mathcal{Q}} (uv)' = \int_0^T \left(\int_{\Omega} [\mathbf{A}u]v - \int_{\Omega} u[\mathbf{A}^*v] \right) = \int_{\mathcal{S}} \phi(\beta'v),\end{aligned}$$

using (2.28) and the equations. The final term of (2.36) (and so also the first term) depends continuously on ϕ topologized in $L^2(\mathcal{S})$ if and only if $\beta'v$ is in $L^2(\mathcal{S})$ so that

$$\begin{aligned}(2.36) \quad \beta_T^* \hat{\omega} &= \beta'v \text{ with } v \text{ given by (2.33)–(2.35) for} \\ \hat{\omega} \text{ in } \mathfrak{D}(\mathbf{B}_T^*) &= \{\hat{\omega} \in L^2(\Omega): \beta'v \in L^2(\mathcal{S})\}.\end{aligned}$$

From Lemma 2.15 one has \mathbf{B}_T^* bounded so $\mathfrak{D}(\mathbf{B}_T^*) = L^2(\Omega)$ for cases N and R; one should even be able to extend \mathbf{B}_T^* as continuous from $[H^{1/2}(\Omega)]^*$. Reversing these considerations, we see that for case D one obtain v in $H^{s+1, (s+1)/2}(\mathcal{Q})$ for $\hat{\omega}$ in $H^s(\Omega)$ with $s = \frac{1}{2} +$ and then $\beta'v$ would be in, say, $L^2((0, T) \rightarrow H^{s-1/2}(\partial\Omega)) \supset L^2(\mathcal{S})$ so \mathbf{B}_T^* will be continuous if interpreted as a map from $H^s(\Omega)$ to $L^2(\mathcal{S})$ for $s > \frac{1}{2}$. Thus, although we have been treating \mathbf{B}_T and \mathbf{B}_T^* as unbounded operators in case D, we can supplement Lemma 2.15 with the observation that

$$(2.37) \quad \begin{array}{l} \text{In case D we can re-interpret } \mathbf{B}_T \text{ as continuous from } L^2(\mathcal{S}) \\ \text{to } [H_0^s(\Omega)]^* \text{ (any } s > \frac{1}{2}), \end{array}$$

as soon as we notice that \mathbf{B}_T has been maximally defined, so

$$(2.38) \quad \mathbf{B}_T^{**} = \mathbf{B}_T \quad \text{and} \quad \mathfrak{M} := \mathfrak{N}(\mathbf{B}_T)^\perp = \overline{\mathfrak{N}(\mathbf{B}_T^*)}.$$

Clearly, (2.38) holds for cases N and R as well.

For bounded μ the representation (2.17)–(2.19) is complete, but for the case of exact control ($\mu = \infty$) one needs a further examination of \mathfrak{M} , for which we must impose a nullcontrollability assumption.

DEFINITION 2.39. Call (\mathbf{A}, β) *nullcontrollable on arbitrary intervals* in $(0, T]$ if for arbitrary τ in $[0, T)$ and arbitrarily small $\varepsilon > 0$ there exists, for any ω_0 in $L^2(\Omega)$, a ϕ_0 in $L^2((\tau, \tau + \varepsilon) \times \partial\Omega)$ such that $u(\tau + \varepsilon) = 0$ for the solution u of

$$\dot{u} = \mathbf{A}u \quad \text{on } (\tau, \tau + \varepsilon) \times \Omega, \quad \beta u = \phi_0 \quad \text{on } (\tau, \tau + \varepsilon) \times \partial\Omega, \quad u(\tau) = \omega_0.$$

In the autonomous case it is clearly sufficient to consider only $\tau = 0$ and ε arbitrarily small in Definition 2.39.

We now obtain a representation for elements of \mathfrak{M} . This corresponds to half of [11, Thm. 5.3] with a more general \mathbf{A} ; compare [11, Remarks 5.5]. This additional generality is not truly applicable in the present state of our knowledge since the nullcontrollability condition has only been verified—so far—for classes of settings to which [11, Remark 5.3] applies directly. For completeness of our discussion of Theorem 1 and in anticipation of more general verification we include the generalization here. For technical convenience we require that β be autonomous while still permitting the coefficients of \mathbf{A} to be time dependent. Note that this requirement on β need only be imposed for Robin conditions: it is automatic for cases D and N.

LEMMA 2.40. *Let $\partial\Omega$ and the coefficients be sufficiently smooth with β autonomous. If (\mathbf{A}, β) is nullcontrollable on arbitrary intervals of $(0, T]$, then each element ψ of \mathfrak{M} has the form:*

$$(2.41) \quad \psi = \beta'v \quad \text{with } v \text{ satisfying (2.30), (2.31), } \quad \psi \in L^2(\mathcal{S}).$$

Proof. By (2.38), ψ in \mathfrak{M} means existence of a sequence $\{\psi_j\}$ in $\mathfrak{R}(\mathbf{B}_T^*)$ such that $\psi_j \rightarrow \psi$ in $L^2(\mathcal{S})$. By (2.36) one then has $\psi_j =: \mathbf{B}_T^* \hat{\omega}_j = \boldsymbol{\beta}' v_j$ with v_j satisfying (2.30), (2.31). Select any $\varepsilon > 0$ and take $\tau = T - \varepsilon$ in Definition 2.39, setting

$$\mathcal{Q}' := (\tau, T) \times \Omega, \quad \mathcal{S}' := (\tau, T) \times \partial\Omega.$$

Then, by the duality theorem [11, Thm. 2.1], the nullcontrollability for (τ, T) implies continuity of the map

$$(2.42) \quad \mathbf{P}_\tau: \boldsymbol{\beta}' v|_{\mathcal{S}'} \mapsto v(\tau): \mathfrak{M}_\tau \rightarrow L^2(\Omega).$$

Here \mathfrak{M}_τ corresponds to \mathfrak{M} but on \mathcal{S}' ; clearly \mathfrak{M}_τ contains the restrictions to \mathcal{S}' of elements of \mathfrak{M} since, as with (2.38), \mathfrak{M}_τ is the closure in $L^2(\mathcal{S}')$ of the set of restrictions to \mathcal{S}' of elements of $\mathfrak{R}(\mathbf{B}_T^*)$. Note that application of [11, Thm. 2.1] requires \mathbf{B}_T to be continuous to the space \mathfrak{B} ; as this space is fairly arbitrary, it can be taken to be, say, $[H^{3/4}(\Omega)]^*$ noting Lemma 2.15 and 2.37. We must also have continuity from $L^2(\Omega)$ to \mathfrak{B} for the solution map (at $t = T$) of the pure initial value problem but this is clear by the usual “evolution operator” theory of (2.1)–(2.3) with $f, \phi = 0$ as an abstract ordinary differential equation on $L^2(\Omega)$.

Thus, convergence of ψ_j in $L^2(\mathcal{S})$ implies the convergence in $L^2(\Omega)$:

$$v_j(\tau) = \mathbf{P}_\tau[\psi_j|_{\mathcal{S}'}] \rightarrow \omega_\tau \in L^2(\Omega).$$

Now let $v = v_\tau$ be the solution of

$$-\dot{v} = \mathbf{A}^* v \quad \text{on } \mathcal{Q}_\tau, \quad \hat{\boldsymbol{\beta}} v = 0 \quad \text{on } \mathcal{S}_\tau, \quad v(\tau) = \omega_\tau$$

where $\mathcal{Q}_\tau := (0, \tau) \times \Omega$, $\mathcal{S}_\tau := (0, \tau) \times \partial\Omega$. For $\boldsymbol{\beta}$ autonomous and sufficiently smooth coefficients, [14, Thm. 1.1] applies to give the estimate

$$(2.43) \quad |v(\sigma)|_{H^2(\Omega)} \leq C |\mathbf{A}(\sigma)v(\sigma)|_{L^2(\Omega)} \leq \frac{C'}{\tau - \sigma} |v(\tau)|_{L^2(\Omega)}$$

for any v satisfying (2.33), (2.34) and $0 < \sigma < \tau$. Thus, $v_j(\sigma) \rightarrow v(\sigma)$ in $H^2(\Omega)$, uniformly in σ on $[0, \sigma']$ if $\sigma' < \tau$. It follows that

$$\psi_j = \boldsymbol{\beta}' v_j \rightarrow \boldsymbol{\beta}' v \quad \text{in } L^\infty([0, \sigma'] \rightarrow H^{1/2}(\partial\Omega)),$$

so $\boldsymbol{\beta}' v$ must coincide on $\mathcal{S}_{\sigma'}$ with $\psi = \lim \psi_j$ and so on $\mathcal{S}_{\sigma'}$, as $\sigma' < \tau$ was arbitrary. It is easy to see that $v_{\tau'}$ is a restriction of v_τ for $0 < \tau' < \tau < T$ so there must be a (maximally defined) function v on \mathcal{Q} of which each v_τ is a restriction. Clearly v satisfies (2.33), (2.34) on \mathcal{Q} and $\psi = \boldsymbol{\beta}' v$ on all of \mathcal{S} . \square

This construction of v gives uniqueness—indeed,

$$v(\tau) = \mathbf{P}_\tau[\psi|_{(\tau, T) \times \partial\Omega}], \quad 0 \leq \tau < T,$$

for any ψ in \mathfrak{M} . Note that nothing can be said about $v(T-)$ and (2.33), (2.34) is to be interpreted for v as holding when restricted to $(0, \tau)$ for each $\tau < T$. For the *completely autonomous* case we also observe that the converse of Lemma 2.40 holds: if ψ satisfies (2.41), then ψ is in \mathfrak{M} (proof as in [11, Remark 5.3]).

With (2.32), (2.36) and Lemma 2.40 in hand, we return to the representation (2.17). We now have that

$$(2.44) \quad \psi_0 = \boldsymbol{\beta}' w_0,$$

where w_0 satisfies (2.34) and

$$(2.45) \quad -\dot{w}_0 = \mathbf{A}^* w_0 + z_0 \quad \text{on } \mathcal{Q},$$

$$(2.46) \quad z_0 := \boldsymbol{\lambda}[u_T - u_0].$$

If $\boldsymbol{\mu}$ is bounded on $L^2(\Omega)$, then (2.19) permits us to adjoin to (2.34), (2.45) the initial condition

$$(2.47) \quad w_0(T) = \boldsymbol{\mu}[\omega_T - u_*(T)] =: \omega_* \in \mathfrak{D}(\mathbf{B}_T^*).$$

Note that even in this case one cannot simply solve (2.44)–(2.47) to obtain ψ_0 (unless $\boldsymbol{\mu} = \mathbf{0}$, the situation considered in [3]) since (2.47) involves $u_*(T) = u_0(T) + \mathbf{B}_T \phi_*$ so ϕ_* appears implicitly.

It will be convenient to put the representation (2.17) in the form of (1.5), eliminating the separate appearance of $\boldsymbol{\kappa}$. Observe that as $\boldsymbol{\kappa}$ is positive definite, by assumption, it has a (unique) positive definite square root $\boldsymbol{\kappa}^{1/2}$. We may then write (2.17) as

$$(2.48) \quad \boldsymbol{\kappa}^{1/2}(\mathbf{1} + \boldsymbol{\kappa}^{-1/2} \mathbf{B}^* \boldsymbol{\lambda} \mathbf{B} \boldsymbol{\kappa}^{-1/2}) \boldsymbol{\kappa}^{1/2} \phi_* = \psi_0$$

so, multiplying through by $\boldsymbol{\kappa}^{-1/2}$ and setting

$$(2.49) \quad \tilde{\phi}_* := \boldsymbol{\kappa}^{1/2} \phi_*, \quad \tilde{\psi}_0 := \boldsymbol{\kappa}^{-1/2} \psi_0, \quad \mathbf{T} := \boldsymbol{\kappa}^{-1/2} \mathbf{B}^* \boldsymbol{\lambda} \mathbf{B} \boldsymbol{\kappa}^{-1/2},$$

we obtain the representation

$$(2.50) \quad (\mathbf{1} + \mathbf{T}) \tilde{\phi}_* = \tilde{\psi}_0 \in L^2(\mathcal{S}).$$

We immediately observe that, from its form (noting the self-adjointness of $\boldsymbol{\kappa}^{-1/2}$), one has

$$(2.51) \quad \mathbf{T} \text{ is a bounded, positive semi-definite operator on } L^2(\mathcal{S}).$$

Our final task of this section is the interpretation of \mathbf{T} . Introducing \mathbf{R} as the solution operator: $z \mapsto w$ for (2.29)–(2.31) so $\mathbf{B}^* = \boldsymbol{\beta}' \mathbf{R}$, we construct \mathbf{T} by the sequence of steps:

- (a) apply $\boldsymbol{\kappa}^{1/2}$ to $\tilde{\phi}$ to obtain $\phi = \boldsymbol{\kappa}^{-1/2} \tilde{\phi}$,
- (b) given ϕ , solve (2.12) to obtain $u := \mathbf{B} \phi$,
- (c) apply $\boldsymbol{\lambda}$ to u to obtain $z := \boldsymbol{\lambda} u$,
- (d) given z , solve (2.29)–(2.31) to obtain $w := \mathbf{R} z$,
- (e) apply $\boldsymbol{\beta}'$ to w to obtain $\psi := \boldsymbol{\beta}' w$,
- (f) apply $\boldsymbol{\kappa}^{-1/2}$ to ψ to obtain $\tilde{\psi} =: \mathbf{T} \tilde{\phi}$.

3. Semigroup formulation. From this point on we assume that the coefficients in (2.4), (2.5) are independent of t . For suitable $s \geq 0$ (how large s could be depends on the smoothness assumed for the coefficients) we can then use the autonomous $(\mathbf{A}, \boldsymbol{\beta})$ to determine a family of operators \mathbf{A}_s :

$$(3.1) \quad \mathbf{A}_s: H^s(\Omega) \supset \mathfrak{D}(\mathbf{A}_s) := \{u \in H^{s+2}(\Omega); \boldsymbol{\beta} u = 0\} \rightarrow H^s(\Omega): u \mapsto \mathbf{A} u.$$

A consequence of the ellipticity of \mathbf{A} and the admitted cases (2.6) for $\boldsymbol{\beta}$ is that [1]:

$$(3.2) \quad \begin{aligned} &\mathbf{A}_s \text{ is the infinitesimal generator of a holomorphic semigroup} \\ &\mathbf{S} = \mathbf{S}_s(\cdot) \text{ acting on } H^s(\Omega) \end{aligned}$$

for suitable s . This is given in, e.g., [1] for $s = 0$. For $s > 0$, note that $\mathbf{S}_0(t)(t > 0)$ takes $L^2(\Omega)$ into $\mathfrak{D}(\mathbf{A}_0^k) \subset H^s(\Omega)$, so it acts on $H^s(\Omega)$ by restriction and is still holomorphic in t —this will just be $\mathbf{S}_s(t)$. (Note that we can only consider $s < \hat{m} + 1/2$: otherwise

$\mathfrak{D}(\mathbf{A}_s)$, given by (3.1), would not be dense in $H^s(\Omega)$ [2] so \mathbf{A}_s could not be the infinitesimal generator.) We similarly introduce \mathbf{A}_s^* given by

$$(3.3) \quad \mathbf{A}_s^*: H^s(\Omega) \supset \mathfrak{D}(\mathbf{A}_s^*) := \{v \in H^{s+2}(\Omega): \hat{\beta}v = 0\} \rightarrow H^s(\Omega): v \rightarrow \mathbf{A}^*v$$

and let $\mathbf{S}^* = \mathbf{S}_s^*(\cdot)$ be the holomorphic semigroup generated by \mathbf{A}_s^* . (One could also consider certain negative s by duality.) Note that, in general ($s \neq 0$), \mathbf{A}_s^* will not truly be the adjoint of \mathbf{A}_s nor need $\mathbf{S}_s^*(t)$ be the adjoint of $\mathbf{S}_s(t)$.

For treating the mixed initial-boundary value problem, we must also introduce the *Green's operator* $\mathbf{G} = \mathbf{G}_s$ (again, for suitable $s \geq 0$) where $\mathbf{G}_s\phi = u$ means

$$(3.4) \quad \mathbf{A}u = 0 \quad \text{on } \Omega, \quad \beta u = \phi \quad \text{at } \partial\Omega$$

for ϕ in $H^s(\partial\Omega)$. (Recall the discussion earlier about the assumed solvability of (2.7) in this section and subsequently.) Note [6] that

$$(3.5) \quad \mathbf{G} = \mathbf{G}_s: H^s(\partial\Omega) \rightarrow H^{s+1/2+\hat{m}}(\Omega) \quad \text{is continuous.}$$

(For convenience we will usually omit the subscripts on \mathbf{A}_s , \mathbf{S}_s , \mathbf{G}_s , \mathbf{A}_s^* , \mathbf{S}_s^* .)

From (2.32) and (2.36) the semigroup representations of \mathbf{R} and \mathbf{B}_T^* are quite standard: we have, allowing for the time-reversal in (2.29) and (2.33),

$$(3.6) \quad w(t) := [\mathbf{R}z](t) = \int_t^T \mathbf{S}^*(r-t)z(r) dr = \mathbf{S}(t'-t)w(t') + \int_t^{t'} \mathbf{S}^*(r-t)z(r) dr$$

for $0 \leq t \leq t' \leq T$ and, similarly, $\mathbf{B}_T^*\hat{\omega} = \beta'v$, so

$$(3.7) \quad [\mathbf{B}_T^*\hat{\omega}](t) = \beta'\mathbf{S}^*(T-t)\hat{\omega} = \beta'\mathbf{S}^*(t'-t)v(t').$$

For \mathbf{B} one has the semigroup representation

$$(3.8) \quad u(t) := [\mathbf{B}\phi](t) = \int_0^t \mathbf{A}\mathbf{S}(t-r)\mathbf{G}\phi(r) dr = \mathbf{S}(t-t')u(t') + \int_{t'}^t \mathbf{A}\mathbf{S}(t-r)\mathbf{G}\phi(r) dr$$

for $0 \leq t' \leq t \leq T$. (Clearly this gives (2.1) while verification of (2.2) is easy for smooth ϕ , after integrating by parts over $[0, T]$, and this extends by continuity.) Finally, for u_0 we obtain

$$(3.9) \quad u_0(t) = \mathbf{S}(t)\omega_0 + \int_0^t \mathbf{S}(t-r)f(r) dr.$$

For a holomorphic semigroup [1], [17], \mathbf{S} , it is standard that $\mathbf{S}(t)$ is uniformly bounded on $[0, T]$ and $\mathbf{A}\mathbf{S}(t) = O(1/t)$ (compare (2.43)). We will repeatedly need a sharper bound for $\mathbf{A}\mathbf{S}(t)$ as an operator from $H^{s'}(\Omega)$ to $H^s(\Omega)$.

LEMMA 3.10. *Let $s > 0$ and $0 \leq \theta \leq 1$ be such that*

$$(3.11) \quad s' := s + 2\theta < \hat{m} + \frac{1}{2}.$$

Then there is a constant K , depending on (\mathbf{A}, β) , s , θ , such that

$$(3.12) \quad |\mathbf{A}\mathbf{S}(t)\omega|_s \leq \frac{K}{|t|^{1-\theta}} |\omega|_{s'} \quad \text{for } \omega \in H^{s'}(\Omega)$$

and similarly for $\mathbf{A}^\mathbf{S}^*(t)$.*

Proof. This is essentially the estimate of [16, Thm. 2(b)] once one observes that, subject to (3.11),

$$(3.13) \quad \mathfrak{D}((- \mathbf{A}_s)^\theta) = [\mathfrak{D}(\mathbf{A}_s), H^s(\Omega)]_{1-\theta} = H^{s'}(\Omega)$$

and similarly for \mathbf{A}_s^* . This is a consequence of the loss of significance of β (resp., $\hat{\beta}$)

acting on $H^{s'}(\Omega)$ as s' drops below $\hat{m} + \frac{1}{2}$. For case D, (3.13) is clear as $H_0^{s'}(\Omega) = H^{s'}(\Omega)$ for $s' < \frac{1}{2}$; see, e.g., [16, Lemma 3(ii)]. This also gives (3.13) for $\hat{m} = 1$ if $s' < \frac{1}{2}$ (which is all we really need—the computations used for case D could also be used for cases N and R) but a more delicate analysis [2], [13] permits $s' < \frac{3}{2}$ for $\hat{m} = 1$. One argument for (3.12) is then to note that $\|A^\rho S(t)\| = O(|t|^{-\rho})$ [1], [8] and that $\mathbf{A}S(t) = \mathbf{A}^{1-\theta}S(t)\mathbf{A}^\theta$ on $\mathfrak{D}(\mathbf{A}^\theta)$ so, for ω in $H^{s'}(\Omega) = \mathfrak{D}(\mathbf{A}_s^\theta)$,

$$|\mathbf{A}S(t)\omega|_s \leq \|\mathbf{A}_s^{1-\theta}S(t)\| |\mathbf{A}_s^\theta\omega|_s \leq K \frac{|\omega|_{s'}}{|t|^{1-\theta}},$$

noting that (3.13) implies $|\mathbf{A}_s^\theta\omega|_s \leq K'|\omega|_{s'}$. The same argument also gives: $\|\mathbf{A}^*\mathbf{S}^*(t)\|_{s' \rightarrow s} = O(1/|t|^{1-\theta})$. \square

Returning to the representation (2.49), (2.50) obtained from Theorem 1, we multiply by \mathbf{T}^k to obtain

$$(3.14) \quad (\mathbf{1} + \mathbf{T})\tilde{\phi}_k = \tilde{\psi}_k, \quad k = 0, 1, \dots,$$

where

$$(3.15) \quad \tilde{\phi}_k := (-\mathbf{T})^k \phi_*, \quad \tilde{\psi}_k := (-\mathbf{T})^k \tilde{\psi}_0$$

and then inductively obtain the partial Neumann series

$$(3.16) \quad \tilde{\phi}_* = \tilde{\psi}_0 + \dots + \tilde{\psi}_{k-1} + \tilde{\phi}_k, \quad k = 1, 2, \dots$$

Since $(\mathbf{1} + \mathbf{T})$ is invertible by (2.51), we may take (3.14) as defining $\tilde{\phi}_k$ and proceed, following (2.52), to obtain each $\tilde{\psi}_k$ from $\tilde{\psi}_{k-1}$ ($k = 1, 2, \dots$) by

$$\begin{aligned} (3.17) \quad & \text{(a) } \bar{\psi}_k(t) := \kappa(t)^{-1/2} \tilde{\psi}_{k-1}(t) = \kappa(t)^{-1} \psi_{k-1}(t), \\ & \text{(b) } u_k(t) := \int_0^t \mathbf{A}S(t-r) \mathbf{G} \bar{\psi}_k(r) dr, \\ & \text{(c) } z_k(t) := -\lambda(t) u_k(t), \\ & \text{(d) } w_k(t) := \int_t^T \mathbf{S}^*(r-t) z_k(r) dr, \\ & \text{(e) } \psi_k(t) := \beta' w_k(t), \\ & \text{(f) } \tilde{\psi}_k(t) := \kappa(t)^{-1/2} \psi_k(t) =: [-\mathbf{T} \tilde{\psi}_{k-1}](t). \end{aligned}$$

Observe that for $k = 0$ (3.17e) would be consistent with the earlier relation (2.44) of w_0, ψ_0 . (Also, (3.17d) would match (2.34), (2.45), (2.47) if $w_0(T) = 0$ in (2.47), for example, if $\mu = 0$, but this is not needed.) The representation of ψ_0 in terms of w_0 satisfying (2.34), (2.45), (2.46) will be used later but for now we use only the knowledge that ψ_0 and so also $\tilde{\psi}_0$ are in $L^2(\mathcal{S})$.

LEMMA 3.18. Let $\mathcal{Y}_\sigma := L^\infty([0, T] \rightarrow H^\sigma(\partial\Omega))$ with $\sigma < 2 - m'$. Then $\tilde{\psi}_1$ is in \mathcal{Y}_σ for cases N and R; $\tilde{\psi}_2$ is in \mathcal{Y}_σ for case D.

Proof. We treat cases N and R together. For these cases $\hat{m} = 1$ so (3.5) gives \mathbf{G}_0 continuous to $H^{3/2}(\Omega)$, hence a fortiori continuous to $H^{s'}(\Omega)$ with $s' = \frac{3}{2} - \delta$. This satisfies (3.11) on taking $s = \bar{s} := \frac{1}{2} - 3\delta$ and $\theta := \frac{1}{2} + \delta$ and we apply Lemma 3.10 to

(3.17a, b) with $k = 1$ to obtain

$$\begin{aligned} |u_1(t)|_{\bar{s}} &\leq \int_0^t K(t-r)^{-1/2+\delta} \|\mathbf{G}\|_{0 \rightarrow s'} (\max \|\kappa^{-1}(r)\|) |\psi_0(r)|_0 dr \\ &\leq K \|\mathbf{G}\| \|\kappa^{-1}\| \left[\int_0^t (t-r)^{-1+2\delta} dr \right]^{1/2} \|\psi_0\|_{L^2(\mathcal{S})}, \end{aligned}$$

so u_1 is in $\mathcal{X}_{\bar{s}} := L^\infty((0, T) \rightarrow H^{\bar{s}}(\Omega))$.

The argument for case D is rather more complicated. Let $\bar{s} := \frac{1}{2} - 9\delta$ and then introduce a bit of numerology:

$$\begin{aligned} (3.19) \quad \hat{\theta} &:= \delta, \quad \tilde{\theta} := \tfrac{1}{2} - 2\delta, \quad \bar{\theta} := 4\delta, \\ \hat{\xi} &:= 1 - \delta, \quad \tilde{\xi} := \tfrac{1}{2} + 2\delta, \quad \bar{\xi} := 1 - 4\delta, \quad (\text{i.e., } \xi := 1 - \theta), \\ \hat{s} &:= \tfrac{1}{2} - 3\delta, \quad s := \tfrac{3}{2} + \delta, \quad \rho := \tilde{\xi} + \bar{\xi} - 1 + \delta = \tfrac{1}{2} - \delta. \end{aligned}$$

We also set

$$(3.20) \quad B(\bar{t}, r) := \int_r^T (\hat{t} - r)^{-\xi} \int_0^{\min\{\hat{t}, \bar{t}\}} (\hat{t} - \tilde{t})^{-\tilde{\xi}} (\bar{t} - \tilde{t})^{-\bar{\xi}} d\tilde{t} d\hat{t}$$

and assert (but defer temporarily the proof) that

$$(3.21) \quad B(\bar{t}, r) \leq K |\bar{t} - r|^{-\rho} \quad \text{for } \bar{t}, r \in [0, T].$$

(Here and later, K, K' will denote generic constants, not necessarily the same at each occurrence.)

For this case, $\hat{m} = 0$ so \mathbf{G}_0 is continuous to $H^{1/2}(\Omega)$ so a fortiori to $H^{s'}(\Omega)$ with $s' := \frac{1}{2} - \delta$, which satisfies (3.11) on setting $s = \hat{s}$, $\theta = \hat{\theta}$. Applying the Lemma (3.10) to (3.17a, b) with $k = 0$ gives, as earlier,

$$(3.22) \quad |u_1(\hat{t})|_s \leq K \int_0^{\hat{t}} (\hat{t} - r)^{-\xi} |\tilde{\psi}_0(r)|_0 dr.$$

Next, take $s' := \hat{s}$, $s := \tilde{s} - 2$, $\theta := \tilde{\theta}$ satisfying (3.11) and apply Lemma 3.10 to (3.17c, d) to obtain, noting [6, Thm. 2.8.3] and the discussion following (2.7),

$$\begin{aligned} (3.23) \quad |w_1(\tilde{t})|_{\tilde{s}} &\leq K' |\mathbf{A}^* w_1(\tilde{t})|_{\tilde{s}-2} \leq K' \int_t^T \|\mathbf{A}^* \mathbf{S}_{\tilde{s}-2}^*(\hat{t} - \tilde{t})\|_{\tilde{s} \rightarrow \tilde{s}-2} \|\mathbf{A}(\hat{t})\|_{\tilde{s}} |u_1(\hat{t})|_{\tilde{s}} d\hat{t} \\ &\leq K \int_t^T (\hat{t} - \tilde{t})^{-\tilde{\xi}} |u_1(\hat{t})|_{\tilde{s}} d\hat{t}. \end{aligned}$$

Finally, with $s := \tilde{s} - \frac{3}{2}$, $\theta = \bar{\theta}$, $s' = \bar{s}$ satisfying (3.11), apply Lemma 3.10 to (3.17e, f, a, b), noting $\tilde{m} = 1$ in case D, to obtain

$$(3.24) \quad |u_2(\bar{t})|_{\bar{s}} \leq K \int_0^{\bar{t}} (\bar{t} - \tilde{t})^{-\bar{\xi}} |w_1(\tilde{t})|_{\tilde{s}} d\tilde{t}.$$

(Note that in obtaining (2.23) we used the continuity of \mathbf{A} on $H^{\tilde{s}}(\Omega)$ and for (3.24) we used the continuity of $\mathbf{B}': H^{\tilde{s}}(\Omega) \rightarrow H^{\tilde{s}-3/2}(\partial\Omega)$ and of $\kappa^{-1/2}$ on $H^{\tilde{s}-3/2}(\partial\Omega)$, following Definition 2.10.)

Substituting (3.22) into (3.33) and that into (3.24), we can interchange the order of integration to obtain

$$(3.25) \quad |u_2(\bar{t})|_{\bar{s}} \leq K \int_0^T B(\bar{t}, r) |\tilde{\psi}_0(r)|_0 dr \leq K \int_0^T |\bar{t} - r|^{-\rho} |\tilde{\psi}_0(r)|_0 dr \leq K \|\tilde{\psi}_0\|_{L^2(\mathcal{S})},$$

using (3.21) and the Cauchy inequality (noting that with $\rho < \frac{1}{2}$, $|\bar{t} - \cdot|^{-\rho}$ is in $L^2(0, T)$, with norm bounded uniformly in \bar{t}). Thus, u_2 is in $\mathfrak{X}_{\bar{s}}$ for case D.

We now return to the verification of (3.21). Let $C(\hat{t}, \bar{t})$ denote the inner integral on the right of (3.20). If $\hat{t} \leq \bar{t}$, then $(\bar{t} - \hat{t}) \geq (\hat{t} - \hat{t})$, $(\bar{t} - \hat{t})$ in the integral so, as $0 < \rho < \bar{\xi}$, one has

$$(\bar{t} - \hat{t})^{-\bar{\xi}} \leq (\hat{t} - \hat{t})^{-\bar{\xi} + \rho} (\bar{t} - \hat{t})^{-\rho}$$

and, as $\bar{\xi} - \bar{\xi} + \rho = -1 + \delta$, one has

$$C(\hat{t}, \bar{t}) \leq \int_0^{\bar{t}} (\hat{t} - \hat{t})^{-1 + \delta} d\hat{t} (\bar{t} - \hat{t})^{-\rho}, \quad \hat{t} \leq \bar{t}.$$

If $\bar{t} \leq \hat{t}$ one similarly obtains

$$C(\hat{t}, \bar{t}) \leq \int_0^{\bar{t}} (\bar{t} - \hat{t})^{-1 + \delta} d\hat{t} (\hat{t} - \bar{t})^{-\rho}, \quad \bar{t} \leq \hat{t}.$$

In either case, then, $C(\hat{t}, \bar{t}) \leq K |\bar{t} - \hat{t}|^{-\rho}$ and substituting this into (3.20) we get

$$(3.26) \quad B(\bar{t}, r) \leq K \int_r^T (\hat{t} - r)^{-\bar{\xi}} |\bar{t} - \hat{t}|^{-\rho} d\hat{t}.$$

Now, if $\bar{t} < r$, one has $|\bar{t} - \hat{t}| = (\hat{t} - \bar{t}) > (r - \bar{t})$ in (3.26), so

$$B(\bar{t}, r) < K \int_r^T (\hat{t} - r)^{-\bar{\xi}} d\hat{t} (r - \bar{t})^{-\rho}, \quad \bar{t} < r,$$

which gives (3.21) in this case. For $r < \bar{t} \leq T$, on the other hand, we split the integral of (3.20) into two parts: $B_+(\bar{t} < \hat{t} < T)$ and $B_-(r < \hat{t} < \bar{t})$. For $\hat{t} > \bar{t} > r$ we have $(\hat{t} - r) > (\hat{t} - \bar{t})$, $(\bar{t} - r)$ so, as $\bar{\xi} > \rho$,

$$(\hat{t} - r)^{\bar{\xi}} (\hat{t} - \bar{t})^{\rho} > (\hat{t} - r)^{\rho} (\hat{t} - \bar{t})^{\bar{\xi}} > (\bar{t} - r)^{\rho} (\hat{t} - \bar{t})^{\bar{\xi}}$$

and

$$B_+ \leq K (\bar{t} - r)^{-\rho} \int_r^T (\hat{t} - \bar{t})^{-\bar{\xi}} d\hat{t}.$$

For $r < \hat{t} < \bar{t}$ we make the substitution $\hat{t} := r + (\bar{t} - r)t$ and get

$$\begin{aligned} B_- &\leq K \int_r^{\bar{t}} (\hat{t} - r)^{-\bar{\xi}} (\bar{t} - \hat{t})^{-\rho} d\hat{t} \\ &= K (\bar{t} - r)^{1 - \bar{\xi} - \rho} \int_0^1 t^{-\bar{\xi}} (1 - t)^{-\rho} dt. \end{aligned}$$

Combining these, noting that $\bar{\xi} < 1$ gives integrability in each case and $(1 - \bar{\xi} - \rho) > -\rho$, one has

$$B(\bar{t}, r) = B_- + B_+ < K (\bar{t} - r)^{-\rho}, \quad \bar{t} > r$$

which again gives (3.21).

Finally we show that u_k in $\mathfrak{X}_{\bar{s}}$ implies $\tilde{\psi}_k$ in \mathfrak{Y}_{σ} . We apply Lemma 3.10 to (3.17c, d) with $s = \bar{s} - 2\delta$, $s' = \bar{s}$, $\theta = 1 - \delta$ to obtain, as for (3.23),

$$|w_k(t)|_{s+2} \leq K' |\mathbf{A}^* w_k|_s \leq K'' \int_t^T (\hat{t} - t)^{-1+\delta} |u_k(\hat{t})|_{\bar{s}} d\hat{t}$$

so w_k is in $\mathfrak{X}_{\bar{s}+2-2\delta}$. The desired result follows with $\sigma = \bar{s} + \frac{3}{2} - m' - 2\delta$ (this may be taken to *define* $\delta > 0$ if σ close to $2 - m'$ was initially specified) on applying \mathbf{B}' , which is continuous from $H^{\sigma+1/2+m'}(\Omega) = H^{\bar{s}+2-2\delta}(\Omega)$ to $H^{\sigma}(\partial\Omega)$, and then $\kappa^{-1/2}$, which is continuous, following Definition 2.10, on \mathfrak{Y}_{σ} . \square

At this point we have obtained the representation (1.5), slightly modified, and Lemma 3.18 will be the key element in step (ii) of the program described there.

4. Analyticity. For the holomorphic semigroups \mathbf{S}, \mathbf{S}^* , there will be a sector of analyticity

$$(4.1) \quad \Sigma_{\bar{\alpha}} := \{t \in \mathbb{C} : |\arg t| < \bar{\alpha}\}, \quad 0 < \bar{\alpha} < \frac{\pi}{2},$$

on which \mathbf{S}, \mathbf{S}^* are analytic.

DEFINITION 4.2. Given a set $\mathcal{R} \subset \mathbb{C}$ (and a specified $\alpha \geq 0$), write $t_0 <_{\mathcal{R}} t_1$ if $t_0, t_1 \in \mathcal{R}$ and there is a Lipschitzian function $t: [0, 1] \rightarrow \mathbb{R}$ with $t(0) = t_0$, $t(1) = t_1$ and $|\arg t'| \leq \alpha$ almost everywhere. The range of such a function is called an \mathcal{R} -suitable path, from t_0 to t_1 . A simply connected $\mathcal{R} \subset \mathbb{C}$ will be α -admissible, if $[0, T] \subset \mathcal{R}$ and, for each t in \mathcal{R} ($t \neq 0, T$), one has $0 <_{\mathcal{R}} t <_{\mathcal{R}} T$.

DEFINITION 4.3. If \mathcal{R} is α -admissible ($0 \leq \alpha < \bar{\alpha}$), then the formulas (3.6)–(3.9) may be used to define \hat{u}_0 and the operators $\hat{\mathbf{R}}, \hat{\mathbf{B}}_T^*, \hat{\mathbf{B}}$, provided that $\hat{z}, \hat{\phi}, \hat{f}$ are suitably analytic in $t \in \mathcal{R}^0$. That is, for each $t \in \mathcal{R}$ the integrals can be taken as complex path integrals along \mathcal{R} -suitable paths, with the results independent of the path and analytic in $t \in \mathcal{R}^0$.

Here, \mathcal{R}^0 denotes the interior (with respect to \mathbb{C}) of \mathcal{R} . Note that if \mathcal{R} has empty interior (e.g., if $\alpha = 0$), then both hypotheses and the conclusions of (4.3), with respect to analyticity are vacuous: $\hat{\mathbf{R}}, \hat{\mathbf{B}}_T^*, \hat{\mathbf{B}}$ reduce to $\mathbf{R}, \mathbf{B}_T^*, \mathbf{B}$, respectively. Note, also that the use of (3.6)–(3.8) ensures commutativity of diagrams like (1.6) for these operators: if z is the restriction of \hat{z} to $[0, T]$, then the restriction of $\hat{\mathbf{R}}\hat{z}$ to $[0, T]$ is just $\mathbf{R}z$, etc. (We will use corresponding letters with and without carets for functions and operators involving \mathcal{R} and $[0, T]$, respectively.)

We also wish to be able to continue to employ the estimate (3.12) for complex t . It is not hard to verify that if $|\alpha| < \bar{\alpha}$, then $e^{i\alpha} \mathbf{A}_s$ is also the infinitesimal generator of a holomorphic semigroup (with sector of analyticity containing $e^{i\alpha} \Sigma_{(\bar{\alpha}-|\alpha|)}$). It follows from this and (3.10) that

$$(4.4) \quad |\mathbf{A}(e^{i\alpha} r)\omega|_s \leq K(\alpha) r^{-1+\theta} |\omega|_{s'}, \quad \text{given } |\alpha| < \bar{\alpha}, r \in \mathbb{R}$$

for s, s', θ satisfying (3.11) and some $K(\alpha)$ but we will need uniformity in α .

LEMMA 4.5. Let $0 \leq \hat{\alpha} < \bar{\alpha}$. Then there exists $K = K_{\hat{\alpha}}$ such that (4.4) holds uniformly in α on $[-\hat{\alpha}, \hat{\alpha}]$ with $K(\alpha)$ replaced by $K_{\hat{\alpha}}$.

Proof. The assertion is that $\|t^{1-\theta} \mathbf{A}(t)\|_{s' \rightarrow s}$ is bounded uniformly on $\bar{\Sigma}_{\hat{\alpha}} \setminus \{0\}$. We know that $t \mapsto t^{1-\theta} \mathbf{A}(t)$ is $[H^{s'} \rightarrow H^s]$ -analytic on $\Sigma_{\bar{\alpha}}$ with $t \mathbf{A}(t)$ bounded on $\bar{\Sigma}_{\hat{\alpha}}$. The Banach space version of the Phragmén–Lindelöf theorem then gives the uniform estimate $K(\alpha) \leq K_{\hat{\alpha}}$ for (4.4) with $K_{\hat{\alpha}} = \max \{K(\hat{\alpha}), K(-\hat{\alpha})\}$. \square

We now return to the program indicated following (1.5). For $\mathcal{R} \subset \mathbb{C}$ and s in \mathbb{R} we can define

$$(4.6) \quad \mathcal{Y}_s(\mathcal{R}) := \{\hat{\phi} \in L^\infty(\mathcal{R} \rightarrow H^s(\partial\Omega)): \hat{\phi} \text{ analytic on } \mathcal{R}^0\}.$$

(In defining L^∞ for \mathcal{R} we interpret the notion of ess sup to ignore sets of *linear* measure 0 in $[0, T]$ and of *planar* measure 0 in the rest of \mathcal{R} .) Since the uniform limit (on \mathcal{R}^0 open in \mathbb{C}) of analytic functions is analytic for vector-valued functions (of course, we now must take $H^s(\partial\Omega)$ to be a space of *complex*-value functions on $\partial\Omega$), one has:

$$\mathcal{Y}_s(\mathcal{R}) \text{ is a closed subspace of } L^\infty(\mathcal{R} \rightarrow H^s(\partial\Omega)).$$

Note that for $\mathcal{R} = [0, T]$, definition of this $\mathcal{Y}_s = \mathcal{Y}_s([0, T])$ coincides with that of Lemma 3.18. We adjoin to Definition 2.10 the assumption that

$$(4.7) \quad \text{The operators } \kappa^{1/2}(t), \kappa^{-1/2}(t), \lambda(t) \text{ depend analytically on } t \text{ in } \mathcal{R}^0, \text{ when considered as bounded operators on } H^s(\partial\Omega), H^s(\Omega) \text{ for suitable } s.$$

We will also require that, for suitable \mathcal{U} open in \mathbb{C} ,

$$(4.8) \quad \hat{z}_0 := \lambda[u_T - u_0] \text{ is bounded uniformly on } [\varepsilon, T] \text{ for } \varepsilon > 0 \text{ and is analytic on } \mathcal{U} \text{ as an } L^2(\Omega)\text{-valued function.}$$

(If $[0, t] \cup \mathcal{U}$ is α -admissible with $0 < \alpha < \bar{\alpha}$ and f is analytic on \mathcal{U} , then (3.9) and the analyticity of \mathbf{S} imply analyticity of u_0 , whence (4.7) and analyticity in t of the target trajectory u_T give (4.8).)

LEMMA 4.9. *Let \mathcal{R} be $\hat{\alpha}$ -admissible ($0 \leq \hat{\alpha} < \bar{\alpha}$) and let Definition 2.10, (3.2) and (4.7) hold. Let $0 \leq \bar{s} < 2 + \hat{m} - m'$ and define $\hat{\mathbf{T}}$ by*

$$(4.10) \quad \hat{\phi} \mapsto u := \hat{\mathbf{B}}\kappa^{-1/2}\hat{\phi} \mapsto \hat{z} := \lambda\hat{u} \mapsto \hat{w} := \hat{\mathbf{R}}\hat{z} \mapsto \kappa^{-1/2}\beta'\hat{w} =: \hat{\mathbf{T}}\hat{\phi}$$

for $\hat{\phi}$ in $\mathcal{Y} = \mathcal{Y}_{\bar{s}}(\mathcal{R})$. Then $\hat{\mathbf{T}}$ is a well-defined compact operator on \mathcal{Y} and (1.6) is commutative.

Proof. By Definition 4.3, the hypotheses ensure that $\hat{\phi}$ analytic on \mathcal{R}^0 implies that \hat{u} (and so \hat{z}) is well defined and is analytic on \mathcal{R}^0 . This analyticity is, of course, a purely local property in \mathcal{R}^0 and—once one notes the analyticity of the functions involved and shows absolute convergence in the appropriate spaces of the integrals—can be shown by the usual methods employed in complex function theory for functions defined by integrals. Similarly one has w —and so $\hat{\mathbf{T}}, \hat{\phi}$ —well defined and analytic. Thus $\hat{\mathbf{T}}$ is well defined and will take \mathcal{Y} into \mathcal{Y} once an L^∞ bound into $H^{\bar{s}}(\partial\Omega)$ is obtained. The commutativity of (1.6) follows by comparing (4.10) to (2.52) and noting the discussion after Definition 4.3.

For $\hat{\phi}$ in \mathcal{Y} one also has $\kappa^{-1/2}\hat{\phi}$ in $\mathcal{Y} \subset L^\infty(\mathcal{R} \rightarrow L^2(\partial\Omega))$ so, by (4.10), (3.8) and Definition 4.3, we have

$$\hat{u}(t) := \int_0^t \mathbf{A}\mathbf{S}(t-r)\mathbf{G}\kappa^{-1/2}(r)\hat{\phi}(r) dr$$

and can apply Lemma 4.5 with $s' := \frac{1}{2} + \hat{m} - \delta$, $s = \hat{s} := \frac{1}{2} + \hat{m} - 3\delta$ and $\theta := 1 - \delta$, satisfying (3.11). Thus,

$$(4.11) \quad |\hat{u}(t)|_{\bar{s}} \leq K_{\hat{\alpha}} \int_0^t |t-r|^{-1+\delta} |\mathbf{G}\kappa^{-1/2}(r)\hat{\phi}(r)|_{s'} |dr|;$$

so, using (3.5) and Definition 2.10 to bound $|\mathbf{G}\kappa^{-1/2}(r)\hat{\phi}(r)|_{s'}$ in terms of $|\hat{\phi}(r)|_{-\delta} \leq K\|\hat{\phi}\|_{\mathcal{Y}}$, one has \hat{u} in

$$\mathcal{X}_{\bar{s}}(\mathcal{R}) := \{\hat{u} \in L^\infty(\mathcal{R} \rightarrow H^{\bar{s}}(\Omega)): \hat{u} \text{ analytic on } \mathcal{R}^0\},$$

with $\hat{s} = \frac{1}{2} + \hat{m} - 3\delta$. By Definition 2.10 and (4.7) we then also have \hat{z} in $\mathcal{X}_{\hat{s}}(\mathcal{R})$. Next, by (4.10), (3.6) and Definition 4.3 we have

$$\mathbf{A}^* \hat{w}(t) := \int_t^T \mathbf{A}^* \mathbf{S}^*(r - \hat{t}) \hat{z}(r) dr$$

and we apply Lemma 4.5 with

$$s' = \hat{s}, \quad s = s^* := \frac{1}{2} + \hat{m} - 5\delta, \quad \theta := 1 - \delta,$$

satisfying (3.11), to obtain

$$(4.12) \quad |\hat{w}(t)|_{s^*+2} \leq K |\mathbf{A}^* \hat{w}(t)|_{s^*} \leq K' \int_t^T |r - t|^{-1+\delta} |\hat{z}(r)|_{\hat{s}} |dr|.$$

Thus, \hat{w} is in $\mathcal{X}_{s^*+2}(\mathcal{R})$ and this, a fortiori, is in $\mathcal{X}_{\hat{s}}(\mathcal{R})$ ($\hat{s} := s^* + 2 - 2\delta = \frac{5}{2} + \hat{m} - 7\delta$) with its range in a specific compact subset of $H^{\hat{s}}(\Omega)$ for $\hat{\phi}$ limited to any bounded subset of \mathcal{Y} .

To show that the map $\hat{\phi} \mapsto \hat{w}$ is compact from \mathcal{Y} to $\mathcal{X}_{\hat{s}}(\mathcal{R})$, it is now sufficient to show a uniform Hölder condition for \hat{w} as an $H^{\hat{s}}(\Omega)$ -valued function on \mathcal{R} and apply the generalized Arzela–Ascoli lemma. Suppose $t <_{\mathcal{R}} t'$. Then, following (3.6),

$$\mathbf{A}^*[\hat{w}(t) - \hat{w}(t')] = [\mathbf{S}^*(t' - t) - \mathbf{1}] \mathbf{A}^* \hat{w}(t') + \int_t^{t'} \mathbf{A}^* \mathbf{S}^*(r - t) \hat{z}(r) dr$$

and we estimate the two terms separately. The first term is comparatively simple, using the estimate (4.12) already obtained for $\hat{w}(t')$:

$$(4.13) \quad \begin{aligned} |[\mathbf{S}^*(t' - t) - \mathbf{1}] \mathbf{A}^* \hat{w}(t')|_{\hat{s}-2} &\leq \left(\int_0^{t'-t} \|\mathbf{A}^* \mathbf{S}^*(r)\|_{s^* \mapsto s-2} |dr| \right) |\mathbf{A}^* \hat{w}(t')|_{s^*} \\ &\leq K \int_0^{t'-t} |r|^{-1+\delta} |dr| K' \|\hat{z}\|_{\hat{s}} \\ &\leq K'' |t' - t|^{\delta} \|\hat{\phi}\|_{\mathcal{Y}} \end{aligned}$$

where Lemma 4.5 has been used with $s := \hat{s} - 2$, $s' := s^*$, $\theta := 1 - \delta$, satisfying (3.11). For the second term one applies Lemma 4.5 with the same s, s', θ to get

$$(4.14) \quad \left| \int_t^{t'} \mathbf{A}^* \mathbf{S}^*(r - t) \hat{z}(r) dr \right|_{\hat{s}-2} \leq K \int_t^{t'} |r - t|^{-1+\delta} |\hat{z}(r)|_{\hat{s}} |dr| \leq K' \int_t^{t'} |r - t|^{-1+\delta} |dr| \|\hat{\phi}\|_{\mathcal{Y}}$$

as \hat{z} is in $\mathcal{X}_{\hat{s}} \subset \mathcal{X}_{s^*}$. The integral here from t to t' is, of course, taken along an \mathcal{R} -suitable path \mathcal{P} . On such a path one has, writing $r = \rho + i\sigma$, that

$$|dr| \leq (\tan \hat{\alpha}) d\rho,$$

so that

$$\int_t^{t'} |r - t|^{-1+\delta} |dr| \leq \int_{\operatorname{Re} t}^{\operatorname{Re} t'} |\rho - \operatorname{Re} t|^{-1+\delta} (\tan \hat{\alpha}) d\rho = \frac{\tan \hat{\alpha}}{\delta} |\operatorname{Re}(t' - t)|^{\delta} \leq K |t' - t|^{\delta}.$$

Substituting this into (4.14) and combining with (4.13) gives

$$(4.15) \quad |\hat{w}(t) - \hat{w}(t')|_{\hat{s}} \leq K |\mathbf{A}^*[\hat{w}(t) - \hat{w}(t')]|_{\hat{s}-2} \leq K' |t' - t|^{\delta} \|\hat{\phi}\|_{\mathcal{Y}}$$

for $t <_{\mathcal{R}} t'$ and similarly if $t' <_{\mathcal{R}} t$. If t, t' in \mathcal{R} are related in neither of these ways one still obtains (4.15) by introducing \bar{t} such that $\bar{t} <_{\mathcal{R}} t$ and $\bar{t} <_{\mathcal{R}} t'$. A geometric

argument from the definition of $\hat{\alpha}$ -admissibility shows that one may always take

$$\bar{t} = t - r e^{i\hat{\alpha}} = t' - r' e^{-i\hat{\alpha}}$$

and that one then has

$$r, r' \leq \frac{|t - t'|}{2 \sin \hat{\alpha}}.$$

(The argument that this \bar{t} is, indeed, in \mathcal{R} —and is connected to t, t' by straight segments in \mathcal{R} —is not difficult but can be avoided by simply restricting consideration to \mathcal{R} of the form

$$(4.16) \quad \mathcal{R} = [0, T] \cup \{t \in \mathbb{C} : |\arg(t - a)|, |\arg(b - t)| \leq \hat{\alpha}\}$$

for some $0 < a < b < T$ and $0 \leq \hat{\alpha} < \bar{\alpha}$.) Then (4.15) for \bar{t}, t and for \bar{t}, t' combine to give it for t, t' (the general case) with a new K' .

The final step in the proof is to go from \hat{w} to $\hat{\mathbf{T}}\hat{\phi} := \kappa^{-1/2}\mathbf{B}'\hat{w}$, a map which is clearly continuous from $\mathcal{X}_{\bar{s}}(\mathcal{R})$ to $\mathcal{Y}_{\bar{s}}(\mathcal{R})$ with

$$\bar{s} := \bar{s} - \frac{1}{2} - m' = 2 + \hat{m} - m' - 7\delta$$

(which we may invert to define $\delta > 0$). \square

COROLLARY 4.17. $(\mathbf{1} + \hat{\mathbf{T}})$ is boundedly invertible on $\mathcal{Y}_{\bar{s}}(\mathcal{R})$.

Proof. Since $\hat{\mathbf{T}}$ is compact, a standard spectral theory result is that -1 can be in the spectrum $\sigma(\hat{\mathbf{T}})$ only if it is an eigenvalue. Suppose this were so and that

$$\hat{\mathbf{T}}\hat{\phi} = -\hat{\phi}, \quad 0 \neq \hat{\phi} \quad \text{in } \mathcal{Y} = \mathcal{Y}_{\bar{s}}(\mathcal{R}).$$

Then, letting $\phi = \mathbf{E}\hat{\phi}$ be the restriction of $\hat{\phi}$ to $[0, T]$, the commutativity of the diagram (1.6) gives $\mathbf{T}\phi = -\phi$. Since \mathbf{T} is known to be positive semidefinite by construction, this can occur only for $\phi = 0$. However, a geometric argument (again avoidable if one restricts \mathcal{R} to the form (4.16)) shows that, if $\mathcal{R} \neq [0, T]$ is $\hat{\alpha}$ -admissible and symmetric with respect to \mathbb{R} in \mathbb{C} , then each component of \mathcal{R}^0 intersects $[0, T]$ and $[0, T] \cup \overline{\mathcal{R}^0}$ contains \mathcal{R} . For symmetric \mathcal{R} one then has $\hat{\phi}|_{[0, T]} = \phi = 0$, implying $\hat{\phi} = 0$ by the analyticity on \mathcal{R}^0 ; for \mathcal{R} not symmetric, a reflection argument across \mathbb{R} permits one to work with its symmetrization which will again be $\hat{\alpha}$ -admissible. Thus, $\hat{\mathbf{T}}\hat{\phi} = -\hat{\phi}$ implies $\hat{\phi} = 0$, so -1 cannot be in $\sigma(\hat{\mathbf{T}})$ and $(\mathbf{1} + \hat{\mathbf{T}})^{-1} : \mathcal{Y} \rightarrow \mathcal{Y}$ is well defined. \square

We have now completed steps (i), (iii), (iv) of the program following (1.5) and next complete step (ii) which was initiated in Lemma 3.18.

LEMMA 4.18. Let \mathcal{R} be $\hat{\alpha}$ -admissible ($0 \leq \alpha < \bar{\alpha}$) with $0, T$ not in $\overline{\mathcal{R}^0}$ and let Definition 2.10, (3.2) and (4.7) hold. Let $\sigma < 2 - m'$ and let (4.8) hold, with \mathcal{U} containing \mathcal{R}^0 . If $\mu = \infty$, assume Definition 2.39, so that (2.41) holds. Then there is a (unique) ψ_2 in $\mathcal{Y} = \mathcal{Y}_{\sigma}(\mathcal{R})$ whose restriction to $[0, T]$ is $\tilde{\psi}_2$; in cases N and R this holds for ψ_1, ψ_1 .

Proof. We suppose w_0 to be already defined on $[0, T]$ satisfying (2.45), (2.46). One then has $w_0(t')$ in $H^s(\Omega)$ for some $s > m' + \frac{1}{2}$ and $0 < t' < T$, given (4.8). Noting Definition 4.3, we define \hat{w}_0 on the rest of \mathcal{R} by the second integral representation (3.6) with t' in $(0, T)$ and then define $\tilde{\psi}_0(t)$ as $\kappa^{-1/2}(t)\mathbf{B}'\hat{w}_0(t)$, using Definition 4.3, (4.7), (4.8) to observe that this gives $\tilde{\psi}_0$ analytic on (a neighborhood of) \mathcal{R}^0 . Using the steps of (3.17), with the integrals taken along $\hat{\alpha}$ -suitable paths, one obtains ψ_2 corresponding to $\tilde{\psi}_2$. Since $\hat{\psi}_2$ will be analytic on a neighborhood of \mathcal{R}^0 , it will be bounded on \mathcal{R}^0 and, combining this with the boundedness on $[0, T]$ given by Lemma 3.18, one has $\hat{\psi}_2$ in \mathcal{Y} . (For cases N and R, Lemma 3.18 gives $\hat{\psi}_1$ in \mathcal{Y} but $\hat{\psi}_2 = -\hat{\mathbf{T}}\hat{\psi}_1$ so, by Lemma 4.9, ψ_2 is also in \mathcal{Y} .) \square

We next note that analyticity in t can also be used to give additional spatial regularity.

LEMMA 4.19. *Let Definition 2.10, (3.2), (4.7) hold and let \mathcal{R} be $\hat{\alpha}$ -admissible with $0 < \hat{\alpha} < \bar{\alpha}$. Let $\partial\Omega$ and the coefficients be “sufficiently smooth.” Then, for $\hat{\phi}$ in $L^2(\mathcal{S})$ and $H^s(\partial\Omega)$ -analytic on \mathcal{R}^0 , we have $\hat{\mathbf{T}}\hat{\phi}$ $H^{s'}(\partial\Omega)$ -analytic on \mathcal{R}^0 for $s' \leq s + 2 + \hat{m} - m'$, s' not an integer. If \hat{z}_0 is $H^s(\Omega)$ -analytic on \mathcal{R}^0 , then $\hat{\psi}_0$ will be $H^{s'}(\partial\Omega)$ -analytic on \mathcal{R}^0 for $s' \leq \tilde{s} + \frac{3}{2} - m'$, s' not an integer.*

Proof. Suppose $\hat{\phi}$ is $H^s(\partial\Omega)$ -analytic ($s > 0$) so $\kappa^{-1/2}\hat{\phi}$ is also $H^s(\partial\Omega)$ -analytic. Certainly this gives at least $L^2(\Omega)$ -analyticity for \hat{u} on \mathcal{R}^0 . By a fundamental property of analytic functions, $d\hat{u}/dt = \hat{u}'$ must then also be $L^2(\Omega)$ -analytic on \mathcal{R}^0 . We now view the equations

$$\mathbf{A}\hat{u} = \hat{u}', \quad \beta\hat{u} = \kappa^{-1/2}\hat{\phi}$$

as an elliptic problem (with t as a parameter). The continuity of the linear map: $\hat{u}(t), \kappa^{-1/2}\hat{\phi}(t) \mapsto \hat{u}(t)$ then gives $H^\sigma(\Omega)$ -analyticity of \hat{u} (hence, also, of \hat{u}') with $\sigma = \min\{s + \frac{1}{2} + \hat{m}, \bar{\sigma} + 2\}$ for \hat{u}' $H^{\bar{\sigma}}(\Omega)$ -analytic. Initially $\bar{\sigma} = 0$, as noted above, but a “bootstrapping” argument gives $\sigma = s + \frac{1}{2} + \hat{m}$. Next, $\hat{w} = \mathbf{R}\hat{z}(\hat{z} = \lambda\hat{u})$ is certainly $L^2(\Omega)$ -analytic on \mathcal{R}^0 (using the semigroup representation) since \hat{z} is. We now view the equations

$$\mathbf{A}^*\hat{w} = -(\hat{w}' + \lambda\hat{u}), \quad \beta\hat{w} = 0$$

as an elliptic problem for $\hat{w}(t)$ and note that continuity of the linear map $[\hat{w}' + \lambda\hat{u}] \mapsto \hat{w}$ gives $H^{\hat{\sigma}+2}(\Omega)$ -analytic on \mathcal{R}^0 with $\hat{\sigma} = \min\{\sigma', \sigma\}$ for \hat{w}' $H^{\sigma'}(\Omega)$ -analytic and \hat{u} (hence $\lambda\hat{u}$) $H^\sigma(\Omega)$ -analytic. Initially, \hat{w} L^2 -analytic gives $\sigma' = 0$ but, bootstrapping, we can obtain $\hat{\sigma} = \sigma = s + \frac{1}{2} + \hat{m}$. Finally, application of β' gives $H^{s'}(\partial\Omega)$ -analyticity of $\hat{\mathbf{T}}\hat{\phi} = \beta'\hat{w}$ with s' as asserted. The argument for $H^s(\partial\Omega)$ -analyticity of $\hat{\psi}_0$ proceeds similarly. \square

With these lemmata in hand, most of the work toward proving the desired regularity result for the optimal control ϕ_* (and for the optimally controlled trajectory u_*) has already been done. The domain of analyticity of ϕ_* will be determined by the domain \mathcal{U} in \mathbb{C} on which $\kappa^{-1/2}, \lambda, \hat{z}_0 := \lambda(u_T - u_0)$ are analytic. With \hat{z}_0 real on \mathcal{Q} , a reflection principle argument shows \mathcal{U} may be taken as symmetric with respect to \mathbb{R} and we then will obtain analyticity in t for ϕ_* in \mathcal{U} given by

$$\begin{aligned} \hat{\mathcal{U}} &:= \cup\{\mathcal{R}^0: \mathcal{U} \supset \mathcal{R} \text{ } \hat{\alpha}\text{-suitable with } \hat{\alpha} < \bar{\alpha}\} \\ (4.20) \quad &= \cup\{\mathcal{R}^0: \mathcal{U} \supset \mathcal{R} \text{ of the form (4.16), } 0 < a < b < T, \hat{\alpha} < \bar{\alpha}\}. \end{aligned}$$

Note that if $(0, T) \subset \mathcal{U}$ then $(0, T) \subset \hat{\mathcal{U}}$.

THEOREM 2. *Let $\kappa^{-1/2}, \lambda, \mu$ satisfy Definition 2.10 with $\kappa^{-1/2}(\cdot), \lambda(\cdot)$ analytic (as operators on $H^\sigma(\partial\Omega), H^{\sigma'}(\Omega)$ for suitable σ, σ') on an open set $\mathcal{U} \subset \mathbb{C}$, symmetric with respect to \mathbb{R} . Let $\hat{z}_0 := \lambda(u_T - u_0)$ be in $L^2(\mathcal{Q})$ and analytic on \mathcal{U} as an $H^s(\Omega)$ -valued function. Let $\partial\Omega$ and the coefficients in (2.4), (2.5) be sufficiently smooth. If $\mu = \infty$, assume nullcontrollability, giving (2.41) and let $[\omega_T - u_0(T)]$ be reachable so (in any case) J , as given by (2.8), (2.9), is not identically infinite on $L^2(\mathcal{S})$. Then the optimal control ϕ_* minimizing J is (the restriction to $[0, T]$ of) a function $H^{s'}(\partial\Omega)$ -analytic on $\hat{\mathcal{U}}$, given by (4.20), for $s' := s + 2 + \hat{m} - m'$ (if this is not an integer and subject to the smoothness assumed for $\partial\Omega$, the coefficients and κ, λ).*

Proof. First take $\tilde{s} < 2 - m'$ and fix $\mathcal{R} \subset \mathcal{U}$ of the form (4.16) with $\hat{\alpha} < \bar{\alpha}$. By Lemma 4.18 we have $\hat{\psi}_2$ in $\mathcal{Y}_{\tilde{s}}(\mathcal{R})$ and by Corollary 4.17 we can obtain $\hat{\phi}_2 := (\mathbf{1} + \hat{\mathbf{T}})^{-1}\hat{\psi}_2$ in $\mathcal{Y}_{\tilde{s}}(\mathcal{R})$. Then, by (2.49), (3.16), ϕ_* is (the restriction to $[0, T]$ of)

$$(4.21) \quad \hat{\phi}_* := \kappa^{-1/2}[\hat{\psi}_0 + \hat{\psi}_1 + \hat{\phi}_2].$$

By Lemma 4.19 $\hat{\psi}_0$ and $\hat{\psi}_1$ are $H^{s'}(\partial\Omega)$ -analytic on \mathcal{R}^0 so, as $\hat{\phi}_2$ is in $\mathcal{Y}_s(\mathcal{R})$, one has $H^s(\partial\Omega)$ -analyticity of $\hat{\phi}_*$. But, by (3.15), $\hat{\phi}_2 = \hat{\mathbf{T}}^2 \mathbf{\kappa}^{1/2} \hat{\phi}_*$ so, using Lemma 4.19 again twice, $\hat{\phi}_2$ is at least $H^{\bar{s}+2}(\partial\Omega)$ -analytic ($s+2+\hat{m}-m' \geq s+1$). This would make $\hat{\phi}_*$ also $H^{\bar{s}+2}(\partial\Omega)$ -analytic on \mathcal{R}^0 if $s+2 \leq s'$, and one iterates this to get $H^{s'}(\partial\Omega)$ -analyticity of $\hat{\phi}_*$ on \mathcal{R}^0 . Since this holds for each such \mathcal{R}^0 one has $H^{s'}(\partial\Omega)$ -analyticity of $\hat{\phi}_*$ on all of \mathcal{U} . Strictly speaking, one has a different $\hat{\phi}_* = \hat{\phi}_*(\mathcal{R})$ for each specification of \mathcal{R} , since $\hat{\phi}_2$ was obtained by inverting $\hat{\mathbf{T}}$ viewed as an operator on $\mathcal{Y}_s(\mathcal{R})$. However, since ϕ_* is unique in $L^2(\mathcal{S})$ and $\hat{\phi}_*(\mathcal{R}_1) = \hat{\phi}_*(\mathcal{R}_2) = \hat{\phi}_*(\mathcal{R}_1 \cap \mathcal{R}_2)$ on $\mathcal{R}_1 \cap \mathcal{R}_2$ (which is again $\hat{\alpha}$ -admissible with $\hat{\alpha} = \min\{\hat{\alpha}_1, \hat{\alpha}_2\}$), one can obtain $\hat{\phi}_*(\hat{\mathcal{U}})$ as a maximally defined extension. \square

5. Infinite horizon problems. In this section we consider (2.1)–(2.6) on $\mathcal{Q} = \mathcal{Q}_\infty := \mathbb{R}^+ \times \Omega$, $\mathcal{S} = \mathcal{S}_\infty := \mathbb{R}^+ \times \partial\Omega$ with the optimality criterion of the same form (2.8), (2.9) considered, but with $\|\cdot\|_\kappa, \|\cdot\|_\lambda$ now taken over $\mathcal{S}_\infty, \mathcal{Q}_\infty$ —and, of course, with the “terminal target” term involving $\|\cdot\|_\mu$ omitted, as there is now no finite T . An example might be a *stabilization* problem:

$$(5.1) \quad J(\phi) := \int_0^\infty e^{-ct} \int_{\partial\Omega} \phi^2 + \lambda \int_0^\infty e^{c't} \int_\Omega (u - u_T)^2,$$

in which $c \geq 0$ implies a discount factor for control cost while $c' \geq 0$ implies a stabilization condition forcing an acceptably rapid exponential decay rate for the trajectory deviation. Our basic assumption on $\|\cdot\|_\kappa, \|\cdot\|_\lambda$ is:

$$(5.2) \quad \mathbf{\kappa}(t), \mathbf{\kappa}^{-1/2}(t), \mathbf{\lambda}(t) \text{ are positive (semi-) definite on } L^2(\partial\Omega), L^2(\Omega), \text{ respectively, and bounded uniformly in } t \in (0, T), \text{ for each } T > 0.$$

It is clear that if, for example, u_T were to grow sufficiently rapidly as $t \rightarrow \infty$, then finiteness of $\|\phi\|_\kappa$ would imply growth conditions for the solution u , which could be incompatible with finiteness of $\|u - u_T\|_\lambda$. On the other hand, even for $u_T \equiv 0$ we might wish to consider (\mathbf{A}, \mathbf{B}) for which $\|u_0\|_\lambda$ is infinite (e.g., $J := \|\phi\|^2 + \|u\|^2$ with ordinary L^2 norms for the standard heat equation in case N) yet have stabilizability with suitable controls; e.g., nullcontrollability would give a reduction to a finite interval by choosing ϕ in $L^2(\mathcal{S}_T)$ so that $u = 0$ for $t = T$ and then, having $\phi, u = 0$ for $t > T$.

As in the earlier case (in which the unboundedness occurred in connection with the terminal target), we again must consider an optimality criterion which need not be continuous or everywhere finite on $L^2(\mathcal{S})$ or even on

$$L^2_\kappa(\mathcal{S}) := \left\{ \phi \in L^2_{\text{loc}}(\mathcal{S}) : \|\phi\|_\kappa^2 := \int_0^\infty \langle \phi(t), \mathbf{\kappa}(t)\phi(t) \rangle_{L^2(\partial\Omega)} dt < \infty \right\}.$$

Our first concern then, paralleling Theorem 1, is the existence, uniqueness and characterization of the optimal control ϕ_* minimizing J .

THEOREM 3. Assume (5.2) and that $(u_0 - u_T)$ is such that J , defined by (2.8), (2.9) omitting the $\|\cdot\|_\mu$ term, is not identically ∞ (i.e., (2.16) holds with L^2 replaced by L^2_κ). Then there exists a unique minimizer ϕ_* of J and a corresponding unique optimal trajectory u_* . For fixed $T > 0$, let

$$(5.3) \quad J_T(\phi) := \|\phi\|_{\kappa, T}^2 + \|u - u_T\|_{\lambda, T}^2 + \|u(T) - \omega_T^*\|_\infty^2,$$

where $\omega_T^* := u_*(T)$ and J_T is defined in terms of restrictions to $\mathcal{S}_T := (0, T) \times \partial\Omega$, $\mathcal{Q}_T := (0, T) \times \Omega$. Then the unique minimizer $\phi_{*, T}$ of J_T is just the restriction to \mathcal{S}_T of ϕ_* .

Proof. Let $\mathfrak{L}_\kappa = L_\kappa^2(\mathcal{S})$ and $\mathfrak{L}_\lambda = L_\lambda^2(\mathcal{Q})$ (for \mathfrak{L}_λ it may be necessary to factor out $\mathfrak{N}(\lambda)$ if $\|\cdot\|_\lambda$ is not a true norm); set $\hat{\mathfrak{L}} := \mathfrak{L}_\kappa \times \mathfrak{L}_\lambda$. Given any $\phi_* \in \mathfrak{D}(J)$, let

$$\Phi_* := [\phi_*, u_* - u_T] \in \hat{\mathfrak{L}},$$

where u_* is the solution of (2.1)–(2.3) associated with the control ϕ_* and noting that $J(\phi_*) = \|\Phi_*\|_{\hat{\mathfrak{L}}}^2$. If $\phi_* + \tilde{h} =: \tilde{\phi}$ is also in $\mathfrak{D}(J)$, then $\tilde{\Phi} := [\tilde{\phi}, \tilde{u} - u_T]$ must also be in $\hat{\mathfrak{L}}$, so

$$\tilde{\Phi} - \Phi_* = [\tilde{h}, \tilde{u} - u_*] = [\tilde{h}, \mathbf{B}\tilde{h}] =: \mathbf{H}\tilde{h}$$

must also be in $\hat{\mathfrak{L}}$. Setting

$$\mathfrak{L}_J := \{h \in \mathfrak{L}_\kappa : \|\mathbf{B}h\|_\lambda < \infty\},$$

with the norm

$$\|h\|_J^2 := \|h\|_\kappa^2 + \|\mathbf{B}h\|_\lambda^2,$$

one has $\mathfrak{D}(J) = \phi_* + \mathfrak{L}_J$ (as a set) with continuity on \mathfrak{L}_J of $h \mapsto J(\phi_* + h)$ —as well as strict convexity and \mathfrak{L}_J -coercivity. Thus J attains its minimum uniquely on $\phi_* + \mathfrak{L}_J$; i.e., on $\mathfrak{D}(J) \subset L_\kappa^2(\mathcal{S})$.

Now let ϕ_* be the minimizer of J and, for fixed $T > 0$, let $\phi_{*,T}$ be the minimizer of J_T ; noting that this exists uniquely by Theorem 1 since the definition of ω_T^* clearly means that it is reachable. Let

$$\hat{\phi}_* := \begin{cases} \phi_{*,T} & \text{on } \mathcal{S}_T & (\text{i.e., for } t < T), \\ \phi_*|_{[T,\infty)} & \text{on } \mathcal{S} \setminus \mathcal{S}_T & (\text{i.e., for } t \geq T) \end{cases}$$

and let $u_{*,T}, \hat{u}_*$ be the trajectories corresponding to $\phi_{*,T}, \hat{\phi}_*$, respectively. Now $\hat{u}_*|_{(0,T)}$ must coincide with $u_{*,T}$, since, by definition, $\hat{\phi}_*|_{(0,T)}$ coincides with $\phi_{*,T}$ and one has uniqueness for (2.1)–(2.3) on \mathcal{Q}_T . Thus, by the definition of J_T ,

$$\hat{u}_*(T) = u_{*,T}(T) = \omega_T^* := u_*(T);$$

so, as $\hat{\phi}_*$ coincides with ϕ_* for $t > T$, \hat{u}_* must coincide with u_* for $t \geq T$ by uniqueness for (2.1)–(2.3) on $[T, \infty) \times \Omega$. Thus

$$\begin{aligned} J(\hat{\phi}_*) &:= \int_0^\infty \langle \hat{\phi}_*, \kappa \hat{\phi}_* \rangle + \int_0^\infty \langle \hat{u}_* - u_T, \lambda(\hat{u}_* - u_T) \rangle \\ &= J_T(\phi_{*,T}) + \int_T^\infty \langle \phi_*, \kappa \phi_* \rangle + \int_T^\infty \langle u_* - u_T, \lambda(u_* - u_T) \rangle \\ &= J_T(\phi_{*,T}) + [J(\phi_*) - J_T(\phi_*|_{(0,T)})], \end{aligned}$$

so that

$$J(\hat{\phi}_*) - J(\phi_*) = J_T(\phi_{*,T}) - J_T(\phi_*|_{(0,T)}).$$

The left side is nonnegative, as ϕ_* minimizes J , and the right side is nonpositive, as $\phi_{*,T}$ minimizes J_T . Thus each side vanishes whence $\phi_*|_{(0,T)} = \phi_{*,T}$ (equivalently, $\hat{\phi}_* = \phi_*$) by uniqueness of the minimizers. \square

THEOREM 4. Let (5.2), (3.2) hold and let $z_0 := \lambda(u_0 - u_T)$ be such that J is not identically ∞ . Let (\mathbf{A}, \mathbf{B}) be nullcontrollable on arbitrary short intervals. Let $\kappa^{-1/2}(t), \lambda(t)$ be analytic in t (as operators on $H^\sigma(\partial\Omega), H^{\sigma'}(\Omega)$ for suitable σ, σ') on an open set \mathcal{U} in \mathbb{C} , symmetric with respect to \mathbb{R} and let z_0 be (the restriction to \mathbb{R}^+ of) a function $H^s(\Omega)$ -analytic on \mathcal{U} . Then the optimal control ϕ_* is (the restriction to \mathbb{R}^+ of) a function

$H^{s'}(\partial\Omega)$ -analytic on

$$(5.4) \quad \mathcal{U} := \{t \in \mathcal{U} : 0 <_{\mathcal{U}} t \text{ (with respect to some } \alpha < \bar{\alpha})\}$$

for $s' := s + 2 + \hat{m} - m'$ (if this is not an integer and subject to the smoothness assumed for $\partial\Omega$, the coefficients and κ, λ).

Proof. This is really a corollary to Theorems 2 and 3. By (5.4) one has

$$\mathcal{U} = \bigcup_{T>0} \mathcal{U}_T, \quad (\mathcal{U}_T \text{ given with respect to } (0, T) \text{ by (4.20)},$$

and, for each $T > 0$, applying Theorem 2 to the minimizer $\phi_{*,T}$ of J_T , given by (5.3), one has the desired analyticity on \mathcal{U}_T of $\hat{\phi}_{*,T}$. Clearly, for $0 < T < T'$ one has:

$$\hat{\phi}_{*,T'}|_{(0,T)} = \phi_{*,T}|_{(0,T)} = \phi_*|_{(0,T)} = \hat{\phi}_{*,T}|_{(0,T)}$$

by Theorem 3, so, by analyticity,

$$\hat{\phi}_{*,T'}|_{\mathcal{U}_T} = \hat{\phi}_{*,T}, \quad 0 < T < T'.$$

Thus, there is a maximally defined extension ϕ_* which is $H^{s'}(\Omega)$ -analytic on \mathcal{U} and $\phi_* = \hat{\phi}_*|_{\mathbb{R}^+}$. \square

There is a certain sense in which a representation of the form (2.49), (2.50) holds in the infinite horizon case. Observe that for h in \mathfrak{H}_J and r in \mathbb{R} we have (2.20), so ϕ_* minimizes J (if and) only if

$$(5.5) \quad 0 = \langle \phi_*, h \rangle_{\kappa} + \langle u_* - u_T, \mathbf{B}_h \rangle_{\lambda} \quad \text{for } h \text{ in } \mathfrak{H}_J.$$

Set $\tilde{\phi}_* := \kappa^{1/2} \phi_*$, $\tilde{h} := \kappa^{1/2} h$ (so \tilde{h} is in $L^2(\mathcal{S})$ for h in \mathfrak{H}_{κ}) and (formally) set $\tilde{\mathbf{B}} := \lambda^{1/2} \mathbf{B} \kappa^{-1/2}$. Then

$$\|\mathbf{B}h\|_{\lambda}^2 = \|\tilde{\mathbf{B}}\tilde{h}\|_{L^2(\mathcal{Q})}^2 \quad \text{for } \tilde{h} \in L^2(\mathcal{S}) \text{ such that } \tilde{\mathbf{B}}\tilde{h} \in L^2(\mathcal{Q})$$

and we interpret $\tilde{\mathbf{B}}$ as an operator (possibly unbounded) from $L^2(\mathcal{S})$ to $L^2(\mathcal{Q})$, noting that \tilde{h} is in $\mathfrak{D}(\tilde{\mathbf{B}})$ if and only if $h := \kappa^{-1/2} \tilde{h}$ is in \mathfrak{H}_J and

$$\|h\|_J^2 = \|\tilde{h}\|_{L^2(\mathcal{S})}^2 + \|\tilde{\mathbf{B}}\tilde{h}\|_{L^2(\mathcal{Q})}^2.$$

Then (5.5) becomes

$$(5.6) \quad \langle \tilde{\phi}_*, \tilde{h} \rangle_{L^2(\mathcal{S})} = \langle \lambda^{1/2}[u_T - u_*], \tilde{\mathbf{B}}\tilde{h} \rangle_{L^2(\mathcal{Q})} \quad \text{for } h \in \mathfrak{D}(\tilde{\mathbf{B}}).$$

If we now assume that $\mathfrak{D}(\tilde{\mathbf{B}})$ is dense in $L^2(\mathcal{S})$ —this can be verified in some cases but it is not clear whether or not this assumption might *always* hold—then the continuity in \tilde{h} over $L^2(\mathcal{S})$ of the left-hand side of (5.6) ($J(\phi_*) < \infty$ implies that $\|\phi_*\|_{\kappa} < \infty$, so $\tilde{\phi}_*$ is in $L^2(\mathcal{S})$) ensures that $\lambda^{1/2}[u_T - u_*]$ must be in the domain of $\tilde{\mathbf{B}}^*$. Thus, (5.6) gives us

$$(5.7) \quad \tilde{\phi}_* = \tilde{\mathbf{B}}^* \lambda^{1/2}[u_T - u_*].$$

Now, if ϕ_1 were any (known) control for which one might know that

$$\lambda^{1/2}[u_T - u_1] \in \mathfrak{D}(\tilde{\mathbf{B}}^*),$$

with u_1 the associated solution of (2.1)–(2.3), then (5.7) gives

$$(\tilde{\phi}_* - \tilde{\phi}_1) + \tilde{\mathbf{B}}^* \lambda^{1/2}(u_* - u_1) = \tilde{\psi}_0 := \tilde{\mathbf{B}}^* \lambda^{1/2}[u_T - u_1]$$

with $\tilde{\phi}_1 := \kappa^{1/2} \phi_1$. Now linearity of (2.1)–(2.3) shows that

$$u_* - u_1 = \mathbf{B}(\phi_* - \phi_1) = \mathbf{B} \kappa^{1/2}(\tilde{\phi}_* - \tilde{\phi}_1),$$

whence

$$(5.8) \quad (\mathbf{I} + \mathbf{T})(\tilde{\phi}_* - \tilde{\phi}_1) = \tilde{\psi}_0 := \kappa^{-1/2} \mathbf{B}^* \boldsymbol{\lambda} [u_T - u_1],$$

with \mathbf{T} as before, provided we can legitimately interpret $\tilde{\mathbf{B}}^* = (\boldsymbol{\lambda}^{1/2} \mathbf{B} \kappa^{-1/2})^*$ as $(\kappa^{-1/2} \mathbf{B}^* \boldsymbol{\lambda}^{1/2})$, etc. Note that (5.8) would reduce to (2.50) if one could take $\phi_1 = 0$, i.e., if we had z_0 in $\mathfrak{D}(\kappa^{-1/2} \mathbf{B}^*)$. Although this derivation of (5.8) has proceeded fairly formally, it seems plausible that an argument along these lines could be developed to prove Theorem 4 more directly, without appealing to Theorem 2, but paralleling it, and perhaps without requiring the hypothesis of nullcontrollability which here seems gratuitous.

Another interesting viewpoint is available for completely autonomous problems in which not only the coefficients of (\mathbf{A}, \mathbf{B}) but also $\kappa(\cdot)$, $\boldsymbol{\lambda}(\cdot)$, f and u_T are independent of t . We consider the case of “pure stabilization” ($u_T = 0, f = 0$), so J has the form

$$(5.9) \quad J(\phi) := \int_0^\infty [|\kappa^{1/2} \phi(t)|_0^2 + |\boldsymbol{\lambda}^{1/2} u(t)|_0^2]$$

with u determined by

$$(5.10) \quad \dot{u} = \mathbf{A}u \quad \text{on } \mathcal{Q}, \quad \beta u = \phi \quad \text{on } \mathcal{S}, \quad u(0) = \omega_0.$$

As ω_0 is now the only extraneous inhomogeneity introduced into (5.9), (5.10), one clearly must have ϕ_* linearly dependent on ω_0 and so, in particular, one may be able to write

$$(5.11) \quad \phi_*(0) = \mathbf{K}\omega_0.$$

Under appropriate hypotheses it can be shown that

$$(5.12) \quad \mathbf{K}: L^2(\Omega) \rightarrow L^2(\partial\Omega) \quad \text{is continuous.}$$

(See, e.g., [12] for an argument, given there in the one-dimensional case, based on consideration of the regularity of ϕ_* along rather different lines from the techniques of this paper.) Given (5.11), (5.12), we observe that the assumption that the problem is completely autonomous means that (5.10) and minimization of J can be “re-started” at any t in *exactly the same form* so that, by a standard embedding argument, (5.11) immediately generalizes to

$$(5.13) \quad \phi_*(t) = \mathbf{K}u_*(t)$$

along the optimally controlled trajectory u_* . (Compare the “decoupling” argument for this feedback in [5].) Substituting (5.13) into (5.10) produces the pure initial value problem

$$(5.14) \quad \dot{u} = \mathbf{A}u \quad \text{on } (0, \infty) \times \Omega,$$

$$(5.15) \quad \beta u = \mathbf{K}u \quad \text{in } L^2(\partial\Omega) \text{ for } t > 0,$$

$$(5.16) \quad u(0) = \omega_0$$

which determines u_* . The solution semigroup $\mathbf{S}_\mathbf{K}(t): \omega_0 \mapsto u(t)$ for (5.14)–(5.16) acts on $L^2(\Omega)$ and has as its infinitesimal generator $\mathbf{A}_\mathbf{K}$ —which is simply \mathbf{A} taken with the domain

$$(5.17) \quad \mathfrak{D}(\mathbf{A}_\mathbf{K}) := \{\omega \in H^2(\Omega): (\beta - \mathbf{K})\omega = 0\} \subset L^2(\Omega).$$

We know that \mathbf{A}_0 ($=\mathbf{A}_0$, as in (3.1), (3.2)) generates a homomorphic semigroup and now observe that, with \mathbf{G} given by (3.4), so that

$$\mathbf{A}\mathbf{G} = \mathbf{0}, \quad \beta\mathbf{G} = \mathbf{1} \quad \text{on } L^2(\partial\Omega),$$

one has

$$(5.18) \quad \mathbf{A}_\mathbf{K} = \mathbf{A}_0(\mathbf{1} - \mathbf{G}\mathbf{K}).$$

($\omega \in \mathfrak{D}(\mathbf{A}_0(\mathbf{1} - \mathbf{G}\mathbf{K}))$ if and only if $(\mathbf{1} - \mathbf{G}\mathbf{K})\omega \in \mathfrak{D}(\mathbf{A}_0)$ if and only if $(\beta - \mathbf{K})\omega = \beta(\mathbf{1} - \mathbf{G}\mathbf{K})\omega = 0$.) Since \mathbf{G} (hence $\mathbf{G}\mathbf{K}$) is continuous to $H^{\hat{m}+1/2}(\Omega)$, one has $\mathbf{G}\mathbf{K}$ compact on $L^2(\Omega)$. Thus [18, Thm. 1] applies to show that $\mathbf{S}_\mathbf{K}$ is also a holomorphic semigroup (and a careful analysis even shows that the domain of analyticity of $\mathbf{S}_\mathbf{K}$ is the same as that of \mathbf{S}_0). It follows that u_* is analytic in t . As $u_*(t)$ is in $\mathfrak{D}(\mathbf{A}_\mathbf{K})$ for t in $\Sigma_{\bar{\alpha}}$, we can apply β to get ϕ_* (now essentially defined as βu_*). Thus ϕ_* must also be analytic on $\Sigma_{\bar{\alpha}}$ —indeed, $H^s(\partial\Omega)$ -analytic for arbitrarily large s (subject to the regularity of $\partial\Omega$, etc.), since $u_*(t)$ will actually be in $\mathfrak{D}(\mathbf{A}_\mathbf{K}^n)$ for $n = 1, 2, \dots$.

Note that this argument (neglecting, for the present, any hypotheses needed to establish (5.12)) does not explicitly require any nullcontrollability hypothesis for (\mathbf{A}, β) , nor would it require boundedness of $\kappa, \kappa^{-1/2}, \lambda$ on $H^\sigma(\partial\Omega), H^{\sigma'}(\Omega)$, except for minimal σ, σ' .

6. Remarks

Remark 6.1. It is clear that the techniques of the present analyses could also be applied to more general cost criteria than given by (2.8)–(2.10). For example, one could certainly let $\|\cdot\|_\kappa$ be modeled on an $H^{s,s'}(\mathcal{S})$ -norm ($s, s' \geq 0$) rather than on the $L^2(\mathcal{S})$ -norm. Without change in κ, λ one could take μ unbounded on $L^2(\Omega)$ without change in the analysis, provided one would still have (2.21). This was not the case for $\mu = \infty$, which required a separate analysis using Lemma 2.40, but would be the case for suitable u_0, ω_T if, e.g., $\|\cdot\|_\mu$ were (dominated by) an $H^s(\Omega)$ -norm since ϕ in $H^{\sigma,\sigma/2}(\mathcal{S})$ for $\frac{1}{2} < \sigma := s - \hat{m} + \frac{1}{2}$, which is dense in $L^2(\mathcal{S})$, would then ensure $\mathbf{B}_T\phi$ in $\mathfrak{D}(\mu)$ ([7, Thm. 4.2.1], etc.). For such μ we must then modify the statement of Theorem 1 to assert that $[u_0(T) - \omega_T]$ is in $\mathfrak{D}([\mu\mathbf{B}_T]^*)$ and suitably interpret $(\mu\mathbf{B}_T)^*$ via (2.33)–(2.35). It is also not hard to verify that for cases N and B (but not, apparently, for case D) we could consider, e.g.,

$$J(\phi) = \int_{\mathcal{S}} \phi^2 + \int_{\mathcal{Q}} \lambda |\nabla(u - u_T)|^2 + \int_{\Omega} \mu |u(T) - \omega_T|^2$$

by the present methods. We also note that analysis along the present lines would also be possible for boundary control problems involving parabolic systems or higher order equations (e.g., taking \mathbf{A} to be $-\Delta^2$) for which one might again have a corresponding holomorphic semigroup.

Remark 6.2. Another modification of the problem would be to consider linear constraints on ϕ of the form:

$$(6.3) \quad \phi(t) \in \mathfrak{L} \quad \text{for } 0 < t < T,$$

where \mathfrak{L} is a suitable closed subspace of $L^2(\partial\Omega)$. In particular, the two most interesting cases seem to be:

- (i) \mathfrak{L} finite dimensional— $\mathfrak{L} := \text{sp}\{c_1, \dots, c_N\}$,
 (6.4) (ii) patch control—an “active” subset $\Gamma_a \subset \partial\Omega$ specified ($\partial\Omega \setminus \Gamma_a =: \Gamma_p$, the “passive” subset) with $\mathfrak{L} := \{\phi \in L^2(\Omega); \phi|_{\Gamma_p} = 0\}$.

Relevant considerations for feedback control corresponding to Remark 6.4(i) have been discussed in, e.g., [15]; observability and controllability questions for Remark 6.4(ii) have been considered in [9], [10], [11].

To adapt the analyses of §§ 2–5 to this case we take $\beta = \beta_{\mathfrak{L}}$ to have \mathfrak{L} as its codomain, so that \mathbf{B} is replaced by its restriction $\mathbf{B}_{\mathfrak{L}}$ to $L^2((0, T) \rightarrow \mathfrak{L})$. Theorem 1 then holds with this modification. We easily verify the identity (2.28) with this adaptation of $\beta_{\mathfrak{L}}$, with $\hat{\beta}$ as before and with $\beta' = \beta'_{\mathfrak{L}}$ now interpreted as the original β' (given by (2.26)) followed by the $L^2(\partial\Omega)$ -orthogonal projection $\mathbf{P}_{\mathfrak{L}}$ onto \mathfrak{L} ; with these interpretations we again have (2.32), (2.36). All the subsequent analysis then proceeds essentially unchanged provided that

$$(6.5) \quad \mathbf{P}_{\mathfrak{L}} : H^{\sigma}(\partial\Omega) \rightarrow H^{\sigma}(\partial\Omega) \quad \text{for suitable } \sigma.$$

In the case of Remark 6.4(i) this just means that the functions c_1, \dots, c_N must each be in $H^{\sigma}(\partial\Omega)$. For Remark 6.4(ii), $\mathbf{P}_{\mathfrak{L}}$ is just restriction to Γ_a and (6.5) will hold if we re-interpret the *codomain* $H^{\sigma}(\partial\Omega)$ in (6.5) as really meaning $H^{\sigma}(\Gamma_a)$; one then would use a corresponding interpretation for the results so that, e.g., Theorem 2 would assert $H^s(\Gamma_a)$ -analyticity for ϕ_* . Note that the exact nullcontrollability hypothesis, used for the consideration of $\mu = \infty$ and of infinite horizon problems, is known [11] to hold for certain situations of the form of Remark 6.4(ii) but cannot be expected to hold for the case of Remark 6.4(i) in more than one space dimension.

Remark 6.6. Throughout the paper we have required “sufficient smoothness” of $\partial\Omega$ and the coefficients of \mathbf{A}, β . Looking back through the arguments, it should be clear that this was used repeatedly, but primarily to ensure the continuity and/or invertibility of the map

$$(6.7) \quad (\mathbf{A}, \beta) : u \mapsto (\mathbf{A}u, \beta u) : H^{\sigma}(\Omega) \rightarrow H^{\sigma-2}(\Omega) \times H^{\sigma-1/2-\hat{m}}(\partial\Omega)$$

and similarly for $(\mathbf{A}^*, \hat{\beta}), \beta'$ for only a few distinct values of σ (all between $\frac{1}{2}$ and $\frac{5}{2}$ apart from the considerations of Lemma 4.19), with the estimates on the inverse corresponding to (3.2). Although [6], [7] impose a uniform hypothesis of C^{∞} regularity, the results are used freely here with the observation that (as is noted there!) consideration of (6.7) for any fixed σ (or bounded range for σ) clearly can only require a “finite degree of regularity.”

Acknowledgments. The original impetus to [12] and so to this work was a conversation with V. J. Mizel. Other discussions which have greatly contributed to this development have been with H. O. Fattorini, J. Henry, I. Lasiecka, V. J. Mizel, D. L. Russell, J. J. Schäffer, M. Sorine, R. Triggiani. Thanks are due to them, to Carnegie-Mellon University for its hospitality and stimulating atmosphere during my sabbatical leave and, for financial support, to the University of Maryland Baltimore County and the U.S. Army Research Office.

REFERENCES

- [1] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.
- [2] D. FUJIWARA, *Concrete characterization of the domains of fractional powers of some elliptic differential operators of the second order*, Proc. Japan. Acad., 43 (1967), pp. 82–86.
- P. GRISVARD, *Caractérisation de quelques espaces d'interpolation*, Arch. Rational Mech. Anal., 25 (1967), pp. 40–63.
- [3] I. LASIECKA, *Boundary control of parabolic systems: Regularity of optimal solutions*, Appl. Math. Optim., 4 (1978), pp. 301–327.
- [4] ———, *Unified theory for abstract parabolic boundary problems—a semigroup approach*, Appl. Math. Optim., 6 (1980), pp. 287–333.

- [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, *Non-Homogeneous Boundary Value Problems and Applications*, vol. I, Springer, New York, 1972.
- [7] ———, *Non-Homogeneous Boundary Value Problems and Applications*, vol. II, Springer, New York, 1972.
- [8] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Lecture Notes 10, Univ. Maryland, 1972.
- [9] D. L. RUSSELL, *Exact boundary value controllability theorems for wave and heat processes in star-complemented regions*, in *Differential Games and Control Theory*, Marcel Dekker, New York, 1974.
- [10] T. I. SEIDMAN, *Observation and prediction for the heat equation, IV: Patch observability and controllability*, this Journal, 15 (1977), pp. 412–427.
- [11] ———, *Exact boundary control for some evolution equations*, this Journal, 16 (1978), pp. 979–999.
- [12] ———, *Boundary control of $u_t = u_{xx} - qu$ and the analyticity of semigroups with distributed conditions*, Séminaires IRIA, 1979, pp. 97–105.
- [13] ———, *Regularity of optimal boundary controls for parabolic equations*, in *Lecture Notes in Control and Information Science*, 28, Bensoussan, Lions, eds., Springer-Verlag, Berlin, 1980, pp. 536–551.
- [14] H. TANABE, *On the equations of evolution in a Banach space*, Osaka J. Math., 12 (1960), pp. 363–376.
- [15] R. TRIGGIANI, *Well-posedness and regularity of boundary feedback parabolic systems*, J. Differential Equations, 36 (1980), pp. 347–362.
- [16] D. WASHBURN, *A bound on the boundary input map for parabolic equations with application to time optimal control*, this Journal, 17 (1979), pp. 652–671.
- [17] K. YOSIDA, *Functional Analysis*, Springer, New York, 1966.
- [18] J. ZABCZYK, *On decomposition of generators*, this Journal, 16 (1978), pp. 523–534.

A SUBGRADIENT ALGORITHM FOR CERTAIN MINIMAX AND MINISUM PROBLEMS—THE CONSTRAINED CASE*

JACQUES CHATELON†, DONALD HEARN‡ AND TIMOTHY J. LOWE§

Abstract. We present an implementable feasible direction subgradient algorithm for minimizing the maximum of a finite collection of functions subject to constraints. It is assumed that each function involved in defining the objective function is the sum of a finite collection of basic convex functions and that the number of different subgradient sets associated with nondifferentiable points of each basic function is finite on any bounded set. Problems involving functions of l_p -norms, such as location and approximation problems, can be put in this form. Conditions are given which guarantee that the algorithm generates a feasible sequence converging to an optimal solution. The results of computational tests on some location problems are included. In these tests we explore the sensitivity of the algorithm to its parameters.

1. Introduction. In this paper, we develop an implementable feasible direction algorithm for solving a class of nondifferentiable nonlinear programming problems of the form

$$(P) \quad \min F(x), \quad x \in C$$

where

$$F(x) \equiv \max \left\{ f_i(x) = \sum_{j=1}^l f_{ij}(x); i = 1, \dots, m \right\},$$

$$f_{ij} \text{ finite, convex, not necessarily differentiable,}$$

$$C \equiv \{x \in R^n; H(x) \leq 0\},$$

$$H(x) \equiv \max \{h_i(x); i = 1, \dots, I\},$$

$$h_i(x) \text{ convex on } R^n.$$

Our present work is related to an earlier paper [2] where we presented an algorithm for solving an unconstrained version of problem (P). The major differences in the algorithms, beyond the obvious due to the constraints, is that we demonstrate that the parameters in the algorithm can be taken to zero, thereby solving the problem optimally, as opposed to near optimally. An earlier version of the algorithm in this paper, where the parameters were fixed, appeared in [1].

Motivation for the work comes from the fact that there has been a proliferation of methods (see, e.g., [4], [7], [8], [14], [15], [18], [21], [22], [26], [28], [31], [34], [35]) designed for special cases of (P) in the location literature. Our view is that these problems have only a "small degree" of nondifferentiability and that a special property (see definition of LFS, below) can be employed to produce a single algorithm which performs effectively on variants of (P). As an example, the well-known location problem of Fermat (and Weber) and the Weiszfeld algorithm [33] for solving it, have received much attention (e.g., [13], [16], [17], [18], [27]). In § 5 we give a simple instance of this problem for which the Weiszfeld algorithm requires thousands of iterations to produce a solution accurate to four digits. Any descent subgradient algorithm with a line search, including ours, can solve this problem to the same accuracy in far fewer iterations. (The exact number depends on line search accuracy.)

* Received by the editors November 13, 1978, and in final revised form July 21, 1981. This work was supported in part by the National Science Foundation under grant 79-25065.

† CSEE, 17 Place Pernet, 75015 Paris, France.

‡ Department of Industrial and Systems Engineering, University of Florida, Gainesville, Florida, 32611.

§ Krannert Graduate School of Management, Purdue University, West Lafayette, Indiana 47907.

Our approach follows that of Dem'yanov and Malozemov [6] who employ subgradients to devise a descent algorithm for minimizing the maximum of continuously differentiable convex functions, i.e., problem (P) with all f_i differentiable and $C = R^n$. For related literature on nondifferentiable optimization methods, the surveys by Mifflin [24], [25] are recommended. He traces, e.g., the development of the descent methods of Lemarechal [19] and Goldstein [10]; the nondescent methods of Shor [32] and Polyak [29], implemented by Held and Karp [11] and Held, Wolfe and Crowder [12]; and the conjugate-type methods of Lemarechal [20] and Wolfe [36]. Mifflin's algorithm [24] for problems with "weakly upper semismooth" functions [25] builds on the notion of generalized gradients introduced by Clarke [3] for Lipschitz functions. While (P) is a special case of that problem type, the algorithm developed here differs in that it is a feasible direction method and is more easily implemented.

Denote n -dimensional Euclidean space as R^n and $\|x\|_p$ as the l_p -norm of $x \in R^n$, where $\|x\|$ is the l_2 -norm. Given a point $x \in R^n$, the Euclidean ball about x of radius $\eta \geq 0$ is $N(x, \eta)$; when $x = 0$ and $\eta = 1$, $B = N(0, 1)$ is the Euclidean unit ball. For a function f defined on R^n let $\partial f(x)$ be the subgradient set of f at x and let $f'(x, d)$ be the directional derivative of f at x in the direction d . Given a set $S \subset R^n$, $\text{Conv}(S)$ is the convex hull of S and $\text{Nr}(S)$ is the element of minimum Euclidean norm in S . Also denote $\partial f(S) \equiv \bigcup \{\partial f(x); x \in S\}$.

For convenience, the functions f_i in problem (P) are assumed to be the sum of exactly l functions f_{ij} , where perhaps for some i some of the f_{ij} functions are identically zero on R^n . Clearly, both F and H are continuous, finite, convex functions on R^n .

Given $\varepsilon \geq 0$ and $\mu \geq 0$, at any $x \in R^n$ define

$$R(x, \varepsilon) \equiv \{i \in \{1, 2, \dots, m\}; f_i(x) \geq F(x) - \varepsilon\},$$

and

$$Q(x, \mu) \equiv \{i \in \{1, 2, \dots, l\}; 0 \geq h_i(x) \geq -\mu\}.$$

With these definitions, we can interpret f_i , $i \in R(x, \varepsilon)$ as an " ε -binding" objective function at x ; and h_i , $i \in Q(x, \mu)$, as a " μ -binding" constraint at x . When $H(x) \leq -\mu$, we say that x is a μ -feasible point and the set of all x in R^n where $H(x) \leq -\mu$ is the μ -feasible set.

Given $x \in R^n$ and $\eta \geq 0$, define

$$G_{ij}(x, \eta) \equiv \{x\} \cup \{y; y \in N(x, \eta), f_{ij} \text{ not differentiable at } y\},$$

$$S_i(x, \eta) \equiv \sum_{j=1}^l \partial f_{ij}(G_{ij}(x, \eta)), \quad i = 1, 2, \dots, m.$$

In addition, let

$$S^1(x, \varepsilon, \eta) \equiv \bigcup \{S_i(x, \eta); i \in R(x, \varepsilon)\}.$$

To handle constraints, we use a procedure somewhat similar to that presented in [24]. At any $x \in R^n$, we consider subgradients of μ -binding constraints by defining $S^2(x, \mu) \equiv \bigcup \{\partial h_i(x); i \in Q(x, \mu)\}$ and letting $S(x, \varepsilon, \mu, \eta) \equiv \text{Conv} \{S^1(x, \varepsilon, \eta) \cup S^2(x, \mu)\}$. For an unconstrained problem, $S(x, \varepsilon, \mu, \eta) = S^1(x, \varepsilon, \eta)$ is precisely the enlargement of the subgradient set considered in [2].

This notation employs three parameters ε , η and μ for generality. However, nothing in the development precludes these parameters from being equal. When this is the case, we write $S_i(x, \varepsilon)$, $S^1(x, \varepsilon)$, etc., where $\varepsilon = \eta = \mu$. This convenience is employed in §§ 2 and 3.

We assume the functions f_{ij} are LFS (locally finitely subdifferentiable) [2], which means that in any closed bounded Euclidean ball, the number of different subgradient sets of f_{ij} corresponding to the points of nondifferentiability, is finite. In [2], we cite several examples of LFS functions, including examples from location theory and linear approximation problems.

Associated with $S(x, \varepsilon, \mu, \eta)$, the function Ψ measures the proximity of $S(\cdot)$ to zero:

$$\Psi(x, \varepsilon, \mu, \eta) \equiv \min \{ \max \{ (g, d); d \in S(x, \varepsilon, \mu, \eta) \}; g \in B \}.$$

It is easily established that

$$\Psi(x, \varepsilon, \mu, \eta) = -\| \text{Nr} (S(x, \varepsilon, \mu, \eta)) \|.$$

We note that Ψ is well defined since S is a nonempty, compact, convex subset of R^n . Further Ψ is always nonpositive. When $\Psi(x, \varepsilon, \mu, \eta) = 0$, we call x a *stationary* point, and any x where $\Psi(x, \varepsilon, \mu, \eta) < 0$ is a *nonstationary* point. In § 4 we give a result concerning stationarity. In what follows in this section, we consider the case where x is nonstationary.

Given a nonstationary point x , the subgradient sets in $S(x, \varepsilon, \mu, \eta)$ relative to the functions f_i ensure that we can find a descent direction, and the subgradient sets ∂h_i , if any, ensure that this descent direction is feasible.

THEOREM 1.1. *If $\Psi(x, \varepsilon, \mu, \eta) < 0$, there exists a feasible descent direction for F at x .*

Proof. Let $g_0 \neq 0$ be the element of minimum norm in $S(x, \varepsilon, \mu, \eta)$, i.e.,

$$0 > \Psi(x, \varepsilon, \mu, \eta) = -\|g_0\| = -\| \text{Nr} (S(x, \varepsilon, \mu, \eta)) \|.$$

Define $d_0 = -g_0/\|g_0\|$. If $H(x) = 0$, then for any i such that $h_i(x) = H(x) = 0$, $\partial h_i(x) \subset S(x, \varepsilon, \mu, \eta)$ so that

$$\begin{aligned} h'_i(x, d_0) &= \max \{ (g, d_0); g \in \partial h_i(x) \} \leq \max \{ (g, d_0); g \in S(x, \varepsilon, \mu, \eta) \} \\ &= -\min \{ (g, -d_0); g \in S(x, \varepsilon, \mu, \eta) \} \\ &= -1/\|g_0\| \min \{ (g, g_0); g \in S(x, \varepsilon, \mu, \eta) \} = -\|g_0\| < 0. \end{aligned}$$

Hence, d_0 is a feasible direction at x . On the other hand, if $H(x) < 0$, the direction d_0 is feasible since the functions h_i are continuous. In either case above, since $\partial F(x) \subset S(x, \varepsilon, \mu, \eta)$ it follows from Theorem 3.3 of [2] that $F'(x, d_0) < 0$. Therefore d_0 is a descent direction for F at x . \square

The remainder of the paper is organized as follows. In § 2 we state our feasible direction subgradient algorithm. Section 3 contains the convergence proof. In the convergence results, we initially assume that a parameter converges to zero. We then give a sufficient condition via an assumption which guarantees the convergence of the parameter to zero. Section 4 contains the stationarity results and § 5 concludes the paper with computational experience.

2. The algorithm. The algorithm exploits the results of the previous section. In order to facilitate a tidy convergence proof, at any step in the algorithm, the parameters ε , μ , and η are identical, where we use the symbol ε to stand for all three parameters. Due to this construct, $Q(x, \varepsilon)$, $G_{ij}(x, \varepsilon)$, $S^1(x, \varepsilon)$, $S(x, \varepsilon)$ and $\Psi(x, \varepsilon)$ are well defined.

As an initialization step, choose β , $0 < \beta < 1$, $\varepsilon_0 > 0$ and x_0 a feasible starting point. Set $k = 0$ and go to Step 1.

Step 1. At x_k , find $F(x_k)$, $R(x_k, \varepsilon_k)$ and $Q(x_k, \varepsilon_k)$.

Calculate $S(x_k, \varepsilon_k)$ and $\Psi(x_k, \varepsilon_k)$.

If $\Psi(x_k, \varepsilon_k) = 0$, go to Step 4.

Otherwise, go to Step 2.

Step 2. If $\Psi(x_k, \varepsilon_k) < -\varepsilon_k$, set $\varepsilon_{k+1} = \varepsilon_k$ and go to Step 3.

If $\Psi(x_k, \varepsilon_k) \geq -\varepsilon_k$, set $\varepsilon_{k+1} = \beta \varepsilon_k$ and go to Step 1.

Step 3. Define g_k as the element of minimum norm in $S(x_k, \varepsilon_k)$ and let $d_k = -g_k/\|g_k\|$. Perform a restricted line search along d_k , finding t_k such that

$$F(x_k + t_k d_k) = \min \{F(x_k + t d_k); t \geq 0, H(x_k + t d_k) \leq 0\}.$$

Set $x_{k+1} = x_k + t_k d_k$ and $k = k + 1$ and return to Step 1.

Step 4. If $\Psi(x_k, 0) = 0$, stop.

Otherwise, set $x_{k+1} = x_k$, $\varepsilon_{k+1} = \beta \varepsilon_k$, and $k = k + 1$ and go to Step 1.

The condition $\Psi(x, 0) = 0$ given in Step 4 implies (under a regularity assumption concerning the feasible region) that x is an optimal solution to problem (P). This follows from Corollary 4.2 given in § 4.

In the next section, under supplementary assumptions, we prove that any limit point, x_* , of the algorithm satisfies $\Psi(x_*, 0) = 0$ and thus is an optimal solution to problem (P).

3. Proof of convergence. In this section, we assume that the algorithm does not stop, but generates an infinite sequence $\{x_k\}$, $k \in K$ converging to some limit x_* . In addition to the assumption that the functions f_{ij} are LFS, we make two additional assumptions in this section. The first (Assumption 3.1) guarantees the existence of a solution to problem (P) and insures that a limit point exists. The second (Assumption 3.2) is sufficient to guarantee that $\{\varepsilon_k\} \rightarrow 0$.

ASSUMPTION 3.1. *There exists some $x_0 \in C$, a starting point for the algorithm, such that the intersection, X , of C with the level set $\{x \in R^n: F(x) \leq F(x_0)\}$ is nonempty and bounded. By continuity of F and H , X is also closed. We further assume that there exists an upper bound, δ , on the norm of any subgradient of any function f_{ij} at any point in C .*

Lemmas 3.1 and 3.2 are general results which hold for any sequence $\{x_k\} \rightarrow x_*$. Lemma 3.2 exploits the LFS property and makes use of results in [2].

LEMMA 3.1. *If $\{x_k\} \rightarrow x_*$, then for any fixed $\varepsilon \geq 0$, $Q(x_k, \varepsilon) \subset Q(x_*, \varepsilon)$, for all k sufficiently large.*

Proof. The lemma follows from the continuity of the functions h_i , $i = 1, \dots, I$. \square

We now show that with $\varepsilon > 0$, fixed, the sets $S(x_k, \varepsilon)$ approximate the set $S(x_*, \varepsilon)$.

LEMMA 3.2. *If $\{x_k\} \rightarrow x_*$, then for any fixed $\varepsilon > 0$, there exists $N_1(\varepsilon)$ such that*

$$S(x_k, \varepsilon) \subset S(x_*, \varepsilon) + \varepsilon \beta, \quad k > N_1(\varepsilon).$$

Proof. By definition,

$$(3.1) \quad S(x_k, \varepsilon) = \text{Conv}(S^1(x_k, \varepsilon) \cup S^2(x_k, \varepsilon)).$$

Since $S^1(x_k, \varepsilon)$ is identical to $S(x_k, \varepsilon, \eta)$ in [2] (with $\eta = \varepsilon$), by [2, Theorem 5.2] there exists $N'_1(\varepsilon)$ such that

$$(3.2) \quad S^1(x_k, \varepsilon) \subset S^1(x_*, \varepsilon) + \varepsilon B, \quad k > N'_1(\varepsilon).$$

Now consider the sets $S^2(\cdot)$ in (3.1). From [30, Corollary 24.5.1] since the h_i are proper convex functions, for each $i \in Q(x_*, \varepsilon)$, there exists $L_i(\varepsilon)$ such that

$$(3.3) \quad \partial h_i(x_k) \subset \partial h_i(x_*) + \varepsilon B, \quad k > L_i(\varepsilon).$$

It follows from the definition of $S^2(\cdot)$, (3.3) and Lemma 3.1 that there exists $N_1''(\varepsilon)$ such that

$$(3.4) \quad S^2(x_k, \varepsilon) \subset S^2(x_*, \varepsilon) + \varepsilon B, \quad k > N_1''(\varepsilon).$$

If we let $N_1(\varepsilon) = \max\{N_1'(\varepsilon), N_1''(\varepsilon)\}$ and use (3.2), (3.3) and (3.4), the result follows. \square

In addition to the condition $\{x_k\} \rightarrow x_*$, $k \in K$, we now assume that $\{\varepsilon_k\} \rightarrow 0$, $k \in K$. We note that from the statement of the algorithm, these conditions imply $\Psi(x_k, \varepsilon_k) \rightarrow 0$, $k \in K$.

COROLLARY 3.3. *If $\{x_k\} \rightarrow x_*$ and $\{\varepsilon_k\} \rightarrow 0$, $k \in K$, then given any $k \in K$ there exists $l(k) \in K$, $l(k) > k$, where*

$$(3.5) \quad S(x_{l(k)}, \varepsilon_k) \subset S(x_*, \varepsilon_k) + \varepsilon_k B.$$

Proof. The corollary follows from Lemma 3.2 and choosing $l(k) \in K$ such that $l(k) > \max\{k, N_1(\varepsilon_k)\}$. \square

The next result shows that the function Ψ evaluated at x_* and on the sequence $\{\varepsilon_k\}$, $k \in K$, converges to 0.

LEMMA 3.4. *If $\{x_k\} \rightarrow x_*$, and $\{\varepsilon_k\} \rightarrow 0$, $k \in K$, then*

$$\lim \Psi(x_*, \varepsilon_k) = 0.$$

Proof. From Corollary 3.3, since $l(k) > k$ for any $k \in K$, it follows that $\varepsilon_{l(k)} \leq \varepsilon_k$, $k \in K$. Using the definition of $S(\cdot)$ and (3.5) gives

$$(3.6) \quad S(x_{l(k)}, \varepsilon_{l(k)}) \subset S(x_{l(k)}, \varepsilon_k) \subset S(x_*, \varepsilon_k) + \varepsilon_k B, \quad k \in K.$$

For fixed $k \in K$, from (3.6),

$$\begin{aligned} \Psi(x_{l(k)}, \varepsilon_{l(k)}) &= \min \{ \max \{ (g, d); g \in S(x_{l(k)}, \varepsilon_{l(k)}) \}; d \in B \} \\ &\leq \min \{ \max \{ (g, d); g \in (S(x_*, \varepsilon_k) + \varepsilon_k B) \}; d \in B \} \\ &\leq \min \{ \max \{ (g, d); g \in S(x_*, \varepsilon_k); d \in B \} \} + \varepsilon_k \\ &= \Psi(x_*, \varepsilon_k) + \varepsilon_k, \end{aligned}$$

or

$$(3.7) \quad \Psi(x_{l(k)}, \varepsilon_{l(k)}) - \varepsilon_k \leq \Psi(x_*, \varepsilon_k).$$

Since (3.7) is true for arbitrary $k \in K$,

$$(3.8) \quad \lim \Psi(x_{l(k)}, \varepsilon_{l(k)}) - \lim \varepsilon_k \leq \lim \Psi(x_*, \varepsilon_k).$$

Since $l(k) \in K$, it follows that $\lim \Psi(x_{l(k)}, \varepsilon_{l(k)}) = 0$. By hypothesis, $\lim \{\varepsilon_k\} = 0$. Further, for any k , $\Psi(x_*, \varepsilon_k) \leq 0$ and thus it follows from (3.8) that

$$\lim \Psi(x_*, \varepsilon_k) = 0. \quad \square$$

Prior to proving that $\Psi(x_*, 0) = 0$, we need the following lemma which exploits the LFS property, generalizing an earlier result in [2].

LEMMA 3.5. *Let f be an LFS function and suppose $\{\varepsilon_k\} \rightarrow 0$. Given any x , there exists $N_2(x)$ such that*

$$\partial f(G(x, \varepsilon_k)) \subset \partial f(G(x, 0)) = \partial f(x), \quad k > N_2(x).$$

Proof. Define $G'(x, \varepsilon_k) = G(x, \varepsilon_k)/x$ and

$$M(k) = \bigcup \{ \partial f(G'(x, \varepsilon_l)); l \geq k \} = \partial f(\bigcup \{ G'(x, \varepsilon_l); l \geq k \}).$$

It follows that, for all k ,

$$(3.9) \quad \partial f(G'(x, \varepsilon_k)) \subset M(k),$$

$$(3.10) \quad \partial f(G(x, \varepsilon_k)) = \partial f(G'(x, \varepsilon_k)) \cup \partial f(x).$$

Since the $M(k)$ are nested (i.e., $k_2 > k_1$ implies $M(k_2) \subset M(k_1)$), if $M(\bar{k}) = \emptyset$ for some \bar{k} , then we can choose $N_2(x) = \bar{k}$ since by (3.9), $\partial f(G'(x, \varepsilon_k)) = \emptyset$ for all $k > N_2(x)$, in which case the conclusion follows from (3.10).

Thus suppose $H(k) \neq \emptyset$ for all k . The set $\cup\{G'(x, \varepsilon_l); l \geq k\}$ is contained in a ball about x and by the LFS property, $M(k)$ is the union of finitely many subgradient sets. Therefore, for sufficiently large k , there exists subgradient sets V^r , $r = 1, 2, \dots, q$ such that

$$M(k) \subset \cup\{V^r, r = 1, \dots, q\},$$

where for each r , there exists an infinite subsequence K_r where $V^r = \partial f(y_k)$ for some $y_k \in G'(x, \varepsilon_k)$, $k \in K_r$. But, the sequence $\{y_k\}$, $k \in K_r$ is on a compact set and since $\{\varepsilon_k\} \rightarrow 0$, it follows that $\{y_k\} \rightarrow x$ (at least on a subsequence).

Since f is a proper convex function, $\partial f(y_k) \subset \partial f(x) + \varepsilon' B$ for any $\varepsilon' > 0$ and sufficiently large $k \in K_r$. However, $V^r = \partial f(y_k)$ is constant on K_r , so, in fact, $V^r \subset \partial f(x)$. Thus, for sufficiently large k ,

$$\partial f(G'(x, \varepsilon_k)) \subset M(k) \subset \cup\{V^r, r = 1, \dots, q\} \subset \partial f(x),$$

or, from (3.10),

$$\partial f(G(x, \varepsilon_k)) \subseteq \partial f(x).$$

The existence of $N_2(x)$ follows. \square

The main theorem now follows.

THEOREM 3.6. *If $\{x_k\} \rightarrow x_*$ and $\{\varepsilon_k\} \rightarrow 0$, $k \in K$, then*

$$\Psi(x_*, 0) = 0.$$

Proof. For sufficiently large k ,

$$(3.11) \quad R(x_*, \varepsilon_k) = R(x_*, 0),$$

$$(3.12) \quad Q(x_*, \varepsilon_k) = Q(x_*, 0).$$

From Lemma 3.5, the definition of $S^1(\cdot)$, the fact that the f_{ij} are LFS, and (3.11), it follows that for sufficiently large k ,

$$(3.13) \quad S^1(x_*, \varepsilon_k) \subset S^1(x_*, 0),$$

and thus from (3.12) and (3.13), for sufficiently large k ,

$$(3.14) \quad \begin{aligned} S(x_*, \varepsilon_k) &= \text{Conv}(S^1(x_*, \varepsilon_k) \cup S^2(x_*, \varepsilon_k)) \\ &\subset \text{Conv}(S^1(x_*, 0) \cup S^2(x_*, 0)) = S(x_*, 0). \end{aligned}$$

But then (3.14) implies that for large k

$$\Psi(x_*, \varepsilon_k) \leq \Psi(x_*, 0).$$

From Lemma 3.4, $\lim \Psi(x_*, \varepsilon_k) = 0$, and so since $\Psi(x_*, 0) = 0$, the conclusion follows. \square

In order to establish Theorem 3.6, we assumed that the sequence $\{\varepsilon_k\}$ converged to 0 as the sequence $\{x_k\}$ converged to x_* . In what follows, we provide a sufficient condition for $\{\varepsilon_k\} \rightarrow 0$ through the following assumption.

ASSUMPTION 3.2. The constraint functions h_i , $i = 1, \dots, I$, are continuously differentiable.

LEMMA 3.7. Suppose $\{\varepsilon_k\} \rightarrow \bar{\varepsilon} > 0$. Then there exist some N_3 and $T > 0$ (where T is independent of k), such that for all $k > N_3$ (with d_k chosen as in Step 3 of the algorithm),

$$H(x_k + td_k) \leq 0,$$

$$F(x_{k+1}) \leq F(x_k + td_k) \quad \text{for all } t \in [0, T].$$

Proof. Since $\varepsilon_k \rightarrow \bar{\varepsilon} > 0$, there exists N_3 such that $\Psi(x_k, \varepsilon_k) < -\varepsilon_k = -\bar{\varepsilon}$, for all $k > N_3$. Thus consider $k > N_3$.

By Assumption 3.2, the h_i are continuously differentiable on R^n and thus from [6, Remark 2, p. 270], for any $i = 1, \dots, I$, there exists $T_{0i} > 0$ such that for all $t \in [0, T_{0i}]$,

$$(3.15) \quad h_i(x + td) = h_i(x) + t(\nabla h_i(x), d) + \sigma_i(x, d; t),$$

where σ_i is a function with the property that $\sigma_i(x, d; t)/t \rightarrow 0$ uniformly in $x \in X$ and $d, \|d\| = 1$, as $t \rightarrow 0$. Because of the uniform convergence of $\sigma_i(x, d; t)/t$, we can choose T'_{0i} , $0 < T'_{0i} \leq T_{0i}$, such that

$$(3.16) \quad t \in [0, T'_{0i}] \quad \text{implies} \quad |\sigma_i(x, d; t)| < \bar{\varepsilon}t/2,$$

for all $x \in X$ and all d where $\|d\| = 1$.

Letting $T'_0 \equiv \min \{T'_{0i}; i = 1, \dots, I\}$, it follows from (3.15) and (3.16) that for any $i = 1, \dots, I$, any $x \in X$ and any $d, \|d\| = 1$, if $t \in [0, T'_0]$, then

$$(3.17) \quad h_i(x + td) \leq h_i(x) + t(\nabla h_i(x), d) + \bar{\varepsilon}t/2.$$

At a point x_k , for any $i = 1, \dots, I$, either $i \in Q(x_k, \bar{\varepsilon})$ or not. If $i \in Q(x_k, \bar{\varepsilon})$, then with $h_i(x_k) \leq 0$ and choosing $d = d_k$ in (3.17),

$$(3.18) \quad h_i(x + td_k) \leq t(\nabla h_i(x_k), d) + \bar{\varepsilon}t/2.$$

Noting that $\nabla h_i(x_k) \in S(x_k, \bar{\varepsilon})$, with $k > N_3$, it follows that $(\nabla h_i(x_k), d_k) \leq \Psi(x_k, \bar{\varepsilon}) \leq -\bar{\varepsilon}$, and so from (3.18),

$$(3.19) \quad h_i(x_k + td_k) \leq -\bar{\varepsilon}t + \bar{\varepsilon}t/2 < 0.$$

Consider any $i \notin Q(x_k, \bar{\varepsilon})$. By uniform continuity of the functions h_i on the compact set C , there exists some $T''_0 > 0$ such that for all x, y in C , $\|x - y\| \leq T''_0$ implies

$$(3.20) \quad |h_i(x) - h_i(y)| < \bar{\varepsilon} \quad \text{for all } i = 1, \dots, I.$$

For any $i \notin Q(x_k, \bar{\varepsilon})$, $h_i(x_k) < -\bar{\varepsilon}$ and so from (3.20), with $t \in [0, T''_0]$,

$$(3.21) \quad h_i(x_k + td_k) \leq h_i(x_k) + \bar{\varepsilon} < -\bar{\varepsilon} + \bar{\varepsilon} = 0.$$

Letting $T = \min \{T'_0, T''_0\} > 0$, from (3.19) and (3.21) and the definition of H ,

$$H(x_k + td_k) \leq 0, \quad t \in [0, T].$$

The second conclusion of the lemma follows since Step 3 of the algorithm determines x_{k+1} , where

$$\begin{aligned} F(x_{k+1}) &= \min \{F(x_k + td_k); t \geq 0, H(x_k + td_k) \leq 0\} \\ &\leq \min \{F(x_k + td_k); t \in [0, T]\}. \end{aligned}$$

□

LEMMA 3.8. *Given any $\varepsilon > 0$ and $\alpha > 0$, there exists $N_4(\varepsilon, \alpha)$ such that for any $k > N_4(\varepsilon, \alpha)$, if $s \in \partial f_i(x_*)$, $i \in R(x_*, 0)$, then there exists an $s' \in S(x_k, \varepsilon)$ and a vector t where*

$$s = s' + t \quad \text{and} \quad \|t\| < \alpha.$$

Proof. The proof follows from [2, Lemma 5.4] and upon noting that $s' \in S^1(x_k, \varepsilon) \subset S(x_k, \varepsilon)$. \square

In the proof of $\{\varepsilon_k\} \rightarrow 0$, we make use of the following result from Cullum, Donath and Wolfe [5].

LEMMA 3.9. *Let F be convex on R^n and suppose that $\{x_k\} \rightarrow x_*$ and $\{d_k\} \rightarrow d_*$, where $F(x_{k+1}) \leq F(x_k + td_k)$, $0 \leq t \leq T$. Then $F'(x_*, d_*) \geq 0$.*

THEOREM 3.10. *If the h_i satisfy Assumption 3.2, then*

$$\{\varepsilon_k\} \rightarrow 0, \quad k \in K.$$

Proof. If not, then $\varepsilon_k \rightarrow \bar{\varepsilon} > 0$. Thus, as in Lemma 3.7,

$$\Psi(x_k, \varepsilon_k) < -\varepsilon_k = -\bar{\varepsilon}, \quad k > N_3.$$

Since the d_k are on a compact set B , there exists d_* such that $\{d_k\} \rightarrow d_*$ in B , and so for some N_5

$$(3.22) \quad l\delta \|d_k - d_*\| < \bar{\varepsilon}/4, \quad k > N_5.$$

The directional derivative of F at x_* is

$$F'(x_*, d_*) = \max \{(g, d_*): g \in \partial F(x_*)\} = (\sum \lambda_j s_j, d_*),$$

where $\sum \lambda_j s_j$ is a convex combination of elements s_j and each $s_j \in \partial f_i(x_*)$ for some $i \in R(x_*, 0)$. Let $N_4(\bar{\varepsilon}, \bar{\varepsilon}/4)$ be such that Lemma 3.8 holds with $\varepsilon = \bar{\varepsilon}$ and $\alpha = \bar{\varepsilon}/4$. Let N_3 be as in Lemma 3.7.

Consider any $k > \max \{N_3, N_4(\varepsilon, \bar{\varepsilon}/4), N_5\}$. Applying Lemma 3.8, we find

$$\begin{aligned} F'(x_*, d_*) &= (\sum \lambda_j s'_j, d_*) + (\sum \lambda_j t_j, d_*) \\ &= (\sum \lambda_j s'_j, d_k) + (\sum \lambda_j s'_j, d_* - d_k) + (\sum \lambda_j t_j, d_*) \\ &\leq (\sum \lambda_j s'_j, d_k) + (\sum \lambda_j s'_j, d_* - d_k) + \|\sum \lambda_j t_j\| \|d_*\| \\ &\leq (\sum \lambda_j s'_j, d_k) + (\sum \lambda_j s'_j, d_* - d_k) + \bar{\varepsilon}/4. \end{aligned}$$

By definition of d_k and since $\Psi(x_k, \varepsilon_k) < -\bar{\varepsilon}$,

$$(\sum \lambda_j s'_j, d_k) \leq \max \{(g, d_k); g \in S(x_k, \bar{\varepsilon})\} = \Psi(x_k, \bar{\varepsilon}) < -\bar{\varepsilon}.$$

Further, $\|s'_j\| \leq l\delta$ and so from (3.22),

$$(\sum \lambda_j s'_j, d_* - d_k) \leq \|\sum \lambda_j s'_j\| \|d_* - d_k\| \leq \bar{\varepsilon}/4.$$

Combining these results gives $F'(x_*, d_*) \leq -\bar{\varepsilon}/2 < 0$, which contradicts Lemma 3.9 (with T as in Lemma 3.7). Thus, $\{\varepsilon_k\} \rightarrow 0$, $k \in K$. \square

4. Stationarity. In this section, we return to the case where the parameters ε , μ , and η can be distinct. This makes the results more general and, in some cases, numerical considerations make this feature desirable. We now discuss the implications of stationarity as defined in § 1.

In the unconstrained problem, information given by the value of Ψ is relatively easy to exploit [2] since the set $S(x, \varepsilon, \eta)$ is derived solely from the f_i functions which are ε -binding. In the constrained problem this is no longer the case, since, to construct

$S(x, \varepsilon, \mu, \eta)$, we also consider the subgradient sets of the μ -binding constraints. Consequently, stationarity in the constrained problem will not always imply a lower bound on the minimum value of F on C , F^* , as is the case with the unconstrained problem. In the presence of constraints, we show in the next theorem that, while it may be possible to obtain a lower bound on F^* , it may be the case that the only implication of stationarity is the emptiness of the interior of the μ -feasible set.

THEOREM 4.1. *If $x \in C$ is such that $\Psi(x, \varepsilon, \mu, \eta) = 0$, then*

(a) *If $H(x) < -\mu$, then*

$$(4.1) \quad F(x) \geq F^* \geq F(x) - \varepsilon - 2l\eta\delta.$$

(b) *If $-\mu \leq H(x) \leq 0$, at least one of the following is true:*

(b1) (4.1) holds,

(b2) $\{z \in R^n; H(z) < -\mu\}$ is empty,

(b3) $F(z) \geq F(x) - \varepsilon - 2l\eta\delta$, for all μ -feasible z .

Proof. Stationarity at x is equivalent to $0 \in S(x, \varepsilon, \mu, \eta)$, which can occur if and only if there exists g_1, \dots, g_q , where $g_i \in \{S^1(x, \varepsilon, \eta) \cup S^2(x, \mu)\}$ and $0 \in \text{Conv}(g_1, \dots, g_q)$. Index the g_i so that $g_i \in S^1(x, \varepsilon, \eta)$ for $1 \leq i \leq q_1$ and $g_i \in S^2(x, \mu)$ for $q_1 + 1 \leq i \leq q$. Thus for i , $1 \leq i \leq q_1$, $g_i = \sum_{j=1}^l g_{ij}$, $g_{ij} \in \partial f_{(i),j}(N(x, \eta))$ where $f_{(i)}(x) \geq F(x) - \varepsilon$ and $\|x - y_i\| \leq \eta$; and for i , $q_1 + 1 \leq i \leq q$, $g_i \in \partial h_{(i)}(x)$, where $0 \geq h_{(i)}(x) \geq -\mu$.

We note in the case where $q_1 = q$ (which certainly holds in case (a)), that $g_i \in S^1(x, \varepsilon, \eta)$, $1 \leq i \leq q$, and thus by Theorem 3.2 of [2], (4.1) holds. Thus suppose $q_1 < q$, in which case $0 = q_1$ or $0 < q_1$.

If $0 = q_1$, then $g_i \in S^2(x, \mu)$ for all i . By the subgradient inequality, $h_{(i)}(z) \geq h_{(i)}(x) + (g_i, z - x)$, $z \in R^n$, and further $h_{(i)}(x) \geq -\mu$ and $H(z) \geq h_{(i)}(z)$. Thus

$$H(z) \geq h_{(i)}(z) \geq -\mu + (g_i, z - x), \quad z \in R^n.$$

Taking the convex combination over all $i = 1, \dots, q$, yields $H(z) \geq -\mu$, $z \in R^n$, which establishes case (b2).

Consider the remaining case, $1 \leq q_1 < q$. As in [2], since each function f_i is Lipschitz,

$$F(z) \geq F(x) - \varepsilon + (g_i, z - x) - 2l\eta\delta, \quad z \in R^n, \quad 1 \leq i \leq q_1;$$

thus for all $z \in R^n$,

$$(4.2) \quad F(z) \geq F(x) - \varepsilon - 2l\eta\delta + \max \{(g_i, z - x) : i = 1, \dots, q_1\}.$$

For any μ -feasible z , and $i = q_1 + 1, \dots, q$,

$$-\mu \geq h_{(i)}(z) \geq h_{(i)}(x) + (g_i, z - x) \geq -\mu + (g_i, z - x),$$

which implies $(g_i, z - x) \leq 0$. Writing zero as a convex combination of all g_i leads to $\max \{(g_i, z - x) : i = 1, \dots, q_1\} \geq 0$ for all μ -feasible z . Thus the final term of (4.2) may be deleted and case (b3) is established. \square

Remark 4.1. The definition of q_1 in the previous proof implies that if a point g_i is in both sets $S^1(x, \varepsilon, \eta)$ and $S^2(x, \mu)$, it should be considered as a point of $S^1(x, \varepsilon, \eta)$. This forces q_1 to be equal to q if possible and thus to obtain the more useful cases (a) or (b1) in which we have bounds on the constrained minimum of F .

Remark 4.2. Cases (b1), (b2), and (b3) are not mutually exclusive. When several different convex combinations of elements in $S^1(x, \varepsilon, \eta)$ and $S^2(x, \mu)$ are zero, (b1), (b2), and (b3) can occur simultaneously.

To illustrate the different situations described in Theorem 4.1 we consider the following example.

Example 4.1. Min $F(x)$, $F(x) \equiv \|x - (.5, -.5)\|$, subject to

$$h_1(x) \equiv \|x_1\|_1 - 1.6 \leq 0,$$

$$h_2(x) \equiv \|x - (1, -1)\|_1 - 1.6 \leq 0.$$

Let $\varepsilon = 0$, $\eta = \sqrt{2} \times 10^{-1}$. It is clear that $\delta = 1$ and $x^* = (.5, -.5)$ is the optimal solution to this problem. For different values of μ , stationarity may have distinct implications.

(a) Let $\mu = 0$. The point $y_0 = (.6, -.4)$ is feasible, with $H(y_0) = \max\{-.6, -.6\} = -.6 < -\mu$. y_0 is also stationary because $(.5, -.5)$ is in $N(y_0, \eta)$. We observe that $F(y_0) = \sqrt{2} \times 10^{-1}$, $F(y_0) - 2\eta\delta = -\sqrt{2} \times 10^{-1}$ and so,

$$\sqrt{2} \times 10^{-1} \geq 0 = F^* = \min \{F(z) : z \in C\} \geq -\sqrt{2} \times 10^{-1}.$$

(b) Let $\mu = 0.6$. y_0 remains stationary with $H(y_0) = -.6 = -\mu$. In this case, in addition to obtaining the lower bound on F^* , $\{z \in R^2 : H(z) < -.6\}$ is empty, and both (b1) and (b2) are obtained. On the other hand, observe that $y_1 = (1, 0)$ is stationary since $H(y_1) = h_1(y_1) = h_2(y_1) = -.6 = -\mu$ so that both points $(1, -1) \in \partial h_1(y_1)$ and $(-1, 1) \in \partial h_2(y_1)$ are in $S(y_1, \varepsilon, \mu, \eta)$. However, the only conclusion is that the set $\{z \in R^2 : H(z) < -.6\}$ is empty.

Now let $\mu = 0.6$, and add a third constraint, $h_3(x) = -x_1 - x_2 \leq 0$, to Example 4.1. Consider $y_2 = (.8, -.2)$. Since $H(y_2) = \max\{-.6, -.6, -.6\} = -.6 = -\mu$, y_2 is feasible. Moreover y_2 is stationary. It is easy to verify that $S(y_2, \varepsilon, \mu, \eta)$ is the square with vertices $(-1, 1)$, $(1, -1)$, $(-1, -1)$, $(1, 1)$. Here the lower bound obtained is only valid on the μ -feasible set, which is the line segment with end points y_2 and $(1, 0)$. With $F(y_2) = 3\sqrt{2} \times 10^{-1}$, this lower bound is positive and equal to $\sqrt{2} \times 10^{-1}$. Hence, at y_2 , both (b2) and (b3) hold.

Remark 4.3. It is clear that case (b2) of Theorem 4.1 is relatively undesirable since nothing can be said about the objective function itself. To discard this case, a constraint qualification, similar to Slater's constraint qualification [23], is necessary. Thus, if the set $\{z \in R^n : H(z) < -\mu\}$ is nonempty, case (b2) cannot occur.

We close this section with a corollary to Theorem 4.1, which establishes that any limit point of the algorithm given in § 2 under the assumptions outlined in § 3 is an optimal solution to problem (P).

COROLLARY 4.2. *If the set $\{z \in R^n : H(z) < 0\}$ is nonempty, and if $\varepsilon = \mu = \eta = 0$ when $\Psi(x_*, \varepsilon, \mu, \eta) = 0$, then x_* solves problem (P).*

5. Computational tests. In this section computational testing of the algorithm on several small location problems is summarized. In our tests the parameters ε_k , μ_k , and η_k are identical to those of § 2. The emphasis, except with the Fermat problem, has been to examine the performance of the algorithm as a function of its parameters ε_0 and β .

All testing was done on a PDP 11/34 minicomputer using double precision Fortran coding. In determining the element of minimum norm in $S(x, \varepsilon)$, Wolfe's algorithm [37] was used on minimax type problems and Gilbert's algorithm [9] was used on minisum type problems. In all problems we chose a tolerance of 10^{-8} for the minimum value of ε . Since the coding is double precision (about 16 digits), the choice of this tolerance implies that we seek single precision accuracy (7 to 8 digits). It is motivated by the fact that ill-conditioned systems of linear equations can be solved to single precision accuracy by elimination methods coded in double precision. In our reported results, an *iteration* is either a line search, or a reduction of ε_k to $\beta\varepsilon_k$ without a line search.

The minicomputer which we used could only measure elapsed time, not CPU time. Elapsed time varies greatly depending on how occupied the system is with various tasks. However, we did make comparisons on the Love minisum problem (below) with an AMDAHL V6 which is roughly equivalent to an IBM 370/165. On that problem, the minicomputer required an average elapsed time of 1.88 seconds per iteration and the AMDAHL required .046 CPU seconds per iteration. The ratio, approximately 40, is in agreement with the relative cycle times of the two machines.

One advantage of feasible direction methods is that they generally do not require very accurate line searches. In all of the results below we used a maximum of three quadratic fits per iteration. A testing of more accurate searches showed the additional time not worth the effort.

A Fermat problem. As previously mentioned, the unconstrained Fermat problem and Weiszfeld's algorithm have received much attention. A colleague, R. L. Francis, has suggested the following version for which the Weiszfeld algorithm converges very slowly:

$$\min_{x \in \mathbb{R}^2} \sum_{i=1}^5 w_i \|a_i - x\|,$$

where $a_1 = (0, 0)$, $a_2 = (1, 0)$, $a_3 = (1, 1)$, $a_4 = (0, 1)$, $a_5 = (100, 100)$ and $w_1 = w_2 = w_3 = w_4 = 1$ and $w_5 = 4$. The solution known to be in the convex hull of the a_i is $x^* = a_5 = (100, 100)$. Weiszfeld's algorithm, starting from $(.5, .5)$, requires over 12,000 iterations to achieve $x = (99.99-, 99.99-)$. (See Katz [13] for a study of the Weiszfeld algorithm convergence rate.) The difficulty is that the algorithm proceeds cautiously along the 45° line joining the starting point with x^* . Along this line the objective function decreases very slowly. With an exact line search, our algorithm finds the solution in at most two iterations starting from any point in the convex hull of the a_i . (No special claim is being made here—any descent algorithm with a line search should do the same.) Even with a crude line search which allows a maximum of three quadratic fits per iteration, the algorithm requires just 71 iterations to obtain $(99.99-, 99.99-)$.

The Eyster et al. [8] minisum problem. This unconstrained problem is

$$\min_{\substack{x_1 \in \mathbb{R}^2 \\ x_2 \in \mathbb{R}^2}} \sum_{r=1}^2 \sum_{s=1}^5 w_{rs} \|x_r - a_s\| + 2\|x_1 - x_2\|,$$

where $a_1 = (0, 0)$, $a_2 = (2, 4)$, $a_3 = (6, 2)$, $a_4 = (6, 10)$, $a_5 = (8, 8)$; and $w_{11} = 4$, $w_{12} = 2$, $w_{13} = 3$, $w_{14} = w_{15} = w_{21} = 0$, $w_{22} = 2$, $w_{23} = 1$, $w_{24} = 3$, $w_{25} = 2$. This problem was solved in [2] with ε fixed at 10^{-5} and required 23 line searches to obtain the objective function value 67.23856 at $x_1 = (2.8400, 2.6866)$ and $x_2 = (5.129, 6.388)$. For comparison, this problem was resolved with the algorithm of § 2 over a range of 40 values for β and the initial ε . Specifically the initial ε was any integral power of 10 between 10^{-1} and 10^{-8} , and β was any like value between 10^{-1} and 10^{-5} . For all values, the same solution was obtained in 24 to 31 iterations. Hence, there was virtually no sensitivity of the algorithm to its parameters on this problem.

A minimax location problem. A constrained problem similar to the Caribbean Islands problem of [2] is

$$\begin{aligned} \min_{\substack{x_1 \in \mathbb{R}^2 \\ x_2 \in \mathbb{R}^2}} F(x_1, x_2) &= \max_{i=1, \dots, 9} [w_{i1} \|a_i - x_1\|_{p_{i1}}, w_{i2} \|a_i - x_2\|_{p_{i2}}, \|x_1 - x_2\|], \\ &\text{subject to} \end{aligned}$$

$$\begin{aligned} \|x_1 - a_1\|^2 &\leq 144, & \|x_1 - a_4\|^2 &\leq 121, \\ \|x_2 - a_1\|^2 &\leq 225, & \|x_2 - a_2\|^2 &\leq 144, \end{aligned}$$

where the constants are given in Table 1.

TABLE 1

<i>i</i>	<i>a_i</i>	<i>w_{i1}</i>	<i>p_{i1}</i>	<i>w_{i2}</i>	<i>p_{i2}</i>
1	(11.4, 11.6)	2.0	2.0	1.0	2.0
2	(35.3, 13.5)	1.0	2.0	2.0	2.0
3	(8.8, 37.2)	1.3*	1.1	.85*	1.4
4	(20.9, 30.6)	1.1*	1.5	1.0	1.9
5	(25.5, 28)	1.5	1.4	1.5	1.2
6	(29.7, 27.7)	1.0	2.0	1.5	2.0
7	(36.2, 27.8)	0.5	1.8	1.0	1.7
8	(45.5, 21.3)	0.5	2.0	0.5	2.0
9	(15.8, 28.2)	0.5	1.1	0.5	1.8

* These values differ from those in [2].

The solution is: objective value = 23.886767, x_1^* is not unique but near (13.794657, 23.004391), and $x_2^* = (24.578965, 18.763441)$.

The third constraint is binding at optimality, and x_1^* can be any point in a small convex region. The three weights indicated with * above were altered from the data in [2] to force more functions to be near-binding near the solution. Table 2 summarizes the results of applying the algorithm for values of ϵ_0 and β between 10^{-1} and 10^{-8} . For $\epsilon_0 \leq 10^{-6}$ and $\beta \leq 10^{-4}$ the performance is almost uniform although the number of iterations tends to increase as ϵ_0 decreases. If ϵ_0 is too small or is decreased too fast because β is too small, the solution accuracy suffers. In the top region ($\epsilon_0 < 10^{-6}$) only one digit of F was correct while for $\beta < 10^{-4}$ the solution obtained always had at least 4 digits of F correct.

TABLE 2
Minimax problem: Number of iterations as a function of ϵ_0, β .

ϵ_0	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}
10^{-8}	5							
10^{-7}	8 (Note 1)	8						
10^{-6}	32	30	31					
10^{-5}	30	29	27	27				
10^{-4}	31	28	29	26	26			
10^{-3}	33	30	29	28	25	27		
10^{-2}	26	21	19	20	21	20	16	
10^{-1}	21 (Note 2)	18	18	22	19 (Note 3)	19	13	13

Starting point $(x_1^0, x_2^0) = (15, 22, 26, 11)$.

Notes: 1. 9% Relative error in F . 2. Single precision accuracy in F and x (8 digits). 3. Relative error in F less than 0.01%.

A *minisum location problem*. Love [22] has posed the problem

$$\min_{\substack{x_1 \in \mathbb{R}^2 \\ x_2 \in \mathbb{R}^2 \\ x_3 \in \mathbb{R}^2}} F(x_1, x_2, x_3) = \sum_{r=1}^3 \sum_{s=1}^5 w_{rs} \|x_r - a_s\|_p + \sum_{1 \leq r < t \leq 3} \|x_r - x_t\|_p$$

subject to

$$x_{31} + x_{32} \leq 3,$$

where $p = 1.78$, and $a_1 = (2, 3)$, $a_2 = (4, 2)$, $a_3 = (5, 4)$, $a_4 = (3, 5)$, $a_5 = (6, 7)$, and $w_{rs} = 1$ (except $w_{13} = w_{15} = 6$ and $w_{21} = 2$). The solution to this problem is: Objective value = 70.271346, $x_1^* = (5, 4) = a_3$, $x_2^* = (2, 3) = a_1$, and $x_3^* = (1.35853, 1.64147)$. Note that F is not differentiable at the solution point with respect to both x_1 and x_2 , and that the constraint on x_3 is binding.

Many runs were made on this problem, and the summary in Table 3 typifies the results using any interior starting point. Note that the sensitivity of the algorithm to ε_0 is significant while the sensitivity to β is slight. For ε_0 between 10^{-3} and 10^{-5} the value of F was correct to four digits and the components of (x_1, x_2, x_3) were correct to from two to five digits.

TABLE 3
Love problem: Number of iterations as a function of ε_0, β .
Maximum percent relative error as a function of ε_0 .

	Number of iterations					Max % relative error	
						F	$\ x\ $
10^{-8}	59					0.0056	8.16
10^{-7}	135	134				0.0029	5.18
10^{-6}	329	297	163			0.0400	1.11
10^{-5}	282	280	280	280		0.0029	0.54
10^{-4}	131	129	128	128	127	0.0026	0.54
10^{-3}	74	79	80	80	80	0.0045	0.55
10^{-2}	45	42	41	40	40	0.0289	0.59
10^{-1}	40	27	26	25	25	0.3123	1.78
	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}		

β

Starting point $(x_1^0, x_2^0, x_3^0) = (0, 0, 0, 0, 0)$.

Using starting points where the constraint holds at equality, the algorithm required about the same number of iterations as shown, but the accuracy of the solution decreased. For the starting point $(0, 0, 0, 0, 3, 0)$, the relative error in F was approximately 0.02% for ε_0 between 10^{-3} and 10^{-5} . It should be noted that the constraint had more influence on the algorithm performance than the nondifferentiability of F . With the constraint deleted, four different runs (with $\varepsilon_0 = 10^{-4}$, 10^{-5} , $\beta = 10^{-3}$ and two starting points) produced, in just 33 or 34 iterations, the same objective value to five digits (56.795) and $(x_1, x_2, x_3) = (4.99997, 4.00000, 3.351, 3.60, 4.02, 3.8963)$. As before, F is not differentiable at $x_1 = (5, 4)$.

6. Summary. One must be cautious to draw conclusions from any test results, but here it seems reasonable to suggest that users of the algorithm could expect maximum relative error of 0.01% in the objective value and 1.0% in $\|x\|$ when $10^{-2} \leq \varepsilon_0 \leq 10^{-4}$ and $10^{-2} \leq \beta \leq 10^{-5}$. If the data of the problem do not warrant higher accuracy than, say 1.0%, in F , then choosing $\varepsilon_0 = 10^{-2}$ should save some computational effort.

REFERENCES

- [1] J. A. CHATELON, D. W. HEARN AND T. J. LOWE, *A feasible direction subgradient algorithm for a class of nondifferentiable optimization problems*, Proc. 18th IEEE Conf. on Decision and Control, IEEE Control Sys. Society, Fort Lauderdale, FL, vol. 1, 1979, pp. 439–444.
- [2] ———, *A subgradient algorithm for certain minimax and minisum problems*, Math. Prog., 15 (1978), pp. 130–145.
- [3] F. H. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.
- [4] L. COOPER, *An extension of the generalized Weber problem*, J. Regional Sci., 8 (1968), pp. 181–197.
- [5] J. CULLUM, W. E. DONATH AND P. WOLFE, *The minimization of certain nondifferentiable sums of eigenvalues of symmetric matrices*, Math. Prog. Stud., 3 (1975), pp. 35–55.
- [6] V. F. DEM'YANOV AND V. N. MALOZEMOV, *Introduction to Minimax*, John Wiley, New York, 1974.
- [7] A. M. EL-SHAIEB, *A new algorithm for locating sources among destinations*, Manag. Sci., 20 (1973), pp. 221–231.
- [8] J. W. EYSTER, J. A. WHITE AND W. W. WIERWILLE, *On solving multifacility location problems using a hyperboloid approximation procedure*, AIIE Trans., 5 (1973), pp. 1–6.
- [9] E. G. GILBERT, *An iterative procedure for computing the minimum of a quadratic form on a convex set*, this Journal, 4 (1966), pp. 61–80.
- [10] A. A. GOLDSTEIN, *Optimization of Lipschitz continuous functions*, Math. Prog., 13 (1977), pp. 14–22.
- [11] M. HELD AND R. M. KARP, *The traveling salesman problem and minimum spanning trees, Part II*, Math. Prog., 1 (1971), pp. 6–25.
- [12] M. HELD, P. WOLFE AND H. P. CROWDER, *Validation of subgradient optimization*, Math. Prog. 6 (1974), pp. 62–88.
- [13] I. N. KATZ, *Local convergence in Fermat's problem*, Math. Prog. 6 (1974), pp. 89–104.
- [14] J. KRARUP AND P. PRUZAN, *Selected families of location problems*, in: Discrete Optimization II, Annals of Discrete Mathematics 5, P. L. Hammer, E. L. Johnson, B. H. Korte, eds., North-Holland, Amsterdam, 1979.
- [15] R. E. KUENNE AND R. M. SOLAND, *Exact and approximate solutions to the multi-source Weber problem*, Math. Prog., 3 (1972), pp. 193–209.
- [16] H. W. KUHN, *On a pair of dual nonlinear programs*, in: Methods of Nonlinear programming, J. Abadie, ed., North Holland, Amsterdam, 1967, pp. 37–54.
- [17] ———, *A note on Fermat's problem*, Math. Prog., 4 (1973), pp. 98–107.
- [18] H. W. KUHN AND R. E. KUENNE, *An efficient algorithm for the numerical solution of the generalized Weber problem in spatial economics*, J. Regional Sci., 4 (1962), pp. 21–34.
- [19] C. LEMARECHAL, *An algorithm for minimizing convex functions*, in: Information Processing, J. L. Rosen, ed., North Holland, Amsterdam, 1974, pp. 552–556.
- [20] ———, *An extension of Davidon methods to non-differentiable functions*, Math. Prog. Study, 3 (1975), pp. 95–109.
- [21] R. F. LOVE, *Locating facilities in three-dimensional space by convex programming*, Naval Res. Logist. Quart., 16 (1969), pp. 503–516.
- [22] ———, *The dual of a hyperbolic approximation to the generalized constrained multifacility location problem with l_p distances*, Manag. Sci., 21 (1974), pp. 22–33.
- [23] O. L. MANGASARIAN, *Nonlinear programming*, McGraw-Hill, New York, 1969.
- [24] R. MIFFLIN, *An algorithm for constrained optimization with semismooth functions*, Math. Oper. Res., 2 (1977), pp. 191–207.
- [25] ———, *Semismooth and semiconvex functions in constrained optimization*, this Journal, 15 (1977), pp. 959–972.
- [26] J. G., MORRIS, *A linear programming solution to the generalized rectangular distance Weber problem*, Naval Res. Logist. Quart., 22 (1975), pp. 155–164.

- [27] L. M. OSTRESH, *Convergence and descent in the Fermat location problem*, Transport Sci., 12 (1978), pp. 153–164.
- [28] A. PLANCHART AND A. P. HURTER, *An efficient algorithm for the solution of the Weber problem with mixed norms*, this Journal, 13 (1975), pp. 650–665.
- [29] B. T. POLYAK, *A general method of solving extremum problem*, Soviet Math. Doklady, 8 (1967), pp. 593–597.
- [30] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, New Jersey, 1970.
- [31] M. K. SCHAEFER AND A. P. HURTER, *An algorithm for the solution of a location problem with metric constraints*, Naval Res. Logist. Quart., 21 (1974), pp. 625–636.
- [32] N. Z. SHOR, *On the structure of algorithms for the numerical solution of optimal planning and design problems*, Dissertation, Cybernetics Inst. AN, Kiev, 1964.
- [33] E. WEISZFELD, *Sur le point pour lequel la somme des distances de n points donnees est minimum*, Tohoku Math. J., 43 (1937), pp. 355–386.
- [34] G. O. WESOLOWSKY AND R. F. LOVE, *The optimal location of new facilities using rectangular distances*, Oper. Res., 19 (1971), pp. 124–130.
- [35] ———, *A nonlinear approximation method for solving a generalized rectangular distance Weber problem*, Manag. Sci., 18 (1972), pp. 656–663.
- [36] P. WOLFE, *A method of conjugate subgradients for minimizing non-differentiable functions*, Math. Prog. Study, 3 (1975), pp. 145–173.
- [37] ———, *Finding the nearest point in a polytope*, Math. Prog., 11 (1976), pp. 128–144.

IDENTIFICATION AND ADAPTIVE CONTROL OF MARKOV CHAINS*

VIVEK BORKAR† AND PRAVIN VARAIYA‡

Abstract. Consider a countable state controlled Markov chain whose transition probability is specified up to an unknown parameter α taking values in a compact metric space A . To each α is associated a prespecified stationary control law $\zeta(\alpha)$. The adaptive control law selects at each time t the control action $\zeta(\alpha_t, x_t)$ where x_t is the state and α_t is the maximum likelihood estimate of α . The asymptotic behavior of this control scheme is investigated for the cases when the true parameter value α_0 does or does not belong to A , and for the case when ζ is chosen to minimize an average cost criterion. The analysis uses an appropriate extension of the notions of recurrence to nonstationary Markov chains.

1. Introduction. We consider a controlled Markov chain $\{x_n, n \geq 0\}$, characterized by

- (i) a countable state space $S = \{1, 2, \dots\}$;
- (ii) a control variable $z(i)$, taking values in a compact separable metric space $Z(i)$, for each i in S ;
- (iii) an unknown parameter α taking values in a compact separable metric space A ;
- (iv) a function $p(i, j; z, \alpha)$, $i, j \in S$, $z \in Z(i)$, $\alpha \in A$, which is the probability of transition from i to j when control z is used and if α is the true parameter.

The following assumptions are made throughout; additional assumptions will be made later as needed.

A1. For each i, j , $p(i, j; z, \alpha)$ is continuous in z, α .

A2. The actual transition probabilities correspond to the parameter value α_0 .

We do *not* assume a priori that α_0 is in A .

A3. If for some z' in $Z(i)$ $p(i, j; z', \alpha_0) = 0$, then $p(i, j; z, \alpha) = 0$ for $z \in Z(i)$, $\alpha \in A$. If $p(i, j; z', \alpha_0) \neq 0$, then there is $\bar{\epsilon} > 0$, independent of i, j, z, α such that $\bar{\epsilon} < p(i, j; z, \alpha) [p(i, j; z, \alpha_0)]^{-1} < (\bar{\epsilon})^{-1}$.

A4. For any fixed values of $\alpha \in A$ and $z(i) \in Z(i)$, for the Markov chain with stationary probabilities $p(i, j; z(i), \alpha)$, S is a communicating class which is positive recurrent.

A control law is any sequence of random variables $\{z_n, n \geq 0\}$ such that

- (i) $z_n \in Z(x_n)$,
- (ii) z_n is measurable with respect to $\mathcal{F}_n = \sigma(x_0, \dots, x_n)$;
- (iii) $P\{x_{n+1} = k | \mathcal{F}_n\} = p(x_n, k; z_n, \alpha_0)$.

The framework above is intended to cover two situations. In the first, $\{z_n\}$ is a deterministic sequence representing a known nonstationarity of the Markov process $\{x_n\}$. In the second, z_n is a random variable chosen on the basis of the known history, \mathcal{F}_n , of the state process in such a way as to satisfy some performance criterion. It is this second situation which will be addressed in § 4. Since this is the main motivation for our work we elaborate a little more. More specifically, our interest is in *adaptive*

* Received by the editors November 16, 1979 and in final revised form June 16, 1981. This research was sponsored in part by the National Science Foundation under grant ENG79-03879 and the Joint Services Electronics Program under contract F49620-79-C-0178.

† Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory, University of California, Berkeley, California 94720. Current address: Department of Applied Mathematics, Twente University of Technology, Postbox 217, 7500 AE Enschede, the Netherlands.

‡ Department of Electrical Engineering and Computer Sciences and the Electrical Research Laboratory, University of California, Berkeley, California 94720.

control laws which are constructed as follows. Suppose that for each α we are given a control function $\zeta(\alpha) = [\zeta(\alpha, 1), \zeta(\alpha, 2), \dots]$ such that if α is the true parameter then the sequence $z_n = \zeta(\alpha, x_n)$ results in a performance which is satisfactory (or optimal for some criterion). Next let α_n be the estimate of α_0 at time n obtained by some estimating scheme. The adaptive control law given by this estimating scheme and the function ζ is the random sequence $z_n = \zeta(\alpha_n, x_n)$, $n \geq 0$. Such an adaptive law seems to have been rigorously explored first in the context of linear systems by Åström and Wittenmark [1] where it was called a self-tuning regulator. A similar scheme for finite state Markov chains was studied by Mandl [2] under the assumption $\alpha_0 \in A$. In [2] the control function ζ was chosen to minimize the time average of the expected cost and a class of estimators for α_0 were considered on the basis of "contrast" functions. This class includes the maximum likelihood estimate (MLE). Mandl established the almost sure convergence of α_n to α_0 and of the time average of the cost to the minimum, by imposing an additional condition which in the case of MLE is the following: for any $\alpha \neq \beta$ in A , there is $i \in S$ such that

$$(1.1) \quad [p(i, 1; z, \alpha), p(i, 2; z, \alpha), \dots] \neq [p(i, 1; z, \beta), p(i, 2; z, \beta), \dots] \quad \text{for all } z \in Z(i).$$

However such an "identifiability" condition may, in applications, be too restrictive as seen from the work on linear systems ([1], [3], see also the discussion in [4]) where (1.1) does not hold and it was observed that the estimates need not converge and, even if they do, not necessarily to the true value α_0 .

The problem we consider is essentially the same as Mandl's in a more general setting: (i) S may be countable, (ii) (1.1) is relaxed and (iii) the control function ζ used to specify the adaptive law is any arbitrary map so that no explicit reference to a cost function is needed. (The case where S and A are both finite has been treated in [4].) Our problem is less general than Mandl's in one sense, namely, we restrict ourselves to MLE and do not consider other estimators.

The paper is organized as follows. Section 2 consists of results on recurrence of controlled Markov chains. These results will be used in §§ 3, 4, 5. Section 3 introduces the likelihood ratio and the MLE for Markov chains and studies their asymptotic properties. These results are applied to adaptive laws in § 4 under the assumption that $\alpha_0 \in A$. The behavior of the MLE when $\alpha_0 \notin A$ is examined in § 5. Some concluding remarks are collected in § 6.

2. Recurrence in controlled Markov chains. We begin with some definitions.

DEFINITION 2.1. Let $\{A_n, n \geq 0\}$ be a sequence of random events and $\{I(A_n)\}$ the corresponding indicator functions. The sequence $\{A_n\}$ is *rare* along a sample path ω if $\lim (1/n) \sum_{m=0}^n I(A_m)(\omega) = 0$, and *frequent* otherwise, i.e., if $\lim (1/n) \sum I(A_n)(\omega) > 0$. $\{A_n\}$ occurs *almost always* if $\{A_n^c\}$ is rare, where $A_n^c = \Omega - A_n$ is the complement of A_n . If $\{A_n\}$ is rare along all sample paths outside a set of zero probability it is *rare a.s.* *Frequent a.s.* and *almost always a.s.* are similarly defined.

DEFINITION 2.2. Let $\{y_n\}$ be a sequence of real numbers and y^* one of its limit points. Then y^* is a *frequent limit point* of $\{y_n\}$ if for every neighborhood O of y^* , the sequence of events $\{y_n \in O\}$ is frequent. A limit point which is not frequent is rare.

The definitions can be extended as follows to (triangular) arrays of events $\{A_{mn}, n \geq m \geq 0\}$. For instance, $\{A_{mn}\}$ is rare along a sample path if $\lim_n (1/n) \sum_m I(A_{mn}) = 0$, and frequent otherwise. The terms "almost always", etc. are defined similarly. To a sequence $\{A_m, m \geq 0\}$ one may associate the array $\{A_{mn} = A_m, n \geq m \geq 0\}$ and the two definitions coincide.

The following elementary lemma is useful for testing whether an array is rare.

LEMMA 2.1. Let $\{a_{mn}\}$ be a nonnegative sequence such that $(1/n) \sum_m a_{mn} \rightarrow 0$. For $\varepsilon > 0$, let k_n^ε be the number of terms in $\{a_{1n}, \dots, a_{nn}\}$ larger than ε , then $n^{-1} k_n^\varepsilon \rightarrow 0$. If $\{a_{mn}\}$ is bounded the converse also holds.

LEMMA 2.2. Let $\{a_n\}$ be a sequence in a compact metric space with metric d . Then:

- (i) The set A^* of limit points of $\{a_n\}$ is compact and $a_n \rightarrow A^*$, i.e., there is a sequence $\{a_n^*\}$ in A^* such that $d(a_n, a_n^*) \rightarrow 0$.
- (ii) $\{a_n\}$ has at least one frequent limit point.
- (iii) The set A^{**} of frequent limit points is compact.
- (iv) For any open set $O \supset A^{**}$, the sequence of events $\{a_n \in O\}$ occurs almost always.

Proof. (i) This is well known.

(ii) Let $\varepsilon > 0$ and cover A^* by finitely many balls of radius ε , say B_1, \dots, B_{m_1} . Then $\{a_n\}$ is eventually in $\bigcup B_i$, and for at least one i , say $i = k$, the sequence $\{a_n \in B_k\}$ is frequent. Cover \bar{B}_k , the closure of B_k , by finitely many balls of radius $\varepsilon/2$, say B_{k1}, \dots, B_{km_2} . Then for some j the sequence $\{a_n \in B_k \cap B_{kj}\}$ is frequent. Cover $\bar{B}_k \cap \bar{B}_{kj}$ by finitely many balls of radius $\varepsilon/4$. Continuing in this way we form a sequence of balls with radius $\varepsilon/2^n$, $n = 1, 2, \dots$ such that the sequence of events that a_n belongs to any of this spheres is frequent. By compactness the centers of these balls have a limit point a^* which it is easy to check must be a frequent limit point of $\{a_n\}$.

(iii) Let $\tilde{a} \in \overline{A^{**}}$, and \tilde{O} a neighborhood of \tilde{a} . Then there exist $a^* \in A^{**}$ and a neighborhood O^* of a^* such that $a^* \in O^* \subset \tilde{O}$. Since $\{a_n \in O^*\}$ is frequent so is the sequence $\{a_n \in \tilde{O}\}$, hence $a \in A^{**}$.

(iv) Let $B = O^c \cap A^*$. Then B is compact and $B \cap A^{**} = \emptyset$. Let O_1, O_2 be disjoint open sets such that $B \subset O_1$, $A^{**} \subset O_2$. For $\varepsilon > 0$ let D'_1, \dots, D'_m be balls of radius ε which cover B . Then $D_1 = D'_1 \cap O_1, \dots, D_m = D'_m \cap O_1$ is a finite cover of B that does not intersect A^{**} . Now if $\{a_n \in O^c\}$ is frequent then so is $\{a_n \in D_k\}$ for some k . Proceeding exactly as in (ii) we can find a frequent limit point in B . This contradicts $B \cap A^{**} = \emptyset$; hence $\{a_n \in O^c\}$ is rare. \square

LEMMA 2.3. Let $\{n_k, k \geq 0\}$ be a sequence of positive integers such that $\overline{\lim} (1/n) \sum_{m=0}^n I(m = n_k \text{ for some } k) > 0$. Let $\{a_n\}$ be a sequence in a compact metric space. Then the subsequence $\{a_{n_k}, k \geq 0\}$ has a limit point which is a frequent limit of $\{a_n\}$.

Proof. Let A^* be the limit points of $\{a_{n_k}\}$. For $\varepsilon > 0$, cover A^* by finitely many balls of radius ε , say B_1, \dots, B_m . Then $a_{n_k} \in \bigcup B_i$ eventually and by the assumption regarding $\{n_k\}$ it follows that the sequence $\{a_n \in \bigcup B_i\}$ is frequent. Proceeding as in the proof of Lemma 2.2 (ii) leads to the result. \square

We can now define recurrence concepts for controlled Markov chains.

DEFINITION 2.3. A state $i \in S$ is said to be *recurrent* along a sample path ω if $x_n(\omega) = i$ infinitely often (i.o.). If the sequence $\{x_n = i\}$ is frequent along ω it is said to be *positive recurrent* along ω ; otherwise it is *null recurrent* along ω .

LEMMA 2.4. Recall the assumptions A1–A4. If $i \in S$ is recurrent on a set D of positive probability, then every state is recurrent a.s. on D . Moreover, for any j, k in S such that $p(j, k; z, \alpha_0) > 0$ for some (hence all) $z \in Z(i)$, the events $\{x_n = j, x_{n+1} = k\}$ occur i.o. a.s. on D .

Proof. Suppose $p(i, i'; z, \alpha_0) > 0$. Since i is recurrent on D and $\min \{p(i, i'; z, \alpha_0) | z \in Z(i)\} > 0$, therefore the set $\{\omega \in D | \text{only finitely many transitions from } i \text{ to } i' \text{ are made along } \omega\}$ has probability zero, and so $\{\omega \in D | \text{infinitely many transitions from } i \text{ to } i' \text{ are made along } \omega\}$ has probability $P(D)$. Hence i' is recurrent. Now by A4 S is a communicating class. Hence by A3, for any $l \in S$ there is a finite path of strictly positive probability from i to l so that a repeated application of the

preceding argument establishes the a.s. recurrence of l on D . The second part of the lemma is proved in a similar manner. \square

LEMMA 2.5. *If $i \in S$ is positive recurrent on a set D of positive probability, then every state is positive recurrent a.s. on D . Moreover, for any j, k in S such that $p(j, k; z, \alpha_0) > 0$ for $z \in Z(i)$, the sequence $\{x_n = j, x_{n+1} = k\}$ is frequent a.s. on D .*

Proof. Suppose $p(i, i'; z, \alpha_0) > 0$. By the martingale stability theorem [6, p. 387],

$$\frac{1}{n} \sum_{m=1}^n [I(x_m = i') - E\{I(x_m = i') | \mathcal{F}_{m-1}\}] = \frac{1}{n} \sum_{m=1}^n [I(x_m = i') - p(x_{m-1}, i'; z_{m-1}, \alpha_0)] \rightarrow 0 \quad \text{a.s.,}$$

and so

$$\overline{\lim} \frac{1}{n} \sum_{m=1}^n I(x_m = i') \geq \left[\min_{Z(i)} p(i, i'; z, \alpha_0) \right] \overline{\lim} \frac{1}{n} \sum_{m=1}^n I(x_m = i) \quad \text{a.s.}$$

By hypothesis, the expression on the right is strictly positive on D . Hence i' is also positive recurrent a.s. on D . By arguing exactly as in the proof of Lemma 2.4 the remaining results may be established. \square

The use of the word "recurrent" in Definition 2.3 is clearly appropriate. Use of the phrases "positive recurrent" and "null recurrent" is justified by the next result.

LEMMA 2.6. *For Markov chains with stationary transition probabilities and a single communicating class the preceding definitions of positive and null recurrence coincide with the usual ones.*

Proof. For fixed $i \in S$ let $y_n = (1/n) \sum_{m=1}^n I(x_m = i)$ so $y_n^{-1} = n[\sum_{m=1}^n I(x_m = i)]^{-1}$. Let τ_0 be the first time i is reached and $\tau_k, k = 1, 2, \dots$, the k th return time for i . If m_n is the number of visits to i up to time n , then

$$\frac{1}{m_n} \sum_{k=0}^{m_n} \tau_k \leq \frac{1}{y_n} \leq \frac{1}{m_n} \sum_{k=0}^{m_n+1} \tau_k.$$

Now τ_0 is finite a.s. and the $\tau_k, k \geq 1$, are independent and identically distributed. Therefore, by the strong law of large numbers,

$$\lim_n \frac{1}{n} \sum_{k=0}^{m_n} \tau_k = \lim_n \frac{1}{m_n} \sum_{k=0}^{m_n} \tau_k = E\tau_1 \quad \text{a.s.,}$$

where the possibility $E\tau_1 = \infty$ is included. Hence $\lim_n y_n = (E\tau_1)^{-1}$ a.s. Therefore $\lim y_n = 0$ iff $E\tau_1 = \infty$ iff i is null recurrent in either sense, and $\lim y_n > 0$ iff $E\tau_1 < \infty$ iff i is positive recurrent in either sense. \square

We now introduce a condition which bears a resemblance to the notion of "tightness" of a family of distributions.

Condition T. There exists a null set N and for each $\varepsilon > 0$ there exists $J_\varepsilon < \infty$ such that $\overline{\lim} (1/n) \sum_{m=0}^n I(x_m = i, i > J_\varepsilon) < \varepsilon$ for every sample path $\omega \notin N$.

The next result is immediate.

LEMMA 2.7. (i) *For a finite state Markov chain Condition T is always satisfied.*

(ii) *For a Markov chain with stationary transition probabilities positive recurrence implies Condition T.*

LEMMA 2.8. *Under Condition T all states are positive recurrent a.s.*

Proof. Suppose $i \in S$ is null recurrent on a set D of positive probability. By Lemma 2.5 all states are null recurrent a.s. on D , i.e., $(1/n) \sum_{m=1}^n I(x_m = j) \rightarrow 0$ a.s. on D for all $j \in S$. Hence for any J

$$\frac{1}{n} \sum_{m=0}^{n-1} I(x_m = j, j \leq J) = \frac{1}{n} \sum_{j=1}^J \sum_{m=0}^{n-1} I(x_m = j) \rightarrow 0 \quad \text{a.s. on } D,$$

so that Condition T cannot hold. \square

To obtain a condition in terms of the transition probabilities $p(i, j; z, \alpha_0)$ which implies condition T we need the following. Let $\zeta = [z(1), z(2), \dots]$ with $z(i) \in Z(i)$ be a fixed control function and consider the control law $\{z_n\}$ with $z_n = z(x_n)$. Such a law $\{z_n\}$, or equivalently ζ , will be called a *stationary law*. Consider the following assumption.

A5. There is a finite number M such that for any stationary law ζ , there is a state s_ζ such that the expected time to hit s_ζ from any i in S is bounded by M .

This is the familiar condition which guarantees the existence of an optimal stationary law in Markov decision processes with the time average cost criterion. (See, e.g., [7, pp. 147–148].) In [8], Federgruen, Hordijk and Tijms have given many equivalent formulations of A5. In particular, a simple modification of their proof shows that under A1–A5, A5 is equivalent to A5'.

A5'. Let $\{\pi(i, \zeta), i \in S\}$ be the invariant probabilities under the stationary law ζ and let $p_{ij}^n(\zeta) = P\{x_n = j | x_0 = i, \zeta \text{ is used}\}$. Then $\lim_n (1/n) \sum_{m=1}^n P_{ij}^n(\zeta) = \pi(j, \zeta)$ uniformly in i, ζ for each j in S .

We can now obtain the following useful results.

LEMMA 2.9. Under A1–A5 the set of probability measures $\{\pi(i, \zeta), i \in S\}$ on S for $\zeta \in Z = \prod_{i \in S} Z(i)$ is tight; i.e., for $\varepsilon > 0$, there exists $J_\varepsilon < \infty$ such that $\sum_{i > J_\varepsilon} \pi(i, \zeta) < \varepsilon$ for all $\zeta \in Z$.

Proof. It is easily seen that for each n, i, j $p_{ij}^n(\zeta)$ is continuous on Z . Since, by A5', $(1/n) \sum_{m=1}^n p_{ij}^m \rightarrow \pi(j, \zeta)$ for each j uniformly in i, ζ , therefore $\pi(j, \zeta)$ is continuous in ζ for each j . Hence, for $J < \infty$, $\sum_{j \leq J} \pi(j, \zeta)$ and $\sum_{j > J} \pi(j, \zeta) = 1 - \sum_{j \leq J} \pi(j, \zeta)$ are both continuous in ζ . Now as $J \rightarrow \infty$, $\sum_{j > J} \pi(j, \zeta)$ decreases monotonically to zero. Since Z is compact under the product topology, it follows by Dini's theorem that $\sum_{j > J} \pi(j, \zeta) \rightarrow 0$ as $J \rightarrow \infty$, uniformly in ζ . Hence $\lim_J \sup_\zeta \sum_{j > J} \pi(j, \zeta) = 0$ and the result follows. \square

LEMMA 2.10. Let $c(i, j, z)$ be a continuous, nonnegative, bounded function defined for $(i, j, z) \in S \times S \times Z(i)$. Then, under A1–A5, for any control law $\{z_n\}$,

$$\overline{\lim}_n \frac{1}{n} \sum_{m=1}^{n-1} c(x_m, x_{m+1}, z_m) \leq \max_{\zeta \in Z} \sum_i \pi(i, \zeta) \sum_j p(i, j; \zeta(i), \alpha_0) c(i, j, \zeta(i)) \quad \text{a.s.}$$

Proof. This follows from well-known results treating $c(\cdot, \cdot, \cdot)$ as the reward function of a Markov decision process (see [7]). \square

LEMMA 2.11. Under A1–A5 Condition T holds.

Proof. By the martingale stability theorem, for any j ,

$$\lim (1/n) \sum_{m=1}^{n-1} [I(x_m = j, j > J) - E\{I(x_m = j, j > J) | \mathcal{F}_{m-1}\}] = 0 \quad \text{a.s.}$$

so

$$\begin{aligned} \overline{\lim}_n \frac{1}{n} \sum_{m=1}^{n-1} I(x_m = j, j > J) &= \overline{\lim}_n \frac{1}{n} \sum_{m=1}^{n-1} E\{I(x_m = j, j > J) | \mathcal{F}_{m-1}\} \\ (2.1) \quad &= \overline{\lim}_n \frac{1}{n} \sum_{m=0}^{n-1} \sum_{j > J} p(x_m, j; z_m, \alpha_0) \\ &\leq \max_{\zeta \in Z} \sum_{i \in S} \pi(i, \zeta) \sum_{j > J} p(i, j; \zeta(i), \alpha_0) \quad \text{a.s.} \end{aligned}$$

The inequality in (2.1) follows from Lemma 2.10 by choosing $c(i, j, z) = \sum_{j > J} p(i, j; z, \alpha_0)$. As in the proof of Lemma 2.9 one can show that

$$(2.2) \quad \lim_J \max_{\zeta} \sum_{j > J} p(i, j; \zeta(i), \alpha_0) = 0 \quad \text{for all } i \text{ in } S.$$

From Lemma 2.9 there exists J_1 such that

$$\sum_{i>J_1} \pi(i, \zeta) < \frac{\varepsilon}{2}, \quad \zeta \in Z.$$

From (2.2) there exists J_2 such that

$$\max_{i \leq J_1} \max_{\zeta} \sum_{j>J_2} p(i, j; \zeta(i), \alpha_0) \leq \frac{\varepsilon}{2J_1},$$

so

$$\sum_{i \in S} \pi(i, \zeta) \sum_{j>J_2} p(i, j; \zeta(i), \alpha_0) \leq \sum_{i \leq J_1} \pi(i, \zeta) \sum_{j>J_2} p(i, j; \zeta(i), \alpha_0) + \sum_{i>J_1} \pi(i, \zeta) < \varepsilon, \quad \zeta \in Z.$$

Using this estimate in (2.1) gives

$$\overline{\lim} \frac{1}{n} \sum I(x_m = j, j > J_2) < \varepsilon \quad \text{a.s.}$$

Let N_ε be the null set where this inequality fails, and let $N = \bigcup_{k=1}^{\infty} N_{\varepsilon/k}$. Then Condition T holds outside of N . \square

LEMMA 2.12. Suppose A1–A5 and let N be the null set on which Condition T fails. Let $\tilde{\omega} \notin N$. Suppose $\{n_k, k \geq 0\}$ is a subsequence of the integers such that

$$(2.3) \quad \overline{\lim} \frac{1}{n} \sum_{m=0}^{n-1} I(m = n_k \text{ for some } k) = \delta > 0.$$

Then there exists $i \in S$ such that the sequence of events $\{m = n_k \text{ for some } k \text{ and } x_m = i, m \geq 0\}$ is frequent along $\tilde{\omega}$.

Proof. Choose $0 < \delta < \delta$ and $\tilde{J} < \infty$ such that

$$(2.4) \quad \overline{\lim} \frac{1}{n} \sum I(x_m = j, j > \tilde{J}) < \tilde{\delta} \quad \text{for } \omega \notin N.$$

Suppose, contrary to the assertion, that

$$\frac{1}{n} \sum_{m=0}^{n-1} I(m = n_k \text{ for some } k, x_m \leq \tilde{J}) \rightarrow 0 \quad \text{along } \tilde{\omega}.$$

Then

$$\overline{\lim} \frac{1}{n} \sum_{m=0}^{n-1} I(m = n_k \text{ for some } k, x_m > \tilde{J}) = \delta > \tilde{\delta} \quad \text{along } \tilde{\omega},$$

and, a fortiori, to

$$\overline{\lim} \frac{1}{n} \sum I(x_m = j, j > \tilde{J}) \geq \delta \quad \text{along } \tilde{\omega},$$

thereby contradicting (2.4). \square

Observe that i in the assertion can be chosen such that $i < \tilde{J}$.

These results form the basis of the proofs in subsequent sections.

3. Likelihood ratios. We recall from [9] some facts about likelihood ratios. Let (Ω, \mathcal{F}, P) be a probability space, and \tilde{P} another probability on (Ω, \mathcal{F}) . Then there exists an integrable function $\Lambda \geq 0$ on (Ω, \mathcal{F}, P) called the *likelihood ratio* of \tilde{P} to P

and a set N with $P(N) = 0$ so that

$$\tilde{P}(A) = \int_A \Lambda dP + \tilde{P}(A \cap N), \quad A \in \mathcal{F}.$$

Λ is also denoted by $d\tilde{P}/dP$. Λ is unique up to P -equivalence and N is unique up to $(P + \tilde{P})$ -equivalence. Define the relations $\tilde{P} \ll P$, $\tilde{P} \perp P$ respectively, by $\tilde{P}(N) = 0$, $\tilde{P}(N) = 1$ (or, equivalently, $\Lambda = 0$ P -a.s.). Let $\tilde{P} \equiv P$ mean $\tilde{P} \ll P$ and $P \ll \tilde{P}$.

Let $\{\mathcal{F}_n, n \geq 0\}$ be an increasing family of σ -fields such that $\mathcal{F} = (\bigvee \mathcal{F}_n)$. Let P_n, \tilde{P}_n be the restrictions of P, \tilde{P} to \mathcal{F}_n and let $\Lambda_n = d\tilde{P}_n/dP_n$. Then $\tilde{P} \ll P$ implies $\tilde{P}_n \ll P_n$. Also $\Lambda_n = E\{\Lambda | \mathcal{F}_n\}$ P -a.s. Thus $(\Lambda_n, \mathcal{F}_n, P)$ is a martingale and $\Lambda_n \rightarrow \Lambda$ P -a.s.

In our problem Ω is the set of all sequences $\{x_n\}$ in S and $\mathcal{F}_n = \sigma(x_0, \dots, x_n)$. Each $\alpha \in A$ defines a probability measure P^α . It is easy to check that

$$\Lambda_n(\alpha) = \frac{dP_n^\alpha}{dP_n^{\alpha_0}} = \prod_{m=0}^{n-1} \frac{p(x_m, x_{m+1}; z_m, \alpha)}{p(x_m, x_{m+1}; z_m, \alpha_0)}.$$

From the above-quoted facts it follows that for each $\alpha \in A$ there is a function $\Lambda(\alpha) = dP^\alpha/dP^{\alpha_0}$ such that $\Lambda_n(\alpha) \rightarrow \Lambda(\alpha)$ a.s. (with respect to P^{α_0}).

LEMMA 3.1. $p(x_n, x_{n+1}; z_n, \alpha)/p(x_n, x_{n+1}; z_n, \alpha_0) \rightarrow 1$ a.s. on the set $\{\Lambda(\alpha) > 0\}$.

Proof. Note that $p(x_n, x_{n+1}; z_n, \alpha)/p(x_n, x_{n+1}; z_n, \alpha_0) = \Lambda_{n+1}(\alpha)/\Lambda_n(\alpha)$ and $\lim \Lambda_n(\alpha) = \lim \Lambda_{n+1}(\alpha) = \Lambda(\alpha)$.

COROLLARY 3.1. Suppose the Markov chain has stationary transition probabilities and is recurrent. Then either $\Lambda(\alpha) = 0$ a.s. and $P^\alpha \perp P^{\alpha_0}$ or $\Lambda(\alpha) = 1$ a.s. and $P^\alpha \equiv P^{\alpha_0}$ in which case the transition probabilities under α, α_0 are identical.

Define $L_n(\alpha) = \ln \Lambda_n(\alpha)$. L_n is the log-likelihood ratio.

DEFINITION 3.1. For each ω let $\alpha_n = \alpha_n(\omega) \in A$ be such that $L_n(\alpha_n) \geq L_n(\alpha)$, $\alpha \in A$. α_n is called the *maximum likelihood estimate* (MLE) at time n . If the maximum value of $L_n(\alpha)$ is achieved at more than one value, we assume that only one of these is chosen according to some prescribed rule which ensures that α_n is \mathcal{F}_n -measurable.

For the remainder of this section assume that A1–A5 hold and $\alpha_0 \in A$.

LEMMA 3.2. $(1/n)L_n(\alpha_n) \rightarrow 0$ a.s.

Proof. $L_n(\alpha_n) \geq L_n(\alpha_0) = 0$. Hence

$$(3.1) \quad \lim \frac{1}{n} L_n(\alpha_n) \geq 0 \quad \text{a.s.}$$

From Lemma 2.10, and for any fixed α in A , an appropriate choice of the function c gives

$$\begin{aligned} \overline{\lim}_n \frac{1}{n} L_n(\alpha) &\leq \max_{\zeta} \sum_i \pi(i, \zeta) \sum_k p(i, k; \zeta(i), \alpha_0) \ln \frac{p(i, k; \zeta(i), \alpha)}{p(i, k; \zeta(i), \alpha_0)} \quad \text{a.s.} \\ &\leq 0 \quad \text{a.s., by Jensen's inequality.} \end{aligned}$$

Let \tilde{A} be a countable dense subset of A . By the preceding inequality there is a null set N outside of which $\overline{\lim} (1/n)L_n(\alpha) \leq 0$ for all $\alpha \in \tilde{A}$. By Lemma 2.11 Condition T holds outside a null set which, we may assume, is included in N . Then for $\varepsilon > 0$ and $\omega \notin N$, there exists J such that

$$(3.2) \quad \overline{\lim}_n \frac{1}{n} \sum I(x_m > J) < \frac{\varepsilon}{8K},$$

where $K > 0$ is any number with $K > |\ln \bar{\varepsilon}|$ and $\bar{\varepsilon}$ is as in A3. Fix α in A . By continuity there exists $\tilde{\alpha} \in \tilde{A}$ such that

$$(3.3) \quad \left| \ln \frac{p(i, k; z, \alpha)}{p(i, k; z, \alpha_0)} - \ln \frac{p(i, k; z, \tilde{\alpha})}{p(i, k; z, \alpha_0)} \right| < \frac{\varepsilon}{2},$$

for all $i, k \leq J$, and $z \in Z(i)$. Hence, for sufficiently large n ,

$$\begin{aligned} \left| \frac{1}{n} L_n(\alpha) - \frac{1}{n} L_n(\tilde{\alpha}) \right| &\leq \frac{1}{n} \sum_{m=0}^{n-1} \left| \ln \frac{p(x_m, x_{m+1}; z_m, \alpha)}{p(x_m, x_{m+1}; z_m, \alpha_0)} - \ln \frac{p(x_m, x_{m+1}; z_m, \tilde{\alpha})}{p(x_m, x_{m+1}; z_m, \alpha_0)} \right| \\ &\leq \frac{1}{n} \sum \left| \ln \frac{p(x_m, x_{m+1}; z_m, \alpha)}{p(x_m, x_{m+1}; z_m, \alpha_0)} - \ln \frac{p(x_m, x_{m+1}; z_m, \tilde{\alpha})}{p(x_m, x_{m+1}; z_m, \alpha_0)} \right| \\ &\quad \cdot [I(x_m > J \text{ or } x_{m+1} > J) + I(x_m \leq J \text{ and } x_{m+1} \leq J)] \leq \varepsilon, \end{aligned}$$

by (3.2), (3.3). Since $\varepsilon > 0$ is arbitrary and $\overline{\lim} (1/n) L_n(\tilde{\alpha}) \leq 0$, $\omega \notin N$ it follows that

$$(3.4) \quad \overline{\lim} \frac{1}{n} L_n(\alpha) \leq 0 \quad \text{for all } \alpha \in A \text{ and } \omega \notin N.$$

Now suppose there is $\omega \notin N$ and $\varepsilon > 0$ and a subsequence $\{\alpha_{n_k}\}$ of $\{\alpha_n(\omega)\}$ such that

$$\frac{1}{n_k} L_{n_k}(\alpha_{n_k}) \geq \varepsilon \quad \text{for all } k.$$

Since A is compact we may suppose $\alpha_{n_k} \rightarrow \alpha^*$. Then, an argument similar to the above may be employed to show that

$$\overline{\lim}_k \frac{1}{n_k} L_{n_k}(\alpha^*) \geq \varepsilon,$$

thereby contradicting (3.4). Hence it must be that $\overline{\lim} (1/n) L_n(\alpha_n) \leq 0$ for $\omega \notin N$ which, with (3.1), proves the assertion. \square

THEOREM 3.1. *There exists a null set N such that, for every $\omega \notin N$, $\varepsilon > 0$, $i \in S$ and $J < \infty$, the triangular array of events $\{A_{mn}, n \geq m \geq 0\}$ is rare, where*

$$A_{mn} = \left\{ I(x_m = i) \left| \ln \frac{p(i, k; z_m, \alpha_n)}{p(i, k; z_m, \alpha_0)} \right| > \varepsilon \text{ for some } k \leq J \right\}.$$

Proof. For each $\alpha \in A$ let

$$G_n(\alpha) = \sum_{k \in S} p(x_n, k; z_n, \alpha_0) \ln \frac{p(x_n, k; z_n, \alpha)}{p(x_n, k; z_n, \alpha_0)}.$$

(As usual, we set the last term to 0 if $p(x_n, k; z_n, \alpha_0) = 0$.) It is easy to show that $G_n(\alpha)$ is continuous in α . Let

$$B_n(\alpha) = \sum_{m=0}^{n-1} G_m(\alpha).$$

By the martingale stability theorem $\lim (1/n)[L_n(\alpha) - B_n(\alpha)] = 0$ a.s. Let \tilde{A} be a countable dense subset of A and N_1 a null set outside which $\lim (1/n)[L_n(\tilde{\alpha}) - B_n(\tilde{\alpha})] = 0$ for every $\tilde{\alpha} \in \tilde{A}$. Let N_2 be the null set outside which Condition T holds. Let $N = N_1 \cup N_2$.

Let $\omega \notin N$ and consider the sequence $(1/n)[L_n(\alpha_n) - B_n(\alpha_n)]$ along ω . This sequence is bounded. Let h^* be any of its limit points. Let $\{n_k, k \geq 0\}$ be a subsequence such that $(1/n_k)[L_{n_k}(\alpha_{n_k}) - B_{n_k}(\alpha_{n_k})] \rightarrow h^*$. We may also suppose that $\alpha_{n_k} \rightarrow \alpha^*$ for some α^* in A .

We will show that $h^* = 0$.

Let $\delta > 0$. Let J be such that $\overline{\lim} (1/n) \sum^{n-1} I(x_m > J) > \delta$. Define

$$h(i, j; z, \alpha) = \ln \frac{p(i, j; z, \alpha)}{p(i, j; z, \alpha_0)} - \sum_k p(i, k; z, \alpha_0) \ln \frac{p(i, k; z, \alpha)}{p(i, k; z, \alpha_0)}.$$

Then $h(i, j; z, \alpha)$ is continuous in α . Let $\tilde{\alpha} \in \tilde{A}$ be such that $|h(i, j; z, \alpha_{n_k}) - h(i, j; z, \tilde{\alpha})| < \delta$ for all $i, j \leq J, z \in Z(i)$ and k sufficiently large. Note that

$$L_n(\alpha) - B_n(\alpha) = \sum^{n-1} h(x_m, x_{m+1}; z_m, \alpha),$$

so

$$\begin{aligned} & \overline{\lim}_{n_k} \frac{1}{n_k} |L_{n_k}(\alpha_{n_k}) - B_{n_k}(\alpha_{n_k}) - [L_{n_k}(\tilde{\alpha}) - B_{n_k}(\tilde{\alpha})]| \\ &= \overline{\lim} \left| \frac{1}{n_k} \sum [h(x_m, x_{m+1}; z_m, \alpha_{n_k}) - h(x_m, x_{m+1}; z_m, \tilde{\alpha})] \right| \\ &\leq \overline{\lim}_{n_k} \frac{1}{n_k} \sum |[h(x_m, x_{m+1}; z_m, \alpha_{n_k}) - h(x_m, x_{m+1}; z_m, \tilde{\alpha})] I(\max(x_m, x_{m+1}) > J)| \\ &\quad + \overline{\lim}_{n_k} \frac{1}{n_k} \sum |[h(x_m, x_{m+1}; z_m, \alpha_{n_k}) - h(x_m, x_{m+1}; z_m, \tilde{\alpha})] I(\max(x_m, x_{m+1}) \leq J)| \\ &\leq 4K \overline{\lim}_{n_k} \frac{1}{n_k} \sum I(\max(x_m, x_{m+1}) > J) + \delta \leq (4K + 1)\delta, \end{aligned}$$

where $K = |\ln \bar{\varepsilon}|$, $\bar{\varepsilon}$ as in assumption A3. We then have

$$|h^*| = |h^* - \lim_{n_k} \frac{1}{n_k} [L_{n_k}(\tilde{\alpha}) - B_{n_k}(\tilde{\alpha})]| \leq (4K + 1)\delta,$$

but $\delta > 0$ is arbitrary, so $h^* = 0$. Since h^* was any limit point, this implies

$$\lim_{n} \frac{1}{n} [L_n(\alpha_n) - B_n(\alpha_n)] = 0,$$

and so, by Lemma 3.2,

$$(3.5) \quad \frac{1}{n} B_n(\alpha_n) = \frac{1}{n} \sum^{n-1} G_m(\alpha_n) \rightarrow 0.$$

For $n \geq m \geq 0$ rewrite $G_m(\alpha_n)$ as

$$-G_m(\alpha_n) = \sum_i I(x_m = i) \sum_k p(i, k; z_m, \alpha_n) \frac{p(i, k; z_m, \alpha_0)}{p(i, k; z_m, \alpha_n)} \ln \frac{p(i, k; z_m, \alpha_0)}{p(i, k; z_m, \alpha_n)}.$$

The convexity of the function $x \ln x$ and Jensen's inequality imply that each of the summands is nonnegative. Hence, for each i in S ,

$$\frac{1}{n} \sum^{n-1} I(x_m = i) \sum_k p(i, k; z_m, \alpha_n) \frac{p(i, k; z_m, \alpha_0)}{p(i, k; z_m, \alpha_n)} \ln \frac{p(i, k; z_m, \alpha_0)}{p(i, k; z_m, \alpha_n)} \rightarrow 0.$$

By Lemma 2.1, for $\varepsilon > 0$, the array of events

$$\left\{ I(x_m = i) \sum_k p(i, k; z_m, \alpha_n) \frac{p(i, k; z_m, \alpha_0)}{p(i, k; z_m, \alpha_n)} \ln \frac{p(i, k; z_m, \alpha_0)}{p(i, k; z_m, \alpha_n)} > \varepsilon \right\}$$

is rare. Now by assumption A3, for fixed i and J , the set of positive numbers $\{p(i, k, z, \alpha) | k \leq J, z \in Z(i), \alpha \in A \text{ and } p(i, k, z, \alpha) \neq 0\}$ is bounded away from zero. Invoking again the strict convexity of $x \ln x$ and Jensen's inequality gives the result.

COROLLARY 3.2. *There exists a null set N such that for every $\omega \notin N$ and limit point α^* of $\{\alpha_n(\omega)\}$, $\varepsilon > 0$, $i \in S$ and $J < \infty$, $\lim_{n \rightarrow \infty} (1/n) \sum_{m=0}^{n-1} I(A_m) = 0$, where*

$$A_n = \left\{ I(x_n = i) \left| \ln \frac{p(i, k; z_n, \alpha^*)}{p(i, k; z_n, \alpha_0)} \right| > \varepsilon \text{ for some } k \leq J \right\}.$$

Proof. This follows in a similar manner from the fact if $\alpha_{n_k} \rightarrow \alpha^*$, then $\lim (1/n_k) B_{n_k}(\alpha^*) = \lim (1/n_k) B_{n_k}(\alpha_{n_k}) = 0$. \square

The next result is an immediate corollary of the preceding result.

THEOREM 3.2. *Suppose the Markov chain $\{x_n\}$ has stationary transition probabilities and is positive recurrent. Then there is a null set N such that for every $\omega \notin N$ and limit α^* of $\{\alpha_n(\omega)\}$, the transition probabilities under α^* and α_0 coincide.*

Proof. Let N_0 be the null set in Corollary 3.2. For each $i \in S$, put $c(x_m, x_{m+1}; z_m) = 1 - I(x_m = i)$ in Lemma 2.10 to get

$$(3.6) \quad \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{m=0}^{n-1} I(x_m = i) \geq \pi(i, \zeta) > 0 \quad \text{a.s.}$$

for some $\zeta \in Z$. (In the present context Z contains only one element.) Let N_i be the null set where the inequality in (3.6) fails. Let $\omega \notin N = N_0 \cup (\cup_i N_i)$ and α^* a limit point of $\{\alpha_n(\omega)\}$. Suppose there are i, j such that $p(i, j, \alpha^*) \neq p(i, j, \alpha_0)$ and so

$$\left| \ln \frac{p(i, j, \alpha^*)}{p(i, j, \alpha_0)} \right| > \varepsilon > 0.$$

But then the sequence of events

$$\left\{ I(x_m = i) \ln \frac{p(i, j, \alpha^*)}{p(i, j, \alpha_0)} > \varepsilon \right\} = \{I(x_m = i)\}$$

contradicts Corollary 3.2. \square

The application of Theorem 3.1 and Corollary 3.2 to adaptive control appears in the next section.

4. Adaptive control. Throughout this section it is assumed that A1–A5 hold and $\alpha_0 \in A$. We also assume given for each α in A a stationary law $\xi(\alpha) = [\xi(\alpha, 1), \xi(\alpha, 2), \dots]$. The actual law is given by the adaptive law $z_n = \xi(\alpha_n, x_n)$, $n \geq 0$. The particular choice of $\xi(\alpha)$ is not relevant in most of the subsequent discussion, but we shall consider the interesting case when $\xi(\alpha)$ is chosen to minimize, assuming α is the true parameter, the average cost of the form $\lim_{n \rightarrow \infty} (1/n) \sum_{m=0}^{n-1} c(x_m, x_{m+1}, z_m)$.

In [4] we considered the case where A, S are both finite and the following result was obtained which is stronger than what seems possible in the more general setting considered here.

THEOREM 4.1 [4]. *Under the scheme above, there exist a random variable α^* and a random time $N < \infty$ a.s. such that for almost all ω , $\alpha_n(\omega) = \alpha^*(\omega)$, $n \geq N(\omega)$ and $p(i, j; \xi(\alpha^*(\omega), i), \alpha^*(\omega)) = p(i, j; \xi(\alpha^*(\omega), i), \alpha_0)$ for i, j in S . Moreover, if $\xi(\alpha)$ minimizes the average cost under α , then the true cost $\lim_{n \rightarrow \infty} (1/n) \sum_{m=0}^{n-1} c(x_m, x_{m+1}, z_m) = J(\alpha^*)$ a.s. where $J(\alpha)$ is the cost corresponding to the stationary law $\xi(\alpha)$.*

Thus in the finite case the adaptive law is "stable" in the sense that the parameter estimate α_n , and the average cost $(1/n) \sum_{m=0}^{n-1} c(x_m, x_{m+1}, z_m)$ converge. However, the

limiting cost $J(\alpha^*)$ may exceed $J(\alpha_0)$ which is the minimum possible cost. (For an example see [4].)

To see what is possible in the more general setting we need the following definition.

DEFINITION 4.1. For a sample path ω , a limit point α^* of $\{\alpha_n(\omega)\}$ and a state $i \in S$, the pair (i, α^*) is said to be *frequent* along ω if for each neighborhood O of α^* , the sequence of events $\{x_n = i, \alpha_n \in O\}$ is frequent along ω . A limit point α^* of $\{\alpha_n(\omega)\}$ is *regular* along ω if there is an i in S such that (i, α^*) is frequent along ω .

LEMMA 4.1. *There is a null set N such that if $\omega \notin N$, the set of regular limit points is dense in the set of frequent limit points of $\{\alpha_n(\omega)\}$.*

Proof. Let N be the null set in Lemma 2.12, $\omega \notin N$, and α^* a frequent limit point of $\{\alpha_n(\omega)\}$ which is not regular. Let O be any neighborhood of α^* and $\{n_k\}$ the maximal subsequence such that $\alpha_{n_k}(\omega) \in O$. By Lemma 2.12 there is an i in S such that the sequence of events $\{\alpha_{n_k} \in O, x_{n_k} = i\}$ is frequent. By Lemma 2.3 there is a subsequence $\{\tilde{n}_k\}$ of $\{n_k\}$ and $\tilde{\alpha}$ in the closure of O such that $\{\alpha_{\tilde{n}_k} \in O, x_{\tilde{n}_k} = i\}$ is frequent and $\lim \alpha_{\tilde{n}_k} = \tilde{\alpha}$. Evidently $\tilde{\alpha}$ is regular. Since O is arbitrary, the result follows. \square

THEOREM 4.2. *Suppose $\xi(\alpha, i)$ is continuous in α for each i . Then there exists a null set N such that for every $\omega \notin N$, limit point α^* of $\{\alpha_n(\omega)\}$ and $i \in S$ such that (i, α^*) is frequent along ω ,*

$$p(i, j; \xi(\alpha^*, i), \alpha^*) = p(i, j; \xi(\alpha^*, i), \alpha_0) \quad \text{for all } j \in S.$$

Proof. Let N be the null set in Theorem 3.1 and $\omega \notin N$. Suppose in contradiction that for some j $p(i, j; \xi(\alpha^*, i), \alpha^*) \neq p(i, j; \xi(\alpha^*, i), \alpha_0)$. Since ξ is continuous there is a neighborhood O of α^* and $\varepsilon > 0$ such that

$$\left| \ln \frac{p(i, j; \xi(\alpha, i), \tilde{\alpha})}{p(i, j; \xi(\alpha, i), \alpha_0)} \right| > \varepsilon,$$

whenever $\alpha, \tilde{\alpha}$ are in O . So

$$\begin{aligned} & \overline{\lim} \frac{1}{n} \sum^{n-1} I\left(x_m = i \text{ and } \left| \ln \frac{p(i, j; \xi(\alpha_m, i), \alpha_n)}{p(i, j; \xi(\alpha_m, i), \alpha_0)} \right| > \varepsilon\right) \\ & \geq \overline{\lim} \frac{1}{n} \sum^{n-1} I(x_m = i, \alpha_m \in O, \alpha_n \in O) > 0, \end{aligned}$$

since the sequence $\{x_m = i, \alpha_m \in O\}$ is frequent. But this contradicts Theorem 3.1. \square

This result is clearly weak in comparison with Theorem 4.1. To obtain stronger conclusions it is necessary to modify the adaptive control law through randomization. We study two such schemes.

Randomization of control values. We impose another assumption.

A6.1. For any $\alpha \neq \beta$ in A , there exists $i \in S$ such that for every open set $O \subset Z(i)$ there exists $z \in O$ for which

$$(4.1) \quad [p(i, 1; z, \alpha), p(i, 2; z, \alpha), \dots] \neq [p(i, 1; z, \beta), p(i, 2; z, \beta), \dots].$$

It is worth comparing this assumption with Mandl's identifiability assumption (1.1). Whereas the latter requires that (4.1) holds for all $z \in Z(i)$, A6.1 requires that it hold only for a dense subset of $Z(i)$. Suppose $Z(i)$ is subset of R^n as is usually the case. Then for $\alpha \neq \beta$ equality will hold in (4.1) for a set of $z \in R^n$ of dimension less than n and then A6.1 is likely to hold even when (1.1) does not.

Consider now the following random perturbation of the given adaptive law ξ . For each i let μ_i be a probability measure on $Z(i)$ which assigns positive values to every open set. Pick $\varepsilon_i > 0$ small, and for each $z \in Z(i)$ let $B(i, z)$ be the open ball of

radius ε_i and center z . Suppose at time n , $\alpha_n = \alpha$ is the MLE and $x_n = i$. Then the control z_n is chosen from $B(i, \xi(\alpha, i))$ by an independent experiment corresponding to the restriction of μ_i to the set $B(i, \xi(\alpha, i))$. Let $\mathcal{G}_n = \sigma(x_0, z_0, \dots, x_n, z_n)$ and $\mathcal{G}'_n = \sigma(x_0, z_0, \dots, x_n, z_n, x_{n+1})$. Then

$$P(x_{n+1} = j | \mathcal{G}_n) = p(x_n, j; z_n, \alpha_0),$$

$$P(z_{n+1} \in C | \mathcal{G}'_n) = \mu_{x_{n+1}}(C) [\mu_{x_{n+1}}(B(x_{n+1}, \xi(\alpha_{n+1}, x_{n+1})))^{-1}]$$

for every open set $C \subset B(x_{n+1}, \xi(\alpha_{n+1}, x_{n+1}))$. The results obtained previously continue to hold if we use \mathcal{G}_n in place of \mathcal{F}_n . The control law $\{z_n\}$ is called an $\{\varepsilon_i\}$ -randomization of ξ .

THEOREM 4.3. *Under any $\{\varepsilon_i\}$ -randomization of ξ , $\alpha_n \rightarrow \alpha_0$ a.s.*

Proof. Let $\tilde{Z}(i)$ be a countable dense subset of $Z(i)$ and $\tilde{\mathcal{B}}_i$ the set of all open balls in $Z(i)$ with rational radius and center in $\tilde{Z}(i)$. By the martingale stability theorem there is a null set N_1 outside which

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{m=0}^{n-1} [I(z_m \in \tilde{B}, x_m = i) - E\{I(z_m \in \tilde{B}, x_m = i) | \mathcal{G}'_{m-1}\}] = 0$$

for every i and \tilde{B} which is a union of finitely many elements of $\tilde{\mathcal{B}}(i)$. As in the proof of Theorem 3.2 (see (3.6)), there is a null set N_2 outside which $\lim_{n \rightarrow \infty} (1/n) \sum_{m=0}^{n-1} I(x_m = i) > 0$ for every $i \in S$. Finally let N_3 be the null set in Corollary 3.2.

Let $\omega \notin N = N_1 \cup N_2 \cup N_3$ and $\alpha^* \neq \alpha_0$ a limit point of $\{\alpha_n(\omega)\}$. By A6.1 there exists $i \in S$ such that for every open set $0 \subset Z(i)$ there exist $z \in 0$ and $j \in S$ for which $p(i, j; z, \alpha^*) \neq p(i, j; z, \alpha_0)$. Let $\{n_k, k \geq 0\}$ be the maximal subsequence for which $x_{n_k}(\omega) = i$ for every k . By Lemma 2.2 the sequence $\{z_{n_k}\}$ converges to the compact set of its limit points $Z^* \subset Z(i)$. For each $z^* \in Z^*$, let $O(z^*)$ be the open ball with radius $\varepsilon_i/4$ and center z^* . By A6.1 there exist $z \in O(z^*)$ and $j = j(z^*)$ in S such that $p(i, j; \tilde{z}, \alpha^*) \neq p(i, j; \tilde{z}, \alpha_0)$. By A1, there is $\tilde{B}(z^*) \in \tilde{\mathcal{B}}(i)$ such that $\tilde{z} \in \tilde{B}(z^*) \subset O(z^*)$ and

$$\left| \ln \frac{p(i, j; z, \alpha^*)}{p(i, j; z, \alpha_0)} \right| > \delta(z^*) > 0, \quad z \in \tilde{B}(z^*).$$

The family $\{O(z^*), z^* \in Z^*\}$ is an open cover of Z^* . Let $O(z_1^*), \dots, O(z_M^*)$ be a finite subcover. Let $O = \bigcup_m O(z_m^*)$, $\delta^* = \min_m \delta(z_m^*)$, $\tilde{B} = \bigcup_m \tilde{B}(z_m^*)$, $\mu^* = \min_m \mu_i(\tilde{B}(z_m^*))$, $J^* = \max_m j(z_m^*)$.

Now, for some $K < \infty$, $\xi(\alpha_{n_k}, i) \in O$ for $k \geq K$. Hence for $k \geq K$ there is a z_m^* such that $\tilde{B}(z_m^*) \subset B(\xi(\alpha_{n_k}, i))$. Let \hat{z}_{n_k} be this z_m^* . Then by (3.6)

$$\begin{aligned} & \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{m=0}^{n-1} I(z_m \in \tilde{B}, x_m = i) \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{m=0}^{n-1} E\{I(z_m \in \tilde{B}, x_m = i) | \mathcal{G}'_{m-1}\} \\ &\geq \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{m=K}^{n-1} \mu_i(\tilde{B}(\hat{z}_{n_k})) [\mu_i(B(\xi(\alpha_{n_k}, i)))]^{-1} I(m = n_k \text{ for some } k) \\ &\geq \mu^* \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{m=K}^{n-1} I(x_m = i) > 0. \end{aligned}$$

Hence the sequence

$$A_n = \left\{ I(x_n - i) \left| \ln \frac{p(i, j; z_n, \alpha^*)}{p(i, j; z_n, \alpha_0)} \right| > \delta \text{ for some } j \leq J^* \right\}$$

satisfies $\lim (1/n) \sum^{n-1} I(A_m) > 0$, contradicting Corollary 3.2. \square

Let $c(i, j, z)$ be a nonnegative bounded cost function which is continuous in $z \in Z(i)$ for each i, j . For every stationary $\zeta \in Z$ and $\alpha \in A$, let

$$V(\zeta, \alpha) = \sum_i \pi_\alpha(i, \zeta) \sum_j p(i, j; \zeta(i), \alpha) c(i, j, \zeta(i)),$$

where $\{\pi_\alpha(i, \zeta)\}$ are the stationary probabilities corresponding to the transition probabilities $\{p(i, j; \zeta(i), \alpha)\}$. Thus $V(\zeta, \alpha)$ is the cost incurred by the control law $z_n = \zeta(x_n)$ if α is the true parameter.

We assume henceforth that A5 holds for all α .

LEMMA 4.2. $V(\zeta, \alpha)$ is continuous in ζ and α .

Proof. As in the proof of Lemma 2.9 we see that $\pi_\alpha(i, \zeta)$ is continuous in ζ, α for each i . Given $\delta > 0$, it follows from Lemma 2.9 and the boundedness of c that there exists J such that

$$V(\zeta, \alpha) \geq \sum_{i=1}^J \pi_\alpha(i, \zeta) \sum_{j=1}^\infty p(i, j; \zeta(i), \alpha) c(i, j, \zeta(i)) \geq V(\zeta, \alpha) - \delta.$$

Now $\sum_{j=1}^n p(i, j; \zeta(i), \alpha)$ is continuous in ζ and α and converges monotonically, hence uniformly, to 1 as $n \rightarrow \infty$ and so, increasing J if necessary, we get

$$V(\zeta, \alpha) \geq \sum_{i=1}^J \pi_\alpha(i, \zeta) \sum_{j=1}^J p(i, j; \zeta(i), \alpha) c(i, j, \zeta(i)) \geq V(\zeta, \alpha) - 2\delta.$$

The term in the middle is continuous in ζ and α and since $\delta > 0$ is arbitrary it follows that $V(\zeta, \alpha)$ is continuous as well. \square

Suppose now that the given adaptive law ξ is such that for each α ,

$$V(\xi(\alpha), \alpha) = V(\alpha) = \min_{\zeta \in Z} V(\alpha, \zeta).$$

We wish to show that if $\{z_n\}$ is an $\{\varepsilon_i\}$ -randomization of ξ then its cost can be made arbitrarily close to $V(\alpha_0)$ by choosing $\varepsilon_i > 0$ sufficiently small.

THEOREM 4.4. Suppose $V(\xi(\alpha_0), \alpha_0) < V(\zeta, \alpha_0)$ when $\zeta \neq \xi(\alpha_0)$; i.e., $\xi(\alpha_0)$ is the unique optimal stationary control law. For any $\delta > 0$, there exists $\varepsilon > 0$ such that if $\{z_n\}$ is an $\{\varepsilon\}$ -randomization of ξ , then

$$V(\alpha_0) \leq \overline{\lim} \frac{1}{n} \sum_{m=0}^{n-1} c(x_m, x_{m+1}, z_m) \leq V(\alpha_0) + \delta \quad \text{a.s.}$$

Proof. By Lemma 4.2 there exists an open set O in Z with $\xi(\alpha_0) \in O$ such that $V(\alpha_0) \leq V(\xi, \alpha_0) < V(\alpha_0) + \delta$ for $\zeta \in O$. Since $Z = \prod_i Z(i)$ has the product topology we may suppose that O has the form

$$O = \prod_{i=1}^m B(i) \times \prod_{i=m+1}^\infty Z(i)$$

for some $m < \infty$ and where $B(i)$ is a ball of radius ε and center $\xi(\alpha_0, i)$. Let $\{z_n\}$ be an $\{\varepsilon\}$ -randomization of ξ . By Theorem 4.3, $\alpha_n \rightarrow \alpha_0$ a.s. and so, from Lemma 4.2 and

the uniqueness of $\xi(\alpha_0)$ it follows that $\xi(\alpha_n) \rightarrow \xi(\alpha_0)$ a.s. Hence there exists a random time $N < \infty$ a.s. such that

$$\xi(\alpha_n) \in O, \quad n \geq N \quad \text{a.s.}$$

By Lemma 2.10 it follows that

$$\overline{\lim} \frac{1}{n} \sum_{m=0}^{n-1} c(x_m, x_{m+1}, z_m) \leq \sup_{\xi \in O} V(\xi, \alpha_0) \leq V(\alpha_0) + \delta. \quad \square$$

Randomization of parameter estimates. We consider an alternative perturbation of the given adaptive law ξ . For the finite case a similar randomization is proposed in [10]. We replace A6.1 by the following.

A6.2. For every $\alpha \neq \beta$ in A and i in S and every neighborhood O of α there is an open set $\tilde{O} \subset O$ such that, for every $\tilde{\alpha} \in \tilde{O}$,

$$[p(i, 1; \xi(\tilde{\alpha}, i), \alpha), p(i, 2; \xi(\tilde{\alpha}, i), \alpha), \dots] \neq [p(i, 1; \xi(\tilde{\alpha}, i), \beta), p(i, 2; \xi(\tilde{\alpha}, i), \beta), \dots].$$

Let ν be a probability measure on A which assigns positive values to every open set. Pick $\gamma > 0$ small and let $B(\alpha)$ denote the ball of radius γ and center α . Let α_n be the MLE and x_n the state at time n . Then the control z_n is chosen to be $z_n = \xi(\tilde{\alpha}_n, x_n)$ where $\tilde{\alpha}_n$ is selected from $B(\alpha_n)$ by an independent experiment corresponding to the restriction of ν to $B(\alpha_n)$. We call the control law $\{z_n\}$ a γ -randomization of ξ .

THEOREM 4.5. Under $\{z_n\}$, α_0 is the only frequent limit point of $\{\alpha_n\}$ almost surely.

Proof. Let N_1 be the null set in Theorem 3.1. Let \mathcal{B} be the family of all open balls with rational radii and centers in a countable dense subset of A . Let N_2 be the null set outside which

$$\lim \frac{1}{n} \sum_{m=0}^{n-1} [I(x_m = i, \tilde{\alpha}_m \in \tilde{B}, \alpha_m \in B) - E\{I(x_m = i, \tilde{\alpha}_m \in \tilde{B}, \alpha_m \in B) | \mathcal{G}'_{m-1}\}] = 0$$

for all \tilde{B}, B in \mathcal{B} and i in S .

Let $N = N_1 \cup N_2$, $\omega \notin N$, and $\alpha^* \neq \alpha_0$ a frequent limit point of $\{\alpha_n(\omega)\}$. Because of Lemma 4.1 we can assume that α^* is regular. Let i in S be such that (i, α^*) is frequent along ω . Let O be the ball of radius $\gamma/2$ with center α^* . By A6.2 it follows that there is \tilde{B} in \mathcal{B} , with $\tilde{B} \subset O$, and j in S such that for $\tilde{\alpha} \in \tilde{B}$

$$\left| \ln \frac{p(i, j; \xi(\tilde{\alpha}, i), \alpha^*)}{p(i, j; \xi(\tilde{\alpha}, i), \alpha_0)} \right| > \delta$$

for some $\delta > 0$. Hence by A1 there is a neighborhood B in \mathcal{B} of α^* , with $B \subset O$, such that

$$\left| \ln \frac{p(i, j; \xi(\tilde{\alpha}, i), \alpha)}{p(i, j; \xi(\tilde{\alpha}, i), \alpha_0)} \right| > \frac{\delta}{2}$$

for $\tilde{\alpha} \in \tilde{B}$, $\alpha \in B$. So

$$\begin{aligned} & \overline{\lim} \frac{1}{n} \sum_{m=0}^{n-1} I\left\{x_m = i \text{ and } \left| \ln \frac{p(i, j; \xi(\tilde{\alpha}_m, i), \alpha_n)}{p(i, j; \xi(\tilde{\alpha}_m, i), \alpha_0)} \right| > \frac{\delta}{2}\right\} \\ & \geq \overline{\lim} \frac{1}{n} \sum_{m=0}^{n-1} I(x_m = i, \tilde{\alpha}_m \in \tilde{B}, \alpha_n \in B) \\ & \geq \overline{\lim} \frac{1}{n} \sum_{m=0}^{n-1} I(x_m = i, \tilde{\alpha}_m \in \tilde{B}, \alpha_m \in B) \end{aligned}$$

$$\begin{aligned}
&= \overline{\lim} \frac{1}{n} \sum_{m=0}^{n-1} E\{I(x_m = i, \tilde{\alpha}_m \in \tilde{B}, \alpha_m \in B) | \mathcal{G}'_{m-1}\} \\
&= \overline{\lim} \frac{1}{n} \sum_{m=0}^{n-1} \frac{\nu(\tilde{B})}{\nu(B(\alpha_m))} I(x_m = i, \alpha_m \in B) \quad \text{since } \tilde{B} \subset B(\alpha_m) \text{ if } \alpha_m \in B \\
&\geq \nu(\tilde{B}) \overline{\lim} \frac{1}{n} \sum_{m=0}^{n-1} I(x_m = i, \alpha_m \in B) > 0.
\end{aligned}$$

But this contradicts Theorem 3.1. \square

Suppose now that $V(\xi(\alpha), \alpha) = V(\alpha)$ as in Theorem 4.4.

THEOREM 4.6. *Suppose $\xi(\alpha_0)$ is the unique optimal stationary control law under α_0 . For any $\delta > 0$ there exists $\gamma > 0$ such that if $\{z_n\}$ is a γ -randomization of ξ , then*

$$V(\alpha_0) \leq \overline{\lim} \frac{1}{n} E \sum_{m=0}^{n-1} c(x_m, x_{m+1}, z_m) \leq V(\alpha_0) + \delta.$$

Proof. Let

$$v = \sup_{\zeta \in \hat{Z}} \sum_i \pi(i, \zeta) \sum_j p(i, j; \zeta(i), \alpha_0) c(i, j, \zeta(i)),$$

where $\hat{Z} = \prod \hat{Z}(i)$ and $\hat{Z}(i) = \{\zeta(\alpha, i) | \alpha \in B(\alpha_0)\}$. Because of A5, [7, Corollary 6.20, p. 149] applies, and so there is a bounded function $h(i)$, $i \in S$ such that for all i in S

$$v + h(i) = \sup_{z \in \hat{Z}(i)} \left\{ \sum_j p(i, j; z, \alpha_0) [c(i, j, z) + h(j)] \right\}.$$

Now let $\{z_n\}$ be a γ -randomization of ξ , and let T be the random set of time instances such that $z_n \in \hat{Z}(x_n)$. By Theorem 4.5 and Lemma 2.2 (iv) the sequence of events $\{n \notin T\}$ is rare. Now

$$v + h(x_m) \geq \sum_j p(x_m, j; z_m, \alpha_0) [c(x_m, j, z_m) + h(j)], \quad m \in T.$$

Let δ_m be such that

$$v + h(x_m) = \sum_j p(x_m, j; z_m, \alpha_0) [c(x_m, j, z_m) + h(j)] + \delta_m, \quad m \notin T.$$

Using these relations and taking expectation gives

$$v \geq -Eh(x_m) + Eh(x_{m+1}) + Ec(x_m, x_{m+1}, z_m) + E\delta_m I(m \notin T).$$

So,

$$\begin{aligned}
v &\geq \frac{1}{n} \sum_{m=0}^{n-1} E\{-h(x_m) + h(x_{m+1}) + c(x_m, x_{m+1}, z_m)\} + \Delta_n \\
&= \frac{1}{n} \sum_{m=0}^{n-1} Ec(x_m, x_{m+1}, z_m) + \frac{1}{n} [Eh(x_n) - Eh(x_0)] + \Delta_n, \\
\Delta_n &= \frac{1}{n} \sum_{m=0}^{n-1} E\delta_m I(m \notin T).
\end{aligned}$$

The sequence $\{n \notin T\}$ is rare and so $\Delta_n \rightarrow 0$ as $n \rightarrow \infty$. Hence

$$v \geq \overline{\lim} \frac{1}{n} \sum_{m=0}^{n-1} Ec(x_m, x_{m+1}, z_m).$$

By choosing the radius γ of the ball $B(\alpha_0)$ small enough v can be bounded by $V(\alpha_0) + \delta$. \square

We close this section with the remark that A6.1 may be replaced by A6.3 which is more similar to A6.2.

A6.3. For every $\alpha \neq \beta$ in A and i in S and every neighborhood O of $\xi(\alpha, i)$ there exists $z \in Z(i)$ for which (4.1) holds.

A heuristic discussion of A6.1–A6.3 is deferred to § 6.

5. Inadequate parameter sets. So far most of the results were derived under the assumption $\alpha_0 \in A$. In § 3, a crucial role is played by this assumption in the proof of Lemma 3.2 where we use the fact that $\Lambda_n(\alpha_n) \geq \Lambda_n(\alpha_0) = 1 > 0$. This observation motivates the following definition.

DEFINITION 5.1. The parameter set A is *adequate* if $P\{\Lambda(\alpha) = 0 \text{ for all } \alpha\} = 0$, or equivalently, if $P\{\sup_\alpha \Lambda(\alpha) > 0\} = 1$; otherwise A is *inadequate*.

Evidently if $\alpha_0 \in A$ then A is adequate. The following results are proved along the same lines as in § 3.

THEOREM 5.1. *If A is adequate then Theorem 3.1 and Corollary 3.2 continue to hold.*

COROLLARY 5.1. *If A is inadequate then the conclusions of Theorem 3.1 and Corollary 3.2 hold outside the set $N = \{\omega \mid \Lambda(\alpha, \omega) = 0 \text{ for all } \alpha\}$. (Note that $P(N) > 0$.)*

From now on we consider the case when A is inadequate. Suppose initially that there are no control parameters so that α merely indexes a stationary transition probability $p(i, j; \alpha)$. Suppose further that under each α all states are positive recurrent. Then it is easy to establish the following result which states that the MLE α_n converges to a subset of A consisting of parameter values which are “closest” to α_0 in a well-defined sense.

THEOREM 5.2. *α_n converges almost surely to the subset of parameter values which maximize*

$$D(\alpha) = \sum_i \pi_{\alpha_0}(i) \sum_j p(i, j; \alpha_0) \ln \frac{p(i, j; \alpha)}{p(i, j; \alpha_0)},$$

where $\{\pi_{\alpha_0}(i)\}$ are the invariant probabilities corresponding to $\{p(i, j; \alpha_0)\}$.

The case when the transition probabilities do depend on a control parameter, discussed next, is considerably more complicated. For simplicity we assume that $Z(i) = \{z_1^i, z_2^i, \dots, z_L^i\}$ contains L elements. The general case can be worked out in a similar way though the details are cumbersome.

Let A1–A5 hold and let N be the set on which Condition T fails. N is null by Lemma 2.11. Let $\omega \notin N$ and let

$$q_n(i, j, l) = \frac{1}{n} \sum_{m=0}^{n-1} I(x_m = i, x_{m+1} = j, z_m = z_l^i).$$

Then $0 \leq q_n \leq 1$ and $\sum_{i,j,l} q_n(i, j, l) = 1$. Thus q_n defines a (random) probability measures on triples (i, j, l) . Let $\rho_n = \{q_n(i, j, l)\}$ denote the vector with components $q_n(i, j, l)$. We think of ρ_n as an element of the normed space l_1 .

LEMMA 5.1. *There is a null set N such that if $\omega \notin N$, the l_1 closure of the sequence $\{\rho_n(\omega)\}$ is compact, and each of its limit points is itself a probability. The set of frequent limit points of $\{\rho_n(\omega)\}$ is compact and for any open neighborhood O of this set the sequence $\{\rho_n(\omega) \notin O\}$ is rare.*

Proof. Let N_1 be the set on which Condition T fails, and $\omega \notin N_1$. By Lemma 2.11 there exist J_1 and M_1 such that

$$\frac{1}{n} \sum_{i=1}^{n-1} I(x_m = i, i > J_1) < \varepsilon, \quad n > M_1$$

and so

$$(5.1) \quad \sum_{i \leq J_1} \sum_j \sum_l q_n(i, j, l) \geq 1 - \varepsilon, \quad n > M_1.$$

By the martingale stability theorem there is a null set N_2 outside which, for each i and J_2 ,

$$\lim_n \left| \sum_{j \geq J_2} \sum_l q_n(i, j, l) - \frac{1}{n} \sum_{i=1}^{n-1} E\{I(x_m = i, x_{m+1} \geq J_2) | \mathcal{F}_m\} \right| = 0,$$

equivalently,

$$\lim_n \left| \sum_{j \geq J_2} \sum_l q_n(i, j, l) - \sum_{j \geq J_2} p(i, j; z_m, \alpha_0) \frac{1}{n} \sum_{i=1}^{n-1} I(x_m = i) \right| = 0.$$

Hence if $\omega \notin N_2$, there exist J_2 and M_2 such that

$$\sum_{j \geq J_2} \sum_l q_n(i, j, l) \leq \varepsilon + \max_l \sum_{j \geq J_2} p(i, j; z_l^i, \alpha_0) \leq 2\varepsilon, \quad n > M_2.$$

From this and (5.1) we see that for any $\omega \notin N = N_1 \cup N_2$, and $\varepsilon > 0$, there exist J, M such that

$$\sum_{i, j \leq J} \sum_l q_n(i, j, l) \geq 1 - \varepsilon, \quad n > M.$$

By [5, p. 338], $\{\rho_n(\omega)\}$ is conditionally compact in l_1 . Let $\rho^* = \{q^*(i, j, l)\}$ be a limit point. Suppose $\rho_{n_k} \rightarrow \rho^*$ in l_1 . Evidently ρ^* is a probability. The remaining assertions follow from Lemma 2.2. \square

COROLLARY 5.2. *If $c(i, j, l)$ is any bounded function then*

$$\lim_k \sum_{i, j, l} c(i, j, l) q_{n_k}(i, j, l) = \sum_{i, j, l} c(i, j, l) q^*(i, j, l).$$

As before let $\{\pi(i, \zeta)\}$ denote the invariant probabilities under the stationary law $\zeta \in Z$. Let $q_\zeta(i, j, l) = \pi(i, \zeta) p(i, j; \zeta(i), \alpha_0) I(\zeta(i) = z_l^i)$. Let $q_\zeta = \{q_\zeta(i, j, l)\}$. Consider $\{q_\zeta | \zeta \in Z\}$ as a subset of l_1 and let G be its closed convex hull.

LEMMA 5.2. *There exists a null set N such that if $\omega \notin N$ then every limit point of $\{\rho_n\}$ belongs to G .*

Proof. Let $\varepsilon > 0$. By Lemma 2.9 there is J_1 such that $\sum_{i \leq J_1} \pi(i, \zeta) \geq 1 - \varepsilon$ for all ζ . Let $J > J_1$ such that $\sum_{j \leq J} p(i, j; \zeta(i), \alpha_0) \geq 1 - \varepsilon$ for $i \leq J_1$ and all ζ . Then

$$\sum_{i, j \leq J} \sum_l q_\zeta(i, j, l) = \sum_{i \leq J_1} \pi(i, \zeta) \sum_{j \leq J} p(i, j; \zeta(i), \alpha_0) \geq (1 - \varepsilon)^2.$$

Hence $\{q_\zeta, \zeta \in Z\}$ is conditionally compact in l_1 . Therefore G is weakly compact in l_1 by [5, p. 434], and hence by [5, p. 338] G is compact.

For each J let l_∞^J be the finite-dimensional subspace of all functions $c(i, j, l)$ in

l_∞ such that $c(i, j, l) = 0$ for $i > J$ or $j > J$. Then there is a null set N such that, for $\omega \notin N$ and for every J and every c in l_∞^J ,

$$(5.2) \quad \overline{\lim} \frac{1}{n} \sum_{m=0}^{n-1} c(x_m, x_{m+1}, z_m) \leq \max_{\zeta} \sum_i \pi(i, \zeta) \sum_j p(i, j; \zeta(i), \alpha_0) c(i, j, \zeta(i)).$$

The existence of N follows from Lemma 2.10 by first showing (5.2) for a countable dense subset of $\bigcup_J l_\infty^J$.

Augment N if needed so that Lemma 5.1 applies. Let $\omega \notin N$, and let $\rho^* = \{q^*(i, j, l)\}$ be a limit point of $\{\rho_n(\omega)\}$. From Corollary 5.2 and (5.2),

$$(5.3) \quad \begin{aligned} \sum_{i,j,l} c(i, j, l) q^*(i, j, l) &\leq \max_{\zeta} \sum_i -(i, \zeta) \sum_j p(i, j; \zeta(i), \alpha_0) c(i, j, \zeta(i)) \\ &= \max_{g \in G} \sum_{i,j,l} c(i, j, l) g(i, j, l) \end{aligned}$$

for every J and c in l_∞^J . Now suppose $c \in l_\infty$. Since $G \cup \{\rho^*\}$ is compact in l_1 therefore, for every $\delta > 0$, there exists $J < \infty$ such that

$$(5.4) \quad \left| \sum_{i,j,l} c(i, j, l) g(i, j, l) - \sum_{i,j \leq J} \sum_l c(i, j, l) g(i, j, l) \right| < \delta$$

for $g \in G \cup \{\rho^*\}$. From this it follows that (5.3) holds for all $c \in l_\infty$. Finally if $\rho^* \notin G$ then by the separation theorem there is c in l_∞ such that

$$\sum_{i,j,l} c(i, j, l) [q^*(i, j, l) - g(i, j, l)] > 0 \quad \text{for all } g \text{ in } G,$$

which contradicts (5.3). \square

LEMMA 5.3. Let N be as in Lemma 5.1 and $\omega \notin N$. Suppose ρ_{n_k} converges to ρ^* along ω . Let $\rho_{\tilde{n}_k}$, $k \geq 0$ be another subsequence such that $|n_k - \tilde{n}_k| \leq M < \infty$ for all k . Then $\rho_{\tilde{n}_k}$ also converges to ρ^* along ω .

Proof. For any i, j, l , if $n_k \geq \tilde{n}_k$,

$$\begin{aligned} q_{\tilde{n}_k}(i, j, l) - q_{n_k}(i, j, l) &\leq \left(\frac{1}{\tilde{n}_k} - \frac{1}{n_k} \right) \sum_{m=0}^{n_k-1} I(x_m = i, x_{m+1} = j, z_m = z^i) \\ &\leq \left(\frac{1}{\tilde{n}_k} - \frac{1}{n_k} \right) n_k = \frac{n_k}{\tilde{n}_k} - 1. \end{aligned}$$

Also,

$$q_{n_k}(i, j, l) - q_{\tilde{n}_k}(i, j, l) \leq \frac{1}{n_k} \sum_{m=\tilde{n}_k}^{n_k-1} I(x_m = i, x_{m+1} = j, z_m = z^i) \leq \frac{M}{\tilde{n}_k}.$$

Thus $|q_{n_k}(i, j, l) - q_{\tilde{n}_k}(i, j, l)| \rightarrow 0$ as $k \rightarrow \infty$. \square

To describe the asymptotic behavior of $\{\rho_n\}$ we need another concept. Let $\{a_n, n \geq 0\}$ be a sequence in a metric space. Let \tilde{O}, O be open sets with $\tilde{O} \subset \bar{\tilde{O}} \subset O$. Let

$$m_k(\tilde{O}) = \min \{n > m_{k-1}(\tilde{O}) | a_{n-1} \notin \tilde{O}, a_n \in \tilde{O}\}$$

be the k th time a_n enters \tilde{O} after leaving it. Let

$$n_k(O) = \min \{n > m_k(\tilde{O}) | a_n \notin O\}, \quad l_k(O) = \max \{n \leq m_k(\tilde{O}) - 1 | a_n \notin O\}.$$

We say that $\{a_n\}$ drifts slowly if for any open sets O, \tilde{O} with $\tilde{O} \subset \bar{\tilde{O}} \subset O$, the sequences $\{n = n_k(O) \text{ for some } k\}$ and $\{n = l_k(O) \text{ for some } k\}$ are both rare.

THEOREM 5.3. *There exists a null set N such that if $\omega \notin N$ then $\{\rho_n(\omega)\}$ drifts slowly considered as a sequence in l_1 .*

Proof. Let N be as in Lemma 5.1 and $\omega \notin N$. Let O, \tilde{O} be open sets in l_1 with $\tilde{O} \subset \tilde{\tilde{O}} \subset O$, and define l_k, m_k, n_k as previously with $a_n = \rho_n(\omega)$. We may suppose $\rho_{m_k} \in \tilde{\tilde{O}}$ for infinitely many k because otherwise there is nothing to prove. We claim first that

$$(5.5) \quad m_k - l_k \rightarrow \infty \quad \text{as } k \rightarrow \infty.$$

For, suppose in contradiction that there is a subsequence k_i , with

$$m_{k_i} - l_{k_i} \leq M < \infty \quad \text{for all } i.$$

Now, by Lemma 5.1, there exists a subsequence, denoted again by $\{k_i\}$ such that

$$\lim_i \rho_{m_{k_i}} = \rho^* \quad \text{and} \quad \lim_i \rho_{l_{k_i}} = \tilde{\rho}.$$

Clearly $\rho^* \in \tilde{\tilde{O}}$ and $\tilde{\rho} \notin O \supset \tilde{\tilde{O}}$ so that $\rho^* \neq \tilde{\rho}$, contradicting Lemma 5.3. Thus (5.5) must hold. Hence either $l_{k+1} = l_k$ for all k from some \bar{k} on, in which case the claim is obvious, or $l_{k+1} > l_k$ infinitely often. In the latter case, let $\{k_i\}$ be the maximal subsequence of $\{k\}$ such that $l_{k_{i+1}} > l_{k_i}$. Then by (5.5) and the definition of l_{k_i} , $l_{k_{i+1}} - l_{k_i} \rightarrow \infty$, implying $l_{k_i}/i \rightarrow \infty$. Hence

$$\overline{\lim} \frac{1}{n} \sum_{m=1}^n I(m = l_k \text{ for some } k) = \overline{\lim} \frac{1}{n} \sum_{m=1}^n I(m = l_{k_i} \text{ for some } i) = \overline{\lim} \frac{i}{l_{k_i}} = 0.$$

Let $k(m) = \max \{k | l_k \leq m\}$. Then

$$\frac{1}{n} \sum_{i=1}^m I(i = l_k \text{ for some } k) \leq \frac{1}{k(m)} \sum_{i=1}^{k(m)} I(i = l_k \text{ for some } k) = \frac{k(m)}{l_{k(m)}} \rightarrow 0 \quad \text{as } m \rightarrow \infty. \quad \square$$

In summary the results of this section show that when $\alpha_0 \notin A$ (more precisely, when A is inadequate) and for any control law, the relative frequencies ρ_n of the various state and control combinations converge to a tight set of probabilities which is the convex hull of the set G of invariant probabilities under all stationary control laws. The sequence ρ_n may, however, drift slowly.

In [4] we have given an example which shows that ρ_n may not converge almost surely. In that example, $z_n = \xi(\alpha_n, x_n)$ is an adaptive law constructed in such a way that, as ρ_n begins to converge to some ρ^* and hence α_n to some α^* , the corresponding control values $z_n = \xi(\alpha^*, x_n)$ are such that the likelihood ratio is maximized at some other parameter value $\tilde{\alpha} \neq \alpha^*$ and so α_n begins to drift slowly to $\tilde{\alpha}$. But at $\tilde{\alpha}$, the control values $\xi(\tilde{\alpha}, x_n)$ are such that the likelihood ratio is maximized at α^* . Thus the MLE α_n keeps switching more and more slowly between α^* and $\tilde{\alpha}$.

The following conjecture seems plausible.

Conjecture. For α, β in A define

$$M(\alpha, \beta) = \sum_i \pi(i, \zeta(\alpha)) \sum_j p(i, j; \zeta(\alpha, i), \alpha_0) \ln \frac{p(i, j; \zeta(\alpha, i), \beta)}{p(i, j; \zeta(\alpha, i), \alpha_0)}.$$

Then a sufficient condition for the MLE α_n to converge almost surely to some A -valued random variable is

$$(5.6) \quad M(\alpha, \alpha) > M(\alpha, \beta)$$

for all $\alpha, \beta \in A$ such that $\alpha \neq \beta$.

If true, the above conjecture suggests the following choice of $\zeta(\alpha)$ for each α . Choose $\zeta(\alpha)$, from among the strategies that are near optimal under α , to make (5.6) hold for as many $\beta \neq \alpha$ as possible. This is a manifestation of the trade-off between identification and optimality considerations in the choice of inputs. Thus the control in the adaptive scheme fulfills a dual purpose—to ensure good convergence properties for the estimates (with, of course, convergence to the true parameter value if possible) and to satisfy the optimality criteria as closely as possible. Clearly, this discussion is only heuristic.

6. Discussion. The approach adopted here puts greater emphasis on “time domain” or sample path behavior and many of the concepts introduced can be seen as analogues of certain ensemble concepts, viz., rare events are analogues of null sets, Condition T is an analogue of tightness, etc. The reduced dependence on ensemble averages makes this approach more suitable for nonstationary processes which are asymptotically well behaved.

Many of the concepts introduced in § 2, such as recurrence and positive recurrence, can be extended to more general spaces such as an arbitrary Borel state space, and it seems reasonable to expect that similar results will hold.

Assumptions A6.1, A6.2 and A6.3 have the common objective of overcoming the limitations of Theorem 4.2 which says that at the limiting values of the parameter estimates, the frequent limits of control values are such that these control values cannot distinguish between different limits of the parameter estimates. This cannot occur if for each $\alpha \neq \alpha_0$ we frequently use a control to distinguish between α and α_0 . This can be achieved, as in A6.1 and A6.3, by a small randomization if for each α the set of control values which cannot distinguish between α and α_0 is “thin”, i.e., has empty interior. A6.2 permits an analogous randomization in parameter space. The latter appears more appealing for practical problems even though the result is slightly weaker.

The case when $\alpha_0 \notin A$ is practically important since models used for identification and control are approximations of the true system. The results presented here are very incomplete and much needs to be done.

Acknowledgment. The authors are grateful to the referee for catching some embarrassing mistakes in the previous version and for suggesting several improved proofs.

REFERENCES

- [1] K. ÅSTRÖM AND B. WITTENMARK, *On self-tuning regulators*, Automatica, 9 (1973), pp. 185–199.
- [2] P. MANDL, *Estimation and control in Markov chains*, Adv. Appl. Prob., 6 (1974), pp. 40–60.
- [3] L. LJUNG AND B. WITTENMARK, *Asymptotic properties of self-tuning regulators*, Rep. 7407, Dept. Automatic Control, Lund Inst. Technology, Lund, Sweden, 1974.
- [4] V. BORKAR AND P. VARAIYA, *Adaptive control of Markov chains, I: Finite parameter set*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 953–957.
- [5] N. DUNFORD AND J. SCHWARTZ, *Linear Operators Part I*, John Wiley, New York, 1964.
- [6] M. LOËVE, *Probability Theory*, 2nd ed., Van Nostrand, Princeton NJ, 1960.
- [7] S. ROSS, *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco, 1970.
- [8] A. FEDERGRUEN, A. HORDIJK AND H. C. TIJMS, *A note on simultaneous recurrence conditions on a set of denumerable stochastic matrices*, Rep. BW 85/77, Dept. Operations Research, Mathematisch Centrum, Amsterdam, 1977; J. Appl. Prob., 15 (1978), pp. 356–373.
- [9] J. NEVEU, *Discrete-Parameter Martingales*, North-Holland, Amsterdam, 1975.
- [10] B. DOSHI AND S. E. SHREVE, *Randomized self-tuning control of Markov chains*, preprint, Dept. of Mathematical Sciences, Univ. of Delaware, Newark, 1979; J. Appl. Prob., 17 (1980), pp. 726–734.
- [11] L. BREIMAN, *Probability*, Addison-Wesley, Reading, MA, 1968.

POINT CONTROL WITH STATE CONSTRAINTS*

L. W. WHITE†

Abstract. Controllability results are presented for pseudoparabolic problems with spacial domains in \mathbb{R}^p with $p = 1, 2$ or 3 , as well as compatibility conditions for state constraints of noncontrollable problems. A concrete problem is considered and a regularity result is given for the optimal control.

1. Introduction. In this note we consider a control problem with an underlying equation of pseudoparabolic (or parabolic) type. Control is exerted through a given point a in Ω , an open parallelopiped in \mathbb{R}^p with $p = 1, 2$ or 3 . We denote the space-time cylinder by $Q = \Omega \times (0, T)$ and its lateral boundary by $\Sigma = \partial\Omega \times (0, T)$. For simplicity we take

$$(1) \quad (1 - \Delta)y_t - \Delta y = v(t)\delta(x - a) \quad \text{in } Q,$$

$$(2) \quad y(x, 0; v) = 0 \quad \text{in } \Omega,$$

$$(3) \quad y(x, t; v) = 0 \quad \text{on } \Sigma$$

as the governing equations. Along with (1)–(3), we consider the problem

$$(4) \quad \begin{aligned} &\text{minimize } \|v\|_{L^2(0, T)}^2 \\ &\text{subject to } v \in L^2(0, T), \\ &\quad \|y(T; v) - z\|_{L^2(\Omega)} \leq \rho, \end{aligned}$$

where z is a given function in $L^2(\Omega)$ and $\|z\|_{L^2(\Omega)} > \rho > 0$.

Pseudoparabolic problems such as (1)–(3) arise in various contexts where higher order accuracy is required of the model such as in the study of second order fluids [10], in heat conduction [3] and in the consolidation of clays [9]. We refer to [2] for an extensive bibliography. Here we give a detailed application taken from the study of flow in porous media [1] in order to provide a physical setting in light of which the problem formulated in (1)–(4) may be interpreted.

Equations such as (1)–(3) arise in the modeling of seepage in fissured rock; see [1]. Here Ω is fissured rock consisting of pores and permeable blocks which in general are separated by a system of fissures. We associate with each point in Ω a pressure p_1 representing the average pressure of the liquid in the fissures in the neighborhood of the point and p_2 the average pressure of the liquid in the pores in the neighborhood of the given point.

Similarly, two velocities \bar{v}_1 and \bar{v}_2 are associated with the seepage at each point. The vector \bar{v}_1 of the seepage velocity of the liquid along the fissures is determined as follows: the projection of this vector in some particular direction is equal to the flow of the liquid through the cross-section of the fissures of a small zone passing through the given point perpendicular to the given direction, divided by the density of the liquid and the total area of this zone. The velocity \bar{v}_2 is defined similarly.

* Received by the editors October 21, 1980, and in revised form July 10, 1981. This work was supported in part by the National Science Foundation under grant MCS-7902037, and by the Institut National de Recherche en Informatique et en Automatique.

† Department of Mathematics, Energy Resources Center, University of Oklahoma, Norman, Oklahoma 73019.

It is characteristic of fissured rock that the flow of liquid proceeds essentially along the fissures so that the flow velocity of the liquid through the blocks is negligibly small as compared to the seepage along the fissures. It is assumed, however, that fissures are sufficiently narrow and the velocity sufficiently small so that the liquid motion along the fissures will be inertialess, fulfilling Darcy's law

$$(5) \quad \bar{v}_1 = -\frac{k_1}{\mu} \text{grad } p_1,$$

where k_1 is the permeability of the system of fissures and μ is the viscosity of the liquid.

Another characteristic feature of the nonsteady-state motion of a liquid in fissured rock is the transfer of liquid between blocks and fissures. This is given in [1] by

$$(6) \quad V = \frac{\alpha}{\mu} (p_2 - p_1),$$

where V is the volume of liquid which flows from the blocks into the fissures per unit time and unit volume of rock and α is a dimensionless characteristic of the fissured rock. Hence, the mass q of liquid flowing from pores to fissures per unit time, per unit volume is

$$(7) \quad q = \frac{\rho\alpha}{\mu} (p_2 - p_1),$$

where ρ is density of the liquid.

Now conservation of mass in the fissures may be written (see [1])

$$(8) \quad \frac{\partial m_1 \rho}{\partial t} + \text{div } \rho \bar{v}_1 - q = 0,$$

where m_1 is the ratio of volume of cavity space occupied by the fissures to the total volume of the rock. Since a very small part of the total volume is contained in the fissures, we may neglect the first term. Assuming the liquid is slightly compressible and neglecting small higher-order terms, we obtain

$$(9) \quad k_1 \Delta p_1 + \alpha (p_2 - p_1) = 0.$$

Similarly, conservation of mass in the pores is given by

$$(10) \quad \frac{\partial m_2 \rho}{\partial t} + \text{div } \rho \bar{v}_2 + q = 0.$$

Here, however, with lower permeability the fluid velocity in the blocks is small. Hence, ignoring higher-order terms, we have

$$(11) \quad \frac{\partial m_2 \rho}{\partial t} + q = 0.$$

Further, it is assumed [1] that the liquid in the blocks participates in supporting upper strata. This is in contrast to liquid in the fissures where influence of the pressure p_1 on porosity of the blocks can be disregarded as compared to the effect of the pressure p_2 . Thus, it can be assumed [1] that

$$(12) \quad dm_2 = \beta_{c_2} dp_2$$

where β_{c_2} is the coefficient of compressibility of the blocks. From the slight compressibility

$$(13) \quad \rho = \rho_0 + \beta \delta p,$$

it follows that [1]

$$(14) \quad (\beta_{c_2} + m_0\beta) \frac{\partial p_2}{\partial t} + \frac{\alpha}{\mu} (p_2 - p_1) = 0.$$

Using (9) to eliminate p_2 , we have an expression for the pressure p_1 in the fissures

$$(15) \quad \frac{\partial p_1}{\partial t} - \varepsilon \frac{\partial \Delta p_1}{\partial t} = \eta \Delta p_1,$$

where $\varepsilon = k_1/\alpha$ and $\eta = k_1/\mu(\beta_{c_2} + m_0\beta)$.

From the above derivation, problem (1)–(4) provides for the control of the average pressure y in the fissures by a point forcing term. The control is to be realized in such a way as to restrict the pressure $y(T)$ at a given time T to within a prescribed tolerance ρ of a desired fissure pressure z . This is to be accomplished using the minimum, according to the criterion, control v .

In order for problem (4) to have a solution, the set

$$U(\rho) = \{v \in L^2(0, T): \|y(T; v) - z\|_{L^2(\Omega)} \leq \rho\}$$

must be nonempty. Hence, we consider the controllability of problem (1)–(3). That is, we seek to determine whether the set given by

$$(16) \quad Y(T) = \{y(\cdot, T; v): v \in L^2(0, T)\}$$

is dense in $L^2(\Omega)$. Here we establish conditions for controllability of (1)–(3) for $\Omega \subset \mathbb{R}^p$ with $p = 1, 2$ or 3 . This is in contrast with the parabolic case in which controllability holds only in the case $p = 1$. If (1)–(3) is controllable, the $U(\rho)$ is nonempty for any $\rho > 0$. In the case that (1)–(3) is not controllable, we give here a compatibility condition depending on z in order for $U(\rho) \neq \emptyset$.

After having given criteria for the formulation of (4), we determine regularity of the optimal control u_0 . For ease we devote our attention to the case $\Omega = (0, 1) \subset \mathbb{R}^1$, pointing out extensions for $\Omega \subset \mathbb{R}^p$. Furthermore, we indicate results for the parabolic case and for more general equations.

2. Controllability of (1)–(3). We determine conditions under which $\overline{Y(T)} = L^2(\Omega)$. Hence, let $\xi \in L^2(\Omega)$ have the property that for every $v \in L^2(0, T)$,

$$(17) \quad (\xi, y(T; v))_{L^2(\Omega)} = 0.$$

Hence, we seek conditions under which (17) implies $\xi = 0$ in $L^2(\Omega)$. As is usual [6], [7], [11], we introduce the adjoint problem

$$(18) \quad -(1 - \Delta)p_t - \Delta p = 0 \quad \text{in } Q,$$

$$(19) \quad p(\cdot, T) = (1 - \Delta)^{-1} \xi(\cdot) \quad \text{in } \Omega,$$

$$(20) \quad p(x, t) = 0 \quad \text{on } \Sigma.$$

Multiplying (1) by p and integrating by parts, we see that (17) implies that for all $v \in L^2(0, T)$,

$$(21) \quad (p(a, \cdot), v)_{L^2(0, T)} = 0.$$

Hence, it follows that

$$(22) \quad p(a, t) = 0$$

for almost all t in $(0, T)$.

We consider the implications of (22). Changing variables, for convenience, $\tau = T - t$ with $q(\cdot, \tau) = p(\cdot, t)$, (18)–(20) become

$$(23) \quad (1 - \Delta)q_t - \Delta q = 0 \quad \text{in } Q,$$

$$(24) \quad q(\cdot, 0) = (1 - \Delta)^{-1} \xi(\cdot) \quad \text{in } \Omega,$$

$$(25) \quad q(x, \tau) = 0 \quad \text{on } \Sigma.$$

Now for $\Omega = (0, 1)$ with $\Delta = \partial^2 / \partial x^2$ and zero Dirichlet conditions at 0 and 1, we consider (23)–(25) in terms of the Fourier sine series. Set $\xi = \sum_{k=1}^{\infty} \xi_k \sin(k\Pi x)$. The solution of (23)–(25) may be given by

$$(26) \quad q(x, t) = \sum_{k=1}^{\infty} \frac{\xi_k}{1 + k^2 \Pi^2} e^{-\mu_k t} \sin(k\Pi x),$$

where $\mu_k = (k^2 \Pi^2) / (1 + k^2 \Pi^2)$. We note that q is the uniform limit of functions continuous in x and τ . Hence, we may evaluate $q(a, t)$ to have

$$(27) \quad q(a, t) = \sum_{k=1}^{\infty} \frac{\xi_k}{1 + k^2 \Pi^2} e^{-\mu_k t} \sin(k\Pi a)$$

for all $t > 0$. From (22) we have $q(a, t) = 0$ for all $t \in (0, T)$. In fact, $q(a, t)$ is analytic for $t > 0$ since from (27) it is the uniform limit of analytic functions. Hence, in fact, $q(a, t) = 0$ for all $t > 0$.

Since convergence is uniform in (27), we may take the Laplace transform to obtain

$$(28) \quad \hat{q}(a, s) = \sum_{k=1}^{\infty} \frac{\xi_k}{1 + k^2 \Pi^2} \sin(k\Pi a) \frac{1}{s + \mu_k}$$

with $\hat{q}(a, s) = 0$ for all $s > 0$. Define the function

$$f(z) = \sum_{k=1}^{\infty} \frac{\xi_k}{1 + k^2 \Pi^2} \sin(k\Pi a) \frac{1}{z + \mu_k}.$$

Now $f(z)$ is a meromorphic function with poles at $-\mu_k$. Furthermore, $f(z) = 0$ for z real and positive. We conclude then that $f(z) \equiv 0$ and has zero residues. Thus, we have

$$(29) \quad \xi_k \sin(k\Pi a) = 0$$

for all k . It is an immediate consequence of (29) that if $a \in (0, 1)$ is an irrational number, then $\xi_k = 0$ for every $k \in \mathbb{N}$. Hence, ξ is zero in $L^2(0, 1)$ and $Y(T) = L^2(0, 1)$.

THEOREM 1. *If $a \in (0, 1)$ is irrational, then (1)–(3) is controllable.*

In the case a is rational, say n/m , then (29) implies $\xi_k = 0$ only for $k \neq lm$ for any l in \mathbb{N} . Otherwise, $\sin(lm\Pi n/m) = \sin(lm\Pi) = 0$. Hence, if $\xi = \sum_{l=1}^{\infty} \xi_l \sin(lm\Pi x)$, then $\xi \perp Y(T)$ and $\xi \neq 0$. Thus, we may describe the orthogonal complement of $Y(T)$ and gain information about $Y(T)$ itself.

THEOREM 2. *If $a = n/m$, then $Y(T)^\perp$ is the span of the functions $\{\sin(lm\Pi x)\}_{l=1}^{\infty}$. Hence, $\overline{Y(T)}$ is the closure of the space of finite linear combinations of the functions $\{\sin k\Pi x\}_{k \neq lm}$.*

Now in the case of \mathbb{R}^p , with $p = 2$ or 3 , an analogous proof remains valid with, say, Ω a rectangle or a parallelopiped with sides having lengths that are independent over the integers; cf. [4], [5]. Hence, we have the following.

THEOREM 3. *If $\Omega \subseteq \mathbb{R}^p$, with $p = 2$ or 3 , and Ω is a rectangle or rectangular solid with sides having lengths that are independent over the integers and if $a \in \Omega$ is a p -tuple with components that are irrational when divided by the length of their respective sides, then the initial value problem (1)–(3) is controllable.*

Remark 4. These results are in contrast to those for parabolic equations which are controllable only for the case $\Omega \subset \mathbb{R}$ [7]. The proof here is true for higher dimensions because the factors $1/(1 + \lambda_k^2)$ in (26) (where $\lambda_k^2 = k^2 \Pi^2$) imply uniform convergence for $\xi \in L^2(\Omega)$. This, of course, is due to the presence of the term $1 - \Delta$ in (1).

3. A control problem. In this section we consider control of (1)–(3) by means of the following minimization problem with state constraints.

$$(30) \quad \begin{aligned} & \text{minimize } \|v\|_{L^2(0,T)}^2 \\ & \text{subject to } v \in L^2(0, T), \\ & \|y(T; v) - z\|_{L^2(\Omega)} \leq \rho, \end{aligned}$$

where $z \in L^2(\Omega)$. Hence, we have the set of admissible controls

$$U(\rho) = \{v \in L^2(0, T) : \|y(T; v) - z\|_{L^2(\Omega)} \leq \rho\}.$$

In order to have a meaningful problem, it is necessary that U be nonempty. An immediate consequence of Theorems 1 and 3 is the following.

COROLLARY 5. *If $a \in (0, 1)$ is irrational then $U(\rho)$ is nonempty for any $\rho > 0$.*

COROLLARY 6. *Let Ω be a rectangle or rectangular solid satisfying the assumption in Theorem 3. Then $U(\rho)$ is nonempty for any $\rho > 0$ whenever all the components of a are irrational.*

Suppose now that for (1)–(3) the number a is rational. A similar argument holds for a being a p -tuple with one or more rational components. In this case it has been shown in Theorem 2 that $Y(T)$ is a proper subspace of $L^2(\Omega)$. Now certainly in order for $U(\rho)$ to be nonempty, the condition

$$(31) \quad \rho > d = \text{minimum}_{y \in Y(T)} \|y - z\|_{L^2(\Omega)}$$

must be satisfied.

To fix ideas let $z \in L^2(0, 1)$ and $a = n/m$. Then we may express z in terms of a Fourier series $z = \sum_{k=1}^{\infty} \zeta_k \sin(k\Pi x)$. Furthermore, we may write

$$z = \sum_{k \neq lm} \zeta_k \sin(k\Pi x) + \sum_{l=1}^{\infty} \zeta_{lm} \sin(lm\Pi x).$$

Thus, it follows that

$$\begin{aligned} d &= \left\| z - \sum_{k \neq lm} \zeta_k \sin(k\Pi x) \right\|_{L^2(0,1)} \\ &= \left\| \sum_{l=1}^{\infty} \zeta_{lm} \sin(lm\Pi x) \right\|_{L^2(0,1)}, \end{aligned}$$

so that

$$(32) \quad d = \left(\sum_{l=1}^{\infty} \zeta_{lm}^2 \right)^{1/2},$$

and we have

$$(33) \quad d = \frac{1}{2} \left[\sum_{l=1}^{\infty} (z, \sin(lm \Pi x))_{L^2(0,1)}^2 \right]^{1/2}.$$

Hence, we have the following.

THEOREM 7. *Let $z \in L^2(\Omega)$. If ρ satisfies $\rho > d$, where d is given by (33), then $U(\rho)$ is nonempty. That is, there exists $v \in L^2(0, T)$ such that $\|y(T; v) - z\|_{L^2(0,1)} < \rho$.*

Having determined conditions to guarantee that $U(\rho)$ is nonempty, we now consider problem (30).

LEMMA 8. *The admissible set $U(\rho)$ is a nonempty closed convex set in $L^2(0, T)$.*

Proof. That $U(\rho)$ is nonempty is a consequence of Theorem 7. Certainly $U(\rho)$ is convex from the linearity of the map $v \rightarrow y(T; v)$ and the triangle inequality.

Let $v_n \rightarrow v$ in $L^2(0, T)$. Again using (18)–(20) with $\xi \in L^2(\Omega)$ and multiplying (1) by p , we see that

$$(34) \quad (\xi, y(T; v_n))_{L^2(\Omega)} = (p(a, \cdot), v_n)_{L^2(0,T)}$$

for all v_n . In the limit we have

$$(\xi, y(T; v))_{L^2(\Omega)} = (p(a, \cdot), v)_{L^2(0,T)}$$

so that $y(T; v_n) \rightarrow y(T; v)$ weakly in $L^2(\Omega)$. Now there is a sequence of convex combinations $\sum \theta_{n_i} y(T; v_{n_i}) = y(T; \sum \theta_{n_i} v_{n_i})$ that converge strongly to $y(T; v)$. Thus, it follows that

$$\begin{aligned} \|y(T; v) - z\|_{L^2(\Omega)} &\leq \|\sum \theta_{n_i} y(T; v_{n_i}) - z\|_{L^2(\Omega)} \\ &\leq \sum \theta_{n_i} \|y(T; v_{n_i}) - z\|_{L^2(\Omega)}. \end{aligned}$$

Hence, we have $\|y(T; v) - z\|_{L^2(\Omega)} \leq \rho$ since $v_{n_i} \in U(\rho)$, and $v \in U(\rho)$ so that $U(\rho)$ is closed.

We immediately have the following.

THEOREM 9. *There exists a unique solution u_0 of problem (30).*

We now determine regularity properties of u_0 . First, we note as a consequence of the controllability and compatibility results that there exists $v \in L^2(0, T)$ such that $\|y(T; v) - z\|_{L^2(\Omega)} < \rho$. This implies, cf. [11], the existence of a positive Lagrange multiplier (see, also, [8]).

THEOREM 10. *If the conditions of Corollary 5 or 6 hold, or if the compatibility condition of Theorem 7 holds, there exists a positive number λ such that for*

$$\Lambda(v) = \|v\|_{L^2(0,T)}^2 + \lambda (\|y(T; v) - z\|_{L^2(\Omega)}^2 - \rho^2),$$

the following holds:

$$(35) \quad \|u_0\|_{L^2(0,T)}^2 = \min_{v \in U(\rho)} \|v\|_{L^2(0,T)}^2 = \min_{v \in U(\rho)} \Lambda(v) = \Lambda(u_0).$$

As a corollary of Theorem 10, we have:

COROLLARY 11.

$$\|y(T; u_0) - z\|_{L^2(\Omega)} = \rho.$$

Calculating the variation of Λ (cf. [11]) we see that $\Lambda(u_0)(v) = 0$ for all $v \in L^2(0, T)$ so that

$$(36) \quad 0 = (u_0, v)_{L^2(0,T)} + \lambda (y(T; u_0) - z, y(T; v))_{L^2(\Omega)}.$$

By introducing the adjoint problems (18)–(20) with $\xi = y(T; u_0) - z$, we obtain

$$(u_0 + \lambda p(a, \cdot), v)_{L^2(0, T)} = 0$$

for all $v \in L^2(0, T)$. Hence, we see that

$$(37) \quad u_0(t) = -\lambda p(a, t)$$

for almost all t in $(0, T)$. But $t \rightarrow p(\cdot, t)$ is an analytic map of $(0, T)$ into $H_0^1(\Omega) \cap H^2(\Omega)$. Thus, we have the following as a consequence of (37).

THEOREM 12. *The optimal control u_0 is equal almost everywhere in $(0, T)$ to an analytic function.*

Remark 13. The existence of a positive Lagrange multiplier λ is dependent upon knowing that there is a v satisfying $\|y(T; v) - z\|_{L^2(\Omega)} < \rho$. Hence, compatibility and controllability conditions play an important role here as well. The result clearly holds for Ω a rectangle or a parallelepiped with sides parallel to the coordinate axes.

Remark 14. Similar arguments to those in the controllability proof remain true for Ω a more general domain and for $-\Delta$ replaced by a symmetric uniformly strongly elliptic operator $L(x)$. The ability to explicitly determine the $Y(T)$ depends on knowing the eigenfunctions and their zeros for L on Ω .

REFERENCES

- [1] G. I. BARENBLATT, I. U. P. ZHELTOV AND I. N. KOCHIVA, *Basic concepts in the theory of seepage of homogeneous liquids in fissured rocks [strata]*, J. Appl. Math. and Mech., 24 (1960), pp. 852–864.
- [2] R. W. CARROLL AND R. E. SHOWALTER, *Singular and Degenerate Cauchy Problems*, Academic Press, New York, 1976.
- [3] P. CHEN AND M. GURTIN, *On a theory of heat conduction involving two temperatures*, Z. Angew. Math. Phys., 19 (1968), pp. 614–627.
- [4] H. O. FATTORINI, *On complete controllability of linear systems*, J. Differential Equations, 3 (1967), pp. 391–402.
- [5] ———, *Boundary control systems*, this Journal, 6 (1968), pp. 349–385.
- [6] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, S. K. Mitter, trans., Springer-Verlag, New York, 1971.
- [7] ———, *Function Spaces and Optimal Control of Distributed Systems*, College de France, 1977.
- [8] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.
- [9] D. TAYLOR, *Research on Consolidation of Clays*, MIT Press, Cambridge, MA, 1952.
- [10] T. W. TING, *Certain non-steady flows of second-order fluids*, Arch. Rat. Mech. Anal., 14 (1963), pp. 1–26.
- [11] L. W. WHITE, *Control problems governed by a pseudoparabolic partial differential equation*, Trans. Amer. Math. Soc., 250 (1979), pp. 235–246.

ON SOME PROPERTIES OF CONDITIONAL MOMENTS IN NONLINEAR FILTERING*

JACOB HAMMER†

Abstract. The nonlinear filtering problem of diffusion processes embedded in additive white noise is considered. It is shown that the paths of all conditional moments of the measurement function can be causally calculated when the path of its first conditional moment is known. The formulas involved in this calculation are independent of the specific parameters of the information process.

In addition, asymptotic properties of the nonlinear filter, as the signal to noise ratio approaches infinity, are also considered. It is shown that, asymptotically, the deviation from Gaussian properties is of the order of the noise to signal ratio at most.

1. Introduction. The optimal filtering problem of diffusion processes embedded in additive white noise is stated as follows. Let (Ω, \mathcal{F}, P) be a probability space on the set Ω , with complete σ -field \mathcal{F} , and probability measure P . Further, let R denote the real numbers, and let $[0, T]$ be the set of all points $t \in R$ satisfying $0 \leq t \leq T$. We assume that there exists a Brownian motion $B: (\Omega, \mathcal{F}, P) \times [0, T] \rightarrow R$. Next, let $x: (\Omega, \mathcal{F}, P) \times [0, T] \rightarrow R$ be a diffusion process, given by the following stochastic differential equation:

$$(1.1) \quad \begin{aligned} dx_t &= m(x_t, t) dt + \sigma(x_t, t) dB_t, & t \in [0, T], \\ x_{t=0} &= x_0, \end{aligned}$$

where $m, \sigma: R \times [0, T] \rightarrow R$. We assume that the functions m, σ satisfy, for every $(x, t) \in R \times [0, T]$, the uniform Lipschitz (or linear growth) condition

$$(1.2) \quad m^2(x, t) + \sigma^2(x, t) \leq K(1 + x^2)$$

for a suitable constant $K \geq 0$. We also assume that the initial condition x_0 is stochastically independent of B_t for every $t \in [0, T]$ and satisfies, for all integers $k \geq 0$, $E\{x_0^{2k}\} < \infty$ (where $E\{\cdot\}$ denotes the expectation). Under these conditions (see Gikhman and Skorokhod [1972]), the solution x_t of (1.1) is almost surely unique, almost surely continuous in t , and, for every integer $k \geq 0$ and every $t \in [0, T]$, we have that $E\{x_t^{2k}\} < \infty$. We shall refer to x as the *information process*.

Next, let W be a standard Brownian motion on $(\Omega, \mathcal{F}, P) \times [0, T]$, and assume that, for all $t \in [0, T]$, $\{W_t\}$ is stochastically independent of $\{B_t\}$. Also, let $g: R \rightarrow R$ be a twice continuously differentiable function satisfying, for every $t \in [0, T]$, the following conditions: (i) $E\{g^{2k}(x_t)\} < \infty$ for every integer $k \geq 0$; (ii) $E\{[g'(x_t)]^2\} < \infty$, and (iii) $E\{[g''(x_t)]^2\} < \infty$, where g' (resp. g'') denotes the first (resp. second) derivative of g . Finally, let λ be a positive real number. Then, the *measurement process* y is defined as

$$(1.3) \quad \begin{aligned} dy_t &= \lambda g(x_t) dt + dW_t, & t \in [0, T], \\ y_{t=0} &= y_0, \end{aligned}$$

* Received by the editors March 30, 1981, and in revised form August 5, 1981.

† Center for Mathematical System Theory, University of Florida, Gainesville, Florida 32611. This work was done while the author was with the Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa, Israel. The present revision was supported in part by U.S. Army research grant DAAG29-80 C0050 and US Air Force grant AFOSR76-3034D through the Center for Mathematical System Theory, University of Florida, Gainesville.

where the initial condition y_0 is stochastically independent of both B_t and W_t for all $t \in [0, T]$. The function g will be called the *measurement function*. If g is the identity function (that is, $g(x) = x$), then we say that (1.3) is a *linear measurement process*.

Having defined the information process x and the measurement process y , we can now state the classical nonlinear filtering problem. Let $f: R \rightarrow R$ be a twice continuously differentiable function, and assume that, for every $t \in [0, T]$, we have that $E\{[f(x_t)]^2\} < \infty$, $E\{[f'(x_t)]^2\} < \infty$ and $E\{[f''(x_t)]^2\} < \infty$. Also, let $y_0^t := \{y_\theta \mid \theta \in [0, t]\}$ denote a sample of the measurement process during the time interval $[0, t]$. Then, given y_0^t , an estimate $\hat{f}(x_t)$ for $f(x_t)$ is sought such that, for any other estimate $\tilde{f}(x_t)$, the following holds: $E\{[f(x_t) - \hat{f}(x_t)]^2\} \leq E\{[f(x_t) - \tilde{f}(x_t)]^2\}$.

The solution to the nonlinear filtering problem is well known, and is given by the conditional expectation

$$\hat{f}(x_t) = E\{f(x_t) \mid y_0^t\},$$

conditioned on the σ -field generated by y_0^t . This solution can also be represented in the following form (see Fujisaki, Kallianpur and Kunita [1972]):

$$\begin{aligned} \hat{f}(x_t) - \hat{f}(x_0) = & \int_0^t [\widehat{m(x_u, u)f'(x_u)} + \frac{1}{2}\sigma^2(x_u, u)f''(x_u)] du \\ (1.4) \quad & + \lambda \int_0^t [\widehat{f(x_u)g(x_u)} - \hat{f}(x_u)\hat{g}(x_u)] dv_u, \end{aligned}$$

where ν is the *innovation process* and is given by

$$(1.5) \quad dv_t = dy_t - \lambda \hat{g}(x_t) dt.$$

It is also known (see op. cit.) that the *innovation process* ν is a *standard Brownian motion*.

Various aspects of the nonlinear filtering problem are considered in Stratonovich [1960], Kushner [1967], Zakai [1969] and [1975], Kailath [1969], Jazwinski [1970], Fujisaki, Kallianpur and Kunita [1972] and others. In the present paper we examine the nonlinear filtering problem using a linear derivative-type operator, which we call the *martingale derivative*. We summarize below the main results.

Let $\hat{g}_t^k := E\{g^k(x_t) \mid y_0^t\}$, where k is a positive integer, be the k th conditional moment of the measurement function. Also, assume that the path \hat{g}_0^1 of a specific sample of the first conditional moment is known. Then, we show that the paths of all other conditional moments $(\hat{g}_0^k)^t$, $k = 2, 3, \dots$, (related to the same sample y_0^t) can be causally calculated using the path \hat{g}_0^1 only. Moreover, the formulas relating $(\hat{g}_0^k)^t$ and \hat{g}_0^1 are independent of the functions $m(x, t)$ and $\sigma(x, t)$ determining the information process x in (1.1). Thus, this calculation can be performed even in cases where a detailed description of the information process is not available.

As an additional application of the martingale derivative, we consider the asymptotic behavior of the nonlinear filter as the constant λ in (1.3) approaches infinity. Informally, this is equivalent to the consideration of filters under conditions of high signal to noise ratio. We show that, as $\lambda \rightarrow \infty$, the conditional probability measure of $g(x_t)$, conditioned on the σ -field generated by y_0^t , approaches a Gaussian measure at the "rate" of $1/\lambda$ at least.

The paper is organized as follows. In § 2 we define the martingale derivative, and in § 3 we apply it to obtain a representation of all conditional moments in terms of martingale derivatives of the first one. The paper is concluded in § 4 with a

consideration of some asymptotic properties as the signal to noise ratio approaches infinity.

2. The martingale derivative. The martingale derivative is a linear operator defined on a certain class of functions of unbounded variation. Its properties are similar to the properties of the usual derivative. First we establish our notation. Let (Ω, \mathcal{F}, P) be a complete probability space. Also, let $\{\mathcal{F}_t\}$, $t \in [0, T]$, be an increasing family of complete σ -fields, which is continuous from the right and satisfies $\mathcal{F}_T \subset \mathcal{F}$. We denote by W a standard Brownian motion on $(\Omega, \mathcal{F}, P; \mathcal{F}_t)$, $0 \leq t \leq T$. Further, let f_t and h_t be square integrable semimartingales of the form

$$(2.1) \quad f_t = a_t + \int_0^t \phi_s dW_s, \quad h_t = b_t + \int_0^t \psi_s dW_s,$$

where ϕ_t and ψ_t are well measurable with respect to the family $\{\mathcal{F}_t\}$, $\int_0^T E\{\phi_s^2\} ds < \infty$ and $\int_0^T E\{\psi_s^2\} ds < \infty$ and, finally, a_t and b_t are adapted to \mathcal{F}_t and differentiable on $[0, T]$. Denote by $[f, h]_t := \int_0^t \phi_s \psi_s ds$ the cross quadratic variation of f and h (e.g., Meyer [1975, Ch. 3]). We next state a formal definition of the martingale derivative, and, immediately afterwards, we give an interpretation of this definition in the particular case which is of main interest to us. (We note that $[h, h]_t$ is almost surely strictly increasing in t if and only if $\psi_t \neq 0$ almost surely for all $t \in [0, T]$.)

DEFINITION 2.2. Let f_t and h_t be as in (2.1), and assume that the quadratic variation $[h, h]_t$ is almost surely strictly increasing in t . Then, the *martingale derivative* f_t^h of f_t with respect to h_t is

$$f_t^h := \lim_{\substack{\Delta \rightarrow 0 \\ \Delta > 0}} \frac{[f, h]_t - [f, h]_{t-\Delta}}{[h, h]_t - [h, h]_{t-\Delta}},$$

on the optional σ -field on $\Omega \times [0, T]$.

Assume now in (2.1) that, for all $t \in [0, T]$, the processes ϕ_t and ψ_t are almost surely continuous and $\psi_t \neq 0$ almost surely. Then, we have that

$$f_t^h = \frac{\phi_t}{\psi_t}, \quad \text{and} \quad f_t^W = \phi_t.$$

In cases where it is possible to iterate the martingale derivative, we shall adopt the following notation:

$$(2.3) \quad f_t^{\{(k+1)h\}} := (f_t^{\{k\}h})^h, \quad f_t^{\{0\}h} := f_t,$$

where $k \geq 0$ is an integer.

Next, we list a series of simple properties of the martingale derivative, showing that it obeys the usual differentiation rules. To this end, we let f_1, \dots, f_n, h, l be a set of semimartingales of the form (2.1), and assume that both of the quadratic variations $[h, h]_t$ and $[l, l]_t$ are almost surely strictly increasing in t . Further, we let c_t be a stochastic process adapted to $\{\mathcal{F}_t\}$ and almost surely differentiable with respect to t , for every $t \in [0, T]$. Finally, we let $H: R^n \times [0, T] \rightarrow R$ be a function twice continuously differentiable in its first n arguments, and differentiable in the last one. Then, using the Ito differential formula (e.g., Gikhman and Skorokhod [1972]), the following (2.4) to (2.8) can readily be verified:

$$(2.4) \quad (f_1 + cf_2)_t^h = f_{1,t}^h + c_t f_{2,t}^h,$$

$$(2.5) \quad (f_1 f_2)_t^h = f_{1,t}^h f_{2,t} + f_{1,t} f_{2,t}^h,$$

$$(2.6) \quad \text{if } f_{2,t} \neq 0 \text{ almost surely for every } t \in [0, T], \text{ then } \left(\frac{f_1}{f_2}\right)_t^h = \frac{f_{1,t}^h f_{2,t} - f_{1,t} f_{2,t}^h}{(f_{2,t}^2)^h},$$

$$(2.7) \quad f_{1,t}^l = f_{1,t}^h h_t^l,$$

$$(2.8) \quad [H(f_{1,t}, f_{2,t}, \dots, f_{n,t}, t)]^h = \sum_{i=1}^n \left[\frac{\partial H(f_{1,t}, \dots, f_{n,t}, t)}{\partial f_{i,t}} \right] f_{i,t}^h.$$

Thus, the usual rules of calculus apply to the martingale derivative.

3. A representation result for conditional moments. We consider the nonlinear filtering problem with information process (1.1) and measurement process (1.3). In this section we let $\lambda = 1$ in (1.3). As before, we denote by $g(x)$ the measurement function and, for every integer $k \geq 0$, $\widehat{g}_t^k := E\{g^k(x_t) | y_0^t\}$ is its k th conditional moment. We examine first the relation between the conditional moments of $g(x_t)$ and the martingale derivatives, with respect to the innovation process ν , of \widehat{g}_t . It turns out that \widehat{g}_t^k is a polynomial function of the martingale derivatives $\widehat{g}_t, \widehat{g}_t^\nu, \dots, \widehat{g}_t^{\{k-1\}\nu}$. Moreover, the k th martingale derivative $\widehat{g}_t^{\{k\}\nu}$ can also be expressed as a polynomial function of the conditional moments $\widehat{g}_t, \dots, \widehat{g}_t^{k+1}$. Thus, we encounter the interesting situation where an infinite set of polynomial equations has a polynomial solution.

We next define two families of multivariable polynomials, $P_k(x_1, \dots, x_k)$ and $G_k(x_1, \dots, x_{k+1})$, where $k = 0, 1, \dots$, by the following recursive formulas:

$$(3.1) \quad \begin{aligned} P_0 &= 1, \\ P_{k+1}(x_1, \dots, x_{k+1}) &= x_1 P_k(x_1, \dots, x_k) + \sum_{i=1}^k \left[\frac{\partial P_k(x_1, \dots, x_k)}{\partial x_i} \right] x_{i+1}, \end{aligned}$$

and

$$(3.2) \quad \begin{aligned} G_0(x_1) &= x_1, \\ G_{k+1}(x_1, \dots, x_{k+2}) &= \sum_{i=1}^{k+1} \left[\frac{\partial G_k(x_1, \dots, x_{k+1})}{\partial x_i} \right] (x_{i+1} - x_i x_1). \end{aligned}$$

Now, we can state the following:

THEOREM 3.3. *Given a nonlinear filtering problem with information process (1.1) and measurement process (1.3) (where $\lambda = 1$), the following hold:*

- (i) $\widehat{g}_t^k = P_k(\widehat{g}_t, \dots, \widehat{g}_t^{\{k-1\}\nu})$;
- (ii) $\widehat{g}_t^{\{k\}\nu} = G_k(\widehat{g}_t, \dots, \widehat{g}_t^{k+1})$.

Proof. We remark that, since $\lambda = 1$, it follows by (1.4) that the martingale derivative of the k th conditional moment is $\widehat{g}_t^{k\nu} = \widehat{g}_t^{k+1} - \widehat{g}_t^k \widehat{g}_t$. We use this formula in the proof of (i) and (ii).

(i) Evidently, $\widehat{g}_t^0 = P_0 = 1$, and we assume, by induction, that $\widehat{g}_t^n = P_n(\widehat{g}_t, \dots, \widehat{g}_t^{\{n-1\}\nu})$ for an integer $n \geq 0$. Then using (2.8), we obtain that

$$\begin{aligned} \widehat{g}_t^{n+1} &= \widehat{g}_t \widehat{g}_t^n + \widehat{g}_t^{n\nu} = \widehat{g}_t P_n(\widehat{g}_t, \dots, \widehat{g}_t^{\{n-1\}\nu}) + \sum_{i=0}^{n-1} \left[\frac{\partial P_n}{\partial \widehat{g}_t^{\{i\}\nu}} \right] (\widehat{g}_t^{\{i\}\nu})^\nu \\ &= \widehat{g}_t P_n(\widehat{g}_t, \dots, \widehat{g}_t^{\{n-1\}\nu}) + \sum_{i=0}^{n-1} \left[\frac{\partial P_n}{\partial \widehat{g}_t^{\{i+1\}\nu}} \right] \widehat{g}_t^{\{i+1\}\nu}. \end{aligned}$$

Hence, $\widehat{g}_t^{n+1} = P_{n+1}(\widehat{g}_t, \dots, \widehat{g}_t^{\{n\}\nu})$, and (i) follows.

(ii) Again, identically, $\hat{g}_t^{(0)\nu} = \hat{g}_t = G_0(\hat{g}_t)$ and, by induction, we assume that (ii) holds for an integer $n \geq 0$. Then, still using (2.8), we obtain

$$\begin{aligned} \hat{g}_t^{(n+1)\nu} &= [G_n(\hat{g}_t, \dots, \widehat{g_t^{n+1}})]^\nu = \sum_{i=1}^{n+1} \left[\frac{\partial G_n}{\partial \widehat{g_t^i}} \right] (\widehat{g_t^i})^\nu \\ &= \sum_{i=1}^{n+1} \left[\frac{\partial G_n}{\partial \widehat{g_t^i}} \right] (\widehat{g_t^{i+1}} - \widehat{g_t^i} \hat{g}_t) = G_{n+1}(\hat{g}_t, \dots, \widehat{g_t^{n+2}}). \end{aligned}$$

Hence, (ii) holds for $n+1$, and our proof is concluded. \square

Let f_t be a semimartingale of the form (2.1). In general, the calculation of the martingale derivative f_t^W involves the use of both a sample of f_t and the corresponding sample of W_t . However, in case of the nonlinear filtering problem, the situation turns out to be different. We next show that, for every $t \in [0, T]$, the martingale derivative \hat{g}_t^ν of \hat{g}_t is completely determined by the sample \hat{g}_0^ν . No explicit information on the sample ν_0^ν is required. In fact, more is true.

THEOREM 3.4. *Let $g(x)$ be the measurement function, and ν the innovation process for the filtering problem described by (1.1) and (1.3). Then, for every integer $k \geq 0$ and for every $t \in [0, T]$, the k th martingale derivative $\hat{g}_t^{(k)\nu}$ is determined by the sample $\hat{g}_{t-\alpha}^\nu$, where $0 < \alpha < t$.*

Proof. First we note that Theorem 3.3(ii) implies that all iterated martingale derivatives $\hat{g}_t^{(k)\nu}$, $k = 0, 1, \dots$, are almost surely continuous for all $t \in [0, T]$. For the sake of simplicity, we shall consider below only *continuous samples*. The null sets on which continuity does not hold can be dealt with by a standard “countable diagonal set” method.

By the definition of the martingale derivative, we have that $[\hat{g}^{(n)\nu}, \hat{g}]_t - [\hat{g}^{(n)\nu}, \hat{g}]_{t-\alpha} = \int_{t-\alpha}^t \hat{g}_s^{(n+1)\nu} \hat{g}_s^\nu ds$. Hence it follows by continuity that, for every $s \in [t-\alpha, t]$, the quantity $h_s^n := \hat{g}_s^{(n+1)\nu} \hat{g}_s^\nu$ is determined by the paths $(\hat{g}^{(n)\nu})_{t-\alpha}^t$ and $\hat{g}_{t-\alpha}^\nu$. We show next, by recursion, that this fact implies that all martingale derivatives can be calculated as required.

First, in case $n = 0$, we obtain $h_s^0 = (\hat{g}_s^\nu)^2$. Now, by (1.4), $\hat{g}_t^\nu = \lambda(\widehat{g_t^2} - \hat{g}_t^2)$ so that, since $\lambda > 0$, it follows by the Jensen inequality that $\hat{g}_t^\nu \geq 0$ for all $t \in [0, T]$. Consequently, \hat{g}_s^ν is determined by h_s^0 , and thus the path $(\hat{g}^\nu)_{t-\alpha}^t$ is determined by the path $\hat{g}_{t-\alpha}^\nu$.

Further, by recursion, we assume that, for some integer $n \geq 0$, the path $(\hat{g}^{(n)\nu})_{t-\alpha}^t$ is determined by $\hat{g}_{t-\alpha}^\nu$. Then, for every $s \in [t-\alpha, t]$, the quantity h_s^n is clearly determined by $\hat{g}_{t-\alpha}^\nu$. Hence, if $\hat{g}_s^\nu \neq 0$, then the martingale derivative $\hat{g}_s^{(n+1)\nu}$ is determined by $\hat{g}_{t-\alpha}^\nu$. Thus, it remains to consider the case $\hat{g}_s^\nu = 0$, which we next do. By continuity, it follows that there exists an element $\delta > 0$ such that either (i) $\hat{g}_u^\nu = 0$ for all $u \in [s-\delta, s]$; or (ii) $\hat{g}_u^\nu \neq 0$ for all $u \in [s-\delta, s)$. But then, in subcase (i) we have $\hat{g}_u^{(k)\nu} = 0$ for all $u \in [s-\delta, s]$ and $k = 2, 3, \dots$. In subcase (ii), we have already shown that $\hat{g}_u^{(n+1)\nu}$ is determined by $\hat{g}_{t-\alpha}^\nu$ for all $u \in [s-\delta, s)$. Hence it follows, by continuity, that $\hat{g}_s^{(n+1)\nu}$ is determined as well. \square

Combining Theorems 3.3(i) and 3.4, we directly obtain the following:

COROLLARY 3.5. *Let $g(x)$ be the measurement function for the filtering problem of (1.1) with measurement (1.3). Then, for every integer $k \geq 0$ and for every $t \in [0, T]$, the conditional moment \hat{g}_t^k is determined by the path $\hat{g}_{t-\alpha}^\nu$, where $0 < \alpha < t$.*

It is interesting to note that the formulas involved in the calculation of the conditional moment \hat{g}_t^k from the first conditional moment path $\hat{g}_{t-\alpha}^\nu$, as described in Theorems 3.3 and 3.4, are independent of the functions $m(x, t)$ and $\sigma(x, t)$ which determine the information process (1.1). In fact, it can be shown that the same

calculation scheme is valid for information processes more general than (1.1) as well.

Before concluding this section, we note that Corollary 3.5 can be generalized in the following sense. Let $f: R \rightarrow R$ be a continuous function, measurable with respect to the σ -field induced by g on R . Then, for every $t \in [0, T]$, $\hat{f}(x_t)$ is determined by the path $\hat{g}_{t-\alpha}^t$, where $0 < \alpha < t$.

4. Some asymptotic properties. We start with a closer examination of the polynomials $G_n(x_1, \dots, x_{n+1})$ of (3.2). Let z be a random variable on the probability space (Ω, \mathcal{F}, P) . We denote by \bar{z}^i the i th moment of z . We next show that the polynomials G_n are closely related to the Gaussian probability law, as follows.

LEMMA 4.1. *A random variable z has a Gaussian distribution function if and only if, for every integer $n \geq 2$, $G_n(\bar{z}, \dots, \bar{z}^n, \bar{z}^{n+1}) = 0$.*

Proof. Assume first that z is a Gaussian random variable. Then, all moments of z are determined by the first two moments \bar{z} and \bar{z}^2 , so that, for every integer $i \geq 0$, $\bar{z}^i = \bar{z}^i(\bar{z}, \bar{z}^2)$. We let $\sigma^2 := (z - \bar{z})^2 \neq 0$, and consider the Gaussian random variable z_ε defined as follows: $\bar{z}_\varepsilon = \bar{z} + \varepsilon$ and $(z_\varepsilon - \bar{z}_\varepsilon)^2 = \sigma^2$. Then, we have

$$\left. \frac{d\bar{z}_\varepsilon^i}{d\varepsilon} \right|_{\varepsilon=0} = (2\pi\sigma^2)^{-1/2} \left. \frac{d}{d\varepsilon} \left\{ \int_{-\infty}^{\infty} u^i \exp \left[\frac{(u - \bar{z} - \varepsilon)^2}{2\sigma^2} \right] du \right\} \right|_{\varepsilon=0} = \frac{(\bar{z}^{i+1} - \bar{z}^i \bar{z})}{\sigma^2}.$$

Now, since the Gaussian distribution is symmetric, and since $G_2(\bar{z}, \bar{z}^2, \bar{z}^3) = (z - \bar{z})^3$, we clearly have that $G_2(\bar{z}, \bar{z}^2, \bar{z}^3) = 0$. By induction, we assume now that there is an integer $n \geq 2$ such that, for every Gaussian random variable z_1 , $G_n(\bar{z}_1, \dots, \bar{z}_1^{n+1}) = 0$. In particular, it follows then that, in the case $z_1 = z_\varepsilon$, $G_n(\bar{z}_\varepsilon, \dots, \bar{z}_\varepsilon^{n+1}) = 0$ for every ε . Hence, $dG_n(\bar{z}_\varepsilon, \dots, \bar{z}_\varepsilon^{n+1})/d\varepsilon = 0$ for every ε as well. Now, by a direct calculation, we have $0 = dG_n(\bar{z}_\varepsilon, \dots, \bar{z}_\varepsilon^{n+1})/d\varepsilon|_{\varepsilon=0} = (1/\sigma^2)[G_{n+1}(\bar{z}, \dots, \bar{z}^{n+2})]$, which implies the necessity of our assertion.

Conversely, if $G_n(\bar{z}, \dots, \bar{z}^{n+1}) = 0$ for every $n = 2, 3, \dots$, then, since G_n is monic in \bar{z}^{n+1} , it follows that all moments \bar{z}^i of z are determined by the first two moments \bar{z} and \bar{z}^2 . Moreover, by our previous discussion it is clear that the functions $\bar{z}^i = \bar{z}^i(\bar{z}, \bar{z}^2)$ thus obtained are identical to those for the Gaussian case. Hence, z has a Gaussian characteristic function, and our proof concludes. \square

Motivated by Lemma 4.1, we shall call the polynomials G_n of (3.2) the *Gaussian polynomials*.

Example. The first Gaussian polynomials are as follows:

$$G_0 = \bar{z}, \quad G_1 = \overline{(z - \bar{z})^2}, \quad G_2 = \overline{(z - \bar{z})^3}, \quad G_3 = \overline{(z - \bar{z})^4} - 3\left[\overline{(z - \bar{z})^2}\right]^2.$$

We return now to the nonlinear filtering problem of the information process (1.1) with the measurement process (1.3). Let $\hat{P}_t: R \rightarrow [0, 1]$ denote the conditional probability distribution of $g(x_t)$, conditioned on the σ -field generated by the measurement process y_t^0 . Then, we have $\hat{g}_t^n = \int u^n d\hat{P}_t(u)$. As usual, we shall say that \hat{P}_t is *symmetric* if it satisfies the following. For every function $f: R \rightarrow R$ that satisfies $f(x - \hat{g}_t) = -f(\hat{g}_t - x)$ for all $x \in R$, one has that $\int f(u) d\hat{P}_t(u) = 0$. As a direct consequence of Theorem 3.3 and Lemma 4.1, we can now show that if \hat{P}_t is symmetric, then it is necessarily Gaussian. This is proved in the following:

PROPOSITION 4.2. *The conditional probability measure \hat{P}_t is almost surely symmetric for all $t \in [0, T]$ if and only if it is almost surely Gaussian for all $t \in [0, T]$.*

Proof. Assume first that \hat{P}_t is symmetric. Then evidently, $(\hat{g}_t - \hat{g}_t)^3 = 0$ almost surely, so that $G_2(\hat{g}_t, \hat{g}_t^2, \hat{g}_t^3) = 0$ almost surely for all $t \in [0, T]$. But then, since by

Theorem 3.3, $\hat{g}_t^{\{2\}\nu} = G_2(\hat{g}_t, \hat{g}_t^2, \hat{g}_t^3)$, we have that $\hat{g}_t^{\{2\}\nu} = 0$ almost surely for all $t \in [0, T]$. Hence, also $\hat{g}_t^{\{n\}\nu} = 0$ almost surely for all integers $n \geq 2$ and all $t \in [0, T]$. Again, by Theorem 3.3, this implies that $G_n(\hat{g}_t, \dots, \hat{g}_t^{n+1}) = 0$ for all $n \geq 2$ and all $t \in [0, T]$, so that, by Lemma 4.1, \hat{P}_t is Gaussian. The converse direction is immediate. \square

We consider next the asymptotic behavior of the conditional probability measure \hat{P}_t as the constant λ in (1.3) approaches infinity. Explicitly, we shall show that, as $\lambda \rightarrow \infty$, the probability law determined by \hat{P}_t approaches the Gaussian probability law at the rate of $1/\lambda$ at least. To this end, we need the following notation. Let $f: [0, T] \rightarrow \mathbb{R}$ be a function, which implicitly depends on λ . We shall say that $f \sim 1/\lambda$ if, for almost every $t \in [0, T]$, the following holds: For every $\alpha > 0$, $\lim_{\lambda \rightarrow \infty} \lambda^{1-\alpha} f(t) = 0$.

As before, we let $g(x)$ be the measurement function, and denote $\hat{g}_t^i = E\{g^i(x_t) | y_0^t\}$. Also, G_n , $n = 0, 1, 2, \dots$, are the Gaussian polynomials defined in (3.2). Clearly, as $\lambda \rightarrow \infty$, the conditional probability measure \hat{P}_t degenerates into a deterministic measure. Thus, we expect by Lemma 4.1 that, for $n \geq 1$, one should have $\lim_{\lambda \rightarrow \infty} E[G_n(\hat{g}_t, \dots, \hat{g}_t^{n+1})] = 0$. In fact, the following stronger result is valid.

THEOREM 4.3. *Given the nonlinear filtering problem of the information process (1.1) with measurement process (1.3), the following holds true: For every $0 \leq \delta < 1$,*

$$E^{1/(1+\delta)} |G_n(\hat{g}_t, \dots, \hat{g}_t^{n+1})|^{(1+\delta)} \sim \frac{1}{\lambda},$$

where $n = 1, 2, 3, \dots$, and $t \in [0, T]$.

Proof. We first note that, if the condition $\lambda = 1$ in Theorem 3.3 is relaxed, then, for all $k = 0, 1, \dots$ and $t \in [0, T]$, we have

$$\hat{g}_t^{\{k+1\}\nu} = \lambda^{k+1} G_{k+1}(\hat{g}_t, \dots, \hat{g}_t^{k+2}) = \lambda^k \sum_{i=1}^{k+1} \left(\frac{\partial G_k}{\partial \hat{g}_t^i} \right) \hat{g}_t^{i\nu},$$

where the last equality follows by (3.2) and (1.4). Then, applying the Minkowski inequality (e.g., Loève [1963, Ch. 3]), we obtain that, for every $0 \leq \delta < 1$,

$$\begin{aligned} f_t &:= E^{1/(1+\delta)} |\lambda G_{k+1}(\hat{g}_t, \dots, \hat{g}_t^{k+2})|^{(1+\delta)} = E^{1/(1+\delta)} |\lambda^{-k} \hat{g}_t^{\{k+1\}\nu}|^{(1+\delta)} \\ &\leq \sum_{i=1}^{k+1} E^{1/(1+\delta)} \left[\left| \frac{\partial G_k}{\partial \hat{g}_t^i} \right|^{(1+\delta)} |\hat{g}_t^{i\nu}|^{(1+\delta)} \right]. \end{aligned}$$

Applying now the Hölder inequality (e.g., Loève [1963, Ch. 3]), with exponents $2/(1-\delta)$ and $2/(1+\delta)$, to each summand in the above sum, we obtain

$$(\alpha) \quad f_t \leq \sum_{i=1}^{k+1} E^{(1-\delta)/2(1+\delta)} \left| \frac{\partial G_k}{\partial \hat{g}_t^i} \right|^{2(1+\delta)/(1-\delta)} E^{1/2} \{\hat{g}_t^{i\nu}\}^2.$$

Now, $\partial G_k / \partial \hat{g}_t^i$ is a polynomial in $\hat{g}_t, \dots, \hat{g}_t^{k+1}$ and, by our assumptions, $E\{g(x_t)\}^{2n} < \infty$ for all integers $n \geq 0$ and for every $t \in [0, T]$. Also, all relevant quantities are almost surely continuous functions of t on the compact interval $[0, T]$ (and we also have $1-\delta > 0$). It follows then that there exists a constant $M' \geq 0$ such that, for every $i = 1, \dots, k+1$ and for all $t \in [0, T]$,

$$(\beta) \quad E^{(1-\delta)/2(1+\delta)} \left| \frac{\partial G_k}{\partial \hat{g}_t^i} \right|^{2(1+\delta)/(1-\delta)} \leq M'.$$

Further, by (1.4), we have

$$\int_0^t \widehat{g_s^i} d\nu_s = \widehat{g_t^i} - \int_0^t \left[\widehat{im(x_s, s)g^{i-1}(x_s)g'(x_s)} + \left(\frac{i(i-1)}{2} \right) \widehat{\sigma^2(x_s, s)g^{i-2}(x_s)g'(x_s)} + \left(\frac{i}{2} \right) \widehat{\sigma^2(x_s, s)g^{i-1}(x_s)g''(x_s)} \right] ds$$

for all integers $i \geq 1$. We now square both sides of the last equation, and consider the expectation of the resulting quantities. On the left-hand side we obtain $\int_0^t E\{\widehat{g_s^i}^2\} ds$. Also, by an application of the Hölder inequality and in view of our assumptions on (1.1) and (1.3), it follows that the expectation of the squared right-hand side is bounded by a constant $M'' \geq 0$. Thus, we have, for all $i = 1, \dots, k+1$, that $\int_0^T E\{\widehat{g_s^i}^2\} ds \leq M''$.

But then, for every $\alpha > 0$, $\lim_{\lambda \rightarrow \infty} \int_0^T \lambda^{-\alpha} E\{\widehat{g_s^i}^2\} ds = 0$, so that

$$(\gamma) \quad \lim_{\lambda \rightarrow \infty} \lambda^{-\alpha} E\{\widehat{g_t^i}^2\} = 0$$

for almost all $t \in [0, T]$. Finally, substituting (β) and (γ) into (α) , it follows that, for every $\alpha > 0$,

$$\lim_{\lambda \rightarrow \infty} E^{1/(1+\delta)} |\lambda^{1-\alpha} G_{k+1}(\widehat{g}_0, \dots, \widehat{g_t^{k+2}})^{1+\delta}| \leq \lim_{\lambda \rightarrow \infty} \left[\sum_{i=1}^{k+1} M'(\lambda^{-\alpha} E^{1/2}\{\widehat{g_t^i}^2\}) \right] = 0$$

for almost every $t \in [0, T]$, proving our assertion. \square

Consider now Theorem 4.3 in the case of linear measurement, that is, when $g(x) = x$. In this case, substituting $n = 1$ and $\delta = 0$, and noting that $G_1 = \widehat{x_t^2} - \widehat{x_t}^2$, we obtain that $E\{(x_t - \widehat{x_t})^2\} \sim 1/\lambda$, which is in accordance with the upper bound of Zakai and Ziv [1972]. Thus, Theorem 4.3 is a generalization of that result.

We conclude this section by showing that the conditional probability measure \hat{P}_t can be replaced by a Gaussian measure, up to an error of the "order" of $1/\lambda$. To this end, we let Π_t , for every $t \in [0, T]$, be the Gaussian measure determined by its first two moments as follows: $\int x d\Pi_t(x) = \widehat{g_t}$ and $\int x^2 d\Pi_t(x) = \widehat{g_t^2}$. Given a function $f: R \rightarrow R$, we shall denote by $\widehat{f_t} := \int f(x) d\Pi_t(x)$ its expectation with respect to Π_t . We now have the following:

THEOREM 4.4. *Given the nonlinear filtering problem of the information process (1.1) with measurement process (1.3), the following holds for every $0 \leq \delta < 1$:*

$$E^{1/(1+\delta)} |\widehat{g_t^i} - \widehat{g_t^i}|^{(1+\delta)} \sim \frac{1}{\lambda},$$

where $i = 0, 1, 2, \dots$, and $t \in [0, T]$.

Proof. The cases $i = 0, 1$, and 2 are clearly implied by the construction of the probability measure Π_t . The proof proceeds by induction. Assume that the theorem holds for all integers $i = 1, \dots, n$. Now by Lemma 4.1, $G_n(\widehat{g}_0, \dots, \widehat{g_t^{n+1}}) = 0$ for all integers $n \geq 2$, so that, since G_n is monic in $\widehat{g_t^{n+1}}$, we have that $\widehat{g_t^{n+1}} - \widehat{g_t^{n+1}} = G_n(\widehat{g}_0, \widehat{g_t^2}, \dots, \widehat{g_t^n}, \widehat{g_t^{n+1}})$ for all integers $n \geq 2$. The following calculation is intended to replace the arguments $\widehat{g_t^n}, \dots, \widehat{g_t}$ in the last expression by $\widehat{g_t^n}, \dots, \widehat{g_t}$, and to compute the error caused by this manipulation. To this end, we represent $G_n(\widehat{g}_0, \dots, \widehat{g_t^n}, \widehat{g_t^{n+1}}) = A_t + \widehat{g_t^{n+1}} B_t$, where A_t and B_t are suitable polynomials in $\widehat{g_t^{n+1}}, \widehat{g_t^n}, \dots, \widehat{g_t}$. By our assumptions on (1.1) and (1.3), it follows that A_t and B_t have all their moments bounded.

Now, let δ' be such that $\delta < \delta' < 1$. Then, using the Minkowski and Hölder inequalities, we obtain:

$$\begin{aligned} E^{1/(1+\delta)} |\widehat{g}_t^{n+1} - \widehat{g}_t^{n+1}|^{(1+\delta)} &= E^{1/(1+\delta)} |A_t + \widehat{g}_t^n B_t + (\widehat{g}_t^n - \widehat{g}_t^n) B_t|^{(1+\delta)} \\ &\leq E^{1/(1+\delta)} |A_t + \widehat{g}_t^n B_t|^{(1+\delta)} + E^{1/(1+\delta')} |\widehat{g}_t^n - \widehat{g}_t^n|^{(1+\delta')} E^{1/\gamma} |B_t|^\gamma, \end{aligned}$$

where $\gamma := (1+\delta)(1+\delta')/(\delta'-\delta)$. Applying now the induction assumption, and the fact that all moments of B_t are bounded, it follows that

$$E^{1/(1+\delta)} |\widehat{g}_t^{n+1} - \widehat{g}_t^{n+1}|^{(1+\delta)} \leq E^{1/(1+\delta)} |A_t + \widehat{g}_t^n B_t|^{(1+\delta)} + f_t,$$

where $f_t \sim 1/\lambda$.

By a similar procedure, we replace, for all $i = 1, \dots, n$, all appearances of \widehat{g}_t^i by \widehat{g}_t^i , retaining the corresponding errors f_t . Thus, after a finite number of steps, we obtain

$$E^{1/(1+\delta)} |\widehat{g}_t^{n+1} - \widehat{g}_t^{n+1}|^{(1+\delta)} \leq E^{1/(1+\delta)} |G_n(\widehat{g}_t^{n+1}, \widehat{g}_t^n, \dots, \widehat{g}_t)|^{(1+\delta)} + h_t,$$

where $h_t \sim 1/\lambda$. But then, it follows by Theorem 4.3, that $E^{1/(1+\delta)} |\widehat{g}_t^{n+1} - \widehat{g}_t^{n+1}|^{(1+\delta)} \sim 1/\lambda$, concluding our proof. \square

Finally, we note that Theorem 4.4 can be directly extended to the case of polynomials in $g(x)$ and, also, to functions which are limits, in a suitable sense, of such polynomials.

Acknowledgment. The use of the quadratic variation in the proofs of the present paper was suggested by Moshe Zakai, Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa, Israel. The author is most grateful to him for this suggestion, and for many valuable and stimulating discussions on mentioned and related topics.

REFERENCES

- M. FUJISAKI, G. KALLIANPUR AND H. KUNITA [1972], *Stochastic differential equations for the nonlinear filtering problem*, Osaka J. Math., 9, pp. 19–40.
- I. I. GIKHMAN AND A. V. SKOROKHOD [1972], *Stochastic Differential Equations*, Springer-Verlag, Berlin.
- A. H. JAZWINSKI [1970], *Stochastic Processes and Filtering Theory*, Academic Press, New York.
- T. KAILATH [1969], *A general likelihood formula for random signals in Gaussian noise*, Trans. IEEE, IT-15, pp. 350–361.
- H. J. KUSHNER [1967], *Dynamical equations for optimal nonlinear filtering*, J. Differential Equations, 2, pp. 179–190.
- M. LOËVE [1963], *Probability Theory*, Van Nostrand, Princeton NJ.
- P. A. MEYER [1966], *Probabilité et potentiel*, Hermann, Paris.
- , [1975], *Un cours sur les intégrales stochastiques*, Lecture notes in Mathematics, Springer-Verlag, Berlin.
- R. I. STRATONOVICH [1960], *Conditional Markov processes*, Theory Prob. Appl., 5, pp. 156–178.
- M. ZAKAI [1969], *On the optimal filtering of diffusion processes*, Z. Wahrsch. Verw. Geb., 11, pp. 230–243.
- , [1975], *Lecture notes on nonlinear filtering*, Technion-Israel Institute of Technology, Haifa, Israel. (In Hebrew).
- M. ZAKAI AND J. ZIV [1972], *Lower and upper bounds on the optimal filtering error of certain diffusion processes*, Trans. IEEE, IT-18, pp. 325–331.

APPROXIMATION SCHEMES FOR THE LINEAR-QUADRATIC OPTIMAL CONTROL PROBLEM ASSOCIATED WITH DELAY EQUATIONS*

KARL KUNISCH†

Abstract. The linear regulator problem for delay equations is discussed. We propose a (theoretical) solution involving Riccati integral equations and then axiomatically discuss a general approximation scheme. The details are given for spline and averaging approximations.

1. Introduction and notation. The problem of approximating delay-differential equations by sequences of either ordinary differential equations or algebraic equations has stimulated research for over fifteen years now. However, it was not until quite recently that convergence proofs in an operator-theoretic framework were given; see [2] and the references given there.

In this paper we address a specific problem of the above type, namely the approximation of the regulator problem of minimizing a quadratic cost-functional subject to a delay—or more generally a functional differential equation (FDE). This question has also attracted attention for quite some time. In [15], [17] Ross and Flugge-Lotz and Solimon and Ray specify certain approximation schemes leaving open the question of convergence. In today's terminology their methods would be called averaging projections or linear interpolating spline schemes [2], [10]. Not only does the question of approximation of the linear-quadratic control problem for FDE present difficulties, but the theoretic development of existence of solutions, deriving a feedback law and discussing an operator Riccati equation are challenging as well. Delfour treats these theoretic aspects in [6], [7] and proves convergence of the averaging scheme, discretizing space and time variables. We also cite [7] as a reference on the literature on the linear-quadratic optimal control problem for FDE up to 1977.

In the present paper we develop a general theory for the above-mentioned problem, which we subsequently apply to the spline and averaging approximation schemes. The theoretical aspects are greatly facilitated by a recent paper of Gibson [8] in which an abstract linear-quadratic optimal control problem is treated in a general Hilbert space; it was observed that the Riccati operators satisfy two (almost) equivalent Riccati integral equations, one of which coincides with the one used by Delfour in [7], the other one ((2.17) in this paper), although implicitly present, was not dealt with in [7]. It should be noted that in our presentation the treatment of the original problem (as opposed to the approximating ones) is based solely on integral equations. A second important feature is that we avoid using the infinitesimal generator of the adjoint of the solution semigroup associated with the FDE. All the estimates depend heavily on the fact that even in the abstract formulation of the FDE (see (2.4)), the control term enters only as an operator with finite-dimensional range.

* Received by the editors May 16, 1980, and in revised form June 8, 1981. This research was supported in part by the Air Force Office of Scientific Research under contract AF-AFOSR 76-3092C and in part by the U.S. Army Research Office under grant ARO-DAAG 29-79-C-0161.

† Institut für Mathematik, Technische Universität, Graz, Austria, and Division of Applied Mathematics, Lefschetz Center for Dynamical Systems, Brown University, Providence, Rhode Island 02912.

Many of the technicalities here arise from the fact that we not only prove convergence of optimal controls, trajectories, payoffs, etc., but also give some error bounds. This leads to an essential difficulty which is described at length in Remark 2.1.

The paper is organized in the following way. Section 2 contains the statement of the problem and its (theoretical) solution. Then a sequence of approximating problems is specified and the convergence results are stated, leaving technical proofs to the Appendix. In § 3, we first show how the results of § 2 can be used for spline approximation schemes. For linear and cubic splines we give all the details, demonstrating convergence of the linear spline scheme and quadratic convergence on certain subspaces of the cubic spline scheme. Averaging projection schemes are discussed in § 4; the approximating equations in this case turn out to coincide with those proposed in [4], [15] and [17].

Most of the notation that is used throughout the paper is quite standard. For a closed interval $I \subset (-\infty, \infty)$, a Banach space X with norm $|\cdot|_X$ and $p \geq 1$, the equivalence class of measurable functions $x: I \rightarrow X$ with $\int_I |x(s)|_X^p ds < \infty$ is denoted by $L^p(I; X)$. $|\cdot|_{L^p(I; X)}$ or simply $|\cdot|_{L^p}$ is the notation for the usual norm in $L^p(I; X)$. The space of continuous functions on I with values in X endowed with the supremum norm is denoted by $C(I; X)$ and $C^k(I; X)$, $k = 1, 2, \dots$, stands for the space of X -valued continuous functions which possess k continuous derivatives on I . $W^{k,2}(I; X)$, $k = 1, 2, \dots$, is the space of $(k-1)$ -times continuously differentiable functions whose $(k-1)$ st derivative is absolutely continuous with derivative in $L^2(I; X)$; $|\cdot|_{W^{k,2}(I; X)}$ denotes any one of the commonly employed $W^{k,2}$ -norms. The space of all essentially bounded and strongly measurable functions from I to X is denoted by $\mathcal{B}_\infty(I; X)$. In the special case of $I = [-r, 0]$, $0 < r < \infty$ and $X = \mathbb{R}^n$ we shall abbreviate the notation of the function spaces by L^2 , C^k , $W^{k,2}$, etc.

For Banach spaces X and Y , the set of all bounded linear operators from X to Y is denoted by $\mathcal{L}(X, Y)$ and for $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ we simply write $\mathbb{R}^{n \times m}$. For $A \in \mathcal{L}(X, Y)$ the strong operator norm is denoted by $\|A\|_{\mathcal{L}(X, Y)}$. A^* stands for the Hilbert space adjoint of an operator A from a Hilbert space H to H . \mathbb{R}^n is endowed with the Euclidean norm $|\cdot|_{\mathbb{R}^n}$, and $(\cdot, \cdot)_{\mathbb{R}^n}$ stands for the usual inner product in \mathbb{R}^n . For elements in $\mathbb{R}^{n \times m}$ we use the spectral norm. Wherever the contents permit we drop the subscript of a norm, simply using $|\cdot|$ for the norm of elements of a Banach space and $\|\cdot\|$ for that of operators between Banach spaces.

The state space of our presentation will be $\mathbb{R}^n \times L^2(-r, 0; \mathbb{R}^n)$ with the norm

$$|(\eta, \phi)|_{Z_\rho} = \left(|\eta|_{\mathbb{R}^n} + \int_{-r}^0 \rho(s) |\phi(s)|^2 ds \right)^{1/2},$$

where the weighting function $\rho: [-r, 0] \rightarrow \mathbb{R}$ is a piecewise continuous and positive function. We denote by Z (or Z_ρ where necessary) the space $\mathbb{R}^n \times L^2(-r, 0; \mathbb{R}^n)$ together with the weighted norm. The symbol $\langle \cdot, \cdot \rangle_{Z_\rho}$ stands for the natural inner product in Z and P_1, P_2 denote the projections of Z onto its first and second components respectively. The need for weighting the inner product and norm will not become apparent before § 3 and we shall commonly drop the subscript Z_ρ . \mathcal{C}^k and $\mathcal{W}^{k,2}$ stand for subspaces of Z given by $\{(\phi(0), \phi) | \phi \in C^k\}$ and $\{(\phi(0), \phi) | \phi \in W^{k,2}\}$ respectively.

A family $V(t, s)$ of operators in $\mathcal{L}(Z, Z)$ with $t_0 \leq s \leq t \leq t_f$ is called evolution operator if $V(s, s)z = z$, if $V(t, s)z = V(t, r)V(r, s)z$ and if $t \rightarrow V(t, s)z$ is continuous for all $z \in Z$ and all s, t with $t_0 \leq s \leq r \leq t \leq t_f$. The derivative of a function x is denoted by \dot{x} or also x' , and, finally, for $x: [-r, \alpha] \rightarrow \mathbb{R}^n$, $\alpha > 0$, the symbol x_t , $0 \leq t \leq \alpha$ stands for the function $[-r, 0] \rightarrow X$ given by $x_t(s) = x(t+s)$ for $s \in [-r, 0]$.

2. Approximation of the linear-quadratic control problem. For $(\eta, \phi) = z \in Z$ and $(t_0, t_f) \in \mathbb{R} \times \mathbb{R}$ we consider the functional differential equation (FDE)

$$(2.1) \quad \begin{aligned} \dot{x}(t) &= L(t, x_t) + f(t) \quad \text{for } t_0 \leq s \leq t \leq t_f, \\ x(s) &= \eta, \quad x_s = \phi \end{aligned}$$

where

$$(2.2) \quad L(t, \phi) = \sum_{i=0}^l A_i(t) \phi(-r_i) + \int_{-r}^0 A_{-1}(t, s) \phi(s) ds.$$

Here we let $0 = r_0 < r_1 < \dots < r_l = r$ and the matrix-valued functions A_i , for $i = -1, \dots, l$, are considered as operators in $A_i \in C(t_0, t_f; \mathbb{R}^{n \times n})$, for $i = 0, \dots, l$, and $A_{-1} \in C(t_0, t_f; L^2(-r, 0; \mathbb{R}^{n \times n}))$, respectively, and $f \in L^2(t_0, t_f; \mathbb{R}^n)$.

We also need to restrict our attention to the homogeneous problem

$$(2.3) \quad \begin{aligned} \dot{x}(t) &= L(t, x_t) \quad \text{for } t_0 \leq s \leq t \leq t_f, \\ x(s) &= \eta, \quad x_s = \phi. \end{aligned}$$

It is quite well known [10] that solutions to (2.1) and (2.3) exist and that they do not depend on the representative of an equivalence class $\phi \in L^2$. We shall denote the solutions by $x(\cdot, s; z, f)$ and $x(\cdot, s; z)$, respectively, dropping arguments if the context permits us to do so. Let $T(t, s): Z \rightarrow Z$ be the solution operator associated with (2.3), i.e.,

$$T(t, s)z = (x(t, s; z), x_t(\cdot, s; z)) \quad \text{for } t_0 \leq s \leq t \leq t_f.$$

Then $T(t, s)$ is an evolution operator on $\Delta = \{(t, s) | t_0 \leq s \leq t \leq t_f\}$.

In the next lemma the weighting function for the norm of Z is chosen identically 1.

LEMMA 2.1. *Exponential bounds on $T(t, s)$ are given by*

$$|T(t, s)z|_{Z_1} \leq M e^{\omega(t-s)} |z|_{Z_1} \quad \text{for } (t, s) \in \Delta,$$

where

$$M = \left(1 + \sum_{i=1}^l \sup_{t \in [t_0, t_f]} \|A_i(t)\| \right)^{1/2}$$

and

$$\omega = M^2 + \sup_{t \in [t_0, t_f]} \|A_{-1}(t, \cdot)\|,$$

with $A_{-1}(t, \cdot)$ considered as an element in $L^2(-r, 0; \mathbb{R}^{n \times n})$.

For the proof see [14, Thms. 2.1 and 3.5].

We return to (2.1) and recall the following variation of constants formula.

LEMMA 2.2. *If for $z_0 \in Z$ we define $z(t, s; z_0) \in Z$ by*

$$z(t, s; z_0) = (x(t, s; z_0, f), x_t(\cdot, s; z_0, f)) \quad \text{for } (t, s) \in \Delta$$

then

$$(2.4) \quad z(t, s; z_0) = T(t, s)z_0 + \int_s^t T(t, \sigma)(f(\sigma), 0) d\sigma \quad \text{for } (t, s) \in \Delta.$$

This result is proved in [6, Thm. 3.1] and in the autonomous case it also follows trivially from [2], [5], [11].

In this paper, we shall consider the following optimal control problem:

Find $u \in L^2(t_0; t_f; \mathbb{R}^m)$ which minimizes

$$(P) \quad \begin{aligned} J(t_0, \eta, \phi, u) = & (Fx(t_f), x(t_f))_{\mathbb{R}^n} + \int_{t_0}^{t_f} (D(t)x(t), x(t))_{\mathbb{R}^n} dt \\ & + \int_{t_0}^{t_f} (C(t)u(t), u(t))_{\mathbb{R}^m} dt \end{aligned}$$

subject to

$$\begin{aligned} \dot{x}(t) &= L(t, x_t) + B(t)u(t), \quad t_0 \leq t \leq t_f, \\ x(t_0) &= \eta, x_{t_0} = \phi, \quad \text{where } (\eta, \phi) \in Z \quad \text{and } t_0, t_f \text{ are given.} \end{aligned}$$

In the notation of the cost functional J we let $x(t)$ stand for $x(t, t_0; \eta, \phi, B(t)u(t))$. The assumptions on F, D, C and B are the following:

$$(2.5) \quad \begin{aligned} F &\in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n), \text{ selfadjoint, nonnegative,} \\ D &\in \mathcal{B}_\infty(t_0, t_f; \mathbb{R}^{n \times n}), \text{ selfadjoint, nonnegative,} \\ C &\in \mathcal{B}_\infty(t_0, t_f; \mathbb{R}^{m \times m}), \text{ selfadjoint, } C(t) \geq c > 0 \\ &\text{for some } c > 0 \text{ and almost all } t, \\ B &\in \mathcal{B}_\infty(t_0, t_f; \mathbb{R}^{n \times m}). \end{aligned}$$

For the presentation of the approximation results we choose a sequence of closed linear subspaces $\{Z^N\}_{N=1}^\infty$ of Z and orthogonal projections

$$P^N : Z \rightarrow Z^N \quad \text{for } N = 1, 2, \dots$$

We shall also use the operator $Q_0 : \mathbb{R}^n \rightarrow Z$ given by

$$Q_0 \eta = (\eta, 0).$$

Of course, Q_0 can be represented as an $n \times n$ Z -valued matrix by

$$Q_0 = \begin{pmatrix} (1, 0) & 0 \\ & \ddots \\ 0 & (1, 0) \end{pmatrix}$$

where 0 stands for the zero-element in Z . In general, we shall not distinguish between the operator Q_0 and its representation. With this notation (2.4) can be written in the form

$$(2.6) \quad z(t, s; z_0) = T(t, s)z_0 + \int_s^t T(t, \sigma)Q_0 f(\sigma) d\sigma.$$

Motivated by earlier work on approximation of FDE [2], [5], [10], we may impose the following hypotheses:

- (H1) *There exists a family of evolution operators $T^N(t, s) : Z \rightarrow Z$, for $N = 1, 2, \dots$ and $(t, s) \in \Delta$ such that*
- i) $\|T^N(t, s)\| \leq \bar{M} e^{\bar{\omega}(t-s)}$ for some $\bar{M} > 0, \bar{\omega} \in \mathbb{R}$.
 - ii) $T^N(t, s)Z^N \subset Z^N$ for all $(t, s) \in \Delta$.

iii) *There exists a real-valued function $\bar{\rho}$ such that*

$$|T(t, s)z - T^N(t, s)z| \leq \bar{\rho}(N, z).$$

Of course in the examples that we have in mind $\bar{\rho}$ will tend to 0 at a certain rate as N goes to ∞ ; the dependence of $\bar{\rho}$ on z will also indicate possible dependence on derivatives of z (compare § 3).

(H2) $\lim_{N \rightarrow \infty} P^N z = z$ for all $z \in Z$.

To get estimates on the rate of convergence we need to introduce a family of operators $Q^N: \mathbb{R}^n \rightarrow Z$, which act as “smoothing operators” for Q_0 .

(H3) *There exists a sequence of linear operators $Q^N: \mathbb{R}^n \rightarrow Z, N = 1, 2, \dots$, such that*

i) $Q^N \mathbb{R}^n \subset Z^N$.

ii) $\|Q^N - Q_0\|_{\mathcal{L}(\mathbb{R}^n; Z)} \leq \rho_Q(N)$ for some real-valued function ρ_Q .

iii) $\|Q^N\|_{\mathcal{L}(\mathbb{R}^n; Z)} \leq q$ for some $q \geq 1$, independent of N .

Throughout this section we assume (H1)–(H3) to hold. A possible candidate for Q^N is the matrix whose columns are the orthogonal projections of the columns of Q_0 onto Z^N . Notice that (H3i) implies that there exist a matrix $Q_0^N \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ and a function valued matrix $Q_1^N \in L^2(-r, 0; \mathbb{R}^{n \times n})$ such that $((Q_0^N)_j, (Q_1^N)_i) \in Z^N$ for $j = 1, \dots, n$, where $(E)_j$ stands for the j th column of a matrix E . In the examples that we have in mind Q^N can always be chosen as a diagonal matrix, with diagonal elements in $\mathbb{R} \times L^2(-r, 0; \mathbb{R})$ approximating $(1, 0) \in \mathbb{R} \times L^2(-r, 0; \mathbb{R})$. The need for introducing the family Q^N to obtain estimates on the rate of convergence will become apparent from the analysis below. The underlying problem, however, can be explained for real-valued functions on $[-1, 0]$.

Remark 2.1. To demonstrate the need for introducing the family of operators Q^N let $g_0: [-1, 0] \rightarrow \mathbb{R}$ be given by

$$g_0(s) = \begin{cases} 1 & \text{for } s = 0, \\ 0 & \text{for } s \in [-1, 0). \end{cases}$$

It is not hard to find a sequence of functions $g_N: [-1, 0] \rightarrow \mathbb{R}$, such that $(\alpha) g_N(0) = 1$; $(\beta) |g_N|_{L^2} = O(1/N^\rho)$ for some $\rho > 0$; $(\gamma) g_N \in W^{1,2}(-1, 0; \mathbb{R})$; $(\delta) |g_N|_{L^1} \leq M_1$ for some M_1 independent of N . In fact, we may take

$$g_N(t) = \begin{cases} Nt + 1 & \text{for } t \in \left[-\frac{1}{N}, 0\right], \\ 0 & \text{otherwise;} \end{cases}$$

however, $|g_N|_{L^2}$ diverges like \sqrt{N} . For functions in $W^{2,2}(-1, 0; \mathbb{R})$ we analyze the question more precisely: there exists *no* family of functions $\{g_N\}$ such that

- (2.7)
$$\begin{aligned} &\text{i) } \lim_N g_N(0) = 1, \\ &\text{ii) } \lim_N |g_N|_{L^2} = 0, \\ &\text{iii) } g_N \in W^{2,2}(-1, 0; \mathbb{R}), \\ &\text{iv) } |g_N|_{L^2} \leq M_1 \text{ and } |\ddot{g}_N|_{L^2} \leq M_2, \text{ both uniformly in } N. \end{aligned}$$

Proof. Assuming (2.7i–iii), we argue that (2.7iv) cannot hold. We first show that $\lim_N g_N(-1) = 0$. For suppose there exists a subsequence, again denoted by g_N such that $g_N(-1) \geq \alpha > 0$, for all N . (The case $\alpha < 0$ is treated similarly.) Then

$$g_N(\varepsilon - 1) = g_N(-1) + \int_{-1}^{\varepsilon-1} \dot{g}_N(s) ds \geq \alpha - \sqrt{\varepsilon} M_1.$$

Therefore, there exists $\varepsilon_0 > 0$ and $\bar{\alpha} > 0$ such that $g_N(\varepsilon - 1) \geq \bar{\alpha} > 0$ for all N and $\varepsilon \in [0, \varepsilon_0]$. This contradicts (2.7ii). Next, we verify that $\lim_N \dot{g}_N(-1) = 0$. If not, there exists a subsequence, again denoted by g_N such that $\dot{g}_N(-1) \geq \tilde{\alpha} > 0$, for all N (the case $\tilde{\alpha} < 0$ is treated similarly). Then

$$g_N(\varepsilon - 1) = g_N(-1) + \varepsilon \dot{g}_N(-1) + \int_{-1}^{\varepsilon-1} (\varepsilon - 1 - s) \ddot{g}_N(s) ds \geq g_N(-1) + \varepsilon \left(\tilde{\alpha} - M_2 \sqrt{\frac{\varepsilon}{3}} \right),$$

so that there exist constants $\tilde{\varepsilon}_0 > 0$ and $k > 0$ such that $g_N(\varepsilon - 1) \geq g_N(-1) + \varepsilon k$, for all N and $\varepsilon \in [0, \tilde{\varepsilon}_0]$, which again contradicts (2.7ii). In a similar way one can show that $\lim_N \ddot{g}_N(0) = \infty$. Since the left-hand side in the next estimate tends to ∞ ,

$$|\dot{g}_N(0) - \dot{g}_N(-1)| \leq \left(\int_{-1}^0 |\ddot{g}_N(s)|^2 ds \right)^{1/2},$$

we see that (2.7iv) is violated and hence the proof of the above claim is completed.

There is yet another way of considering properties (2.7i-iv), which is interesting from the point of view of spline analysis. We let $\tilde{s}_N \in W^{2,2}(-1, 0; \mathbb{R})$ denote the unique cubic Hermite spline function given by

$$(2.8) \quad \begin{aligned} \tilde{s}'_N(t_j^N) &= \tilde{s}_N(t_j^N) = 0 \quad \text{for } j = 2, \dots, N, \\ \tilde{s}_N(t_1^N) &= 0, \quad \tilde{s}'(t_1^N) = \beta, \\ \tilde{s}_N(0) &= 1, \quad \tilde{s}'_N(0) = \alpha, \end{aligned}$$

for a partition $t_j^N = -j/N$, $j = 0, \dots, N$ of $[-1, 0]$. A simple calculation shows that

$$\tilde{s}_N(t) = \begin{cases} (-2N^3 + (\alpha + \beta)N^2)t^3 + (2\alpha N + \beta N - 3N^2)t^2 + \alpha t + 1 & \text{for } t \in [t_1^N, 0], \\ 0 & \text{otherwise.} \end{cases}$$

We recall that the variational problem of finding the function $v \in W^{2,2}(-1, 0; \mathbb{R})$ satisfying (2.8) and minimizing $|\ddot{v}|_{L^2}$ is exactly the cubic Hermite spline \tilde{s}_N ; but $|\tilde{s}_N''|_{L^2}$ diverges like $N^{3/2}$. Of course, a similar negative result can be shown for cubic spline functions. To relate the above observations to the operator Q_0 we suppose $r = n = 1$ and $Z^N \subset \{(\phi(0), \phi) | \phi \in W^{2,2}(-1, 0; \mathbb{R})\}$. Then we have demonstrated that there does not exist a sequence of operators $Q^N : \mathbb{R} \rightarrow Z^N$ such that $\lim_N \|Q^N - Q_0\|_{\mathcal{L}(\mathbb{R}, Z)} = 0$ and such that $\|d(P_2 Q^N)\|_{\mathcal{L}(\mathbb{R}, L^2(-1, 0; \mathbb{R}))}$ and $\|d^2(P_2 Q^N)\|_{\mathcal{L}(\mathbb{R}, L^2(-1, 0; \mathbb{R}))}$ are uniformly bounded in N ; here d denotes the differentiation operator.

To explain the significance of the above negative result, we recall that in order to get good convergence results in spline analysis, the $(L^2$ - and Chebyshev-norms of the) derivatives of the approximated function play an essential role [13], [16]. In the next section we shall apply the general result of this section to specific spline approximation schemes, and it is no surprise that again the convergence of $T^N(t, s)z$ to $T(t, s)z$ depends on the smoothness of z (see [5]). A brief look at (2.4) or (2.6) indicates that this will cause severe difficulties, since under the integral the operator $T(t, s)$ always acts on a discontinuous function. The special form of the integral will help to get the convergence result. But for estimates of the rate of convergence of control, state, payoff and Riccati operator, certain uniformities in the convergence of Q^N to Q_0 would be needed to estimate $\int_s^t T^N(t, \sigma) Q^N f(\sigma) d\sigma$, the candidate for approximating the integral term in (2.6). In the case of cubic spline approximations, for example, one would look for a sequence of bounded linear operators Q^N , which in the one-dimensional case ought to consist of elements satisfying (2.7).

To get better results than just convergence, in spite of the above difficulties, we shall use the following simple technique, which we explain by using g_0 : For some desired accuracy ε determine a function g_N (Q^N or $P^N Q_0$ later on in this paper) such that $g_N(0) = 1$ and $|g_N|_{L^2} \leq \varepsilon$. Moreover, g_N will be chosen in such a way that it suits our smoothness requirements for the specific situation and such that $|g_{N+\mu}|_{L^2} \leq \varepsilon$ for $\mu = 1, 2, \dots$. Then, once N is fixed, we can expect that $T^{N+\mu}(t, s)g_N$ converges at a certain rate as $\mu \rightarrow \infty$, depending on the approximation scheme, i.e., on Z^N , and on the smoothness of g_N . This ends Remark 2.1. \square

We return to the development of the theory begun prior to this long remark and aim at an “abstract” formulation of problem (P) in the space Z . We shall need the operators \mathcal{B} and $\mathcal{B}^N \in \mathcal{B}_\infty(t_0, t_f; \mathcal{L}(\mathbb{R}^m, Z))$, \mathcal{D} and $\mathcal{D}^N \in \mathcal{B}_\infty(t_0, t_f; \mathcal{L}(Z, Z))$ and \mathcal{F} and $\mathcal{F}^N \in \mathcal{L}(Z, Z)$ given by

$$\begin{aligned}\mathcal{B}(t) &= Q_0 B(t) \quad \text{and} \quad \mathcal{B}^N(t) = Q^N B(t), \\ \mathcal{F}(\eta, \phi) &= (F\eta, 0) \quad \text{for } (\eta, \phi) \in Z, \quad F^N = P^N \mathcal{F} P^N, \\ \mathcal{D}(t)(\eta, \phi) &= (D(t)\eta, 0) \quad \text{for } (\eta, \phi) \in Z, \quad \mathcal{D}^N(t) = P^N \mathcal{D}(t) P^N.\end{aligned}$$

LEMMA 2.3. *The operators \mathcal{B} and \mathcal{B}^N , \mathcal{D} and \mathcal{D}^N , \mathcal{F} and F^N satisfy the same properties as B , D and F in (2.5), respectively. Moreover,*

$$\begin{aligned}\mathcal{B}^*(t)(\eta, \phi) &= B^*(t)\eta \quad \text{for } (\eta, \phi) \in Z \text{ and } t_0 \leq t \leq t_f, \\ (\mathcal{B}^N)^*(t)(\eta, \phi) &= B^*(t)(Q_0^N)^*\eta + \int_{-r}^0 B^*(t)(Q_1^N)^*(s)\phi(s) ds,\end{aligned}$$

and

$$(\mathcal{B}^N)^* \text{ and } \mathcal{B}^* \in \mathcal{B}_\infty(t_0, t_f; \mathcal{L}(Z, \mathbb{R}^m)).$$

Proof. We shall only verify the representation of $(\mathcal{B}^N)^*$. So let $v \in \mathbb{R}^m$, and $(\eta, \phi) \in Z$ be arbitrary; then

$$\begin{aligned}\langle \mathcal{B}^N v, (\eta, \phi) \rangle &= (Q_0^N B(t)v, \eta)_{\mathbb{R}^n} + \int_{-r}^0 (Q_1^N(s)B(t)v, \phi(s))_{\mathbb{R}^m} ds \\ &= (v, B^*(t)(Q_0^N)^*\eta)_{\mathbb{R}^m} + \int_{-r}^0 (v, B^*(t)(Q_1^N)^*(s)\phi(s))_{\mathbb{R}^m} ds \\ &= (v, B^*(t)[(Q_0^N)^*\eta + \int_{-r}^0 (Q_1^N)^*(s)\phi(s)]_{\mathbb{R}^m}.\end{aligned} \quad \square$$

Next we introduce the family of approximating optimal control problems in which the original problem (\mathcal{P}) is imbedded. Let

$$\begin{aligned}T^0(t, s) &= T(t, s), \quad P^0 = I, \quad \mathcal{B}^0(t) = \mathcal{B}(t), \quad Q^0 = Q_0, \\ \mathcal{D}^0 &= \mathcal{D}, \quad F^0 = \mathcal{F},\end{aligned}$$

and consider

For $t_0 \in \mathbb{R}$, $t_f \in \mathbb{R}$ and $z \in Z$ given, minimize

$$\begin{aligned}J(t_0, P^{N+\mu} z, u) &= \langle F^{N+\mu} z^{N,\mu}(t_f), z^{N,\mu}(t_f) \rangle \\ (\mathcal{P}^{N,\mu}) \quad &+ \int_{t_0}^{t_f} (\langle \mathcal{D}^{N+\mu}(t) z^{N,\mu}(t), z^{N,\mu}(t) \rangle + (C(t)u(t), u(t))) dt\end{aligned}$$

over $u \in L^2(t_0, t_f; \mathbb{R}^m)$ subject to

$$(2.9) \quad z^{N,\mu}(t) = T^{N+\mu}(t, t_0)P^{N+\mu}z + \int_{t_0}^t T^{N+\mu}(t, \eta)\mathcal{B}^N(\eta)u(\eta) d\eta \quad \text{for } t_0 \leq t \leq t_f.$$

For $\mu = N = 0$ and by Lemma 2.2 we have $(\mathcal{P}^{0,0}) = (\mathcal{P})$. Some motivation for the double limit process in $(\mathcal{P}^{N,\mu})$ was already given in Remark 2.1. Since $\mathcal{B}(t) = Q_0 B(t)$ contains the “jump operator” Q_0 which is extremely formidable for higher order convergence of the states ([16, p. 42], [5, Remark 3.3]) we introduce the family of operators Q^N , which will play the role of smoothing and approximating Q_0 . For each N we may then consider a second approximation in μ from which we expect to obtain rate of convergence results. In the estimates that follow we will always be able to separate the influence of the two limit processes, one essentially depending on how well Q^N approximates Q_0 (compare (H3ii)) and the other one on the approximation of $T(t, s)z$ by $T^N(t, s)z$ (compare (H1iii)). The consequence of the first limit process will be dealt with by tedious calculations, whereas for the latter, one may appeal to known results on the approximation of $T(t, s)$ which were obtained by employing the Trotter–Kato theorem from semigroup theory. We will come back to a discussion of the use of this double limit process in Remark 2.4. A similar double limit process was used in [9], where approximation schemes for neutral functional differential equations were derived. There, as a consequence of the specific features of neutral functional differential equations it was needed essentially in the proof of convergence and its necessity was also demonstrated by numerical examples [9, Ex. 3]. In the present paper the double limit process will only be needed for the rate of convergence results, whereas for convergence alone it is superfluous and one can put $\mu = 0$ in this case. We now address ourselves to solving $(\mathcal{P}^{N,\mu})$, $N, \mu = 0, 1, \dots$ and to the question of the behavior of the solution of $(\mathcal{P}^{N,\mu})$ as $N, \mu \rightarrow \infty$. Problems $(\mathcal{P}^{N,\mu})$ are special cases of the linear quadratic optimal control problem considered in [8], from which we have the following information about $(\mathcal{P}^{N,\mu})$.

Under the assumptions on F, D and C the unique optimal controls are the solutions of

$$J'(t_0, P^{N+\mu}z(t_0), u)v = 0 \quad \text{for all } v \in L^2(t_0, t_f; \mathbb{R}^m),$$

where

$$J'(t_0, P^{N+\mu}z(t_0), u)v$$

denotes the Fréchet derivative of J at u applied to v . Therefore, after some calculations one finds that the optimal controls $\tilde{u}^{N,\mu}$ are given by

$$(2.10) \quad \tilde{u}^{N,\mu}(t) = -((V_{t_0}^{N,\mu})^{-1}W_{t_0}^{N,\mu}z)(t) \quad \text{a.e. in } [t_0, t_f],$$

where

$$V_{t_0}^{N,\mu} \in \mathcal{L}(L^2(t_0, t_f; \mathbb{R}^m), L^2(t_0, t_f; \mathbb{R}^m))$$

and

$$W_{t_0}^{N,\mu} \in \mathcal{L}(Z, L^2(t_0, t_f; \mathbb{R}^m))$$

and

$$(2.11) \quad V_{t_0}^{N,\mu} = C + (\mathcal{B}^N)^*(\mathcal{F}_{t_0}^{N+\mu})^* \mathcal{D}^{N+\mu} \mathcal{F}_{t_0}^{N+\mu} \mathcal{B}^N + (\mathcal{B}^N)^*(\mathcal{F}_{t_0}^{N+\mu})^* F^{N+\mu} \mathcal{F}_{t_0}^{N+\mu} \mathcal{B}^N,$$

$$(2.12) \quad W_{t_0}^{N,\mu} = (\mathcal{B}^N)^*(\mathcal{F}_{t_0}^{N+\mu})^* \mathcal{D}^{N+\mu} T_{t_0}^{N+\mu} + (\mathcal{B}^N)^*(\mathcal{F}_{t_0}^{N+\mu})^* F^{N+\mu} T^{N+\mu}(t_f, t_0).$$

Here

$$\begin{aligned}\mathcal{T}_{t_0}^{N+\mu} &\in \mathcal{L}(L^2(t_0, t_f; Z), L^2(t_0, t_f; Z)), \\ \mathcal{F}_{t_0}^{N+\mu} &\in \mathcal{L}(L^2(t_0, t_f; Z), Z), \\ T_{t_0}^{N+\mu} &\in (Z, L^2(t_0, t_f; Z))\end{aligned}$$

are defined by

$$\begin{aligned}(\mathcal{T}_{t_0}^{N+\mu}\phi)(s) &= \int_{t_0}^s T^{N+\mu}(s, \eta)\phi(\eta) d\eta \quad \text{for } \phi \in L^2(t_0, t_f; Z), \\ (\mathcal{F}_{t_0}^{N+\mu}\phi) &= (\mathcal{T}_{t_0}^{N+\mu}\phi)(t_f), \\ ((\mathcal{T}_{t_0}^{N+\mu})^*\phi)(t) &= \int_t^{t_f} (T^{N+\mu})^*(\eta, t)\phi(\eta) d\eta, \\ ((\mathcal{F}_{t_0}^{N+\mu})^*\tilde{z})(t) &= (T^{N+\mu})^*(t_f, t)\tilde{z} \quad \text{for } \tilde{z} \in Z, \\ (T_{t_0}^{N+\mu}\tilde{z})(t) &= T^{N+\mu}(t, t_0)\tilde{z}.\end{aligned}$$

Consider for a moment the optimal control problems $(\mathcal{P}^{N,\mu})$ with $J(t_0, P^{N+\mu}z, u)$ replaced by $J(s, P^{N+\mu}z^{N,\mu}(s), u)$, $t_0 \leq s \leq t_f$; if we let $\tilde{u}_s^{N,\mu}$ denote the corresponding optimal control, then it is clear from the above that

$$(2.13) \quad (\tilde{u}_s^{N,\mu})(t) = -((V_s^{N,\mu})^{-1}W_s^{N,\mu}P^{N+\mu}z)(t) \quad \text{a.e. in } [s, t_f],$$

if $V_s^{N,\mu} \in \mathcal{L}(L^2(s, t_f; \mathbb{R}^m), L^2(s, t_f; \mathbb{R}^m))$ and $W_s^{N,\mu} \in \mathcal{L}(Z, L^2(s, t_f; \mathbb{R}^m))$ are defined analogously to $V_{t_0}^{N,\mu}$ and $W_{t_0}^{N,\mu}$ in (2.11) and (2.12), respectively. For $z \in Z$ the optimal trajectories $S^{N,\mu}(t, s)P^{N+\mu}z$ corresponding to $J(s, P^{N+\mu}z, u)$ are then given by

$$(2.14) \quad \begin{aligned}S^{N,\mu}(t, s)P^{N+\mu}z &= T^{N+\mu}(t, s)P^{N+\mu}z \\ &\quad - \int_s^t T^{N+\mu}(t, \eta)\mathcal{B}^N(\eta)((V_s^{N,\mu})^{-1}W_s^{N,\mu}P^{N+\mu}z)(\eta) d\eta.\end{aligned}$$

We extend $S^{N,\mu}(t, s)$ to Z and let $S^{N,\mu}(t, s)z$ be given by (2.14) with $P^{N,\mu}z$ replaced by z . From the results in [8] it follows that $S^{N,\mu}(t, s)$ is an evolution operator in Δ on Z for each N, μ , and moreover that $\tilde{u}^{N,\mu}$ is also given by

$$(2.15) \quad \tilde{u}^{N,\mu}(t) = -C^{-1}(t)(\mathcal{B}^N(t))^*\Pi^{N,\mu}(t)S^{N,\mu}(t, t_0)P^{N+\mu}z \quad \text{a.e.,}$$

with

$$(2.16) \quad \begin{aligned}\Pi^{N,\mu}(t)z &= (T^{N+\mu})^*(t_f, t)F^{N+\mu}S^{N,\mu}(t_f, t)z \\ &\quad + \int_t^{t_f} (T^{N+\mu})^*(\eta, t)\mathcal{D}^{N+\mu}(\eta)S^{N,\mu}(\eta, t)z d\eta \quad \text{for } t_0 \leq t \leq t_f, \quad z \in Z.\end{aligned}$$

The basis for the numerical approximation scheme will be (2.10), (2.14) and (2.16) together with the following Riccati integral equation for $\Pi^{N,\mu}$:

$$(2.17) \quad \begin{aligned}\Pi^{N,\mu}(t)z &= (T^{N+\mu})^*(t_f, t)F^{N+\mu}T^{N+\mu}(t_f, t)z \\ &\quad + \int_t^{t_f} (T^{N+\mu})^*(\eta, t) \\ &\quad \cdot [\mathcal{D}^{N+\mu}(\eta) - \Pi^{N,\mu}(\eta)\mathcal{B}^N(\eta)C^{-1}(\eta)(\mathcal{B}^N(\eta))^*\Pi^{N,\mu}(\eta)]T^{N+\mu}(\eta, t)z d\eta.\end{aligned}$$

Since Z is separable, so that $T^*(\cdot, \cdot)$ is strongly measurable, (2.17) is also a direct consequence of the results in [8]; moreover

$$\Pi^{N,\mu}(t) \text{ is nonnegative and selfadjoint.}$$

To establish the approximation results we adopt the following:

DEFINITIONS.

- (i) 0 as a superscript may be dropped.
- (ii) $b = \sup_{t \in [t_0, t_f]} \|B(t)\|$.
- (iii) $\tilde{c} = \sup_{t \in [t_0, t_f]} \|C(t)\|$.
- (iv) $d = \sup_{t \in [t_0, t_f]} \|D(t)\|$.
- (v) $f = \|F\|_{\mathcal{L}(Z, Z)}$.
- (vi) We assume that $M \leq \bar{M}$, $\omega \leq \bar{\omega}$, and put $\tilde{M} = \bar{M} e^{\bar{\omega}(t_f - t_0)}$.

Moreover we note that

- (vii) By (H1i–iii) there exists a real-valued function ρ such that

$$|T(t, s)z - T^N(t, s)P^N z| \leq \rho(N, z) \quad \text{uniformly in } \Delta;$$

indeed $\rho(N, z) = \bar{\rho}(N, z) + \tilde{M}|P^N z - z|$.

- (viii) By (H1iii) there exists a real-valued function $\tilde{\rho}$ such that

$$\|T(t, s)Q^N - T^{N+\mu}(t, s)Q^N\|_{\mathcal{L}(\mathbb{R}^n, Z)} \leq \tilde{\rho}(\mu + N, Q^N);$$

indeed one can let

$$\tilde{\rho}(\mu + N, Q^N) = \sqrt{n} \max_{j=1, \dots, n} \bar{\rho}(N + \mu, ((Q_0^N)_j, (Q_1^N)_j)).$$

- (ix) The constants K_i to be used below depend on the following variables of $(\mathcal{P}^{N,\mu})$:

$$K_i = K_i(n, A_i, f, d, b, c, q, t_0, t_f),$$

and are calculated explicitly in the proofs. We shall also use constants k_i whose values are given explicitly in the theorems.

LEMMA 2.4.

$$(a) \quad \|W_{t_0}^{N,\mu}\|_{\mathcal{L}(Z; C(t_0, t_f; \mathbb{R}^m))} \leq bq\tilde{M}^2(d(t_f - t_0) + f) \stackrel{\text{def}}{=} k_0$$

$$(b) \quad \|(V_{t_0}^{N,\mu})^{-1}\|_{\mathcal{L}(L^2(t_0, t_f; \mathbb{R}^m), L^2(t_0, t_f; \mathbb{R}^m))} \leq c^{-1}.$$

(c) There exist constants K_1 and K_2 such that for all $z \in Z$

$$\begin{aligned} & \sup_{t \in [t_0, t_f]} |(W_{t_0} z)(t) - (W_{t_0}^{N,\mu} P^{N+\mu} z)(t)|_{\mathbb{R}^m} \\ & \leq \rho_Q(N)|z|k_1 + \tilde{\rho}(N + \mu, Q^N)|z|K_1 + \rho(N + \mu, z)K_2 + \|P^{N+\mu}Q_0 - Q_0\||z|k_2, \end{aligned}$$

where

$$k_1 = b\sqrt{n}\tilde{M}^2[(2\bar{\omega})^{-1}d + f],$$

$$k_2 = bq\sqrt{n}\tilde{M}^2[d(t_f - t_0)(2\bar{\omega})^{-1} + f].$$

(d) There exists a constant K_3 such that, for all $w \in L^2(t_0, t_f; \mathbb{R}^m)$,

$$\begin{aligned} & \sup_{t \in [t_0, t_f]} |((V_{t_0} - V_{t_0}^{N,\mu})w)(t)| \leq \rho_Q(N)k_3|w|_{L^2} + \tilde{\rho}(N + \mu, Q^N)K_3|w|_{L^2} \\ & + \|Q_0 - P^{N+\mu}Q_0\|k_4|w|_{L^2}, \end{aligned}$$

where

$$k_3 = b^2 \sqrt{n}(1+q) \tilde{M}^2 \left(\frac{f}{\sqrt{2\bar{\omega}}} + \frac{d}{\sqrt{\bar{\omega}^3}} \right)$$

and

$$k_4 = b^2 \sqrt{n} q^2 \tilde{M}^2 \left(\frac{d}{\sqrt{2\bar{\omega}^3}} + \frac{f}{\sqrt{2\bar{\omega}}} \right).$$

Remark 2.2. It is simple to check that the same estimates as in the previous lemma also hold for $V_s^{N,\mu} \in \mathcal{L}(L^2(s, t_f; \mathbb{R}^m), L^2(s, t_f; \mathbb{R}^m))$ and $W_s^{N,\mu} \in \mathcal{L}(Z, L^2(s, t_f; \mathbb{R}^m))$, for $s \in [t_0, t_f]$.

THEOREM 2.1. For the optimal controls $\tilde{u}^{N,\mu}$ and \tilde{u} we have the following L^2 estimate:

$$\begin{aligned} |\tilde{u} - \tilde{u}^{N,\mu}|_{L^2(t_0, t_f; \mathbb{R}^m)} &\leq c^{-1}(t_f - t_0)^{1/2} [\rho_Q(N)|z|k_1 + \tilde{\rho}(N + \mu, Q^N)|z|K_1 \\ &\quad + \rho(N + \mu, z)K_2 + \|P^{N+\mu}Q_0 - Q_0\||z|k_2] \\ &\quad + c^{-2}(t_f - t_0)k_0|z|[\rho_Q(N)k_3 + \tilde{\rho}(N + \mu, Q^N)K_3 + \|Q_0 - P^{N+\mu}Q_0\|k_4]. \end{aligned}$$

Proof. The proof, using Lemma 2.4, follows from the following simple estimate:

$$\begin{aligned} |\tilde{u} - \tilde{u}^{N,\mu}|_{L^2} &\leq |V_{t_0}^{-1}W_{t_0}z - (V_{t_0}^{N,\mu})^{-1}W_{t_0}^{N,\mu}P^{N+\mu}z|_{L^2} \\ &= |V_{t_0}^{-1}W_{t_0}z - V_{t_0}^{-1}W_{t_0}^{N,\mu}P^{N+\mu}z|_{L^2} \\ &\quad + |V_{t_0}^{-1}W_{t_0}^{N,\mu}P^{N+\mu}z - (V_{t_0}^{N,\mu})^{-1}W_{t_0}^{N,\mu}P^{N+\mu}z|_{L^2} \\ &\leq c^{-1}|W_{t_0}z - W_{t_0}^{N,\mu}P^{N+\mu}z|_{L^2} + |(V_{t_0}^{N,\mu})^{-1}(V_{t_0} - V_{t_0}^{N,\mu})(V_{t_0})^{-1}W_{t_0}^{N,\mu}P^{N+\mu}z|_{L^2} \\ &\leq c^{-1}|W_{t_0}z - W_{t_0}^{N,\mu}P^{N+\mu}z|_{L^2} \\ &\quad + c^{-2}\|V_{t_0} - V_{t_0}^{N,\mu}\|_{\mathcal{L}(L^2(t_0, t_f; \mathbb{R}^m), L^2(t_0, t_f; \mathbb{R}^m))} k_0(t_f - t_0)^{1/2}|z|. \quad \square \end{aligned}$$

Remark 2.3. Although it is not difficult to show that the controls $\tilde{u}^{N,\mu}$, $N, \mu = 0, 1, \dots$, are continuous if $B(\cdot)$ and $C(\cdot)$ are, and that $V^{N,\mu}$ is invertible in $\mathcal{L}(C(t_0, t_f; \mathbb{R}^m), C(t_0, t_f; \mathbb{R}^m))$ it does not seem possible to find a uniform bound on $\|(V^{N,\mu})^{-1}\|_{\mathcal{L}(C(t_0, t_f; \mathbb{R}^m), C(t_0, t_f; \mathbb{R}^m))}$. We shall, however, consider the question of uniform convergence of the controls in Theorem 2.4.

Remark 2.4. The use of Theorem 2.1 will be demonstrated for the case where the subspaces Z^N are chosen as subspaces of spline functions, for example. Then, if one is merely interested in convergence of the optimal controls $\tilde{u}^{N,\mu}$ one may put $\mu = 0$ and Theorem 2.1 will guarantee L^2 -convergence of $\tilde{u}^{N,0}$ to \tilde{u} (see Corollary 2.1 below). However, if the initial data z in Z are picked sufficiently smooth one would expect to find higher order estimates on the rate of convergence. Unfortunately, even for smooth initial data, one still has to deal with the “jump” operator Q_0 used in the variation-of-constants formula (compare (2.4), $(\mathcal{P}^{N,\mu})$ and (2.12)): For any given ε one can use the explicit formulas for k_i and (H2), (H3ii) to determine $N = N(\varepsilon) > 0$, such that for $\mu = 1, 2, \dots$

$$\begin{aligned} \|Q_0 - Q^N\| [c^{-1}(t_f - t_0)^{1/2}|z|k_1 + c^{-2}(t_f - t_0)k_0|z|k_3] \\ + \|P^{N+\mu}Q_0 - Q_0\| [c^{-1}(t_f - t_0)^{1/2}|z|k_2 + c^{-2}(t_f - t_0)k_0|z|k_4] < \varepsilon. \end{aligned}$$

Fixing N , Theorem 2.1 guarantees that the approximating optimal controls $\tilde{u}^{N,\mu}$ converge as $\mu \rightarrow \infty$ at a rate given by $\tilde{\rho}(N + \mu, Q^N)$ and $\rho(N + \mu, z)$ into the previously fixed ε -neighborhood of the optimal control \tilde{u} . Similar remarks apply for the optimal

trajectories, the payoff functional and the Riccati operators in the following theorems; again the constants k_i multiplying $\|Q_0 - Q^N\|$ and $\|Q_0 - P^{N+\mu}Q_0\|$ are stated explicitly.

Remark 2.5. It can be seen easily that Theorem 2.1 remains true if $\tilde{u}^{N,\mu}$ and \tilde{u} are replaced by $\tilde{u}_s^{N,\mu}$ and \tilde{u}_s as defined in (2.13). For the optimal trajectories $S(t, s)z$, with $(t, s) \in \Delta$, we have the following estimate.

THEOREM 2.2.

(a) *For the approximating optimal trajectories we have*

$$\|S^{N,\mu}(t, s)\| \leq k_5 \quad \text{for all } (t, s) \in \Delta \text{ and } \mu, N = 0, 1, 2, \dots,$$

where $k_5 = \tilde{M}(1 + bq(t_f - t_0)c^{-1}k_0)$.

(b) *There exists a constant K_4 such that*

$$\begin{aligned} |S(t, s)z - S^{N,\mu}(t, s)P^{N+\mu}z| &\leq \rho(N + \mu, z) + \tilde{\rho}(N + \mu, Q^N)|z|K_4 \\ &\quad + |\tilde{u} - \tilde{u}^{N,\mu}|_{L^2}k_6 + |z|\rho_Q(N)k_7, \end{aligned}$$

where

$$k_6 = b\tilde{M}\frac{1}{\sqrt{2\tilde{\omega}}}, \quad k_7 = k_6c^{-1}k_0(t_f - t_0)^{1/2}.$$

THEOREM 2.3. *For all $y, z \in Z$ and $t \in [t_0, t_f]$*

$$(a) \quad \|\Pi^{N,\mu}(t)\| \leq \tilde{M}k_5(f + d(t_f - t_0)).$$

$$\begin{aligned} (b) \quad \langle \Pi(t)z - \Pi^{N,\mu}(t)P^{N+\mu}z, y \rangle &\leq (d(t_f - t_0) + f)\{\tilde{\rho}(N + \mu, y)|z|k_5 \\ &\quad + \tilde{M}|y|[\rho(N + \mu, z) + |z|\tilde{\rho}(N + \mu, Q^N)K_4 \\ &\quad + |\tilde{u} - \tilde{u}^{N,\mu}|_{L^2}k_6 + |z|\rho_Q(N)k_7 + |z|k_5\|Q_0 - P^{N+\mu}Q_0\|]\}. \end{aligned}$$

For $\mu = 0$ we have the following corollary to Theorems 2.1–2.3.

COROLLARY 2.1. *There exist constants K_5, K_6, K_7 such that*

$$\begin{aligned} (a) \quad &|\tilde{u} - \tilde{u}^{N,0}|_{L^2(t_0, t_f; \mathbb{R}^m)} \leq K_5[\tilde{\rho}(N, Q_0)|z| + \rho(N, z) + \rho_Q(N)|z|], \\ (b) \quad &|S(t, s)z - S^{N,0}(t, s)P^Nz| \leq K_6[\tilde{\rho}(N, Q_0)|z| + \rho(N, z) + \rho_Q(N)|z|], \\ (c) \quad &\langle \Pi(t)z - \Pi^{N,0}P^Nz, y \rangle \leq K_7[\tilde{\rho}(N, y)|z| + \rho(N, z)|y| + |y||z|(\rho_Q(N) + \tilde{\rho}(N, Q_0))]. \end{aligned}$$

Proof. By Theorem 2.1 we have

$$(2.18) \quad |\tilde{u} - \tilde{u}^{N,0}|_{L^2(t_0, t_f; \mathbb{R}^m)} \leq \tilde{K}_7[\rho_Q(N)|z| + \tilde{\rho}(N, Q^N)|z| + \rho(N, z) + \|P^NQ_0 - Q_0\||z|].$$

Since $Q^N\mathbb{R}^n \subset Z^N$ and since P^N is an orthogonal projection, $|P^NQ_0\eta - Q_0\eta| < \|Q^N - Q_0\|\|\eta\|$ for all η , which implies that

$$(2.19) \quad \|P^NQ_0 - Q_0\| < \rho_Q(N).$$

Also, for $(t, s) \in \Delta$ we find

$$\begin{aligned} (2.20) \quad \tilde{\rho}(N, Q^N) &\leq \|T(t, s)(Q^N - Q_0)\| + \|(T(t, s) - T^N(t, s))Q_0\| + \|T^N(t, s)(Q_0 - Q^N)\| \\ &\leq 2\tilde{M}e^{\tilde{\omega}(t-s)}\rho_Q(N) + \tilde{\rho}(N, Q_0). \end{aligned}$$

Estimates (2.18)–(2.20) imply (a). To prove (b) and (c) one uses (2.19), (2.20) and (a) of this Corollary together with Theorem 2.2 and Theorem 2.3, respectively. \square

COROLLARY 2.2. *For the payoff J the following estimate holds:*

$$\begin{aligned} |J(t_0, P^{N+\mu}z, \tilde{u}^{N,\mu}) - J(t_0, z, \tilde{u})| \\ \leq [(t_f - t_0)d + f]k_5|z|\{2[\rho(N + \mu, z) + \tilde{\rho}(N + \mu, Q^N)]|z|K_4 \\ + |\tilde{u} - \tilde{u}^{N,\mu}|_{L^2}k_6 + |z|\rho_Q(N)k_7\} + \|P^{N+\mu}Q_0 - Q_0\|k_5|z| \\ + 2|\tilde{u} - \tilde{u}^{N,\mu}|_{L^2}k_0c^{-1}(t_f - t_0)^{1/2}\tilde{c}|z|. \end{aligned}$$

Finally, we discuss the convergence of $\tilde{u}^{N,\mu}$ to \tilde{u} in the supremum-norm. For the sake of a simpler representation we restrict our attention to the case $\mu = 0$.

THEOREM 2.4. *If $\lim_{N \rightarrow \infty} \rho_Q(N) = 0$ and $\lim_{N \rightarrow \infty} \tilde{\rho}(N, z) = 0$ for each $z \in Z$, then*

$$\lim_{N \rightarrow \infty} \sup_{t \in [t_0, t_f]} |\tilde{u}(t) - \tilde{u}^{N,0}(t)| = 0.$$

Because of the engineering importance of feedback control, we return to the feedback control law (2.15). First we notice that the nonnegative and selfadjoint operator $\Pi^{N,\mu}(t) \in \mathcal{L}(Z, Z)$ can be written as a matrix of operators

$$\Pi^{N,\mu}(t) = \begin{pmatrix} \Pi_{00}^{N,\mu}(t) & \Pi_{01}^{N,\mu}(t) \\ \Pi_{10}^{N,\mu}(t) & \Pi_{11}^{N,\mu}(t) \end{pmatrix}, \quad t_0 \leq t \leq t_f,$$

where $\Pi_{00}^{N,\mu}$ is a real symmetric and nonnegative $n \times n$ matrix and $\Pi_{10}^{N,\mu}(t) \in \mathcal{L}(\mathbb{R}^n, L^2)$ can be realized by a real square integrable $n \times n$ matrix function $\Pi_{10}^{N,\mu}(t, \cdot)$ on $[-r, 0]$. Moreover, $\Pi_{10}^{N,\mu}(t)^* = \Pi_{01}^{N,\mu}(t)$ and $\Pi_{11}^{N,\mu}(t)$ is a nonnegative selfadjoint operator in $\mathcal{L}(L^2, L^2)$. Applying Lemma 2.3 and (2.15) we can therefore give the optimal control law for problem (P) by

$$u(t) = -C^{-1}(t)B^*(t)\Pi_{00}(t)x(t) - C^{-1}(t)B^*(t) \int_{-r}^0 \Pi_{10}^*(t, \sigma)x_t(\sigma) d\sigma.$$

A similar feedback control law holds for the problems $(\mathcal{P}^{N,\mu})$. The following corollary is a direct consequence of Theorem 2.3. By e_i we denote the i th unit vector in \mathbb{R}^n .

COROLLARY 2.3. (a) *If $\lim_{N \rightarrow \infty} \rho_Q(N) = 0$ and $\lim_{N \rightarrow \infty} \tilde{\rho}(N, z) = 0$ for each $z \in Z$, then we find for the elements in the matrix representation of $\Pi(t)$ that $\lim_{N \rightarrow \infty} \Pi_{00}^{N,0}(t) = \Pi_{00}(t)$ in $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$, $\text{weak-}\lim_{N \rightarrow \infty} \Pi_{11}^{N,0}(t)y = \Pi_{11}(t)y$ in L^2 and $\lim_{N \rightarrow \infty} \Pi_{01}^{N,0}(t)y = \Pi_{01}(t)y$ in \mathbb{R}^n for each $y \in L^2$, with all limits holding uniformly in $t \in [t_0, t_f]$. More precisely, there exists a constant K_9 , such that for all $x, y \in L^2$ and $v \in \mathbb{R}^n$ we have*

$$\begin{aligned} |\Pi_{00}^{N,\mu}(t)v - \Pi_{00}(t)v|_{L^2} &\leq K_9[(\max_i \tilde{\rho}(N + \mu, (e_i, 0)) + \tilde{\rho}(N + \mu, Q^N) \\ &\quad + \rho_Q(N) + \|Q_0 - P^{N+\mu}Q_0\|)|v| + \rho(N + \mu, (v, 0))], \\ |(\Pi_{11}^{N,\mu}(t)y - \Pi_{11}(t)y, x)_{L^2}| \\ &\leq K_9[\tilde{\rho}(N + \mu, (0, x))|y|_{L^2} + \rho(N + \mu, (0, y))|x|_{L^2} \\ &\quad + \tilde{\rho}(N + \mu, Q^N)|y|_{L^2} + \rho_Q(N)|y|_{L^2} + \|Q_0 - P^N Q_0\||y|_{L^2}], \end{aligned}$$

$$\begin{aligned} |\Pi_{01}^{N,\mu}(t)y - \Pi_{01}(t)y| &\leq K_9[(\max_i \tilde{\rho}(N + \mu, (e_i, 0)) + \tilde{\rho}(N + \mu, Q^N) + \rho_Q(N) \\ &\quad + \|Q_0 - P^{N+\mu}Q_0\|)|y|_{L^2} + \rho(N + \mu, (0, y))], \end{aligned}$$

uniformly in $t \in [t_0, t_f]$.

(b) *There exists a constant K_{10} such that for each $z \in Z$ and $t \in [t_0, t_f]$*

$$\begin{aligned} & |C^{-1}(t)(\mathcal{B}^N(t))^* \Pi^{N,\mu}(t)z - C^{-1}(t)\mathcal{B}^*(t)\Pi(t)z| \\ & \leq K_{10}[(\max_i \bar{\rho}(N + \mu, (e_i, 0)) + \tilde{\rho}(N + \mu, Q^N) \\ & \quad + \rho_Q(N) + \|Q_0 - P^{N+\mu}Q_0\|)|z| + \rho(N + \mu, z)]. \end{aligned}$$

Proof. We note that Theorem 2.3(b) clearly holds if $P^{N+\mu}z$ is replaced by z . The above estimates then follow by simple applications of the triangle inequality. \square

We now turn to the final question of this section and denote by $\hat{z}^{N,\mu}(t)$ the trajectory which results from applying the optimal feedback laws associated with the approximating problems $(\mathcal{P}^{N,\mu})$ to the original system:

$$(2.21) \quad \hat{z}^{N,\mu}(t) = T(t, t_0)z + \int_{t_0}^t T(t, \eta)\mathcal{B}(\eta)[-C^{-1}(\eta)(\mathcal{B}^N(\eta))^* \Pi^{N,\eta}(\eta)]\hat{z}^{N,\eta}(\eta) d\eta.$$

for $t \in [t_0, t_f]$. Will the solutions $\hat{z}^{N,\mu}(t)$ of the closed loop system (2.21) converge to $S(t, t_0)$, the optimal trajectory of the original problem (P)? This is indeed the case, as will be shown in the next theorem. To relate (2.21) to the original problem (P) we recall that $\hat{z}^{N,\mu}(t)$ is connected to the function $\hat{x}^{N,\mu} : [t_0 - r, t_f] \rightarrow \mathbb{R}^n$ satisfying

$$(2.22) \quad \begin{aligned} \frac{d}{dt} \hat{x}^{N,\mu}(t) &= L(t, \hat{x}_t^{N,\mu}) + B(t)[-C^{-1}(t)(\mathcal{B}^N(t))^* \Pi^{N,\mu}(t)](\hat{x}^{N,\mu}(t), \hat{x}_t^{N,\mu}) \\ (\hat{x}^{N,\mu}(t_0), \hat{x}_{t_0}^{N,\mu}) &= z, \end{aligned}$$

via $(\hat{x}^{N,\mu}(t), \hat{x}_t^{N,\mu}) = \hat{z}^{N,\mu}(t)$. This can be seen from Lemma 2.2 with $f(t) = -B(t)C^{-1}(t)(\mathcal{B}^N(t))^* \Pi^{N,\mu}(t)\hat{z}^{N,\mu}(t)$.

Let $\mathcal{M} = \{S(t, t_0)z | t \in [t_0, t_f]\}$, which is a compact subset of Z . If $\lim_{N \rightarrow \infty} \bar{\rho}(N, z) = 0$ for each $z \in Z$ then (H1) and (H2) imply that $\lim_{N \rightarrow \infty} (N, \bar{z}) = 0$ uniformly in $\bar{z} \in \mathcal{M}$.

THEOREM 2.5. *There exists a constant K_{11} such that*

$$\begin{aligned} |S(t, t_0)z - \hat{z}^{N,\mu}(t)| &\leq K_{11}[(\max_i \bar{\rho}(N + \mu, (e_i, 0)) + \tilde{\rho}(N + \mu, Q^N) \\ & \quad + \|Q_0 - P^{N+\mu}Q_0\| + \rho_Q(N))|z| + \sup_{z \in \mathcal{M}} \rho(N + \mu, \bar{z})], \end{aligned}$$

for all $t \in [t_0, t_f]$.

Proof. Let $\kappa = \tilde{M}bc^{-1}$. Then we have the following inequalities:

$$\begin{aligned} |S(t, t_0)z - \hat{z}^{N,\mu}(t)| &= \int_{t_0}^t |T(t, \sigma)\mathcal{B}(\sigma)C^{-1}(\sigma) \\ & \quad \cdot [\mathcal{B}^*(\sigma)\Pi(\sigma)S(\sigma, t_0)z - (\mathcal{B}^N(\sigma))^* \Pi^{N,\mu}(\sigma)\hat{z}^{N,\mu}(\sigma)] d\sigma \\ &\leq \kappa \int_{t_0}^t |[\mathcal{B}^*(\sigma)\Pi(\sigma) - (\mathcal{B}^N(\sigma))^* \Pi^{N,\mu}(\sigma)]S(\sigma, t_0)z| d\sigma \\ & \quad + \kappa \int_{t_0}^t |(\mathcal{B}^N(\sigma))^* \Pi^{N,\mu}(\sigma)(S(\sigma, t_0)z - \hat{z}^{N,\mu}(\sigma))| d\sigma \\ &\leq \kappa \int_{t_0}^t |[\mathcal{B}^*(\sigma)\Pi(\sigma) - (\mathcal{B}^N(\sigma))^* \Pi^{N,\mu}(\sigma)]S(\sigma, t_0)z| d\sigma \\ & \quad + \kappa b q \tilde{M} k_5 (f + d(t_f - t_0)) \int_{t_0}^t |S(\sigma, t_0)z - \hat{z}^{N,\mu}(\sigma)| d\sigma, \end{aligned}$$

where in the last estimate we used Theorem 2.3(a). An application of the Gronwall lemma and Corollary 2.3(b) imply the result. \square

Let us denote the optimal control of the closed loop system (2.21) by

$$\hat{u}^{N,\mu}(t) = -C^{-1}(t)(\mathcal{B}^N(t))^* \Pi^{N,\mu}(t) \hat{z}^{N,\mu}(t).$$

As before the optimal control of the original problem (P) is \tilde{u} . The cost functional corresponding to (2.21) is given by

$$\begin{aligned} \hat{J}^{N,\mu}(t_0, z, u) &= \langle \mathcal{F} \hat{z}^{N,\mu}(t_f), \hat{z}^{N,\mu}(t_f) \rangle \\ &+ \int_{t_0}^{t_f} (\langle \mathcal{D}(t) \hat{z}^{N,\mu}(t), \hat{z}^{N,\mu}(t) \rangle + (C(t) \hat{u}^{N,\mu}(t), \hat{u}^{N,\mu}(t))) dt. \end{aligned}$$

COROLLARY 2.4. *There exists a constant K_{12} such that for all $t \in [t_0, t_f]$ we have the estimates*

$$\begin{aligned} |\tilde{u}(t) - \hat{u}^{N,\mu}(t)| &\leq \nu(N, \mu), \\ |J(t_0, z, \tilde{u}) - \hat{J}^{N,\mu}(t_0, z, \hat{u}^{N,\mu})| &\leq \nu(N, \mu), \end{aligned}$$

with

$$\begin{aligned} \nu(N, \mu) &= K_{12}[(\max_i \bar{\rho}(N + \mu, (e_i, 0)) + \tilde{\rho}(N + \mu, Q^N) \\ &+ \rho_Q(N) + \|Q_0 - P^{N+\mu} Q_0\|) |z| + \sup_{\bar{z} \in \mathcal{M}} \rho(N + \mu, \bar{z})]. \end{aligned}$$

The proof of this corollary follows directly from Corollary 2.3 and Theorem 2.5 by repeated use of triangle inequalities and will therefore not be given here.

Remark 2.6. We draw the reader's attention to the fact that all the convergence results were obtained avoiding any specific information about the adjoint evolution operator or its generator. This is quite important, since the properties of the adjoint evolution operator are unfavorable to constructing approximation schemes: if $\mathcal{A}^*(t)$ denotes the infinitesimal generator of the adjoint evolution operator, then

$$\bigcap_{t \geq t_0} \text{Dom}(\mathcal{A}^*(t)) \text{ need not be dense in } Z$$

(see [6]), and for autonomous FDE $\text{Dom}(\mathcal{A}^*)$ consists of all elements $(\eta, \phi) \in Z$ with ϕ absolutely continuous on $[-r_i, r_{i-1}]$, for $i = 1, \dots, l$, and with jumps at r_i determined by A_i^* [18]. It would therefore be quite difficult to find a general procedure for constructing a sequence of operators $T_*^N(t, s)$ such that (H1) holds and which in addition satisfies properties analogous to (H1) with $T^N(t, s)$ and $T(t, s)$ replaced by $T_*^N(t, s)$ and $T(t, s)^*$, respectively. Indeed, we shall see shortly that $Z^N \subset \text{Dom}(\mathcal{A}(t))$ is a very convenient property for showing that $T^N(t, s)$ converges to $T(t, s)$, but the analogous hypothesis $Z^N \subset \text{Dom}(\mathcal{A}^*(t))$ will generally not be satisfied for any of the schemes that will be discussed in the following two sections. If, of course, strong convergence of the family $T_*^N(t, s)$ to $T(t, s)^*$ is assumed, then it is shown in [8] in the context of general evolution equations that the approximating Riccati equations converge strongly rather than weakly. Moreover, [8] contains approximation theorems that are very similar in spirit to the results of this paper. But no rates of convergence are discussed and here we concentrate on the specific features of approximating the regulator problem for functional differential equations. Whereas our approximation schemes lead to a discretization of the space (i.e., delay) variable only—see (2.4) and

(3.1)—a discretization scheme for space and time variable simultaneously was discussed in [7].

3. Spline approximation schemes. In this section we apply the results of the previous one to subspaces of spline functions. There are three subsections: (α) Generalities, (β) Linear-spline functions, (γ) Cubic-spline functions.

(α) Generalities. Spline approximations for FDE have been developed in [5] and we shall use these results here. Throughout, we assume (2.3) to be autonomous, so that A_i , $i = -1, 0, \dots$, are independent of t . We recall that in this case the solution evolution operator becomes a semigroup via $T(t, s)z = T(t - s)z$ for $z \in Z$ and $(t, s) \in \Delta$, whose infinitesimal generator \mathcal{A} is given by $\mathcal{A}(\phi(0), \phi) = (L(\phi), \dot{\phi})$, where $\text{Dom}(\mathcal{A}) = \{(\eta, \phi) | \phi \in W^{1,2}(-r, 0; \mathbb{R}^n), \phi(0) = \eta\}$. We specify a weighting function for the norm of Z by

$$g(s) = j \quad \text{for } s \in [-r_{l-j+1}, -r_{l-j}], \quad \text{for } j = 1, \dots, l.$$

Obviously Z_1 and Z_g are equivalent Hilbert spaces, since $l^{-1/2}|(\eta, \phi)|_{Z_g} \leq |(\eta, \phi)_{Z_1}| \leq |(\eta, \phi)|_{Z_g}$. We continue to drop the subscript Z_g if only the set-theoretic or topological structure of Z is important. The need for introducing a weighting function when studying functional differential equations with semigroup theoretic means is well known since [19]; it guarantees dissipativity of \mathcal{A} in Z_g . Dissipativity is essential in the proof of the following theorem, for example, and means that $\langle z, \mathcal{A}z \rangle_{Z_g} \leq \omega |z|_{Z_g}$ for some $\omega \geq 0$ and all $z \in Z$. For a more detailed discussion from which it can be seen that not the exact shape of g but only the jumps at $-r_i$ are essential, we refer to [2, p. 185]. Next we repeat a general result from [5], and call $\{Z^N, P_g^N, \mathcal{A}^N\}$, $N = 1, 2, \dots$ an approximation scheme if $\{Z^N\}$ is a sequence of closed linear subspaces of Z_g , $\{P_g^N\}$ is the sequence of orthogonal projections, $P_g^N: Z_g \rightarrow Z^N$ and $\{\mathcal{A}^N\}$ is a sequence of operators $Z_g \rightarrow Z^N$.

THEOREM 3.1. *Let $\{Z^N, P_g^N, \mathcal{A}^N\}$ be an approximation scheme satisfying*

- (i) $Z^N \subset \text{Dom}(\mathcal{A})$, $N = 1, 2, \dots$;
- (ii) $\mathcal{A}^N = P_g^N \mathcal{A} P_g^N$, $N = 1, 2, \dots$;
- (iii) (a) $\lim_{N \rightarrow \infty} P_g^N z = z$ in Z for all $z \in Z$,
 (b) for some integer $k \geq 1$ we have $\lim_{N \rightarrow \infty} L(\psi^N) = L(\psi)$ in \mathbb{R}^n and $\lim_{N \rightarrow \infty} (\psi^N)' = \psi'$ in L^2 for all $\psi \in C^k$, where ψ^N is defined by $P_g^N \hat{\psi} = (\psi^N(0), \psi^N)$.

Then each \mathcal{A}^N is the infinitesimal generator of a C_0 -semigroup $T^N(t)$, $t \geq 0$, such that

$$T^N(t)Z^N \subset Z^N, \quad N = 1, 2, \dots, \quad t \geq 0,$$

$$\lim_{N \rightarrow \infty} T^N(t)z = T(t)z \quad \text{in } Z,$$

$$\|T^N(t)\|_{Z_g} \leq e^{\tilde{\omega}t}, \quad \|T(t)\|_{Z_g} \leq e^{\tilde{\omega}t},$$

where

$$\tilde{\omega} = \frac{l+1}{2} + \|A_0\| + \frac{1}{2} \sum_{i=1}^l \|A_i\|^2 + \frac{1}{2} \int_{-r}^0 \|A(s)\|^2 ds.$$

We carefully avoided \mathcal{A}^* , the adjoint of \mathcal{A} ; however we shall need

LEMMA 3.1. *The infinitesimal generators \mathcal{A}_N^* of $(T^N(t))^*$, the adjoint semigroup of $T^N(t)$ generated by $P_g^N \mathcal{A} P_g^N$, are given by*

$$\text{Dom}(\mathcal{A}_N^*) = Z$$

and

$$\mathcal{A}_N^* = (P_g^N \mathcal{A} P_g^N)^*.$$

Proof. Since $\mathcal{A}P_g^N$ is closed and defined on Z it is bounded; therefore $P_g^N \mathcal{A} P_g^N$ is bounded and so is $(P_g^N \mathcal{A} P_g^N)^*$, and $\text{Dom}(P_g^N \mathcal{A} P_g^N)^* = Z$. The second claim follows from general semigroup theory [12, p. 277]. \square

We need one more condition on the subspaces Z^N :

$\dim Z^N = k_N < \infty$ and for each N there exists a nontrivial sequence $\mu_k = \mu_k(N)$ such that $Z^N \subset Z^{N+\mu_k(N)}$ for $k = 1, 2, \dots$.

We now turn to a discussion of the variation-of-constants formula (2.9), the feedback laws (2.13) and (2.15) and the Riccati integral equation (2.17); the assumptions used in the rest of this section are (H3), (H4) and the assumptions of Theorem 3.1.

The fact that by (H3) the columns of Q^N are in Z^N , together with (H4) and $T^N(t)Z^N \subset Z^N$ imply that for each N there exist integers μ such that the right-hand side of (2.9) is in the finite dimensional subspace $Z^{N+\mu}$, and therefore

$$(3.1) \quad \begin{aligned} \dot{z}^{N,\mu}(t) &= \mathcal{A}^{N+\mu} z^{N,\mu}(t) + Q^N B(t) u(t), \quad t_0 \leq t \leq t_f, \\ z^{N,\mu}(t_0) &= P_g^{N+\mu} z. \end{aligned}$$

By a similar argument we find that $\Pi^{N,\mu}(\cdot)$ satisfies in $Z^{N+\mu}$ the Riccati differential equation

$$(3.2) \quad \begin{aligned} \frac{d\Pi^{N,\mu}(t)}{dt} &= -(\mathcal{A}^{N+\mu})^* \Pi^{N,\mu}(t) - \Pi^{N,\mu}(t) \mathcal{A}^{N+\mu} \\ &\quad - [\mathcal{D}^{N+\mu}(t) - \Pi^{N,\mu}(t) \mathcal{B}^N(t) C^{-1}(t) (\mathcal{B}^N)^*(t) \Pi^{N,\mu}(t)] \quad \text{for } t_0 \leq t \leq t_f, \\ \Pi^{N,\mu}(t_f) &= F^{N+\mu}. \end{aligned}$$

We also recall the feedback law

$$(3.3) \quad \tilde{u}^{N,\mu}(t) = -C^{-1}(t) (\mathcal{B}^N)^*(t) \Pi^{N,\mu}(t) S^{N,\mu}(t, t_0) P_g^{N+\mu} z,$$

where $S^{N,\mu}(\cdot, t_0) P_g^{N+\mu} z$ is the optimal trajectory corresponding to $(\mathcal{P}^{N,\mu})$.

To approximate (\mathcal{P}) by the finite dimensional problems $(\mathcal{P}^{N,\mu})$ we yet have to express the various operators in (3.1)–(3.3) with respect to some bases in Z^N .

Remark 3.1. For the reader who cares to follow the calculations carried out in this subsection, or the potential applicant of the resulting finite-dimensional linear quadratic control problem, it might be helpful to think of \mathbb{R}^n -vectors as $n \times n$ diagonal-valued matrices.

For each $N = 1, 2, \dots$, we now choose a basis $(\hat{\beta}_1^N, \dots, \hat{\beta}_{k_N}^N)$ of Z^N . From $Z^N \subset \text{Dom}(\mathcal{A})$ it follows that $\hat{\beta}_i^N = (\beta_i^N(0), \beta_i^N)$ for $i = 1, \dots, k_N$, with $\beta_i^N \in W^{1,2}$. We shall need the matrix functions

$$\beta^N = (\beta_1^N, \dots, \beta_{k_N}^N)$$

and

$$\hat{\beta}^N = (\hat{\beta}_1^N, \dots, \hat{\beta}_{k_N}^N).$$

Each element $z^N \in Z^N$ can be expressed as

$$z^N = \hat{\beta}^N \alpha^N \quad \text{for } \alpha^N = \text{col}(\alpha_1^N, \dots, \alpha_{k_N}^N) \in \mathbb{R}^{k_N}$$

or in terms of elements as

$$z^N = \left(\sum_{i=1}^{k_N} \beta_i^N(0) \alpha_i^N, \sum_{i=1}^{k_N} \beta_i^N \alpha_i^N \right).$$

The matrix representation of $\mathcal{A}^N : Z^N \rightarrow Z^N$, denoted by A^N and the coordinate vector of $P_g^N z$, for $z = (\eta, \phi) \in Z$ have been calculated in [5]. To present this result, which is a simple consequence of $P_g^N(\eta, \phi) - (\eta, \phi) \perp Z_g^N$, we define the matrices

$$J^N = \langle \hat{\beta}^N, \hat{\beta}^N \rangle_{Z_g} \stackrel{\text{def}}{=} \beta^N(0)^* \beta^N(0) + \int_{-r}^0 \beta^N(s)^* \beta^N(s) g(s) ds,$$

$$h^N(\eta, \phi) = \langle \hat{\beta}^N, (\eta, \phi) \rangle_{Z_g} \stackrel{\text{def}}{=} \beta^N(0)^* \eta + \int_{-r}^0 \beta^N(s)^* \phi(s) g(s) ds$$

and

$$(3.4) \quad H^N = h^N(L(\beta^N), \dot{\beta}^N) = \beta^N(0)^* L(\beta^N) + \int_{-r}^0 \beta^N(s)^* \dot{\beta}^N(s) g(s) ds.$$

Then, if $P_g^N(\eta, \phi) = \hat{\beta}^N \alpha^N$, the coordinate vector α^N is given by

$$(3.5) \quad \alpha^N = (J^N)^{-1} h^N(\eta, \phi)$$

and the matrix representation of \mathcal{A}^N by

$$(3.6) \quad A^N = (J^N)^{-1} H^N.$$

If for any N one chooses μ satisfying (H4), then the columns of Q^N are in $Z^{N+\mu}$ and there exists a vector $\delta^{N,\mu} \in \mathbb{R}^{k_{N+\mu}}$ such that

$$(3.7) \quad Q^N = \hat{\beta}^{N+\mu} \delta^{N,\mu}.$$

Since we think of the approximation in N as chosen by the user according to some desired accuracy which can be achieved by fixing N sufficiently large and then by letting $\mu \rightarrow \infty$, the following formula will be useful:

$$(3.8) \quad \delta^{N,\mu} = (J^{N+\mu})^{-1} \langle \hat{\beta}^{N+\mu}, \hat{\beta}^N \rangle_{Z_g} \delta^{N,0}.$$

Next, we turn to $(Q^{N+\mu})^* : Z^{N+\mu} \rightarrow \mathbb{R}^n$. For $\hat{\phi}^{N+\mu} = (\phi^{N+\mu}(0), \phi^{N+\mu}) \in Z^{N+\mu}$, define $\gamma^{N+\mu} \in \mathbb{R}^{k_{N+\mu}}$, and $\eta \in \mathbb{R}^n$ by

$$\hat{\phi}^{N+\mu} = \hat{\beta}^{N+\mu} \gamma^{N+\mu} \quad \text{and} \quad (Q^{N+\mu})^* \hat{\phi}^{N+\mu} = \eta.$$

Then for all $x \in \mathbb{R}^n$

$$\langle Q^{N+\mu} x, \hat{\phi}^{N+\mu} \rangle_{Z_g} = (x, (Q^{N+\mu})^* \hat{\phi}^{N+\mu})$$

or, equivalently

$$\langle \hat{\beta}^{N+\mu} \delta^{N,\mu} x, \hat{\phi}^{N+\mu} \gamma^{N+\mu} \rangle_{Z_g} = (x, \eta).$$

We use Remark 3.1 and easily deduce

$$(\delta^{N,\mu})^* J^{N+\mu} \gamma^{N+\mu} = \eta.$$

Therefore, the matrix representation $[(Q^N)^*]$ of Q^{N*} is given by

$$(3.9) \quad [(Q^N)^*] = (\delta^{N,\mu})^* J^{N+\mu} = (\delta^{N,0})^* \langle \hat{\beta}^N, \hat{\beta}^{N+\mu} \rangle_{Z_g}.$$

Let A_*^N denote the matrix representation of $(P_g^N \mathcal{A} P_g^N)^*$; then

$$(3.10) \quad A_*^N = (J^N)^{-1} (A^N)^* J^N.$$

We notice, of course, that if $\hat{\beta}_1, \dots, \hat{\beta}_{k_N}$ were an orthonormal basis then A_*^N would equal $(A^N)^*$. To find the matrix representation $[D^N(t)]$ of $\mathcal{D}^N(t) = P_g^N \mathcal{D}(t) P_g^N$ we let $\hat{\phi}^N = (\phi^N(0), \phi^N) \in Z^N$ and define $\alpha^N \in \mathbb{R}^{k_N}$ and $\gamma^N \in \mathbb{R}^{k_N}$ by

$$\hat{\phi}^N = \hat{\beta}^N \alpha^N \quad \text{and} \quad \mathcal{D}^N(t) \hat{\phi}^N = \hat{\beta}^N \gamma^N.$$

Therefore

$$\mathcal{D}^N(t) \hat{\phi}^N = P_g^N \mathcal{D}(t) \hat{\phi}^N = P_g^N (D(t) \phi^N(0), 0)$$

and by (3.5)

$$\gamma^N = (J^N)^{-1} h^N(D(t) \phi^N(0), 0).$$

But

$$h^N(D(t) \phi^N(0), 0) = h^N(D(t) \beta^N(0) \alpha^N, 0) = \tilde{D}^N(t) \alpha^N,$$

where

$$\tilde{D}^N(t) = \beta^N(0)^* D(t) \beta^N(0),$$

so that

$$\gamma^N = (J^N)^{-1} \tilde{D}^N(t) \alpha^N$$

or

$$(3.11) \quad [D^N(t)] = (J^N)^{-1} \tilde{D}^N(t).$$

In a similar manner we see that the representation $[F^N]$ of $F^N = P_g^N \mathcal{F} P_g^N$ is given by

$$[F^N] = (J^N)^{-1} \tilde{F}^N,$$

where

$$\tilde{F}^N = \beta^N(0)^* F \beta^N(0).$$

To write equations (3.1)–(3.3) in terms of the coordinates with respect to the chosen basis, we let $\pi^{N,\mu}(t)$ denote the matrix representation of the operators $\Pi^{N,\mu}(t): Z^N \rightarrow Z^N$ and let $w^{N,\mu}(t) = w^{N,\mu}(t; u)$ and $w_0^{N,\mu}$ be defined by

$$z^{N,\mu}(t) = \hat{\beta}^{N+\mu} w^{N,\mu}(t) \quad \text{and} \quad P_g^{N+\mu} z = \hat{\beta}^{N+\mu} w_0^{N,\mu}.$$

Since the matrix representation of $(\mathcal{B}^N)^*$ is given by $B^* \delta^{N,\mu*} J^{N+\mu}$, we find that (3.2) is equivalent to

$$\begin{aligned} \frac{d}{dt} \pi^{N,\mu}(t) &= -A_*^{N+\mu} \pi^{N,\mu}(t) - \pi^{N,\mu}(t) A^{N+\mu} - [D^{N+\mu}(t)] \\ &\quad + \pi^{N,\mu}(t) \delta^{N,\mu} B(t) C(t)^{-1} B^*(t) \delta^{N,\mu*} J^{N+\mu} \pi^{N,\mu}(t) \end{aligned}$$

with $\pi^{N,\mu}(t_f) = [F^{N+\mu}]$. It can be seen from (3.10) that $\pi^{N,\mu}$ will not be symmetric in general, even though $\Pi^{N,\mu}$ is selfadjoint. We therefore define the nonnegative, symmetric matrix $\hat{\pi}^{N,\mu} = J^{N+\mu} \pi^{N,\mu}$. After premultiplying the last equation by $J^{N+\mu}$, we

find that (3.1)–(3.3) is equivalent to

$$\begin{aligned}
 \dot{w}^{N,\mu}(t) &= A^{N+\mu} w^{N,\mu}(t) + \delta^{N,\mu} B(t) u^{N,\mu}(t) \quad \text{for } t_0 \leq t \leq t_f, \\
 w^{N,\mu}(t_0) &= w_0^{N,\mu}, \\
 \frac{d}{dt} \hat{\pi}^{N,\mu}(t) &= -(A^{N+\mu})^* \hat{\pi}^{N,\mu}(t) - \hat{\pi}^{N,\mu}(t) A^{N+\mu} \\
 &\quad - \tilde{D}^{N+\mu}(t) + \hat{\pi}^{N,\mu}(t) \delta^{N,\mu} B(t) C(t)^{-1} B^*(t) \delta^{N,\mu*} \hat{\pi}^{N,\mu}(t), \\
 \hat{\pi}^{N,\mu}(t_f) &= \tilde{F}^{N+\mu}, \\
 u^{N,\mu}(t) &= -C^{-1}(t) B^*(t) \delta^{N,\mu*} \hat{\pi}^{N+\mu}(t) w^{N,\mu}(t).
 \end{aligned}
 \tag{3.12}$$

We close this subsection with a final remark on the choice of the operator Q^N .

Remark 3.2. The natural possibilities of choosing Q^N are

(α) either take $Q^N = P_g^N Q_0$, which by (3.5) implies that

$$\delta^{N,0} = (J^N)^{-1} \beta^N(0)^*,$$

(β) or, if the subspaces are chosen as spline functions, to take the representation of $Q^N: \mathbb{R}^n \rightarrow Z$ as the interpolating spline, which at the knot $t=0$ takes the value I (identity matrix) and 0 (zero-matrix) on the other knots.

The choice between (α) and (β) has to be made on the ground of getting the best convergence for $\tilde{\rho}(N, Q^N)$. Condition (H3) is checked in essentially the same manner for (α) and (β).

(β) Linear spline functions. We begin this subsection with a brief discussion on the rate of convergence of the approximating semigroups $T^N(t)$ constructed in Theorem 3.1. In [5] it is shown that general semigroup theory provides the following estimate:

There exists a constant $\tilde{M} = \hat{M}(t_f, A_i, \lambda_0)$ such that

$$\begin{aligned}
 |(T^N(t) - T(t))z| &\leq \hat{M} \{ |(\mathcal{A}^N - \mathcal{A})z| \\
 &\quad + \int_0^{t_f} |[\mathcal{A}^N - \mathcal{A}]T(s)(\lambda_0 I - \mathcal{A})z| ds + |[\mathcal{A}^N - \mathcal{A}]T(t)z| \}
 \end{aligned}
 \tag{3.13}$$

for all $t \in [0, t_f]$ and $z \in \text{Dom}(\mathcal{A}^2)$.

To give estimates on the rate of convergence for our problem we use (3.13) together with results from the theory of spline functions to estimate $\mathcal{A}^N \rightarrow \mathcal{A}$. In spline analysis the estimate on the rate of convergence of an interpolating spline always contains higher order derivatives of the function that it interpolates. This constitutes an essential problem for choosing Q^N and estimating $\tilde{\rho}(N, Q^N)$, since (H3) does not allow us to pick the representative of Q^N arbitrarily smooth.

Although estimate (3.13) (which was derived from general semigroup theory) might be too weak for the special case of FDE, it nevertheless clearly indicates that the “jump” operator Q_0 needs extra treatment. For spline approximation of neutral functional differential equations this has turned out to be essential, both in theory and in numerical work [9].

For the rest of this subsection, we choose $\mu = 0$ in $(\mathcal{D}^{N,\mu})$ and let $N = 0, 1, 2, \dots$.

We denote by $Z_1^N = \{(\phi(0), \phi) \in \mathcal{C} | \phi \text{ first order spline with knots at } t_j^N, j = \{0, \dots, N\}, \text{ where } t_j^N = -jr/N, j = 0, \dots, N\}$.

P_1^N stands for the orthogonal projection $Z_g \rightarrow Z_1^N$, $N = 1, \dots$. It is proved in [5] that the approximation scheme $\{Z_1^N, P_1^N, P_1^N \mathcal{A} P_1^N\}$ satisfies the hypotheses of Theorem 3.1 and that $\dim(Z_1^N) = n(N+1)$. Therefore, for each z there exists a real-valued

function $\bar{\rho}_1(N, z)$ such that

$$(3.14) \quad \lim_{N \rightarrow \infty} \bar{\rho}_1(N, z) = 0 \quad \text{and} \quad |T(t)z - T^N(t)z| \leq \bar{\rho}_1(N, z).$$

This and the estimates to follow hold on the interval $[0, t_f]$. By Theorem 3.1 and the above inequality, (H1) is trivially satisfied. Using the triangle inequality, interpolating spline functions and [16, Thm. 2.4] it follows by a simple density argument that for all $z \in Z$

$$(3.15) \quad |P_1^N z - z| \leq \tilde{\rho}_1(N, z), \quad \text{with} \quad \lim_{\mu \rightarrow \infty} \tilde{\rho}_1(N, z) = 0,$$

so that (H2) is verified. The operators Q_1^N are chosen as

$$Q_1^N = P_1^N Q_0,$$

or in terms of their representation

$$(Q_1^N)_j = P_1^N(e_j, 0),$$

where $e_j, j = 1, \dots, n$ stands for the n unit vectors in \mathbb{R}^n , and 0 for the zero function. (H3i) holds trivially and a short calculation gives

$$(3.16) \quad \|Q_1^N - Q_0\|_{\mathcal{L}(\mathbb{R}^n; Z)} \leq \sqrt{n} \max_{e_i} \tilde{\rho}_1(\mu, (e_i, 0)),$$

and

$$\|Q_1^N\|_{\mathcal{L}(\mathbb{R}^n; Z)} \leq 1.$$

Thus (H3) is verified.

Finally

$$\begin{aligned} \tilde{\rho}(N, Q_1^N) &= \|T(t)Q_1^N - T^N(t)Q_1^N\| \\ &= \|T(t)P_1^N Q_0 - T(t)Q_0\| + \|T(t)Q_0 - T^N(t)P_1^N Q_0\| \\ &\leq e^{\tilde{\omega}t_f} \|P_1^N Q_0 - Q_0\| \\ &\quad + \sqrt{n} \max_{e_i} (\bar{\rho}_1(N, (e_i, 0)) + e^{\tilde{\omega}t_f} \tilde{\rho}_1(N, (e_i, 0))) \\ &\leq 2 e^{\tilde{\omega}t_f} \sqrt{n} (\max_{e_i} \bar{\rho}_1(N, (e_i, 0)) + \max_{e_i} \tilde{\rho}_1(N, (e_i, 0))). \end{aligned} \quad (3.17)$$

Estimates (3.14)–(3.17) are exactly those needed for the convergence results of control, state, payoff and Riccati operators in § 2.

By (3.13) we know that on subspaces of Z determined by $\text{Dom}(\mathcal{A}^k)$, $k > 0$, $\bar{\rho}_1$ will actually go to zero with a rate given by convergence of the generators. But this is always at the expense of $\bar{\rho}_1$ not only depending on z but also on (at least) the L^2 -norm of its second derivative. So even if we dispense of (H3i) for a moment, high order convergence of $\|T(t)Q_0 - T^N(t)Q_0\|$ to zero seems quite unlikely in the light of Remark 2.1.

(γ) Cubic spline functions. In this subsection the general results of § 2 are used to discuss subspaces Z_3^N of Z given by

$$Z_3^N = \{(\phi(0), \phi) \in \mathcal{C}^2 \mid \phi \text{ is a cubic spline with knots at } t_j^N, j = 0, \dots, N\},$$

where again $t_j^N = -jr/N$, $j = 0, \dots, N$, and $P_3^N: Z_g \rightarrow Z_3^N$ are the orthogonal projections. It is quite simple to verify that the approximation scheme $\{Z_3^N, P_3^N, \mathcal{A}_3^N\}$ with $\mathcal{A}_3^N = P_3^N \mathcal{A} P_3^N$ satisfies the conditions of Theorem 3.1 with $\dim Z_3^N = n(N+3)$ and that for $\mu = 0$ we can derive results similar to subsection (β) . (H1) is therefore trivially satisfied. Here, however, we restrict our attention to the question of rate of convergence on subspaces of Z .

For $k = 1, 2, \dots$, we introduce

$$\mathcal{D}^k = \{(\phi(0), \phi) \in \mathcal{W}^{k,2} \mid \phi \in W^{k+1,2}, \phi^{(i)}(0) = L(\phi^{(i-1)}), i = 1, \dots, k\}.$$

Notice that \mathcal{D}^k is the domain of the infinitesimal generator of the solution semigroup of the autonomous equation (2.3) if considered in the Banach space $\mathcal{W}^{k,2}$, (with its natural norm).

In particular, this implies that

$$\mathcal{D}^k \text{ is dense in } \mathcal{W}^{k,2}, \quad k = 1, 2, \dots.$$

Moreover,

$$\text{if } z \in \mathcal{D}^k \text{ then } z \in \text{Dom}(\mathcal{A}^{k+1}).$$

In [5] it is proved that for $\hat{\psi} = (\psi(0), \psi)$

$$(3.18) \quad |T^N(t)\hat{\psi} - T(t)\hat{\psi}| \leq \bar{\rho}_3(N, \hat{\psi}) = O\left(\frac{1}{N^3}\right)$$

for $\hat{\psi} \in \mathcal{D}^5$, where $O(1/N^3)$ depends on $\psi^{(4)}$, and from [16, Thm. 6.9] it follows that

$$(3.19) \quad |P^N \hat{\psi}_1 - \hat{\psi}_1| \leq O\left(\frac{1}{N^3}\right) \quad \text{for } \psi_1 \in W^{3,2}.$$

Therefore, for cubic spline approximation and for $\psi \in \mathcal{D}^6$ the generic function ρ of § 2 satisfies

$$(3.20) \quad \rho_3(N, \hat{\psi}) = O\left(\frac{1}{N^3}\right).$$

We define the operators approximating Q_0 by

$$\hat{Q}_3^N = ((Q_3^N(0), Q_3^N)_1, \dots, (Q_3^N(0), Q_3^N)_n),$$

where Q_3^N is the $n \times n$ function-valued matrix

$$Q_3^N = \begin{pmatrix} s_3^N & & 0 \\ & \ddots & \\ 0 & & s_3^N \end{pmatrix};$$

s_3^N can be chosen very conveniently as the unique $C^2(-r, 0; \mathbb{R})$ function, given by

$$\begin{aligned} s^N(t_1^N) &= \dot{s}^N(t_1^N) = s^N(t_j^N) = 0 \quad \text{for } j = 1, \dots, N, \\ s^N(t_0^N) &= 1. \end{aligned}$$

Notice that this choice of s_3 lets the diagonal of the representative of Q_3^N become a basis function of Z^N (possibly after multiplying with some scalar) in the commonly chosen basis of cubic B -splines.

The function s_3 can be explicitly represented as

$$s_3(t) = \begin{cases} \left(\frac{N}{r}\right)^3 \left(t + \frac{r}{N}\right)^3 & \text{for } t \in \left[-\frac{r}{N}, 0\right], \\ 0 & \text{otherwise,} \end{cases}$$

and a short calculation yields

$$(3.21) \quad \rho_Q(N) = \|Q_0 - Q_3^N\|_{\mathcal{L}(\mathbb{R}^n, Z)} = O\left(\frac{1}{\sqrt{N}}\right)$$

and

$$(3.22) \quad \|Q_0 - P^N Q_0\|_{\mathcal{L}(\mathbb{R}^n, Z)} = O\left(\frac{1}{\sqrt{N}}\right).$$

The special form of \hat{Q}_3^N and (3.21) imply (H3i) and ii). (H3iii) is verified easily. If $\mu + N$ is some multiple of N then (H4) holds and (H2) is a consequence of (3.19) and density of $\mathcal{W}^{3,2}$ in Z . Finally bounds on $\tilde{\rho}(N + \mu, Q_3^N)$ are given via the following lemmas. For $\hat{\phi} = (\phi(0), \phi) \in \mathcal{C}$ let ϕ_I^N denote the interpolating cubic spline function defined by

$$\begin{aligned} \phi_I^N(t_j^N) &= \phi(t_j^N) \quad \text{for } j = 0, \dots, N, \\ (\phi_I^N)'(0) &= (\phi_I^N)'(-r) = 0. \end{aligned}$$

LEMMA 3.2. For all $\hat{\phi} \equiv (\phi(0), \phi) \in \mathcal{D}^2$ we have

$$(3.23) \quad |\mathcal{A}_3^N \hat{\phi} - \mathcal{A}_3 \hat{\phi}| = O\left(\frac{1}{N^2}\right) |\phi^{(3)}|_{L^2},$$

where $O(1/N^2)$ depends only on L and $\phi^{(3)}$ denotes the third derivative of ϕ .

Proof. Since P_3^N is an orthogonal projection

$$|P_3^N \mathcal{A} \hat{\phi} - \mathcal{A} \hat{\phi}|_{Z_R} = \min_{z \in Z_3^N} |z - \mathcal{A} \hat{\phi}|_{Z_R} \leq \sqrt{l} |(\dot{\phi})_I^N - \dot{\phi}|_{L^2},$$

where we used the fact that $\hat{\phi} \in \mathcal{D}^1$ and \sqrt{l} is a consequence of the weighting function g . The last estimate, [16, Thm. 4.5] and $\phi \in \mathcal{D}^2$ imply

$$(3.24) \quad |P_3^N \mathcal{A} \hat{\phi} - \mathcal{A} \hat{\phi}| = O\left(\frac{1}{N^2}\right) |\phi^{(3)}|_{L^2}.$$

We now turn to estimate $\mathcal{A} P_3^N \hat{\phi} - \mathcal{A} \hat{\phi}$ and let $P_3^N \hat{\phi} = (\phi^N(0), \phi^N)$. Then

$$\begin{aligned} |(\phi^N - \phi)'|_{L^2} &\leq |(\phi^N - \phi_I^N)'|_{L^2} + |(\phi_I^N - \phi)'|_{L^2} \\ &\leq C_1 N |\phi^N - \phi_I^N|_{L^2} + O\left(\frac{1}{N^2}\right) |\phi^{(3)}|_{L^2}, \end{aligned}$$

where the first term is estimated by the Schmidt inequality [16, p. 7], the second by [16, Thm. 6.9], and C_1 is a constant independent of N and ϕ .

The last inequality implies

$$(3.25) \quad |D(\phi^N - \phi)|_{L^2} = O\left(\frac{1}{N^2}\right) |\phi^{(3)}|_{L^2}.$$

A similar calculation gives

$$\sup_{s \in [-r, 0]} |\phi(s) - \phi^N(s)| = O\left(\frac{1}{N^2}\right) |\phi^{(3)}|_{L^2},$$

and therefore, since L is a bounded linear function from $C(-r, 0; \mathbb{R}^n) \rightarrow \mathbb{R}^n$,

$$(3.26) \quad |L(\phi - \phi^N)| \leq O\left(\frac{1}{N^2}\right) |\phi^{(3)}|_{L^2}.$$

(3.24)–(3.26) are used in the final estimate

$$\begin{aligned} |\mathcal{A}_3^N \hat{\phi} - \mathcal{A} \hat{\phi}| &\leq |\mathcal{A}_3^N \hat{\phi} - P_3^N \mathcal{A} \hat{\phi}| + |P_3^N \mathcal{A} \hat{\phi} - \mathcal{A} \hat{\phi}| \\ &\leq |\mathcal{A} P_3^N \hat{\phi} - \mathcal{A} \hat{\phi}| + |P_3^N \mathcal{A} \hat{\phi} - \mathcal{A} \hat{\phi}| \leq O\left(\frac{1}{N^2}\right) |\phi^{(3)}|_{L^2}, \end{aligned}$$

which ends the proof. \square

LEMMA 3.3. For all N , and $t \in [0, t_f]$

$$\|(T^\mu(t) - T(t))Q_3^N\|_{\mathcal{L}(\mathbb{R}^n; \mathcal{Z})} = O\left(\frac{1}{\mu^2}\right) |s_3^N|_{W^{3,2}(-r, 0; \mathbb{R})}.$$

Proof. For an arbitrary j let $\hat{q}^N = (q^N(0), q^N) = (Q_3^N)_j$. Notice that $\hat{q}^N \in \mathcal{W}^{3,2}(-r, 0; \mathbb{R}^n)$, so that by the density of \mathcal{D}^3 in $W^{3,2}(-r, 0; \mathbb{R}^n)$ there exists a sequence of functions $\hat{q}_n = (q_n(0), q_n) \in \mathcal{D}^3$ such that

$$(3.27) \quad q_n \rightarrow q^N \quad \text{in } W^{3,2}.$$

We turn to estimating $|T^\mu(t)\hat{q}_n - T(t)\hat{q}_n|$ first. Since $q_n \in W^{4,2}$ there exists a constant k_1 , depending only on L and t_f such that

$$(3.28) \quad |T(t)\hat{q}_n|_{W^{3,2}} \leq k_1 |q_n|_{W^{3,2}} \quad \text{for } t \in [0, t_f].$$

Using the fact that $\hat{q}_n \in \mathcal{D}^3$, it is easy to check that $(\lambda_0 I - \mathcal{A})\hat{q}_n \in \mathcal{D}^2$, for some fixed $\lambda_0 > \tilde{\omega}$. Therefore, there exists another constant \tilde{k}_2 , depending only on L and t_f such that

$$(3.29) \quad |T(t)(\lambda_0 I - \mathcal{A})\hat{q}_n|_{W^{3,2}} \leq \tilde{k}_2 |q_n|_{W^{3,2}}.$$

If we use (3.28) and (3.29) together with (3.23) in (3.13), we get

$$|T^\mu(t)\hat{q}_n - T(t)\hat{q}_n| = O\left(\frac{1}{\mu^2}\right) |q_n|_{W^{3,2}},$$

where the $O(1/\mu^2)$ -term is independent of q_n and $t \in [0, t_f]$. The last estimate, together with

$$\begin{aligned} |T^\mu(t)\hat{q}^N - T(t)\hat{q}^N| &\leq |T^\mu(t)\hat{q}^N - T^\mu(t)\hat{q}_n| + |T^\mu(t)\hat{q}_n - T(t)\hat{q}_n| + |T(t)\hat{q}_n - T(t)\hat{q}^N| \\ &\leq 2e^{\tilde{\omega}t_f} |\hat{q}^N - \hat{q}_n| + O\left(\frac{1}{\mu^2}\right) |q^N|_{W^{3,2}} + O\left(\frac{1}{\mu^2}\right) |q_n - q^N|_{W^{3,2}}, \end{aligned}$$

implies the claim.

The estimates (3.18)–(3.22) and Lemma 3.3 are exactly those estimates which are needed to apply the results of § 2 and essentially establish that cubic spline approximations to the linear-quadratic optimal control problem (P) are $O(1/\mu^2)$ convergent for trajectories, controls and payoffs, if the initial data are chosen from certain subspaces.

Remark 3.3. In this section we have discussed linear and cubic spline approximations in detail. Since these two schemes have already proved to be useful in numerical calculations (see [3], [5], [9]), we have given preference to studying two specific schemes rather than the general case for odd order spline approximations. The results in [5] also demonstrate that cubic spline approximations lead to much better results than linear splines or averaging approximations in the approximation of $T(t)$ not only theoretically but also on the computer. Of course, the basis elements of cubic splines are much more coupled than those of linear splines or averaging approximations and consequently cubic splines require more numerical efforts. Quintic splines have not been used for approximating functional differential equations yet. For higher order odd spline functions of degree $2N - 1$ one can use (3.13) together with error estimates well known in the theory of spline functions [13, p. 113] to show $O(2N - 1)$ -convergence of the unperturbed states $T^N(t)z$ to $T(t)z$. In the approximation of the optimal controls, optimal trajectories, etc., one power gets lost due to the additional difficulties arising from Q_0 in the variation of constants formula and convergence will basically be of order $O(2N - 2)$ in addition to the need to deal with the double limit process. We shall investigate our results numerically and expect to get good results for linear and cubic splines.

4. The averaging approximation scheme. When applying the results of § 2 to averaging approximation schemes, the approximating state and Riccati equations are found to be of particularly simple structure; moreover, for the class of problems under consideration, we find exactly those equations approximating problem (P) that were first proposed in [15] and [17]; they were also discussed in [4].

For any positive integer N we partition the interval $[-r, 0]$ into the subintervals $[t_j^N, t_{j-1}^N]$ with $t_j^N = -jr/N$, for $j = 0, \dots, N$. Let χ_j^N denote the characteristic function of $[t_j^N, t_{j-1}^N)$ for $j = 1, \dots, N$. Then the averaging approximation Z_{av}^N of Z are defined by

$$Z_{av}^N = \left\{ (\eta, \phi) \mid \eta \in \mathbb{R}^n, \phi = \sum_{j=1}^N v_j^N \chi_j^N, v_j^N \in \mathbb{R}^n \right\}.$$

We note that $(\eta, 0) \in Z_{av}^N$ for each $\eta \in \mathbb{R}^n$. It is simple to calculate the orthogonal projection $P_{av}^N: Z \rightarrow Z_{av}^N$; indeed for $(\eta, \phi) \in Z$ we have

$$(4.1) \quad P_{av}^N(\eta, \phi) = \left(\eta, \sum_{j=1}^N \phi_j^N \chi_j^N \right), \quad \text{where } \phi_j^N = \frac{N}{r} \int_{t_j^N}^{t_{j-1}^N} \phi(s) ds.$$

A scheme for approximating $T(t)$ using the subspaces Z_{av}^N has been derived in [2]. This "averaging approximation scheme" is described next. Again, we assume that the matrices A_i in (2.2) are independent of t and define a sequence of operators $\mathcal{A}_{av}^N: Z \rightarrow Z_{av}^N$ by

$$\mathcal{A}_{av}^N(\eta, \phi) = \left(A_0 \eta + \sum_{i=1}^l \sum_{j=1}^N A_i \phi_j^N \chi_j^N(-r_i) + \sum_{j=1}^N \frac{r}{N} D_j^N \phi_j^N, \sum_{j=1}^N \frac{N}{r} (\phi_{j-1}^N - \phi_j^N) \chi_j^N \right),$$

where

$$\phi_0^N = \eta, \quad \phi_j^N = \frac{N}{r} \int_{t_j^N}^{t_{j-1}^N} \phi(s) ds, \quad D_j^N = \frac{N}{r} \int_{t_j^N}^{t_{j-1}^N} A_{-1}(s) ds, \quad j = 1, \dots, N.$$

It was shown in [2] that if we fix the weight functions $\rho \equiv 1$, $\mathcal{A}_{\text{av}}^N$ generates semigroups $T_{\text{av}}^N(t)$ such that

$$\begin{aligned} \|T_{\text{av}}^N(t)\| &\leq M^* e^{\omega^* t} \quad \text{for } t \geq 0, \text{ with } M^* = M^*(A_i) \text{ and } \omega^* = \omega^*(A_i), \\ (4.2) \quad T_{\text{av}}^N Z_{\text{av}}^N &\subset Z_{\text{av}}^N, \quad \lim_N P_{\text{av}}^N z = z, \quad \text{for all } z \in Z, \\ |T(t)z - T_{\text{av}}^N(t)z| &\leq \bar{\rho}_{\text{av}}(N, z) \quad \text{with } \lim_N \bar{\rho}_{\text{av}}(N, z) = 0, \\ &\quad \text{for } t \text{ in compact subsets of } [0, \infty), \end{aligned}$$

so that (H1) and (H2) of § 2 are satisfied. The operators Q^N are chosen as

$$(4.3) \quad Q^N = P_{\text{av}}^N Q_0 = Q_0,$$

which, of course, implies that (H3) is trivially satisfied. By (4.1) we also have

$$(4.4) \quad F^N = \mathcal{F}, \quad \mathcal{D}^N = \mathcal{D} \quad \text{and} \quad \mathcal{B}^N = \mathcal{B} = Q_0 B.$$

Now the estimates of § 2 can be applied; for the optimal controls, for example, we get by Corollary 2.1

$$(4.5) \quad |\tilde{u} - \tilde{u}^{N,0}|_{L^2(t_0, t_f; \mathbb{R}^m)} \leq K_5^{\text{av}} [\tilde{\rho}_{\text{av}}(N, Q_0)|z| + \rho_{\text{av}}(N, z)],$$

and similarly

$$(4.6) \quad |S(t)z - S_{\text{av}}^{N,0}(t)P_{\text{av}}^N z| \leq K_6^{\text{av}} [\tilde{\rho}_{\text{av}}(N, Q_0)|z| + \rho_{\text{av}}(N, z)] \quad \text{for } t_0 \leq t \leq t_f,$$

$$(4.7) \quad \begin{aligned} &|\langle \Pi(t)z - \Pi^{N,0}(t)P^N z, y \rangle| \\ &\leq K_7^{\text{av}} [\bar{\rho}_{\text{av}}(N, y)|z| + \rho_{\text{av}}(N, z)|y| + |y||z|\tilde{\rho}(N, Q_0)], \end{aligned}$$

and by Theorem 2.2

$$(4.8) \quad |J(t_0, P^N z, \tilde{u}^N) - J(t_0, z, \tilde{u})| \leq K_8^{\text{av}} [\rho_{\text{av}}(N, z)|z| + \tilde{\rho}_{\text{av}}(N, Q_0)|z|^2],$$

so that in view of (4.2), convergence of optimal controls, optimal states and payoff, as well as weak convergence of the Riccati operators are guaranteed.

Finally, we give the form of the approximating state and Riccati equations. We use e_0^N, \dots, e_N^N defined by

$$e_0^N = (1, 0), \quad e_j^N = (0, \chi_j^N), \quad j = 1, \dots, N$$

as a basis for Z^N . Since $T_{\text{av}}^N(t)$ leaves Z_{av}^N invariant, (2.9) is equivalent to

$$(4.9) \quad \begin{aligned} \dot{z}^{N,0}(t) &= A_{\text{av}}^N z^{N,0}(t) + (B(t)u^N(t), 0), \\ z^{N,0}(t_0) &= P_{\text{av}}^N(\eta, \phi), \end{aligned}$$

which, in turn, is equivalent to

$$(4.10) \quad \begin{aligned} \dot{w}^N(t) &= A_{\text{av}}^N w^N(t) + \text{col}(Bu^N(t), 0, \dots, 0) \\ w^N(t_0) &= \text{col}(\eta, \phi_1^N, \dots, \phi_N^N), \end{aligned}$$

where $z^{N,0}(t) = \sum_{j=0}^N w_j^N(t) e_j^N$, ϕ_j^N is defined in (4.1), and

$$A_{av}^N = \begin{bmatrix} A_0 & \frac{r}{N} D_1^N & 0 & \cdots & 0 & \frac{r}{N} D_{N-1}^N & A_l + \frac{r}{N} D_N^N \\ \frac{N}{r} I & -\frac{N}{r} I & 0 & \cdots & 0 & 0 & 0 \\ 0 & \frac{N}{r} I & -\frac{N}{r} I & 0 & \cdots & 0 & 0 \\ \vdots & 0 & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \frac{N}{r} I & -\frac{N}{r} I \end{bmatrix}$$

here I is the $n \times n$ identity matrix. If we let π_{av}^N denote the matrix representation of $\Pi^{N,0}: Z^N \rightarrow Z^N$, then we find by (2.17) that $\hat{\pi}_{av}^N := J_{av}^N \pi_{av}^N$ satisfies on $[t_0, t_f]$ the matrix Riccati equation

$$\begin{aligned} \frac{d}{dt} \hat{\pi}_{av}^N(t) &= -(A_{av}^N)^* \hat{\pi}_{av}^N(t) - \hat{\pi}_{av}^N(t) A_{av}^N - \tilde{D}_{av}^N + \hat{\pi}_{av}^N(t) [B(t)] C(t)^{-1} [B(t)]^* \hat{\pi}_{av}^N(t), \\ (4.11) \quad \hat{\pi}_{av}^N(t_f) &= \tilde{F}_{av}^N, \end{aligned}$$

where $(A_{av}^N)^*$, as before, denotes the transpose of A_{av}^N and

$$\tilde{D}_{av}^N = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}, \quad \tilde{F}_{av}^N = \begin{pmatrix} F & 0 \\ 0 & 0 \end{pmatrix}, \quad [B(t)] = \text{col}(B(t), 0, \dots, 0),$$

\tilde{D}_{av}^N and \tilde{F}_{av}^N being of dimension $n(N+1) \times n(N+1)$ and $[B(t)]$ of dimension $n(N+1) \times n$. In (4.11) the matrix J_{av}^N is given by $J_{av}^N = \text{diag}(I, (r/N)I, \dots, (r/N)I)$.

The optimal feedback law becomes

$$(4.12) \quad u^N(t) = -C^{-1} [B(t)]^* \hat{\pi}_{av}^N(t) w^N(t).$$

Equations (4.10)–(4.12) completely describe the approximating linear regulator problem on the finite interval $[t_0, t_f]$.

Appendix. Here we give the proofs of those results that were not verified in § 2. In addition to the conventions specified there, we let $P^N z = z^N$ for $z \in Z$ and $N = 1, 2, \dots$.

Proof of Lemma 2.4.

(a) We use (2.12) to find the following estimate for $W_{t_0}^{N,\mu}$:

$$\begin{aligned} |W^{N,\mu} z(t)|_{\mathbb{R}^m} &= |(\mathcal{B}^N(t))^* \int_{t_0}^{t_f} (T^{N+\mu}(\eta, t))^* \mathcal{D}^{N+\mu}(\eta) T^{N+\mu}(\eta, t_0) z \, d\eta| \\ &\quad + |(\mathcal{B}^N(t))^* (T^{N+\mu}(t_f, t))^* F^{N+\mu} T^{N+\mu}(t_f, t_0) z| \\ &\leq bq(t_f - t_0) \tilde{M}^2 |z| + bq f \tilde{M}^2 |z| \\ &= bq \tilde{M}^2 |z| (d(t_f - t_0) + f). \end{aligned}$$

(b) The conditions on C, D, F and (2.11) imply (b) after a short calculation.

(c) For $t \in [t_0, t_f]$ we have

$$\begin{aligned}
 & |(W_0 z)(t) - (W_0^{N,\mu} z^{N+\mu})(t)|_{\mathbb{R}^m} |\mathcal{B}(t)^* \int_t^{t_f} T^*(\eta, t) \mathcal{D}(\eta) T(\eta, t_0) z \, d\eta \\
 & \quad - \mathcal{B}^N(t)^* \int_t^{t_f} T^{N+\mu}(\eta, t)^* \mathcal{D}^{N+\mu}(\eta) T^{N+\mu}(\eta, t_0) z^{N+\mu} \, d\eta| \\
 & \quad + |\mathcal{B}(t)^* T(t_f, t)^* \mathcal{F} T(t_f, t_0) z - \mathcal{B}^N(t)^* T^{N+\mu}(t_f, t)^* F^{N+\mu}(t_f, t_0) z^{N+\mu}| \\
 & = |\mathbf{I}_w(t)| + |\mathbf{II}_w(t)|.
 \end{aligned}$$

The terms \mathbf{I}_w and \mathbf{II}_w are now estimated separately. Let $\tau(t) = \text{sgn } \mathbf{I}_w(t)$; then

$$\begin{aligned}
 |\mathbf{I}_w(t)|_{\mathbb{R}^m} & \leq (\mathbf{I}_w(t), \tau(t))_{\mathbb{R}^m} \\
 & = \left(\int_t^{t_f} T(\eta, t)^* \mathcal{D}(\eta) T(\eta, t_0) z \, d\eta, \mathcal{B}(t) \tau(t) \right) \\
 & \quad - \left(\int_t^{t_f} T^{N+\mu}(\eta, t)^* \mathcal{D}^{N+\mu}(\eta) T^{N+\mu}(\eta, t_0) z^{N+\mu} \, d\eta, \mathcal{B}^N(t) \tau(t) \right) \\
 & = \left(\int_t^{t_f} T(\eta, t)^* \mathcal{D}(\eta) T(\eta, t_0) z \, d\eta, (\mathcal{B}(t) - \mathcal{B}^N(t)) \tau(t) \right) \\
 & \quad + \left(\int_t^{t_f} T(\eta, t)^* \mathcal{D}(\eta) T(\eta, t_0) z \, d\eta \right. \\
 & \quad \left. - \int_t^{t_f} T^{N+\mu}(\eta, t)^* \mathcal{D}(\eta) T^{N+\mu}(\eta, t_0) z^{N+\mu} \, d\eta, \mathcal{B}^N(t) \tau(t) \right) \\
 & \quad + \left(\int_t^{t_f} T^{N+\mu}(\eta, t)^* (\mathcal{D}(\eta) - \mathcal{D}^{N+\mu}(\eta)) T^{N+\mu}(\eta, t_0) z^{N+\mu} \, d\eta, \mathcal{B}^N(t) \tau(t) \right) \\
 & \leq d\bar{M}^2 \int_t^{t_f} e^{\tilde{\omega}(\eta-t)} e^{\tilde{\omega}(\eta-t_0)} |z| \, d\eta \, b\rho_Q(N) \sqrt{n} \\
 & \quad + \int_t^{t_f} \langle \mathcal{D}(\eta) (T(\eta, t_0) z - T^{N+\mu}(\eta, t_0) z^{N+\mu}), T(\eta, t) \mathcal{B}^N(t) \tau(t) \rangle_Z \, d\eta \\
 & \quad + \int_t^{t_f} \langle \mathcal{D}(\eta) T^{N+\mu}(\eta, t_0) z^{N+\mu}, T(\eta, t) \mathcal{B}^N(t) \tau(t) - T^{N+\mu}(\eta, t) \mathcal{B}^N(t) \tau(t) \rangle_Z \, d\eta \\
 & \quad + \int_t^{t_f} \langle (\mathcal{D}(\eta) - \mathcal{D}^{N+\mu}(\eta)) T^{N+\mu}(\eta, t_0) z^{N+\mu}, T^{N+\mu}(\eta, t) \mathcal{B}^N(t) \tau(t) \rangle_Z \, d\eta \\
 & \leq bd\sqrt{n}\bar{M}^2 |z| \rho_Q(N) (e^{2\tilde{\omega}(t_f-t_0)} - 1) (2\tilde{\omega})^{-1} \\
 & \quad + bqd\sqrt{n}\bar{M} (e^{\tilde{\omega}(t_f-t_0)} - 1) \tilde{\omega}^{-1} [\rho(N+\mu, z) + \tilde{\rho}(N+\mu, Q^N) |z|] \\
 & \quad + bqd\sqrt{n}\bar{M}^2 (t_f - t_0) (2\tilde{\omega})^{-1} |z| \|P^{N+\mu} Q_0 - Q_0\|.
 \end{aligned}$$

Let $e(t) = \text{sgn } \mathbf{II}_w(t)$; then

$$\begin{aligned}
 |\mathbf{II}_w(t)| & \leq (\mathbf{II}_w(t), e(t))_{\mathbb{R}^m} \\
 & \leq (\mathcal{B}(t)^* T(t_f, t)^* \mathcal{F} T(t_f, t_0) z - \mathcal{B}^N(t)^* T(t_f, t)^* \mathcal{F} T(t_f, t_0) z, e(t)) \\
 & \quad + (\mathcal{B}^N(t)^* T(t_f, t)^* \mathcal{F} T(t_f, t_0) z - \mathcal{B}^N(t)^* T^{N+\mu}(t_f, t)^* \mathcal{F} T^{N+\mu}(t_f, t_0) z^{N+\mu}, e(t)) \\
 & \quad + (\mathcal{B}^N(t)^* T^{N+\mu}(t_f, t)^* (\mathcal{F} T^{N+\mu}(t_f, t_0) z^{N+\mu} - F^{N+\mu} T^{N+\mu}(t_f, t_0) z^{N+\mu}), e(t))
 \end{aligned}$$

$$\begin{aligned}
&= \langle T^*(t_f, t) \mathcal{F} T(t_f, t) z, (\mathcal{B}(t) - \mathcal{B}^N(t)) e(t) \rangle \\
&\quad + \langle \mathcal{F} T(t_f, t_0) z, T(t_f, t) \mathcal{B}^N(t) e(t) \rangle - \langle \mathcal{F} T^{N+\mu}(t_f, t_0) z^{N+\mu}, T^{N+\mu}(t_f, t) \mathcal{B}^N(t) e(t) \rangle \\
&\quad + \langle (\mathcal{F} - P^{N+\mu} \mathcal{F}) T^{N+\mu}(t_f, t_0) z^{N+\mu}, T^{N+\mu}(t_f, t) \mathcal{B}^N(t) e(t) \rangle \\
&= bf\sqrt{n} |z| \tilde{M}^2 \rho_Q(N) \\
&\quad + \langle \mathcal{F} T(t_f, t_0) z - \mathcal{F} T^{N+\mu}(t_f, t_0) z^{N+\mu}, T(t_f, t) \mathcal{B}^N(t) e(t) \rangle \\
&\quad + \langle \mathcal{F} T^{N+\mu}(t_f, t_0) z^{N+\mu}, (T(t_f, t) - T^{N+\mu}(t_f, t)) \mathcal{B}^N(t) e(t) \rangle \\
&\quad + \|\mathcal{F} - P^{N+\mu} \mathcal{F}\| \tilde{M}^2 bq\sqrt{n} |z| \\
&\leq bf\sqrt{n} |z| \tilde{M}^2 \rho_Q(N) + bq\sqrt{n} f \tilde{M} [\rho(N + \mu, z) + |z| \tilde{\rho}(N + \mu, Q^N)] \\
&\quad + bq\sqrt{n} |z| \tilde{M}^2 f \|P^{N+\mu} Q_0 - Q_0\|.
\end{aligned}$$

The two inequalities together imply

$$\begin{aligned}
&\sup_{t \in [t_0, t_f]} (|I_w(t)| + |II_w(t)|) \\
&\leq \rho_Q(N) bf\sqrt{n} \tilde{M}^2 |z| [(2\bar{\omega})^{-1} d + f] + \tilde{\rho}(N + \mu, Q^N) bq\sqrt{n} \tilde{M} |z| (d\bar{\omega}^{-1} + f) \\
&\quad + \rho(N + \mu, z) bq\sqrt{n} \tilde{M} (d\bar{\omega}^{-1} + f) \\
&\quad + \|P^{N+\mu} Q_0 - Q_0\| bq\sqrt{n} \tilde{M}^2 |z| [d(t_f - t_0)(2\bar{\omega})^{-1} + f].
\end{aligned}$$

(d) For $w \in L^2(t_0, t_f; \mathbb{R}^m)$ we next estimate

$$\begin{aligned}
&|((V_0 - V_0^{N,\mu})w)(t)| \\
&= |\mathcal{B}(t)^* T(t_f, t)^* \mathcal{F} \int_t^{t_f} T(t_f, \sigma) \mathcal{B}(\sigma) w(\sigma) d\sigma \\
&\quad - \mathcal{B}^N(t)^* T^{N+\mu}(t_f, t)^* F^{N+\mu} \int_{t_0}^{t_f} T^{N+\mu}(t_f, \sigma) \mathcal{B}^N(\sigma) w(\sigma) d\sigma| \\
&\quad + |\mathcal{B}(t)^* \int_t^{t_f} T(\eta, t)^* \mathcal{D}(\eta) \int_{t_0}^t T(\eta, \sigma) \mathcal{B}(\sigma) w(\sigma) d\sigma d\eta \\
&\quad - \mathcal{B}^N(t)^* \int_t^{t_f} T^{N+\mu}(\eta, t)^* \mathcal{D}^{N+\mu}(\eta) \int_{t_0}^t T^{N+\mu}(\eta, \sigma) \mathcal{B}^N(\sigma) w(\sigma) d\sigma d\eta| \\
&= |I_v(t) + II_v(t)|.
\end{aligned}$$

Again, we estimate I_v and II_v separately. We let $e(t) = \text{sgn } I_v(t)$; then

$$\begin{aligned}
|I_v(t)| &= (\mathcal{B}(t)^* T(t_f, t)^* \mathcal{F} \int_{t_0}^{t_f} T(t_f, \sigma) \mathcal{B}(\sigma) w(\sigma) d\sigma \\
&\quad - \mathcal{B}^N(t)^* T(t_f, t)^* \mathcal{F} \int_{t_0}^{t_f} T(t_f, \sigma) \mathcal{B}^N(\sigma) w(\sigma) d\sigma, e(t)) \\
&\quad + (\mathcal{B}^N(t)^* T(t_f, t)^* \mathcal{F} \int_{t_0}^{t_f} T(t_f, \sigma) \mathcal{B}^N(\sigma) w(\sigma) d\sigma, e(t)) \\
&\quad - (\mathcal{B}^N(t)^* T^{N+\mu}(t_f, t)^* \mathcal{F} \int_{t_0}^{t_f} T^{N+\mu}(t_f, \sigma) \mathcal{B}^N(\sigma) w(\sigma) d\sigma, e(t)) \\
&\quad + (\mathcal{B}^N(t)^* T^{N+\mu}(t_f, t)^* (\mathcal{F} - F^{N+\mu}) \int_{t_0}^{t_f} T^{N+\mu}(t_f, \sigma) \mathcal{B}^N(\sigma) w(\sigma) d\sigma, e(t))
\end{aligned}$$

$$\begin{aligned}
 &= \langle T(t_f, t)^* \mathcal{F} \int_t^{t_f} T(t_f, \sigma) \mathcal{B}(\sigma) w(\sigma) d\sigma, (\mathcal{B}(t) - \mathcal{B}^N(t))e(t) \rangle \\
 &\quad + \langle T(t_f, t)^* \mathcal{F} \int_{t_0}^{t_f} T(t_f, \sigma) (\mathcal{B}(\sigma) - \mathcal{B}^N(\sigma)) w(\sigma) d\sigma, \mathcal{B}^N(t)e(t) \rangle \\
 &\quad + \langle \mathcal{F} \int_{t_0}^{t_f} T(t_f, \sigma) \mathcal{B}^N(\sigma) w(\sigma) d\sigma, T(t_f, t) \mathcal{B}^N(t)e(t) \rangle \\
 &\quad - \langle \mathcal{F} \int_{t_0}^{t_f} T^{N+\mu}(t_f, \sigma) \mathcal{B}^N(\sigma) w(\sigma) d\sigma, T^{N+\mu}(t_f, t) \mathcal{B}^N(t)e(t) \rangle \\
 &\quad + \langle (\mathcal{F} - F^{N+\mu}) \int_{t_0}^{t_f} T^{N+\mu}(t_f, \sigma) \mathcal{B}^N(\sigma) w(\sigma) d\sigma, T^{N+\mu}(t_f, t) \mathcal{B}^N(t)e(t) \rangle \\
 &\leq b^2 f \sqrt{n} \bar{M}^2 (1+q) \rho_Q(N) e^{\bar{\omega}(t_f-t)} \left[\frac{1}{2\bar{\omega}} (e^{2\bar{\omega}(t_f-t_0)} - 1) \right]^{1/2} |w|_{L^2} \\
 &\quad + \langle \mathcal{F} \int_{t_0}^{t_f} T(t_f, \sigma) \mathcal{B}^N(\sigma) w(\sigma) d\sigma, (T(t_f, t) - T^{N+\mu}(t_f, t)) \mathcal{B}^N(t)e(t) \rangle \\
 &\quad + \langle \mathcal{F} \int_{t_0}^{t_f} (T(t_f, \sigma) - T^{N+\mu}(t_f, \sigma)) \mathcal{B}^N(\sigma) w(\sigma) d\sigma, T^{N+\mu}(t_f, t) \mathcal{B}^N(t)e(t) \rangle \\
 &\quad + \|\mathcal{F} - P^{N+\mu} \mathcal{F}\| b^2 q^2 \sqrt{n} \bar{M}^2 e^{\bar{\omega}(t_f-t_0)} \left[\frac{e^{2\bar{\omega}(t_f-t_0)} - 1}{2\bar{\omega}} \right]^{1/2} |w|_{L^2} \\
 &\leq b^2 f \sqrt{n} \tilde{M}^2 (1+q) \rho_Q(N) \frac{1}{\sqrt{2\bar{\omega}}} |w|_{L^2} \\
 &\quad + \bar{M} b^2 q f \sqrt{n} \tilde{\rho}(N+\mu, Q^N) |w|_2 \left[\left(\frac{e^{2\bar{\omega}(t_f-t_0)} - 1}{2\bar{\omega}} \right)^{1/2} + e^{\bar{\omega}(t_f-t_0)} (t_f - t_0)^{1/2} \right] \\
 &\quad + \|Q_0 - P^{N+\mu} Q_0\| f b^2 q^2 \sqrt{n} \tilde{M}^2 \frac{1}{\sqrt{2\bar{\omega}}} |w|_{L^2}.
 \end{aligned}$$

Finally, with $v(t) = \text{sgn } \Pi_v(t)$ we get

$$\begin{aligned}
 |\Pi_v(t)| &= (\Pi_v(t), v(t)) \\
 &= \left(\mathcal{B}(t)^* \int_t^{t_f} T(\eta, t)^* \mathcal{D}(\eta) \int_{t_0}^\eta T(\eta, \sigma) \mathcal{B}(\sigma) w(\sigma) d\sigma d\eta \right. \\
 &\quad \left. - \mathcal{B}^N(t)^* \int_t^{t_f} T^{N+\mu}(\eta, t)^* \mathcal{D}^{N+\mu}(\eta) \int_{t_0}^\eta T^{N+\mu}(\eta, \sigma) \mathcal{B}^N(\sigma) w(\sigma) d\sigma d\eta, v(t) \right) \\
 &= \left(\mathcal{B}(t)^* \int_t^{t_f} T(\eta, t)^* \mathcal{D}(\eta) \int_{t_0}^\eta T(\eta, \sigma) \mathcal{B}(\sigma) w(\sigma) d\sigma d\eta \right. \\
 &\quad \left. - \mathcal{B}^N(t)^* \int_t^{t_f} T(\eta, t)^* \mathcal{D}(\eta) \int_{t_0}^\eta T(\eta, \sigma) \mathcal{B}^N(\sigma) w(\sigma) d\sigma d\eta, v(t) \right) \\
 &\quad + \left(\mathcal{B}^N(t)^* \int_t^{t_f} T(\eta, t)^* \mathcal{D}(\eta) \int_{t_0}^\eta T(\eta, \sigma) \mathcal{B}^N(\sigma) w(\sigma) d\sigma d\eta \right. \\
 &\quad \left. - \mathcal{B}^N(t)^* \int_t^{t_f} T^{N+\mu}(\eta, t)^* \mathcal{D}(\eta) \int_{t_0}^\eta T^{N+\mu}(\eta, \sigma) \mathcal{B}^N(\sigma) w(\sigma) d\sigma d\eta, v(t) \right)
 \end{aligned}$$

$$\begin{aligned}
& + \left(\mathcal{B}^N(t)^* \int_t^{t_f} T^{N+\mu}(\eta, t)^* \right. \\
& \quad \times (\mathcal{D}(\eta) - \mathcal{D}^{N+\mu}(\eta)) \int_{t_0}^{\eta} T^{N+\mu}(\eta, \sigma) \mathcal{B}^N(\sigma) w(\sigma) d\sigma d\eta, v(t) \Big) \\
& = \left\langle \int_t^{t_f} T(\eta, t)^* \mathcal{D}(\eta) \int_{t_0}^{\eta} T(\eta, \sigma) (\mathcal{B}(\sigma) - \mathcal{B}^N(\sigma)) w(\sigma) d\sigma d\eta, \mathcal{B}(t) v(t) \right\rangle \\
& \quad + \left\langle \int_t^{t_f} T(\eta, t)^* \mathcal{D}(\eta) \int_{t_0}^{\eta} T(\eta, \sigma) \mathcal{B}^N(\sigma) w(\sigma) d\sigma d\eta, (\mathcal{B}(t) - \mathcal{B}^N(t)) v(t) \right\rangle \\
& \quad + \int_t^{t_f} \left\{ \left\langle \mathcal{D}(\eta) \int_{t_0}^{\eta} T(\eta, \sigma) \mathcal{B}^N(\sigma) w(\sigma) d\sigma, (T(\eta, t) - T^{N+\mu}(\eta, t)) \mathcal{B}^N(t) v(t) \right\rangle \right. \\
& \quad \left. + \left\langle \mathcal{D}(\eta) \int_{t_0}^{\eta} (T(\eta, \sigma) - T^{N+\mu}(\eta, \sigma)) \mathcal{B}^N(\sigma) w(\sigma) d\sigma, \right. \right. \\
& \quad \left. \left. T^{N+\mu}(\eta, t) \mathcal{B}^N(t) v(t) \right\rangle \right\} d\eta \\
& \quad + \int_t^{t_f} \left\langle (\mathcal{D}(\eta) - \mathcal{D}^{N+\mu}(\eta)) \right. \\
& \quad \left. \times \int_{t_0}^{\eta} T^{N+\mu}(\eta, \sigma) \mathcal{B}^N(\sigma) w(\sigma) d\sigma, T^{N+\mu}(\eta, t) \mathcal{B}^N(t) v(t) \right\rangle d\eta \\
& \leq b^2 d \sqrt{n} \bar{M}^2 \rho_Q(N) |w|_{L^2} \frac{1}{\sqrt{2\bar{\omega}^3}} [e^{2\bar{\omega}(t_f-t_0)} - 1] (1+q) \\
& \quad + \int_t^{t_f} d \left(\int_{t_0}^{\eta} \bar{M} e^{\bar{\omega}(\eta-\sigma)} b q |w(\sigma)| d\sigma \right) \tilde{\rho}(N+\mu, Q^N) b q \sqrt{n} d\eta \\
& \quad + \int_t^{t_f} d \left(\int_{t_0}^{\eta} \tilde{\rho}(N+\mu, Q^N) b q |w(\sigma)| d\sigma \right) \bar{M} e^{\bar{\omega}(\eta-t)} b q \sqrt{n} d\eta \\
& \quad + d \|Q_0 - P^{N+\mu} Q_0\| \int_t^{t_f} \int_{t_0}^{\eta} \bar{M}^2 e^{\bar{\omega}(\eta-\sigma)} b^2 q^2 |w(\sigma)| d\sigma e^{\bar{\omega}(\eta-t)} \sqrt{n} d\eta \\
& \leq b^2 d \sqrt{n} \bar{M}^2 \rho_Q(N) |w|_{L^2} \frac{1}{\sqrt{2\bar{\omega}^3}} [e^{2\bar{\omega}(t_f-t_0)} - 1] (1+q) \\
& \quad + db^2 q^2 \bar{M} \sqrt{n} \tilde{\rho}(N+\mu, Q^N) \int_t^{t_f} \int_{t_0}^{\eta} e^{\bar{\omega}(\eta-\sigma)} |w(\sigma)| d\sigma d\eta \\
& \quad + \int_t^{t_f} \int_{t_0}^{\eta} e^{\bar{\omega}(\eta-t)} |w(\sigma)| d\sigma d\eta \\
& \quad + db^2 g^2 \bar{M}^2 \sqrt{n} \|Q_0 - P^{N+\mu} Q_0\| \int_t^{t_f} \int_{t_0}^{\eta} e^{\bar{\omega}(\eta-\sigma)} e^{\bar{\omega}(\eta-t)} |w(\sigma)| d\sigma d\eta \\
& \leq b^2 d \sqrt{n} \bar{M}^2 \rho_Q(N) |w|_{L^2} \frac{1}{\sqrt{2\bar{\omega}^3}} (1+q) \\
& \quad + db^2 q^2 \bar{M} \sqrt{n} \tilde{\rho}(N+\mu, Q^N) |w|_{L^2} \left[\frac{1}{\sqrt{2\bar{\omega}^2}} e^{\bar{\omega}(t_f-t_0)} + \frac{(t_f-t_0)^{1/2}}{\bar{\omega}} (e^{\bar{\omega}(t_f-t_0)} - 1) \right] \\
& \quad + db^2 q^2 \bar{M}^2 \sqrt{n} \|Q_0 - P^{N+\mu} Q_0\| |w|_{L^2} \left[\frac{1}{\sqrt{2\bar{\omega}^3}} (e^{2\bar{\omega}(t_f-t_0)} - 1) \right].
\end{aligned}$$

The bounds on I_v and II_v imply

$$\begin{aligned} & \sup_{t \in [t_0, t_f]} |((V_0 - V_0^{N, \mu})w)(t)| \\ & \leq \rho_Q(N) b^2 \sqrt{n} \tilde{M}^2 (1+q) |w|_{L^2} \left(\frac{f}{\sqrt{2\bar{\omega}}} + \frac{d}{\sqrt{2\bar{\omega}^3}} \right) \\ & \quad + \tilde{\rho}(N + \mu, Q^N) b^2 \sqrt{n} \tilde{M} q^2 |w|_{L^2} \left(f \left(\frac{1}{\sqrt{2\bar{\omega}}} + (t_f - t_0)^{1/2} \right) + d \left(\frac{1}{\sqrt{2\bar{\omega}^3}} + \frac{(t_f - t_0)^{1/2}}{\bar{\omega}} \right) \right) \\ & \quad + \|Q_0 - P^{N+\mu} Q_0\| b^2 q^2 \tilde{M}^2 \sqrt{n} |w|_{L^2} \left[\frac{d}{\sqrt{2\bar{\omega}^3}} + \frac{f}{\sqrt{2\bar{\omega}}} \right]. \end{aligned}$$

This completes the proof. \square

Proof of Theorem 2.2.

$$\begin{aligned} \text{(a)} \quad |S^{N, \mu}(t, s)z| & \leq |T^{N+\mu}(t, s)z| + \left| \int_s^t T^{N+\mu}(t, \eta) \mathcal{B}^N(\eta) ((V_s^{N+\mu})^{-1} W_s^{N, \mu} z)(\eta) d\eta \right| \\ & \leq \tilde{M} |z| + \tilde{M} b q \int_s^t |((V_s^{N, \mu})^{-1} W_s^{N, \mu} z)(\eta)| d\eta \\ & \leq \tilde{M} |z| + \tilde{M} b q (t_f - t_0)^{1/2} c^{-1} (t_f - t_0)^{1/2} k_0 |z| \\ & = \tilde{M} |z| (1 + b q (t_f - t_0) c^{-1} k_0). \end{aligned}$$

(b)

$$\begin{aligned} & |S(t, s)z - S^{N, \mu}(t, s)P^{N+\mu}z| \\ & \leq |T(t, s)z - T^{N+\mu}(t, s)z^{N+\mu}| \\ & \quad + \int_s^t |T(t, \eta) \mathcal{B}(\eta) (V_s^{-1} W_s z)(\eta) - T(t, \eta) \mathcal{B}(\eta) ((V_s^{N, \mu})^{-1} W_s^{N, \mu} z^{N+\mu})(\eta)| d\eta \\ & \quad + \int_s^t |T(t, \eta) \mathcal{B}(\eta) ((V_s^{N, \mu})^{-1} W_s^{N, \mu} z^{N+\mu})(\eta) \\ & \quad \quad - T^{N+\mu}(t, \eta) \mathcal{B}^N(\eta) ((V_s^{N, \mu})^{-1} W_s^{N, \mu} z^{N+\mu})(\eta)| d\eta \\ & \leq \rho(N + \mu, z) + \int_s^t \tilde{M} e^{\bar{\omega}(t-\eta)} b | (V_s^{-1} W_s z)(\eta) - ((V_s^{N, \mu})^{-1} W_s^{N, \mu} z^{N+\mu})(\eta) | d\eta \\ & \quad + b \left(\int_s^t \|T(t, \eta) Q_0 - T^{N+\mu}(t, \eta) Q^N\|^2 d\eta \right)^{1/2} \left(\int_s^t |((V_s^{N, \mu})^{-1} W_s^{N, \mu} z^{N+\mu})(\eta)|^2 d\eta \right)^{1/2} \\ & \leq \rho(N + \mu, z) + b |\tilde{u} - \tilde{u}^{N, \mu}|_{L^2} \tilde{M} \left(\int_s^t e^{2\bar{\omega}(t-\eta)} d\eta \right)^{1/2} \\ & \quad + b \left[\left(\int_s^t \|T(t, \eta) Q_0 - T(t, \eta) Q^N\|^2 d\eta \right)^{1/2} \right. \\ & \quad \quad \left. + \left(\int_s^t \|T(t, \eta) Q^N - T^{N+\mu}(t, \eta) Q^N\|^2 d\eta \right)^{1/2} \right] c^{-1} |z| k_0 (t_f - t_0)^{1/2} \end{aligned}$$

$$\begin{aligned}
&\leq \rho(N + \mu, z) + b|\tilde{u} - \tilde{u}^{N,\mu}|_{L^2} \tilde{M} \left(\frac{1}{2\tilde{\omega}} (e^{2\tilde{\omega}(t_f - t_0)} - 1) \right)^{1/2} \\
&\quad + bc^{-1}|z|k_0(t_f - t_0)^{1/2} \left[\tilde{M}\rho_Q(N) \left(\int_s^t e^{2\tilde{\omega}(t-\eta)} d\eta \right)^{1/2} + \tilde{\rho}(N + \mu, Q^N)(t_f - t_0)^{1/2} \right] \\
&= \rho(N + \mu, z) + b|\tilde{u} - \tilde{u}^{N,\mu}|_{L^2} \tilde{M} \frac{1}{\sqrt{2\tilde{\omega}}} + bc^{-1}|z|k_0(t_f - t_0)^{1/2} \tilde{M}\rho_Q(N) \frac{1}{\sqrt{2\tilde{\omega}}} \\
&\quad + bc^{-1}|z|k_c(t_f - t_0)\tilde{\rho}(N + \mu, Q^N) \\
&= \rho(N + \mu, z) + \tilde{\rho}(N + \mu, Q^N)|z|K_4 + |\tilde{u} - \tilde{u}^{N,\mu}|_{L^2} k_6 + |z|\rho_Q(N)k_7,
\end{aligned}$$

where

$$K_4 = bc^{-1}k_0(t_f - t_0), \quad k_6 = b\tilde{M} \frac{1}{\sqrt{2\tilde{\omega}}}, \quad k_7 = bc^{-1}k_0(t_f - t_0)\tilde{M} \frac{1}{\sqrt{2\tilde{\omega}}}.$$

Proof of Theorem 2.3.

(a) By (2.16) we have for $t \in [t_0, t_f]$

$$\begin{aligned}
|\Pi^{N,\mu}(t)z| &\leq |T^{N+\mu}(t_f, t)^* F^{N+\mu} S^{N,\mu}(t_f, t)z| \\
&\quad + \left| \int_t^{t_f} T^{N+\mu}(\eta, t)^* \mathcal{D}^{N+\mu}(\eta) S^{N,\mu}(\eta, t)z d\eta \right| \\
&\leq \tilde{M}fk_5|z| + \tilde{M}dk_5|z|(t_f - t_0).
\end{aligned}$$

(b)

$$\begin{aligned}
&\langle \Pi(t)z - \Pi^{N,\mu}(t)z^{N+\mu}, y \rangle \\
&= \langle T(t_f, t)^* \mathcal{F}S(t_f, t)z - T^{N+\mu}(t_f, t)^* F^{N+\mu} S^{N,\mu}(t_f, t)z^{N+\mu}, y \rangle \\
&\quad + \int_t^{t_f} \langle T(\eta, t)^* \mathcal{D}(\eta)S(\eta, t)z - T^{N+\mu}(\eta, t)^* \mathcal{D}^{N+\mu}(\eta)S^{N,\mu}(\eta, t)z^{N+\mu}, y \rangle d\eta \\
&\leq \langle \mathcal{F}S(t_f, t)z, T(t_f, t)y - T^{N+\mu}(t_f, t)y \rangle \\
&\quad + \langle \mathcal{F}S(t_f, t)z - \mathcal{F}S^{N,\mu}(t_f, t)z^{N+\mu}, T^{N+\mu}(t_f, t)y \rangle \\
&\quad + \langle (\mathcal{F} - F^{N+\mu})S^{N,\mu}(t_f, t)z^{N+\mu}, T^{N+\mu}(t_f, t)y \rangle \\
&\quad + \int_t^{t_f} \langle (\mathcal{D}(\eta)S^{N,\mu}(\eta, t)z, T(\eta, t)y - T^{N+\mu}(\eta, t)y) \\
&\quad \quad + \langle \mathcal{D}(\eta)S(\eta, t)z - \mathcal{D}(\eta)S^{N,\mu}(\eta, t)z^{N+\mu}, T^{N+\mu}(\eta, t)y \rangle \rangle d\eta \\
&\quad + \int_t^{t_f} \langle \mathcal{D}(\eta)S^{N,\mu}(\eta, t)z^{N+\mu} - \mathcal{D}^{N+\mu}(\eta)S^{N,\mu}(\eta, t)z^{N+\mu}, T^{N+\mu}(\eta, t)y \rangle d\eta \\
&\leq fk_5|z|\tilde{\rho}(N + \mu, y) \\
&\quad + (f + d(t_f - t_0))[\tilde{M}|y|(\rho(N + \mu, z) + \rho(N + \mu, Q^N))|z|K_4 \\
&\quad \quad \quad + |\tilde{u} - \tilde{u}^{N,\mu}|_{L^2} k_6 + |z|\rho_Q(N)k_7] \\
&\quad + \|Q_0 - P^{N+\mu}Q_0\|fk_5|z|\tilde{M}|y| + dk_5|z|(t_f - t_0)\tilde{\rho}(N + \mu, y) \\
&\quad + \|Q_0 - P^{N+\mu}\|d(t_f - t_0)k_5|z|\tilde{M}|y|
\end{aligned}$$

$$\begin{aligned}
 &= [f + d(t_f - t_0)][k_5|z|\bar{\rho}(N + \mu, y) + k_5\tilde{M}|y||z|\|Q_0 - P^{N+\mu}Q_0\| \\
 &\quad + \tilde{M}|y|[\rho(N + \mu, z) + \tilde{\rho}(N + \mu, Q^N)|z|K_4 \\
 &\quad + |\tilde{u} - \tilde{u}^{N,\mu}|_{L^2}k_6 + |z|\rho_Q(N)k_7]].
 \end{aligned}$$

Proof of Corollary 2.2.

$$\begin{aligned}
 &|J(t_0, P^{N+\mu}z, \tilde{u}^{N,\mu}) - J(t_0, z, \tilde{u})| \\
 &\leq |\langle F^{N+\mu}z^{N,\mu}(t_f), z^{N,\mu}(t_f) \rangle - \langle \mathcal{F}z(t_f), z(t_f) \rangle| \\
 &\quad + \int_{t_0}^{t_f} |\langle \mathcal{D}^{N+\mu}(t)z^{N,\mu}(t), z^{N,\mu}(t) \rangle - \langle \mathcal{D}(t)z(t), z(t) \rangle| dt \\
 &\quad + \int_{t_0}^{t_f} |(C(t)\tilde{u}^{N,\mu}(t), \tilde{u}^{N,\mu}(t)) - (C(t)u(t), u(t))| dt \\
 &\leq |\langle P^{N+\mu}\mathcal{F}z^{N,\mu}(t_f) - \mathcal{F}z^{N,\mu}(t_f), z^{N,\mu}(t_f) \rangle| \\
 &\quad + (FP_1(z^{N,\mu}(t_f) - z(t_f)), P_1z^{N,\mu}(t_f)) + (FP_1z(t_f), P_1(z^{N,\mu}(t_f) - z(t_f))) \\
 &\quad + \int_{t_0}^{t_f} |\langle \mathcal{D}^{N+\mu}(t)z^{N,\mu}(t) - \mathcal{D}(t)z^{N,\mu}(t), z^{N,\mu}(t) \rangle| \\
 &\quad + \langle \mathcal{D}(t)(z^{N,\mu}(t) - z(t)), z^{N,\mu}(t) \rangle + \langle \mathcal{D}(t)z(t), z^{N,\mu}(t) - z(t) \rangle| dt \\
 &\quad + \int_{t_0}^{t_f} |(C(t)(\tilde{u}(t) - \tilde{u}^{N,\mu}(t)), \tilde{u}(t)) + (C(t)\tilde{u}^{N,\mu}(t), \tilde{u}(t) - \tilde{u}^{N,\mu}(t))| dt \\
 &= ((t_f - t_0)d + f)k_5|z|\{2[\rho(N + \mu, z) + \tilde{\rho}(N + \mu, Q^N)|z|K_4 \\
 &\quad + |\tilde{u} - \tilde{u}^{N,\mu}|_{L^2}k_6 + |z|\rho_Q(N)k_7] + \|P^{N+\mu}Q_0 - Q_0\|k_5|z|\} \\
 &\quad + 2|\tilde{u} - \tilde{u}^{N,\mu}|_{L^2}c^{-1}k_0(t_f - t_0)^{1/2}\tilde{c}|z|,
 \end{aligned}$$

which completes the proof. \square

Proof of Theorem 2.4. By Remarks 2.4 and (2.10) it follows that \tilde{u} is continuous. Moreover, by (2.15) we have

$$|\tilde{u}(t) - \tilde{u}^{N,0}(t)| = |C(t)^{-1}\mathcal{B}(t)^*(\Pi(t)S(t, t_0)z - \Pi^{N,0}(t)S^{N,0}(t, t_0)P^N z)|.$$

Let $\varepsilon^N(t) = \text{sgn } C(t)^{-1}\mathcal{B}(t)^*(\Pi(t)S(t, t_0)z - \Pi^{N,0}(t)S^{N,0}(t, t_0)P^N z)$; then

$$|\tilde{u}(t) - \tilde{u}^{N,0}(t)| \leq \langle \Pi(t)S(t, t_0)z - \Pi^{N,0}S^{N,0}(t, t_0)P^N z, \mathcal{B}(t)(C(t)^{-1})^*\varepsilon^N(t) \rangle.$$

By Corollary 2.1(c) and after an inspection of (2.16) and the proof of Theorem 2.3(b), it now follows that

$$|\tilde{u}(t) - \tilde{u}^{N,0}(t)| \leq \sup_{\xi} K_7[\bar{\rho}(N, \xi)|z| + \rho(N, z)|\xi| + |\xi||z|(\rho_Q(N) + \tilde{\rho}(N, Q_0))]$$

where the supremum is taken over $\{\mathcal{B}(t)(C(t)^{-1})^*\varepsilon^N(t) \mid t \in [t_0, t_f], N = 1, 2, \dots\}$, a relatively compact subset of Z . (H1), (H2), together with definitions (vii) and (viii) in § 2, and a simple compactness argument imply the result. \square

Acknowledgments. The author would like to thank Professor H. T. Banks for encouragement to work on this problem and for various stimulating discussions. I am also indebted to the referee whose comments led to Theorem 2.5 and Corollary 2.4.

Note added in proof. Professors H. T. Banks and G. I. Rosen have used some of the approximation schemes of this paper for actual computations. Their results will be documented in a forthcoming technical report.

REFERENCES

- [1] H. T. BANKS, *Approximation of nonlinear functional differential equation control systems*, J. Optim. Theory Appl., 29, 383–407.
- [2] H. T. BANKS AND J. A. BURNS, *Hereditary control problems: Numerical methods based on averaging approximations*, this Journal, 16 (1978), pp. 169–208.
- [3] H. T. BANKS, J. A. BURNS AND E. M. CLIFF, *A comparison of numerical methods for identification and optimization problems involving control systems with delays*, LCDS Tech. Rep. 79-7, Lefschetz Center for Dynamical Systems, Brown Univ., Providence, RI, 1979.
- [4] H. T. BANKS, J. A. BURNS, E. M. CLIFF AND P. R. THRIFT, *Numerical solutions of hereditary control problems via an approximation technique*, Tech. Rep. 75-6, Lefschetz Center for Dynamical Systems, Brown Univ., Providence, RI, 1975.
- [5] H. T. BANKS AND F. KAPPEL, *Spline approximations for functional differential equations*, J. Differential Equations, 34 (1979), pp. 496–522.
- [6] M. C. DELFOUR, *State theory of linear hereditary differential systems*, J. Math. Anal. Appl., 60 (1977), pp. 8–35.
- [7] ———, *The linear quadratic optimal control problem for hereditary differential systems: Theory and numerical solutions*, Appl. Math. Optim., 3 (1977), pp. 101–162.
- [8] J. S. GIBSON, *The Riccati integral equations for optimal control problems on Hilbert spaces*, this Journal, 17 (1979), pp. 537–565.
- [9] F. KAPPEL AND K. KUNISCH, *Spline approximations for neutral functional differential equation*, SIAM J. Numer. Anal., 18 (1981), pp. 1058–1080.
- [10] F. KAPPEL AND W. SCHAPPACHER, *Nonlinear functional differential equations and abstract integral equations*, Proc. Roy. Soc. Edinburgh Sect. A, 84 (1979), pp. 71–91.
- [11] K. KUNISCH AND W. SCHAPPACHER, *Variation of constants formulas for partial differential equations with delay*, Nonlinear Anal. Theory Meth. Appl., 5 (1981), 123–142.
- [12] R. S. PHILLIPS, *The adjoint semigroup*, Pacific J. Math., 5 (1955), pp. 269–283.
- [13] P. M. PRENTER, *Splines and Variational Methods*, Wiley-Interscience, New York, 1975.
- [14] D. C. REBER, *A finite difference technique for solving optimization problems governed by linear functional differential equations*, J. Differential Equations, 32 (1979), pp. 192–232.
- [15] D. W. ROSS AND I. FLUGGE-LOTZ, *An optimal control problem for systems with differential-difference equation dynamics*, this Journal, 7 (1969), pp. 609–623.
- [16] M. H. SCHULTZ, *Spline Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
- [17] M. A. SOLIMAN AND W. H. RAY, *Optimal feedback control for linear-quadratic systems having time delays*, Internat. J. Control, 15 (1972), pp. 609–627.
- [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] G. F. WEBB, *Functional differential equations and nonlinear semigroups in L^p -spaces*, J. Differential Equations, 20 (1976), pp. 71–89.

LEARNING ALGORITHMS FOR TWO-PERSON ZERO-SUM STOCHASTIC GAMES WITH INCOMPLETE INFORMATION: A UNIFIED APPROACH*

S. LAKSHMIVARAHAN† AND K. S. NARENDRA‡

Abstract. This paper extends recent results [Lakshmivarahan and Narendra, Math. Oper. Res., 6 (1981), pp. 379–386] in two-person zero-sum sequential games in which the players use learning algorithms to update their strategies. It is assumed that neither player knows (i) the set of strategies available to the other player or (ii) the mixed strategy used by the other player or its pure realization at any stage. The outcome of the game depends on chance and the game is played sequentially. The distribution of the random outcome as a function of the pair of pure strategies chosen by the players is also unknown to them. It is shown that if the players use a learning algorithm of the reward-penalty type, with proper choice of certain parameters in the algorithm, the expected value of the mixed strategies for both players can be made arbitrarily close to optimal strategies.

1. Introduction. Analysis of games with incomplete information [4] has received considerable attention in the game-theoretic literature [4]–[9], [13]. However, with the exception of the work by Suppes and Atkinson [10] most of the papers on the learning approach to this class of games have been confined to the literature on systems theory [1]–[3]. The recent monograph by Tsetlin [1] well summarizes the work of the Russian authors in this direction. Papers [14], [15] represent related work on this topic. In spite of this wide-spread interest, the problem of finding conditions under which a given learning algorithm will converge to the well-established game-theoretic solutions has remained open till recently. In [17] Lakshmivarahan and Narendra have shown that if the underlying game matrix has a saddle point in pure strategies and if the players use a class of learning algorithms known as linear reward-inaction algorithms [2], [3], then they will converge to the optimal pure strategies with probability as close to unity as desired.

In this paper we consider two-person zero-sum games with incomplete information where the game matrix need not have a saddle point in pure strategies. At any stage each player is allowed to use mixed strategies. The players update their mixed strategies using a linear reward-penalty algorithm (instead of the linear reward-inaction algorithm discussed in [17]). It is shown that using a penalty term which is very small compared to the reward term and by proper choice of the step length parameter, the expected value of the mixed strategy used by either player asymptotically can be made arbitrarily close to the optimal strategy dictated by game theory.

Linear reward-penalty algorithms have been extensively investigated in the literature on mathematical psychology [18]–[21]. Norman [18]–[20] has abstracted the properties of these algorithms and has developed the theory of Markov processes that evolve by small steps. This latter theory provides the necessary mathematical framework for our analysis. The linear reward-penalty algorithm with small penalty term has been recently developed in [16] for the problem of a game against nature to attain ϵ -optimality. The linear reward-inaction algorithm used in [17] and the linear reward-penalty algorithm used in this paper constitute the prototype of two distinct

* Received by the editors February 6, 1980, and in revised form Sept. 8, 1981. This research was supported in part by the National Science Foundation under grant 03664.

† School of Electrical Engineering and Computer Science, University of Oklahoma, Norman, Oklahoma 73019.

‡ Department of Engineering and Applied Science, Yale University, New Haven, Connecticut 06520.

classes of learning algorithms as known today [26]. In fact, the theory and the methods used to analyze the properties of these two algorithms are entirely different and have evoked enormous interest in the study of these algorithms independent of any application such as zero-sum games, etc. (see [2], [3] and [26]).

In § 2 we formally state the problem and in § 3 present a proof of our main theorem. A comparison with related results along with some comments and examples are given in § 4.

2. Statement of the problem. We consider a zero-sum game between two players A and B in which both of the players have two pure strategies each. On the k th play, let player A use a mixed strategy $\mathbf{p}(k) = (p_1(k), p_2(k))$ where $p_s(k)$ is the probability of choosing the s th pure strategy. Similarly, player B uses $\mathbf{q}(k) = (q_1(k), q_2(k))$. The outcome of the game depends on chance and on the choice of the pure strategies of both players. There are only two possible values that the random outcome can take: $+1$, called unit gain for A (loss for B) and -1 called unit loss for A (gain for B). The game is played repeatedly and after each play both the players observe only the random outcome. At any stage each player is unaware of the mixed strategy used by the other player or its pure realization. In fact, they do not even know the distribution of the random outcome as a function of the strategy pair, or indeed what strategies are available to the other player.

Using a learning algorithm each player increases (decreases) the probability of choosing a particular pure strategy if that strategy was chosen on the previous play and resulted in a unit gain (loss). All the other probabilities are adjusted to keep the total probability equal to unity. The actual algorithm used in this paper may be described as follows:

Linear reward-penalty algorithm (L_{R-P}). Let θ_t^A and θ_t^B ($t = 1, 2$) be constants such that $0 < \theta_t^A, \theta_t^B < 1$. At the k th play, let the player A choose the i th pure strategy (that is, i th pure strategy is the sample realization of the mixed strategy $\mathbf{p}(k)$) and player B choose the j th pure strategy. The mixed strategy $\mathbf{p}(k+1)$ to be used by player A at stage $k+1$ is defined by

$$(1) \quad \left. \begin{aligned} p_i(k+1) &= p_i(k) + \theta_1^A [1 - p_i(k)], \\ p_s(k+1) &= p_s(k) - \theta_1^A p_s(k), \quad s \neq i \end{aligned} \right\} \quad \text{if } A \text{ received unit gain on } k\text{th play,}$$

$$\left. \begin{aligned} p_i(k+1) &= p_i(k) - \theta_2^A p_i(k) \\ p_s(k+1) &= p_s(k) + \theta_2^A p_s(k), \quad s \neq i \end{aligned} \right\} \quad \text{if } A \text{ received unit loss on } k\text{th play.}$$

The mixed strategy $\mathbf{q}(k+1)$ for player B is defined exactly in the same way but by replacing \mathbf{p} by \mathbf{q} , i by j , θ_t^A by θ_t^B ($t = 1, 2$) and A by B .

Let the pair (i, j) denote the play at any stage, where i and j are the pure strategies chosen by A and B respectively. Also in the sequel by gain and loss we will refer to player A 's gain and loss. For the play (i, j) , let $d_{ij}(c_{ij} = 1 - d_{ij})$ be the probability of gain (loss). It is assumed that $0 < d_{ij} < 1$. Let D denote the matrix $[d_{ij}]$. Since player A tries to increase the gain and player B to decrease the gain, D is essentially the game matrix.¹

Let both the players use the L_{R-P} algorithm with arbitrary but fixed initial mixed strategies $\mathbf{p}(0)$ and $\mathbf{q}(0)$, where $0 \leq p_s(0) \leq 1$ and $0 \leq q_s(0) \leq 1$, ($s = 1, 2$). Clearly, $\mathbf{p}(k)$ and $\mathbf{q}(k)$ as defined by (1) are random vectors. It can be seen that $\{\mathbf{p}(k), \mathbf{q}(k)\}$ $k \geq 0$

¹ The actual game matrix is $G = [g_{ij}] = [2d_{ij} - 1]$.

is a stationary Markov process while $\{p(k)\}$ and $\{q(k)\}$ individually are nonstationary Markov processes.

Define

$$(2) \quad \eta(k) \triangleq E[p(k)]DE[q^T(k)],$$

where $E[\cdot]$ is the unconditional expectation taken over all possible choices of pure strategies by either player and all the random outcomes prior to the k th play and T denotes transpose.

Let $\theta_1^A = \theta\beta_A$, $\theta_2^A = \theta\alpha_A$, $\theta_1^B = \theta\beta_B$ and $\theta_2^B = \theta\alpha_B$. The constant $\beta_A(\beta_B)$ is called the reward parameter for $A(B)$, and $\alpha_A(\alpha_B)$ is called the penalty parameter for $A(B)$. It is assumed $0 < \alpha_A < \beta_A < 1$ and $0 < \alpha_B < \beta_B < 1$. The constant $0 < \theta < 1$ essentially controls the step size. Let V denote the (von Neumann) value of the game (corresponding to the matrix D). Our main result is summarized in the following.

THEOREM 1. *For every $\varepsilon > 0$, $0 < \beta_A < 1$, $0 < \beta_B < 1$ there exist $0 < \alpha_A^* < \beta_A$, $0 < \alpha_B^* < \beta_B$ and $0 < \theta^* < 1$ such that for all $0 < \alpha_A < \alpha_A^*$, $0 < \alpha_B < \alpha_B^*$ and $0 < \theta < \theta^*$*

$$\lim_{k \rightarrow \infty} |\eta(k) - V| < \varepsilon.$$

Stated in words, if either player uses the L_{R-P} algorithm (1), by proper choice of penalty parameters and step size parameters, the expected probability of gain can be made as close to the value of the game as desired. It will become clear in the course of the proof of this theorem that α_A^* , α_B^* and θ^* depend only on ε and not on the initial states $p(0)$ and $q(0)$. Theorem 1 further demonstrates the robustness of the method in the sense that both the players can individually choose their reward and penalty parameters in such a way that their asymptotic expected probability of gain can be made arbitrarily close to the value of the game.

3. Analysis of the game. Let $S = [0, 1] \times [0, 1]$, $I = \{1, 2\}$, $E_1 = \{\text{unit gain, unit loss}\}$ and $E = I \times I \times E_1$. Define $P(k) = p_1(k)$ and $Q(k) = q_1(k)$ and let $p(k) = (P(k), Q(k))$. Clearly $\{p(k)\}$ $k \geq 0$ is a Markov process with stationary transition function with S as its state space. E is called the event space. If $s = (P, Q) \in S$, $e = (i, j, e_1) \in E$, then the learning algorithm (1) defines a mapping $T: S \times E \rightarrow S$, where

$$s(k+1) = T[s(k), e(k)].$$

The event probability distribution is given below:

Let

$$K[e, s] = \text{Prob}[e | s = (P, Q) \in S]$$

where $e \in E$. If $e = (1, 2, e_1)$, then

$$K[e, s] = \begin{cases} P(1-Q)d_{12} & \text{if } e_1 = \text{unit gain,} \\ P(1-Q)C_{12} & \text{if } e_1 = \text{unit loss} \end{cases}$$

and similarly for other $e \in E$.

Let $d(s_1, s_2)$ refer to Euclidean distance in S .

DEFINITION 1. A learning algorithm T is said to be *distance diminishing* [19] if

$$d(s'_1, s'_2) < d(s_1, s_2) \quad \text{for all } e \in E,$$

where $s'_t = T[s_t, e]$, $t = 1, 2$.

DEFINITION 2. A state $s \in S$ is said to be an *absorbing state* [19] if

$$T[s, e] = s \quad \text{for all } e \in E.$$

Otherwise, it is called *nonabsorbing*.

If T corresponds to the L_{R-P} algorithm since $0 < d_{ij} < 1$ for $i, j = 1, 2$ and $0 < \theta_i^A, \theta_i^B < 1, i = 1, 2$, it can be verified by direct computation that T is distance diminishing and that all the states in S are nonabsorbing.

Remark 1. The quadruple (S, E, T, K) is called a *random system with complete connection* (RSCC) [18], [22]. In the sequel by “algorithm T ” we will implicitly refer to the associated RSCC.

We now state without proof an important proposition due to Norman [18, Chapt. 3, § 3.6, Thm. 6.1].

PROPOSITION 1. *A learning algorithm T which is distance diminishing and has no absorbing states is ergodic.*

An immediate consequence of this proposition is that the probability distribution of the process $p(k) = (P(k), Q(k))$ for large k is independent of the initial values of $P(0)$ and $Q(0)$. In the following our aim is to characterize the asymptotic distribution of $p(k)$.

Let $p = (P, Q)$ and $\delta p(k) = (\delta P(k), \delta Q(k))$ where $\delta z(k) = z(k+1) - z(k)$. It can be shown that

$$(P.1) \quad E[\delta p(k) | p(k) = p] = \theta W(p),$$

where

$$\begin{aligned} W(p) &= (W_1(p), W_2(p)), \\ W_i(p) &= W_i^R(p) + W_i^P(p), \quad i = 1, 2, \\ W_1^R(p) &= \beta_A P(1-P)[c_2^A(Q) - c_1^A(Q)], \\ W_2^R(p) &= \beta_B Q(1-Q)[c_2^B(P) - c_1^B(P)], \\ W_1^P(p) &= \alpha_A [(1-P)^2 c_2^A(Q) - P^2 c_2^A(Q)], \\ W_2^P(p) &= \alpha_B [(1-Q)^2 c_2^B(P) - Q^2 c_2^B(P)], \\ C_i^A(Q) &= (1 - d_{i2}) - Q(d_{i1} - d_{i2}), \quad i = 1, 2, \\ C_i^B(P) &= d_{2i} + P(d_{1i} - d_{2i}), \quad i = 1, 2. \end{aligned}$$

Similarly,

$$(P.2) \quad E\{[\delta p(k) - W(p)]^T [\delta p(k) - W(p)] | p(k) = p\} = \theta^2 s(p), \quad \text{where } s(p) = a(p) - W^T(p)W(p) \text{ and } a(p) = E[\delta p^T(k)\delta p(k) | p(k) = p].$$

The elements of the matrix $a(p)$ can be easily computed. As all the states are nonabsorbing it follows that $s(p)$ is positive definite uniformly for all $p \in S$. Further,

$$(P.3) \quad E\{|\delta p(k)|^3 | p(k) = p\} = O(\theta^3) \text{ uniformly for all } p \in S, \text{ where } |\cdot| \text{ is the norm corresponding to the metric } d \text{ on } S.$$

$$(P.4) \quad W(p) \text{ has a bounded Lipschitz derivative in } S.$$

$$(P.5) \quad s(p) \text{ is Lipschitz in } S.$$

$$(P.3)-(P.5) \text{ can be easily checked by routine arguments and we omit the details.}$$

Let $P_{\text{opt}} = (P_{\text{opt}}, Q_{\text{opt}})$ be the state that corresponds to the optimal strategy for both players. For purposes of easy presentation in the following we distinguish two cases.

Case 1. The game matrix has no saddle point in pure strategies. In this case without loss of generality let us assume

$$(3) \quad d_{11} > \max\{d_{12}, d_{21}\}, \quad d_{22} > \max\{d_{12}, d_{21}\}.$$

It can be seen that

$$(4) \quad P_{\text{opt}} = \frac{d_{22} - d_{21}}{L}, \quad Q_{\text{opt}} = \frac{d_{22} - d_{12}}{L},$$

where $L = (d_{11} + d_{22}) - (d_{12} + d_{21})$.

Case 2. The game matrix has a saddle point in pure strategies. In this case we shall assume that

$$(5) \quad d_{21} < d_{11} < d_{12},$$

from which we have $P_{\text{opt}} = 1$ and $Q_{\text{opt}} = 1$.

Remark 2. All the other relations between the elements of the matrix D can be reduced to (3) or (5) by suitable relabeling of players and/or strategies.

The following Lemmas 1 and 2 are crucial for our analysis.

LEMMA 1. *For every $\varepsilon > 0$, there exist $0 < \alpha_A^* < \beta_A$ and $0 < \alpha_B^* < \beta_B$ such that for all $0 < \alpha_A < \alpha_A^*$, $0 < \alpha_B < \alpha_B^*$ there exists a unique $p^* = (P^*, Q^*)$ such that*

$$(a) \quad W(p^*) = 0,$$

and

$$(b) \quad |p^* - p_{\text{opt}}| < \varepsilon.$$

Proof. Consider Case 1 when D has no saddle point in pure strategies. Define

$$C_{21}^A(Q) = C_2^A(Q) - C_1^A(Q) = (d_{12} - d_{22}) + LQ.$$

Since $L > 0$, $C_{21}^A(Q)$ is an increasing function of Q with

$$C_{21}^A(Q) \begin{cases} < 0 & \text{if } Q < 1_{\text{opt}}, \\ = 0 & \text{if } Q = Q_{\text{opt}}, \\ > 0 & \text{if } Q > Q_{\text{opt}}. \end{cases}$$

For each fixed Q $W_1(p)$ is quadratic in P and $W_1(p) = \alpha_A C_2^A(Q) > 0$ at $P = 0$ and $W_1(p) = -\alpha_A C_1^A(Q) < 0$ at $P = 1$. From this fact and from Lemma A in Appendix A it follows that there exists a unique $p(Q) = (P(Q), Q)$ for each fixed Q such that $W_1(p(Q)) = 0$ and

$$P(Q) \begin{cases} < \frac{1}{2} & \text{if } Q < Q_{\text{opt}}, \\ = \frac{1}{2} & \text{if } Q = Q_{\text{opt}}, \\ > \frac{1}{2} & \text{if } Q > Q_{\text{opt}}. \end{cases}$$

Similarly

$$C_{21}^B(P) = (d_{22} - d_{21}) - LP$$

and $C_{21}^B(P)$ is a decreasing function of P with

$$C_{21}^B(P) \begin{cases} > 0 & \text{if } P < P_{\text{opt}}, \\ = 0 & \text{if } P = P_{\text{opt}}, \\ < 0 & \text{if } P > P_{\text{opt}}. \end{cases}$$

Since $W_2(p)$ is quadratic in Q for each fixed P with $W_2(p) = \alpha_B C_2^B(P) > 0$ at $Q = 0$, $W_2(p) = -\alpha_B C_1^B(P) < 0$ at $Q = 1$, there exists a unique $p(P) = (P, Q(P))$ such that $W_2(p(P)) = 0$ and by Lemma A

$$Q(P) \begin{cases} > \frac{1}{2} & \text{if } P < P_{\text{opt}}, \\ = \frac{1}{2} & \text{if } P = P_{\text{opt}}, \\ < \frac{1}{2} & \text{if } P > P_{\text{opt}}. \end{cases}$$

The functions $P(Q)$ and $Q(P)$ when plotted on the unit square have a unique intersection $p^* = (P^*, Q^*)$ such that $W(p^*) = 0$. This proves part (a) of Lemma 1.

To prove part (b) first observe that $W_1(p)$ and $W_2(p)$ are continuous in α_A and α_B , and if $\alpha_A = \alpha_B = 0$ then

$$W_1(p) = 0 \quad \text{all along} \begin{cases} P = 0, \\ P = 1, \\ Q = Q_{\text{opt}}, \end{cases}$$

$$W_2(p) = 0 \quad \text{all along} \begin{cases} Q = 0 \\ Q = 1, \\ P = P_{\text{opt}}. \end{cases}$$

From this we obtain

$$W(p) = 0 \quad \text{at } p = (0, 0), (0, 1), (1, 0), (1, 1) \text{ and } (P_{\text{opt}}, Q_{\text{opt}}).$$

Again from Lemma A in Appendix A, by choosing α_A small, $P(Q)$ can be made close to zero for all $Q < Q_{\text{opt}}$ and close to unity for all $Q > Q_{\text{opt}}$. Similarly, by choosing α_B small, $Q(P)$ can be made close to unity for all $P < P_{\text{opt}}$ and close to zero for all $P > P_{\text{opt}}$. This in turn implies that for any given $\varepsilon > 0$ there exist $0 < \alpha_A^* < \beta_A$ and $0 < \alpha_B^* < \beta_B$ such that for all $0 < \alpha_A < \alpha_A^*$ and $0 < \alpha_B < \alpha_B^*$, we obtain $|p^* - p_{\text{opt}}| < \varepsilon$.

The proof of Lemma 1 for Case 2 follows along the same lines and we omit the details. \square

Denoting the Hessian of $W(p)$ as $W'(p)$ we have

$$W'(p) = \begin{bmatrix} \frac{\delta W_1(p)}{\delta P} & \frac{\delta W_1(p)}{\delta Q} \\ \frac{\delta W_2(p)}{\delta P} & \frac{\delta W_2(p)}{\delta Q} \end{bmatrix},$$

where

$$\frac{\delta W_1(p)}{\delta P} = \beta_A [1 - 2P] [C_2^A(Q) - C_1^A(Q)] - 2\alpha_A [(1 - P)C_2^A(Q) + PC_1^A(Q)],$$

$$\frac{\delta W_1(p)}{\delta Q} = \beta_A LP(1 - P) + \alpha_A [(1 - P)^2(d_{22} - d_{21}) - P^2(d_{12} - d_{11})]$$

$$\frac{\delta W_2(p)}{\delta P} = -\beta_B LQ(1 - Q) + \alpha_B [(1 - Q)^2(d_{12} - d_{22}) - Q^2(d_{11} - d_{21})],$$

$$\frac{\delta W_2(p)}{\delta Q} = \beta_B [1 - 2Q] [C_2^B(P) - C_1^B(P)] - 2\alpha_B [(1 - Q)C_2^B(P) + QC_1^B(P)],$$

where L is defined in (4) above.

LEMMA 2. $W'(p^*)$ is negative definite where p^* is as defined in Lemma 1.

Proof.

Case 1. From Lemma 1 it follows that by proper choice of α_A and α_B we can make $P^* \simeq P_{\text{opt}}$ and $Q^* \simeq Q_{\text{opt}}$. With this choice it can be seen that $C_i^A(Q^*) \simeq C_i^A(Q_{\text{opt}})$ and $C_i^B(P^*) \simeq C_i^B(P_{\text{opt}})$. Further, it can be checked that in Case 1

$$C_2^A(Q_{\text{opt}}) - C_1^A(Q_{\text{opt}}) = 0 \quad \text{and} \quad C_2^B(P_{\text{opt}}) - C_1^B(P_{\text{opt}}) = 0.$$

Thus, at the point p^* , $W'(p^*)$ can be well approximated by

$$W'(p^*) \approx \begin{bmatrix} -2\alpha_A C_2^A(Q_{\text{opt}}) & \frac{(\beta_A + \alpha_A)}{L}(d_{11} - d_{12})(d_{22} - d_{21}) \\ -\frac{(\beta_B + \alpha_B)}{L}(d_{11} - d_{22})(d_{22} - d_{12}) & -2\alpha_B C_2^B(P_{\text{opt}}) \end{bmatrix}.$$

Let u and v denote the trace and determinant of $W'(p^*)$. Then

$$\begin{aligned} u &= -2[\alpha_A C_2^A(Q_{\text{opt}}) + \alpha_B C_2^B(P_{\text{opt}})] < 0, \\ v &= 4\alpha_A \alpha_B D_2^A(Q_{\text{opt}}) C_2^B(P_{\text{opt}}) + \frac{(\beta_A + \alpha_A)(\beta_B + \alpha_B)}{L^2} E, \end{aligned}$$

where $E = (d_{11} - d_{21})(d_{11} - d_{12})(d_{22} - d_{12})(d_{22} - d_{21})$.

From (3) it follows that $E > 0$ and hence $v > 0$. Also it can be verified that $u^2 - 4v < 0$. It follows [23] that the eigenvalues of $W'(p^*)$ are complex conjugates with negative real parts.

Case 2. It follows from Lemma 1 that $P^* \approx 1$ and $Q^* \approx 1$. $W'(p^*)$ in this case can be well approximated by

$$(6) \quad W'(p^*) = \begin{bmatrix} -\beta_A(d_{11} - d_{21}) - 2\alpha_A(1 - d_{11}) & -\alpha_A(d_{12} - d_{11}) \\ -\alpha_A(d_{11} - d_{21}) & -\beta_B(d_{12} - d_{11}) - 2\alpha_B d_{11} \end{bmatrix}.$$

From (5) it follows that $u < 0$ and $v > 0$ but $u^2 - 4v$ is positive only for very small values of α_A and α_B . Again it follows [23] that the eigenvalues of $W'(p^*)$ are real and negative or complex conjugates with negative real parts. \square

Examples of exact computation of $W'(p^*)$ and its eigenvalues given in § 4 further illustrate Lemma 2.

Proof of Theorem 1. Combining (P.1)–(P.5) and Lemmas 1 and 2, it follows that the process $\{p(k)\}$ $k \geq 0$ satisfies all the conditions of Theorem B of Appendix B. Hence from the conclusion (b) of that theorem we obtain

$$(7) \quad E[p(k) | p(0) = p] = f(k\theta) + O(\theta)$$

for all $k \geq 0$ uniformly for all $p \in S$, where

$$(8) \quad f'(t) = W(f(t)), \quad f(0) = p(0) = p.$$

From the properties of $W(\cdot)$ it follows that the differential equation (8) is uniformly asymptotically stable in S [24], that is,

$$(9) \quad f(t) \rightarrow p^* \quad \text{as } t \rightarrow \infty.$$

Also from conclusion (c) of Theorem B in Appendix B, we obtain that the normalized random vector $\theta^{-1/2}[P(k) - f(k\theta)]$ converges in distribution. This, in turn, implies that $E[p(k)]$ converges [25], that is,

$$(10) \quad \lim_{k \rightarrow \infty} E[p(k)] \text{ exists.}$$

From Lemma 1, (7), (9) and (10) it follows that for any $\delta > 0$ there exists a $0 < \theta^* < 1$ such that for all $0 < \theta < \theta^*$

$$(11) \quad \lim_{k \rightarrow \infty} |E[p(k)] - p_{\text{opt}}| < \delta.$$

Substituting (11) in (2) we obtain

$$(12) \quad \lim_{k \rightarrow \infty} |\eta(k) - V| < \delta |h[d_{ij}]|,$$

where $h[d_{ij}]$ is a bounded function of the elements of the matrix D . Now for any given $\varepsilon > 0$, we can choose $\delta > 0$ such that the right-hand side of (12) is less than ε . This concludes the proof of Theorem 1. \square

4. Examples and comments

4.1. Examples. In this subsection we illustrate the foregoing analysis with different examples. In each example, the variation of P^* , Q^* and the eigenvalues (λ_1 and λ_2) of $W'(p^*)$ as a function of α are given. All the calculations are done in double precision arithmetic. These examples further illustrate the accuracy of the approximations in Lemma 2 for characterizing the nature of the eigenvalues of $W'(p^*)$.

Example 1. Consider an instance of case 1 where the matrix D is given below.

$$D = \begin{bmatrix} .6 & .2 \\ .35 & .9 \end{bmatrix}.$$

It can be checked that $P_{\text{opt}} = 0.5789$ and $Q_{\text{opt}} = 0.7368$. Table 1 gives the variation of P^* , Q^* , λ_1 and λ_2 for different values of $\alpha_A = \alpha_B = \alpha$ and for $\beta_A = \beta_B = \beta$.

TABLE 1

β	α	P^*	Q^*	λ_1		λ_2	
				Re (λ_1)	Im (λ_1)	Re (λ_2)	Im (λ_2)
0.20	0.0	0.5789	0.7368	0.0	0.0413	0.0	-0.0413
0.20	0.001	0.5255	0.7384	-0.00131	0.0415	-0.00131	-0.0415
0.20	0.005	0.5469	0.7412	-0.00652	0.0424	-0.00652	-0.0424
0.20	0.01	0.5169	0.7400	-0.01280	0.0436	-0.01280	-0.0436
0.20	0.05	0.4713	0.7257	-0.02470	0.0463	-0.02470	-0.0463

Example 2. In this example we consider two different matrices both having saddle-points in pure strategies:

$$D_1 = \begin{bmatrix} .6 & .8 \\ .35 & .9 \end{bmatrix}, \quad D_2 = \begin{bmatrix} .7 & .9 \\ .6 & .8 \end{bmatrix}.$$

D_1 has column dominance and D_2 has both row and column dominance. Tables 2 and 3 give the variation of P^* , Q^* , λ_1 and λ_2 for D_1 and D_2 respectively.

TABLE 2
 $\beta_A = \beta_B = \beta$, $\alpha_A = \alpha_B = \alpha$

β	α	P^*	Q^*	λ_1		λ_2	
				Re (λ_1)	Im (λ_1)	Re (λ_2)	Im (λ_2)
0.2	0.0	1.00	1.00	-0.0500	0.0	-0.0400	0.0
0.2	0.001	0.99196	0.98548	-0.0489	0.0	-0.0406	0.0
0.2	0.002	0.98385	0.97187	-0.0478	0.0	-0.0415	0.0
0.2	0.005	0.95932	0.93599	-0.0446	0.00286	-0.0446	-0.00286
0.2	0.01	0.91885	0.88954	-0.0455	0.00518	-0.0455	-0.00518

TABLE 3
 $\beta_A = \beta_B = \beta, \alpha_A = \alpha_B = \alpha$

β	α	P^*	Q^*	λ_1		λ_2	
				Re (λ_1)	Im (λ_1)	Re (λ_2)	Im (λ_2)
0.2	0.0	1.0	1.0	-0.02000	0.0	-0.04000	0.0
0.2	0.001	0.98539	0.98234	-0.02002	0.0	-0.04003	0.0
0.2	0.002	0.97154	0.96637	-0.02010	0.0	-0.04010	0.0
0.2	0.005	0.93408	0.92094	-0.02030	0.0	-0.04080	0.0
0.2	0.01	0.88300	0.85766	-0.02090	0.0	-0.04300	0.0

4.2. Comments.

(1) When $\alpha_A = \alpha_B = 0$ the equation $W(p) = 0$ has multiple solutions and Theorem B of Appendix B is not applicable. Thus uniqueness of the solution dictates that α_A and α_B both be greater than zero.

(2) When $\alpha_A = \alpha_B = 0$ (or $\theta_2^A = \theta_2^B = 0$) algorithm (1) is called a linear reward-inaction algorithm. It was shown in [17] that if the game matrix D has a saddle point, the use of such algorithms by both players will result in their picking their optimal pure strategies asymptotically with probabilities arbitrarily close to 1. When the game matrix D has no saddle point in pure strategies $W(p_{\text{opt}}) = 0$ and the eigenvalues of $W'(p_{\text{opt}})$ are purely imaginary. This explains the oscillations of $\eta(k)$ around the von Neumann value reported in [15].

(3) When one of the players (say B) uses a fixed pure or mixed strategy, the problem discussed in this paper is equivalent to a game against nature [2], [3], [16]. Let B pick the two pure strategies with probabilities Q and $1 - Q$ where $0 \leq Q \leq 1$. By the proper choice of α [16] player A can make the probability of gain as close to $\max_i \{Qd_{i1} + (1 - Q)d_{i2}\}$ as possible.

Further, it can be shown that

$$\max_i \{Qd_{i1} + (1 - Q)d_{i2}\} \geq V$$

with equality holding when $Q = Q_{\text{opt}}$ —when B uses his optimal strategy and when V is the von Neumann value of the game. In other words, player A stands to gain if player B uses a fixed strategy.

(4) When the game matrix has no saddle point in pure strategies (as in Example 1) the importance of the penalty parameter is evident from Table 1. Larger values of α result in greater distances between p^* and p_{opt} . In other words a compromise has to be made between speed of convergence of $\eta(k)$ on the one hand and accuracy (in the sense of the nearness of the limiting value of $\eta(k)$ to the von Neumann value of the game) on the other.

(5) In this paper it is assumed that there is no transfer of information between the players during the evolution of the game. The manner in which players can exploit such information while using learning schemes needs further investigation.

Appendix A. Define

$$h_1(x) \triangleq [(1 - x)^2 c_2 - x^2 c_1],$$

$$h_2(x) \triangleq x(1 - x)(c_2 - c_1) + h_1(x),$$

$$h_3(x) \triangleq x(1 - x)(c_2 - c_1) + bh_1(x),$$

where $x \in [0, 1]$, $0 < C_1$, $C_2 < 1$, $0 < b < 1$. The following Lemma A provides the basic idea for the proof of Lemma 1.

LEMMA A. *There exist unique $\lambda_i \in (0, 1)$ ($i = 1, 2, 3$) such that $h_i(\lambda_i) = 0$ and:*

- (a) $\lambda_3 \geq \lambda_2 \geq \lambda_1 \geq \frac{1}{2}$ according as $c_2 \geq c_1$.
- (b) λ_i is an increasing function of $(c_2 - c_1)$, for all $i = 1, 2, 3$.
- (c) as $b \rightarrow 0$, $\lambda_3 \rightarrow 1$ or 0 according as $c_2 > c_1$ or $c_2 < c_1$.

Proof. Let $c_2 - c_1 = \delta$. Note that $\delta \in (-1, 1)$.

We have

$$(A1) \quad h_1(x) = c_1(1 - 2x) + \delta(1 - x)^2.$$

As $h_1(0) = c_2 > 0$; $h_1(1) = -c_1 < 0$, $h_1(\frac{1}{2}) \geq 0$ according as $\delta \geq 0$ and $h_1(x)$ is quadratic in x , there exists a unique $\lambda_1 \in (0, 1)$ and $\lambda_1 \geq \frac{1}{2}$ according as $\delta \geq 0$ such that $h_1(\lambda_1) = 0$. From (A1) it further follows that λ_1 is an increasing function of δ .

Consider

$$(A2) \quad h_2(x) = c_1(1 - 2x) + \delta(1 - x):$$

we have $h_2(0) = c_2 > 0$; $h_2(1) = -c_1 < 0$; $h_2(\lambda_1) = \lambda_1(1 - \lambda_1)\delta$ and $h_2(x)$ is linear in x . Thus there exists a unique $\lambda_2 \in (0, 1)$ such that $h_2(\lambda_2) = 0$ and $\lambda_2 \geq \lambda_1$ according as $\delta \geq 0$. Also, the claim λ_2 is an increasing function of δ follows from (A2). Similarly, we have

$$(A3) \quad \begin{aligned} h_3(x) &= h_2(x) - (1 - b)h_1(x), \\ h_3(0) &= bc_2 > 0, \quad h_3(1) = -bc_1 < 0, \\ h_3(\lambda_2) &= -(1 - b)h_1(\lambda_2) > 0. \end{aligned}$$

As $h_3(x)$ is quadratic in x , there exists a $\lambda_3 \in (0, 1)$ such that $h_3(\lambda_3) = 0$ and $\lambda_3 \geq \lambda_2$ accordingly as $\delta \geq 0$. Further, λ_2 increasing with δ implies λ_3 is also increasing with δ . To prove (c), let $\delta > 0$ and redefine $h_3(x)$ as

$$(A4) \quad h_3[b, x] = h_2(x) - (1 - b)h_1(x)$$

and let λ_3^b be such that $h_3[b, \lambda_3^b] = 0$, to emphasize the dependence on b . If $b' < b$, then it can be seen that

$$(A5) \quad h_3[b', \lambda_3^b] = h_1(\lambda_3^b)(b' - b) > 0.$$

From (A4) and (A5) it follows that $\lambda_3^{b'} > \lambda_3^b$ if $b > b'$. Thus λ_3 increases as b decreases and by making b sufficiently small, λ_3^b can be made as close to 1 as desired. By similar arguments, we can see that

$$\lambda_3 \rightarrow 0 \quad \text{as } b \rightarrow 0, \quad \text{when } \delta < 0. \quad \square$$

Appendix B. In this appendix, for the sake of easy reference, we quote a theorem from [18]–[20] that provides the basic convergence result that is relevant for our analysis. For each $\theta \in (0, 1)$, let $\{x_k^\theta\}$ be a stationary Markov process where $x_k^\theta \in I$ for all θ and $k \geq 0$ where I is a subset of the M -dimensional space R^M . The parameter θ is an index of the magnitude of $\delta x_k^\theta = x_{k+1}^\theta - x_k^\theta$ and we are concerned with the asymptotic behavior of x_k^θ as $\theta \rightarrow 0$ and $k\theta \rightarrow \infty$.

Assumptions.

- A1. $E\{\delta x_k^\theta | x_k = y\} = \theta W(y) + O(\theta^2)$.
- A2. $E\{[\delta x_k^\theta - W(y)][\delta x_k^\theta - W(y)]^T | x_k = y\} = \theta^2 s(y) + o(\theta^2)$.
- A3. $E\{|\delta x_k^\theta|^3 | x_k = y\} = O(\theta^3)$,

where $|\cdot|$ is a norm defined in I and all the orders of magnitudes are uniform in y .

A4. $W(y)$ has bounded Lipschitz derivative.

A5. $s(y)$ is Lipschitz in I .

Let

$$\mu_k(\theta, x) \triangleq E\{x_k^\theta | x_0^\theta = x\} \quad \text{and} \quad \omega_k(\theta, x) \triangleq E\{[x_k^\theta - \mu_k(\theta, x)][x_k^\theta - \mu_k(\theta, x)]^T | x_0^\theta = x\}.$$

Let $W'(y)$ denote the Hessian of $W(y)$ and (x, y) denote an inner product in R^M .

The following theorem summarizes the behavior of x_k when $\theta \rightarrow 0$ and $k\theta \rightarrow \infty$.

THEOREM B. *If, in addition to Assumptions A1–A5,*

- (i) *I is compact,*
- (ii) *there exists an unique $\lambda \in I$ such that $W(\lambda) = 0$, that is, $(y - \lambda, w(\lambda)) < 0$ for all $y \neq \lambda$ and $y \in I$,*
- (iii) *$(z, W'(y)z) < 0$ for all $z \in R^M$ and $z \neq 0$,*

then the following conclusions are true:

- (a) *$\omega_k(\theta, x) = O(\theta)$ uniformly for all $x \in I$ and $k \geq 0$.*
- (b) *For any $x \in I$, the (vector) differential equation*

$$1) \quad f'(t) = W[f(t)]$$

has an unique solution $f(t)$ where $f(0) = x$ and

$$2) \quad \mu_k(\theta, x) = f(k\theta) + O(\theta) \text{ uniformly for all } x \in I \text{ and } k \geq 0.$$

(c) *The (matrix) differential equation*

$$1) \quad g'(t) = W'[f(t)]g(t) + g(t)W'^T[f(t)] + s[f(t)]$$

has an unique solution $g(t)$ with $g(0) = 0$:

$$2) \quad \mathcal{L}(Z_k^\theta) \sim N(0, g(t)) \quad \text{as } \theta \rightarrow 0, k\theta \rightarrow t < \infty \quad \text{where } Z_k^\theta \triangleq [x_k^\theta - f(k\theta)]\theta^{-1/2}$$

($\mathcal{L}(x)$ refers to the distribution of x and $N(a, b)$ refers to the normal distribution with mean a and covariance matrix b) and as $\theta \rightarrow 0$, $k\theta \rightarrow \infty$, $g(\infty)$ is obtained as the unique solution of the system of linear equations.

$$3) \quad W'(\lambda)g(\infty) + g(\infty)W'^T(\lambda) + S(\lambda) = 0.$$

The theorem as stated above is a combination of many theorems from Norman [18]–[20].

REFERENCES

- [1] M. L. TSETLIN, *Automaton Theory and Modelling of Biological Systems*, Academic Press, New York, 1973.
- [2] K. S. NARENDRA AND M. A. L. THATHACHAR, *Learning automata—A survey*, IEEE Trans. Systems, Man and Cybernetics, 4 (1974) pp. 323–334.
- [3] K. S. NARENDRA AND S. LAKSHMIVARAHAN, *Learning automata—A critique*, Cybernetics and Information Sciences—Special Issue on Learning Automata, 1 (1978), pp. 53–66.
- [4] J. C. HARSANYI, *Games with incomplete information played by Bayesian players—Part I*, Management Sci., 14 (1967), pp. 159–182.
- [5] H. E. SCARF AND L. S. SHAPLEY, *Games with partial information*, in Contributions to the Theory of Games, Vol. III, Kuhn and Tucker, eds., Princeton University Press, Princeton, NJ, 1957, pp. 213–231.
- [6] R. J. AUMANN AND M. MASCHLER, *Repeated games with incomplete information*, Rep. ACDA/ST-116, Mathematica, Princeton, NJ, 1968, Chapter 10, pp. 287–303.
- [7] E. KOHLBERG, *Optimal strategies in repeated games with incomplete information*, Internat. J. Game Theory, 4 (1976), pp. 7–24.

- [8] J. P. PONSSARD AND S. ZAMIR, *Zero-sum sequential games with incomplete information*, *ibid.*, 2 (1974), pp. 99–110.
- [9] S. ZAMIR, *On the notion of the value for games with infinitely many stages*, *Ann. Statist.* 1 (1973), pp. 791–796.
- [10] P. SUPPES AND R. C. ATKINSON, *Markov Learning Models for Multi-Person Interaction*, Stanford University Press, Stanford, CA, 1960.
- [11] V. P. CRAWFORD, *Learning the optimal strategy in a zero-sum game*, *Econometrica*, 42 (1974), pp. 885–891.
- [12] J. VON NEUMANN AND O. MORGENSTERN, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, NJ, 1944.
- [13] A. P. SANGHVI AND M. J. SOBEL, *Bayesian games as stochastic processes*, *Internat. J. Game Theory*, 5 (1976), pp. 1–22.
- [14] B. CHANDRASEKARAN AND D. W. C. SHEN, *On stochastic automata games*, *IEEE Trans. Systems Science and Cybernetics*, 5 (1969), pp. 145–146.
- [15] R. VISWANATHAN AND K. S. NARENDRA, *Games of stochastic automata*, *IEEE Trans. Systems and Cybernetics*, 4 (1974), pp. 131–135.
- [16] S. LAKSHMIVARAHAN, *ϵ -optimal learning algorithms—non-absorbing barrier type*, Tech. Rep. EECS-7901, Feb. 1979, School of Electrical Engineering and Computing Sciences, University of Oklahoma, Norman.
- [17] S. LAKSHMIVARAHAN AND K. S. NARENDRA, *Learning algorithms for two-person zero sum stochastic games with incomplete information*, S & IS Rep. 7712, April 1978, Dept. Engineering and Applied Science, Yale University, New Haven, CT; *Math. Oper. Res.* 6 (1981), pp. 379–386.
- [18] M. F. NORMAN, *Markov Processes and Learning Models*, Academic Press, New York, 1973.
- [19] ———, *Some convergence theorems for stochastic learning models with distance diminishing operators*, *J. Math. Psych.*, 5 (1968), pp. 61–101.
- [20] ———, *A central limit theorem for Markov processes that move by small steps*, *Ann. Probab.* 2 (1974), pp. 1065–1074.
- [21] R. R. BUSH AND F. MOSTELLER, *Stochastic Models for Learning*, John Wiley, New York, 1955.
- [22] M. IOSIFESCU AND R. THEODORESCU, *Random Processes and Learning*, Springer-Verlag, New York, 1969.
- [23] A. BLAQUIÈRE, *Non-linear System Analysis*, Academic Press, New York, 1966, Chap. 3.
- [24] J. P. LASALLE AND S. LEFSHETZ, *Stability by Liapunov's Direct Method, With Applications*, Academic Press, New York, 1961.
- [25] W. FELLER, *An Introduction to Probability Theory and Applications Vol. II*, John Wiley, New York, Chap. VIII.
- [26] S. LAKSHMIVARAHAN, *Learning Algorithms: Theory and Applications*, Springer-Verlag, New York, 1981.

ON MINIMAL SPLITTING SUBSPACES AND STOCHASTIC REALIZATIONS*

ARTHUR E. FRAZHO†

Abstract. Using shift operator techniques, the set of all minimal splitting subspaces for a stationary, scalar valued random process is given. Each minimal splitting subspace is shown to naturally define a stochastic forward and backward realization. These realizations are related through duality and can be infinite-dimensional.

1. Introduction. The role of minimal splitting subspaces in stochastic realization theory has been well stressed in [19], [20], [22] and elsewhere. The backward shift operator has played a substantial role in linear operator theory [5], [8], [13], [14], [15], [17], [23], [24] and deterministic realization theory [4], [10], [11], [12], [16]. In this note the backward shift operator and the concept of a minimal splitting subspace is used to develop stochastic realizations for a wide sense stationary process y . First, [6, Thm. (3.1.5)] is used to provide a complete characterization of all minimal splitting subspaces for y . This offers a new interpretation to some of the results in [19]; see § 2. Then it is shown that each minimal splitting subspace naturally defines a forward stochastic realization and a backward stochastic realization. These realizations are obtained directly from the backward shift operator. They are the stochastic counterparts of the shift realizations used in deterministic systems theory [4], [10], [11], [12], [16].

To begin, some notation is established. Throughout, $y(n)$ is a scalar valued, discrete-time, wide sense stationary, second order, centered process. The Hilbert space generated by the closed linear span of $\{y(n) : -\infty < n < \infty\}$ is denoted by \mathcal{H} . The inner product on \mathcal{H} is $(x, y) = Ex\bar{y}$, where E is the mathematical expectation. The *future* of $[past\ of]$ y is defined by $\mathcal{H}^+ \doteq \bigvee_{n \geq 0} y(n)$ [$\mathcal{H}^- \doteq \bigvee_{n \leq 0} y(n)$], respectively. If \mathcal{U} is a linear subspace of \mathcal{H} (or any Hilbert space), then $P(h|\mathcal{U})$ is the orthogonal projection of h on \mathcal{U} . The shift operators $U(n)$ is the uniquely determined group of unitary operators on \mathcal{H} defined by $U(n)y(k) \doteq y(n+k)$. Note $U(n+m) = U(n)U(m)$ for all integers n, m .

We say that $\mathcal{X}_1 \subseteq \mathcal{H}$ is a *splitting subspace* for y if $\mathcal{H}^- \vee \mathcal{X}_1$ is an irreducible invariant subspace for $U(-1)$ and

$$(1) \quad P(\mathcal{H}^+|\mathcal{H}^- \vee \mathcal{X}_1) = P(\mathcal{H}^+|\mathcal{X}_1).$$

\mathcal{S} is an *irreducible invariant subspace* for an isometry W on \mathcal{H} if \mathcal{S} is an invariant subspace for W and $\bigcap_{n \geq 0} W^n \mathcal{S} = \{0\}$, [8], [13], [14], [23]. The usual definition of a splitting subspace does not require that $\mathcal{H}^- \vee \mathcal{X}_1$ be an irreducible invariant subspace [19], [20], [22]. This extra assumption has been incorporated in our definition to simplify some of the forthcoming calculations. It has no effect on the final result. The irreducibility implies that the process $U(n)\mathcal{X}_1 \doteq \{U(n)x : x \in \mathcal{X}_1\}$ is purely nondeterministic, i.e. $\bigcap_{n \leq 0} [\bigvee_{i=-\infty}^n U(i)\mathcal{X}_1] \subseteq \bigcap_{n \leq 0} U(n)(\mathcal{H}^- \vee \mathcal{X}_1) = \{0\}$. Furthermore, the additional assumption eliminates many trivial splitting subspaces that would otherwise occur. For instance, under our definition, $\mathcal{X}_1 = \mathcal{H}$ is not a splitting subspace. Finally, a splitting subspace \mathcal{X}_1 is called *minimal* if \mathcal{X}_1 contains no strictly proper splitting subspace.

* Received by the editors April 15, 1980, and in revised form September 1981.

† Purdue University, School of Aeronautics and Astronautics, West Lafayette, Indiana 47907.

2. Functional representations. At this point it becomes advantageous to work in the transform domain. This setting allows us to apply the shift operator techniques of [6], [13], [24] to obtain all the minimal splitting subspace for y .

To begin, some standard terminology is established. The closure of a set \mathcal{X} is denoted by $\text{cl } \mathcal{X}$. Let $Z = \{z : |z| = 1\}$ be the unit circle in the complex plane and dm the normalized Lebesgue measure on Z . The set of all square integrable, Lebesgue measurable functions $f(z)$ on $[0, 2\pi]$ is denoted by L^2 . Multiplication by z on L^2 is defined by $M_z f \doteq zf$ if f is in L^2 . Since $z = e^{it}$ for $t \in [0, 2\pi]$, the complex conjugate of z equals the inverse of z , i.e., $\bar{z} = 1/z$ and $M_{\bar{z}} = (M_z)^{-1}$. The Hardy space H^2 is the subspace of L^2 consisting of all functions $f(z)$ such that $\int z^n f dm = 0$ for all $n > 0$, [14], [17], [24]. An *inner function* ψ is an element in H^2 such that $|\psi(z)| = 1$ a.e. An excellent reference on inner functions is [17]. If ψ is an inner function, then H_ψ is the subspace defined by $H_\psi \doteq H^2 \ominus \psi H^2$. Following [24], we call θ in H^2 *outer* if $\bigvee_{n \geq 0} z^n \theta = H^2$. This is equivalent to the usual definition of an outer function given in [17].

The *backward shift operator* is the co-isometry on H^2 defined by

$$(2) \quad U_+^* f = \frac{f - f_0}{z} = P(\bar{z}f | H^2) \quad (f \in H^2),$$

where $f_0 \doteq \int f dm$. If \mathcal{X} is an invariant subspace for U_+^* , then $U_+^*|_{\mathcal{X}}$ is the restriction of U_+^* to \mathcal{X} . The following modified version of [6, Thm. (3.1.5)] will play an important role in our theory. The proof is identical to the one in [6].

THEOREM 1. *Let f be in H^2 . The vector $U_+^* f$ is noncyclic for U_+^* if and only if there exists a g in H^2 and an inner function Φ such that $f(z) = \bar{g}(z)\Phi(z)$ a.e. Further, if the inner part of g and Φ are relatively prime, then the factorization is unique up to a constant of modulus one, and*

$$(3) \quad \bigvee_{n > 0} U_+^{*n} f = H^2 \ominus \Phi H^2 \doteq H_\Phi.$$

Using this factorization, it is easy to verify:

COROLLARY 1. *Let f be in H^2 and f_e be its outer factor. Then*

- (i) *f is cyclic for U_+^* if and only if $U_+^* f$ is cyclic for U_+^* .*
- (ii) *$U_+^* f$ is cyclic for U_+^* if and only if $U_+^* f_e$ is cyclic for U_+^* .*

Throughout, θ is the outer spectral factor for y . It is the uniquely determined outer function in H^2 such that $Ey(n)\overline{y(m)} = \int \bar{z}^{(n-m)} \theta \bar{\theta} dm$. Let Y be the unitary operator mapping L^2 onto \mathcal{H} defined by $Y\bar{z}^n \theta = y(n)$ for $-\infty < n < \infty$. Clearly, $YM_z = U(-1)Y$. The processes $\bar{z}^n \theta$ and $y(n)$ are unitarily equivalent. Since θ is outer, the past of $\bar{z}^n \theta$ is H^2 . (Recall $\bigvee_{n \geq 0} z^n \theta = H^2$.) The future of $\bar{z}^n \theta$ is $\bigvee_{n > 0} \bar{z}^n \theta$. Further, $YH^2 = \mathcal{H}^-$ and $Y[\bigvee_{n > 0} \bar{z}^n \theta] = \mathcal{H}^+$. Therefore, $\mathcal{X}[Y\mathcal{H}]$ is a splitting subspace for $\bar{z}^n \theta[y]$, respectively, if $H^2 \vee \mathcal{X}$ is an irreducible invariant subspace for M_z and

$$(4) \quad P\left(\bigvee_{n > 0} \bar{z}^n \theta | H^2 \vee \mathcal{X}\right) = P\left(\bigvee_{n > 0} \bar{z}^n \theta | \mathcal{X}\right).$$

$Y\mathcal{X}$ is a minimal splitting subspace for y if and only if \mathcal{X} is a minimal splitting subspace for $\bar{z}^n \theta$. In this paper we characterize all the minimal splitting subspaces for y by finding all the minimal splitting subspaces for the equivalent process $\bar{z}^n \theta$. This begins with

LEMMA 1. *Let $\mathcal{S} \supseteq H^2$. If \mathcal{S} is an irreducible invariant subspace for M_z , then there exists an inner function ψ such that $\mathcal{S} = \psi H^2$. The function ψ is uniquely determined up to a constant factor of modulus one.*

Proof. The hypothesis guarantees that $\mathcal{S} = fH^2$. The function is unique up to a constant of modulus one and $|f(z)| = 1$, a.e., [8], [13], [23]. Using $\mathcal{S} \supseteq H^2$, $fh = 1$ for some h in H^2 . Thus, $\bar{f} = h$ is in H^2 . Setting $\psi = \bar{f}$ completes the proof. (Since f has modulus one, the complex conjugate of f equals the inverse of f , i.e., $\bar{f} = 1/f$.)

LEMMA 2. *If ψ is an inner function, then*

$$(5) \quad P(\bar{z}^n h | \bar{\psi}H^2) = \bar{\psi}[U_+^{*n}(h\psi)] \quad (h \in H^2 \text{ and } n \geq 0).$$

Proof. Let $\bar{\psi}f = P(\bar{z}^n h | \bar{\psi}H^2)$, where f is in H^2 . Then $\bar{z}^n h - \bar{\psi}f$ is orthogonal to $\bar{\psi}H^2$. Equivalently, $\bar{z}^n h\psi - f$ is orthogonal to H^2 . Thus, $f = P(\bar{z}^n h\psi | H^2)$. Since $P(\bar{z}^n q | H^2) = U_+^{*n}q$ for any q in H^2 , the proof is complete.

Using Lemma 1 and (4), (5), we can prove

PROPOSITION 1. *Let $Y\mathcal{X}$ be a splitting subspace for y . The space \mathcal{X} uniquely determines an inner function ψ up to a constant of modulus one such that $H^2 \vee \mathcal{X} = \bar{\psi}H^2$. Furthermore,*

$$(6) \quad \text{cl } P\left(\bigvee_{n>0} \bar{z}^n \theta | \mathcal{X}\right) = \text{cl } P\left(\bigvee_{n>0} \bar{z}^n \theta | \bar{\psi}H^2\right) = \bar{\psi}\left[\bigvee_{n>0} U_+^{*n}(\theta\psi)\right].$$

Proposition 1 has a converse: Each inner function ψ determines at least one splitting subspace. For instance, $Y(\bar{\psi}H^2)$ is a splitting subspace.

For the moment, let θ be cyclic for U_+^* and $Y\mathcal{X}$ be a splitting subspace for y . By Proposition 1, (6) and Corollary 1, part (ii):

$$(7) \quad \bar{\psi}H^2 = \bar{\psi}\left[\bigvee_{n>0} U_+^{*n}(\theta\psi)\right] \subseteq \mathcal{X} \subseteq H^2 \vee \mathcal{X} = \bar{\psi}H^2.$$

Hence $\mathcal{X} = \bar{\psi}H^2$. The set of all splitting subspaces for y is: $\{Y\bar{\psi}H^2 | \psi \text{ is inner}\}$. Note $H^2 \subseteq \bar{\psi}H^2$ for all inner functions ψ . Therefore, the only minimal splitting subspace when θ is cyclic is $\mathcal{X}^- = YH^2$.

Now assume that θ is noncyclic for U_+^* . Let $\theta = \bar{g}\Phi$ be the unique factorization given in Theorem 1. (Unique means unique up to a constant of modulus one.) For each splitting subspace $Y\mathcal{X}$, there is an inner function ψ such that $H^2 \vee \mathcal{X} = \bar{\psi}H^2$. Clearly $\theta\psi = \bar{g}\Phi\psi = \bar{g}_e\bar{g}_i\Phi\psi$. (Throughout, $g_i[g_e]$ is the inner [outer] factor for g , respectively.) The common factors of g_i and ψ cancel. Thus, $\theta\psi = \bar{h}\Phi\delta$, where h is in H^2 , $\psi = \xi\delta$, and ξ is the greatest common inner divisor of ψ and g_i . This factorization of $\theta\psi$ is unique by Theorem 1. By Proposition 1, (3), (6) and Theorem 1,

$$(8) \quad \mathcal{X} \supseteq \text{cl } P\left(\bigvee_{n>0} \bar{z}^n \theta | \mathcal{X}\right) = \bar{\psi}H_{\Phi\delta} = \bar{\psi}[H_\delta \oplus \delta H_\Phi] = \bar{\psi}H_\delta \oplus \bar{\xi}H_\Phi.$$

([1, Lemma (3.1)] is used in the above calculation. It states $H_{\Phi\delta} = H_\delta \oplus \delta H_\Phi$ when Φ and δ are inner functions.) By (8) every splitting subspace \mathcal{X} for $\bar{z}^n\theta$ contains a space of the form $\bar{\xi}H_\Phi$, where ξ divides g_i . The following proposition shows that these are precisely the minimal splitting subspaces.

PROPOSITION 2. *Let θ be the outer spectral factor for the process $y(n)$. If θ is cyclic for the backward shift operator, then \mathcal{X}^- is the only minimal splitting subspace for y . Furthermore, assume θ is noncyclic for the backward shift and $\theta = \bar{g}\Phi$ is the unique factorization given in Theorem 1. Then the set of all minimal splitting subspaces for y is: $\{Y\bar{\xi}H_\Phi : \xi \text{ is an inner divisor of } g_i\}$.*

Proof. (ii) In light of (8), it remains to verify that $\bar{\xi}H_\Phi$ is a minimal splitting subspace for $\bar{z}^n\theta$. First it is shown that $\bar{\xi}H_\Phi$ is a splitting subspace. Since ξ divides g_i , the unique factorization of $\theta\xi$ given by Theorem 1 is $\theta\xi = \bar{h}\Phi$, where h is in H^2 .

By (3), (5) and Theorem 1,

$$(9) \quad \text{cl } P\left(\bigvee_{n>0} \bar{z}^n \theta | \bar{\xi} H^2\right) = \bar{\xi} H_\Phi.$$

Equation (9) can also be used to demonstrate that $H^2 \vee \bar{\xi} H_\Phi = \bar{\xi} H^2$:

$$(10) \quad \begin{aligned} \bar{\xi} H^2 &= P(L^2 | \bar{\xi} H^2) = P\left(\bigvee_{-\infty < n < \infty} \bar{z}^n \theta | \bar{\xi} H^2\right) \\ &= P(H^2 | \bar{\xi} H^2) \vee P\left(\bigvee_{n>0} \bar{z}^n \theta | \bar{\xi} H^2\right) = H^2 \vee \bar{\xi} H_\Phi. \end{aligned}$$

From (9) and (10), it follows that $\bar{\xi} H_\Phi$ is a splitting subspace.

Finally, it is shown that $\bar{\xi} H_\Phi$ is minimal. Let $\mathcal{W} \subseteq \bar{\xi} H_\Phi$ be a splitting subspace and ξ_1 be the corresponding inner function. By (10), $\bar{\xi}_1 H^2 = H^2 \vee \mathcal{W} \subseteq H^2 \vee \bar{\xi} H_\Phi = \bar{\xi} H^2$. Hence ξ_1 divides $\bar{\xi}$. Since ξ divides g_i , the unique factorization of $\theta \xi_1$ given by Theorem 1 is $\theta \xi_1 = \bar{h}_1 \Phi$, where h_1 is in H^2 . By (3), (6), Theorem 1 and the fact that \mathcal{W} is a splitting subspace,

$$(11) \quad \mathcal{W} \supseteq \text{cl } P\left(\bigvee_{n>0} \bar{z}^n \theta | \mathcal{W}\right) = \text{cl } P\left(\bigvee_{n>0} \bar{z}^n \theta | \bar{\xi}_1 H^2\right) = \bar{\xi}_1 H_\Phi.$$

Without loss of generality, set $\mathcal{W} = \bar{\xi}_1 H_\Phi$. (Recall $\bar{\xi}_1 H_\Phi$ is a splitting subspace.) By (11), $\bar{\xi}_1 H_\Phi \subseteq \bar{\xi}_1 \bar{\xi}_2 H_\Phi$, where $\xi_1 \xi_2 = \xi$. Thus, $\xi_2 H_\Phi \subseteq H_\Phi$. To complete the proof, it is sufficient to show that $\mathcal{W}_2 \doteq H_\Phi \ominus \xi_2 H_\Phi$ is zero. By [1, Lemma (3.1)],

$$(12) \quad \xi_2 H_\Phi \subseteq H_\Phi \subseteq H_{\Phi \xi_2} = H_{\xi_2} \oplus \xi_2 H_\Phi.$$

The functions ξ_2 and Φ are relatively prime. This follows because ξ_2 divides ξ , ξ divides g_i , and g_i , Φ are relatively prime, from Theorem 1. Equation (12) yields $\mathcal{W}_2 \subseteq H_{\xi_2} \cap H_\Phi = \{0\}$. The equality is a consequence of ξ_2 and Φ being prime (see [24, p. 123]), which completes the proof.

COROLLARY 2. *If \mathcal{X}_1 and \mathcal{X}_2 are two minimal splitting subspaces for y , then the dimension of \mathcal{X}_1 equals the dimension of \mathcal{X}_2 . The process y admits a finite-dimensional minimal splitting subspace if and only if θ is rational.*

Proof. The first part is obvious. The second follows from the fact that the dimension of $\bigvee_{n>0} U_+^{*n} \theta$ is finite if and only if θ is rational.

Example 1. Let θ be the outer function given by

$$(13) \quad \theta(z) = \prod_{k=1}^m (1 - \bar{\alpha}_k z) \cdot \prod_{k=1}^n (1 - \bar{\beta}_k z)^{-1}.$$

Suppose that the numerator and denominator have no common factors. Since θ is outer, all the poles and zeros of θ are outside of the unit disc [17]. A simple calculation gives

$$(14) \quad \theta = x \bar{z}^{(n-m)} \prod_{k=1}^n \overline{(1 - \bar{\beta}_k z)}^{-1} \cdot \prod_{k=1}^m \overline{B_{\alpha_k}(1 - \bar{\alpha}_k z)} \cdot \prod_{k=1}^n B_{\beta_k},$$

where x is a constant of modulus one and B_γ is the Blaschke factor with respect to γ [17]. Therefore, $\theta = \bar{g} \Phi$, where

$$(15) \quad \begin{aligned} \Phi &= z^{m-n} \prod_{k=1}^n B_{\beta_k} \quad \text{if } m > n, & \Phi &= \prod_{k=1}^n B_{\beta_k} \quad \text{if } m \leq n, \\ g_i &= z^{n-m} \prod_{k=1}^m B_{\alpha_k} \quad \text{if } m \leq n, & g_i &= \prod_{k=1}^m B_{\alpha_k} \quad \text{if } m > n. \end{aligned}$$

Note that g_i and Φ are relatively prime. Hence, the factorization is unique. The set of all minimal splitting subspaces for y is $Y\xi H_\Phi$, where ξ divides g_i .

In the noncyclic case the set of all minimal splitting subspaces for y have also been characterized in [19]. Aside from the fact that they work in continuous time, their approach is quite different. It relies heavily upon the geometry of splitting subspaces. To complete this section, we compare our approach to theirs.

Let us briefly summarize their results. First two spectral factors W_* and W^* for y are found. W_* is the outer spectral factor and W^* is a stable spectral factor determined by the frame space. Then two inner functions j_* and j^* are defined. j_* is the inner part of W^* and $j^* \doteq W^*/\bar{W}_*$. Using the geometry of splitting subspaces they prove [19, Thm. 3]: The set of all minimal splitting subspaces for y is unitarily equivalent to $\{jP(j^*H_2^-|H_2^+): j \text{ is an inner divisor of } j_*\}$. The Hardy space of analytic functions in the right [left] half plane is denoted by $H_2^+[H_2^-]$, respectively. It can be shown that $P(j^*H_2^-|H_2^+) = H_j^* \doteq H_2^+ \ominus j^*H_2^+$. Using this, their theorem becomes: The set of all minimal splitting subspaces for y is unitarily equivalent to $\{jH_{j^*}: j \text{ is an inner divisor of } j_*\}$. Multiplying the above spaces by \bar{j}_* yields: The set of all minimal splitting subspaces for y is unitarily equivalent to $\{\bar{j}H_{j^*}: j \text{ is an inner divisor of } j_*\}$. We will show the connection between [19, Thm. 3] and our Proposition 2.

We did not use the geometry of splitting subspaces. Our approach depends upon the backward shift operator and Theorem 1. We have shown that the set of all minimal splitting subspaces for y is unitarily equivalent to $\{\bar{\psi}H_\Phi: \psi \text{ divides } \bar{g}_i\}$. Here g_i and Φ are obtained from the factorization $\theta = \bar{g}\Phi$. Note that our approach only requires the knowledge of θ . The functions g_i , Φ are obtained by factoring θ . Lindquist-Picci use two spectral factors W_* and W^* . Their factor W^* was calculated by using the frame space. We can obtain their W^* by factoring W_* . A continuous time version of Theorem 1 verifies that W_* can be uniquely factored into $W_* = \bar{g}\Phi$. It can be shown that $\Phi = j^*$, $W^* = g$ and $j_* = g_i$. This completely exposes the relationship between [19, Thm. 3] and our Proposition 2. In [19] they have arrived at the functions $j_* = g_i$ and $j^* = \Phi$ by using splitting subspace techniques. Here we obtain the functions g_i , Φ by using the backward shift operator and factoring θ . Example 1 demonstrates that this factorization is easy to perform if θ is rational.

Finally, note that [19] does not treat the cyclic case. They have eliminated this possibility by assuming $\text{cl } P(\mathcal{H}^+|\mathcal{H}^-) \neq \mathcal{H}^-$. A cyclic θ occurs more often than one might suspect. In [6] it was shown that the set of all noncyclic vectors for the backward shift is in the first category. So, in some sense, the number of cyclic vectors far exceeds the number of noncyclic vectors.

3. State space realizations. In this section it is shown that the previous theory easily leads to a state space realization of y . For each minimal splitting subspace, a forward and a backward realization of y is given. Throughout the rest of the paper, θ is noncyclic and $\theta = \bar{g}\Phi$ is its unique factorization.

H_Φ is the minimal splitting subspace which is completely contained in the past of $\bar{z}''\theta$. Each h in H_Φ defines a wide sense stationary process by $\bar{z}''h$ for $-\infty < n < \infty$. In the time domain this process is: $x(n) = Y\bar{z}''h$. The processes determined by H_Φ admit a natural forward state space representation:

$$(16) \quad \begin{aligned} \bar{z}h &= (U_+^*|H_\Phi)h + \bar{z}h_0 \quad (h \in H_\Phi), \\ \theta &= P(\theta|H_\Phi) + \Phi\bar{g}_0. \end{aligned}$$

(Recall $f_0 \doteq (f)_0 \doteq \int f dm$ if f is in L^2 .) The first equation in (16) follows from (2) and the fact that H_Φ is an invariant subspace for U_+^* . The second equation is the orthogonal

decomposition of θ . Note that $P(\theta|\Phi H^2) = \Phi \bar{g}_0$ by using $\theta = \bar{g}\Phi$. In the time domain (16) becomes

$$(17) \quad x(1) = \tilde{A}x(0) + \tilde{B}u(1), \quad y(0) = \tilde{C}x(0) + \tilde{D}u(0),$$

where $Y\bar{z}^n = u(n)$, $\tilde{A} = Y(U_+^*|H_\Phi)Y^{-1}$, $\tilde{B}Y = Yh_0$, $\tilde{C} = P(\cdot|YH_\Phi)$ and $\tilde{D}Y = Y\Phi\bar{g}_0$. Since $u(n)$, $x(n)$ and $y(n)$ are all wide sense stationary, (17) formally becomes

$$(18) \quad x(n+1) = \tilde{A}x(n) + \tilde{B}u(n+1), \quad y(n) = \tilde{C}x(n) + \tilde{D}u(n).$$

System (18) is called the *restricted forward realization for y determined by H_Φ* . By Corollary 2, system (18) is finite dimensional if and only if θ is rational.

Notice that \tilde{A} is unitarily equivalent to $U_+^*|H_\Phi$. By [5], [24] $\tilde{\Phi}$ is the characteristic function for \tilde{A} . If f is in L^2 , then $\tilde{f}(z) \doteq \tilde{f}(\bar{z})$ a.e. The spectrum and all the invariant subspaces of \tilde{A} can be determined directly from $\tilde{\Phi}$. The spectrum of \tilde{A} is the union of the closure of the zero set of $\tilde{\Phi}$ and the closed support of the singular measure determined by $\tilde{\Phi}$. Each inner factor of $\tilde{\Phi}$ uniquely determines an invariant subspace for \tilde{A} . Furthermore, $\tilde{\Phi}$ is the minimal function for \tilde{A} . A minimal function is a generalization of the concept of a minimal polynomial used in linear algebra. In particular, this implies $\tilde{\Phi}(\tilde{A}) = 0$. For other important aspects of minimal and characteristic functions see [5], [24]. Therefore, the factorization $\theta = \bar{g}\Phi$ determines all minimal splitting subspaces for y and a Markov representation of the form (18) such that $\tilde{\Phi}$ is the minimal and characteristic function for \tilde{A} . Finally, it is noted that this factorization also plays an important role in deterministic realization theory [4], [10], [11], [12]. They also arrive at a shift realization.

Example 2. System (18) is an abstract realization of y . It degenerates into the usual autoregressive model when θ is rational. For instance, let $\theta = 1/d$, where $d = (1 - \gamma z - \beta z^2 - \alpha z^3)$ and $\alpha \neq 0$. Applying U_+^* recursively, see (2), gives $H_\Phi \doteq \bigvee_{n \geq 0} U_+^{*n}\theta = \text{span}\{1/d, z/d, z^2/d\}$. Choose $h^1 = z^2/d$, $h^2 = z/d$ and $h^3 = 1/d$ as a basis for H_Φ . Then $U_+^*h^1 = h^2$, $U_+^*h^2 = h^3$, $U_+^*h^3 = \alpha h^1 + \beta h^2 + \gamma h^3$, $h_0^1 = 0$, $h_0^2 = 0$, $h_0^3 = 1$.

By (16),

$$(19) \quad \begin{bmatrix} \bar{z}h^1 \\ \bar{z}h^2 \\ \bar{z}h^3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \alpha & \beta & \gamma \end{bmatrix} \begin{bmatrix} h^1 \\ h^2 \\ h^3 \end{bmatrix} + \bar{z} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \theta = h^3.$$

Converting to the time domain gives

$$(20) \quad \begin{bmatrix} x_1(n+1) \\ x_2(n+1) \\ x_3(n+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \alpha & \beta & \gamma \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \\ x_3(n) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} u(n+1),$$

$$y(n) = [0, 0, 1]x(n).$$

Note that (20) is only a matrix representation for the forward realization. The 3×3 matrix in (20) is similar to $U_+^*|H_\Phi$. The operator Y is used to obtain unitary equivalence.

The above method can be applied to any rational function $\theta = n/d$ to obtain a matrix representation for the restricted forward realization of y determined by H_Φ . If the polynomials n and d have no common factors, then $H_\Phi \doteq \bigvee_{i \geq 0} U_+^{*i}\theta = \text{span}\{1/d, z/d, \dots, z^k/d\}$, where k is the maximum degree of $\{n, d\}$ minus 1. Using the obvious basis and (16), one obtains an autoregressive model for y , where \tilde{A} is a

companion matrix. This agrees with the results of [22]. It is important to note that the above method works even if n and d have common factors. In this case one simply computes a basis for $H_\Phi \doteq \bigvee_{i>0} U_+^{*i} \theta$ and then proceeds as before. In fact, the roots of n and d are needed only when one is computing the characteristic or minimal function $\tilde{\Phi}$ for $U_+^*|H_\Phi$ (see Example 1).

A backward stochastic realization of y is obtained by compressing M_z to H_Φ . The compression of M_z to H_Φ is the linear operator S_Φ mapping H_Φ into H_Φ defined by $S_\Phi h \doteq P(zh|H_\Phi)$ when h is in H_Φ [5], [14]. Note S_Φ is the adjoint of $U_+^*|H_\Phi$. The orthogonal decomposition of zh is $zh = S_\Phi h + P(zh|\Phi H^2)$ for h in H_Φ . Using $P(zh|\Phi H^2) = \Phi(z\tilde{\Phi}h)_0$ gives

$$(21) \quad \begin{aligned} zh &= S_\Phi h + \Phi(z\tilde{\Phi}h)_0 \quad (h \in H_\Phi), \\ \theta &= P(\theta|H_\Phi) + \Phi\tilde{g}_0. \end{aligned}$$

In the time domain, (21) formally becomes

$$(22) \quad x(n-1) = Ax(n) + Bu(n), \quad y(n) = Cx(n) + Du(n),$$

where $AY = YS_\Phi$, $BY = Y\Phi(z\tilde{\Phi}h)_0$, etc. System (22) is called the *restricted backward stochastic realization for y determined by H_Φ* . The characteristic and minimal function for A is Φ [5], [24]. To summarize,

PROPOSITION 3. *Let θ be the outer spectral factor for y and let $\theta = \tilde{g}\Phi$ be its unique decomposition. The process y admits a forward [backward] stochastic realization of the form (18) [22] where $\tilde{\Phi}[\Phi]$ is the characteristic and minimal function for $\tilde{A}[A]$, respectively.*

Remark 1. A restricted forward and backward stochastic realization can be obtained for any minimal splitting subspace. For instance, if h is in $\tilde{\xi}H_\Phi$, where ξ is an inner divisor of g , then

$$(23) \quad \begin{aligned} \tilde{z}h &= \tilde{\xi}(U_+^*|H_\Phi)\xi h + \tilde{z}\tilde{\xi}(\xi h)_0 \quad (h \in \tilde{\xi}H_\Phi), \\ \theta &= P(\theta|\tilde{\xi}H_\Phi) + \tilde{\xi}\Phi(\xi\tilde{g})_0. \end{aligned}$$

Transforming to the time domain,

$$(24) \quad x(n+1) = \tilde{A}_\xi x(n) + \tilde{B}_\xi u(n+1), \quad y(n) = \tilde{C}_\xi x(n) + \tilde{D}_\xi u(n),$$

where \tilde{A}_ξ is unitarily equivalent to $M_\xi U_+^* M_\xi| \tilde{\xi}H_\Phi$. Hence \tilde{A}_ξ is unitarily equivalent to $U_+^*|H_\Phi$. The characteristic and minimal function for \tilde{A} is $\tilde{\Phi}$ [5], [24]. System (24) is called the *restricted forward stochastic realization for y determined by $\tilde{\xi}H_\Phi$* .

All restricted forward stochastic realizations determined by any minimal splitting subspace are unitarily equivalent. Two forward [backward] systems $\{\tilde{A}_\xi, \tilde{B}_\xi, \tilde{C}_\xi, \tilde{D}_\xi\}$, $[\{\tilde{A}_{\xi_1}, \tilde{B}_{\xi_1}, \tilde{C}_{\xi_1}, \tilde{D}_{\xi_1}\}]$ are *unitarily equivalent* if there exists a unitary operator V from $Y\tilde{\xi}H_\Phi$ onto $Y\tilde{\xi}_1H_\Phi$ such that $V\tilde{A}_\xi = \tilde{A}_{\xi_1}V$ and $V\tilde{B}_\xi = \tilde{B}_{\xi_1}$. In this definition no stipulations on the \tilde{C}_ξ and \tilde{D}_ξ terms are given, because these operators are always derived from the appropriate orthogonal decomposition. In fact, the unitary operator V intertwining these two forward systems is $V = YM_{\xi_1}M_\xi Y^{-1}$. In a similar manner one can obtain the restricted backward stochastic realization for y determined by $\tilde{\xi}H_\Phi$. (In fact, the one step model for $\tilde{\xi}H_\Phi$ is: $zh = \tilde{\xi}S_\Phi \xi h + \tilde{\xi}\Phi(z\tilde{\Phi}\xi h)_0$, $\theta = P(\theta|\tilde{\xi}H_\Phi) + \tilde{\xi}\Phi(\xi\tilde{g})_0$ when $h \in \tilde{\xi}H_\Phi$.) As expected, all restricted backward realizations are unitarily equivalent.

4. Duality and stochastic realizations. Clearly there is a dual relationship between the forward and backward stochastic realizations for y . More precisely, in this section it is shown that the restricted forward [backward] realizations for $y(n)$ are unitarily equivalent to the restricted backward [forward] realizations for $y(-n)$, respectively.

Let $w(n) = y(-n)$ for all n . The outer spectral factor for the process w is $\tilde{\theta}$. Recall that θ is outer if and only if $\tilde{\theta}$ is outer [24]. Furthermore, $\tilde{\theta} = \tilde{g}\tilde{\Phi}$ is the unique factorization for $\tilde{\theta}$ given by Theorem 1. The set of all minimal splitting subspaces for w is: $\{W\tilde{\xi}H_{\Phi}; \xi \text{ is an inner divisor of } \tilde{g}_i\}$. Here W is the unitary operator mapping L^2 onto \mathcal{H} defined by $Wz^n\tilde{\theta} = w(n)$.

Let J be the unitary operator from L^2 onto L^2 defined by $Jf(z) = \bar{z}\Phi(z)f(\bar{z})$ a.e. In [9] it is shown that J maps H_{Φ} onto H_{Φ} and intertwines $U_+^*|H_{\Phi}$ with S_{Φ} :

$$(25) \quad JH_{\Phi} = H_{\Phi} \quad \text{and} \quad J(U_+^*|H_{\Phi}) = S_{\Phi}(J|H_{\Phi}).$$

The unitary operator $J\tilde{g}_i$ mapping L^2 onto L^2 is the operator that ‘‘reverses time’’. (A continuous time version of this operator has been used in [19].) The operator $J\tilde{g}_i$ maps the past of [future of] $\bar{z}^n\tilde{\theta}$ onto the future of [past of] $\bar{z}^n\theta$, respectively. More precisely,

PROPOSITION 4. *Let θ be the outer spectral factor for y and $\theta = \bar{g}\Phi$ be its unique factorization. Then*

- (i) $J\tilde{g}_i[\tilde{g}_iH_{\Phi}] = H_{\Phi}$;
- (ii) $J\tilde{g}_iH^2 = \bigvee_{n>0} \bar{z}^n\theta$;
- (iii) $J\tilde{g}_i[\bigvee_{n>0} \bar{z}^n\tilde{\theta}] = H^2$.

Proof. Part (i) follows from (25). Notice that $|\theta| = |g|$ a.e. Hence $\theta = g_e$, the outer part of g . Using $\tilde{\theta} = \tilde{g}\tilde{\Phi} = \tilde{\theta}\tilde{g}_i\tilde{\Phi}$, it is easy to verify that $J\tilde{g}_i\bar{z}^n\tilde{\theta} = \bar{z}^{(n+1)}\theta$ for $-\infty < n < \infty$. Since θ and $\tilde{\theta}$ are both outer, the last equality implies (ii) and (iii). (Recall that θ is outer if and only if $\bigvee_{n \geq 0} z^n\theta = H^2$.)

From the above proposition, it is clear that $J\tilde{g}_i$ maps the *future minimal splitting* subspace \tilde{g}_iH_{Φ} for $\bar{z}^n\tilde{\theta}$ onto the minimal splitting subspace H_{Φ} contained in the *past* of $\bar{z}^n\theta$. Consider the state equation for the restricted forward realization of $\bar{z}^n\tilde{\theta}$ determined by \tilde{g}_iH_{Φ} :

$$(26) \quad \bar{z}f = \tilde{g}_i(U_+^*|H_{\Phi})\tilde{g}_if + \bar{z}\tilde{g}_i(\tilde{g}_if)_0 \quad (f \in \tilde{g}_iH_{\Phi}).$$

In the time domain, (26) formally becomes

$$(27) \quad x(n+1) = \tilde{A}_F x(n) + \tilde{B}_F u(n+1).$$

For simplicity we will ignore the output terms in (26), (27), which correspond to the appropriate orthogonal decomposition. Here \tilde{A}_F is unitarily equivalent to $U_+^*|H_{\Phi}$. The characteristic and minimal function for A_F is Φ (note $\tilde{\Phi} = \Phi$). Applying $J\tilde{g}_i$ to (26) yields the state equation (21). To verify this set $h = J\tilde{g}_if$, $f = \tilde{g}_i h'$ and use (25):

$$J\tilde{g}_i\bar{z}\tilde{g}_i(\tilde{g}_if)_0 = \Phi(h')_0 = \Phi(J^{-1}h)_0 = \Phi(\bar{z}\tilde{\Phi}h(\bar{z}))_0 = \Phi(\bar{z}\Phi(\bar{z})h(\bar{z}))_0 = \Phi(z\tilde{\Phi}h)_0.$$

Thus, (27) and (22) are unitarily equivalent. A dual argument on the backward realization gives

PROPOSITION 5. *Let $w(n) = y(-n)$ for all n . The restricted forward [backward] stochastic realization for w determined by \tilde{g}_iH_{Φ} is unitarily equivalent to the restricted backward [forward] stochastic realization for y determined by H_{Φ} , respectively.*

Since all forward [backward] realizations for y determined by any minimal splitting subspace are unitarily equivalent (see Remark 1), we get

COROLLARY 3. *All restricted forward [backward] stochastic realizations for y are unitarily equivalent to all restricted backward [forward] stochastic realizations for w , respectively.*

We have derived a shift realization for a stochastic process $y(n)$. We shall show some similarities and differences between deterministic and stochastic realizations.

Consider the following system:

$$(\Sigma) \quad x(n+1) = Ax(n) + Bu(n), \quad y(n) = Cx(n) + Du(n),$$

where A, B, C, D are bounded linear operators on the appropriate space. For simplicity $u(n)$ and $y(n)$ are assumed to be scalar-valued sequences. System Σ is denoted by $\{A, B, C, D\}$. The transfer function Ω_Σ for Σ is defined by $\Omega_\Sigma \doteq D + zC(I - zA)^{-1}B$. Without loss of generality, it is assumed that Σ is stable, i.e., Ω_Σ is in H^2 . The deterministic realization problem is: Given Ω in H^2 , find a minimal system $\Sigma = \{A, B, C, D\}$ such that Ω is the transfer function for Σ , i.e., $\Omega = \Omega_\Sigma$. Following [4], [10], [11], [12], [16], a solution to this problem is:

$$(28) \quad \begin{aligned} x(n+1) &= (U_+^*|H_\Phi)x(n) + (U_+^*\Omega)u(n), \\ y(n) &= P_0x(n) + \Omega_0u(n), \end{aligned}$$

where Φ is the inner function defined by $H_\Phi = \bigvee_{n>0} U_+^{*n}\Omega$. The operator P_0 mapping H_Φ into the complex number is given by $P_0f \doteq (f)_0 \doteq \int f dm$.

Let us compare the stochastic realization (16) with the deterministic realization (28). Both contain operators of the form $U_+^*|H_\Phi$. The inner function $\tilde{\Phi}$ is the minimal and characteristic function for $U_+^*|H_\Phi$. The spectrum and all the invariant subspaces for $U_+^*|H_\Phi$ are determined by $\tilde{\Phi}$. However, (16) and (28) are different realizations. The operator P_0 does not correspond to $P(\theta|H_\Phi)$ and $\bar{z}h_0$ does not correspond to $U_+^*\Omega$. Furthermore, stochastic realizations evolve forward and backward in time. Deterministic realizations only move forward in time. Here Theorem 1 was used to prove Proposition 2 and classify all minimal splitting subspaces. In deterministic realization theory, Theorem 1 is used to compute Φ in $H_\Phi = \bigvee_{n>0} U_+^{*n}\Omega$. Theorem 1 is also used to study series and parallel coupling of linear systems [12]. Finally, for deterministic realizations, one computes $H_\Phi \doteq \bigvee_{n>0} U_+^{*n}\Omega$ where Ω can be any arbitrary function in H^2 . Our $H_\Phi \doteq \bigvee_{n>0} U_+^{*n}\theta$ is always determined by an outer function θ . We have demonstrated several important differences between stochastic and deterministic realizations. In a future paper on vector processes, it will be shown that this gap is far wider than one might suspect.

REFERENCES

- [1] P. R. AHERN AND D. N. CLARK, *On functions orthogonal to invariant subspaces*, Acta Math., 124 (1970), pp. 191–204.
- [2] H. AKAIKE, *Stochastic theory of minimal realizations*. IEEE Trans. Automatic Control, AC-19 (1974), pp. 667–674.
- [3] F. A. BADAWI, A. LINDQUIST AND M. PAVON, *A stochastic realization approach to the smoothing problem*, IEEE Trans. Automatic Control, AC-24 (1979), pp. 878–888.
- [4] J. S. BARAS AND R. W. BROCKETT, *H^2 -functions and infinite dimensional realization theory*, this Journal, 13 (1975), pp. 221–241.
- [5] R. G. DOUGLAS, *Canonical models*, in Topics in Operator Theory, C. Pearcy, ed., Mathematical Surveys, 13, American Mathematical Society, Providence, RI, 1974.
- [6] R. G. DOUGLAS, H. S. SHAPIRO AND A. L. SHIELDS, *Cyclic vectors and invariant subspaces for the backward shift*. Ann. Inst. Fourier, Grenoble, 20 (1971), pp. 37–76.
- [7] P. FAURRE, *Stochastic realization algorithms*, in System Identification: Advances and Case Studies, R. K. Mehra and D. G. Lainiotis, eds., Academic Press, New York, 1976.
- [8] P. A. FILLMORE, *Notes on Operator Theory*. Van Nostrand, New York, 1970.
- [9] P. A. FUHRMANN, *On the corona theorem and its applications to spectral problems in Hilbert space*. Trans. Amer. Math. Soc., 132 (1968), pp. 55–66.
- [10] ———, *On realization of linear systems and applications to some questions on stability*, Math. Systems Theory, 8 (1974), pp. 132–141.

- [11] ———, *Realization theory in Hilbert space for a class of transfer functions*, J. Functional Anal., 18 (1975), pp. 338–349.
- [12] ———, *On series and parallel coupling of a class of discrete time infinite-dimensional systems*, this Journal, 14 (1976), pp. 339–358.
- [13] P. R. HALMOS, *Shifts on Hilbert spaces*, J. Reine Angew. Math., 208 (1961), pp. 102–112.
- [14] ———, *A Hilbert Space Problem Book*, Van Nostrand, New York, 1967.
- [15] H. HELSON, *Lectures on Invariant Subspaces*, Academic Press, New York, 1964.
- [16] J. W. HELTON, *Discrete time systems, operator models and scattering theory*, J. Functional Analysis, 16 (1974), pp. 15–38.
- [17] K. HOFFMAN, *Banach Spaces of Analytic Functions*, Prentice-Hall, Englewood Cliffs, NJ, 1962.
- [18] A. LINDQUIST AND G. PICCI, *On the stochastic realization problem*, this Journal, 17 (1970), pp. 365–389.
- [19] ———, *A Hardy space approach to the stochastic realization problem*, Proc. 1978 Conference on Decision and Control, San Diego, CA, Jan. 1979.
- [20] A. LINDQUIST, G. PICCI AND R. RUCKEBUSCH, *On minimal splitting subspaces and Markovian representations*, Math. Syst. Theory, 12 (1979), pp. 271–279.
- [21] M. PAVON, *Stochastic realization and invariant directions of the matrix Riccati equation*, this Journal, 18 (1980), pp. 155–180.
- [22] Y. A. ROZANOV, *On two selected topics connected with stochastic systems theory*, Appl. Math. Optim., 3 (1976), pp. 73–80.
- [23] D. SARASON, *Invariant subspaces*, in Topics in Operator Theory, C. Pearcy, ed., Mathematical Surveys 13, American Mathematical Society, Providence, RI, 1974.
- [24] B. SZ-NAGY AND C. FOIAS, *Harmonic Analysis of Operators on Hilbert Space*, North-Holland, Amsterdam, 1970.

TANGENT SETS' CALCULUS AND NECESSARY CONDITIONS FOR EXTREMALITY*

CORNELIU URSESCU†

Abstract. In this paper we deduce some relations concerning the tangent sets to the inverse image of a set under a function and we use these relations to derive necessary conditions for extremality in the form of a multiplier rule.

1. Introduction. We have considered in [23, p. 151] the tangent sets $L_{X_0}(x_0)$, $l_{X_0}(x_0)$, $k_{X_0}(x_0)$ and $K_{X_0}(x_0)$, where X_0 is a subset and x_0 is a point both of the same linear topological space X (throughout this paper “linear space” means “real linear space” and “topological space” means “Hausdorff topological space”). The definitions of these tangent sets are as follows:

$$L_{X_0}(x_0) = \{x \in X; \exists U \in \mathcal{U}, \exists r > 0, \forall s \in (0, r), \forall u \in U, x_0 + s(x + u) \in X_0\},$$

$$l_{X_0}(x_0) = \{x \in X; \exists U \in \mathcal{U}, \forall r > 0, \exists s \in (0, r), \forall u \in U, x_0 + s(x + u) \in X_0\},$$

$$k_{X_0}(x_0) = \{x \in X; \forall U \in \mathcal{U}, \exists r > 0, \forall s \in (0, r), \exists u \in U, x_0 + s(x + u) \in X_0\},$$

$$K_{X_0}(x_0) = \{x \in X; \forall U \in \mathcal{U}, \forall r > 0, \exists s \in (0, r), \exists u \in U, x_0 + s(x + u) \in X_0\},$$

where \mathcal{U} is the family of all neighborhoods of the origin in X . Then we have established several inclusions relating to the tangent sets $L_{f^{-1}(Y_0)}(x_0)$, $l_{f^{-1}(Y_0)}(x_0)$, $k_{f^{-1}(Y_0)}(x_0)$ and $K_{f^{-1}(Y_0)}(x_0)$ where X_0 and Y_0 are nonempty subsets of the linear topological spaces X and Y , respectively, f is a function from X_0 into Y , x_0 is a point of X_0 and $f^{-1}(Y_0)$ denotes the inverse image of Y_0 under f . With that end in view we have defined the differentiability of f at x_0 and, supposing f to be differentiable at x_0 , we have established in [23, pp. 154–155] the following relations:

- (a) $L_{X_0}(x_0) \cap (D_f(x_0))^{-1}(L_{Y_0}(f(x_0)))$
 $\subseteq L_{f^{-1}(Y_0)}(x_0)$
 $\subseteq L_{X_0}(x_0) \cap (D_f(x_0))^{-1}(k_{Y_0}(f(x_0))),$
- (b) $l_{X_0}(x_0) \cap (D_f(x_0))^{-1}(L_{Y_0}(f(x_0)))$
 $\subseteq l_{f^{-1}(Y_0)}(x_0)$
 $\subseteq l_{X_0}(x_0) \cap (D_f(x_0))^{-1}(K_{Y_0}(f(x_0))),$
- (c) $k_{X_0}(x_0) \cap (D_f(x_0))^{-1}(L_{Y_0}(f(x_0)))$
 $\subseteq k_{f^{-1}(Y_0)}(x_0)$
 $\subseteq k_{X_0}(x_0) \cap (D_f(x_0))^{-1}(k_{Y_0}(f(x_0))),$
- (d) $(D_f(x_0))^{-1}(L_{Y_0}(f(x_0))) \subseteq K_{f^{-1}(Y_0)}(x_0) \subseteq (D_f(x_0))^{-1}(K_{Y_0}(f(x_0))),$
- (e) $L_{X_0}(x_0) \cap (D_f(x_0))^{-1}(l_{Y_0}(f(x_0)))$
 $\subseteq L_{X_0}(x_0) \cap l_{f^{-1}(Y_0)}(x_0)$
 $\subseteq L_{X_0}(x_0) \cap (D_f(x_0))^{-1}(K_{Y_0}(f(x_0))),$

* Received by the editors August 9, 1979, and in final revised form July 1, 1981.

† Institutul de Matematică, Universitatea Al. I. Cuza, 6600 Iași, Romania.

$$\begin{aligned}
 (f) \quad & k_{X_0}(x_0) \cap (D_f(x_0))^{-1}(l_{Y_0}(f(x_0))) \\
 & \subseteq k_{X_0}(x_0) \cap K_{f^{-1}(Y_0)}(x_0) \\
 & \subseteq k_{X_0}(x_0) \cap (D_f(x_0))^{-1}(K_{Y_0}(f(x_0))),
 \end{aligned}$$

where $D_f(x_0)$ denotes the differential of f at x_0 .

The inclusions $L_{Y_0}(f(x_0)) \subseteq l_{Y_0}(f(x_0)) \subseteq K_{Y_0}(f(x_0))$ and $L_{Y_0}(f(x_0)) \subseteq k_{Y_0}(f(x_0)) \subseteq K_{Y_0}(f(x_0))$ suggest that the best improvements we can obtain are the following relations:

$$\begin{aligned}
 (A) \quad & L_{X_0}(x_0) \cap (D_f(x_0))^{-1}(L_{Y_0}(f(x_0))) = L_{f^{-1}(Y_0)}(x_0), \\
 (B) \quad & l_{f^{-1}(Y_0)}(x_0) \subseteq l_{X_0}(x_0) \cap (D_f(x_0))^{-1}(l_{Y_0}(f(x_0))), \\
 (C) \quad & k_{f^{-1}(Y_0)}(x_0) = k_{X_0}(x_0) \cap (D_f(x_0))^{-1}(k_{Y_0}(f(x_0))), \\
 (D) \quad & (D_f(x_0))^{-1}(k_{Y_0}(f(x_0))) \subseteq K_{f^{-1}(Y_0)}(x_0), \\
 (E) \quad & L_{X_0}(x_0) \cap (D_f(x_0))^{-1}(l_{Y_0}(f(x_0))) = L_{X_0}(x_0) \cap l_{f^{-1}(Y_0)}(x_0), \\
 (F) \quad & k_{X_0}(x_0) \cap K_{f^{-1}(Y_0)}(x_0) = k_{X_0}(x_0) \cap (D_f(x_0))^{-1}(K_{Y_0}(f(x_0))).
 \end{aligned}$$

In what follows we shall prove each of these relations under appropriate hypotheses (see Theorems 1, 2, 3 and 4). The basic tool in this attempt is that of a generalized concave function (see Definition 1). Then we shall define the notion of a generalized extremal point (see Definition 2), and we shall derive necessary conditions for extremality in form of a multiplier rule (see Theorem 5). Finally we shall apply these results to programming problems (see Definitions 3 and 4, and Theorems 6, 7 and 8).

Before we continue our exposition, let us remember that: the set $K_{X_0}(x_0)$ is closely related to a tangency concept which was first introduced by Bouligand [4, p. 215] and Severi [19, p. 99]; the differential $D_f(x_0)$ was first introduced by Severi [20, p. 10]; the set $L_{X_0}(x_0)$ is due to Dubovickii and Miljutin [6, p. 399]; the right-hand side of the relationship (d) was proved by Varaiya [24, p. 287].

2. On algebraic and topological interior and closure. Let X be a linear space and let X_0 be a subset of X . We shall denote by $\text{ain } X_0$ and $\text{acl } X_0$ the *algebraic interior* and the *algebraic closure* of X_0 , respectively, i.e.,

$$\begin{aligned}
 \text{ain } X_0 &= \{x_0 \in X; \forall x \in X, \exists r > 0, \forall s \in (0, r), x_0 + sx \in X_0\}, \\
 \text{acl } X_0 &= \{x_0 \in X; \exists x \in X, \forall r > 0, \exists s \in (0, r), x_0 + sx \in X_0\}.
 \end{aligned}$$

Obviously $\text{comp ain } X_0 = \text{acl comp } X_0$ and $\text{comp acl } X_0 = \text{ain comp } X_0$, where $\text{comp } X_0$ denotes the complement of X_0 with respect to X . We have also $\text{ain } X_0 \subseteq X_0 \subseteq \text{acl } X_0$.

The definition of the set $\text{ain } X_0$ is due to Klee [13, p. 445]. He also considered in [13, p. 448] the set

$$\text{lin } X_0 = \{x_0 \in X; \exists x \in X, \exists r > 0, \forall s \in (0, r), x_0 + sx \in X_0\}.$$

Clearly $X_0 \subseteq \text{lin } X_0 \subseteq \text{acl } X_0$. Moreover if X_0 is convex, then $\text{lin } X_0 = \text{acl } X_0$.

LEMMA 1. Let $\emptyset \neq X_1 \subseteq X_2 \subseteq X$ and let $t_1X_1 + t_2X_2 \subseteq X_1$ whenever $t_1 > 0$, $t_2 > 0$ and $t_1 + t_2 = 1$. Then $\text{ain } X_1 = \text{ain } X_2$ and $\text{acl } X_1 = \text{acl } X_2$.

Proof. Since $X_1 \subseteq X_2$, we have $\text{ain } X_1 \subseteq \text{ain } X_2$ and $\text{acl } X_1 \subseteq \text{acl } X_2$. Before we prove the inverse inclusions let us make some notation. Denote $S^* = \{s/(1+s); s \in S\}$ for $S \subseteq (0, +\infty)$. Denote $S_1 = \{s > 0; x_0 + sx_1 \in X_1\}$ and $S_2 = \{s > 0; x_0 + sx_2 \in X_2\}$ for $x_0 \in X$, $x_1 \in X$ and $x_2 \in X$. We assert that $(S_2)^* \subseteq S_1$ whenever $x_0 + x_1 \in X_1 + x_2$. Indeed

if $s \in S_2$, then

$$\begin{aligned} x_0 + (s(1+s))x_1 &= (1/(1+s))x_0 + (s/(1+s))(x_0 + x_1) \in (1/(1+s))x_0 + (s/(1+s))(X_1 + x_2) \\ &= (s/(1+s))X_1 + (1/(1+s))(x_0 + sx_2) \subseteq (s/(1+s))X_1 \\ &\quad + (1/(1+s))X_2 \subseteq X_1, \end{aligned}$$

hence $s/(1+s) \in S_1$ and $(S_2)^* \subseteq S_1$.

Now let us turn to the proof of the lemma. First let $x_0 \in \text{ain } X_2$, let $x_1 \in X$ and consider $x_2 \in X$ such that $x_0 + x_1 \in X_1 + x_2$ (recall that $\emptyset \neq X_1$). According to the definition of the set $\text{ain } X_2$, $(0, r) \subseteq S_2$ for some $r > 0$. Because of the definition of the set $(S_2)^*$, $(0, r) \subseteq (S_2)^*$ for some $r > 0$. Consequently $(0, r) \subseteq S_1$ for some $r > 0$, $x_0 \in \text{ain } X_1$, $\text{ain } X_2 \subseteq \text{ain } X_1$ and $\text{ain } X_1 = \text{ain } X_2$. Second let $x_0 \in \text{acl } X_2$. According to the definition of the set $\text{acl } X_2$, there is $x_2 \in X$ such that $\emptyset \neq (0, r) \cap S_2$ for every $r > 0$. Consider $x_1 \in X$ such that $x_0 + x_1 \in X_1 + x_2$ (recall that $\emptyset \neq X_1$). Because of the definition of the set $(S_2)^*$, $\emptyset \neq (0, r) \cap (S_2)^*$ for every $r > 0$. Consequently $\emptyset \neq (0, r) \cap S_1$ for every $r > 0$, $x_0 \in \text{acl } X_1$, $\text{acl } X_2 \subseteq \text{acl } X_1$ and $\text{acl } X_1 = \text{acl } X_2$. \square

Next we shall transpose Lemma 1 in the setting of linear topological spaces. Suppose X is a linear topological space. We shall denote by $\text{tin } X_0$ and $\text{tcl } X_0$ the *topological interior* and the *topological closure* of X_0 , respectively. Obviously $\text{comp tin } X_0 = \text{tcl comp } X_0$ and $\text{comp tcl } X_0 = \text{tin comp } X_0$. We have also $\text{tin } X_0 \subseteq \text{ain } X_0 \subseteq X_0 \subseteq \text{acl } X_0 \subseteq \text{tcl } X_0$.

LEMMA 2. *Let $\emptyset \neq X_1 \subseteq X_2 \subseteq X$ and let $t_1 X_1 + t_2 X_2 \subseteq X_1$ whenever $t_1 > 0$, $t_2 > 0$ and $t_1 + t_2 = 1$. Then $\text{tin } X_1 = \text{tin } X_2$ and $\text{tcl } X_1 = \text{tcl } X_2$.*

Proof. It follows from Lemma 1 that $\text{tin ain } X_1 = \text{tin ain } X_2$ and $\text{tcl acl } X_1 = \text{tcl acl } X_2$. But $\text{tin ain } X_1 \subseteq \text{tin } X_1 \subseteq \text{tin } X_2 = \text{tin tin } X_2 \subseteq \text{tin ain } X_2$ and $\text{tcl acl } X_1 \subseteq \text{tcl tcl } X_1 = \text{tcl } X_1 \subseteq \text{tcl } X_2 \subseteq \text{tcl acl } X_2$, which proves the lemma. \square

Let us recall some well-known results which can be easily derived using Lemmas 1 and 2. If X_0 is convex and $\emptyset \neq \text{ain } X_0$, then $\text{ain ain } X_0 = \text{ain acl } X_0$ and $\text{acl ain } X_0 = \text{acl acl } X_0$. This follows from Lemma 1 since $\emptyset \neq \text{ain } X_0 \subseteq \text{acl } X_0$ and $t_1 \text{ain } X_0 + t_2 \text{acl } X_0 = \text{ain } (t_1 X_0) + \text{acl } (t_2 X_2) = \text{ain } (t_1 X_0) + \text{lin } (t_2 X_0) \subseteq \text{ain } (t_1 X_0 + t_2 X_0) \subseteq \text{ain } X_0$ whenever $t_1 > 0$, $t_2 > 0$ and $t_1 + t_2 = 1$.

Further, if X_0 is convex and $\emptyset \neq \text{tin } X_0$, then $\text{tin } X_0 = \text{ain } X_0$, $\text{acl } X_0 = \text{tcl } X_0$, $\text{tin } X_0 = \text{tin tcl } X_0$ and $\text{tcl tin } X_0 = \text{tcl } X_0$ (see Klee [13, p. 448]). This follows from Lemmas 1 and 2 since $t_1 \text{tin } X_0 + t_2 \text{tcl } X_0 = \text{tin } (t_1 X_0) + \text{tcl } (t_2 X_0) \subseteq \text{tin } (t_1 X_0 + t_2 X_0) \subseteq \text{tin } X_0$ whenever $t_1 > 0$, $t_2 > 0$ and $t_1 + t_2 = 1$.

3. Convex and concave functions. Let X and Y be linear spaces, let X_0 and Y_0 be nonempty subsets of X and Y , respectively, and let f be a function from X_0 into Y . We shall denote by $\text{ep}_{Y_0}(f)$ and $\text{hp}_{Y_0}(f)$ the Y_0 -epigraph and the Y_0 -hypograph of f , respectively, i.e.,

$$\text{ep}_{Y_0}(f) = \{(x, f(x) + y); x \in X_0, y \in Y_0\},$$

$$\text{hp}_{Y_0}(f) = \{(x, f(x) - y); x \in X_0, y \in Y_0\}.$$

DEFINITION 1. We shall say that the function f is Y_0 -convex if the set $\text{ep}_{Y_0}(f)$ is convex. We shall say that the function f is Y_0 -concave if the set $\text{hp}_{Y_0}(f)$ is convex.

Observe that $\text{ep}_{Y_0}(f) = \text{hp}_{(-Y_0)}(f)$, hence f is Y_0 -convex if and only if it is $(-Y_0)$ -concave. In the sequel we shall consider only concave functions.

LEMMA 3. *The function f is Y_0 -concave if and only if the sets X_0 and Y_0 are convex and $t_1 f(x_1) + t_2 f(x_2) - Y_0 \subseteq f(t_1 x_1 + t_2 x_2) - Y_0$ whenever $x_1 \in X_0$, $x_2 \in X_0$, $t_1 \geq 0$, $t_2 \geq 0$ and $t_1 + t_2 = 1$.*

Proof “only if”. Let $x_1 \in X_0, x_2 \in X_0, y_1 \in Y_0, y_2 \in Y_0, y \in Y_0, t_1 \geq 0, t_2 \geq 0$ and $t_1 + t_2 = 1$. Since $(x_1, f(x_1) - y_1) \in \text{hp}_{Y_0}(f)$ and $(x_2, f(x_2) - y_2) \in \text{hp}_{Y_0}(f)$, we have $(t_1 x_1 + t_2 x_2, t_1 f(x_1) + t_2 f(x_2) - t_1 y_1 - t_2 y_2) = t_1(x_1, f(x_1) - y_1) + t_2(x_2, f(x_2) - y_2) \in \text{hp}_{Y_0}(f)$ (see Definition 1); hence $t_1 x_1 + t_2 x_2 \in X_0$ and $t_1 f(x_1) + t_2 f(x_2) - t_1 y_1 - t_2 y_2 \in f(t_1 x_1 + t_2 x_2) - Y_0$. First of all this means that X_0 is convex. Further, taking $x_1 = x_2$ we obtain that Y_0 is convex. Finally, taking $y_1 = y$ and $y_2 = y$ we complete the proof of the “only if” part.

“if”. Let $(x_1, y_1) \in \text{hp}_{Y_0}(f), (x_2, y_2) \in \text{hp}_{Y_0}(f), t_1 \geq 0, t_2 \geq 0$ and $t_1 + t_2 = 1$. Since $y_1 \in f(x_1) - Y_0$ and $y_2 \in f(x_2) - Y_0$, we have $t_1 y_1 + t_2 y_2 \in t_1 f(x_1) + t_2 f(x_2) - (t_1 Y_0 + t_2 Y_0) \subseteq f(t_1 x_1 + t_2 x_2) - Y_0$. Hence $t_1(x_1, y_1) + t_2(x_2, y_2) = (t_1 x_1 + t_2 x_2, t_1 y_1 + t_2 y_2) \in \text{hp}_{Y_0}(f)$, the set $\text{hp}_{Y_0}(f)$ is convex and the function f is Y_0 -concave (see Definition 1), which proves the “if” part. \square

If $0 \in Y_0$ and $Y_0 + Y_0 \subseteq Y_0$, then the function f is Y_0 -concave if and only if the sets X_0 and Y_0 are convex (note that in this case Y_0 is a cone, i.e., $ty \in Y_0$ for all $t > 0$ and $y \in Y_0$) and $t_1 f(x_1) + t_2 f(x_2) \in f(t_1 x_1 + t_2 x_2) - Y_0$ whenever $x_1 \in X_0, x_2 \in X_0, t_1 \geq 0, t_2 \geq 0$ and $t_1 + t_2 = 1$. We recognize here a definition of Hurwicz [11, p. 68]. The following considerations will clear up the connection between Definition 1 and the definition of Hurwicz.

Consider a nonempty subset Y_1 of Y and denote $Y_2 = \bigcap_{y_1 \in Y_1} (Y_1 - y_1)$. We assert that: $0 \in Y_2; Y_2 + Y_2 \subseteq Y_2$; the function f is Y_1 -concave if and only if it is Y_2 -concave and the set Y_1 is convex. Before we justify these assertions we note that $y_2 \in Y_2$ if and only if $y_2 + Y_1 \subseteq Y_1$. Thus, obviously, $0 \in Y_2$. Further, $(Y_2 + Y_2) + Y_1 = Y_2 + (Y_2 + Y_1) \subseteq Y_2 + Y_1 \subseteq Y_1$, hence $Y_2 + Y_2 \subseteq Y_2$. Finally, according to Lemma 3, a function f is Y_1 -concave if and only if the sets X_0 and Y_1 are convex (hence the set Y_2 is convex, too) and $f(t_1 x_1 + t_2 x_2) - t_1 f(x_1) - t_2 f(x_2) \in Y_2$ whenever $x_1 \in X_0, x_2 \in X_0, t_1 \geq 0, t_2 \geq 0$ and $t_1 + t_2 = 1$, which substantiates our assertions.

Subsequently we shall study some properties of concave functions in linear topological spaces. Suppose X and Y are linear topological spaces. We shall be concerned mainly with the set $\text{tin}(f(X_0) - Y_0)$.

LEMMA 4. Let f be Y_0 -concave. If $\emptyset \neq \text{tin } Y_0$, then $\text{tin}(f(X_0) - Y_0) = f(X_0) - \text{tin } Y_0$. Moreover if $\emptyset \neq \text{tin } X_0$, then $\text{tin}(f(X_0) - Y_0) = f(\text{tin } X_0) - \text{tin } Y_0$.

Proof. Let $\emptyset \neq \text{tin } Y_0$. Clearly $\emptyset \neq f(X_0) - \text{tin } Y_0 \subseteq f(X_0) - \text{tcl } Y_0$. Further if $t_1 > 0, t_2 > 0$ and $t_1 + t_2 = 1$, then

$$\begin{aligned} t_1(f(x_1) - \text{tin } Y_0) + t_2(f(x_2) - \text{tcl } Y_0) &= \text{tin}(t_1(f(x_1) - Y_0)) + \text{tcl}(t_2(f(x_2) - Y_0)) \\ &\subseteq \text{tin}(t_1(f(x_1) - Y_0) + t_2(f(x_2) - Y_0)) \\ &\subseteq \text{tin}(f(t_1 x_1 + t_2 x_2) - Y_0) \\ &= f(t_1 x_1 + t_2 x_2) - \text{tin } Y_0 \end{aligned}$$

whenever $x_1 \in X_0$ and $x_2 \in X_0$ (see Lemma 3). Hence $t_1(f(X_0) - \text{tin } Y_0) + t_2(f(X_0) - \text{tcl } Y_0) \subseteq f(X_0) - \text{tin } Y_0$. According to Lemma 2, we have $\text{tin}(f(X_0) - \text{tin } Y_0) = \text{tin}(f(X_0) - \text{tcl } Y_0)$. But $\text{tin}(f(X_0) - \text{tin } Y_0) \subseteq f(X_0) - \text{tin } Y_0 \subseteq \text{tin}(f(X_0) - Y_0) \subseteq \text{tin}(f(X_0) - \text{tcl } Y_0)$, which proves the first part of the lemma.

Let, in addition, $\emptyset \neq \text{tin } X_0$. Clearly $\emptyset \neq f(\text{tin } X_0) - \text{tin } Y_0 \subseteq f(X_0) - \text{tcl } Y_0$. Further if $t_1 > 0, t_2 > 0$ and $t_1 + t_2 = 1$, then, arguing as above, we get $t_1 x_1 + t_2 x_2 \in \text{tin } X_0$ and $t_1(f(x_1) - \text{tin } Y_0) + t_2(f(x_2) - \text{tcl } Y_0) \subseteq f(t_1 x_1 + t_2 x_2) - \text{tin } Y_0$ whenever $x_1 \in \text{tin } X_0$ and $x_2 \in X_0$, hence $t_1(f(\text{tin } X_0) - \text{tin } Y_0) + t_2(f(X_0) - \text{tcl } Y_0) \subseteq f(\text{tin } X_0) - \text{tin } Y_0$. According to Lemma 2, we have $\text{tin}(f(\text{tin } X_0) - \text{tin } Y_0) = \text{tin}(f(X_0) - \text{tcl } Y_0)$. But $\text{tin}(f(\text{tin } X_0) - \text{tin } Y_0) \subseteq f(\text{tin } X_0) - \text{tin } Y_0 \subseteq \text{tin}(f(X_0) - Y_0) \subseteq \text{tin}(f(X_0) - \text{tcl } Y_0)$, which proves the second part of the lemma. \square

Let us state now a slight generalization of a well-known result (see Fan, Glicksberg and Hoffman [7, p. 622]) which can be easily derived using Lemma 4. If f is Y_0 -concave and $\emptyset \neq \text{tin } Y_0$, then either the equation $f(x) \in \text{tin } Y_0$ has a solution, or there is a linear continuous nonzero function α from Y into R such that $\sup_{x \in X_0} \alpha(f(x)) \leq \inf_{y \in Y_0} \alpha(y)$; the two alternatives exclude each other. This follows from Lemma 4 which implies that one and only one of the following alternatives holds: either $\emptyset \neq f^{-1}(\text{tin } Y_0)$, or $0 \notin \text{tin } (f(X_0) - Y_0)$. Since the set $f(X_0) - Y_0$ is convex and has a nonempty topological interior, the conclusion follows applying a separation theorem.

Another variant of the above result is the following. If f is Y_0 -concave, $\emptyset \neq \text{tin } X_0$ and $\emptyset \neq \text{tin } Y_0$, then either the system $x \in \text{tin } X_0, f(x) \in \text{tin } Y_0$ has a solution, or there is α verifying the above properties, and the two alternatives exclude each other.

LEMMA 5. *Let f be Y_0 -concave and let $0 \in \text{tin } (f(X_0) - Y_0)$. If $\emptyset \neq \text{tin } Y_0$, then $\text{tin } f^{-1}(\text{tin } Y_0) = \text{tin } f^{-1}(\text{tcl } Y_0)$ and $\text{tcl } f^{-1}(\text{tin } Y_0) = \text{tcl } f^{-1}(\text{tcl } Y_0)$.*

Proof. Let $\emptyset \neq \text{tin } Y_0$. It follows from Lemma 4 that $0 \in f(X_0) - \text{tin } Y_0$, hence $\emptyset \neq f^{-1}(\text{tin } Y_0) \subseteq f^{-1}(\text{tcl } Y_0)$. If $x_1 \in f^{-1}(\text{tin } Y_0)$, $x_2 \in f^{-1}(\text{tcl } Y_0)$, $t_1 > 0$, $t_2 > 0$ and $t_1 + t_2 = 1$, then $0 \in t_1(f(x_1) - \text{tin } Y_0) + t_2(f(x_2) - \text{tcl } Y_0) \subseteq f(t_1x_1 + t_2x_2) - \text{tin } Y_0$ (see the proof of Lemma 4), i.e., $t_1x_1 + t_2x_2 \in f^{-1}(\text{tin } Y_0)$. Consequently $t_1f^{-1}(\text{tin } Y_0) + t_2f^{-1}(\text{tcl } Y_0) \subseteq f^{-1}(\text{tin } Y_0)$ whenever $t_1 > 0$, $t_2 > 0$ and $t_1 + t_2 = 1$, and the conclusion follows from Lemma 2.

If the sets Y_0 and $\text{gr } (f)$, where $\text{gr } (f)$ denotes the graph of f , are cones, then the set $f(X_0) - Y_0$ is a cone too, and the relation $0 \in \text{tin } (f(X_0) - Y_0)$ is equivalent to the relationship $f(X_0) - Y_0 = Y$. This will be the case later when we shall use the function $d_f(x_0)$ or $D_f(x_0)$ instead of f , and the set $k_{Y_0}(f(x_0))$ or $K_{Y_0}(f(x_0))$ instead of Y_0 (see also Lemmas 6 and 9).

4. Some properties of tangent sets and differentiable functions. Let X be a linear topological space. We shall denote by \mathcal{U} the family of all neighborhoods of the origin in X . Let X_0 be a subset of X and let x_0 be a point of X . We have given in the introduction of the present paper the definitions of the tangent sets $L_{X_0}(x_0)$, $l_{X_0}(x_0)$, $k_{X_0}(x_0)$ and $K_{X_0}(x_0)$. Some elementary properties of these tangent sets are discussed below.

Obviously $\text{comp } L_{X_0}(x_0) = K_{\text{comp } X_0}(x_0)$, $\text{comp } l_{X_0}(x_0) = k_{\text{comp } X_0}(x_0)$, $\text{comp } k_{X_0}(x_0) = l_{\text{comp } X_0}(x_0)$ and $\text{comp } K_{X_0}(x_0) = L_{\text{comp } X_0}(x_0)$. We have also $L_{X_0}(x_0) \subseteq l_{X_0}(x_0) \subseteq K_{X_0}(x_0)$ and $L_{X_0}(x_0) \subseteq k_{X_0}(x_0) \subseteq K_{X_0}(x_0)$. If, in addition, X_0 is convex, then $L_{X_0}(x_0) = l_{X_0}(x_0)$ and $k_{X_0}(x_0) = K_{X_0}(x_0)$.

The set $K_{X_0}(x_0)$ is closely related to the concept of a "tangent ray" to X_0 at x_0 (see Bouligand [4, p. 215], Severi [19, p. 99] and the references therein). It is not difficult to prove that a ray from x_0 is tangent to X_0 at x_0 if and only if it is a subset of the set $x_0 + K_{X_0}(x_0)$. The "contingent" to X_0 at x_0 (see Bouligand [5, p. 42]) is the family of all tangent rays to X_0 at x_0 . During the 1930s a great deal of research was carried out over this concept. Most of the reviews on the vast literature existing for the contingent can be easily found under the subject "direkte Infinitesimalgeometrie" in *Zentralblatt für Mathematik* (volumes 5 to 41). Nowadays the contingent has become a dictionary term (see Naas and Schmid [17, p. 282]), but people are hardly mindful of it.

The definitions of the sets $L_{X_0}(x_0)$ and $K_{X_0}(x_0)$ are from Dubovickii and Miljutin [6, p. 399]. Next we shall review some equivalent definitions of the sets $k_{X_0}(x_0)$ and $K_{X_0}(x_0)$ in order to point out their relationship to other known sets.

Since

$$K_{X_0}(x_0) = \{x \in X; \forall U \in \mathcal{U}, \forall r > 0, \emptyset \neq (x_0 + (0, r)(x + U)) \cap X_0\},$$

we get

$$K_{X_0}(x_0) = \{x \in X; \forall U \in \mathcal{U}, \emptyset \neq (x_0 + (0, +\infty)(x + U)) \cap (x_0 + U) \cap X_0\}.$$

This equality holds since, given $x \in X$, for every $\bar{U} \in \mathcal{U}$ there are $\tilde{U} \in \mathcal{U}$ and $\tilde{r} > 0$ such that $(0, \tilde{r})(x + \tilde{U}) \subseteq ((0, +\infty)(x + \bar{U})) \cap \bar{U}$; conversely, for every $\tilde{U} \in \mathcal{U}$ and $\tilde{r} > 0$ there is $\bar{U} \in \mathcal{U}$ such that $((0, +\infty)(x + \bar{U})) \cap \bar{U} \subseteq (0, \tilde{r})(x + \tilde{U})$ (we have to consider $\bar{U} \in \mathcal{U}$ such that $\bar{U} + (1/\tilde{r})\bar{U} \subseteq \tilde{U}$ and then we have to apply appropriately the inclusion $aA \cap bB \subseteq (ab/(a+b))(A+B)$, where $a > 0$, $b > 0$, $A \subseteq X$ and $B \subseteq X$).

Consequently, if X is metrizable, then the following three statements are equivalent: $x \in K_{X_0}(x_0)$; there are sequences $(x_n)_{n \in \mathbb{N}} \subseteq X_0$ and $(s_n)_{n \in \mathbb{N}} \subseteq (0, +\infty)$ such that $\lim_{n \rightarrow \infty} s_n = 0$ and $\lim_{n \rightarrow \infty} (1/s_n)(x_n - x_0) = x$ (see Hestenes [10, p. 170]); there are sequences $(s_n)_{n \in \mathbb{N}} \subseteq (0, +\infty)$ and $(x_n)_{n \in \mathbb{N}} \subseteq X_0$ such that $\lim_{n \rightarrow \infty} x_n = x_0$ and $\lim_{n \rightarrow \infty} s_n(x_n - x_0) = x$ (see Abadie [1, p. 33] and Whitney [25, p. 211]).

The set $k_{X_0}(x_0)$ can be characterized sometimes using functions $v: (0, \rho) \rightarrow X$, where ρ is possibly dependent on v . Namely, denoting

$$Y = \{v: (0, \rho) \rightarrow X; \lim_{s \rightarrow 0} v(s) = 0\}$$

we obtain the inclusion

$$\{x \in X; \exists v \in Y, \forall s \in (0, \rho), x_0 + s(x + v(s)) \in X_0\} \subseteq k_{X_0}(x_0),$$

but the corresponding equality does not hold unless X is a particular space, for example, a metrizable one. In this case the following two statements are equivalent: $x \in k_{X_0}(x_0)$; there is a function $\chi: (0, \rho) \rightarrow X_0$ such that

$$\lim_{s \rightarrow 0} (1/s)(\chi(s) - x_0) = x \quad (\text{see Girsanov [9, p. 39]}).$$

Finally if X is normable and $\delta(x_0)$ is the distance from the point x_0 to the set X_0 , i.e., $\delta(x_0) = \inf \{\|x - x_0\|; x \in X_0\}$, then we have

$$k_{X_0}(x_0) = \{x \in X; \lim_{s \rightarrow 0} \delta(x_0 + sx)/s = 0\},$$

$$K_{X_0}(x_0) = \{x \in X; \liminf_{s \rightarrow 0} \delta(x_0 + sx)/s = 0\}$$

(see Federer [8, p. 435]).

LEMMA 6. *The sets $L_{X_0}(x_0)$ and $l_{X_0}(x_0)$ are open cones. The sets $k_{X_0}(x_0)$ and $K_{X_0}(x_0)$ are closed cones.*

Proof. The properties of the sets $L_{X_0}(x_0)$ and $K_{X_0}(x_0)$ are well known so we omit their proof. We shall discuss only the sets $l_{X_0}(x_0)$ and k_{X_0} which are rather new.

Let $x \in l_{X_0}(x_0)$. According to the definition of the set $l_{X_0}(x_0)$, there are $U \in \mathcal{U}$ and a function $s: (0, +\infty) \rightarrow (0, +\infty)$ such that $s(r) \in (0, r)$ and $x_0 + s(r)(x + U) \subseteq X_0$ for every $r > 0$. If $t > 0$, then $tU \in \mathcal{U}$, $s(rt)/t \in (0, r)$ and $x_0 + (s(rt)/t)(tx + tU) = x_0 + s(rt)(x + U) \subseteq X_0$ for every $r > 0$, hence $tx \in l_{X_0}(x_0)$ and $l_{X_0}(x_0)$ is a cone. Further, denoting $U_u = U - u$ for $u \in X$, we have $U_u \in \mathcal{U}$ and $x_0 + s(r)(x + u + U_u) = x_0 + s(r)(x + U) \subseteq X_0$ for every $r > 0$ and $u \in \text{tin } U$, hence $x + u \in l_{X_0}(x_0)$ for every $u \in \text{tin } U$ and $l_{X_0}(x_0)$ is open.

Because of the equality $k_{X_0}(x_0) = \text{comp } l_{\text{comp } X_0}(x_0)$, it follows that $k_{X_0}(x_0)$ is a closed cone. \square

We get from Lemma 6 that $L_{X_0}(x_0) \subseteq \text{tin } k_{X_0}(x_0)$, $l_{X_0}(x_0) \subseteq \text{tin } K_{X_0}(x_0)$, $\text{tcl } L_{X_0}(x_0) \subseteq k_{X_0}(x_0)$ and $\text{tcl } l_{X_0}(x_0) \subseteq K_{X_0}(x_0)$. We shall use later assumptions like $L_{X_0}(x_0) = \text{tin } k_{X_0}(x_0)$ and $l_{X_0}(x_0) = \text{tin } K_{X_0}(x_0)$. Note that if X_0 is convex and $\emptyset \neq \text{tin } X_0$, then $\emptyset \neq L_{X_0}(x_0) = \text{tin } k_{X_0}(x_0)$ and $\text{tcl } L_{X_0}(x_0) = k_{X_0}(x_0)$ (see Lobry [15, p. 31]).

LEMMA 7. *The following three relations are equivalent: $x_0 \in \text{tin } X_0$; $L_{X_0}(x_0) = X$; $l_{X_0}(x_0) = X$. The following three relations are equivalent, too: $x_0 \in \text{tcl } X_0$; $\emptyset \neq k_{X_0}(x_0)$; $\emptyset \neq K_{X_0}(x_0)$.*

Proof. First let $x_0 \in \text{tin } X_0$. Then for every $x \in X$ there are $U \in \mathcal{U}$ and $r > 0$ such that $x_0 + (0, r)(x + U) \subseteq X_0$; hence $x \in L_{X_0}(x_0)$ and $L_{X_0}(x_0) = X$. Further let $L_{X_0}(x_0) = X$. Then obviously $l_{X_0}(x_0) = X$. Finally let $l_{X_0}(x_0) = X$. Since $0 \in l_{X_0}(x_0)$, it follows from the definition of the set $l_{X_0}(x_0)$ that there are $U \in \mathcal{U}$ and $s > 0$ such that $x_0 + sU = x_0 + s(0 + U) \subseteq X_0$. But $sU \in \mathcal{U}$, hence $x_0 \in \text{tin } X_0$ and the first part of the lemma is proved. The second part follows by negation from the first one with $\text{comp } X_0$ instead of X_0 . \square

We shall discuss now some formulas involving tangent sets in Cartesian product spaces. Let X^1 and X^2 be linear topological spaces, let X_0^1 and X_0^2 be subsets of X^1 and X^2 , respectively, and let x_0^1 and x_0^2 be points of X^1 and X^2 , respectively.

LEMMA 8. *Let $X = X^1 \times X^2$, $X_0 = X_0^1 \times X_0^2$ and $x_0 = (x_0^1, x_0^2)$. Then the following relations hold:*

$$L_{X_0}(x_0) = L_{X_0^1}(x_0^1) \times L_{X_0^2}(x_0^2),$$

$$l_{X_0}(x_0) \subseteq l_{X_0^1}(x_0^1) \times l_{X_0^2}(x_0^2),$$

$$k_{X_0}(x_0) = k_{X_0^1}(x_0^1) \times k_{X_0^2}(x_0^2),$$

$$K_{X_0}(x_0) \subseteq K_{X_0^1}(x_0^1) \times K_{X_0^2}(x_0^2).$$

Proof. Denote by \mathcal{U}^1 and \mathcal{U}^2 the families of all neighborhoods of the origins in X^1 and X^2 , respectively. As is well known, $U \in \mathcal{U}$ if and only if there are $U^1 \in \mathcal{U}^1$ and $U^2 \in \mathcal{U}^2$ such that $U^1 \times U^2 \subseteq U$.

First let $x = (x^1, x^2) \in L_{X_0}(x_0)$. According to the definition of the set $L_{X_0}(x_0)$, there are $U \in \mathcal{U}$ and $r > 0$ such that $x_0 + (0, r)(x + U) \subseteq X_0$. Consider $U^1 \in \mathcal{U}^1$ and $U^2 \in \mathcal{U}^2$ such that $U^1 \times U^2 \subseteq U$. Then $x_0^1 + (0, r)(x^1 + U^1) \subseteq X_0^1$ and $x_0^2 + (0, r)(x^2 + U^2) \subseteq X_0^2$, hence $x^1 \in L_{X_0^1}(x_0^1)$, $x^2 \in L_{X_0^2}(x_0^2)$, $x \in L_{X_0^1}(x_0^1) \times L_{X_0^2}(x_0^2)$ and $L_{X_0}(x_0) \subseteq L_{X_0^1}(x_0^1) \times L_{X_0^2}(x_0^2)$. Conversely, let $x = (x^1, x^2) \in L_{X_0^1}(x_0^1) \times L_{X_0^2}(x_0^2)$. According to the definitions of the sets $L_{X_0^1}(x_0^1)$ and $L_{X_0^2}(x_0^2)$, there are $U^1 \in \mathcal{U}^1$, $r^1 > 0$, $U^2 \in \mathcal{U}^2$ and $r^2 > 0$ such that $x_0^1 + (0, r^1)(x^1 + U^1) \subseteq X_0^1$ and $x_0^2 + (0, r^2)(x^2 + U^2) \subseteq X_0^2$. Denote $U = U^1 \times U^2$ and $r = \min(r^1, r^2)$. Then $x_0 + (0, r)(x + U) \subseteq X_0$, hence $x \in L_{X_0}(x_0)$, $L_{X_0^1}(x_0^1) \times L_{X_0^2}(x_0^2) \subseteq L_{X_0}(x_0)$ and the first relation is proved.

Now let $x = (x^1, x^2) \in l_{X_0}(x_0)$. According to the definition of the set $l_{X_0}(x_0)$, there is $U \in \mathcal{U}$ such that for every $r > 0$ there is $s \in (0, r)$ such that $x_0 + s(x + U) \subseteq X_0$. Consider $U^1 \in \mathcal{U}^1$ and $U^2 \in \mathcal{U}^2$ such that $U^1 \times U^2 \subseteq U$, let $r > 0$ and consider $s \in (0, r)$ as above. Then $x_0^1 + s(x^1 + U^1) \subseteq X_0^1$ and $x_0^2 + s(x^2 + U^2) \subseteq X_0^2$; hence $x^1 \in l_{X_0^1}(x_0^1)$, $x^2 \in l_{X_0^2}(x_0^2)$, $x \in l_{X_0^1}(x_0^1) \times l_{X_0^2}(x_0^2)$ and the second relation is proved.

Let now $x = (x^1, x^2) \in k_{X_0}(x_0)$, let $U^1 \in \mathcal{U}^1$ and let $U^2 \in \mathcal{U}^2$. Denote $U = U^1 \times U^2$. According to the definition of the set $k_{X_0}(x_0)$, there is $r > 0$ such that $\emptyset \neq (x_0 + s(x + U)) \cap X_0$ for every $s \in (0, r)$. Then $\emptyset \neq (x_0^1 + s(x^1 + U^1)) \cap X_0^1$ and $\emptyset \neq (x_0^2 + s(x^2 + U^2)) \cap X_0^2$ for every $s \in (0, r)$, hence $x^1 \in k_{X_0^1}(x_0^1)$, $x^2 \in k_{X_0^2}(x_0^2)$, $x \in k_{X_0^1}(x_0^1) \times k_{X_0^2}(x_0^2)$ and $k_{X_0}(x_0) \subseteq k_{X_0^1}(x_0^1) \times k_{X_0^2}(x_0^2)$. Conversely, let $x = (x^1, x^2) \in k_{X_0^1}(x_0^1) \times k_{X_0^2}(x_0^2)$ and let $U \in \mathcal{U}$. Consider $U^1 \in \mathcal{U}^1$ and $U^2 \in \mathcal{U}^2$ such that $U^1 \times U^2 \subseteq U$. According to the definitions of the sets $k_{X_0^1}(x_0^1)$ and $k_{X_0^2}(x_0^2)$, there are $r^1 > 0$ and $r^2 > 0$ such that $\emptyset \neq (x_0^1 + s(x^1 + U^1)) \cap X_0^1$ for every $s \in (0, r^1)$, and $\emptyset \neq$

$(x_0^2 + s(x^2 + U^2)) \cap X_0^2$ for every $s \in (0, r^2)$. Denote $r = \min(r^1, r^2)$. Then $\emptyset \neq (x_0 + s(x + U)) \cap X_0$ for every $s \in (0, r)$, hence $x \in k_{X_0}(x_0)$, $k_{X_0^1}(x_0^1) \times k_{X_0^2}(x_0^2) \subseteq k_{X_0}(x_0)$ and the third relation is proved.

Finally let $x = (x^1, x^2) \in K_{X_0}(x_0)$, let $U^1 \in \mathcal{U}^1$, let $U^2 \in \mathcal{U}^2$ and let $r > 0$. Denote $U = U^1 \times U^2$. According to the definition of the set $K_{X_0}(x_0)$, $\emptyset \neq (x_0 + (0, r)(x + U)) \cap X_0$. Then $\emptyset \neq (x_0^1 + (0, r)(x^1 + U^1)) \cap X_0^1$ and $\emptyset \neq (x_0^2 + (0, r)(x^2 + U^2)) \cap X_0^2$; hence $x^1 \in K_{X_0^1}(x_0^1)$, $x^2 \in K_{X_0^2}(x_0^2)$, $x \in K_{X_0^1}(x_0^1) \times K_{X_0^2}(x_0^2)$, and the last relation is proved. \square

Suppose now X_0 is a nonempty subset of X and x_0 is a point of X_0 . Let Y be a linear topological space. We shall denote by \mathcal{V} the family of all neighborhoods of the origin in Y . Let f be a function from X_0 into Y . We have defined in [22, p. 200] the differentiability and the differential of f at x_0 as follows.

The function f is said to be differentiable at the point x_0 if for every $x \in K_{X_0}(x_0)$ there is $y \in Y$ with the following property: for every $V \in \mathcal{V}$ there are $U \in \mathcal{U}$ and $r > 0$ such that $f(x_0 + s(x + u)) \in f(x_0) + s(y + V)$ whenever $s \in (0, r)$, $u \in U$ and $x_0 + s(x + u) \in X_0$. It can be shown that if f is differentiable at x_0 , then for every $x \in K_{X_0}(x_0)$ there is a unique $y \in Y$ with the above property. The function just obtained, from $K_{X_0}(x_0)$ into Y , is said to be the differential of f at x_0 and is denoted by $D_f(x_0)$.

The function $D_f(x_0)$ is closely related to the concept of "directional derivative" of f at x_0 (see Severi [20, p. 10]). It is not difficult to prove that the directional derivative of f at x_0 along a tangent ray to X_0 at x_0 is just $D_f(x_0)(x)$ where x is the direction of the ray. It seems that Severi was the first to extend the idea of differentiation to functions whose domains are not necessarily open.

It can be shown (see [22, p. 201]) that $D_f(x_0)$, if it exists, is continuous on $K_{X_0}(x_0)$. As a consequence, supposing $X_0 \subseteq x_0 + K_{X_0}(x_0)$, we deduce that f is differentiable at x_0 if and only if there is a continuous function $D_f(x_0)$ from $K_{X_0}(x_0)$ into Y with the following property: for every $x \in K_{X_0}(x_0)$ and $V \in \mathcal{V}$ there are $U \in \mathcal{U}$ and $r > 0$ such that $f(x_0 + s(x + u)) \in f(x_0) + s(D_f(x_0)(x + u) + V)$ whenever $s \in (0, r)$, $u \in U$ and $x_0 + s(x + u) \in X_0$. Thus, in the case where $x_0 \in \text{tin } X_0$, differentiability of f at x_0 together with additivity of $D_f(x_0)$ coincides with differentiability in the sense of Bastiani [3, p. 18], and this one is equivalent (see Averbuh and Smoljanov [2, p. 92]) with differentiability in the sense of Michal [16, p. 340].

We have proved in [22, p. 201] that $\text{gr}(D_f(x_0)) = K_{\text{gr}(f)}(x_0, f(x_0))$. We shall denote by $d_f(x_0)$ the restriction of the function $D_f(x_0)$ to the set $k_{X_0}(x_0)$ and we shall show that $d_f(x_0)$ verifies a similar equality.

LEMMA 9. *Let f be differentiable at x_0 . Then $\text{gr}(d_f(x_0)) = k_{\text{gr}(f)}(x_0, f(x_0))$.*

Proof. Let $(x, y) \in \text{gr}(d_f(x_0))$, i.e., $x \in k_{X_0}(x_0)$ and $y = D_f(x_0)(x)$, and let $W \in \mathcal{W}$, where \mathcal{W} denotes the family of all neighborhoods of the origin in $X \times Y$. Consider $U \in \mathcal{U}$ and $V \in \mathcal{V}$ such that $U \times V \subseteq W$. According to the definition of the function $D_f(x_0)$, there are $\tilde{U} \in \mathcal{U}$ and $\tilde{r} > 0$ such that $f(x_0 + s(x + u)) \in f(x_0) + s(y + V)$ whenever $s \in (0, \tilde{r})$, $u \in \tilde{U}$ and $x_0 + s(x + u) \in X_0$. Denote $\tilde{U} = U \cap \tilde{U}$. According to the definition of the set $k_{X_0}(x_0)$, there is $\tilde{r} > 0$ such that $\emptyset \neq (x_0 + s(x + \tilde{U})) \cap X_0$ for every $s \in (0, \tilde{r})$. Denote $r = \min(\tilde{r}, \tilde{r})$ and let $s \in (0, r)$. Since $s \in (0, \tilde{r})$, there is $u \in \tilde{U}$ such that $x_0 + s(x + u) \in X_0$. Denote $v = (1/s)(f(x_0 + s(x + u)) - f(x_0)) - y$. Since $s \in (0, \tilde{r})$ and $u \in \tilde{U}$, we have $v \in V$. Consequently $(u, v) \in W$ and $(x_0, f(x_0)) + s((x, y) + (u, v)) = (x_0 + s(x + u), f(x_0 + s(x + u))) \in \text{gr}(f)$, hence $(x, y) \in k_{\text{gr}(f)}(x_0, f(x_0))$ and $\text{gr}(d_f(x_0)) \subseteq k_{\text{gr}(f)}(x_0, f(x_0))$. Conversely let $(x, y) \in k_{\text{gr}(f)}(x_0, f(x_0))$. Since $\text{gr}(f) \subseteq X_0 \times Y$, we have $x \in k_{X_0}(x_0)$ (see Lemma 8), and since $k_{\text{gr}(f)}(x_0, f(x_0)) \subseteq K_{\text{gr}(f)}(x_0, f(x_0))$, we have $(x, y) \in \text{gr}(D_f(x_0))$; hence $(x, y) \in \text{gr}(d_f(x_0))$, $k_{\text{gr}(f)}(x_0, f(x_0)) \subseteq \text{gr}(d_f(x_0))$ and the lemma is proved. \square

5. Calculus of tangent sets. Let X and Y be linear topological spaces, let X_0 and Y_0 be nonempty subsets of X and Y , respectively, let f be a function from X_0 into Y and let x_0 be a point of X_0 . We shall suppose that $f(x_0) \in \text{tcl } Y_0$. Thus both $k_{Y_0}(f(x_0))$ and $K_{Y_0}(f(x_0))$ are nonempty sets (see Lemma 7). Now we are able to state and prove the main results concerning the relations (A)–(F).

THEOREM 1. *Let f be differentiable at x_0 , let $d_f(x_0)$ be $k_{Y_0}(f(x_0))$ -concave and let $d_f(x_0)(k_{X_0}(x_0)) - k_{Y_0}(f(x_0)) = Y$. If $\emptyset \neq L_{Y_0}(f(x_0)) = \text{tin } k_{Y_0}(f(x_0))$, then (A) and (C) hold. Moreover if $L_{X_0}(x_0) = \text{tin } k_{X_0}(x_0)$, then $L_{f^{-1}(Y_0)}(x_0) = \text{tin } k_{f^{-1}(Y_0)}(x_0)$. Moreover if $\emptyset \neq L_{X_0}(x_0)$, then $\emptyset \neq L_{f^{-1}(Y_0)}(x_0)$.*

Proof. Let $\emptyset \neq L_{Y_0}(f(x_0)) = \text{tin } k_{Y_0}(f(x_0))$. Using the definition of the function $d_f(x_0)$, we can write relations (a) and (c) as follows:

$$\begin{aligned} L_{X_0}(x_0) \cap (d_f(x_0))^{-1}(L_{Y_0}(f(x_0))) &\subseteq L_{f^{-1}(Y_0)}(x_0) \subseteq L_{X_0}(x_0) \cap (d_f(x_0))^{-1}(k_{Y_0}(f(x_0))), \\ (d_f(x_0))^{-1}(L_{Y_0}(f(x_0))) &\subseteq k_{f^{-1}(Y_0)}(x_0) \subseteq (d_f(x_0))^{-1}(k_{Y_0}(f(x_0))). \end{aligned}$$

Since the sets $L_{X_0}(x_0)$ and $L_{f^{-1}(Y_0)}(x_0)$ are open and the sets $k_{f^{-1}(Y_0)}(x_0)$ and $k_{Y_0}(f(x_0))$ are closed (see Lemma 6), we have

$$\begin{aligned} L_{f^{-1}(Y_0)}(x_0) &\subseteq L_{X_0}(x_0) \cap \text{tin } (d_f(x_0))^{-1}(k_{Y_0}(f(x_0))), \\ \text{tcl } (d_f(x_0))^{-1}(L_{Y_0}(f(x_0))) &\subseteq k_{f^{-1}(Y_0)}(x_0), \end{aligned}$$

and we deduce from Lemma 5 that

$$\begin{aligned} \text{tin } (d_f(x_0))^{-1}(k_{Y_0}(f(x_0))) &\subseteq (d_f(x_0))^{-1}(L_{Y_0}(f(x_0))), \\ (d_f(x_0))^{-1}(k_{Y_0}(f(x_0))) &\subseteq \text{tcl } (d_f(x_0))^{-1}(L_{Y_0}(f(x_0))). \end{aligned}$$

Consequently

$$\begin{aligned} L_{X_0}(x_0) \cap (d_f(x_0))^{-1}(L_{Y_0}(f(x_0))) &= L_{f^{-1}(Y_0)}(x_0), \\ k_{f^{-1}(Y_0)}(x_0) &= (d_f(x_0))^{-1}(k_{Y_0}(f(x_0))); \end{aligned}$$

and the relations (A) and (C) are proved.

Further let $L_{X_0}(x_0) = \text{tin } k_{X_0}(x_0)$. Since $L_{f^{-1}(Y_0)}(x_0) \subseteq \text{tin } k_{f^{-1}(Y_0)}(x_0) \subseteq \text{tin } k_{X_0}(x_0) \cap (d_f(x_0))^{-1}(L_{Y_0}(f(x_0)))$, we have $L_{f^{-1}(Y_0)}(x_0) = \text{tin } k_{f^{-1}(Y_0)}(x_0)$.

Further let $\emptyset \neq L_{X_0}(x_0)$. We deduce from Lemma 4 that $d_f(x_0)(L_{X_0}(x_0)) - L_{Y_0}(f(x_0)) = Y$. Consequently $\emptyset \neq L_{f^{-1}(Y_0)}(x_0)$, and the theorem is proved. \square

The following three theorems can be proved in a similar manner.

THEOREM 2. *Let f be differentiable at x_0 , let $D_f(x_0)$ be $k_{Y_0}(f(x_0))$ -concave and let $D_f(x_0)(K_{X_0}(x_0)) - k_{Y_0}(f(x_0)) = Y$. If $\emptyset \neq L_{Y_0}(f(x_0)) = \text{tin } k_{Y_0}(f(x_0))$, then (A) and (D) hold.*

THEOREM 3. *Let f be differentiable at x_0 , let $d_f(x_0)$ be $K_{Y_0}(f(x_0))$ -concave and let $d_f(x_0)(k_{X_0}(x_0)) - K_{Y_0}(f(x_0)) = Y$. If $\emptyset \neq l_{Y_0}(f(x_0)) = \text{tin } K_{Y_0}(f(x_0))$, then (E) and (F) hold. Moreover if $L_{X_0}(x_0) = \text{tin } k_{X_0}(x_0)$, then $L_{X_0}(x_0) \cap l_{f^{-1}(Y_0)}(x_0) = L_{X_0}(x_0) \cap \text{tin } K_{f^{-1}(Y_0)}(x_0)$. Moreover, if $\emptyset \neq L_{X_0}(x_0)$, then $\emptyset \neq L_{X_0}(x_0) \cap l_{f^{-1}(Y_0)}(x_0)$.*

THEOREM 4. *Let f be differentiable at x_0 , let $D_f(x_0)$ be $K_{Y_0}(f(x_0))$ -concave and let $D_f(x_0)(K_{X_0}(x_0)) - K_{Y_0}(f(x_0)) = Y$. If $\emptyset \neq l_{Y_0}(f(x_0)) = \text{tin } K_{Y_0}(f(x_0))$, then (B) and (E) hold.*

The hypotheses of the above theorems have a common structure which we will explain using Theorem 1. They contain *general regularity conditions* such as “ f is differentiable at x_0 , $d_f(x_0)$ is $k_{Y_0}(f(x_0))$ -concave and $d_f(x_0)(k_{X_0}(x_0)) - k_{Y_0}(f(x_0)) = Y$ ”

and *particular regularity conditions* such as " $\emptyset \neq L_{Y_0}(f(x_0)) = \text{tin } k_{Y_0}(f(x_0))$ ". The problem in the tangent sets' calculus is to maintain the same general regularity conditions and to discover other particular regularity conditions in order to obtain relation (C) or (D) or (F), at least. This the author intends to do in a future paper.

6. Extremal points and necessary conditions for extremality. Let X and Y be topological spaces, let X_0 and Y_0 be nonempty subsets of X and Y respectively, and let f be a function from X_0 into Y .

DEFINITION 2. We shall say that a point x_0 of X_0 is Y_0 -*extremal* for the function f if $f(x_0) \in \text{tcl } Y_0$ and $x_0 \notin \text{tcl } f^{-1}(Y_0)$.

This concept of extremality is similar to that given by Neustadt in [18, p. 59]. The second relation in Definition 2 says that there is a neighborhood U_0 of x_0 in X such that $U_0 \cap f^{-1}(Y_0) = \emptyset$, i.e., $f(x) \notin Y_0$ whenever $x \in X_0 \cap U_0$. Thus every local solution of every programming problem, with scalar or vector objective and without or with constraints, becomes an extremal point for suitably chosen f and Y_0 . We shall explain this fact in detail in the next section.

Suppose now X and Y are linear topological spaces. Taking Lemma 7 into account, every result stating relation (C) or (D) or (F) can be transformed into a result stating necessary conditions for extremality. Because of the symmetry we have chosen the theorem stating relation (C).

Let x_0 be a point of X_0 . The theorem below contains necessary conditions for extremality in form of a multiplier rule.

THEOREM 5. Let x_0 be Y_0 -extremal for f , let f be differentiable at x_0 and let $d_f(x_0)$ be $k_{Y_0}(f(x_0))$ -concave. If $\emptyset \neq L_{Y_0}(f(x_0)) = \text{tin } k_{Y_0}(f(x_0))$, then there is a linear continuous nonzero function α from Y into \mathbb{R} such that:

$$\alpha(d_f(x_0)(x)) \leq 0 \quad \forall x \in k_{X_0}(x_0),$$

$$\alpha(y) \geq 0 \quad \forall y \in k_{Y_0}(f(x_0)).$$

Proof. Denote $Z_0 = d_f(x_0)(k_{X_0}(x_0)) - k_{Y_0}(f(x_0))$. The set Z_0 is a convex (see Lemma 3) cone (see Lemmas 6 and 9) and has a nonempty topological interior. Thus the conclusion is equivalent to the fact that $Z_0 \neq Y$. Suppose by contradiction that $Z_0 = Y$. Then $\emptyset \neq k_{f^{-1}(Y_0)}(x_0)$ (see Theorem 1), hence $x_0 \in \text{tcl } f^{-1}(Y_0)$ (see Lemma 7) contradicting the Y_0 -extremality of x_0 (see Definition 2) and thereby completing the proof. \square

Note that if Y_0 is convex, then the relation $\alpha(y) \geq 0$ for all $y \in k_{Y_0}(f(x_0))$ is equivalent to the relation $\alpha(y) \geq \alpha(f(x_0))$ for all $y \in Y_0$, and this one is equivalent to the pair of relations $\alpha(f(x_0)) = 0$ and $\alpha(y) \geq 0$ for all $y \in Y_0$ provided Y_0 is a convex cone (see Neustadt [18, p. 64]).

7. Applications to programming problems. Let X be a topological space and let X_0 be a nonempty subset of X .

Let Y be a topological space, let f be a function from X_0 into Y , and let \leq be a nonempty binary relation in Y . The pair (f, \leq) will be termed an *objective*.

DEFINITION 3. We shall say that a point x_0 of X_0 is *locally minimal* for the objective (f, \leq) if there is a neighborhood U_0 of x_0 in X such that $f(x) = f(x_0)$ whenever $x \in X_0 \cap U_0$ and $f(x) \leq f(x_0)$. We shall say that a point x_0 of X_0 is *locally maximal* for the objective (f, \leq) if there is a neighborhood U_0 of x_0 in X such that $f(x_0) = f(x)$ whenever $x \in X_0 \cap U_0$ and $f(x_0) \leq f(x)$.

Let Z be a topological space, let g be a function from X_0 into Z , and let Z_0 be a nonempty subset of Z . The pair (g, Z_0) will be termed a *constraint*.

DEFINITION 4. We shall say that a point x_0 of X_0 is *locally minimal* for the objective (f, \leq) subject to the constraint (g, Z_0) if $g(x_0) \in Z_0$ and there is a neighborhood U_0 of x_0 in X such that $f(x) = f(x_0)$ whenever $x \in X_0 \cap U_0$, $g(x) \in Z_0$ and $f(x) \leq f(x_0)$. We shall say that a point x_0 of X_0 is *locally maximal* for the objective (f, \leq) subject to the constraint (g, Z_0) if $g(x_0) \in Z_0$ and there is a neighborhood U_0 of x_0 in X such that $f(x_0) = f(x)$ whenever $x \in X_0 \cap U_0$, $g(x) \in Z_0$ and $f(x_0) \leq f(x)$.

Observe that locally minimal points with respect to a binary relation are locally maximal points with respect to the inverse of that binary relation. In the sequel we shall consider only locally maximal points. Similar maximal notions were considered by Hurwicz in [11, p. 67].

Let x_0 be a point of X_0 , denote $Y_0 = \{y \in Y; f(x_0) < y\}$, where $y_1 < y_2$ means that $y_1 \leq y_2$ and $y_1 \neq y_2$, and suppose that $f(x_0) \in \text{tcl } Y_0$. Such an assumption is verified if, for example, the pair (Y, \leq) is an ordered linear topological space with strictly positive elements.

THEOREM 6. *The point x_0 is locally maximal for the objective (f, \leq) if and only if x_0 is Y_0 -extremal for f . The point x_0 is locally maximal for the objective (f, \leq) subject to the constraint (g, Z_0) if and only if $g(x_0) \in Z_0$ and x_0 is $Y_0 \times Z_0$ -extremal for $f \times g$.*

Proof. It follows from Definitions 3 and 4 that x_0 is locally maximal for the objective (f, \leq) if and only if $x_0 \notin \text{tcl } f^{-1}(Y_0)$; x_0 is locally maximal for the objective (f, \leq) subject to the constraint (g, Z_0) if and only if $g(x_0) \in Z_0$ and $x_0 \notin \text{tcl } (f^{-1}(Y_0) \cap g^{-1}(Z_0))$. Since $f(x_0) \in \text{tcl } Y_0$ and $Z_0 \subseteq \text{tcl } Z_0$, the conclusion follows from Definition 2. \square

Finally, suppose X , Y and Z are linear topological spaces. The result below represents a reformulation of Theorem 5 via Theorem 6.

THEOREM 7. *Let x_0 be locally maximal for the objective (f, \leq) , let f be differentiable at x_0 and let $d_f(x_0)$ be $k_{Y_0}(f(x_0))$ -concave. If $\emptyset \neq L_{Y_0}(f(x_0)) = \text{tin } k_{Y_0}(f(x_0))$, then there is a linear continuous nonzero function α from Y into \mathbb{R} such that:*

$$\begin{aligned}\alpha(d_f(x_0)(x)) &\leq 0 \quad \forall x \in k_{X_0}(x_0) \\ \alpha(y) &\geq 0 \quad \forall y \in k_{Y_0}(f(x_0)).\end{aligned}$$

The next result contains the necessary conditions of John [12, p. 188] and can be derived at once from Theorem 5 using Theorem 6 and Lemma 8.

THEOREM 8. *Let x_0 be locally maximal for the objective (f, \leq) subject to the constraint (g, Z_0) , let f and g be differentiable at x_0 , let $d_f(x_0)$ be $k_{Y_0}(f(x_0))$ -concave and let $d_g(x_0)$ be $k_{Z_0}(g(x_0))$ -concave. If $\emptyset \neq L_{Y_0}(f(x_0)) = \text{tin } k_{Y_0}(f(x_0))$ and $\emptyset \neq L_{Z_0}(g(x_0)) = \text{tin } k_{Z_0}(g(x_0))$, then there are two linear continuous not both zero functions α from Y into \mathbb{R} and β from Z into \mathbb{R} such that:*

$$\begin{aligned}\alpha(d_f(x_0)(x)) + \beta(d_g(x_0)(x)) &\leq 0 \quad \forall x \in k_{X_0}(x_0), \\ \alpha(y) &\geq 0 \quad \forall y \in k_{Y_0}(f(x_0)), \\ \beta(z) &\geq 0 \quad \forall z \in k_{Z_0}(g(x_0)).\end{aligned}$$

Note that α is not zero in Theorem 8 if $d_g(x_0)(k_{X_0}(x_0)) - k_{Z_0}(g(x_0)) = Z$. Note also that this equality is equivalent (see Lemma 4) to the relation $\emptyset \neq (d_g(x_0))^{-1}(L_{Z_0}(g(x_0)))$, which seems to be a Slater condition (see Slater [21]), and implies (see Theorem 1) the equality $k_{g^{-1}(Z_0)}(x_0) = (d_g(x_0))^{-1}(k_{Z_0}(g(x_0)))$, which seems to be a Kuhn–Tucker constraint qualification (see Kuhn and Tucker [14, p. 483]).

REFERENCES

- [1] J. ABADIE, *Problèmes d'optimisation*, Institut Blaise Pascal, Paris, 1965.
- [2] V. I. AVERBUH AND O. G. SMOLJANOV, *Različnye opredelenija proizvodnoi v lineinyh topologičeskikh prostranstvakh* (Different definitions of derivative in linear topological spaces) *Uspehi Mat. Nauk*, 23 (4) (1968), pp. 67–114. (In Russian).
- [3] A. BASTIANI, *Applications différentiables et variétés différentiables de dimension infinie*, *J. Analyse Math.*, 13 (1964), pp. 1–114.
- [4] G. BOULIGAND, *Sur l'existence des demi-tangentes à une courbe de Jordan*, *Fund. Math.*, 15 (1930), pp. 215–218.
- [5] ———, *Sur quelques points de méthodologie géométrique*, *Rev. Gén. Sci. Pures Appl.*, 41 (1930), pp. 39–43; 366–371; 599–603.
- [6] A. JA. DUBOVICKIĬ AND A. A. MILJUTIN, *Zadači na èkstreum pri naličii ograničeniei* (Extremal problems with constraints), *Z. Vyčisl. Mat. i Mat. Fiz.*, 5 (1965), pp. 395–453. (In Russian).
- [7] K. FAN, I. GLICKSBERG AND A. J. HOFFMAN, *Systems of inequalities involving convex functions*, *Proc. Amer. Math. Soc.*, 8 (1957), pp. 617–622.
- [8] H. FEDERER, *Curvature measures*, *Trans. Amer. Math. Soc.*, 93 (1959), pp. 418–491.
- [9] I. V. GIRSANOV, *Lekcii po matematičeskoj teorii èkstremaľnyh zadač*, Moscow University, Moscow, 1970 (in Russian) = *Lectures on mathematical theory of extremum problems*, in *Lecture Notes in Economics and Mathematical Systems* 67, Springer, New York, 1972.
- [10] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, London, Sydney, 1966.
- [11] L. HURWICZ, *Programming in linear spaces*, in *Studies in Linear and Nonlinear Programming*, K. J. Arrow, L. Hurwicz and H. Uzawa, eds., Stanford University Press, Stanford, CA, 1958, pp. 38–102.
- [12] F. JOHN, *Extremum problems with inequalities as subsidiary conditions*, in *Studies and Essays Presented to R. Courant on his 60th Birthday, January 8, 1948*, K. O. Friedrichs, O. E. Neugebauer and J. J. Stoker, eds., Interscience, New York, 1948, pp. 187–204.
- [13] V. KLEE, *Convex sets in linear spaces*, *Duke Math. J.*, 18 (1951), pp. 443–466.
- [14] H. W. KUHN AND A. W. TUCKER, *Nonlinear programming*, in *Proc. 2nd Berkeley Symp. on Math. Statistics and Probability*, J. Neyman, ed., Univ. California Press, Berkeley and Los Angeles, 1951, pp. 481–492.
- [15] C. LOBRY, *Etude géométrique des problèmes d'optimisation en présence de contraintes*, Thèse, Grenoble, June 1967.
- [16] A. D. MICHAL, *Differential calculus in linear topological spaces*, *Proc. Nat. Acad. Sci. U.S.A.*, 24 (1938), pp. 340–342.
- [17] J. NAAS AND H. L. SCHMID, *Mathematisches Wörterbuch*, Band I, Akademie Verlag GMBH, Berlin, B. G. Teubner Verlagsgesellschaft, Leipzig, 1961.
- [18] L. W. NEUSTADT, *A general theory of extremals*, *J. Comput. System. Sci.*, 3 (1969), pp. 57–92.
- [19] F. SEVERI, *Su alcune questioni di topologia infinitesimale*, *Ann. Soc. Polon. Math.*, 9 (1930), pp. 97–108.
- [20] ———, *Sulla differenziabilità totale delle funzioni di piu variabili reali*, *Ann. Mat. Pura Appl.*, 13 (1935), pp. 1–35.
- [21] M. SLATER, *Lagrange multipliers revisited: A contribution to nonlinear programming*, Cowles Commission Discussion Paper, Mathematics 403, November 1950.
- [22] C. URSESCU, *Sur une généralisation de la notion de différentiabilité*, *Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur.*, 54 (1973), pp. 199–204.
- [23] ———, *Tangent sets and differentiable functions*, in *Mathematical Control Theory*, S. Dolecki, C. Olech and J. Zabczyk, eds., Polish Scientific Publ., Warsaw, 1976, pp. 151–155.
- [24] P. P. VARAIYA, *Nonlinear programming in Banach space*, *SIAM J. Appl. Math.*, 15 (1967), pp. 284–293.
- [25] H. WHITNEY, *Local properties of analytic varieties*, in *Differential and Combinatorial Topology (A Symposium in Honor of Marston Morse)*, S. S. Cairns, ed., Princeton Univ. Press, Princeton, NJ, 1965, pp. 205–244.

CONTROLLABILITY FOR DISTRIBUTED BILINEAR SYSTEMS*

J. M. BALL,[†] J. E. MARSDEN[‡] AND M. SLEMROD[§]

Abstract. This paper studies controllability of systems of the form $dw/dt = \mathcal{A}w + p(t)\mathcal{B}w$ where \mathcal{A} is the infinitesimal generator of a C^0 semigroup of bounded linear operators $e^{\mathcal{A}t}$ on a Banach space X , $\mathcal{B} : X \rightarrow X$ is a C^1 map, and $p \in L^1([0, T]; \mathbb{R})$ is a control. The paper (i) gives conditions for elements of X to be accessible from a given initial state w_0 and (ii) shows that controllability to a full neighborhood in X of w_0 is impossible for $\dim X = \infty$. Examples of hyperbolic partial differential equations are provided.

1. Introduction. The purpose of this paper is to discuss controllability for abstract evolution equations of the form

$$(1.1) \quad \dot{w}(t) = \mathcal{A}w(t) + p(t)\mathcal{B}(w(t)),$$

$$(1.2) \quad w(0) = w_0,$$

where \mathcal{A} generates a C^0 semigroup of bounded linear operators on a (possibly complex) Banach space X , $\mathcal{B} : X \rightarrow X$ is a C^1 map, and $p \in L^1([0, T]; \mathbb{R})$ is a control defined on a specified interval $[0, T]$. Usually we assume that \mathcal{B} is linear, so that (1.1) is *bilinear* in the pair (p, w) ; note that even in this case the solution w of (1.1), (1.2) is a *nonlinear function of p* . A motivating example is the rod equation

$$(1.3) \quad u_{tt} + u_{xxxx} + p(t)u_{xx} = 0, \quad 0 < x < 1,$$

with hinged end conditions

$$(1.4) \quad u = u_{xx} = 0 \quad \text{at } x = 0, 1,$$

which can be put in the form (1.1) by setting $w = \begin{pmatrix} u \\ u_t \end{pmatrix}$ with $X = (H^2(0, 1) \cap H_0^1(0, 1)) \times L^2(0, 1)$. Here the control $p(t)$ is the axial load.

The main tool used in our analysis is the generalized inverse function, or “local onto” theorem. In finite dimensions, the well-known controllability results for bilinear systems have been obtained in this way (see, for example, Brockett [1972] and Hermes [1974]). In infinite dimensions, however, new phenomena arise. Perhaps the most interesting of these is our result (Theorem 3.6) which shows that for \mathcal{B} linear and $\dim X = \infty$, the set of states accessible from w_0 for $p \in L_{\text{loc}}^1([0, \infty); \mathbb{R})$, $1 < r \leq \infty$, has *dense complement* in X . Hence we can *never* expect to control to an open neighborhood of w_0 for controls in L_{loc}^1 . (Using L^1 controls doesn’t help, at least for examples such as (1.3), (1.4); see Theorem 5.5.) This stands in direct contrast to the available positive results on controllability when $\dim X < \infty$.

* Received by the editors March 24, 1981, and in revised form September 25, 1981.

[†] Department of Mathematics, Heriot-Watt University, Edinburgh, United Kingdom EH14 4AS. The research of this author was supported in part by the U.S. Army Research Office under contract DAAG29-79-C-0086, the National Science Foundation under grant MCS-78-06718 and a United Kingdom Science Research Council Fellowship.

[‡] Department of Mathematics, University of California, Berkeley, California 94720. The research of this author was supported in part by the U.S. Army Research Office under contract DAAG29-79-C-0086 and the National Science Foundation under grant MCS-78-06718.

[§] Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, New York 12181. The research of this author was supported in part by the National Science Foundation under grant MCS-79-02773 and by the Air Force Office of Scientific Research, Air Force Systems Command, United States Air Force under contract/grant AFOSR-81-0172.

Given the impossibility of controlling the system (1.1) to a full neighborhood of w_0 with p 's in L' , we investigate two alternative procedures. One approach generalizes an idea of Hermes [1979]; we show that it is often possible to *control with respect to finite-dimensional observations* in a neighborhood of w_0 . Our second idea is based upon the concept of *approximate controllability*, i.e., we identify a dense subset of X , depending on w_0 and t , to which $w(t)$ belongs, and show that with respect to a strengthened topology one can control to a neighborhood of $e^{\mathcal{A}t}w_0$ (the "free solution" of (1.1), (1.2) corresponding to $p \equiv 0$) in this set, provided t is suitably chosen. For (1.3), (1.4) we prove that $t > 0$ can be taken arbitrarily small, whereas for the wave equation

$$(1.5) \quad u_{tt} - u_{xx} + p(t)u = 0, \quad 0 < x < 1,$$

with either the boundary conditions

$$u = 0 \quad \text{at } x = 0, 1,$$

or the boundary conditions

$$u = 0 \quad \text{at } x = 0, \quad u + \alpha u_x = 0 \quad \text{at } x = 1, \quad \alpha > 0,$$

t has to exceed some number $T > 0$. This study of local approximate controllability involves technicalities concerning nonharmonic Fourier series in the spirit of Russell [1967] and Ball and Slemrod [1979]. The delicacy of these questions has the unfortunate consequence that we have only been able to obtain positive results in cases, such as these described above, in which (1.1) is an abstract hyperbolic equation that is "diagonal"; i.e., is reducible to an infinite set of uncoupled ordinary differential equations (each, of course, containing the control $p(t)$). Since we have to control infinitely many ordinary differential equations simultaneously, however, the problem is still not trivial. Nevertheless, our assumptions exclude some important nondiagonal examples such as (1.3) with clamped end conditions

$$u = u_x = 0 \quad \text{at } x = 0, 1.$$

In special cases, such as (1.3), (1.4), our local approximate controllability theory leads to a global approximate controllability result; thus, for example, for suitable initial data, we prove that the attainable set for (1.3), (1.4) is dense in X .

The paper is divided into six sections. Section 2 assembles the machinery for studying (1.1), (1.2) in the form of various abstract existence and smoothness theorems. Section 3 provides an abstract controllability theorem and the result on noncontrollability mentioned above. In § 4 we discuss the general theory of control with respect to finite-dimensional observers. In § 5 we consider abstract hyperbolic equations, apply the theory of § 4 to this case, and develop our theory of approximate controllability. We conclude in § 6 with specific applications to partial differential equations, such as (1.3), (1.4).

2. Abstract existence and smoothness theorems. In this section we give some basic results on nonlinear evolution equations which will be useful in our later analysis. Let X be a Banach space with norm $\|\cdot\|$, let \mathcal{A} generate a C^0 semigroup of bounded linear operators on X , and let $\mathcal{B}: X \rightarrow X$ be a C^k mapping, $k \geq 1$. Let $Z(T)$ be a Banach space continuously and densely included in $L^1([0, T]; \mathbb{R})$, where $T > 0$ is given.

For a given $w_0 \in X$ and $p \in Z(T)$, consider the initial value problem associated with (1.1) written in integrated form, i.e.,

$$(2.1) \quad w(t) = e^{\mathcal{A}t}w_0 + \int_0^t e^{\mathcal{A}(t-s)}p(s)\mathcal{B}(w(s)) \, ds.$$

Solutions of (2.1) are often called “mild solutions” of (1.1), (1.2). The question as to when solutions of (2.1) are actually solutions of (1.1) is discussed in Remark 2.7 at the end of this section.

PROPOSITION 2.1. *For each $w_0 \in X$, and $p \in Z(T)$ there exists t_0 , $0 < t_0 \leq T$, such that (2.1) has a unique solution $w \in C([0, t_0]; X)$.*

Proof. Let $\mathcal{F} = \{w \in C([0, t_0]; X) \mid \|w(t) - w_0\| \leq R\}$, and define $T_p : \mathcal{F} \rightarrow C([0, t_0]; X)$ by

$$(T_p w)(t) = e^{\mathcal{A}t} w_0 + \int_0^t e^{\mathcal{A}(t-s)} p(s) \mathcal{B}(w(s)) ds.$$

Since $\|e^{\mathcal{A}t}\| \leq M e^{\beta t}$ for positive constants β, M , an easy estimate shows that T maps \mathcal{F} to \mathcal{F} provided

$$\|e^{\mathcal{A}t} w_0 - w_0\| + M e^{\beta t_0} C \int_0^{t_0} |p(s)| ds \leq R, \quad 0 \leq t \leq t_0,$$

where C is such that $\|Bw\| \leq C$ for $\|w - w_0\| \leq R$. This condition is achieved for R, t_0 sufficiently small via the continuity of \mathcal{B} , $e^{\mathcal{A}t} w_0$ and the fact that $p \in L^1([0, T]; \mathbb{R})$. Similarly, T_p is a contraction map of \mathcal{F} to \mathcal{F} provided that

$$KM e^{\beta t_0} \int_0^{t_0} |p(s)| ds < 1,$$

where K is a Lipschitz constant for \mathcal{B} on the ball $\|w - w_0\| \leq R$. Again this holds for R and t_0 sufficiently small. The result now follows from the contraction mapping principle. \square

Of course the above proposition is a special case of many more general results on existence and uniqueness of solutions to semilinear evolution equations (see, for example, Segal [1963], Pazy [1974], Balakrishnan [1976] and Tanabe [1979b]). The point for us here is that use of the contraction mapping principle leads to other important features of the solution map w , as we now see.

PROPOSITION 2.2. *Fix $p_0 \in Z(T)$. Then there exist an open neighborhood U of p_0 in $Z(T)$ and $t_0 > 0$ such that for any $p \in U$, (2.1) has a unique solution $w(t; p, w_0)$, $0 \leq t \leq t_0$. Moreover $w(t; p, w_0)$ is a C^k map from U to $C([0, t_0]; X)$.*

Proof. The proof of Proposition 2.1 shows that if R and t_0 are sufficiently small and p is close enough to p_0 in L^1 -norm then T_p is a uniform contraction. Also, T_p is a C^k function of w and p on the interior of \mathcal{F} , so that the C^1 result follows from Hale [1969, Thm. 3.2, p. 7]. The C^k result is then obtained by induction. \square

COROLLARY 2.3. *The map $w(t_0; \cdot, w_0) : U \rightarrow X$ is C^k .*

Proof. This follows from the chain rule, Proposition 2.2 and the fact that the map $w(\cdot) \mapsto w(t_0)$ is smooth (since it is continuous and linear from $C([0, t_0]; X)$ to X). \square

In the same way we see that the solution $w(t; \cdot, \cdot)$ is a C^k function of w_0 and p . However, in this paper we are primarily concerned with differentiability in p . The proof of the theorem in Hale [1969] cited above shows that the derivative can be obtained by formally linearizing. Thus we get the following result.

COROLLARY 2.4. *The (Fréchet) derivative $D_p w(t; p_0, w_0) \cdot p$ of $w(t; p, w_0)$ with respect to p at p_0 in the direction p is the unique solution of the equation*

$$(2.2) \quad \begin{aligned} D_p w(t; p_0, w_0) \cdot p &= \int_0^t e^{\mathcal{A}(t-s)} p(s) \mathcal{B}(w(s; p_0, w_0)) ds \\ &+ \int_0^t e^{\mathcal{A}(t-s)} p_0(s) D\mathcal{B}(w(s; p_0, w_0)) D_p w(s; p_0, w_0) \cdot p ds. \end{aligned}$$

Here $D\mathcal{B}(w(s; p_0, w_0))$ denotes the Fréchet derivative of \mathcal{B} at $w(s; p_0, w_0)$. In particular, at $p_0 = 0$, $D_p w(t; 0, w_0) \cdot p$ is given explicitly by

$$(2.3) \quad D_p w(t; 0, w_0) \cdot p = \int_0^t e^{\mathcal{A}(t-s)} p(s) \mathcal{B}(e^{\mathcal{A}s} w_0) ds.$$

Next we show that solutions are globally defined under a sublinear growth condition.

THEOREM 2.5. *If there are constants C and K such that $\|\mathcal{B}(x)\| \leq C + K\|x\|$ for all $x \in X$, then (2.1) has solutions defined for $0 \leq t \leq T$. These solutions are unique within the class $C([0, T]; X)$. Moreover, the solution $w(t; p, w_0)$ is a C^k function of $p \in Z(T)$ and $w_0 \in X$ with (Fréchet) derivative in p given by (2.2) (or (2.3) if $p_0 = 0$).*

The proof is based on the following version of Gronwall's inequality (see, for example, Carroll [1969, p. 124]).

LEMMA 2.6. *Let $p \in L^1([a, b]; \mathbb{R})$ and let $v \in L^\infty([a, b]; \mathbb{R})$ with $v \geq 0$. If there exists a constant $C \geq 0$ such that for all $t \in [a, b]$*

$$v(t) \leq C + \int_0^t |p(s)| v(s) ds,$$

then

$$v(t) \leq C \exp \left(\int_a^t |p(s)| ds \right).$$

Proof of Theorem 2.5. Suppose $w(t)$ solves (2.1) and is defined for $0 \leq t < a \leq T$. Then

$$\|w(t)\| \leq M e^{\beta a} \left(\|w_0\| + \int_0^t |p(s)| (C + K\|w(s)\|) ds \right),$$

and so, assuming $K > 0$ without loss of generality, we get

$$\|w(t)\| \leq (M e^{\beta a} \|w_0\| + CK^{-1}) \exp \left(M e^{\beta a} K \int_0^t |p(s)| ds \right) - CK^{-1} \leq C_1.$$

Therefore, for $s, t \in [0, a)$ we have

$$\begin{aligned} \|w(t) - w(s)\| &\leq \|e^{\mathcal{A}t} w_0 - e^{\mathcal{A}s} w_0\| + \left\| \int_s^t e^{\mathcal{A}(t-\tau)} p(\tau) \mathcal{B}(w(\tau)) d\tau \right\| \\ &\leq \|e^{\mathcal{A}t} w_0 - e^{\mathcal{A}s} w_0\| + M e^{\beta a} (C + KC_1) \int_s^t |p(\tau)| d\tau. \end{aligned}$$

Thus $\lim_{t \rightarrow a-} w(t)$ exists, so that by Proposition 2.1 $w(t)$ can be continued beyond $t = a$. Hence solutions are defined for $0 \leq t \leq T$.

For global uniqueness, we use the standard argument: suppose $w(t)$ and $\bar{w}(t)$ solve (2.1) for $0 \leq t \leq T$. Let $S = \{a \in [0, T] \mid w(t) = \bar{w}(t) \text{ for } t \in [0, a]\}$. The local uniqueness assertion in Proposition 2.1 shows that S is relatively open in $[0, T]$. If $a_n \in S$ and $a_n \rightarrow a \leq T$ then $a \in S$ since $\lim_{n \rightarrow \infty} w(a_n) = \lim_{n \rightarrow \infty} \bar{w}(a_n)$. Thus S is closed, so that $S = [0, T]$.

Thus there is a globally defined semiflow $F_t^p(w_0)$, $F_t^p(\cdot): \mathbb{R}^+ \times X \rightarrow X$, which depends parametrically on p . Proposition 2.2 shows that $F_t^p(w_0)$ is C^k in p and w_0 for t sufficiently small. Let $\tilde{S} = \{a \in [0, T] \mid F_t^p(w_0) \text{ is } C^k \text{ in } (w_0, p) \text{ for } t \in [0, a]\}$. We claim that \tilde{S} is open. Indeed, if $a \in \tilde{S}$ and k is small,

$$F_{a+h}^p(w_0) = F_h^p(F_a^p(w_0))$$

is C^k in p and w_0 , because by Proposition 2.2 $F_h^p(w)$ is C^k in p and w for w near $F_a^{p_0}(w_0)$. The local uniformity of the time interval on which Proposition 2.2 holds shows that \tilde{S} is closed, and hence $\tilde{S} = [0, T]$.

Thus we have shown that $w(t; p, w_0)$ is C^k in p and w_0 . By differentiating (2.1) we obtain (2.2). \square

Remark 2.7. Suppose $w_0 \in D(A)$ and $p \in C^1([0, T]; \mathbb{R})$. Then $w(t) \in D(A)$ and $w(t)$ is differentiable and satisfies (1.1). This assertion follows from Segal [1963, Lemma 3.1] or from Tanabe [9, p. 102]. If merely $w_0 \in X$ and $p \in L^1([0, T]; \mathbb{R})$ then w is a "weak solution" of (1.1) (see Balakrishnan [1976] and Ball [1977]).

3. An abstract controllability theorem and a negative result. Define the linear operator $L_T : Z(T) \rightarrow X$ by

$$L_T p = \int_0^T e^{\mathcal{A}(T-s)} p(s) \mathcal{B}(e^{\mathcal{A}s} w_0) ds.$$

Then by (2.3) we have

$$(3.1) \quad D_p w(T; 0, w_0) \cdot p = L_T p.$$

A natural consequence of Theorem 2.5 is the following.

THEOREM 3.1. *Let \mathcal{A} be the infinitesimal generator of a C^0 semigroup of bounded linear operators on the Banach space X , and let $\mathcal{B} : X \rightarrow X$ be a C^k map, $k \geq 1$, which satisfies $\|\mathcal{B}x\| \leq C + K\|x\|$ for all $x \in X$, where C and K are constants. Suppose that $\text{Range}(L_T) = X$. Then there is an $\varepsilon > 0$ such that $w(T; p, w_0) = h$ for some $p \in Z(T)$, provided $\|h - e^{\mathcal{A}T} w_0\| < \varepsilon$.*

This result follows easily from the (generalized) inverse function theorem; a convenient reference is Luenberger [1969, p. 240]. The p that controls w_0 to hit h will be in a neighborhood of zero in $Z(T)$.

We note that if \mathcal{A} generates a group, surjectivity of L_T is equivalent to surjectivity of $\hat{L}_T : Z(T) \rightarrow X$, where

$$(3.2) \quad \hat{L}_T p = \int_0^T e^{-\mathcal{A}s} p(s) \mathcal{B}(e^{\mathcal{A}s} w_0) ds.$$

A major difficulty with Theorem 3.1 is that it is not usually an easy matter to check the surjectivity of L_T (or \hat{L}_T). In fact, as we shall prove in Theorem 3.6, if $\dim X = \infty$, L_T will not in general be surjective, though it may have dense range. This prevents us from applying Theorem 3.1 to partial differential equations.

We now present a basic criterion for L_T to have dense range.

PROPOSITION 3.2. *Suppose that*

$$\langle l, e^{\mathcal{A}(t-s)} \mathcal{B}(e^{\mathcal{A}s} w_0) \rangle = 0$$

for all $s, 0 \leq s \leq T$, where $l \in X^$ (the dual space of X), implies $l = 0$. Then $\text{Range}(L_T)$ is dense in X .*

Proof. $\text{Range}(L_T)$ is dense if the only $l \in X^*$ annihilating the range is $l = 0$. But

$$\langle l, L_T p \rangle = \int_0^T \langle l, e^{\mathcal{A}(T-s)} \mathcal{B}(e^{\mathcal{A}s} w_0) \rangle p(s) ds.$$

If this vanishes for all $p \in Z(T)$, then the continuous function $\langle l, e^{\mathcal{A}(T-s)} \mathcal{B}(e^{\mathcal{A}s} w_0) \rangle$ must vanish. This follows because $Z(T)$ is dense in $L^1([0, T]; \mathbb{R})$. Our hypothesis then gives $l = 0$.

Remark 3.3. If \mathcal{B} is linear and \mathcal{A} is a bounded linear operator, then

$$e^{-\mathcal{A}s} \mathcal{B} e^{\mathcal{A}s} w_0 = \mathcal{B} w_0 + s[\mathcal{A}, \mathcal{B}] w_0 + \frac{s^2}{2} [\mathcal{A}, [\mathcal{A}, \mathcal{B}]] w_0 + \cdots$$

(i.e., the Campbell–Baker–Hausdorff formula), where $[\mathcal{A}, \mathcal{B}] = -\mathcal{A}\mathcal{B} + \mathcal{B}\mathcal{A}$. From Proposition 3.2, we see that $\text{Range}(L_T)$ is dense in X for all $T > 0$ if the closure of the span of $\mathcal{B}w_0, [\mathcal{A}, \mathcal{B}]w_0, [\mathcal{A}, [\mathcal{A}, \mathcal{B}]]w_0, \dots$ is dense in X .

The next two well-known controllability results now follow for $X = \mathbb{R}^n$ and \mathcal{B} linear.

COROLLARY 3.4 (Hermes [1974], Lobry [1970]). *Assume $X = \mathbb{R}^n$ and that $\dim \text{span} \{\mathcal{B}w_0, [\mathcal{A}, \mathcal{B}]w_0, [\mathcal{A}, [\mathcal{A}, \mathcal{B}]]w_0, \dots\} = n$. Then for every $T > 0$ there is an $\varepsilon_T > 0$ with the property that if $\|e^{\mathcal{A}T} w_0 - h\| < \varepsilon_T$, we can find a $p \in Z(T)$ such that $w(T; p, w_0) = h$.*

Here one can choose $Z(T) = L^q([0, T]; \mathbb{R})$ for any q , $1 \leq q \leq \infty$, or $Z(T) = C^k([0, T]; \mathbb{R})$, for example.

COROLLARY 3.5 (Lobry [1970], Jurdjevic and Quinn [1978]). *Let the hypotheses of Corollary 3.4 hold. Assume $e^{\mathcal{A}t} w_0$ is almost periodic. Then for any $k \geq 0$, there exist $T > 0$ and $\varepsilon > 0$ such that $\|h - w_0\| < \varepsilon$ implies $w(T; p, w_0) = h$ for some $p \in C^k([0, T]; \mathbb{R})$.*

Proof. Let $T_1 > 0$ be fixed and let $\varepsilon_{T_1} > 0$ be as in Corollary 3.4. We show that if $\|h - w_0\| < \varepsilon_{T_1}/2$, then there exists $\tau > 0$ such that $w(T_1 + \tau; p, w_0) = h$ for some $p \in C^k([0, T_1 + \tau]; \mathbb{R})$. First, by the almost periodicity of $e^{\mathcal{A}t} w_0$, there exists $\tau > 0$ such that

$$\|e^{\mathcal{A}\tau} w_0 - e^{-\mathcal{A}T_1} w_0\| < \frac{\varepsilon_{T_1}}{2} \|e^{\mathcal{A}T_1}\|^{-1}.$$

We run (2.1) from time $t = 0$ until $t = \tau$ with $p \equiv 0$, so that $w(\tau) = e^{\mathcal{A}\tau} w_0$. By Corollary 3.4, we can hit h in additional time T_1 , using a C^k control which vanishes together with its first k derivatives at τ , provided $\|e^{\mathcal{A}T_1} w(\tau) - h\| < \varepsilon_{T_1}$. But this is true, since

$$\begin{aligned} \|e^{\mathcal{A}T_1} w(\tau) - h\| &= \|e^{\mathcal{A}T_1} e^{\mathcal{A}\tau} w_0 - h\| = \|e^{\mathcal{A}T_1} (e^{\mathcal{A}\tau} w_0 - e^{-\mathcal{A}T_1} w_0 + e^{-\mathcal{A}T_1} w_0) - h\| \\ &\leq \|e^{\mathcal{A}T_1}\| \|e^{\mathcal{A}\tau} w_0 - e^{-\mathcal{A}T_1} w_0\| + \|w_0 - h\| < \varepsilon_{T_1}. \end{aligned} \quad \square$$

In the case $\dim X = \infty$ things are quite different. Specifically, we shall now show that for a large class of spaces $Z(T)$, the map $w(T; \cdot, w_0) : Z(T) \rightarrow X$ will never cover an open neighborhood of $e^{\mathcal{A}T} w_0$ (and consequently L_T cannot be onto). Thus, for these $Z(T)$'s, Theorem 3.1 will be vacuous unless $\dim X < \infty$.

THEOREM 3.6. *Let X be a Banach space with $\dim X = \infty$. Let \mathcal{A} generate a C^0 semigroup of bounded linear operators on X and let $\mathcal{B} : X \rightarrow X$ be a bounded linear operator. Let $w_0 \in X$ be fixed and let $w(t; p, w_0)$ denote the unique solution of (2.1) for $p \in L^1_{\text{loc}}([0, \infty); \mathbb{R})$. If $T > 0$ and $p_n \rightarrow p$ weakly in $L^1([0, T]; \mathbb{R})$, then $w(\cdot; p_n, w_0) \rightarrow w(\cdot; p, w_0)$ strongly in $C([0, T]; X)$. Moreover, the set of states accessible from w_0 defined by*

$$S(w_0) = \bigcup_{\substack{t \geq 0 \\ p \in L^1_{\text{loc}}([0, \infty); \mathbb{R}) \\ r > 1}} w(t; p, w_0)$$

is contained in a countable union of compact subsets of X , and in particular has dense complement.

Proof. Let $p_n \rightarrow p$ weakly in $L^1([0, T]; \mathbb{R})$. Write $w_n(t) = w(t; p_n, w_0)$, $w(t) = w(t; p, w_0)$, and $z_n(t) = w_n(t) - w(t)$. Then

$$w_n(t) = e^{\mathcal{A}t} w_0 + \int_0^t p_n(s) e^{\mathcal{A}(t-s)} \mathcal{B} w_n(s) ds$$

and

$$w(t) = e^{\mathcal{A}t} w_0 + \int_0^t p(s) e^{\mathcal{A}(t-s)} \mathcal{B} w(s) ds,$$

so that

$$(3.3) \quad z_n(t) = \int_0^t [p_n(s) - p(s)] e^{\mathcal{A}(t-s)} \mathcal{B} w(s) ds + \int_0^t p_n(s) e^{\mathcal{A}(t-s)} \mathcal{B} z_n(s) ds.$$

We now need the following:

LEMMA 3.7. *Let*

$$\varepsilon_n = \sup_{t \in [0, T]} \left\| \int_0^t [p_n(s) - p(s)] e^{\mathcal{A}(t-s)} \mathcal{B} w(s) ds \right\|.$$

Then $\lim_{n \rightarrow \infty} \varepsilon_n = 0$.

Proof of Lemma 3.7. Suppose the lemma is false. Then there exist $\varepsilon > 0$, a subsequence $\{p_\mu\}$ of $\{p_n\}$ and a sequence $\{t_\mu\} \subset [0, T]$, $t_\mu \rightarrow t \in [0, T]$, such that for all μ

$$(3.4) \quad \left\| \int_0^{t_\mu} [p_\mu(s) - p(s)] e^{\mathcal{A}(t_\mu-s)} \mathcal{B} w(s) ds \right\| > \varepsilon.$$

We can suppose without loss of generality that either $t_\mu \leq t$ for all μ , or $t_\mu \geq t$ for all μ . In the case $t_\mu \leq t$ let

$$c_\mu = \sup_{s \in [0, t_\mu]} \|(e^{\mathcal{A}(t_\mu-s)} - e^{\mathcal{A}(t-s)}) \mathcal{B} w(s)\|.$$

The joint continuity of the map $(x, \tau) \mapsto e^{\mathcal{A}\tau} x$ and the continuity of $w(\cdot)$ together imply that $c_\mu \rightarrow 0$ as $\mu \rightarrow \infty$. Hence

$$(3.5) \quad \begin{aligned} \lim_{\mu \rightarrow \infty} \left\| \int_0^{t_\mu} [p_\mu(s) - p(s)] (e^{\mathcal{A}(t_\mu-s)} - e^{\mathcal{A}(t-s)}) \mathcal{B} w(s) ds \right\| \\ \leq \lim_{\mu \rightarrow \infty} c_\mu \int_0^{t_\mu} |p_\mu(s) - p(s)| ds = 0. \end{aligned}$$

Furthermore, since $p_\mu \rightarrow p$ weakly in $L^1([0, T]; \mathbb{R})$, $|p_\mu - p|$ is uniformly equi-integrable over $[0, T]$ (see Dunford and Schwartz [1964, pp. 293–294]), and hence

$$(3.6) \quad \lim_{\mu \rightarrow \infty} \left\| \int_{t_\mu}^t [p_\mu(s) - p(s)] e^{\mathcal{A}(t-s)} \mathcal{B} w(s) ds \right\| \leq \text{const} \cdot \lim_{\mu \rightarrow \infty} \int_{t_\mu}^t |p_\mu(s) - p(s)| ds = 0.$$

Combining (3.5) and (3.6), we deduce that

$$(3.7) \quad \lim_{\mu \rightarrow \infty} \left\| \int_0^{t_\mu} [p_\mu(s) - p(s)] e^{\mathcal{A}(t_\mu-s)} \mathcal{B} w(s) ds - \int_0^t [p_\mu(s) - p(s)] v(s) ds \right\| = 0,$$

where $v(s)$ is defined by $v(s) = e^{\mathcal{A}(t-s)} \mathcal{B} w(s)$. A similar argument shows that (3.7) holds if $t_\mu \geq t$ for all μ .

Let $\rho = \sup_{\mu} \int_0^t |p_{\mu}(s) - p(s)| ds$. Since $v \in C([0, T]; X)$ there exists a step function g such that $\|g - v\|_{L^{\infty}([0, T]; X)} < \varepsilon/4\rho$. Suppose $g(s) = \sum_{j=1}^M \chi_{I_j}(s) e_j$, where the I_j are disjoint intervals and $e_j \in X$. Then

$$\int_0^t [p_{\mu}(s) - p(s)]g(s) ds = \sum_{j=1}^M \int_{I_j \cap [0, t]} [p_{\mu}(s) - p(s)] ds e_j,$$

which tends to zero as $\mu \rightarrow \infty$ from the weak convergence of p_{μ} . Therefore

$$(3.8) \quad \left\| \int_0^t [p_{\mu}(s) - p(s)]v(s) ds \right\| \leq \frac{\varepsilon}{4\rho} \int_0^t |p_{\mu}(s) - p(s)| ds + \left\| \int_0^t [p_{\mu}(s) - p(s)]g(s) ds \right\| \leq \frac{\varepsilon}{2}$$

for large enough μ . We now combine (3.8) with (3.4) and (3.7) to reach a contradiction, which proves the lemma. \square

Continuation of proof of Theorem 3.6. From (3.3) we have

$$\|z_n(t)\| \leq \varepsilon_n + \int_0^t |p_n(s)| \|e^{\mathcal{A}(t-s)}\| \|\mathcal{B}\| \|z_n(s)\| ds \leq \varepsilon_n + C \int_0^t |p_n(s)| \|z_n(s)\| ds,$$

where C is a positive constant independent of $t \in [0, T]$. By Gronwall's inequality

$$\|z_n(t)\| \leq \varepsilon_n \exp \left(C \int_0^t |p_n(s)| ds \right),$$

which by the lemma tends to zero uniformly in $[0, T]$ as $n \rightarrow \infty$. This proves the first part of the theorem.

To prove the second part, given positive integers m, n and r , define

$$S_{mnr}(w_0) = \bigcup_{\substack{t \in [0, m] \\ \|p\|_{L^{1+1/r}([0, m]; \mathbb{R})} \leq n}} w(t; p, w_0).$$

Let $w(t_i; p_i, w_0) \in S_{mnr}(w_0)$. Since $L^{1+1/r}([0, m]; \mathbb{R})$ is reflexive there exist subsequences $\{t_{\mu}\} \subset [0, n]$ and $\{p_{\mu}\} \subset L^{1+1/r}([0, m]; \mathbb{R})$, such that $t_{\mu} \rightarrow t$ and $p_{\mu} \rightarrow p$ weakly in $L^{1+1/r}([0, m]; \mathbb{R})$. By the first part of the theorem, $w(t_{\mu}; p_{\mu}, w_0) \rightarrow w(t; p, w_0)$ in X . Hence $S_{mnr}(w_0)$ is precompact in X . But $S(w_0) \subset \bigcup_{m,n,r=1}^{\infty} S_{mnr}(w_0)$ so that $S(w_0)$ is contained in a countable union of compact sets.

Since $\dim X = \infty$, $S_{mnr}(w_0)$ is nowhere dense. By the Baire category theorem, $S(w_0)$ has dense complement. \square

Remark 3.8. The theorem leaves open the question of whether

$$\{w(t; p, w_0); t \geq 0, p \in L^1_{\text{loc}}([0, \infty); \mathbb{R})\}$$

has dense complement. We show in Theorem 5.5 that this holds in an important special case.

4. Finite-dimensional observability. In this section we consider the restricted problem of trying to control only a finite-dimensional projection of the state variable $w(t; p, w_0)$; i.e., we try to control only a “finite number of modes.” This problem was discussed originally by Hermes [1979], and our first result is analogous to his.

THEOREM 4.1. *Let \mathcal{A}, \mathcal{B} be as in Theorem 3.1. Suppose $G: X \rightarrow \mathbb{R}^n$ is a bounded linear map. Suppose that for given $T > 0$ and $\lambda \in (\mathbb{R}^n)^*$,*

$$\langle \lambda, G e^{\mathcal{A}(T-s)} \mathcal{B}(e^{\mathcal{A}s} w_0) \rangle = 0$$

for all $s, 0 \leq s \leq T$ implies $\lambda = 0$. Then there is an $\varepsilon_T > 0$ such that $\|q - G e^{\mathcal{A}T} w_0\|_{\mathbb{R}^n} < \varepsilon_T$ implies $Gw(T; p, w_0) = q$ for some $p \in Z(T)$.

Proof. The derivative of the map $p \mapsto Gw(t; p, w_0)$ from $Z(T)$ to the range of G , evaluated at $p = 0$ is the operator GL_T . To show this is surjective, let $\lambda \in (\mathbb{R}^n)^*$ and

assume λ annihilates the range of GL_T . An argument similar to the proof of Proposition 3.2 shows that $\lambda = 0$.

COROLLARY 4.2. *Let \mathcal{A} , \mathcal{B} and G be as in Theorem 4.1, where G is now assumed to be surjective. Suppose the hypothesis of Proposition 3.2 holds. Then there is an $\varepsilon_T > 0$ such that*

$$\|q - Ge^{\mathcal{A}T}w_0\|_{\mathbb{R}^n} < \varepsilon_T \text{ implies } Gw(T; p, w_0) = q \text{ for some } p \in Z(T).$$

Proof. Set $l = G^*\lambda$, where G^* is the adjoint of G , and use Theorem 4.1. \square

The usefulness of Corollary 4.2 is that it applies to *all* surjective bounded maps $G: X \rightarrow \mathbb{R}^n$, n arbitrary.

COROLLARY 4.3. *Assume that either the hypotheses of Theorem 4.1 or those of Corollary 4.2 hold for some $T_1 > 0$ and that $e^{\mathcal{A}t}$ is a group with $e^{\mathcal{A}t}w_0$ an almost periodic function of t . Then for any $k \geq 0$ there exist $T > 0$ and $\varepsilon_T > 0$ such that $\|q - Gw_0\|_{\mathbb{R}^n} < \varepsilon_T$ implies $Gw(T; p, w_0) = q$ for some $p \in C^k([0, T]; \mathbb{R})$.*

Proof. This is very similar to the proof of Corollary 3.5. \square

We note that the above results could be extended to nonlinear $G \in C^1(X; \mathbb{R}^n)$ in the obvious way.

One approach to trying to obtain full state controllability might be to solve an infinite sequence of finite-dimensional controllability problems by letting $n \rightarrow \infty$. This possibility will be precluded by Theorem 3.6. More specifically, we note:

COROLLARY 4.4. *Let $\{X_n\}$ be an increasing sequence of subspaces of X , with $\dim X_n = n$ for each n such that $\text{Closure}(\bigcup_{n=1}^{\infty} X_n) = X$, and with corresponding continuous projections G_n of X onto X_n having uniformly bounded norms. If*

$$H = \{h \in X; \text{there exist } T > 0, r > 1 \text{ and } \{p_n\} \subset L^r([0, T]; \mathbb{R}) \text{ such that } G_n w(T; p_n, w_0) = G_n h \text{ and } \|p_n\|_{L^r([0, T]; \mathbb{R})} \leq \text{const (independent of } n), n = 1, 2, \dots\},$$

then H has dense complement in X .

Proof. Let $h \in H$. Then there exists a corresponding sequence $\{p_n\} \subset L^r([0, T]; \mathbb{R})$, $r > 1$. Since $\{p_n\}$ is bounded, there exists a subsequence, also denoted by $\{p_n\}$, such that $p_n \rightarrow p$ weakly in $L^1([0, T]; \mathbb{R})$. Now

$$\begin{aligned} \|w(T; p, w_0) - h\| &\leq \|w(t; p, w_0) - G_n w(T; p, w_0)\| \\ (4.1) \quad &+ \|G_n w(T; p, w_0) - G_n w(T; p_n, w_0)\| \\ &+ \|G_n w(T; p_n, w_0) - G_n h\| + \|G_n h - h\|. \end{aligned}$$

Since the G_n are projections having uniformly bounded norms the first and last terms on the right-hand side of (4.1) tend to zero as $n \rightarrow \infty$. By hypothesis the third term is identically zero. As to the second term, $w(T; p_n, w_0) \rightarrow w(T; p, w_0)$ by Theorem 3.6 and $\|G_n\| \leq \text{const}$, so that this tends to zero also. Hence $h = w(T; p, w_0)$ and so H is a subset of the attainable set $S(w_0)$, which by Theorem 3.6 has dense complement. \square

In practical terms Corollary 4.4 says that, in general, approximation of the problem $w(T; p, w_0) = h$ by a sequence of finite-dimensional problems will inevitably lead to the need for ever larger controls p_n as $n \rightarrow \infty$. In this sense, finite-dimensional approximations can be misleading for control of the full problem.

5. Abstract hyperbolic equations. We now investigate systems of the form

$$(5.1) \quad \ddot{u} + Au + p(t)Bu = 0,$$

$$(5.2) \quad u(0) = u_0 \in D(A^{1/2}), \quad \dot{u}(0) = u_1 \in H,$$

where A is a positive definite self-adjoint operator with dense domain $D(A)$ in the real Hilbert space H , B is a bounded linear operator from $D(A^{1/2})$ to H , and p is a real-valued control. The inner product in H is denoted (\cdot, \cdot) . We suppose that A^{-1} is compact, and that A has simple eigenvalues λ_n^2 , $n = 1, 2, \dots$, where $0 < \lambda_1 < \lambda_2 < \dots$. Then there exists a corresponding complete orthonormal basis $\{\phi_n\}$ of eigenfunctions: $A\phi_n = \lambda_n^2\phi_n$, $(\phi_n, \phi_m) = \delta_{mn}$.

To investigate controllability of (5.1) we could rewrite (5.1) in first order form

$$w = \begin{pmatrix} u \\ \dot{u} \end{pmatrix}, \quad \mathcal{A} = \begin{pmatrix} 0 & I \\ -A & 0 \end{pmatrix}, \quad \mathcal{B} = \begin{pmatrix} 0 & 0 \\ -B & 0 \end{pmatrix}$$

and set $X = D(A^{1/2}) \times H$ with inner product

$$\langle (u_1, u_2), (v_1, v_2) \rangle_X = (\mathcal{A}^{1/2}u_1, A^{1/2}v_1) + (u_2, v_2).$$

With this set-up, we see that \mathcal{A} generates a C^0 group of isometries on X and the hypotheses of Theorem 2.5 are satisfied. Controllability then hinges on the operator \hat{L}_T . To facilitate computations, however, it is advantageous to introduce a different first order form. We therefore set up a complex structure in a way that is standard for Hamiltonian systems (see Chernoff and Marsden [1974, § 2.7]).

Let \mathcal{H} denote the complexified Hilbert space $H \oplus iH$ with inner product defined by

$$\langle x_1 + iy_1, x_2 + iy_2 \rangle_{\mathcal{H}} = (x_1, x_2) + (y_1, y_2) + i[(y_1, x_2) - (x_1, y_2)]$$

for $x_1, x_2, y_1, y_2 \in H$. The map $\psi: X \rightarrow \mathcal{H}$ defined by

$$\psi(u_1, u_2) = A^{1/2}u_1 + iu_2$$

is an isometry. Let $z = A^{1/2}u + i\dot{u}$, so that (5.1), (5.2) become

$$(5.3) \quad i\dot{z} = A^{1/2}z + p(t)BA^{-1/2}\operatorname{Re} z,$$

$$(5.4) \quad z(0) = z_0,$$

where

$$(5.5) \quad z_0 = A^{1/2}u_0 + iu_1 \in \mathcal{H}.$$

Of course, in (5.3) $p(t)$ is still *real*. Writing $\hat{\mathcal{A}} = -iA^{1/2}$ (regarded as a complex operator) and $\hat{\mathcal{B}} = -iBA^{-1/2}\operatorname{Re}$ (a real-linear bounded operator from \mathcal{H} into \mathcal{H}), we see that the hypotheses of Theorem 2.5 are satisfied.

The basis $\{\phi_n\}$ of H may also be regarded as a basis of \mathcal{H} . For any $z \in \mathcal{H}$, let $\{z_n\}$ be the (complex) components of z relative to this basis, i.e.,

$$(5.6) \quad z = \sum_{n=1}^{\infty} z_n \phi_n,$$

so that $\{z_n\} \in l_2$. Thus we have

$$(5.7) \quad e^{\hat{\mathcal{A}}s}z = \sum_{n=1}^{\infty} z_n e^{-i\lambda_n s} \phi_n.$$

Let $B_{mn} = (B\phi_m, \phi_n)$, so that the B_{mn} are real and

$$B\phi_m = \sum_{n=1}^{\infty} B_{mn}\phi_n.$$

Thus (5.7) gives

$$\hat{\mathcal{B}} e^{\hat{\mathcal{A}}s}z = -i \sum_{m,n=1}^{\infty} \frac{B_{mn}}{\lambda_m} \operatorname{Re} (e^{-i\lambda_m s} z_m) \phi_n,$$

and so

$$(5.8) \quad e^{-\hat{\mathcal{A}}s} \hat{\mathcal{B}} e^{\hat{\mathcal{A}}s} z = -\frac{i}{2} \sum_{m,n=1}^{\infty} \frac{B_{mn}}{\lambda_m} (e^{i(\lambda_n - \lambda_m)s} z_m + e^{i(\lambda_n + \lambda_m)s} \bar{z}_m) \phi_n.$$

5.1. Riesz bases.

DEFINITION. A sequence of elements $\{\omega_j\}_{j=1}^{\infty}$ of a (real or complex) Hilbert space Z is called a *Riesz basis* of Z if every $\theta \in Z$ has a unique expansion

$$\theta = \sum_{j=1}^{\infty} a_j \omega_j$$

that is convergent in Z , and

$$C_1 \sum_{m=1}^{\infty} |a_j|^2 \leq \|\theta\|^2 \leq C_2 \sum_{j=1}^{\infty} |a_j|^2$$

for absolute positive constants C_1, C_2 .

We collect together some useful facts concerning Riesz bases.

LEMMA 5.1. *Let $\{\omega_j\}$ be a Riesz basis of Z , and let $\{e_j\}$ be any complete orthonormal basis of Z . Then:*

(i) *the formula $T(\sum_{j=1}^{\infty} a_j e_j) = \sum_{j=1}^{\infty} a_j \omega_j$ defines an isomorphism*

$$T: Z \rightarrow Z;$$

(ii) *for any $\theta \in Z$,*

$$\sum_{j=1}^{\infty} |(\theta, \omega_j)|^2 \leq \|T^*\|^2 \|\theta\|^2;$$

(iii) *given any sequence $\{a_n\} \in l_2$ there exists a unique solution $\theta \in Z$ of the equations*

$$(5.9) \quad (\theta, \omega_j) = a_j, \quad j = 1, 2, \dots$$

Proof. For a proof of (i) see Gohberg and Krein [1969, p. 310]. To prove (ii) note that $(\theta, \omega_j) = (\theta, T e_j) = (T^* \theta, e_j)$, so that

$$\sum_{j=1}^{\infty} |(\theta, \omega_j)|^2 = \|T^* \theta\|^2 \leq \|T^*\|^2 \|\theta\|^2.$$

Finally, the equations (5.9) are equivalent to

$$(T^* \theta, e_j) = a_j,$$

and thus have the unique solution

$$\theta = (T^*)^{-1} \sum_{j=1}^{\infty} a_j e_j. \quad \square$$

A useful criterion for the construction of a Riesz basis is as follows.

THEOREM 5.2. *Let $0 = \mu_0 < \mu_1 < \mu_2 < \dots$, $\mu_{-k} = -\mu_k$, and suppose that*

$$\lim_{k \rightarrow \infty} (\mu_{k+1} - \mu_k) \geq \gamma > 0.$$

Then for any $T > 2\pi/\gamma$ the functions $\{e^{i\mu_k t}\}_{k=-\infty}^{\infty}$ may be extended to a Riesz basis of $L^2([0, T]; \mathbb{C})$.

Proof. Let S denote the closed linear span of the set of functions $\{e^{i\mu_k t}\}$ in $L^2([0, T]; \mathbb{C})$. It follows from Ball and Slemrod [1979, Thm. 2.1] (the essential idea is due to Ingham) that for any finite sum

$$f(t) = \sum_{|k| \leq N} a_k e^{i\mu_k t},$$

we have

$$C_1 \sum_{|k| \leq N} |a_k|^2 \leq \frac{1}{T} \int_0^T |f(t)|^2 dt \leq C_2 \sum_{|k| \leq N} |a_k|^2.$$

It follows that any $f \in S$ has a unique expansion

$$f(t) = \sum_{k=-\infty}^{\infty} a_k e^{i\mu_k t}$$

convergent in $L^2([0, T]; \mathbb{C})$, and that

$$C_1 \sum_{k=-\infty}^{\infty} |a_k|^2 \leq \frac{1}{T} \int_0^T |f(t)|^2 dt \leq C_2 \sum_{k=-\infty}^{\infty} |a_k|^2.$$

Let $\{e_j\}$ be an orthonormal basis of S^\perp . It follows readily that $\{e_j\} \cup \{e^{i\mu_k t}\}$ is a Riesz basis of $L^2([0, T]; \mathbb{C})$. \square

The above discussion is a slightly different presentation of results summarized in Russell [1967].

5.2. Finite-dimensional observers. We now employ Theorem 4.1 to discuss when (5.1) is controllable relative to a finite-dimensional observer.

THEOREM 5.3. *Assume the initial data u_0, u_1 in (5.2) satisfy*

$$(i) \quad B_{nn}[(u_0, \phi_n)^2 + (u_1, \phi_n)^2] \neq 0, \quad n = 1, 2, \dots$$

and that $T > 0$ is such that

$$(ii) \quad \{e^{2i\lambda_n s}\}_{n=1}^{\infty} \cup \{e^{i(\lambda_p - \lambda_q)s}, e^{i(\lambda_p + \lambda_q)s} | p \neq q \text{ and } B_{pq} \neq 0\}$$

can be extended to a Riesz basis of $L^2([0, T]; C)$.

Then (5.3) satisfies the hypotheses of Proposition 3.2. In particular, for any $T_1 \geq T$ and bounded surjective maps $G_1: D(A^{1/2}) \rightarrow \mathbb{R}^m$, $G_2: H \rightarrow \mathbb{R}^n$, there exists ε_{T_1} such that if

$$\|q_1 - G_1 u(T_1; 0, u_0, u_1)\|_{\mathbb{R}^m} < \varepsilon_{T_1}, \quad \|q_2 - G_2 \dot{u}(T_1; 0, u_0, u_1)\|_{\mathbb{R}^n} < \varepsilon_{T_1},$$

then

$$G_1 u(T_1; p, u_0, u_1) = q_1, \quad G_2 \dot{u}(T_1; p, u_0, u_1) = q_2$$

for some $p \in Z(T_1)$. Here $u(t; p, u_0, u_1)$ is the solution of (5.1), (5.2).

Proof. Let $l = \sum_{n=1}^{\infty} l_n \phi_n$ be an arbitrary element of \mathcal{H} . Then $\langle l, e^{-\hat{\mathcal{A}}s} \hat{\mathcal{B}} e^{\hat{\mathcal{A}}s} z_0 \rangle_{\mathcal{H}}$ may be computed for (5.3) by using (5.8). Specifically we have

$$(5.10) \quad \begin{aligned} 2i \langle e^{-\hat{\mathcal{A}}s} \hat{\mathcal{B}} e^{\hat{\mathcal{A}}s} z_0, l \rangle_{\mathcal{H}} &= \sum_{n=1}^{\infty} \bar{l}_n (z_{0n} + \bar{z}_{0n} e^{2i\lambda_n s}) \frac{B_{nn}}{\lambda_n} \\ &+ \sum_{\substack{m \neq n \\ m, n=1}}^{\infty} \bar{l}_n (z_{0m} e^{i(\lambda_n - \lambda_m)s} + \bar{z}_{0m} e^{i(\lambda_n + \lambda_m)s}) \frac{B_{mn}}{\lambda_m}, \end{aligned}$$

where z_0 is given by (5.5) and where

$$z_0 = \sum_{n=1}^{\infty} z_{0n} \phi_n,$$

so that $z_{0n} = \lambda_n(u_0, \phi_n) + i(u_1, \phi_n)$. Thus, if $\langle e^{-\hat{\mathcal{A}}s} \hat{\mathcal{B}} e^{\hat{\mathcal{A}}s} z_0, l \rangle_{\mathcal{H}} = 0$ for all s such that $0 \leq s \leq T_1$, the right-hand side of (5.10) will equal zero on $[0, T]$. By assumption (ii) the coefficients of $\{e^{2i\lambda_n s}\}$ vanish; that is,

$$\frac{\bar{l}_n \bar{z}_{0n} B_{nn}}{\lambda_n} = 0 \quad \text{for } n = 1, 2, \dots$$

By (i) this implies $l_n = 0$ for $n = 1, 2, \dots$, and hence $l = 0$. Therefore, the hypothesis of Proposition 3.2 is satisfied, and by Corollary 4.2 the result follows. \square

COROLLARY 5.4. *Assume the hypotheses of Theorem 5.3 are satisfied, and let G_1, G_2 be bounded surjective linear maps, $G_1: D(A^{1/2}) \rightarrow \mathbb{R}^m$, $G_2: H \rightarrow \mathbb{R}^n$. Then for any $k \geq 0$ there exist $T_1 > 0$ and $\varepsilon_{T_1} > 0$ such that*

$$\|q_1 - G_1 u_0\|_{\mathbb{R}^m} < \varepsilon_{T_1}, \quad \|q_2 - G_2 u_1\|_{\mathbb{R}^n} < \varepsilon_{T_1},$$

imply

$$G_1 u(T_1; p, u_0, u_1) = q_1, \quad G_2 \dot{u}(T_1; p, u_0, u_1) = q_2$$

for some $p \in C^k([0, T_1]; \mathbb{R})$.

Proof. The result follows immediately from Theorem 5.3 and Corollary 4.3. \square

Hypothesis (ii) of Theorem 5.3 is difficult to verify unless $B_{pq} = 0$ for $p \neq q$. Sufficient conditions for it to hold may be deduced from Theorem 5.2, but they are not revealing except in the case just mentioned.

5.3. Approximate controllability. In this subsection we study approximate controllability, in a sense to be made precise, of (5.1), (5.2). As above we work with the equivalent first order system

$$(5.11) \quad \dot{z} = \hat{\mathcal{A}}z + p(t)\hat{\mathcal{B}}z$$

where $\hat{\mathcal{A}} = -iA^{1/2}$, $\hat{\mathcal{B}} = -iBA^{-1/2}$ Re. In addition, to simplify matters we make the assumption

$$(D1) \quad B_{mn} = b_m \delta_{mn}$$

for nonzero constants b_m , where δ_{mn} is the Kronecker delta. Since (D1) implies that $B_{mn} = 0$ for $m \neq n$, we shall refer to (D1) as the *diagonal case*.

Writing

$$z(t) = \sum_{n=1}^{\infty} z_n(t) \phi_n,$$

we see that in the diagonal case, (5.11) reduces to the infinite system of uncoupled ordinary differential equations

$$(5.12) \quad \dot{z}_n = -i\lambda_n z_n - ip(t) \frac{b_n}{\lambda_n} \operatorname{Re} z_n, \quad n = 1, 2, \dots$$

The corresponding initial conditions are

$$(5.13) \quad z_n(0) = z_{0n}.$$

We note that the fact that $BA^{-1/2}$ is a bounded linear operator from $H \rightarrow H$ is equivalent to the condition

$$(5.14) \quad \left\{ \frac{b_n}{\lambda_n} \right\} \in l_{\infty}.$$

We first strengthen Theorem 3.6 in the diagonal case by showing that even when L^1 controls are allowed, exact controllability is in general impossible.

THEOREM 5.5. *Given $\{z_{0n}\} \in l_2$, the set*

$$\bigcup_{\substack{t \geq 0 \\ p \in L^1_{\text{loc}}([0, \infty); \mathbb{R})}} \{z_n(t; p, z_0)\}$$

is contained in a countable union of compact sets of l_2 , and thus has dense complement.

Here, $\{z_n(t; p, z_0)\}$ denotes the unique mild solution of (5.12), (5.13) with $z_0 = \{z_{0n}\}$. Consequently the attainability set $\{u(t; p, u_0, u_1), u_t(t; p, u_0, u_1) \mid t \geq 0, p \in L^1_{\text{loc}}([0, \infty))\}$ is contained in the countable union of compact sets in $D(A^{1/2}) \times H$ and so has a dense complement.

Proof. Since

$$z_n(t) = e^{-i\lambda_n t} z_{0n} = i \frac{b_n}{\lambda_n} \int_0^t e^{-i\lambda_n(t-s)} p(s) \operatorname{Re} z_n(s) ds,$$

it follows that

$$|z_n(t)| \leq |z_{0n}| + \left| \frac{b_n}{\lambda_n} \right| \int_0^t |p(s)| |z_n(s)| ds,$$

and hence, by Gronwall's inequality and (5.14)

$$|z_n(t)| \leq |z_{0n}| \exp \left(\kappa \int_0^t |p(s)| ds \right),$$

where $\kappa = \|\{b_n/\lambda_n\}\|_{l_\infty}$. Thus $\{z_n(t)\} \in \bigcup_{N=1}^\infty S_N(z_0)$ for any $t \geq 0$ and $p \in L^1_{\text{loc}}([0, \infty); \mathbb{R})$, where S_N is defined by

$$S_N(z_0) = \{\{a_n\} \in l_2 : |a_n| \leq N |z_{0n}|\}.$$

The result now follows from the next lemma.

LEMMA 5.6. $S_N(z_0)$ is a compact subset of l_2 .

*Proof.*¹ Let $a^{(r)} \in S_N(z_0)$, $r = 1, 2, \dots$. Then

$$\sum_{n=1}^\infty |a_n^{(r)}|^2 \leq N^2 \sum_{n=1}^\infty |z_{0n}|^2 = N^2 \|z_0\|_{l_2}^2.$$

So some subsequence $a^{(\mu)} \rightarrow a$ weakly in l_2 , which implies in particular that $a_n^{(\mu)} \rightarrow a_n$ for each n . Also, given $\varepsilon > 0$

$$\sum_{n=M}^\infty |a_n^{(\mu)}|^2 \leq N^2 \sum_{n=M}^\infty |z_{0n}|^2 < \varepsilon$$

for M sufficiently large. Therefore $\sum_{n=1}^\infty |a_n^{(\mu)}|^2 \rightarrow \sum_{n=1}^\infty |a_n|^2$, and so $a^{(\mu)} \rightarrow a$ strongly in l_2 . Hence $S_N(z_0)$ is precompact. Since $S_N(z_0)$ is closed, the lemma is proved. \square

We now make the following additional assumption

$$(D2) \quad \frac{b_n}{\lambda_n} = c + \gamma_n \quad \text{for some } c \in \mathbb{R} \text{ and } \{\gamma_n\} \in l_2.$$

We write $P(t) = \int_0^t p(s) ds$ and make the following change of variables (motivated by averaging):

$$(5.15) \quad \zeta_n = \frac{\lambda_n}{b_n} \left[\frac{z_n}{z_{0n}} \exp i \left(\lambda_n t + \frac{b_n}{2\lambda_n} p(t) \right) - 1 \right].$$

Substitution of (5.15) into (5.12) yields

$$(5.16) \quad \dot{\zeta}_n(t) = -i \frac{p(t)}{2} \frac{\bar{z}_{0n}}{z_{0n}} \left(\frac{b_n}{\lambda_n} \bar{\zeta}_n(t) + 1 \right) \exp \left[2i \left(\lambda_n t + \frac{b_n}{2\lambda_n} p(t) \right) \right],$$

$$(5.17) \quad \zeta_n(0) = 0.$$

¹ This lemma follows from Dunford and Schwartz [1964, p. 338]. We have included the proof for completeness.

The following existence and differentiability theorem gives conditions under which the solution $\{\zeta_n(t)\}$ of (5.16), (5.17) belongs to l_2 , and thus gives more precise information on the attainable set (but under stronger hypotheses) than Theorem 5.5.

THEOREM 5.7. *Suppose $\{z_{0n}\} \in l_2$, $z_{0n} \neq 0$ for all $n = 1, 2, \dots$, and that $\{e^{2i\lambda_n t}\}$ can be extended to a Riesz basis of $L^2([0, l]; \mathbb{R})$ for some $l > 0$. Let $p \in L^2_{\text{loc}}([0, \infty); \mathbb{R})$. Then (5.16), (5.17) have a unique absolutely continuous solution $\zeta_n = \zeta_n(t; p)$ defined for all $t \geq 0$, and $\{\zeta_n(\cdot; p)\} \in C([0, T]; l_2)$ for $0 < T \leq l$. Furthermore, the mapping $p \mapsto \{\zeta_n(T; p)\}$ is C^1 from $L^2([0, T]; \mathbb{R})$ to l_2 for each $0 < T \leq l$, and*

$$(5.18) \quad D_p\{\zeta_n(T; 0)\} \cdot p = -\frac{i}{2} \frac{\bar{z}_{0n}}{z_{0n}} \int_0^T p(t) \exp(2i\lambda_n t) dt.$$

Proof. We write (5.16), (5.17) in integrated form:

$$(5.19) \quad \zeta_n(t) = -\frac{i}{2} \int_0^t p(s) \frac{\bar{z}_{0n}}{z_{0n}} \left(\frac{b_n}{\lambda_n} \bar{\zeta}_n(s) + 1 \right) \exp \left[2i \left(\lambda_n s + \frac{b_n}{2\lambda_n} p(s) \right) \right] ds.$$

We can solve these equations in a manner similar to Theorem 2.5, but for variety we shall adopt a standard device to get existence on an arbitrary time interval in a single step. Let $0 < T \leq l$. For any $\delta \geq 0$ the norm

$$\|\zeta\|_\delta = \sup_{t \in [0, T]} e^{-\delta t} \|\zeta(t)\|_{l_2}$$

on $X_T = C([0, T]; l_2)$ is equivalent to the usual one, namely $\|\cdot\|_0$. For $\zeta \in X_T$ define

$$((J_p \zeta)(t))_n = -\frac{i}{2} \int_0^t p(s) \frac{\bar{z}_{0n}}{z_{0n}} \left(\frac{b_n}{\lambda_n} \bar{\zeta}_n(s) + 1 \right) \exp \left[2i \left(\lambda_n s + \frac{b_n}{2\lambda_n} p(s) \right) \right] ds.$$

Then for $0 \leq \tau \leq t \leq T$

$$(5.20) \quad \begin{aligned} & \sum_{n=1}^{\infty} |(J_p \zeta(t) - J_p(\zeta(\tau)))_n|^2 \\ &= \frac{1}{4} \sum_{n=1}^{\infty} \left| \int_{\tau}^t p(s) \frac{\bar{z}_{0n}}{z_{0n}} \left(\frac{b_n}{\lambda_n} \bar{\zeta}_n(s) + 1 \right) \exp \left[2i \left(\lambda_n s + \frac{b_n}{2\lambda_n} p(s) \right) \right] ds \right|^2 \\ &\leq \frac{\kappa^2}{2} \sum_{n=1}^{\infty} \left(\int_{\tau}^t |p(s)| |\bar{\zeta}_n(s)| ds \right)^2 + \frac{1}{2} \sum_{n=1}^{\infty} \left| \int_{\tau}^t p(s) \exp \left[2i \left(\lambda_n s + \frac{b_n}{2\lambda_n} p(s) \right) \right] ds \right|^2, \end{aligned}$$

where, as before, $\kappa = \|\{b_n/\lambda_n\}\|_{L^\infty}$. But

$$\frac{\kappa^2}{2} \sum_{n=1}^{\infty} \left(\int_{\tau}^t |p(s)| |\bar{\zeta}_n(s)| ds \right)^2 \leq \frac{\kappa^2 T}{2} \left(\int_{\tau}^t |p(s)|^2 ds \right) \|\zeta\|_0^2$$

while

$$\begin{aligned} & \frac{1}{2} \sum_{n=1}^{\infty} \left| \int_{\tau}^t p(s) \exp \left[2i \left(\lambda_n s + \frac{b_n}{2\lambda_n} p(s) \right) \right] ds \right|^2 \\ &= \frac{1}{2} \sum_{n=1}^{\infty} \left| \int_{\tau}^t p(s) \exp(icP(s)) \exp(2i\lambda_n s) [1 + i\gamma_n p(s) + o(|\gamma_n|)] ds \right|^2 \\ &\leq C \int_{\tau}^t |p(s)|^2 ds \left[1 + \sum_{n=1}^{\infty} |\gamma_n|^2 \right], \end{aligned}$$

where C is a constant (depending on p), and where we have applied Lemma 5.1 (ii) to

the function $\theta(s) = \chi_{[\tau, t]}(s)p(s) \exp(icP(s))$ with $Z = L^2([0, l]; \mathbb{C})$. From (5.20) we thus deduce that J_p maps χ_T into itself.

Let $\zeta, \eta \in X_T$. Then

$$\begin{aligned} e^{-\delta t} \left(\sum_{n=1}^{\infty} |(J_p \zeta(t) - J_p \eta(t))_n|^2 \right)^{1/2} &\leq \frac{\kappa^2}{2} e^{-\delta t} \left(\sum_{n=1}^{\infty} \left(\int_0^t |p(s)| |\zeta_n(s) - \eta_n(s)| ds \right)^2 \right)^{1/2} \\ &\leq \frac{\kappa}{2} \left(\int_0^t |p(s)|^2 ds \right)^{1/2} \int_0^t e^{\delta(s-t)} ds \|\zeta - \eta\|_{\delta}. \end{aligned}$$

Hence J_p is a uniform contraction with respect to the norm $\|\cdot\|_{\delta}$ provided δ is sufficiently large. Calculations similar to those above show that J_p is C^1 in p . The result then follows as in Propositions 2.1, 2.2. \square

It is now easy to prove a local approximate controllability result.

THEOREM 5.8. Suppose $\{z_{0n}\} \in l_2$, $z_{0n} \neq 0$, $b_n \neq 0$, for all $n = 1, 2, \dots$, and that $\{1, e^{\pm 2i\lambda_n t}\}$ can be extended to a Riesz basis of $L^2([0, l]; \mathbb{C})$ for some $l > 0$. Then there exists $\varepsilon_l > 0$ such that if $\|h\|_{l_2} + |\theta| < \varepsilon_l$ where $h \in l_2$ and $\theta \in \mathbb{R}$, then

$$(5.21) \quad \frac{\lambda_n}{b_n} \left(\frac{z_n(l)}{z_{0n}} \exp \left[i \left(\lambda_n l + \frac{b_n}{2\lambda_n} \theta \right) \right] - 1 \right) = h_n, \quad n = 1, 2, \dots,$$

for some $p \in L^2([0, l]; \mathbb{R})$ with $\int_0^l p(t) dt = \theta$.

Proof. Consider the map $Q : L^2([0, l]; \mathbb{R}) \rightarrow l_2 \times \mathbb{R}$ defined by

$$Q(p) = \left(\{\zeta_n(l; p)\}, \int_0^l p(t) dt \right).$$

By (5.18),

$$D_p Q(0) \cdot p = \left(\left\{ -\frac{i}{2} \frac{\bar{z}_{0n}}{z_{0n}} \int_0^l p(t) \exp(2i\lambda_n t) dt \right\}, \int_0^l p(t) dt \right).$$

Since Q is C^1 by Theorem 5.7 it suffices to show that $D_p Q(0)$ is surjective. Let $\{a_n\} \in l_2$, $\alpha \in \mathbb{R}$. Write $b_n = 2i(z_{0n}/\bar{z}_{0n})a_n$. By Lemma 5.1 (iii) we can solve the equations

$$\begin{aligned} \int_0^l q(t) \exp(2i\lambda_n t) dt &= b_n, & \int_0^l q(t) \exp(-2i\lambda_n t) dt &= \bar{b}_n, & n = 1, 2, \dots, \\ \int_0^l q(t) dt &= \alpha \end{aligned}$$

for $q \in L^2([0, l]; \mathbb{C})$. Setting $p(t) = \operatorname{Re} q(t)$ we see that $D_p Q(0)$ is surjective. \square

Remark 5.9. Suppose that $\{1, e^{\pm 2i\lambda_n t}, \phi_1(t), \dots, \phi_N(t)\}$ can be extended to a Riesz basis of $L^2([0, l]; \mathbb{C})$, where $\phi_i \in L^2([0, l]; \mathbb{R})$, $1 \leq i \leq N$. Then the proof shows that we can find a $p \in L^2([0, l]; \mathbb{R})$ such that (5.21) holds, $\int_0^l p(t) dt = \theta$, and $\int_0^l p(t) \phi_i(t) dt = \theta_i$, $1 \leq i \leq N$, provided that

$$\|h\|_{l_2} + |\theta| + \sum_{i=1}^N |\theta_i|$$

is sufficiently small. Thus, the more deficient the set $\{1, e^{\pm 2i\lambda_n t}\}$ is, the more controls there are such that (5.22) holds. If $\{1, e^{\pm 2i\lambda_n t}\}$ is already a Riesz basis, then p is unique.

COROLLARY 5.10. Suppose $\{z_{0n}\} \in l_2$ with $b_n \neq 0$, $z_{0n} \neq 0$ for all $n = 1, 2, \dots$, and

$$\lim_{n \rightarrow \infty} (\lambda_{n+1} - \lambda_n) \geq \nu > 0.$$

Then given any $T > (\pi/\nu)$ there exists $\varepsilon_T > 0$ such that for any $h \in l_2$, $\theta \in \mathbb{R}$, with $\|h\|_{l_2} + |\theta| < \varepsilon_T$, there is a $p \in L^2([0, T]; \mathbb{R})$ such that

$$\frac{\lambda_n}{b_n} \left(\frac{z_n(T)}{z_{0n}} \exp \left[i \left(\lambda_n T + \frac{b_n}{2\lambda_n} \theta \right) \right] - 1 \right) = h_n, \quad n = 1, 2, \dots$$

and $\int_0^T p(t) dt = \theta$.

Furthermore, if λ_n/σ is an integer for all n and some $\sigma > 0$, then there exists an $\varepsilon > 0$ such that if $\|h\|_{l_2} + |\theta| < \varepsilon$ then there is a $p \in L^2([0, 2\pi/\sigma]; \mathbb{R})$ such that

$$\frac{z_n(2\pi/\sigma)}{z_{0n}} = \exp \left(\frac{-ib_n\theta}{2\lambda_n} \right) \left(1 + \frac{b_n h_n}{\lambda_n} \right), \quad n = 1, 2, \dots$$

and $\int_0^{2\pi/\sigma} p(t) dt = \theta$.

Proof. The first part follows immediately from Theorems 5.2, 5.8. The second part is then obvious. \square

Remarks 5.11. 1. In Corollary 5.10 there exist infinitely many families of possible controls p . This follows from the fact that by Theorem 5.2 $\{1, e^{\pm 2i\lambda_n t}\}$ can be extended to a Riesz basis of $L^2([0, A]; \mathbb{C})$ for any $\pi/\nu < A < T$, so that there are infinitely many linearly independent real functions in the orthogonal complement of the subspace of $L^2([0, T]; \mathbb{C})$ spanned by $\{1, e^{\pm 2i\lambda_n t}\}$, and Remark 5.9.

2. The set of $z = \sum_{n=1}^{\infty} z_n \phi_n \in \mathcal{H}$ such that for some $\theta \in \mathbb{R}$

$$\zeta_n = \frac{\lambda_n}{b_n} \left(\frac{z_n}{z_{0n}} \exp \left[i \left(\lambda_n T + \frac{b_n}{2\lambda_n} \theta \right) \right] - 1 \right)$$

belongs to the ball $\|\zeta\|_{l_2} < \varepsilon$ is compact (use $|z_n| = (|b_n \zeta_n / \lambda_n| + 1) |z_{0n}| \leq (C\varepsilon^{1/2} + 1) |z_{0n}|$ and Lemma 5.6). Hence the results of Theorem 5.8 and Corollary 5.10 do not say that we can control in finite time to points of a dense subset of some neighborhood of $e^{\mathcal{A}T} z(0)$ in \mathcal{H} . To prove such an approximate controllability result we would need to extend Theorem 5.8 by allowing ε_l to be arbitrarily large.

We now show how Corollary 5.10 can be applied to prove a *global* approximate controllability theorem. We restrict attention to the case when $e^{\mathcal{A}t}$ is periodic.

THEOREM 5.12. Suppose that $z_0 = \{z_{0n}\} \in l_2$ with $z_{0n} \neq 0$ for all $n = 1, 2, \dots$, and let λ_n/σ be an integer for all n and some $\sigma > 0$. Then for any $h \in l_2$ with $1 + (b_n/\lambda_n)h_n \neq 0$ for all n , and any $\theta \in \mathbb{R}$, there exist a positive integer m and a control $p \in L^2([0, 2m\pi/\sigma]; \mathbb{R})$ such that

$$\frac{z_n(2m\pi/\sigma)}{z_{0n}} = \exp \left(-\frac{ib_n\theta}{2\lambda_n} \right) \left(1 + \frac{b_n}{\lambda_n} h_n \right), \quad n = 1, 2, \dots$$

Proof. Let

$$A = \left\{ (h, \theta) \in l_2 \times \mathbb{R} \mid z_n \left(\frac{2m\pi}{\sigma} \right) = \exp \left(-\frac{b_n\theta}{2\lambda_n} \right) \left(1 + \frac{b_n}{\lambda_n} h_n \right) z_{0n} \text{ for all } n \right\},$$

some positive integer m , and some $p \in L([0, 2m\pi/\sigma]; \mathbb{R})$ and

$$B = \left\{ (h, \theta) \in l_2 \times \mathbb{R} \mid 1 + \frac{b_n}{\lambda_n} h_n \neq 0 \text{ for all } n \right\}.$$

We show that $A = B$. By the backwards uniqueness of solutions to (5.13) and the assumption $z_{0n} \neq 0$ for all n we see that $A \subset B$. It therefore suffices to show that (i) A is open, (ii) $\partial A \cap B$ is empty, and (iii) B is arcwise connected.

To prove (i), let $(h, \theta) \in A$, so that

$$z_n\left(\frac{2m\pi}{\sigma}\right) = \exp\left(-\frac{ib_n\theta}{2\lambda_n}\right)\left(1 + \frac{b_n}{\lambda_n}h_n\right)z_{0n}, \quad n = 1, 2, \dots$$

for some m and $p \in L^2([0, 2m\pi/\sigma]; \mathbb{R})$. We apply Corollary 5.10, with initial data

$$\tilde{z}_{0n} = \exp\left(-\frac{ib_n\theta}{2\lambda_n}\right)\left(1 + \frac{b_n}{\lambda_n}h_n\right)z_{0n},$$

to deduce the following assertion: if

$$(5.22) \quad \|g\|_{l_2} + |\alpha| < \varepsilon$$

then there exists $p \in L^2([0, 2(m+1)\pi/\sigma]; \mathbb{R})$ such that

$$z_n\left(\frac{2(m+1)\pi}{\sigma}\right) = \exp\left(-\frac{ib_n\alpha}{2\lambda_n}\right)\left(1 + \frac{b_n}{\lambda_n}g_n\right)\tilde{z}_{0n}, \quad n = 1, 2, \dots$$

But if $\|h - \tilde{h}\|_{l_2}$ and $|\theta - \tilde{\theta}|$ are sufficiently small then

$$g_n \equiv \frac{\tilde{h}_n - h_n}{1 + (b_n/\lambda_n)h_n} \quad \text{and} \quad \alpha \equiv \tilde{\theta} - \theta$$

satisfy (5.22) (note that $\{h_n\} \in l_2$ implies that $|1 + (b_n/\lambda_n)h_n| \geq k > 0$), and so for the corresponding p we have

$$z_n\left(\frac{2(m+1)\pi}{\sigma}\right) = \exp\left(-\frac{ib_n\tilde{\theta}}{2\lambda_n}\right)\left(1 + \frac{b_n}{\lambda_n}\tilde{h}_n\right)z_{0n}, \quad n = 1, 2, \dots$$

Thus A is open.

Suppose that $(h, \theta) \in \partial A \cap B$. We show that the time reversibility properties of (5.1) lead to a contradiction. Let

$$w_{0n} = \exp\left(-\frac{ib_n\theta}{2\lambda_n}\right)\left(1 + \frac{b_n}{\lambda_n}h_n\right)z_{0n}.$$

By Corollary 5.10, if (5.22) holds, there exists $q \in L^2([0, 2\pi/\sigma]; \mathbb{R})$ with $\int_0^{2\pi/\sigma} q(t) dt = \alpha$, such that the solution of

$$\dot{v}_n(t) = -i\lambda_n v_n(t) - iq(t)\frac{b_n}{\lambda_n} \operatorname{Re} v_n(t), \quad v_n(0) = \bar{w}_{0n},$$

satisfies

$$v_n\left(\frac{2\pi}{\sigma}\right) = \exp\left(-\frac{ib_n\alpha}{2\lambda_n}\right)\left(1 + \frac{b_n}{\lambda_n}\bar{g}_n\right)\bar{w}_{0n}, \quad n = 1, 2, \dots$$

Hence

$$\tilde{z}_n(t) \equiv \bar{v}_n\left(\frac{2\pi}{\sigma} - t\right)$$

satisfies

$$(5.23) \quad \begin{aligned} \dot{\tilde{z}}_n(t) &= -i\lambda_n \tilde{z}_n(t) - iq\left(\frac{2\pi}{\sigma} - t\right)\frac{b_n}{\lambda_n} \operatorname{Re} \tilde{z}_n(t), \\ \tilde{z}_n(0) &= \exp\left(\frac{ib_n\alpha}{2\lambda_n}\right)\left(1 + \frac{b_n}{\lambda_n}g_n\right)w_{0n}, \\ \tilde{z}_n\left(\frac{2\pi}{\sigma}\right) &= w_{0n}. \end{aligned}$$

Since $(h, \theta) \in \partial A$, there exists a sequence $(h^{(r)}, \theta^{(r)}) \in A$ with $(h^{(r)}, \theta^{(r)}) \rightarrow (h, \theta)$ in $l_2 \times \mathbb{R}$. Define

$$\alpha = \theta - \theta^{(r)} \quad \text{and} \quad g_n = \frac{h_n^{(r)} - h_n}{1 + (b_n/\lambda_n)h_n},$$

for some fixed r large enough for (5.22) to hold. For this r there exist m and $p \in L^2([0, 2m\pi/\sigma]; \mathbb{R})$ such that

$$z_n\left(\frac{2m\pi}{\sigma}\right) = \exp\left(-\frac{ib_n\theta^{(r)}}{2\lambda_n}\right)\left(1 + \frac{b_n}{\lambda_n}h_n^{(r)}\right)z_{0n} = \exp\left(\frac{ib_n\alpha}{2\lambda_n}\right)\left(1 + \frac{b_n}{\lambda_n}g_n\right)w_{0n}.$$

Extending p to be $q(2(m+1)\pi/\sigma - t)$ on $[2m\pi/\sigma, 2(m+1)\pi/\sigma]$ we see that by (5.23)

$$z_n\left(\frac{2(m+1)\pi}{\sigma}\right) = w_{0n}.$$

Hence $(h, \theta) \in A$, a contradiction. This proves (ii).

To prove (iii), note that if $(h, \theta) \in B$ then $|(b_n/\lambda_n)h_n| < 1$ for $n > N$, say. Let $h^N = (h_1, \dots, h_N, 0, \dots)$. The arc $t \mapsto (h^N + t(h - h^N), t\theta)$, $t \in [0, 1]$ connects (h, θ) to $(h^N, 0)$ and lies in B . But $(h^N, 0)$ can be connected to $(0, 0)$ by an arc in B of the form $(s, 0)$ where $s \in \mathbb{R}^N$ and runs from h^N to 0 and avoids $(-\lambda_1/b_1, -\lambda_2/b_2, \dots, -\lambda_N/b_N)$. Thus B is arcwise connected. \square

COROLLARY 5.13. *Let the hypotheses of Theorem 5.12 hold. Then the attainable set*

$$s(z_0) = \bigcup_{\substack{t \geq 0 \\ 0 \in L^2_{loc}([0, \infty); \mathbb{R})}} z(t; p, z_0)$$

is dense in \mathcal{H} .

Proof. The set $\{h \in l_2 \mid 1 + (b_n/\lambda_n)h_n \neq 0 \text{ for all } n\}$ is dense in l_2 . \square

Remark 5.14. Clearly the information provided by Theorem 5.12 implies global controllability with respect to suitable finite-dimensional observers. We leave the precise formulation of these results to the reader.

6. Applications to partial differential equations.

Example 1. *Wave equation with Dirichlet boundary conditions.* Consider the wave equation

$$u_{tt} - u_{xx} + p(t)u = 0, \quad 0 < x < 1,$$

with boundary conditions

$$u = 0 \quad \text{at } x = 0, 1$$

and initial conditions

$$u(x, 0) = u_0(x), \quad u_t(x, 0) = u_1(x), \quad 0 < x < 1.$$

In the notation of (5.1), (5.2) we have

$$A = -\frac{d^2}{dx^2}, \quad B = I, \quad H = L^2(0, 1) = L^2([0, 1]; \mathbb{R}),$$

$$D(A) = H^2(0, 1) \cap H_0^1(0, 1), \quad D(A^{1/2}) = H_0^1(0, 1),$$

$$\lambda_n = n\pi, \quad \phi_n = \sqrt{2} \sin n\pi x, \quad n = 1, 2, \dots,$$

$$(B\phi_n, \phi_m) = \delta_{mn}.$$

We thus see that (D1) holds, and since $b_n = 1$ we have $b_n/\lambda_n = 1/n\pi$ so that (D2) also holds.

As before, we set

$$z(t) = A^{1/2}u(t) + i\dot{u}(t) \quad \text{and} \quad z_0 = A^{1/2}u_0 + iu_1,$$

so that

$$z_{0n} = \lambda_n(u_0, \phi_n) + i(u_1, \phi_n).$$

In this case $\mathcal{H} = L^2(0, 1) \oplus iL^2(0, 1)$. We suppose that $z_0 \in \mathcal{H}$. We note that $\{1, e^{\pm 2i\lambda_n t}\}$ forms a Riesz basis of $L^2([0, 1]; \mathbb{C})$ and can be extended to a Riesz basis of $L^2([0, l]; \mathbb{C})$ for any $l \geq 1$. Then Theorem 5.3, Corollary 5.4, Theorem 5.7 and Theorem 5.8 are all applicable. For example, Theorem 5.3 says that if $z_{0n} \neq 0$ for all n we can control any finite-dimensional projection of the solution to take any value sufficiently close to the projection of the free solution ($p \equiv 0$) at time $T_1 \geq 1$, while Theorem 5.8 holds for any $l \geq 1$.

In particular, Theorem 5.5 shows that the set of $\{u, u_t\}$ in $H_0^1(0, 1) \times L^2(0, 1)$ accessible from $\{u_0, u_1\}$ with controls in $L_{\text{loc}}^r[0, \infty)$, $r \geq 1$, given by

$$S(\{u_0, u_1\}) = \bigcup_{\substack{r \geq 0 \\ p \in L_{\text{loc}}^r([0, \infty); \mathbb{R})}} \{u(t; p, u_0, u_1), u_t(t; p, u_0, u_1)\}$$

has dense complement in $H_0^1(0, 1) \times L^2(0, 1)$. On the other hand, by Theorem 5.12 and Corollary 5.13 we have global approximate controllability: thus the set S of states that can be reached using L^2 controls on a time interval of length at least one is dense in $H_0^1 \times L^2$, provided $z_{0n} \neq 0$, i.e., all modes of the initial data are active.

Example 2. Wave equation with mixed boundary conditions. Consider the wave equation

$$u_{tt} - u_{xx} + p(t)u = 0, \quad 0 < x < 1,$$

with boundary conditions

$$u = 0 \text{ at } x = 0, \quad u - \alpha u_x = 0 \text{ at } x = 1, \quad \alpha > 0 \text{ constant},$$

and initial conditions

$$u(x, 0) = u_0(x), \quad u_t(x, 0) = u_1(x), \quad 0 < x < 1.$$

In the notation of (5.1) and (5.2) we have

$$A = -\frac{d^2}{dx^2}, \quad B = I, \quad H = L^2(0, 1),$$

$$D(A) = \{u \in H^2(0, 1) \mid u = 0 \text{ at } x = 0, u + \alpha u_x = 0 \text{ at } x = 1\},$$

$$D(A^{1/2}) = \{u \in H^1(0, 1) \mid u = 0 \text{ at } x = 0\},$$

$$\tan \lambda_n + \alpha \lambda_n = 0, \quad \phi_n(x) = (\sin \lambda_n x) / \left(\int_0^1 \sin^2 \lambda_n x \right)^{1/2}, \quad n = 1, 2, \dots,$$

and $(B\phi_m, \phi_n) = \delta_{mn}$.

In this case,

$$\lambda_n = \frac{n\pi}{2} + \varepsilon_n(\alpha), \quad n = 1, 2, \dots,$$

where $|\varepsilon_n(\alpha)| \rightarrow 0$ as $n + \alpha \rightarrow \infty$. Thus, since $b_n = 1$, $\{b_n/\lambda_n\} \in l_2$. Hence (D1) and (D2) hold.

As usual, we set

$$z(t) = A^{1/2}u(t) + i\dot{u}(t), \quad z_0 = A^{1/2}u_0 + iu_1,$$

so that

$$z_{0n} = \lambda_n(u_0, \phi_n) + i(u_1, \phi_n).$$

As in Example 1, $\mathcal{H} = L^2(0, 1) \oplus iL^2(0, 1)$, and we let $z_0 \in \mathcal{H}$. Theorem 5.3, Corollary 5.4, Theorem 5.7, Theorem 5.8 and the first part of Corollary 5.10 are all applicable. By Theorem 5.2 $\{e^{\pm 2i\lambda_n t}\}$ can be extended to a Riesz basis of $L^2([0, T]; \mathbb{C})$ for any $T > 2$, so that in the above results the assertions of finite-dimensional or approximate controllability apply to time intervals of length greater than 2. Actually, for α sufficiently large we can take $T_1 \geq 2$ in Theorem 5.3 and $T \geq 2$ in Corollary 5.10. (This is because $\sup_n |\varepsilon_n(\alpha)| = |\varepsilon_1(\alpha)| < \frac{1}{2} \log 2$ for α sufficiently large, so that

$$\sup_n |2\lambda_n - n\pi| < \log 2,$$

which implies by Riesz and Nagy [1955, p. 209] that $\{1, e^{\pm 2i\lambda_n t}\}$ forms a Riesz basis of $L^2(0, 2)$.)

Example 3. Rod equation with hinged ends. Consider the system

$$(6.1) \quad u_{tt} + u_{xxxx} + p(t)u_{xx} = 0, \quad 0 < x < 1,$$

with boundary conditions

$$(6.2) \quad u = u_{xx} = 0 \quad \text{at } x = 0, 1$$

and initial conditions

$$(6.3) \quad u(x, 0) = u_0(x), \quad u_t(x, 0) = u_1(x).$$

In the notation of (5.1), (5.2) we set

$$A = \frac{d^4}{dx^4}, \quad B = \frac{d^2}{dx^2}, \quad H = L^2(0, 1),$$

$$D(A) = \{u \in H^4(0, 1) \mid u, u_{xx} \in H_0^1(0, 1)\},$$

$$D(A^{1/2}) = H^2(0, 1) \cap H_0^1(0, 1), \quad \lambda_n = n^2 \pi^2,$$

$$\phi_n = \sqrt{2} \sin n\pi x, \quad n = 1, 2, \dots,$$

$$(B\phi_m, \phi_n) = 0, \quad n \neq m, \quad (B\phi_n, \phi_n) = -n^2 \pi^2.$$

In this case $b_n/\lambda_n = -1$, so that (D1), (D2) are again satisfied. As usual we write $z(t) = A^{1/2}u(t) + i\dot{u}(t) = \sum_{n=1}^{\infty} z_n(t)\phi_n$, $z_n(0) = z_{0n}$. Note that

$$(6.4) \quad \lim_{n \rightarrow \infty} (\lambda_{n+1} - \lambda_n) = \infty.$$

Theorem 5.2 therefore implies that $\{1, e^{\pm 2i\lambda_n t}\}$ can be extended to a Riesz basis of $L^2([0, T]; \mathbb{C})$ for any $T > 0$. Theorem 5.3 is therefore applicable with any $T_1 > 0$, Corollary 5.4 holds, Theorems 5.7 and 5.8 hold for any $l > 0$, both conclusions of Corollary 5.10 are valid, and Theorem 5.12 and Corollary 5.13 hold. We summarize the approximate controllability results in the following theorem.

THEOREM 6.1. *Let $u_0 \in H^2(0, 1) \cap H_0^1(0, 1)$, $u_1 \in L^2(0, 1)$ and suppose that*

$$z_{0n} \equiv n^2 \pi^2 (u_0, \phi_n) + i(u_1, \phi_n) \neq 0 \quad \text{for all } n = 1, 2, \dots.$$

For any $p \in L^1_{\text{loc}}([0, \infty); \mathbb{R})$ a unique mild solution

$$\{u, \dot{u}\} \in C([0, \infty); X)$$

of (6.1)–(6.3) exists, where $X = (H^2(0, 1) \cap H^1_0(0, 1)) \times L^2(0, 1)$, and if $p \in L^2_{\text{loc}}([0, \infty); \mathbb{R})$ then

$$\left\{ \frac{z_n(t)}{z_{0n}} \exp \left[i \left(\lambda_n t - \frac{1}{2} \int_0^t p(s) ds \right) \right] - 1 \right\} \in C([0, \infty); l_2).$$

Conversely, for any $T > 0$ there exists $\varepsilon_T > 0$ such that if $\|h\|_{l_2} + |\alpha| < \varepsilon_T$ then

$$\frac{z_n(T)}{z_{0n}} e^{i(\lambda_n T - \alpha)} - 1 = h_n, \quad n = 1, 2, \dots$$

for infinitely many $p \in L^2_{\text{loc}}([0, T]; \mathbb{R})$ with $\int_0^T p(t) dt = 2\alpha$. In particular, setting $T = 2/\pi$, there exists $\varepsilon > 0$ such that if $\|h\|_{l_2} + |\alpha| < \varepsilon$ then

$$z_n\left(\frac{2}{\pi}\right) = e^{i\alpha} (1 + h_n) z_{0n}, \quad n = 1, 2, \dots$$

for infinitely many $p \in L^2([0, 2/\pi]; \mathbb{R})$ with $\int_0^{2/\pi} p(t) dt = 2\alpha$. Furthermore, if $(h, \alpha) \in l_2 \times \mathbb{R}$ with $h_n \neq -1$ for all n , there exist a positive integer m and a control $p \in L^2([0, 2m/\pi]; \mathbb{R})$ such that

$$(6.5) \quad z_n\left(\frac{2m}{\pi}\right) = e^{i\alpha} (1 + h_n) z_{0n}, \quad n = 1, 2, \dots,$$

so that the set of states accessible from $\{u_0, u_1\}$ is dense in X .

Remark 6.2. Our method of proof shows that given $\varepsilon > 0$ we can find m and p such that (6.5) holds and $\|p\|_{L^2(I; \mathbb{R})} < \varepsilon$ for any interval $I \subset [0, 2m/\pi]$ of length 1. Of course m will need to be large if ε is small.

Example 4. Rod equation with clamped ends. Consider (6.1) with boundary conditions

$$u = u_x = 0 \quad \text{at } x = 0, 1$$

and initial conditions (6.3). As is well known, this case is much more delicate than (6.1) with hinged boundary conditions (6.2). We now have

$$\begin{aligned} A &= \frac{d^4}{dx^4}, \quad B = \frac{d^2}{dx^2}, \quad H = L^2(0, 1), \\ D(A) &= H^4(0, 1) \cap H^2_0(0, 1), \quad D(A^{1/2}) = H^2_0(0, 1), \\ \cosh \lambda_n^{1/2} \cos \lambda_n^{1/2} &= 1, \quad n = 1, 2, \dots \end{aligned}$$

The usual graphical analysis shows that

$$\lambda_n = (n - \frac{1}{2})^2 \pi^2 + \varepsilon_n,$$

where $\varepsilon_n \rightarrow 0$ as $n \rightarrow \infty$. (Very precise estimates for ε_n are given in Ball and Slemrod [1979].) The corresponding orthonormal eigenfunctions ϕ_n do not satisfy $(B\phi_m, \phi_n) = 0$, $m \neq n$, and so none of the results in § 5.3 are applicable. Furthermore, hypothesis (ii) of Theorem 5.3 does not hold, since $2\lambda_n - (\lambda_p + \lambda_q)$ can be arbitrarily small for arbitrarily large n, p and q (cf. Ball and Slemrod [1979], especially pp. 560, 574). So it is not obvious that (6.1), (6.6) is controllable locally with respect to finite-dimensional observers. It is possible that estimates on the lines of those in the preceding reference

for the λ_n might establish local controllability relative to G of the form

$$G\left(\sum_{n=1}^{\infty} a_n z_n\right) = L\left(\sum_{n=1}^N a_n z_n\right).$$

The only results in this paper applicable to (6.1), (6.6) are the basic existence theorem, Theorem 2.5, which just gives the standard result that for $\{u_0, u_1\} \in D(A^{1/2}) \times H = X$ there exists for each $p \in L^1_{\text{loc}}([0, \infty); \mathbb{R})$ a unique mild solution with initial data $\{u_0, u_1\}$, and Theorem 3.6, which demonstrates the general impossibility of exact controllability using controls $p \in L^r_{\text{loc}}([0, \infty); \mathbb{R})$, $r > 1$.

REFERENCES

- A. V. BALAKRISHNAN [1976], *Applied Functional Analysis*, Applications of Mathematics 3, Springer, New York.
- J. M. BALL [1977], *Strongly continuous semigroups, weak solutions and the variation of constants formula*, Proc. AMS, 63, pp. 370–373.
- J. M. BALL AND M. SLEMROD [1979], *Nonharmonic Fourier series and the stabilization of distributed semi-linear control systems*, Comm. Pure Appl. Math., 32, pp. 555–587.
- R. BROCKETT [1972], *System theory on group manifolds and coset spaces*, SIAM J. Control, 10, pp. 265–284.
- R. W. CARROLL [1969], *Abstract Methods in Partial Differential Equations*, Harper and Row, New York.
- P. CHERNOFF AND J. MARSDEN [1974], *Some Properties of Infinite Dimensional Hamiltonian Systems*, Lecture Notes in Mathematics 425, Springer-Verlag, New York.
- N. DUNFORD AND J. SCHWARTZ [1964], *Linear Operators I*, Wiley-Interscience, New York.
- I. C. GOHBERG AND M. G. KREIN [1969], *Introduction to the theory of linear non-self-adjoint operators*, AMS Transl., vol. 18, American Mathematical Society, Providence, RI.
- J. K. HALE [1969], *Ordinary Differential Equations*, John Wiley, New York.
- H. HERMES [1974], *On local and global controllability*, SIAM J. Control, 12, pp. 252–261.
- H. HERMES [1979], *Local controllability of observables in finite and infinite dimensional nonlinear control systems*, Appl. Math. Optim., 5, pp. 117–125.
- V. JURDEVIC AND J. QUINN [1978], *Controllability and stability*, J. Diff. Eqs., 28, pp. 281–289.
- C. LOBRY [1970], *Contrôlabilité des systèmes non linéaires*, SIAM J. Control, 8, pp. 450–460.
- D. G. LUENBERGER [1969], *Optimization by Vector Space Methods*, John Wiley, New York.
- A. PAZY [1974], *Semi-groups of linear operators and applications to partial differential equations*, Lect. Notes 10, Univ. of Maryland, College Park.
- F. RIESZ AND B. SZ. NAGY [1955], *Functional Analysis*, F. Ungar, New York.
- D. RUSSELL [1967], *Nonharmonic fourier series in the control theory of distributed parameter systems*, J. Math. Anal. Appl., 18, pp. 542–560.
- I. SEGAL [1963], *Nonlinear Semigroups*, Ann. Math., 78, pp. 334–362.
- H. TANABE [1979], *Equations of Evolution*, Pitman, London.

ADDENDUM: THE CIRCLE CRITERION AND THE L^p STABILITY OF FEEDBACK SYSTEMS*

S. MOSSAHEB†

Abstract. In an earlier paper [SIAM J. Control. Optim., 20 (1982), pp. 144–152] it was shown that under certain mild conditions the circle theorem of Sandberg and Zames implies the L^p stability of a broad class of scalar feedback systems for all $p \geq 1$. In this addendum it is shown how such stability theorems can be extended to multivariable systems.

In our earlier paper [3] it has been shown that under certain conditions the celebrated circle theorem of Sandberg and Zames implies L^p stability for all $1 \leq p \leq \infty$ of the solution x of the scalar nonlinear equation:

$$(1) \quad x(t) = f(t) - \int_0^t g(t-s)n(s, x(s)) \, ds.$$

The said conditions are that $(1+t)g(t) \in L^1 \cap L^2$, that for some constants a and b , the nonlinearity $n(t, x)$ is in the sector (a, b) and the hypotheses of the circle theorem hold. It has also been shown that under the same conditions if $p \geq 2$ and $\lim_{t \rightarrow \infty} f(t) = 0$ then $\lim_{t \rightarrow \infty} x(t) = 0$. The idea in [3] was to reduce (1) to the form in which $n(s, x(s)) = k(s)x(s)$ so that in a sense (1) is linearized and then to obtain the final results by examining the properties of the resolvent of $g(t-s)k(s)$. Finally, by using the so-called loop-transformation technique of Zames, the theorems were extended to cover certain unstable forms of g .

The aim of this note is to point out that all the arguments of [3] carry over to vector equations of the form (1). The only perhaps nontrivial modifications are the linearization process of writing $n(s, x(s)) = k(s)x(s)$ for a suitable matrix k and of treating the case of unstable g . Having shown these two facts, we briefly indicate how to use them to deal with vector equations.

Throughout this note r will be a fixed positive integer and $|x|$ is the usual l^2 -norm of the vector x . For any square matrix A we write $\Lambda(A)$ for the square root of the largest eigenvalue of A^*A where A^* is the conjugate transpose of A . Finally, for any p , $1 \leq p \leq \infty$ the L^p_r norm of an r -vector of p -summable functions is

$$\|f\|_p = \left(\sum_{i=1}^r \|f_i\|_p^2 \right)^{1/2}.$$

The nonlinearities which will be considered satisfy a conicity condition of the form (2) below. For more details on the conicity of operators see [5]. For $a > 0$ and any $r \times r$ constant matrix L , let $N(L, a)$ be the set of functions $n(t, x)$ defined and measurable on $\mathbb{R}_+ \times \mathbb{R}^r$ such that $n(t, x)$ is continuous in x for almost all t and

$$(2) \quad |n(t, x) - Lx| \leq a|x| \quad \text{for all } x \in \mathbb{R}^r \text{ and almost all } t.$$

LEMMA 1. *Let $n \in N(L, a)$ and let $x(t)$ be an r -vector of measurable functions on $(0, \infty)$. Then there exists an $r \times r$ matrix $k(t)$ of measurable functions such that $n(t, x(t)) = k(t)x(t)$ and $\Lambda(k(t) - L) \leq a$, for almost all t .*

* This Journal, 20 (1982), pp. 144–152. Received by the editors August 5, 1981.

† Postgraduate School of Studies in Control Engineering, University of Bradford, Bradford, West Yorkshire, BD7 1DP, United Kingdom.

Proof. For any $y \in \mathbb{R}^r \setminus \{0\}$, put $N(t, y) = n(t, y) - Ly$. Let N_i be the i th element of N , and let $m_{ij} = y_i N(t, y) / |y|^2$. Let $M(t, y)$ be the $r \times r$ matrix whose ij th element is m_{ij} . Clearly, M is continuous on $\mathbb{R}^r \setminus \{0\}$ for almost all t and $N(t, y) = M(t, y)y$. Observe that $M(t, y) = N(t, y)z$, where z is the row vector whose i th element is $y_i / |y|^2$. Thus, $M^*M = |N(t, y)|^2 z^*z$. Since M^*M has unit rank, $r-1$ of its eigenvalues are zero and by considering its trace we have $\Lambda^2(M) = |N(t, y)|^2 / |y|^2$, which by assumption implies $\Lambda(M) \leq a$. Now let T_1 be the set of points at which x vanishes and let T_2 be its complement. Let k_0 be any $r \times r$ matrix with $\Lambda(k_0) \leq a$. Put $k(t) = M(t, x(t)) + L$ on T_2 and $k(t) = k_0$ on T_1 . Clearly, $k(t)$ satisfies the equality and inequality of the lemma and its measurability follows from that of $x(t)$ and the continuity of $M(t, y)$ for $y \neq 0$ and almost all $t > 0$.

In dealing with unstable cases of g , we shall need the following. Write $L(1+t)$ for the Banach algebra under convolution of the functions g on $(0, \infty)$ such that $(1+t)g(t) \in L^1$ with norm $\|g\| = \|(1+t)g\|_1$. Let $A(1+t)$ be the algebra obtained from $L(1+t)$ by joining a unit to it. By [1, p. 141] an element g of $A(1+t)$ is invertible if and only if $\inf\{|\hat{g}(s)|: \operatorname{Re} s \geq 0\} > 0$, where \hat{g} is the Laplace transform of g . Let A_1 be the subset of $A(1+t)$ consisting of those g such that $(1+t)g(t) \in L^1 \cap L^2$. From

$$(1+t)(f*g) = \int_0^t (1+t-s)f(t-s)g(s) ds + \int_0^t f(t-s)(sg(s)) ds$$

it follows that A_1 is an algebra and also an ideal of $A(1+t)$ under convolution. Let A'_1 and $A'(1+t)$ be the set of $r \times r$ matrices of functions whose elements are in A_1 and $A(1+t)$, respectively. Finally, let A'_2 be the set of $r \times r$ matrices of functions which can be written as $g_1 + g_2$, where $g_1 \in A'_1$ and the Laplace transform of g_2 is a strictly proper rational function.

DEFINITION.

(i) Two elements N and D of A'_1 are said to be *coprime* if there exist P and Q in A'_1 such that $\hat{N}(s)\hat{P}(s) + \hat{D}(s)\hat{Q}(s) = I$.

(ii) An $r \times r$ matrix g of Laplace transformable functions is said to have a *coprime factorization* (N, D) if N and D are two coprime elements of A'_1 and $\hat{g}(s) = (\hat{D}(s))^{-1}\hat{N}(s)$.

We shall only need the elementary properties of coprime factorizations listed below. The general theory of such factorizations has been used extensively in multivariable systems by several authors, notably F. M. Callier, C. A. Desoer and M. Vidyasagar. For more details the reader is referred to the works of these authors.

(I) Suppose $g \in A'_2$ and let $g = g_1 + g_2$, where $g_1 \in A'_1$ and $\hat{g}_2(s)$ is a matrix of strictly proper rational functions. It follows from [4, Th.1] that g_2 has a coprime factorization in A'_1 given by, say, (N, D) . If $\hat{N}\hat{P} + \hat{D}\hat{Q} = I$ for some P and Q in A'_1 , then putting $\hat{N}_1 = \hat{N} + \hat{D}\hat{g}_1$ and $\hat{Q}_1 = \hat{Q} - \hat{g}_1$ we have $\hat{g}(s) = (\hat{D}(s))^{-1}\hat{N}_1(s)$ and $\hat{N}_1\hat{P} + \hat{D}\hat{Q}_1 = I$. Since A'_1 is closed under convolution, it follows that every element of A'_2 has a coprime factorization in A'_1 .

(II) Let (N, D) be a coprime factorization of $g \in A'_2$, and let L be a constant $r \times r$ matrix. Assume that

$$(3) \quad \inf\{|\det(\hat{D}(s) + \hat{N}(s)L)|: \operatorname{Re} s \geq 0\} > 0.$$

From the characterization of the spectrum of $A(1+t)$, it follows that $D + NL$ has an inverse in $A'(1+t)$. From $(I + \hat{g}L)^{-1} = (\hat{D} + \hat{N}L)^{-1}\hat{D}$ and $(I + \hat{g}L)^{-1}\hat{g} = (\hat{D} + \hat{N}L)^{-1}\hat{N}$ and since A'_1 is an ideal in $A'(1+t)$, it follows that $(I + \hat{g}L)^{-1}$ and $(I + \hat{g}L)^{-1}\hat{g}$ are the Laplace transforms of two elements of A'_1 . Since the elements of \hat{g} , \hat{N} and \hat{D} are

meromorphic and tend to zero as $\operatorname{Re} s \rightarrow 0$ in the closed right-half plane, it is easily seen that (3) is equivalent to the following two conditions:

$$(4) \quad \inf \{ |\det(I + \hat{g}L)| : \operatorname{Re} s \geq 0 \} > 0,$$

$$(5) \quad \det(\hat{D}(s) + \hat{N}(s)L) \neq 0 \quad \text{whenever } \operatorname{Re} s \geq 0 \text{ and } \det \hat{D}(s) = 0.$$

The above conditions are the usual conditions for the stability of linear systems under constant feedback. The reason for having to consider coprime factorizations is to take account of possible polezero cancellations in $(I + \hat{g}L)^{-1}\hat{g}$.

The generalization of the main results of [3] is as follows.

THEOREM. In (1) assume that $f \in L_r^p$ with $1 \leq p \leq \infty$ and $n \in N(L, a)$ for some $a > 0$ and some constant $r \times r$ matrix L . Assume further that either (i) $g \in A_1'$, (4) holds and, moreover,

$$(6) \quad a \sup \{ \Lambda((I + \hat{g}(iw)L)^{-1}\hat{g}(iw)) : w \in \mathbb{R} \} < 1,$$

or (ii) $g \in A_2'$, $L \neq 0$, (4) and (6) holds, and moreover, for some coprime factorization (N, D) of g , (5) holds. Then any measurable solution x of (1) is in L_r^p , and for some constant c independent of x and f , we have $\|x\|_p \leq c\|f\|_p$. Moreover, if $p \geq 2$ and $\lim_{t \rightarrow \infty} f(t) = 0$, then $\lim_{t \rightarrow \infty} x(t) = 0$.

Outline of proof. Let x be any measurable solution of (1). Rewrite the equation as

$$x(t) + \int_0^t g(t-s)Lx(s) ds = f(t) - \int_0^t g(t-s)(n(s, x(s)) - Lx(s)) ds.$$

Let $k(s)$ be as in Lemma 1 and put $k_1(s) = k(s) - L$. Then $\Lambda(k_1(s)) \leq a$ for almost all s . By the assumptions on g and paragraph II above, the inverse g_1 of $I + gL$ exists in A_1' and $g_1 * g = h$ is also in A_1' . Since $g_1 * f \in L_r^p$, it follows that x satisfies the equation

$$x(t) = f_1(t) - \int_0^t h(t-s)k_1(s)x(s) ds$$

with $f_1 \in L_r^p$ and $\|f_1\|_p \leq d\|f\|_p$, where d is independent of f and x . Put $a(t, s) = -h(t-s)k_1(s)$, and let $r(t, s)$ be the $r \times r$ resolvent of $a(t, s)$ [2]. Thus, $r(t, s)$ is the solution of

$$(7) \quad r(t, s) = -a(t, s) + \int_0^t r(t, u)a(u, s) du.$$

It may be proved that r also satisfies

$$(8) \quad r(t, s) = -a(t, s) + \int_0^t a(t, u)r(u, s) du.$$

Premultiplying (7) and postmultiplying (8) by arbitrary row and column vectors α and β respectively, and using the method of the proof of [3, Thm. 1] it can be shown that

$$(9) \quad \sup_t \int_0^\infty |\alpha r(t, s)| ds < \infty, \quad \sup_s \int_0^\infty |r(t, s)\beta| dt < \infty.$$

The only procedural difference is to use the multivariable version of the circle theorem that if in an equation of the form

$$\delta(t) = \eta(t) - \int_0^t \phi(t-s)\theta(s)\delta(s) ds,$$

$\Lambda(\theta(s)) \leq a$ for almost all s , ϕ is a matrix of integrable function and a $\sup \{\Lambda(\hat{\phi}(iw)) : w \in \mathbb{R}\} < 1$. Then $\|\delta\|_2 \leq C\|\eta\|_2$ for some C is independent of δ and η .

Since in (9) α and β are arbitrary, it follows that for any Euclidean matrix norm $|\cdot|$

$$\sup_t \int_0^\infty |r(t, s)| ds < \infty \quad \text{and} \quad \sup_s \int_0^\infty |r(t, s)| dt < \infty.$$

The remaining assertions of the theorem follow by trivial modifications of the proofs of [3].

REFERENCES

- [1] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semi-Groups*, AMS Colloquium Publications, XXXI, American Mathematical Society, Providence, RI, 1957.
- [2] R. K. MILLER, *Non-linear Volterra Integral Equations*, W. A. Benjamin, New York, 1971.
- [3] S. MOSSAHEB, *The circle criterion and the L^p stability of feedback systems*, this Journal, 20 (1982), pp. 144–152.
- [4] M. VIDYASAGAR, *On the use of right-coprime factorizations in distributed feedback systems containing unstable sub-systems*, IEEE Trans., Circuits and Systems. CAS, 25 (1978), pp. 916–921.
- [5] G. ZAMES, *On the input-output stability of non-linear time-varying feedback systems, Parts I and II*, IEEE Trans. Auto. Control, AC-11 (1966), pp. 228–238, pp. 465–477.

ERRATUM: ON POLE ASSIGNMENT FOR A CLASS OF INFINITE DIMENSIONAL SYSTEMS*

AVRAHAM FEINTUCH† AND MOSHE ROSENFELD†

In this note we would like to point out two errors that appeared in the paper. These were pointed out by L. Pandolfi and the authors of [1] who corrected and generalized the main results.

The first error concerned an incorrect quote of a result from [2] which we stated as Theorem B.

The correct statement is

THEOREM B*. *If A is a discrete operator and $\operatorname{Re} \lambda \leq \rho < 0$ for $\lambda \in \sigma(A)$, then the semigroup generated by A is exponentially stable.*

This does not affect the correctness of the other results in the paper.

However, the proof of Theorem 4 as given is incorrect. The c chosen with finitely many nonzero coordinates will not shift all the poles of the new operator but only finitely many. In fact, using a result of [1] we have the following:

THEOREM 4*. *Suppose A is a discrete normal operator with $\{A, b\}$ controllable and $\sigma(A) = \{\lambda_n\}$. Then $\{A + b \otimes c, b\}$ has no common poles with $\{A, b\}$ if and only if $\{A, c\}$ is controllable.*

Proof. Suppose $\{\phi_n\}$ is the orthonormal basis of eigenvectors of A and $b_n = (b, \phi_n)$. It was shown in [1] that $\sigma(A) \cap \sigma(A + b \otimes c) = \emptyset$ if and only if $b_n(c, \phi_n) \neq 0$ for all n . By controllability of $\{A, b\}$, $b_n \neq 0$ for all n . The condition $(c, \phi_n) \neq 0$ for all n is equivalent to $\{A, c\}$ being controllable.

Thus any c all of whose coordinates are nonzero will shift all the finite poles of $\{A, b\}$, and the poles of the new system (as was seen in Theorem 3) are exactly the one-points of the function

$$\rho(z) = \sum_{k=1}^{\infty} \frac{b_k \bar{c}_k}{z - \lambda_k}.$$

The authors of [1] generalize some of the results of our paper to the case where A is not assumed to be normal.

REFERENCES

- [1] T. HAMATSUKE, A. MO'OMEN AND H. AKASHI, *On pole assignment and stabilization for a class of infinite dimensional linear systems*, to appear.
- [2] R. TRIGGIANI, *On the stabilizability problem in Banach space*, J. Math. Anal. Applic., 52 (1975), pp. 383-403.

* This Journal, 16 (1978), pp. 270-276.

† Department of Mathematics, Ben Gurion University of the Negev, BeerSheva 84 120, Israel.

IDENTIFICATION OF NONSTATIONARY DIFFUSION MODEL BY THE METHOD OF SIEVES*

HUNG T. NGUYEN† AND TUAN D. PHAM‡

Abstract. In this paper, we apply Grenander's method of sieves to the problem of estimation of the infinite dimensional parameter in a nonstationary linear diffusion model. We use an increasing sequence of finite dimensional subspaces of the parameter space as the natural sieves on which we maximize the likelihood function. We show that if the dimension of the sieves tends to infinity with the sample size with a rate not too fast then the sequence of restricted maximum likelihood estimators for the parameter is consistent and asymptotically normal.

1. Introduction. Statistical inference for stochastic processes, especially for diffusion processes suitable for describing stochastic dynamical systems corrupted by white noise, has received a good deal of attention recently. Primarily, works in this field are intended to provide new identification techniques to deal with more complex situations in control problems.

In the finite dimensional parameters case, as in detection and estimation in white Gaussian noise, the mathematical groundwork for the approach was developed by Grenander [7]; see, also Van Trees [16]. This paper discusses a model identification technique, based on Grenander's method of sieves, Grenander [8], for a class of systems represented as nonstationary linear diffusion processes in which the parameter to be estimated from data lies in an infinite dimensional space. This sieve-based technique led to the estimation of the whole unknown function at once and not just its values at a fixed point. Note that nonparametric identification for diffusion processes, using kernel estimators which estimate the drift function at each fixed point, is studied in Banon and Nguyen [1] and the implementation and the simulation studies will be presented in a forthcoming paper.

In the case of finite dimensional parameters, the maximum likelihood method is commonly used, (e.g., Feigin [5], Pham [13]). In the case of infinite dimensional parameters, the kernel method has been successful for pointwise density estimation and pointwise estimation of the drift coefficient of a stationary diffusion process (e.g., Nguyen [11], Nguyen and Pham [12], Pham [14]). It is known that the maximum likelihood method has difficulties in the infinite dimensional case since the maximum likelihood solution is generally either not attained or is not consistent. Recently, Grenander [8] has developed a new technique known as method of sieves to handle these difficulties. In this method, for each sample size a sieve which roughly speaking is a suitable subset of the parameter space, is chosen. The likelihood function is then maximised on the sieves yielding a sequence of estimators. The crucial point is the choice of appropriate sieves. Some general results on the existence of sieves leading to estimators with interesting asymptotic properties are given in Geman and Hwang [6] in which the sieves are chosen to be compact sets satisfying rather complicated conditions. The main purpose of this work is to show that in many situations where the parameter lies in a Hilbert space, a sequence of sieves can be chosen simply as an increasing sequence of finite dimensional subspaces of the parameter space. The restricted maximum likelihood estimation on the sieves is shown to be consistent and

* Received by the editors August 8, 1980 and in revised form July 15, 1981.

† Department of Mathematics and Statistics, University of Massachusetts, Amherst, Massachusetts 01003. Currently at: Department of Mathematical Sciences, New Mexico State University, Las Cruces, New Mexico 88003.

‡ Department of Mathematics, University of Indiana, Bloomington, Indiana 47405.

asymptotically normal provided that the dimension of the sieves grows not too fast with respect to the sample size. To obtain these large sample properties of our estimators, we introduce a device similar to the technique of projections (or of orthogonal functions) used by Bosq [2] in density estimation.

The model considered in this paper is the nonstationary linear diffusion model

$$dX(t) = \theta(t)X(t) dt + dW(t), \quad t \geq 0,$$

where $W(t)$, $t \geq 0$, is a Brownian noise and $\theta(\cdot)$ is the unknown function to be estimated on the entire interval $[0, T]$ based on the observation of a sample of n independent trajectories of $X(t)$ on $[0, T]$. This model has been used in engineering. Note that Delebecque and Quadrat [3] have considered the more general nonlinear diffusion model, but only the case of finite dimensional parameters is in fact investigated. Note also that the simpler model

$$dX(t) = \theta(t) dt + dW(t)$$

has been widely studied in the literature (Grenander [8], Ibragimov and Rozanov [9], Liptser and Shiryaev [10]). The last four authors considered the case where $\theta(t)$ is of the form $\theta_1 h_1(t) + \theta_2 h_2(t) + \dots + \theta_k h_k(t)$, where the h_i are known functions and the θ_i are unknown parameters to be estimated. In this case the estimation problem can be solved by regressions analysis. On the other hand, Grenander [8] has considered the estimation of the unknown function θ entirely based on a sample of n observed trajectories of $X(t)$ using precisely his method of sieves. However, he considered a different sieve than ours, and it is not hard to see that our technique, using a simpler sieve, can also be adapted to this model.

2. The model and the estimation of parameter. The considered model is the nonstationary linear diffusion model

$$(2.1) \quad dX(t) = \theta(t)X(t) dt + dW(t), \quad X(0) = x_0,$$

where x_0 is deterministic, $W(t)$ is a Brownian motion with $E\{dW(t)\}^2 = \sigma^2 dt$ and where $\theta(\cdot) \in L^2([0, T], dt)$, $[0, T]$ being the interval of observation of the process. The last condition is needed to ensure the absolute continuity of P_θ with respect to P_0 , where P_θ and P_0 are the probability distributions of $\{X(t), t \in [0, T]\}$ when $\theta(\cdot)$ is the true parameter and when $\theta = 0$, respectively. Indeed, a sufficient condition for this is Liptser and Shiryaev [10, Vol. I, Thm. 7.2, p. 240]

$$P_\theta \left\{ \int_0^T \theta^2(t) X^2(t) dt < +\infty \right\} = 1.$$

This condition will be satisfied if

$$E_\theta \left\{ \int_0^T \theta^2(t) X^2(t) dt \right\} = \int_0^T \theta^2(t) EX^2(t) dt < +\infty.$$

Therefore, if $\theta(\cdot) \in L^2([0, T], dt)$, the above condition will be satisfied provided that $EX^2(t)$ is bounded on $[0, T]$. But by Liptser and Shiryaev [10, Vol. II, Thm. 15.1, p. 135], the mean function $EX(t)$ and the covariance function $\Gamma(t) = \{EX(t) - EX(t)\}^2$

are solutions of the differential equations

$$\begin{aligned}\frac{dEX(t)}{dt} &= \theta(t)EX(t), & EX(0) &= x_0, \\ \frac{d\Gamma(t)}{dt} &= 2\theta(t)\Gamma(t) + \sigma^2, & \Gamma(0) &= 0,\end{aligned}$$

which give

$$EX(t) = x_0 \exp \int_0^t \theta(u) du, \quad \Gamma(t) = \sigma^2 \int_0^t \exp \left\{ 2 \int_s^t \theta(u) du \right\} ds.$$

Thus, $EX(t)$ and $\Gamma(t)$ are bounded on $[0, T]$, and hence so is $EX^2(t)$.

The Radon–Nikodým derivative of P_θ with respect to P_0 is

$$\frac{dP_\theta}{dP_0}(X(\cdot)) = \exp \left\{ \frac{1}{\sigma^2} \int_0^T \theta(t)X(t) dX(t) - \frac{1}{2\sigma^2} \int_0^T \theta^2(t)X^2(t) dt \right\}.$$

We are interested in estimating the function $\theta(\cdot)$ on $[0, T]$ based on a sample $X_1(\cdot), X_2(\cdot), \dots, X_n(\cdot)$ of n independent trajectories of $X(t)$ on $[0, T]$. Here σ^2 is assumed to be known. The log likelihood function is then

$$L_n(\theta) = \sum_{k=1}^n \left\{ \frac{1}{\sigma^2} \int_0^T \theta(t)X_k(t) dX_k(t) - \frac{1}{2\sigma^2} \int_0^T \theta^2(t)X_k^2(t) dt \right\}.$$

We shall use as sieves an increasing sequence V_n of subspaces of $L^2([0, T], dt)$, with finite dimension d_n such that $\bigcup_{n \geq 1} V_n$ is dense in $L^2([0, T], dt)$. The method of sieves consists of maximizing $L_n(\theta)$ on V_n . For this purpose let ϕ_j , $j = 1, 2, \dots$, be a sequence of independent vectors of $L^2([0, T], dt)$ such that $\phi_1, \phi_2, \dots, \phi_{d_n}$ form a basis of V_n for all n . Then for $\theta \in V_n$, $\theta(\cdot) = \sum_{j=1}^{d_n} \theta_j \phi_j(\cdot)$, we have

$$\begin{aligned}L_n(\theta) &= \sum_{k=1}^n \left[\frac{1}{\sigma^2} \int_0^T \left\{ \sum_{j=1}^{d_n} \theta_j \phi_j(t) \right\} X_k(t) dX_k(t) \right. \\ &\quad \left. - \frac{1}{2\sigma^2} \int_0^T \left\{ \sum_{j=1}^{d_n} \theta_j \phi_j(t) \right\}^2 X_k^2(t) dt \right] \\ &= \frac{n}{\sigma^2} \{ B^{(n)} \theta^{(n)} - \frac{1}{2} \theta^{(n)'} A^{(n)} \theta^{(n)} \},\end{aligned}$$

where $\theta^{(n)}$, $B^{(n)}$ and $A^{(n)}$ are the vectors and the matrix with general elements θ_j , $j = 1, 2, \dots, d_n$.

$$B_j^{(n)} = \frac{1}{n} \sum_{k=1}^n \theta_j(t)X_k(t) dX_k(t), \quad j = 1, 2, \dots, d_n$$

and

$$A_{ij}^{(n)} = \frac{1}{n} \sum_{k=1}^n \int_0^T \phi_i^{(t)} \phi_j^{(t)} X_k^2(t) dt, \quad i, j = 1, 2, \dots, d_n.$$

Therefore, the restricted maximum likelihood estimator $\hat{\theta}^{(n)}(\cdot)$ of θ is given by $\hat{\theta}^{(n)}(\cdot) = \sum_{j=1}^{d_n} \hat{\theta}_j^{(n)} \phi_j^{(n)}(\cdot)$, where $\hat{\theta}^{(n)} = (\hat{\theta}_1^{(n)}, \hat{\theta}_2^{(n)}, \dots, \hat{\theta}_{d_n}^{(n)})$ is the solution of $A^{(n)} \hat{\theta}^{(n)} = B^{(n)}$. It can be shown that $A^{(n)}$ is invertible almost surely, so $\hat{\theta}^{(n)} = (A^{(n)})^{-1} B^{(n)}$.

Note that the equation $A^{(n)}\hat{\theta}^{(n)} = B^{(n)}$ is equivalent to

$$(2.2) \quad \begin{aligned} h \in V_n: \frac{1}{n} \sum_{k=1}^n \int_0^T h(t) X_k^2(t) \hat{\theta}^{(n)}(t) dt \\ = \frac{1}{n} \sum_{k=1}^n \int_0^T h(t) X_k(t) dX_k(t), \end{aligned}$$

which characterizes the function $\hat{\theta}^{(n)}(\cdot)$ independent of the basis of V_n .

3. Asymptotic properties of the estimator. To investigate the asymptotic properties of the estimator, it is convenient to introduce an orthonormal basis of V_n relative to a certain scalar product in order to transform $A^{(n)}$ into an almost identity matrix for large n . We observe that by the strong law of large numbers for independent identically distributed (i.i.d.) random variables $A_{ij}^{(n)}$ converges strongly as $n \rightarrow \infty$ to

$$\int_0^T \phi_i(t) \phi_j(t) EX^2(t) dt.$$

Therefore, we shall consider a sequence Ψ_1, Ψ_2, \dots , such that $\Psi_1, \Psi_2, \dots, \Psi_{d_n}$ is an orthonormal basis of V_n in the sense of the scalar product

$$(h, g) \rightarrow \int_0^T h(t)g(t) d\mu(t), \quad d\mu(t) = EX^2(t) dt,$$

which is well defined since $EX^2(t)$ is bounded on $[0, T]$, and hence $L^2([0, T], dt) \subset L^2([0, T], d\mu)$; let $\hat{\xi}_1^{(n)}, \hat{\xi}_2^{(n)}, \dots, \hat{\xi}_{d_n}^{(n)}$ be the co-ordinates of $\hat{\theta}^{(n)}(\cdot)$ in the new basis $\{\Psi_1, \Psi_2, \dots, \Psi_{d_n}\}$ of V_n . By (2.2) the vector $\hat{\xi}^{(n)} = (\hat{\xi}_1^{(n)}, \hat{\xi}_2^{(n)}, \dots, \hat{\xi}_{d_n}^{(n)})$ is the solution of $a^{(n)}\hat{\xi}^{(n)} = b^{(n)}$, where $a^{(n)}$ and $b^{(n)}$ are the matrix and the vector with general elements

$$\begin{aligned} a_{ij}^{(n)} &= \frac{1}{n} \sum_{k=1}^n \int_0^T \Psi_i(t) \Psi_j(t) X_k^2(t) dt, \\ b_i^{(n)} &= \frac{1}{n} \sum_{k=1}^n \int_0^T \Psi_i(t) X_k(t) dX_k(t) \\ &= \frac{1}{n} \sum_{k=1}^n \int_0^T \Psi_i(t) X_k^2(t) \theta(t) dt + \frac{1}{n} \sum_{k=1}^n \int_0^T \Psi_i(t) X_k(t) dW_k(t). \end{aligned}$$

Let $\theta^{(n)}(\cdot) = \sum_{i=1}^{d_n} \xi_i \Psi_i(\cdot)$ be the orthogonal projection of $\theta(\cdot)$ onto V_n in the sense of the above scalar product. Then the first term of the right-hand side can be written as

$$\begin{aligned} \sum_{j=1}^{d_n} a_{ij}^{(n)} \xi_j + \frac{1}{n} \sum_{k=1}^n \int_0^T \Psi_i(t) X_k^2(t) \{\theta(t) - \theta^{(n)}(t)\} dt \\ = \sum_{j=1}^{d_n} a_{ij}^{(n)} \xi_j + \frac{1}{n} \sum_{k=1}^n \int_0^T \Psi_i(t) \{X_k^2(t) EX^2(t)\} \{\theta(t) - \theta^{(n)}(t)\} dt, \end{aligned}$$

since $\int \Psi_i(t) \{\theta(t) - \theta^{(n)}(t)\} EX^2(t) dt = 0$ for all i . Therefore

$$(3.1) \quad a^{(n)}(\hat{\xi}^{(n)} - \xi^{(n)}) = c^{(n)},$$

where $\xi^{(n)}$ and $c^{(n)}$ are vectors with components $\xi_j, j = 1, 2, \dots, d_n$ and

$$c_j^{(n)} = \frac{1}{n} \sum_{k=1}^n \left[\int_0^T \Psi_j(t) \{X_k^2(t) - EX^2(t)\} \{\theta(t) - \theta^{(n)}(t)\} dt + \int_0^T \Psi_j(t) X_k(t) dW_k(t) \right].$$

By the law of large numbers, one would expect that $a^{(n)}$ and $c^{(n)}$ converge to the identity matrix and the zero vector in some sense, and hence, $\hat{\xi}^{(n)}$ would be close to $\xi^{(n)}$ for large n . More precisely, we shall show that under the condition $d_n^2/n \rightarrow 0$ as $n \rightarrow \infty$, $\|\hat{\xi}^{(n)} - \xi^{(n)}\|^2 \rightarrow 0$ in probability as $n \rightarrow \infty$, and under the condition $d_n^3/n \rightarrow 0$ as $n \rightarrow \infty$ for any d_n -vector $\lambda^{(n)}$ such that $\|\lambda^{(n)}\|^2$ is convergent, $\sqrt{n}\lambda^{(n)'}(\hat{\xi}^{(n)} - \xi^{(n)})$ is asymptotically normal. Going back to $\theta^{(n)}(\cdot)$, we obtain the consistency and asymptotic normality of the estimator. Note that the $\hat{\xi}_i^{(n)}$ are not statistics since they depend on the basis $\Psi_1, \Psi_2, \dots, \Psi_{d_n}$ which is defined in terms of the unknown measure μ . These $\hat{\xi}_i^{(n)}$ are introduced only for the purpose of studying the asymptotic properties of $\hat{\theta}^{(n)}(\cdot)$.

To obtain the convergence of $a_{ij}^{(n)}$ and $c_i^{(n)}$, we shall write them in the form

$$(3.2) \quad a_{ij}^{(n)} - \delta_{ij} = \frac{1}{n} \sum_{k=1}^n \left[\int_0^T \Psi_i(t) \Psi_j(t) Z_k(t) dt \right] = \frac{1}{n} \sum_{k=1}^n \eta_{ijk}, \quad \text{say,}$$

$$(3.3) \quad c_i^{(n)} = \frac{1}{n} \sum_{k=1}^n \left[\int_0^T \Psi_i(t) \{\theta(t) + \theta^{(n)}(t)\} Z_k(t) dt + \int_0^T \Psi_i(t) X_k(t) dW_k(t) \right] \\ = \frac{1}{n} \sum_{k=1}^n (\eta_{ik}^{(n)} + \tilde{\eta}_{ik}), \quad \text{say,}$$

where $\delta_{ij} = 0$ if $i \neq j$, $\delta_{ii} = 1$ and $Z_k(t) = X_k^2(t) - EX^2(t)$.

LEMMA 1. For any μ -integrable function h on $[0, T]$,

$$E \left\{ \int_0^T h(t) Z_k(t) dt \right\}^2 \leq 2 \left\{ \int_0^T |h(t)| d\mu(t) \right\}^2.$$

Proof. We have

$$E \left\{ \int_0^T h(t) Z_k(t) dt \right\}^2 \leq \int_0^T \int_0^T h(t) h(s) E \{ Z_k(t) Z_k(s) \} dt ds.$$

We observe that the $X(t)$ process solution of (2.1) is Gaussian; therefore $Y(t) = X(t) - EX(t)$ is a zero mean normal variate and hence $\text{Var} \{Y^2(t)\} = 2[\text{Var} \{Y(t)\}]^2$ and $\text{Cov} \{Y^2(t), Y(t)\} = 0$. Thus,

$$EZ_k^2(t) = \text{Var} \{X^2(t)\} = \text{Var} [Y^2(t) + 2\{EX(t)\}Y(t) + \{EX(t)\}^2] \\ = 2[\text{Var} \{Y(t)\}]^2 + 4\{EX(t)\}^2 \text{Var} \{Y(t)\} \\ \leq 2[\text{Var} \{X(t)\} + \{EX(t)\}^2]^2 = 2\{EX^2(t)\}^2.$$

Therefore, by the Schwarz inequality,

$$|E \{Z_k(t) Z_k(s)\}| \leq \{EZ_k^2(t) EZ_k^2(s)\}^{1/2} \leq 2EX^2(t) EX^2(s),$$

and hence,

$$E \left\{ \int_0^T h(t) Z_k(t) dt \right\}^2 \leq 2 \int_0^T \int_0^T |h(t)| |h(s)| d\mu(t) d\mu(s) \\ = 2 \left\{ \int_0^T |h(t)| d\mu(t) \right\}^2.$$

LEMMA 2. We have for all n

$$E[a_{ij}^{(n)} - \delta_{ij}]^2 \leq \frac{2}{n},$$

$$E\{C_i^{(n)}\}^2 \leq \left[\sigma + \left\{ 2 \int_0^T |\theta(t) - \theta^{(n)}(t)|^2 d\mu(t) \right\}^{1/2} \right]^2 / n.$$

Proof. The first part of the lemma is a direct consequence of [3.2] and Lemma 1. To prove the second part, note that by the triangular inequality

$$\{E(\eta_{ij}^{(n)} + \tilde{\eta}_{ij})^2\}^{1/2} \leq [E\{\eta_{ik}^{(n)}\}^2]^{1/2} + \{E(\tilde{\eta}_{ik}^2)\}^{1/2},$$

where $\eta_{ik}^{(n)}$ and $\tilde{\eta}_{ik}$ are given by the right-hand side of (3.3). But $E\tilde{\eta}_{ik}^2 = \sigma^2$, and by Lemma 1,

$$E\{\eta_{ik}^{(n)}\}^2 \leq 2 \left\{ \int_0^T |\Psi_i(t)| |\theta(t) - \theta^{(n)}(t)| d\mu(t) \right\}^2$$

$$\leq 2 \int_0^T |\theta(t) - \theta^{(n)}(t)|^2 d\mu(t).$$

The result then follows from (3.3).

LEMMA 3. Let $\|M\| = \sup \{\|Mx\|, \|x\| \leq 1\}$ be the operator norm of the matrix M . Then $\|M\|^2 \leq \sum M_{ij}^2$ and

$$\|M^{-1}\| \leq \left[1 + \left\{ \sum_{ij} (M_{ij} - \delta_{ij})^2 \right\}^{1/2} \right]^{-1},$$

provided that $\sum_{ij} (M_{ij} - \delta_{ij})^2 < 1$.

The result $\|M\|^2 \leq \sum_{ij} M_{ij}^2$ is a direct consequence of Schwartz inequality. Now, by assumption, $\|M - I\| < 1$ where I is the identity matrix. Therefore

$$M^{-1} = \sum_{m=0}^{\infty} (I - M)^m,$$

$$\|M^{-1}\| \leq \sum_{m=0}^{\infty} \|M - I\|^m \leq \left[1 - \left\{ \sum_{ij} (M_{ij} - \delta_{ij})^2 \right\}^{1/2} \right]^{-1}.$$

THEOREM 1. Under the condition $d_n^2/n \rightarrow 0$ as $n \rightarrow \infty$, $\|\hat{\xi}^{(n)} - \xi^{(n)}\| \rightarrow 0$ in probability as $n \rightarrow \infty$.

Proof. By (3.1), $\hat{\xi}^{(n)} - \xi^{(n)} = (a^{(n)} - 1c^{(n)})m$ so, using Lemma 3,

$$\|\hat{\xi}^{(n)} - \xi^{(n)}\| \leq \left[1 - \left\{ \sum_{i=1}^{d_n} \sum_{j=1}^{d_n} (a_{ij}^{(n)} - \delta_{ij})^2 \right\}^{1/2} \right]^{-1} \|c^{(n)}\|.$$

Assume $d_n^2/n \rightarrow 0$; then by Lemma 2, as $n \rightarrow \infty$

$$E \left\{ \sum_{i=1}^{d_n} \sum_{j=1}^{d_n} (a_{ij}^{(n)} - \delta_{ij})^2 \right\} \rightarrow 0, \quad E\|c^{(n)}\|^2 = E \left\{ \sum_{i=1}^{d_n} (c_i^{(n)})^2 \right\} \rightarrow 0.$$

Therefore, the two factors of the above right-hand side converge in probability to 1 and to 0 respectively; hence the result.

COROLLARY. Under the conditions $d_n \rightarrow \infty$, $d_n^2/n \rightarrow 0$ as $n \rightarrow \infty$,

$$\lim_{n \rightarrow \infty} \int_0^T |\hat{\theta}^{(n)}(t) - \theta^{(n)}(t)|^2 EX^2(t) dt \rightarrow 0$$

in probability.

Proof. The expression in the corollary is the square of the $L^2([0, T], d\mu)$ norm of $\hat{\theta}^{(n)}(\cdot) - \theta^{(n)}(\cdot)$, and hence can also be given by

$$\sum_{i=1}^{d_n} |\hat{\xi}_i^{(n)} - \xi_i|^2 + \sum_{i=d_{n+1}}^{\infty} \xi_i^2.$$

By Theorem 1 the first term of the above expression converges in probability to zero as $n \rightarrow \infty$. Now $\bigcup_{n \geq 1} V_n$, being dense in $L^2([0, T], dt)$, is also dense in $L^2([0, T], d\mu)$ in the metric of $L^2([0, T], d\mu)$. Therefore the last term of the above expression converges to zero as $n \rightarrow \infty$. The result follows.

Remark. The metric of $L^2([0, T], d\mu)$ is the natural metric for our problem. However, it is not convenient to work with since μ is unknown. One might consider the $L^2([0, T], dt)$ -norm of $\theta^{(n)}(\cdot) - \theta^{(n)}(\cdot)$, but it may fail to converge to zero since the two norms in $L^2([0, T], d\mu)$ and $L^2([0, T], dt)$ are not necessarily equivalent. They are so if $EX^2(t)$ is bounded below by a strictly positive constant which is true if $x_0 \neq 0$. When $x_0 = 0$, we can only show that as $n \rightarrow \infty$,

$$\int_{\varepsilon}^T |\hat{\theta}^{(n)}(t) - \theta^{(n)}(t)|^2 dt \rightarrow 0 \quad \text{in probability,}$$

for every $\varepsilon > 0$. It is probable that one cannot strengthen this result, since if $x_0 = 0$, $X(t)$ will be very small for t near 0, so it is not possible to estimate $\theta(t)$ with great precision unless the drift term $\theta(t)X(t)$ can be estimated with much greater precision.

LEMMA 4. Let $\lambda^{(n)} = (\lambda_1^{(n)}, \lambda_2^{(n)}, \dots, \lambda_{d_n}^{(n)})$ be such that

$$\sum_{i=1}^{d_n} (\lambda_i^{(n)})^2 \rightarrow \lambda^2 \quad \text{as } n \rightarrow \infty.$$

Then the random variable $\sqrt{n} \sum_{i=1}^{d_n} \lambda_i^{(n)} c_i^{(n)}$ is asymptotically normal with zero mean and variance $\lambda^2 \sigma^2$.

Proof. By (3.3),

$$\sqrt{n} \sum_{i=1}^{d_n} \lambda_i^{(n)} c_i^{(n)} = \frac{1}{\sqrt{n}} \sum_{k=1}^n \left[\sum_{i=1}^{d_n} \lambda_i^{(n)} \eta_{ik}^{(n)} + \sum_{i=1}^{d_n} \lambda_i^{(n)} \tilde{\eta}_{ik} \right].$$

But by Lemma 1

$$\begin{aligned} E \left\{ \sum_{i=1}^{d_n} \lambda_i^{(n)} \eta_{ik}^{(n)} \right\}^2 &\leq 2 \left\{ \int_0^T \left| \sum_{i=1}^{d_n} \lambda_i^{(n)} \psi_i(t) \right| |\theta(t) - \theta^{(n)}(t)| d\mu(t) \right\}^2 \\ &\leq 2 \sum_{i=1}^{d_n} \{\lambda_i^{(n)}\}^2 \int_0^T \{\theta(t) - \theta^{(n)}(t)\}^2 d\mu(t). \end{aligned}$$

Since the variance of $\sum_{i=1}^{d_n} \lambda_i^{(n)} \tilde{\eta}_{ik}$ is $\sum_{i=1}^{d_n} \{\lambda_i^{(n)}\}^2 \sigma^2$, the variance of the bracket $[\]$ in the above equality tends to $\lambda^2 \sigma^2$ as $n \rightarrow \infty$. The result then follows from the central limit theorem for triangular arrays of random variables.

THEOREM 2. Let $\lambda^{(n)}$ be as in the Lemma 4 and suppose that $d_n^3/n \rightarrow 0$ as $n \rightarrow \infty$. Then $\sqrt{n} \sum_{i=1}^{d_n} \lambda_i^{(n)} (\hat{\xi}_i^{(n)} - \xi_i)$ is asymptotically normal with zero mean and variance $\lambda^2 \sigma^2$.

Proof. By (3.1), $\hat{\xi}^{(n)} - \xi^{(n)} = (a^{(n)})^{-1} c^{(n)}$, so $\hat{\xi}^{(n)} - \xi^{(n)} - c^{(n)} = (a^{(n)})^{-1} \{I - a^{(n)}\} c^{(n)}$. Therefore, with $\|\cdot\|$ denoting the operator norm or the Euclidean norm, we have

$$|\lambda^{(n)'} (\hat{\xi}^{(n)} - \xi^{(n)} - c^{(n)})| \leq \|\lambda^{(n)}\| \|a^{(n)-1}\| \|a^{(n)} - I\| \|c^{(n)}\|.$$

But by Lemma 2 for all n ,

$$nE\|c^{(n)}\|^2 = n \sum_{i=1}^{d_n} E(c_i^n)^2 \leq (\text{Const.}) d_n,$$

$$E\|a^{(n)} - I\|^2 \leq E\left\{ \sum_{i=1}^{d_n} \sum_{j=1}^{d_n} (a_{ij}^{(n)} - \delta_{ij})^2 \right\} \leq \frac{2d_n^2}{n}.$$

Therefore $\sqrt{n}\|a^{(n)} - I\|\|c^{(n)}\| \rightarrow 0$ in probability as $n \rightarrow \infty$, since $d_n^3/n \rightarrow 0$ by assumption. On the other hand, we have seen in the proof of Theorem 1 that $\|(a^{(n)})^{-1}\| \rightarrow 1$ in probability as $n \rightarrow \infty$. Hence, $\sqrt{n}\lambda^{(n)'}(\hat{\xi}^{(n)} - \xi^{(n)} - c^{(n)}) \rightarrow 0$ in probability as $n \rightarrow \infty$. But $\sqrt{n}\lambda^{(n)'}c^{(n)}$ is asymptotically normal with zero mean and variance $\sigma^2\lambda^2$ by Lemma 4, which gives the result.

COROLLARY. *Let h be a function in $L^2([0, T], dt)$ such that $\int_0^T h^2(t)/EX^2(t) dt < +\infty$. Then under the conditions of Theorem 2,*

$$\sqrt{n} \int_0^T h(t)\{\hat{\theta}^{(n)}(t) - \theta^{(n)}(t)\} dt$$

is asymptotically normal with zero mean and variance

$$\int_0^T h^2(t)/EX^2(t) dt.$$

Proof. By assumption, the function $\tilde{h}(\cdot) = h(\cdot)/EX^2(\cdot)$ is in $L^2([0, T], d\mu)$. Therefore, using the fact that $\hat{\theta}^{(n)}(\cdot) - \theta^{(n)}(\cdot) = \sum_{i=1}^{d_n} (\hat{\xi}_i^{(n)} - \xi_i)\Psi_i(\cdot)$, we have

$$\int_0^T h(t)\{\hat{\theta}^{(n)}(t) - \theta^{(n)}(t)\} dt = \int_0^T h(t)\{\hat{\theta}^{(n)}(t) - \theta^{(n)}(t)\} d\mu(t) = \sum_{i=1}^{d_n} \tilde{h}_i(\hat{\xi}_i^{(n)} - \xi_i),$$

where $\tilde{h}_i = \int_0^T \Psi_i(t)\tilde{h}(t) d\mu(t)$. Now by Parseval's theorem,

$$\lim_{n \rightarrow \infty} \sum_{i=1}^{d_n} \tilde{h}_i^2 = \sum_{i=1}^{\infty} \tilde{h}_i^2 = \int_0^T \tilde{h}^2(t) d\mu(t) = \int_0^T h^2(t)/EX^2(t) dt.$$

The result then follows from Theorem 2.

Remark. In the above corollary, we only prove the asymptotic normality of

$$\sqrt{n} \int_0^T h(t)\{\hat{\theta}^{(n)}(t) - \theta^{(n)}(t)\} dt$$

and not of

$$\sqrt{n} \int_0^T h(t)\{\hat{\theta}^{(n)}(t) - \theta(t)\} dt.$$

However, the latter will have the same asymptotic distribution as the former if

$$\sqrt{n} \int_0^T h(t)\{\theta(t) - \theta^{(n)}(t)\} dt \rightarrow 0, \quad n \rightarrow \infty.$$

The condition is clearly satisfied if the $L^2([0, T], d\mu)$ distance between $\theta(\cdot)$ and $\theta^{(n)}(\cdot)$ converges to 0 as $n \rightarrow \infty$ faster than $n^{-1/2}$. But the $L^2([0, T], d\mu)$ distance is less than a constant, say C , times the $L^2([0, T], dt)$ distance; hence, if δ_n denotes the infimum of the $L^2([0, T], dt)$ distance from $\theta(\cdot)$ to a point of V_n , then the $L^2([0, T], d\mu)$ distance between $\theta(\cdot)$ and $\theta^{(n)}(\cdot)$ is certainly less than $C\delta_n$. Thus the condition is satisfied if $\delta_n \rightarrow 0$ faster than $n^{-1/2}$ as $n \rightarrow \infty$.

Acknowledgment. Our sincere thanks are extended to the referees for valuable comments leading to a substantial improvement in the presentation of the paper.

REFERENCES

- [1] G. BANON AND H. T. NGUYEN, *Recursive estimation in diffusion model*, this Journal, 19 (1981), pp. 676–685.
- [2] D. BOSQ, *Contribution à la théorie de l'estimation fonctionnelle*, Publ. Inst. Stat. Univ., Paris, 1970.
- [3] F. DELEBECQUE AND J. P. QUADRAT, *Identification d'une diffusion stochastique*, Technical report, IRIA #121, Paris, 1975.
- [4] ———, [1978], *Sur l'estimation des caractéristiques locales d'un processus de diffusion avec sauts*, Technical Report IRIA #311, Paris, 1978.
- [5] P. D. FEIGIN, *Maximum likelihood estimation for continuous time stochastic processes*, Adv. Appl. Probab., 8 (1976), pp. 712–736.
- [6] S. GEMAN AND C. R. HWANG, *Nonparametric maximum likelihood estimation by the method of sieves*, Report on Pattern Analysis #80, Brown Univ., Providence, RI, 1979.
- [7] U. GRENANDER, *Stochastic processes and statistical inference*, Arkiv. Math., 1 (1950), pp. 195–277.
- [8] ———, *Abstract Inference*, John Wiley, New York, 1981.
- [9] I. A. IBRAGIMOV AND Y. A. ROZANOV, *Gaussian Random Processes*, Springer, New York, 1978.
- [10] R. S. LIPTSER AND A. M. SHIRYAEV, *Statistics of Random Processes*, Vols. I, II, Springer, New York, 1977, 1978.
- [11] H. T. NGUYEN, *Density estimation in a continuous time stationary Markov process*, Ann. Statist., 9 (1979), pp. 341–348.
- [12] H. T. NGUYEN AND T. D. PHAM, *Law of large numbers for martingales and application to statistics*, C. R. Acad. Sci. Paris, A, 288 (1979), pp. 303–305.
- [13] T. D. PHAM, *Estimation of parameters of a continuous time Gaussian Stationary process with rational spectral density*, Biometrika, 64 (1979), pp. 385–399.
- [14] ———, *Nonparametric estimation of the drift coefficient in the diffusion equation* Tech. rep. Séminaire de Statistique, Univ. Grenoble, France, (1978–1979), pp. 103–124.
- [15] A. LEBRETON, *Sur l'estimation de paramètre dans les modèles différentiels*, Thesis, Univ. Grenoble, France, 1976.
- [16] H. L. VAN TREES, *Detection, Estimation and Modulation Theory*, Vols. I, II, John Wiley, New York, 1968, 1971.

ON SINGULAR IMPLICIT LINEAR DYNAMICAL SYSTEMS*

PIERRE BERNHARD†

Abstract. We investigate properties of existence, unicity, representation, of the (causal) solutions of implicit linear systems (or “generalized systems”) when the underlying matrix pencil is singular. We relate the geometric and the algebraic approaches. The main conclusion is that if the underlying matrix pencil is “column singular” (i.e., has a nonempty set of column minimal indices) the causal solutions, when they exist, can exactly be represented as the output of a classical two-player dynamical system, where the second player accounts for the nonunicity. Properties of the equivalent system are related to those of the singular matrix pencils made with the given matrices.

1. Introduction.

1.1. Problems considered. We study systems given in one of the following two forms, respectively, discrete and continuous:

$$(*) \quad Ey(t+1) = Fy(t) + Gu(t),$$

$$(**) \quad E \frac{dy}{dt}(t) = Fy(t) + Gu(t),$$

with the following definitions:

$y(t) \in \mathbb{R}^m$ is the (fundamental) output of the system,

$u(t) \in \mathbb{R}^p$ is the input.

E and F are $r \times m$ constant matrices, G is a $r \times p$ constant matrix. r is called the *rank* of the system. It may be larger than, equal to or lower than m .

The questions of existence and unicity we shall investigate arise only if E is not invertible (in case $r = m$). We shall also consider problems of representation and canonical forms. We are mainly interested in *singular* systems, where the solution is nonunique. (See Definition 3 and Theorem 2 for a precise statement.)

DEFINITION 1. If $r = m$, the system is called *square*.

PROPOSITION 1. A system (E, F, G) is always equivalent:

- (i) to a system with rank equal to the rank of the composite matrix $[E \ F \ G]$;
- (ii) if this rank is lower than m to a square system.

Proof. (i) If the lines of the composite matrix $[E \ F \ G]$ are not independent, we can always delete redundant equations in $(*)$ or $(**)$.

- (ii) If $r < m$ we can add lines of zeros to them.

HYPOTHESIS. Because of property (i) above, we shall always assume that $\text{rank } [E \ F \ G] = r$.

1.2. Motivation. (i) P.I.D. control. Systems of the form $(**)$ naturally arise when applying output derivative feedback to an ordinary system. There the resulting implicit system is square. The interesting question is its limit behavior when the E matrix is “close” to be singular. A prerequisite to a complete understanding of the resulting “infinite frequency” modes (see [20]) is the present analysis.

(ii) Systems with a linear state or state-control constraint. An equation of the form

$$0 = Cy + Du$$

* Received by the editors May 11, 1981, and in revised form October 20, 1981.

† INRIA, Domaine de Voluceau-Rocquencourt, B.P. 105-78150 Le Chesnay, France.

may be added to a standard system as an extra set of equations resulting in a matrix E made of the identity and lines of zeros. There $r > m$.

(iii) Interconnected systems. The natural statement of the equations of sets of interconnected systems may lead to equations of the type (ii).

(iv) Econometric systems. Econometric systems are almost always of the form (*) (or a more complex one with nonlinear r.h.s.). Most famous among them are Leontief's models, and ARMA models with noninvertible leading coefficient.

(v) Perturbed systems. The perturbed system

$$\dot{x} = Ax + Bu + Cv$$

is equivalent to the implicit system

$$E\dot{x} = EAx + EBu,$$

where E is a matrix of maximum rank such that $EC = 0$. As a matter of fact, for any pair of measurable input functions $(u(\cdot), v(\cdot))$, the solution of the first, explicit, differential equation satisfies the second, implicit, one. Conversely, to any pair of an absolutely continuous $x(\cdot)$ and a measurable $u(\cdot)$ satisfying the second one, corresponds a measurable $v(\cdot)$ such that $x(\cdot)$ satisfies the first one with inputs $u(\cdot)$ and $v(\cdot)$.

(vi) Time reversibility in discrete time systems. Backward projection for a standard discrete system

$$x_{k+1} = Fx_k + Gu_k$$

leads to the study of the backward system

$$F\bar{x}_{k+1} = \bar{x}_k - G\bar{u}_k,$$

where $\bar{x}_{k+1} = x_{-(k+1)}$ and $\bar{u}_k = u_{-k-1}$.

(vii) Operator splitting numerical methods. Solution of the equation

$$Ay = f$$

can be pursued using a recursion of form (*) with $A = E - F$ and $Gu = f = \text{constant}$ (or $Gu_k \rightarrow f$).

(viii) Implicit differential equations. The representation results obtained here may be of some interest for their own sake in the study of implicit linear differential equations.

1.3. Originality. More than ten years ago, Rosenbrock's theory was explicitly devised to address implicit systems of a more complicated type since higher derivatives were allowed as well as derivatives of the control. See a rather complete account in Rosenbrock [14]. Since then the precise type of systems we study have been investigated by Luenberger and coworkers [12], [13], [15]. Beyond problems of existence and unicity, they have considered optimization problems. More recently, papers by Verghese, Kailath and coworkers have dealt with the infinite frequency aspects of these systems [16], [17]. Systems of the form (*) also appear in connection with linear programming, see, for instance, [5].

All the above references deal with the "regular case", i.e., square systems with $\det(zE - F) \neq 0$. In that case, as we shall see, existence implies unicity. Our main emphasis is on the singular case, and the representation of nonunicity. Some works on that topic are due to Campbell. While [4] again deals only with the regular case, [3] considers a very particular instance of the singular case. It is a subcase of our "static nonunicity". Moreover, his application to linear systems is further restricted to the regular case.

While this paper was being typed, the author became aware (through D. Gabay, of Inria) of the work of Wilkinson [21]. It deals with the general singular case but lacks the necessary tools of control theory to give a complete description of the nonunicity via invariants. It essentially covers the method of our paragraph 5.4 without the references to the geometric and transfer function theories.

After this paper was first submitted for publication, several articles appeared on that topic¹, covering both the regular case, see [18], [19] and [20] (which is a more complete account of an earlier publication in the 1979 IEEE CDC, held in 1980), and, more importantly to us, the singular case. See [10] and [11], which rely heavily on an analysis very similar to that of our paragraph 5.4. Reference [1] is also an approach of system theory without unicity.

While [18], and to a smaller extent [17], use some geometrical concepts, the literature has been in most part algebraic in nature. We believe, however, that our § 2 shows that the geometric approach allows a completely elementary treatment of both the regular and singular cases.

1.4. Outline. In § 2 we develop the (elementary) geometric theory of strictly causal discrete systems (*). In the very short § 3, we check that all the results but a minor one carry over to the continuous case. In § 4 we investigate the geometric theory of the causal (but not strictly causal) case. Section 5 is devoted to the algebraic theory, invariants, transfer functions and canonical forms.

2. Discrete time systems, the strictly causal case.

2.1. Causality. We quickly review here what *causality*, or *strict causality*, means for a dynamical system with possibly nonunique solutions. We deal with the discrete system (*), the extension to (**) is straightforward, provided, in the definition of causality, “ $\forall t$ ” be replaced by “for almost all t ”. As a consequence, the difference between causality and strict causality, as given here, vanishes. Strict causality, in the continuous case, will carry an added requirement. See § 3.

Let Ω be the set of admissible control functions, i.e., applications from $[t_0, t_1]$ into \mathbb{R}^m . (Usually, $t_1 = +\infty$.) A *correspondence of solutions* is a set-valued function S from Ω into the set of trajectories, which to each $u(\cdot)$ in Ω associates a set $S(u(\cdot))$ of trajectories $y(\cdot)$ satisfying (*). Let $S_\tau(u(\cdot))$ be the set of the restrictions to $[t_0, \tau]$ of the elements of $S(u(\cdot))$. We recall the following.

DEFINITION. The correspondence S is called *strictly causal* if given $u_1(\cdot)$ and $u_2(\cdot)$ in Ω

$$\text{if } u_1(t) = u_2(t) \quad \forall t < \tau \quad \text{then } S_\tau(u_1(\cdot)) = S_\tau(u_2(\cdot)).$$

S is said *causal* if the conclusion holds provided $u_1(t) = u_2(t)$, for all $t \leq \tau$. (In all the sequel, “strictly causal” may correspondingly be replaced by “causal”.) The set of strictly causal solutions of the system is the maximal strictly causal correspondence of solutions, i.e., the union \bar{S} of all of them. Given $u(\cdot)$ in Ω , a trajectory $y(\cdot)$ is called a *strictly causal solution* if it belongs to $\bar{S}(u(\cdot))$.

A characteristic property of a strictly causal solution is that, in addition to satisfying (*) for all t , it is such that, for all τ in (t_0, t_1) , the system (*) initialized at $y(\tau)$ has strictly causal solutions for every sequence $\{u(t), t \geq \tau\}$. The reader may easily check that this inductive characterization is indeed necessary and sufficient. It will be used hereafter in the proofs.

¹ Some were pointed out to us by a reviewer whom we thank here.

Remark. Restricting oneself to the causal case, as we shall do, amounts precisely to ignoring the “impulsive modes” of the theory as developed in [17], [20]. As a matter of fact, we want to focus here on the nonunicity, not on impulsive modes.

2.2. Existence. We write $\mathcal{E} = \mathcal{R}(E)$ and $\mathcal{G} = \mathcal{R}(G)$, the respective ranges of E and G , as subspaces of $Y = \mathbb{R}^l$. Consider the following relation for a linear subspace \mathcal{V} of Y :

$$(1) \quad F\mathcal{V} \subset E\mathcal{V}.$$

DEFINITION 1. We call *characteristic subspace* of the pair (E, F) the largest subspace \mathcal{V}^* satisfying (1).

PROPOSITION 1. *This subspace exists since $\{0\}$ satisfies (1), and this equation being stable under addition of subspaces, \mathcal{V}^* is the sum of all subspaces that satisfy it. (However, \mathcal{V}^* may be trivial.)*

This space is clearly related to the solution of

$$E\dot{x} = Fx \quad \text{or} \quad x(t+1) = Fx(t)$$

and can be considered a dynamic invariant. It will be seen further that it contains the “generalized eigenvectors” of this system.

In the special case where we are given a two input system

$$x(t+1) = Ax(t) + Bu(t) + Cv(t)$$

and where as in motivation (v) (*) is obtained by taking an injective matrix E such that $\text{Ker } E = \mathcal{R}(C)$:

$$Ex(t+1) = EAx(t) + EBu(t),$$

then (1) translates in

$$EA\mathcal{V} \subset E\mathcal{V},$$

which is equivalent to

$$A\mathcal{V} \subset \mathcal{V} + \mathcal{R}(C).$$

Therefore \mathcal{V} is A invariant mod C and \mathcal{V}^* is then the largest A invariant mod C subspace, i.e., \mathbb{R}^n . The similarity with (A, C) invariance, which was suggested by a reviewer, is further displayed in the algebraic theory. See Remark 7.

THEOREM 1. *The system (*) has a strictly causal solution over an interval of arbitrary length, for any control sequence $u(\cdot)$, if and only if*

$$(2) \quad \mathcal{G} \subset E\mathcal{V}^*,$$

$$(3) \quad y(0) \in \mathcal{V}^*.$$

Proof. (i) *Necessity.* Let t be given. In order for $y(t+1)$ to exist, it is necessary that

$$Fy(t) + Gu(t) \in \mathcal{E},$$

and since this must be true for all $u(t) \in \mathbb{R}^p$, this implies

$$\mathcal{G} \subset \mathcal{E} \quad \text{and} \quad y(t) \in \mathcal{V}^0 = F^{-1}(\mathcal{E}).$$

In order for the last relation to hold for every $u(t-1)$, we need

$$\mathcal{G} \subset E\mathcal{V}^0, \quad y(t-1) \in \mathcal{V}^1 = F^{-1}(E\mathcal{V}^0).$$

Continuing this process, we construct the sequence \mathcal{V}^k by

$$(4) \quad \mathcal{V}^{k+1} = F^{-1}(E\mathcal{V}^k),$$

and we must have for all k ,

$$\mathcal{G} \subset E\mathcal{V}^k, \quad y(t-k) \in \mathcal{V}^k.$$

Necessity follows from the following fact.

PROPOSITION 2. *The sequence \mathcal{V}^k is decreasing and converges to \mathcal{V}^* in no more than m steps.*

Proof. Clearly, $F^{-1}(E\mathcal{V}^0) \subset F^{-1}(\mathcal{G})$, and thus, $\mathcal{V}^1 \subset \mathcal{V}^0$, and so on by induction. However, subspaces can decrease only by losing one dimension, which cannot occur more than m times in \mathbb{R}^n . Let k be the first index such that $\mathcal{V}^{k+1} = \mathcal{V}^k$. The sequence \mathcal{V}^k becomes stationary from this point on, and (4) shows that \mathcal{V}^k satisfies (1). Therefore, $\mathcal{V}^k \subset \mathcal{V}^*$. This establishes the necessity of (2), (3), but not the proposition, which states that $\mathcal{V}^k = \mathcal{V}^*$. This can easily be proved directly but follows also from the sufficiency of (2), (3) that we now establish.

(ii) *Sufficiency.* Let V be a rectangular injective (full column rank) matrix such that $\mathcal{R}(V) = \mathcal{V}^*$ (let $\dim \mathcal{V}^* = n^*$, $V: m \times n^*$). Relations (1) and (2) imply

$$(5) \quad \exists \bar{A}: \quad FV = EV\bar{A},$$

$$(6) \quad \exists \bar{B}: \quad G = EV\bar{B},$$

where \bar{A} is a $n^* \times n^*$ matrix and \bar{B} is $n^* \times p$. We also have that $y(t) \in \mathcal{V}^*$ is equivalent to

$$(7) \quad \exists \xi(t) \in \mathbb{R}^{n^*}: \quad y(t) = V\xi(t).$$

Now, (*) is equivalent to

$$(8) \quad EV\xi(t+1) = EV(\bar{A}\xi(t) + \bar{B}u(t)),$$

which together with (3) has the obvious solution

$$(9) \quad \xi(t+1) = \bar{A}\xi(t) + \bar{B}u(t), \quad y(0) = V\xi(0). \quad \square$$

Remark 1. When (2) is not satisfied, we may restrict u to belong to $\mathcal{U}_{\text{ad}} = G^{-1}(E\mathcal{V}^*)$. In the sequel, condition (2) may always be understood to mean that this reduction has been performed and will always be assumed to hold.

2.3. Unicity.

DEFINITION 2. We call *characteristic kernel* of the pair (E, F) the subspace \mathcal{N} defined by

$$(10) \quad \mathcal{N} = \text{Ker } E \cap \mathcal{V}^*.$$

Let $\dim \mathcal{N} = q$.

DEFINITION 3. The pair (E, F) is said *C-regular* (or more accurately *column regular*) if $q = 0$:

$$(11) \quad \mathcal{N} = \{0\}.$$

THEOREM 2. *Under conditions (2) and (3), the solution to equation (*) is unique, for any $u(\cdot)$, if and only if the system (the pair E, F) is C-regular. Otherwise, the nonunicity is described by the arbitrary choice of the sequence $v(\cdot)$ in equation (14), and (14), (15) constitute a representation of all solutions of (*).*

Proof. (i) *Unicity.* Equation (8) implies (9) only modulo the kernel of EV , which reduces to $\{0\}$ under and only under condition (11).

(ii) *Nonunicity.* If $\mathcal{N} \neq \{0\}$, let us choose a decomposition of \mathcal{V}^* of the form

$$(12) \quad \mathcal{V}^* = \mathcal{M} \oplus \mathcal{N}.$$

To this decomposition we may associate a partition of V of the form

$$(13) \quad V = [M \ N], \quad EV = [EM \ 0].$$

Let us partition accordingly ξ , \bar{A} and \bar{B} in the following way:

$$\xi = \begin{pmatrix} x \\ v \end{pmatrix}, \quad \bar{A} = \begin{pmatrix} A & C \\ \tilde{A} & \tilde{C} \end{pmatrix}, \quad \bar{B} = \begin{pmatrix} B \\ \tilde{B} \end{pmatrix}.$$

By definition EM is injective so that (8) is equivalent to

$$(14) \quad x(t+1) = Ax(t) + Bu(t) + Cv(t),$$

$$(15) \quad y(t) = Mx(t) + Nv(t). \quad \square$$

The nonunicity is therefore described as the effect of an extra input in a classical linear system. We may apply to it the tools of two-player control systems. In that respect it is worthwhile to notice that V being injective knowledge of y is equivalent to the knowledge of both x and v . (This is important, for instance, in discrete capturability theory [2].)

Remark 2. The matrix C may of course be of less than full column rank. If this is the case, by a proper choice of basis we can write

$$C = [C_1 \ 0],$$

accordingly partitioning v in $v' = (v'_1 v'_2)$. Then v_1 must be considered as parametrizing a *dynamic nonunicity* since its effect propagates forward in time through the dynamics, while v_2 parametrizes a *static nonunicity* since it appears only in the output equation (15) (Recall that N is injective.)

The triple (A, B, C) is clearly nonunique. It may be altered through a change of basis within \mathcal{V}^* . This leads to the following fact.

PROPOSITION 3. *The pair (A, C) is uniquely defined up to a transformation of Brunovsky's feedback group (see Kalman [9]).*

Proof. A change of basis within \mathcal{V}^* can be described as

- (i) a change of basis within \mathcal{N} , i.e., on v ;
- (ii) a change of choice of \mathcal{M} within \mathcal{V}^* . Let \tilde{M} generate an alternate $\tilde{\mathcal{M}}$:

$$y = Mx + Nv = \tilde{M}\tilde{x} + N\tilde{v}.$$

The difference $v - \tilde{v}$ depends linearly on y and is null when $y \in \mathcal{N}$, i.e., when $x = 0$. Therefore it depends linearly on x alone:

$$\tilde{v} = Px + v.$$

Using the fact that M is injective, this gives

$$\tilde{x} = Qx,$$

where Q can be calculated as a function of M, \tilde{M}, N and P . Therefore, this is equivalent to a state feedback superimposed on v and a change of basis on x .

- (iii) a change of basis on x alone (which can, of course, undo the previous one). \square

We shall study further the invariants of (A, C) . However, an interesting geometric one at this point will be provided by the following definition.

DEFINITION 4. We call the *neutral subspace* of the pair (E, F) the smallest subspace \mathcal{V}_* that satisfies (1) and contains \mathcal{N} .

PROPOSITION 4. *Such a subspace exists as a consequence of Theorem 5 below, that is, by applying it with $G = 0$.*

THEOREM 3. *Two solutions of (*) corresponding to the same initial point and same sequence $u(\cdot)$ are equal modulo \mathcal{V}_* . \mathcal{V}_* is image by V of the reachable space of the pair (A, C) in (14).*

Proof. By subtraction, two solutions of (*) corresponding to the same initial point and the same sequence $u(\cdot)$ have their differences δy that satisfy

$$\begin{aligned}\delta y(t) &= V\delta\xi(t) + M\delta x(t) + N\delta v(t), & \delta\xi(0) &= 0, \\ \delta x(t+1) &= A\delta x(t) + C\delta v(t), & \delta x(0) &= 0.\end{aligned}$$

Therefore, $\delta x(t)$ belongs to the reachable space of the pair (A, C) . Conversely, any solution of this system remains strictly causal and satisfies

$$(16) \quad E\delta y(t+1) = F\delta y(t), \quad \delta y(0) = 0$$

and can therefore be added to a solution of (*) and still remain a solution.

The fact that \mathcal{V}_* is exactly the (image of) reachable subspace of the pair (A, C) will be a corollary of Theorem 5 below. \square

2.3. Minimality.

DEFINITION 5. We call the *maximum subspace* of the triple (E, F, G) the largest subspace \mathcal{W}^* satisfying

$$(17) \quad F\mathcal{W}^* + \mathcal{G} = E\mathcal{W}^*.$$

PROPOSITION 5. *The subspace \mathcal{W}^* exists; it is a subspace of \mathcal{V}^* and is the limit, attained in no more than m steps of the sequence \mathcal{W}^k defined by*

$$(18) \quad \mathcal{W}^0 = \mathcal{V}^*, \quad \mathcal{W}^{k+1} = E^{-1}(F\mathcal{W}^k + \mathcal{G}) \cap \mathcal{V}^*.$$

Proof. Notice that since $F\mathcal{V}^* + \mathcal{G} \subset E\mathcal{V}^*$ we have

$$(19) \quad E\mathcal{W}^1 = F\mathcal{W}^0 + \mathcal{G} \quad \text{and} \quad \mathcal{W}^1 \subset \mathcal{V}^*.$$

It follows easily that property (19) holds at every step of the algorithm, shifting the indices of \mathcal{W} by an equal number and also that the sequence is decreasing. It therefore has a limit which satisfies (17), of which it is easy to check that it is the largest solution of (17) (which is stable by addition of subspaces).

THEOREM 4. *\mathcal{W}^* is the largest subspace traversed by the asymptotic regime of (*), i.e., for all $k \geq n$; the application $(y(0), u(\cdot)) \mapsto y(k)$ is surjective over \mathcal{W}^* , which is exactly its range.*

Proof. By construction, (3) implies $y(1) \in \mathcal{W}^1$ and, by induction, $y(t) \in \mathcal{W}^t$ with surjectivity. This, with the proposition, proves the theorem. \square

While this result characterizes in some sense the reachable subspace of (*), it is not the most interesting one. As a matter of fact, classical system theory teaches us that the reachable subspace of interest is that which is reachable from the state zero. We therefore proceed with the following.

DEFINITION 6. The minimal subspace of the triple (E, F, G) is the smallest subspace \mathcal{W}_* satisfying (1) and (2) and containing \mathcal{N} .

Remark 3. In the case where $E = I$, this is a classical characterization of the reachable space of (F, G) .

THEOREM 5. \mathcal{W}_* exists and is the limit (in m steps or less) of the same recurrence as in (18) but initialized with $\mathcal{W}_0 = \{0\}$. It is the image by V of the reachable space of the system (14), where both u and v are taken as controls.

Proof. Notice first that property (1) is not stable under intersection and, therefore, the existence of a smallest subspace satisfying it and other conditions is not obvious. Consider the recurrence (18) initialized with $\mathcal{W}_0 = \{0\}$;

$$\mathcal{W}_1 = E^{-1}(\mathcal{G}) \cap \mathcal{V}^*.$$

Because of (2),

$$E\mathcal{W}_1 = \mathcal{G}.$$

The sequence \mathcal{W}_k is clearly increasing and, by induction, satisfies the same sequence of equalities of the form (19) as \mathcal{W}^k . By construction, $\mathcal{N} \subset \mathcal{W}_k$ for all k . Therefore, it has a limit \mathcal{W}_* that satisfies (1), (2) and contains \mathcal{N} .

According to Theorem 2, the image by V of the reachable space of (14) is exactly the reachable space for $y(t)$ from zero. By construction it is the limit of the above recurrence.

That \mathcal{W}_* be the smallest subspace satisfying (1), (2) and containing \mathcal{N} follows from the following lemma.

LEMMA 1. Any subspace \mathcal{W} satisfying (1), (2) and containing \mathcal{N} contains the reachable space of (*) from zero.

Proof. \mathcal{W} is a subspace of \mathcal{V}^* since it satisfies (1). Let us assume that the matrix V has been chosen in such a way that a submatrix W generates \mathcal{W} :

$$V = [L \quad W].$$

Since $\mathcal{N} \subset \mathcal{W}$ we may choose W such that N be a submatrix of it.

We may therefore partition V further in $W = (\tilde{M}, N)$ and therefore

$$V = [L \quad \tilde{M} \quad N],$$

with $M = [L \quad \tilde{M}]$. Now (5) and (6) give

$$(20) \quad FM = EMA, \quad FN = EMC, \quad G = EMB,$$

which, further partitioned according the above partition of M , gives

$$(21) \quad F\tilde{M} = ELA_{12} + E\tilde{M}A_{22}, \quad G = ELB_1 + E\tilde{M}B_2.$$

Now, by hypothesis, \mathcal{W} satisfies (1) and (2) so that there exist \tilde{A}_1, \tilde{A}_2 and \tilde{B} such that

$$(22) \quad F\tilde{M} = E\tilde{M}\tilde{A}_1, \quad FN = E\tilde{M}\tilde{A}_2, \quad G = E\tilde{M}\tilde{B}.$$

If we remember that $[EL \quad E\tilde{M}] = EM$ is injective, comparison of (21) and (22) yield

$$A_{12} = 0, \quad B_1 = 0.$$

This is the standard form for a system whose reachable space is contained in \mathcal{W} . \square

COROLLARY 1. The neutral space \mathcal{V}_* exists and is the reachable space of the pair (A, C) .

Proof. Apply Theorem 5 with $G = 0$. \square

Let now W be a submatrix of V generating \mathcal{W}_* , and let

$$(23) \quad W = [\hat{M} \quad N].$$

As before, there exists \hat{A} , \hat{B} and \hat{C} such that

$$(24) \quad F\hat{M} = E\hat{M}\hat{A}, \quad FN = E\hat{M}\hat{C}, \quad G = E\hat{M}\hat{B}.$$

If the system (*) is initialized at $y(0) \in \mathcal{W}_*$, we can always represent its solution as

$$(25) \quad \hat{x}(t+1) = \hat{A}\hat{x}(t) + \hat{B}u(t) + \hat{C}v(t),$$

$$(26) \quad y(t) = \hat{M}\hat{x}(t) + Nv(t),$$

and this constitutes a minimal representation of system (14) (15) (possibly with a feedback on v if we have changed of choice for \mathcal{M}). It can therefore be considered as a *minimal representation of* (*). It is unique up to a change of basis and a feedback on $v(\cdot)$.

3. Continuous time system. This short section is aimed at checking that all previous results, except Theorem 4 which is not important in the theory, carry over to the continuous case. We keep same notations and same numbers to the theorems.

Strict causality is taken to mean causality plus the fact that to a measurable input corresponds an absolutely continuous output.

3.1. Existence.

Proof of Theorem 1. (i) *Necessity.* Let \mathcal{V} be the subspace generated by those y 's that can be reached by the system. Necessarily, $\dot{y} \in \mathcal{V}$, therefore \mathcal{V} must satisfy (1) and, thus, be included in \mathcal{V}^* and (2).

(ii) *Sufficiency.* Perform exactly as in § 2.2 to end up with

$$\dot{\xi}(t) = \bar{A}\xi(t) + \bar{B}u(t).$$

3.2. Unicity.

Proof of Theorem 2. Unchanged, except for the substitution of an arbitrary measurable time function $v(\cdot)$ to the arbitrary sequence.

Proof of Theorem 3. The proof that two solutions corresponding to the same initial condition and the same control function $u(\cdot)$ differ at each time instant of an element of the image by V of the reachable space of the pair (A, C) is unchanged. The rest of the theorem relies on the next paragraph.

3.3 Minimality. Theorem 4 does not carry over in a simple way. One can prove that

$$y(t) \in y_0 + ty_1 + \cdots + \frac{t^{K-1}}{(K-1)!} y_{K-1} + \mathcal{W}^*,$$

where K is the smallest integer such that $\mathcal{W}^{K+1} = \mathcal{W}^k = \mathcal{W}^*$ and y_k is a sequence satisfying the homogeneous discrete system (16). The proof is a direct consequence of the remark that

$$\dot{y} \in \mathcal{W}^1 \Rightarrow y(t) \in y_0 + \mathcal{W}^1 \Rightarrow \dot{x} \in E^{-1}(F(\mathcal{W}^1 + x_0) + \mathcal{G}) \cap \mathcal{V}^* = \mathcal{W}^2 + x_1$$

and then iterating.

Proof of Theorem 5. Defining \mathcal{W}_* as previously, the algebraic constructions of § 2 remain the same. Moreover, classical system theory teaches us that given a pair $(A, [B \quad C])$ the reachable space is the same for the discrete time system and the continuous time system. Therefore \mathcal{W}_* is still the reachable space of the system. \square

Notice that without the parallel between continuous time and discrete time systems, Theorem 5 would be far less trivial in the continuous time case, since the identification of the limit of the recurrence \mathcal{W}_k with the reachable space relies on a direct study of the system (*).

The corollary carries over unchanged.

4. The nonstrictly causal case. We investigate here existence, unicity and representation of the solution of (*) when $y(t)$ is allowed to depend on past $u(s)$ and on $u(t)$. As for system (**), the same (algebraic) results hold if causality is defined via the existence of a proper transfer function (see § 5) since $y(\cdot)$ may now be non-differentiable. Deriving the results of § 2 from those of this section is obviously possible; however, this would hide the elementary character of § 2 and make the introduction of \mathcal{V}^* very artificial.

4.1. Existence.

THEOREM 5. *There exists a causal solution to (*) over any time interval for any sequence $u(\cdot)$ if and only if*

$$(27) \quad \mathcal{G} \subset E\mathcal{V}^* + F \operatorname{Ker} E = E\mathcal{M} + F \operatorname{Ker} E,$$

$$(28) \quad y_0 \in \mathcal{V}^* + \operatorname{Ker} E = \mathcal{M} + \operatorname{Ker} E.$$

Proof. (i) *Necessity.* Let us arbitrarily write

$$y(t) = z(t) + \varepsilon(t),$$

where

$$\varepsilon(t) \in \operatorname{Ker} E, \quad z(t) \in \mathcal{Z},$$

and \mathcal{Z} is a subspace that we shall choose later on. By an appropriate restriction, we can manage to have $\mathcal{Z} \cap \operatorname{Ker} E = \{0\}$ so that the above decomposition of y is unique. Equation (*) yields

$$(29) \quad Ez(t+1) = Fz(t) + F\varepsilon(t) + Gu(t),$$

so that, given $y(t)$ and $u(t)$, $Ez(t+1)$ is uniquely determined and also z once we restrict \mathcal{Z} to have no intersection with $\operatorname{Ker} E$.

By the same type of induction as in paragraph 2.1, we readily see that we must have

$$(30) \quad F\mathcal{Z} \subset E\mathcal{Z} + F \operatorname{Ker} E,$$

$$(31) \quad \mathcal{G} \subset E\mathcal{Z} + F \operatorname{Ker} E,$$

in order for (29) to have a solution $(z(t+1), \varepsilon(t))$ once $z(t)$, which depends upon the past, and $u(t)$ are given. The result then follows from the following fact.

LEMMA 2. *The largest subspace satisfying (30) is $\mathcal{V}^* + \operatorname{Ker} E$.*

Proof. Notice first that $\mathcal{V}^* + \operatorname{Ker} E$ satisfies (30). Now let \mathcal{Z} satisfy (30) and contain $\operatorname{Ker} E$ (since the maximal one does). Write

$$\mathcal{Z} = \mathcal{V} + \operatorname{Ker} E,$$

and

$$F(\mathcal{V} + \operatorname{Ker} E) \subset E\mathcal{V} + F \operatorname{Ker} E.$$

This implies that

$$\forall a \in \mathcal{V}, \quad \exists \tilde{a} \in \mathcal{V} \quad \text{and } b \in \operatorname{Ker} E \text{ such that } Fa = E\tilde{a} + Fb.$$

Clearly, \tilde{a} and b can be chosen depending linearly on a . Let therefore K generate $\text{Ker } E$. There exists a matrix of appropriate type, such that, for every $a \in \mathcal{V}$

$$Fa = E\tilde{a} + FKP a,$$

thus,

$$F(I - KP)a = E\tilde{a} = E(I - KP)\tilde{a}.$$

Let therefore

$$\tilde{\mathcal{V}} = (I - KP)\mathcal{V}.$$

Clearly,

$$\tilde{\mathcal{V}} + \text{Ker } E = \mathcal{V} + \text{Ker } E = \mathcal{L},$$

but also

$$F\tilde{\mathcal{V}} \subset E\tilde{\mathcal{V}},$$

so that

$$\tilde{\mathcal{V}} \subset \mathcal{V}^*, \quad \mathcal{L} \subset \mathcal{V}^* + \text{Ker } E.$$

This proves the lemma. Notice that to get the unicity of $z(t)$ we must choose $\mathcal{L} = \mathcal{M}$, a complement of $\text{Ker } E$ in \mathcal{V}^* .

(ii) *Sufficiency.* Let \tilde{K} be a matrix whose columns span $\text{Ker } E$ and M be as in § 2. Let

$$(32) \quad y(t) = Mx(t) + \tilde{K}w(t).$$

Condition (27) implies that there exist matrices \tilde{B} and \tilde{P} such that

$$G = EM\tilde{B} + F\tilde{K}\tilde{P}.$$

Now equation (*) can be written equivalently

$$(33) \quad EMx(t+1) = EMx(t) + F\tilde{K}w(t) + EM\tilde{B}u(t) + F\tilde{K}\tilde{P}u(t)$$

so that one possible solution of (*) is, using again (32),

$$(34) \quad x(t+1) = Ax(t) + \tilde{B}u(t),$$

$$(35) \quad y(t) = Mx(t) - \tilde{K}\tilde{P}u(t).$$

(Notice that A is defined using only E and F , as in § 2. However, since the requirement on G has been changed, one should not look for a relation between the matrices G , \tilde{B} , \tilde{P} of this section and G , \tilde{B} in the previous ones.) (34) and (35) together provide a causal solution and end the proof. \square

4.2. Unicity.

THEOREM 7. *The causal solution of (*) under conditions (27), (28) is unique for each sequence $u(\cdot)$ if and only if the pair (E, F) is column regular.*

Proof. We want to find under what conditions (33) has a unique solution $x(t+1)$, $w(t)$, once $x(t)$ and $u(t)$ are given. As a matter of fact, if this is true, since $x(0)$ is uniquely determined by $y(0)$, $x(1)$ and $w(0)$ will be unique and all succeeding y 's will be by induction.

By taking the difference $\delta x(t+1)$, $\delta w(t)$ between two solutions, we are led to the investigation of the nonzero solutions of

$$EM\delta x(t+1) = F\tilde{K}\delta w(t).$$

The only solution is zero if and only if

$$(36) \quad \text{Ker } F \cap \text{Ker } E = \{0\}$$

and

$$(37) \quad E\mathcal{M} \cap F \text{Ker } E = \{0\}.$$

This is so because EM is injective. Therefore, for a nonzero solution, either both sides are zero (but then (36) does not hold) or there is a nonzero element in $E\mathcal{M} \cap F \text{Ker } E$.

Notice that $\mathcal{V}^* = F^{-1}(E\mathcal{M})$, so that

$$(38) \quad \mathcal{N} = F^{-1}(E\mathcal{M}) \cap \text{Ker } E.$$

Now, it can easily be checked that for two subspaces \mathcal{A} and \mathcal{B} and an arbitrary linear operator F , one has

$$F\mathcal{A} \cap F\mathcal{B} = F[(\mathcal{A} + \text{Ker } F) \cap \mathcal{B}].$$

Apply this to (38), noticing that $F^{-1}(E\mathcal{M}) \subset \text{Ker } F$; it becomes

$$(39) \quad F\mathcal{N} = E\mathcal{M} \cap F \text{Ker } E.$$

Notice also that $\text{Ker } F \subset \mathcal{V}^*$, so that

$$(40) \quad \mathcal{N} \supset \text{Ker } E \cap \text{Ker } F.$$

From (39) and (40) we conclude that if (37) or (36) is violated, \mathcal{N} is nontrivial, i.e., the system is not C -regular.

Conversely, if \mathcal{N} is nontrivial and if, moreover, (36) holds, then since $\mathcal{N} \subset \text{Ker } E$, (36) implies

$$\mathcal{N} \cap \text{Ker } F = \{0\},$$

and therefore, $F\mathcal{N}$ has same dimension as \mathcal{N} and (39) shows that (37) is violated. \square

Remark 4. We may again make a distinction between two types of nonuniquity as in Remark 2. In the case (37) holds (but not (36)), the nonuniquity in y involves only $w(t)$ and does not propagate in time. The sequence $x(\cdot)$ is unique. The nonuniquity may be called "static". The dynamic nonuniquity is induced by nonzero elements in $E\mathcal{M} \cap F \text{Ker } E$.

The fact that the unicity condition is the same as in the strictly causal case will be more fully explained by the algebraic theory. It is not a trivial consequence of the fact that it is in both cases a study of nonzero solutions of (16) since y ranges over a larger subspace here.

4.3. Representation. Let us be more precise in representation (32), putting

$$\tilde{K} = [N \quad K]$$

(and with w having now a different meaning)

$$y(t) = Mx(t) + Nv(t) + Kw(t).$$

We also have (recalling that $F\mathcal{N} \subset E\mathcal{M}$)

$$G = EMB + FKP$$

so that (33) can now be written

$$EMx(t+1) = EM(Ax(t) + Bu(t) + Cv(t)) + FK(Pu(t) + w(t)).$$

But now,

$$\mathcal{R}(EM) \cap \mathcal{R}(FK) = \{0\},$$

since any part of $\text{Ker } E$ whose image by F is in EM belongs to \mathcal{V}^* , i.e., to \mathcal{N} , and moreover, since clearly $\text{Ker } F \subset \mathcal{V}^*$, FK is, as well as EM , injective. Therefore, the only solution is

$$w(t) = -Pu(t)$$

or defining $-KP = D$,

$$y(t) = Mx(t) + Du(t) + Nv(t), \quad x(t+1) = Ax(t) + Bu(t) + Cv(t).$$

These equations will be summarized further ((48) to (52)). Notice that those for the strictly causal case are identical to these where we set $D = 0$. Notice also that the same analysis applies to a representation of system (**).

5. Algebraic theory.

5.1. Generalized spectrum and regularity.

DEFINITION 7. We call a *generalized eigenvalue* of the pair (E, F) and associated *generalized eigenvector* a complex number z and a nonzero complex vector ξ of \mathbb{C}^m such that

$$(41) \quad (zE - F)\xi = 0.$$

LEMMA 3. Both the real part and imaginary part of a generalized eigenvector of (E, F) belong to \mathcal{V}^* . Under condition (2) this is also true of the first component (in \mathbb{R}^m) of a generalized eigenvector of the pair $([E \ 0], [F \ G])$.

Proof. Let

$$(42) \quad z = \sigma + i\omega, \quad \xi = \eta + i\zeta$$

be a generalized eigenvalue and eigenvector of (E, F) . Then (41) yields

$$F[\eta \ \zeta] = E[\eta \ \zeta] \begin{pmatrix} \sigma & \omega \\ -\omega & \sigma \end{pmatrix}.$$

Calling \mathcal{X} the subspace generated by $[\eta \ \zeta]$, this reads

$$F\mathcal{X} \subset E\mathcal{X},$$

and according to Proposition 1, this implies $\mathcal{X} \subset \mathcal{V}^*$, hence, the first claim. Keeping the notation (42), let $\varphi \in \mathbb{C}^p$:

$$\varphi = \chi + i\psi$$

constitute with ξ a generalized eigenvector of $([E \ 0], [F \ G])$:

$$(zE - F)\xi - G\varphi = 0.$$

Using (6) and separating again real and imaginary parts, we get

$$F[\eta \ \zeta] = E[\eta \ \zeta] \begin{pmatrix} \sigma & \omega \\ -\omega & \sigma \end{pmatrix} - EV\bar{B}[\chi \ \psi],$$

which gives

$$F\mathcal{X} \subset E\mathcal{X} + E\mathcal{V}^*.$$

Adding $F\mathcal{V}^*$ to the left and using (1), this gives according to Proposition 1 $\mathcal{X} + \mathcal{V}^* \subset \mathcal{V}^*$, and thus, $\mathcal{X} \subset \mathcal{V}^*$. \square

THEOREM 8. *The generalized spectrum of the pair (E, F) is finite if and only if this pair is C -regular (Definition 3). Otherwise, the generalized spectrum is the whole set \mathbb{C} and the rank defect of $zE - F$ is at least q for all z (where $q = \dim \mathcal{N}$).*

Proof. Equations (5) and (6) yield

$$(43) \quad EV(zI_n - \bar{A}) = (zE - F)V.$$

Assume (E, F) is C -regular. Then EV is injective. Let ξ be a generalized eigenvector; we know according to Lemma 3 that there exists a vector ν of \mathbb{C}^{n^*} such that $\xi = V\nu$. Placing this in (41) and using (43) gives

$$(44) \quad EV(zI - \bar{A})\nu = 0,$$

and since EV is injective, z has to be an eigenvalue of \bar{A} , of which there are at most n^* .

To the contrary, assume now the (E, F) is not C -regular. Then (43) yields, partitioning V and \bar{A} as in § 2.3,

$$(45) \quad (zE - F)[M \quad N] = EM[zI_n - A \quad -C].$$

The matrix $[zI - A \quad -C]$ has q fewer lines than columns. Therefore, it has for all z 's a kernel of dimension at least q . V being injective, it gives rise to a kernel of dimension at least q for $(zE - F)$. \square

This theorem is the justification for Definition 3. As a matter of fact, a *pencil of matrices* $(zE - F)$ is said to be column singular if its columns are not independent as polynomials in $\mathbb{R}^n[z]$, a characterization that coincides with Theorem 8. The degrees of the vectors of a polynomial minimal basis [6] of its kernel are called the Kronecker minimal column indices of the pencil. They are invariant under pencil similarity [7]. It also justifies the following definition.

DEFINITION 8. We call an *essential eigenvalue* of the pair (E, F) a complex number z such that

$$\text{rank}(zE - F) < m - q.$$

(It is a root of an invariant factor of the pencil (E, F) .)

As a corollary of Lemma 3 and Theorem 8, we have:

COROLLARY 1. q is the column rank defect of the matrix pencil $(zE - F)$, equal to m minus the size of the largest nonidentically null determinant in this matrix.

- If $r < m$, \mathcal{V}^* is never trivial, the system never C -regular, $q \geq m - r$.
- If $r = m$, \mathcal{V}^* is never trivial, the system is C -regular if and only if $\det(zE - F) \neq 0$.
- If $r > m$, \mathcal{V}^* is nontrivial if and only if the matrix $(zE - F)$ is reducible, i.e., all $m \times m$ determinants have a common root (for this value of z , the columns of $(zE - F)$ are not independent in \mathbb{C}^m). The system is C -regular if and only if one of the $m \times m$ determinants is not identically zero.

Proof. According to Lemma 3, if there exists a generalized eigenvector, \mathcal{V}^* is nontrivial. Conversely, if \mathcal{V}^* is nontrivial, (43) shows that (E, F) has generalized eigenvalues: those of \bar{A} at least. Now a generalized eigenvalue is clearly a complex number z such that the columns of $(zE - F)$ are not independent in \mathbb{C}^m , i.e., no $m \times m$ determinant is different from zero. And if the generalized spectrum of (E, F) is \mathbb{C} , all $m \times m$ determinants are null for all z 's, i.e., identically zero. \square

5.2. Invariants. We first recall a fact of system theory:

PROPOSITION 6. Let (A, C) be a (noncompletely controllable) system. A complete

set of invariants under the feedback group is given by:

- (i) the control invariants of the controllable part;
- (ii) the invariant factors of the uncontrollable part.

THEOREM 9. Given a pair (E, F) , the corresponding system (A, C) is entirely characterized by:

(i) the control invariants of the controllable part of (A, C) , which coincide with the Kronecker minimal column indices of the pencil $(zE - F)$.

(ii) the invariant factors of the uncontrollable part of A , which coincide with the finite invariant factors of the pencil $(zE - F)$.

Proof. Because of Propositions 3 and 6, the elements quoted for the pair (A, C) are indeed a complete set of invariants. There only remains to relate them to the corresponding quantities of the pencil $(zE - F)$.

(i) From Kalman [9] we know that the control invariants of the pair (A_{11}, C_1) are the minimal column indices of the pencil $[zI - A_{11} - C_1]$. From (45) it follows that they are the same as the minimal column indices of the pencil $[zI - A - C]$. As a matter of fact, let

$$\nu(z) = \begin{pmatrix} \nu_1(z) \\ \nu_2(z) \\ \mu(z) \end{pmatrix}$$

be a polynomial vector in $\text{Ker } [zI - A - C]$; this is equivalent to

$$(zI - A_{11})\nu_1(z) - A_{12}\nu_2(z) - C_1\mu(z) = 0, \quad (zI - A_{22})\nu_2(z) = 0.$$

However, $zI - A_{22}$ is a regular pencil, and therefore, $\nu_2(z)$ is identically null (since it is a polynomial, null for all z that are not in the spectrum of A_{22}). Thus, $[\nu'_1(z) \ \mu'(z)]'$ is in the kernel of $[zI - A_{11} - C_1]$. According to Lemma 3, all generalized eigenvectors, and therefore the basis vectors of $\text{Ker } (zE - F)$, can be written as

$$\xi(z) = V\nu(z).$$

Therefore, using (44) we see that to each $\xi(z)$ in $\text{Ker } (zE - F)$ corresponds a $\nu(z)$ in $\text{Ker } [zI - A - C]$ and conversely. Moreover, V being injective, $\xi(z)$ and $\nu(z)$ are of same degree.

(ii) We now show that essential eigenvalues of (E, F) are eigenvalues of A_{22} , with the rank defect of A_{22} equal to that of $(zE - F)$, minus q . Let λ be an essential eigenvalue of (E, F) with a corresponding kernel of dimension $q + k$. According to Lemma 3 and (44), $[\lambda I - A - C]$ has a kernel of dimension $q + k$ in \mathbb{R}^{n^*} , with $n^* = n + q$. Therefore, only $n - k$ of its lines are independent, and this is a fortiori true for $(\lambda I - A)$. Thus, λ is an eigenvalue of A , with an associated eigensubspace of dimension at least k . Now this property is independent of the particular choice of basis within \mathcal{V}^* , and thus, according to Proposition 3, invariant under feedback. Therefore, this eigenvalue and eigensubspace are associated to the uncontrollable part of A .

Conversely, considering the form (45) of (A, C) , we have seen that polynomial vectors in $\text{Ker } [zI - A - C]$ have a zero block in the uncontrollable part of the state space. Thus, to an eigenvalue of A_{22} , with an eigensubspace of dimension k , correspond k generalized eigenvectors (that we shall choose with zero blocks in the first and third parts), independent of each other and of any vector in $\text{Ker } (zE - F)$. Therefore, this complex number is an essential eigenvalue with a column rank defect at least $q + k$.

At this stage, we know that essential eigenvalues of (E, F) are eigenvalues of A_{22} and that the number of Jordan blocks associated to it coincide. There remains

to prove that they are identical in dimension. The technique is the same, using Jordan chains, and only heavier. We shall not go into too much detail. To a Jordan block of $(\lambda E - F)$ corresponds a Jordan chain $\xi_1, \xi_2, \dots, \xi_p$ satisfying

$$\begin{aligned}(\lambda E - F)\xi_1 &= 0, \\(\lambda E - F)\xi_2 &= E\xi_1, \\&\vdots \\(\lambda E - F)\xi_p &= E\xi_{p-1}.\end{aligned}$$

Here p is the size of the Jordan block. There remains to check that all the ξ_i 's are in \mathcal{V}^* and can be chosen independent of the vectors of $\text{Ker}(zE - F)$ at $z = \lambda$. Hence, there are $q + 1$ independent solutions to each of the above equations, and consequently, using a linear combination with total weight one, we can find one with a zero component in \mathcal{N} . Consequently, there corresponds to it a Jordan chain of $\lambda I - A$. Independence modulo $\text{Ker}(zE - F)|_{z=\lambda}$ in \mathbb{R}^m corresponds to independence in \mathbb{R}^n . Therefore, elementary divisors of $(zE - F)$ are elementary divisors of $(zI - A)$ fixed under feedback and, thus, according to Rosenbrock's feedback theorem, elementary divisors of $zI - A_{22}$. The converse proof goes exactly as above. \square

The particularization of the above results to the fact that the eigenvalues of A_{22} coincide with the essential eigenvalues of (E, F) leads to the following definition and corollary.

DEFINITION 9. The system (*) or (**), satisfying (2) or (27), is called *stable* if for every bounded input function $u(\cdot)$, there exists a bounded (causal) output function $y(\cdot)$ from any initial condition.

COROLLARY 1. *The implicit system is stable if and only if the essential eigenvalues of the pair (E, F) are stable (i.e., of modulus less than one or simple and of modulus unity in case (*) and of negative real part or simple imaginary in case (**)).*

Proof. If the condition of the corollary is met, the equivalent system is stabilizable with v with a linear feedback (or, equivalently, can be chosen stable). Therefore, there exists bounded solutions $x(\cdot)$ from any initial condition, with a choice of a bounded function $v(\cdot)$ (zero if the system is chosen stable). Therefore, $y(\cdot)$ as given by (15) or (49) remains bounded for these solutions.

To the contrary, if the condition is not met, there is a mode, uncontrollable with v , which is unstable. Therefore, except for a strict subspace of initial conditions, the solution $x(\cdot)$ will diverge for all choices of $v(\cdot)$. And since the matrix M is injective and has a range \mathcal{M} in direct sum with the range \mathcal{N} of N , $y(\cdot)$ as given by (15), or (49) recalling that $u(\cdot)$ is assumed bounded, will diverge as well for all (causal) solutions. \square

Remark 5. It is impossible to request, for singular systems, that all solutions be bounded in view of Theorem 3.

Remark 6. One may, of course, define in the same way asymptotically stable implicit systems.

Finally, one can clearly define the feedback group for systems (*) or (**) exactly in the same way as for an ordinary system. It clearly preserves existence of a strictly causal solution.

DEFINITION 10. The implicit system is *minimal* if the minimal subspace \mathcal{W}_* coincides with the characteristic subspace \mathcal{V}^* . Then (15) (16) is completely controllable. We have:

THEOREM 10. *Under condition (2), if the implicit system is minimal, a complete set of invariants under the feedback group is provided by the Kronecker minimal indices*

of the matrix pencil $[zE - F \quad -G]$, and they coincide with the control invariants of the system $(A, [B \ C])$.

Proof. The proof is in two steps. First check that the feedback group on the implicit system, combined with the nonunicity pointed out in Proposition 3, translates exactly in the classical feedback group for $(A, [B \ C])$ and that, conversely, the latter generates the former. This is an easy consequence of the fact that V is injective. We leave it to the reader to check. Then using the fact that $(A, [B \ C])$ is by hypothesis completely controllable and Kalman's theorem, we have that its control invariants are a complete set of invariants for the implicit system.

The second step is to identify the control invariants of $(A, [B \ C])$, i.e., according to Kalman [9], the column indices of $[zI - A \quad -B \quad -C]$ with the column indices of $[zE - F \quad -G]$. This is done in the same fashion as in Theorem 9 (i), using the second claim of Lemma 3 and

$$[(zE - F)[M \ N] \quad -G] = EM[zI - A \quad -C \quad -B]. \quad \square$$

5.3. Transfer functions.

THEOREM 11. *There exists a (strictly) causal solution to the system (*) or (**) if and only if there exists a (strictly) proper rational matrix $K(z)$ such that*

$$(46) \quad (zE - F)K(z) = G.$$

Let also $L(z)$ be a proper (not strictly) rational matrix of maximum rank, such that

$$(47) \quad (zE - F)L(z) = 0.$$

Then all solutions of the implicit system are given by

$$Y(z) = K(z)U(z) + L(z)V(z),$$

where $Y(z)$ and $U(z)$ are the z -transforms of $y(\cdot)$ and $u(\cdot)$, respectively, and $V(z)$ is an arbitrary power series of z^{-1} of appropriate dimension.

Proof. Notice first that there exist complex (column) vectors $l_i(z)$ satisfying (47) if and only if the pair (E, F) is not C -regular. It is easy to see (see [7]) that they can be chosen polynomial or, dividing each such column by the highest power of z present in it (since (47) is homogeneous), rational proper. If these degrees are chosen as small as possible, they are the column minimal indices or Kronecker indices of the pencil.

(i) *Necessity.* We know that, if a strictly causal solution exists, it is represented by (14), (15) or in the nonstrictly causal case by the following set (that coincides with the former if we set $D = 0$):

$$(48) \quad x(t+1) = Ax(t) + Bu(t) + Cv(t),$$

$$(49) \quad y(t) = Mx(t) + Du(t) + Nv(t)$$

with the definitions of the matrices A, B, C, D, M and N as

$$(50) \quad F[M \ N] = [EMA \ EMC],$$

$$(51) \quad G + FD = EMB,$$

$$(52) \quad ED = 0, \quad EN = 0.$$

Hence, the formula of the theorem for $Y(z)$ with

$$(53) \quad K(z) = D + M(zI - A)^{-1}B,$$

$$(54) \quad L(z) = N + M(zI - A)^{-1}C.$$

We can calculate

$$(55) \quad (zE - F)K(z) = (zE - F)M(zI - A)^{-1}B - FD.$$

Now (43) still holds with $V = [MN]$. Taking the first blocks in both sides, it comes

$$(56) \quad (zE - F)M(zI - A)^{-1} = EM,$$

and therefore, (55) with (51) yield (46). Similarly, we have with the second block in (43) (or with 50)

$$(zE - F)N = EMC$$

and this together with (56) yields (47).

(ii) *Sufficiency*. Assume the two proper rational matrices $K(z)$ and $L(z)$ exist, satisfying (46) and (47). Consider the rational matrix

$$(57) \quad H(z) = [K(z) \quad L(z)].$$

It can be realized according to standard realization theory, and we partition the last matrix according to the partition of H . There exist therefore matrices A , B , C , D , M and N such that

$$(58) \quad H(z) = [D \quad N] + M(zI - A)^{-1}[B \quad C],$$

and we may choose M , A , B and C such that the system $(M, A, [B \quad C])$ be minimal (i.e., completely controllable and observable). Take equality (46), which holds by hypothesis:

$$(zE - F)(D + M(zI - A)^{-1}B) = G.$$

Expand $(zI - A)^{-1}$ in a series in z^{-1} and equate like powers on both sides. It becomes

$$\begin{aligned} \text{power 1:} \quad & ED = 0, \\ \text{power 0:} \quad & EMB - FD = G, \\ \text{power } -k: \quad & (EMA - FM)A^{k-1}B = 0, \quad k = 1, \dots \end{aligned}$$

We do the same with (47). It becomes

$$\begin{aligned} \text{power 1:} \quad & EN = 0, \\ \text{power 0:} \quad & EMC - FN = 0, \\ \text{power } -k: \quad & (EMA - FM)A^{k-1}C = 0, \quad k = 1, \dots \end{aligned}$$

The “power 1” relations yield (52), “power 0” (51) and the second block of (50). The two “power- k ” together can be written

$$(EMA - FM)[[B \quad C] \quad A[B \quad C] \quad \dots \quad A^{n-1}[B \quad C]] = 0.$$

Since $(A, [B \quad C])$ is taken completely controllable, the right matrix in this equality is surjective, and therefore, we get the first part of (50). Straightforward calculation shows that the solutions (48) (49), subject to (50) (51) (52), satisfy (*) and similarly for the continuous case.

The strictly causal case is a specialization of this one with $D = 0$. \square

Notice also that the theorem yields

$$(59) \quad (zE - F)Y(z) = GU(z),$$

which is the direct z transform of (*) or Laplace transform of (**).

The rational matrix $H(z)$ of (57) can be considered as the generalized transfer function of the implicit (or generalized) system.

5.4. Canonical form. A change of coordinates on y amounts to a right multiplication by an invertible $m \times m$ matrix Q of both E and F . (In case one is interested in an output $Hy(t)$, H should be multiplied to the right by Q also.) The system is not changed either if we replace some or all of the r equations (*) or (**) by independent linear combinations of them, i.e., if we multiply to the left E , F and G by an invertible $r \times r$ matrix P .

Therefore, two implicit systems (H, E, F, G) and (H_1, E_1, F_1, G_1) , where H is an output matrix, will be said to be strictly equivalent if there exist two invertible matrices P and Q of appropriate dimension such that

$$(60) \quad \begin{aligned} H_1 &= HQ, \\ E_1 &= PEQ, & F_1 &= PFQ, \\ G_1 &= PG. \end{aligned}$$

Relations (60) are precisely the definition of equivalence of the pencils $(zE - F)$ and $(zE_1 - F_1)$. We know, therefore, that by a proper choice of matrices P and Q , $(zE - F)$ can be brought into the canonical form described; e.g., in [7].

Let $\alpha_1(z)$ be a polynomial vector of minimum degree, say ε_1 , such that

$$(zE - F)\alpha_1(z) = 0.$$

Let then $\alpha_2(z)$ be a polynomial independent of $\alpha_1(z)$ satisfying the same equality, and so on. The numbers $\varepsilon_1, \dots, \varepsilon_q$ are the column minimal indices. Performing similarly for E' and F' , we get the line minimal indices, say, η_1, \dots, η_l . The canonical form of $(zE - F)$ is block diagonal, made of four types of blocks.

(i) *Blocks L_{ε_i} :* To each ε_i , corresponds a block $\varepsilon_i \times \varepsilon_i + 1$ of the form

$$L_{\varepsilon_i} = \begin{pmatrix} z & -1 & 0 & \cdots & 0 & 0 \\ 0 & z & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & z & -1 \end{pmatrix}.$$

In this basis we obviously have

$$\alpha_i(z) = \begin{pmatrix} 1 \\ z \\ \vdots \\ z^{\varepsilon_i} \end{pmatrix}.$$

We make correspond to it the column $l_i(z)$ of $L(z)$:

$$l_i(z) = \begin{pmatrix} z^{-\varepsilon_i} \\ z^{-\varepsilon_i+1} \\ \vdots \\ 1 \end{pmatrix}.$$

This makes up the matrix $L(z)$ of (47).

Writing equations (*) with this special form for E and F , we immediately see that each such block involves $\varepsilon_i + 1$ coordinates of y . They always have a solution whatever

the coefficients of G in the same lines, and the last coordinate of this subvector of y is free. It corresponds to a coordinate in \mathcal{N} , the ε_i first corresponding to coordinates in \mathcal{V}_* .

As a matter of fact, L_{ε_i} has the rational strictly proper right inverse

$$L_{\varepsilon_i}^r = \begin{pmatrix} z^{-1} & z^{-2} & \cdots & z^{-\varepsilon_i} \\ 0 & z^{-1} & \cdots & z^{-\varepsilon_i+1} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & z^{-1} \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

so that whatever the corresponding lines of G , (46) will have a strictly proper solution for this block, obtained by multiplying these lines of G par $L_{\varepsilon_i}^r$ to the left.

(ii) *Blocks L_{η_i} .* To the row indices η_i correspond blocks L_{η_i} of type $\eta_i + 1 \times \eta_i$ having the form of the transpose of a block L_e .

Writing equations (*) with this block, we see that it involves η_i coordinates of y but that the last line amounts to a recurrence relation between the elements of the sequence $u(\cdot)$. It can be satisfied for all sequences only if the corresponding lines of G are all zero, but then all these coordinates must be and remain zero. They correspond to coordinates in a complement of \mathcal{V}^* in \mathbb{R}^m , and the requirement on G is (part of) condition (2).

Correspondingly, it is a simple task to see, thanks to the triangular form of L_{η_i} , that (46) can be satisfied with a strictly proper block in $K(z)$ if and only if the corresponding lines of G are null, the solution being then zero.

(iii) *Blocks L_{μ_k} .* These are square blocks of type $\mu_k \times \mu_k$ corresponding to the infinite invariant factors of the pencil $(zE - F)$. They are of the form

$$L_{\mu_k} = \begin{pmatrix} -1 & z & 0 & \cdots \\ 0 & -1 & z & \\ \vdots & & \ddots & \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

Again, writing the equations (*) for this block, we see that they involve μ_k coordinates of y , but depending in an anticausal way on the sequence $u(\cdot)$. Therefore, these coordinates also correspond to a complement of \mathcal{V}^* in \mathbb{R}^m , and the corresponding rows of G must be zero for a strictly causal solution to exist.

However, the dependence of y on $u(\cdot)$ is anticausal but not strictly. Therefore, a causal but not strictly causal solution may exist where the first coordinate of the corresponding subvector of y is nonzero but all others zero. The same row in G may be nonzero. This corresponds to the fact that E has a column of zeros in the first column of L_{μ_k} , and the corresponding coordinate of y is therefore in $\text{Ker } E$ but not in \mathcal{N} . We recover conditions (28) and (27).

A complete information is given again looking at (46). As a matter of fact, L_{μ_k} is invertible:

$$L_{\mu_k}^{-1} = \begin{pmatrix} -1 & -z & \cdots & -z^{\mu_k-1} \\ 0 & -1 & \cdots & -z^{\mu_k-2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & -1 \end{pmatrix}$$

so that the only solution of (46) for this block is $L_{\mu_k}^{-1}G_{\mu_k}$, which is anticipatory of $\mu_k - 1$

steps unless G_μ has some zero rows. If its first row is the only nonzero one, then the corresponding block of $K(z)$ is proper but not strictly.

(iv) *Blocks* L_{λ_i} . These are square blocks that together constitute a characteristic matrix

$$zI - A_\lambda,$$

where A_λ is in Jordan form, for example. This clearly corresponds to coordinates of y for which there is a unique strictly causal solution. They are therefore in \mathcal{V}^* but in a complement of \mathcal{V}_* . The corresponding block of $K(z)$ is $(zI - A_\lambda)^{-1}$. The corresponding eigenvalues are the essential eigenvalues of the pair (E, F) .

Remark 7. This kind of link between geometrical concepts and the system pencil was shown for standard systems in Jaffe and Karcanias [8]. Using their characterization our space \mathcal{V}^* appears as a generalization of (A, C) invariant subspaces since it is characterized by the fact that $(zE - F)V$ has only column minimal indices and finite invariant factors. This also clearly shows how to investigate the impulsive behavior of our system (**), or noncausal behavior of (*), by looking at the infinite invariant factors and the associated subspaces.

6. Conclusion. We have a simple theory of singular implicit systems whether they are square, or over- or underdetermined. It should be noted that overdetermination may go along with nonunicity of the solution in a nontrivial way.

The recurrences defining the various subspaces \mathcal{V}^* , \mathcal{W}^* , \mathcal{W}_* , \mathcal{V}_* , provide the basis for finite algorithms, unfortunately rather ill-behaved in terms of robustness in their native form. They involve finding zero determinants and computing right or left inverses, numerically difficult operations. Standard techniques could be applied to improve them (like computing the rank of AA^* , or A^*A , instead of A).

The stage seems to be set to extend a significant part of Rosenbrock's theory to these systems and of its modern developments, in the spirit of Wolovich or Fuhrman. Also, the study of impulsive (or noncausal) behavior seems to be straightforward, using the literature on that topic.

A domain of interest is naturally the use of tools of two-player control systems theory to study the property of implicit systems: making an output sequence unique (decoupling $v(\cdot)$ through feedback), ensuring that all trajectories meet a given subspace at a given instant (capturing the state), or that some do (controllability through v), insuring that all trajectories will do better than a given amount with respect to some criterion (dynamical games), etc.

REFERENCES

- [1] J. D. APLEVICH, *Time domain input output representations of linear systems*, Automatica, 18 (1981) pp. 509-522.
- [2] P. BERNHARD, *Sur la commandabilité des systèmes dynamiques discrets à deux joueurs*, RAIRO, J3 (1972), pp. 53-68.
- [3] S. L. CAMPBELL, *Limit behaviour of solutions of singular difference equations*, Linear Algebra and Appl., 23 (1979), pp. 167-179.
- [4] S. L. CAMPBELL, C. D. MEYER AND N. J. ROSE, *Applications of the Drazin inverse to linear systems of differential equations with singular constant coefficients*, SIAM J. Appl. Math., 31 (1976), pp. 411-425.
- [5] J. J. M. EVERS, *Linear Programming Over an Infinite Horizon*, Tilburg University Press, Academic Book Service of Holland, 1973.
- [6] G. D. FORNEY, JR., *Minimal bases of rational vector spaces, with applications to multivariable linear systems*, this Journal, 13 (1975), pp. 493-520.

- [7] F. R. GANTMACHER, *Théorie des matrices*, tome 2, Dunod, Paris, 1966 (translated from the Russian).
- [8] S. JAFFE AND N. KARCANIAS, *Matrix pencil characterization of almost (A, B) invariant subspaces, a classification of geometric concepts*, Internat. J. Control, 33 (1981), p. 51.
- [9] R. E. KALMAN, *Kronecker invariants and feedback*, Ordinary Differential Equations, 1971 NRL-MRC Conference, L. Weiss, ed., Academic Press, Paris, 1972.
- [10] N. KARCANIAS AND G. E. HAYTON, *State space and transfer function invariant infinite zeros: A unified approach*. Joint Automatic Control Conference, TA4C, Charlottesville, VA, 1981.
- [11] ———, *Generalised autonomous dynamical systems, algebraic duality and geometric theory*. 8th World IFAC Congress, Kyoto, Japan, 1981.
- [12] D. G. LUENBERGER, *Dynamic systems in descriptor form*, IEEE Trans. Automat. Control, AC 22 (1977), pp. 312–321.
- [13] ———, *Time invariant descriptor systems*, Automatica, 14 (1978), pp. 473–480.
- [14] H. H. ROSENBRICK, *State-space and Multivariable Theory*, Nelson, London, 1970.
- [15] D. N. STENGEL, R. E. LARSON, D. G. LUENBERGER AND T. B. CLINE, *A descriptor variable approach to modeling and optimization of large scale systems*, Proc. of the Engineering Foundation Conference on System Engineering for Power: Organizational Forms for Large Scale Systems, vol. 7. Davos, Switzerland, 1979.
- [16] P. VAN DOOREN, G. VERGHESE AND T. KAILATH, *Properties of the system matrix of a generalized state space system*, IEEE 1977 Decision Control Conference, San Diego, CA, 1979.
- [17] G. VERGHESE, *Infinite frequency behavior in generalized dynamical systems*, Ph.D. dissertation, Dept. of Electrical Engineering, Stanford University, Stanford, CA, 1978.
- [18] ———, *Further notes on singular descriptions*, Joint Automatic Control Conference, TA4B, Charlottesville, VA, 1981.
- [19] G. VERGHESE AND T. KAILATH, *Rational matrix structure*, IEEE Trans. Automat. Control AC-26 (1981), pp. 434–439.
- [20] G. VERGHESE, B. LEVY AND T. KAILATH, *Generalized state space for singular systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 811–831.
- [21] J. H. WILKINSON, *Linear differential equations and Kronecker's canonical form*, Recent Advances in Numerical Analysis, C. de Boor and G. Golub, eds., Academic Press, New York, 1979. (Proceedings of a conference held in Madison, WI, 1978.)

SOME COMMENTS ON POSITIVE ORTHANT CONTROLLABILITY OF BILINEAR SYSTEMS*

WILLIAM M. BOOTHBY†

Abstract. Consider the bilinear system $\dot{x} = (A + uB)x$, $x \in \mathbb{R}^n$, and u unrestricted. The system has the property that any solution $x(t)$ with $x(0) \geq 0$ (i.e., all components of x nonnegative) will remain in the positive orthant, $R_+^n = \{x \in \mathbb{R}^n | x \geq 0\}$ for $0 \leq t < \infty$ if and only if B is diagonal and $A = (a_{ij})$ has the property that $a_{ij} \geq 0$ if $i \neq j$. In this note the controllability of solutions from a point of R_+^n to another such point is studied. Some results are given for arbitrary $n > 0$ and detailed results are presented for the case $n = 2$.

1. Introduction. In what follows we consider bilinear systems

$$(*) \quad \dot{x} = (A + uB)x,$$

with $x \in \mathbb{R}^n$ and A, B real $n \times n$ matrices with u denoting an admissible control function, which here will mean a piecewise constant or piecewise continuous function on $[0, \infty)$ into \mathbb{R} .

We will let $x \geq 0$, $(x > 0)$, $A \geq 0$, $(A > 0)$, etc., mean that the given vector or matrix has only nonnegative components (resp. only positive components). The set $R_+^n = \{x \in \mathbb{R}^n | x \geq 0\}$ is called the *positive orthant*; its interior is $\overset{\circ}{R}_+^n = \{x \in \mathbb{R}^n | x > 0\}$. An $n \times n$ matrix A which satisfies $A \geq 0$ clearly has the property $A(R_+^n) \subset R_+^n$, but what is more important for us is the following proposition, which is well known and is easy to verify.

PROPOSITION 1.1. *Let $A = (a_{ij})$ be an $n \times n$ matrix such that $a_{ij} \geq 0$ when $i \neq j$, and let $x(t)$ be a solution of $\dot{x} = Ax$. If $x(0) \in R_+^n$, then for $t \geq 0$ $x(t) \in R_+^n$. Conversely, if some off-diagonal element is negative, there exists a solution which leaves the positive orthant.*

We will find it convenient to call such a matrix A *essentially nonnegative*, or, if the inequalities are strict, *essentially positive*. It is well known that $e^{At} \geq 0$ (resp. > 0) for all $t \geq 0$ if and only if A is essentially nonnegative (resp. positive), see, for example, Bellman [1, p. 176].

We are interested in those bilinear systems $(*)$ which have the property that no matter how the controls are chosen, any solution $x(t)$ which lies in R_+^n at $t = 0$ will remain in R_+^n for all $t \geq 0$. If the controls are unrestricted, e.g., u can take on negative values, then it is clear that a necessary and sufficient condition is as follows.

(1.2). *For unrestricted controls a necessary and sufficient condition for the statement:*

$$x(0) \geq 0 \text{ implies } x(t) \geq 0 \text{ for } t \geq 0$$

to hold is that A be essentially nonnegative and B be a diagonal matrix.

There are many examples of such systems; several are given in the next section. The question we wish to investigate is the following: given a system $(*)$ with A essentially nonnegative and B diagonal and given $x_0 \neq 0$, $x_1 > 0$ in R_+^n , can we choose controls so that there is a solution of $(*)$ with $x(0) = x_0$ and $x(T) = x_1$ for some $T > 0$? If so, we call this property *positive orthant controllability*. We shall henceforth restrict ourselves to the subclass of systems $(*)$ for which this question is meaningful, i.e., A essentially nonnegative and B diagonal. In fact, for the most part we consider only

* Received by the editors April 23, 1981, and in final revised form November 9, 1981.

† Department of Mathematics, Washington University, St. Louis, Missouri 63130. This work was supported by the National Science Foundation under grant ENG 78 22166. Part of these results were presented at the Mathematical Systems Theory Meeting, University of Warwick (England), July 7–11, 1980.

the generic case, an open subset of these (A, B) , for which A is essentially positive, B is nonsingular and $b_i - b_j$ are distinct for $i \neq j$. This latter is a natural condition on B in terms of Lie theory—it is a requirement on $\text{ad}(B)$. It also appears in the Jurdjevic–Kupka study of bilinear systems [5].

2. An example. Given the facts that (1) a very large class of systems are locally approximable by bilinear systems (Krener [7]) (which are in fact the simplest nonlinear systems) and that (2) in many situations it is natural to assume that the variables which describe the state of the system are positive (populations of various species, prices of goods, masses of chemical reagents, probabilities, etc., see Luenberger [9], for example), the questions studied in this note seem to be quite natural. There is also some theoretical interest which stems from the fact that the situation considered here—given local accessibility—is in contrast to that studied in the important work of Jurdjevic and Kupka [5]. There the matrix B satisfied conditions similar to those above ($b_i - b_j$ all distinct), but rather different conditions on A , in particular $a_{1n}a_{n1} < 0$, which preclude $a_{ij} > 0$ for $i \neq j$ but result in controllability on all of $R^n - \{0\}$. This, of course, is in marked contrast to the case considered here where the accessible set of a point in the positive orthant must lie in the positive orthant. This might cast light on necessary conditions for controllability of bilinear systems on R^n .

Probably the most important model which is well known and which satisfies the conditions we have imposed arises in the case in which one is studying estimation theory of various stochastic processes. In [2], for example, Brockett and Clark have shown the importance of determining the accessibility set for the control problem

$$\dot{p} = (A - \tfrac{1}{2}B^2)p + uBp, \quad p(0) = p_0$$

in studying an unnormalized conditional density equation

$$dp = Ap \, dt + Bp \, dy$$

rewritten in Fisk–Stratonovich form as

$$\dot{p} = (A - \tfrac{1}{2}B^2)p \, dt + bp \, d^+y.$$

Here, the vector $p(t)$ takes values in the positive orthant.

We also mention that in [3] Brockett studied in § 3 the reachability set for bilinear systems rather closely related to, but not quite the same as those studied here. Finally, it was pointed out to the author by L. Markus that matrices satisfying conditions similar to those discussed here are widely used in economics, see Markus [10], for example. However, control problems do not seem to have been considered in this context.

3. Some generalities. It is well known from the work of several authors that a necessary condition for controllability of any system of the form $\dot{x} = f(x, u)$, where $f(x, u)$ is a vector field on a manifold depending analytically on controls u as well as the point x , can be given in terms of the rank of Lie algebra \mathfrak{g} generated by these vector fields. The necessary condition is that for each point x of the manifold the vectors of the Lie algebra must span the tangent space at x . The condition is not usually sufficient; however, it is enough to insure that the set of points accessible from x in positive time has a nonempty interior, in whose closure it is contained [6], [12].

For our case where the manifold may be taken to be \dot{R}_+^n and $f(x, u) = (A + uB)x$, $u \in R$, it follows easily from Lie theory that generically this maximum rank condition is satisfied, i.e., for a very large (open and dense) subset of pairs A, B .

We require that A be essentially positive, i.e., all off-diagonal elements of A be positive and that the diagonal matrix B satisfy the following condition for all $i \neq j$,

(3.1). $b_i - b_j = b_k - b_l$ implies $i = k$ and $j = l$ (in particular $b_i - b_j \neq 0$).

We then easily see that the following is true.

PROPOSITION 3.2. *If A, B are $n \times n$ matrices with A essentially positive and B a diagonal matrix satisfying (3.1), then the Lie algebra \mathfrak{g} generated by the set $\{A + uB \mid u \in \mathbb{R}\}$ is $\mathfrak{sl}(n, \mathbb{R})$ if $\text{tr } A = 0$ and $\mathfrak{gl}(n, \mathbb{R})$ otherwise.*

This follows from the facts that B is regular and that A has a nonzero component in each eigenspace belonging to a nonzero eigenvalue of $\text{ad } B$; hence A, B generate $\mathfrak{sl}(n, \mathbb{R})$. From this it is clear that if either $\text{tr } B \neq 0$ or $\text{tr } A \neq 0$, then the algebra generated has dimension n^2 and is thus all of $\mathfrak{gl}(n, \mathbb{R})$.

4. Characteristic roots and vectors. If C is an essentially positive $n \times n$ matrix, let d be its smallest diagonal element. If $d > 0$, then $C > 0$, i.e., C is a positive matrix. If $d \leq 0$, then $C + pI$, $p > |d|$, is a positive matrix whose invariant subspaces, if any, are the same as those of C and whose spectrum is the translate by p of that of C , i.e., λ is a characteristic value of C with characteristic vector v if and only if $\lambda + p$ is a characteristic value of $C + pI$ with the same characteristic vector v . It follows from the Perron–Frobenius theory of positive matrices that the following facts hold:

(4.1). *C has a real characteristic value r to which there corresponds a positive characteristic vector $v > 0$. The value r is a simple root of the characteristic equation. If $w > 0$ is any other positive characteristic vector, then w is a multiple of v . The root r corresponds to the dominant root $r + p$ of $C + pI$; hence any other characteristic root λ satisfies the inequality $r + p > |\lambda + p|$.*

Since for each real u , $A + uB$ is essentially positive if A is essentially positive, then (4.1) holds. As noted earlier, we assume throughout this note that A is essentially positive, i.e., all off-diagonal elements are strictly positive. This is the generic case (and already involves enough problems!). Although $A + uB$ is essentially positive for each choice of u , there may or may not be choices of u such that $A + uB$ is a positive (or even nonnegative) matrix. Let $a_{ii} + ub_i$ denote the i th diagonal element of $A + uB$. Graphically we consider n lines $v = a_{ii} + ub_i$ on the uv -plane. Whether or not there is a nonempty set of values of u such that all diagonal elements are positive depends on the intercept a_{ii} and slope b_i of these lines. Of course, for u large, it is the slopes b_i which are important. Thus, if b_1, \dots, b_n all have the same sign, then for some choice of u , $A + uB > 0$. In this case, as we can see from Gershgorin's theorem, for suitable choice of u , we can make the (dominant) characteristic value r associated with the characteristic vector $v > 0$ mentioned in (4.1) take on a positive value, $r > 0$, or a negative value, $r < 0$. To see this we first quote the theorem.

THEOREM 4.2 (Gershgorin). *Let $C = (c_{ij})$ be an $n \times n$ matrix and for each i , $1 \leq i \leq n$, let $\rho_i = \sum_{j \neq i} |c_{ij}|$. Let \mathcal{D}_i given by $\mathcal{D}_i = \{z \in \mathbb{C} \mid |z - a_{ii}| \leq \rho_i\}$ be a disk in the complex plane. Then each characteristic root is in one of the disks \mathcal{D}_i ; and if the disks are pairwise disjoint, then there is exactly one in each \mathcal{D}_i . In particular, for a real matrix, if the \mathcal{D}_i are pairwise disjoint, the characteristic roots are real and distinct.*

From this we can derive the following proposition for our case.

PROPOSITION 4.3. *If the b_i , $i = 1, \dots, n$, are pairwise distinct, then we may find values of u such that $A + uB$ has n distinct real characteristic values. If the b_i additionally all have the same sign, say positive, then for all sufficiently large u (resp. sufficiently negative u), the characteristic values are all positive (resp. all negative).*

Proof. If $\rho_i(C)$ denotes the radius, $\rho_i = \sum_{j \neq i} c_{ij}$ of the i th disk \mathcal{D}_i of the matrix C , then note that $\rho_i(A + uB) = \rho_i(A)$ for $i = 1, \dots, n$. On the other hand, the center of

the disk \mathcal{D}_i for $A + uB$ is $a_{ii} + ub_i$, and hence, $|(a_{ii} + a_{jj}) + u(b_i - b_j)|$ is the distance between centers. This distance is at least equal to $|u||b_i - b_j| - |a_{ii} - a_{jj}|$ and can be made as large as we wish by choosing $|u|$ large enough. Since the radii ρ_i are independent of u , we see that taking $|u| > |u_0|$ for some $u_0 > 0$ will guarantee that for all $i \neq j$, $\mathcal{D}_i \cap \mathcal{D}_j = \emptyset$. This proves the first statement. The second follows from this since we can make all $a_{ii} + ub_i$ positive, or all of these diagonal elements negative, by a suitable choice of u .

We remark that if the b_i do not all have the same sign, even though they are pairwise unequal, we may not be able in some cases to change the sign of the dominant characteristic value v by our choice of u .

Example 4.4.

$$A = \begin{pmatrix} 1 & 2 \\ 4 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} -1 & 0 \\ 0 & +1 \end{pmatrix}, \quad A + uB = \begin{pmatrix} 1-u & 2 \\ 4 & -1+u \end{pmatrix}.$$

A has characteristic values $\lambda_1 = 3$ and $\lambda_2 = -3$ for which corresponding characteristic vectors are, say $(1, 1)'$ and $(-1, 2)'$. In general, the characteristic polynomial for $A + uB$ is

$$f(t) = t^2 - [(u-1)^2 + 8],$$

and hence, the dominant characteristic root is positive for all u . However, the slope of the corresponding positive characteristic vector or direction is given by $\frac{1}{2}\{\sqrt{(u-1)^2 + 8} + (u-1)\}$, which goes from 0 to ∞ as u goes from $-\infty$ to $+\infty$. Thus, for a fixed value of u , a typical picture of the flow lines of $\dot{x} = (A + uB)x$ is shown in Fig. 1. By varying u the characteristic vector v can be made to take on any direction in the open positive quadrant. It would appear that since the flow always tends outward in the quadrant, positive quadrant controllability would not be possible in this case. However, it is possible by varying u to make the vector $(A + uB)x$ for x on the x_1 -axis or on the x_2 -axis take any direction pointing into the quadrant, which makes it somewhat less clear that controllability cannot be achieved in this case. This is settled by constructing a Lyapunov function in § 6.

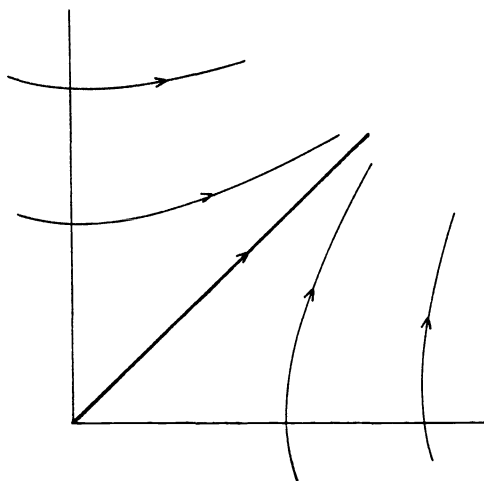


FIG. 1

5. The two-dimensional case. As examples of some possible qualitatively distinct types of behavior of solutions of bilinear systems satisfying our conditions, we consider the $n = 2$ case, where it is possible to distinguish several different patterns determined by open conditions on the values of $A = (a_{ij})$ and b_1, b_2 . We shall suppose that $a_{12} > 0$ and $a_{21} > 0$ (A essentially positive) and B nonsingular with $b_1 \neq b_2$. For convenience of notation, let $A_u = A + uB$, then $A_0 = A$. The characteristic polynomial of A_u is then

$$(5.1) \quad f_u(t) = t^2 - (\operatorname{tr} A_u)t + \det A_u,$$

and its discriminant is $D = (\operatorname{tr} A_u)^2 - 4 \det A_u$, which reduces to

$$(5.2) \quad D = [(b_1 + b_2)u + (a_{11} + a_{22})]^2 + 4a_{12}a_{21}.$$

Thus $D > 0$, and for each u A_u has two distinct real roots, say λ_1, λ_2 with $\lambda_1 > \lambda_2$. In fact, $\lambda_1, \lambda_2 = \frac{1}{2}[a_{11} + a_{22} + u(b_1 + b_2) \pm \sqrt{D}]$ with the $+$ corresponding to the dominant root λ_1 . An easy computation shows that the characteristic vectors belonging to these roots lie on the two lines

$$(5.3) \quad x_2 = \frac{1}{2a_{12}}[(a_{22} - a_{11}) + u(b_2 - b_1) \pm \sqrt{D}]x_1.$$

Let $a_{22} - a_{11} + u(b_2 - b_1) = p(u)$, then $D = (p(u))^2 + 4a_{12}a_{21}$ so the slope of these lines is

$$(5.4) \quad \frac{1}{2a_{12}}[p(u) \pm \sqrt{(p(u))^2 + 4a_{12}a_{21}}],$$

with $+$ corresponding to the dominant root λ_1 and $-$ to λ_2 . It is clear that the slope of the line corresponding to λ_1 is positive and the slope of the line corresponding to λ_2 is negative. Moreover, by varying u from $-\infty$ to $+\infty$, the slope of the first line goes through all values in the interval $(0, \infty)$.

Even in the two-dimensional case, we cannot give a universal criterion for controllability—our more modest goal is to show the existence of an open (i.e., large) subset in the space of pairs (A, B) in which we have positive orthant controllability and another open subset on which we do not have controllability. For the latter case the ideas of this section suggested a result which holds for arbitrary n (Theorem 6.1), so the discussion is postponed to § 6.

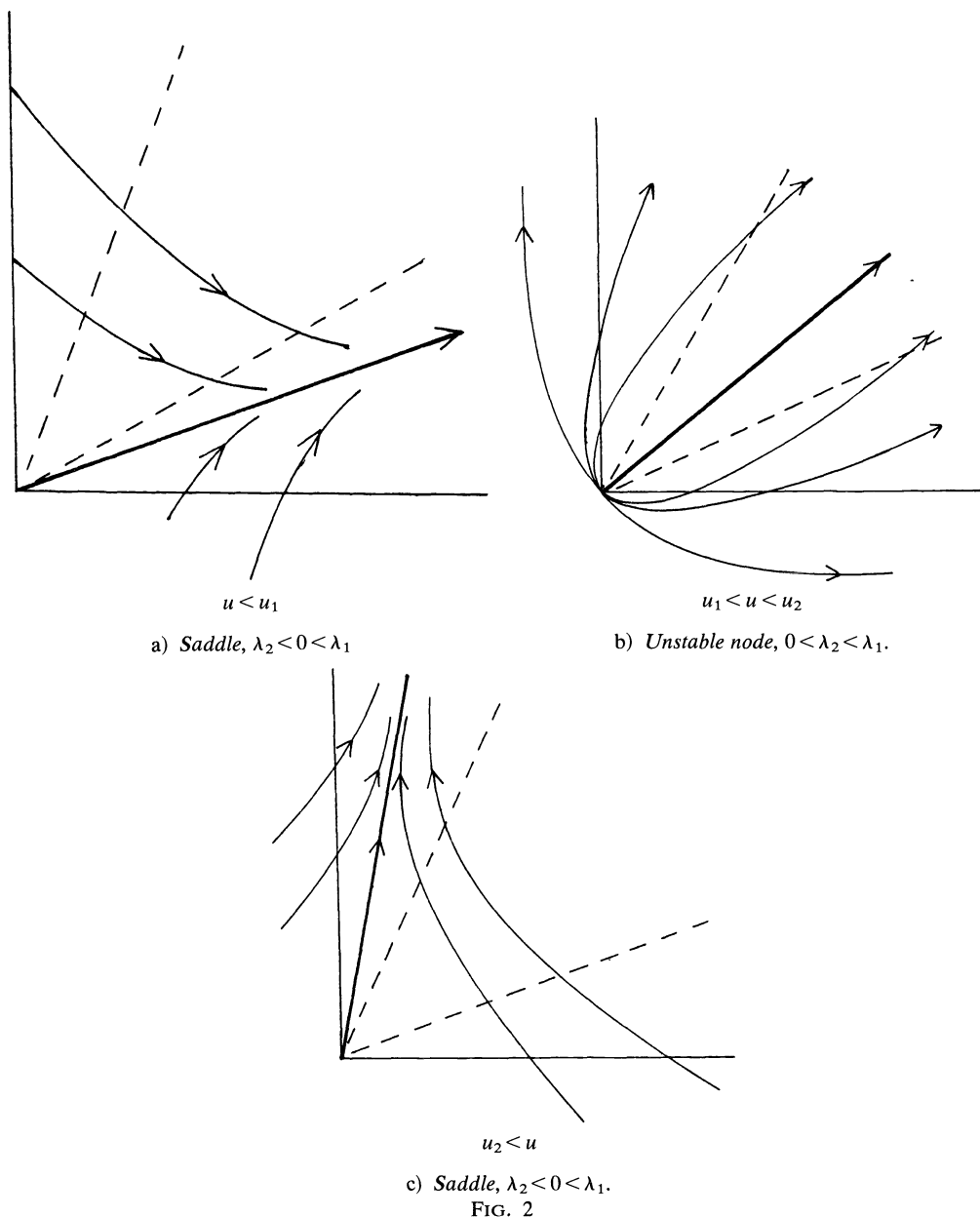
Intuitively, it appears that there is a greater possibility of controllability if the dominant root λ_1 changes sign at least once as u varies from $-\infty$ to $+\infty$. This can only happen if the quadratic polynomial $\det(A + uB) = \lambda_1\lambda_2$ has at least one real root. For this reason we will restrict to the open subset of (A, B) on which its discriminant \mathcal{D} is positive, where

$$(5.5) \quad \mathcal{D} = (a_{11}b_2 + a_{22}b_1)^2 - 4b_1b_2 \det A.$$

When $\mathcal{D} > 0$ it is easily verified that exactly one of the following cases occurs as u goes from $-\infty$ to $+\infty$.

- (a) λ_1 is always positive, λ_2 goes from negative to positive, then back to negative;
- (b) λ_2 is always negative but λ_1 goes from positive to negative, then back to positive;
- (c) λ_1 and λ_2 have the same sign and both change to the opposite sign.

In all cases let u_1, u_2 denote the u -values at which the sign change occurs, i.e., at which $\lambda_1\lambda_2 = \det(A + uB) = 0$. We illustrate these cases with figures in which the characteristic vectors corresponding to λ_1 when $u = u_1$ and $u = u_2$ are shown by dashed



lines. The flow patterns corresponding to typical u in the intervals $(-\infty, u_2)$, (u_1, u_2) and $(u_2, +\infty)$ are shown in each case, the first case (a) in Figs. 2a, 2b and 2c. Of the three cases, this one seems to be the poorest candidate for positive controllability. However, I do not have a proof that it is not controllable.

In case (b) we have two positive characteristic vectors (on the dashed lines) for which $\lambda_1 = 0$ with $\lambda_1 < 0$ for characteristic vectors between them. The pattern of flow lines for $u < u_1$, $u_1 < u < u_2$ and $u_2 < u$ are shown in Figs. 3a, 3b, 3c.

In this case it can be seen that we have controllability in the first quadrant. To see this we must show that $x_0 \in R_+^2$, $x_0 \neq 0$, can be steered to any interior point x_1 of the first quadrant. The dotted lines divide the quadrant into three sectors with flow patterns as shown in Fig. 3, where u is chosen so that the dominant characteristic

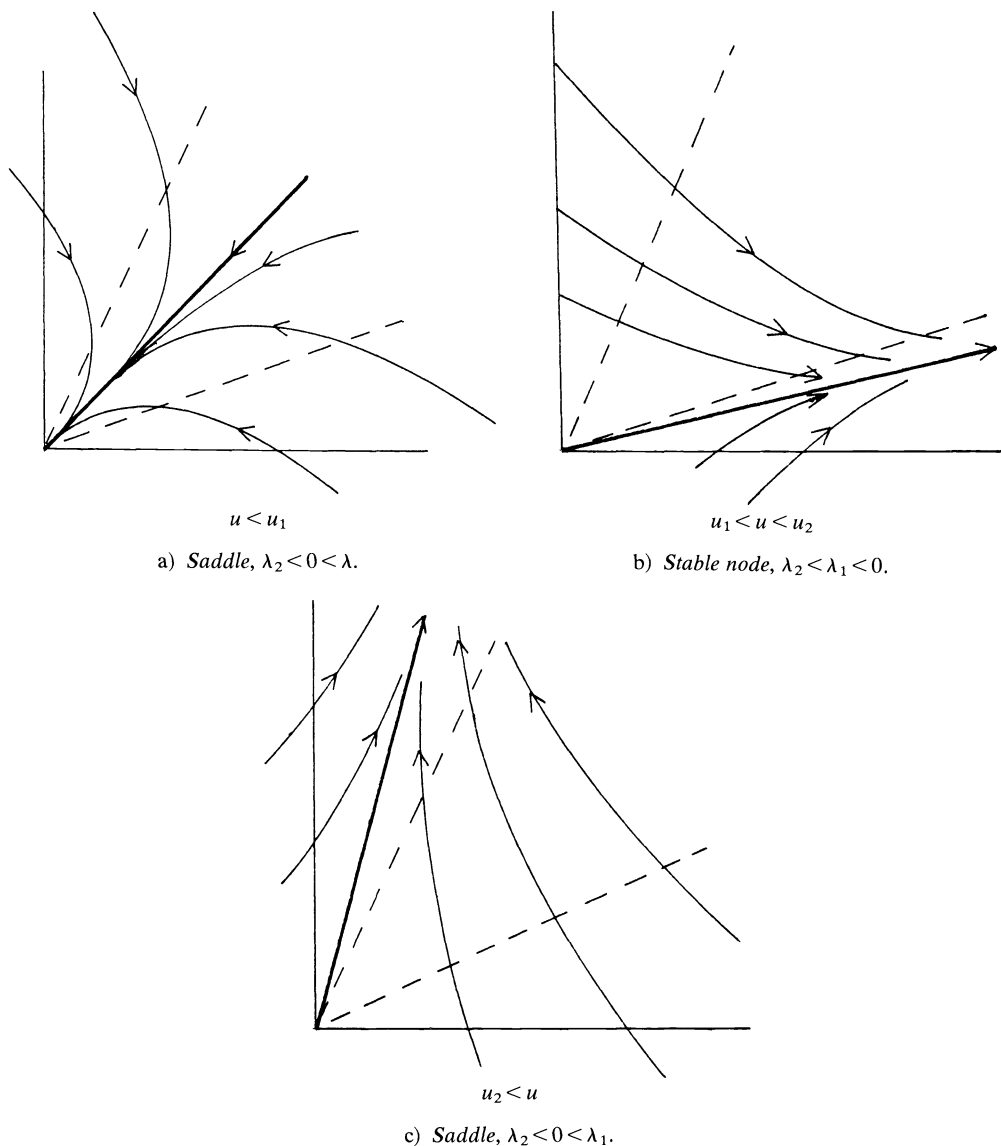


FIG. 3

vector v is in the first, second or third sector (numbered counterclockwise). Suppose x_1 is in sector III, then it will lie on a flow line passing through sector II (Fig. 4a) which in turn will be cut by a flow line corresponding to λ_1 negative which can be chosen to cut the first flow line and carry x_0 to it, at least if x_0 is far enough out to lie on the opposite side from the origin of the first flow line (Fig. 4b). If it is not, we can easily move it out to such a position in a first step by choosing u so that $\lambda_1 > 0$ as in Fig. 3c, for example.

If x_1 does not, in fact, lie in sector III, it does lie on a flow line (taking $\lambda_1 > 0$ as in Fig. 3a or 3c) which comes in from sector III, so we can add one more step to the above path. Thus, we must switch controls at most four times to pass from x_0 to x_1 .

Finally, we wish to look at case (c) in which both λ_1 and λ_2 change sign as u goes through its range of values. The corresponding flow patterns are illustrated in Figs. 5a, 5b, 5c with the dashed lines corresponding to the values $u = u_1$ and $u = u_2$.

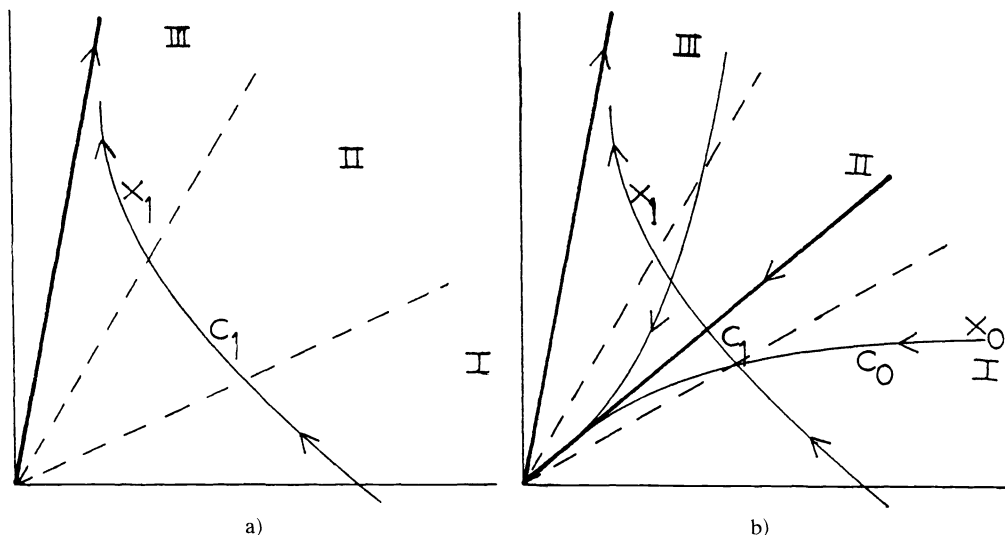


FIG. 4

By arguments similar to the previous case, we can establish controllability in the first quadrant in this case also. Roughly speaking, there is a flow line of type shown in Fig. 4b or 4c passing to any x_1 and we can carry any x_0 , using $u < u_1$ as in Fig. 4a, to a point on this flow line which moves along it to x_1 .

In summary, there do exist fairly large, i.e., open, subsets of our generic pairs A, B for which we have controllability in the first octant. In fact, the space of pairs A, B is divided by $\mathcal{D} = 0$ into two open sets: $\mathcal{D} > 0$ and $\mathcal{D} < 0$. On the former we have controllability in cases (b) and (c)—thus, in an open subset. When $\mathcal{D} < 0$ we shall see below that the system is not controllable. (Corollary 6.9).

It would, of course, be interesting to see if something similar is true in the case $n = 3$. However, a geometric analysis of the type above is much more difficult, and the author knows of no analytic approach which could be used here.

6. Some noncontrollable cases. In this section we will establish the existence of an open subset of pairs A, B with A essentially positive and B diagonal which can be shown to be noncontrollable. We do this by establishing the existence of a function $\Phi(x) = \Phi(x_1, \dots, x_n)$ on R^n which is monotone along each solution curve of $\dot{x} = Ax + uBx$ for a special class of A, B .

THEOREM 6.1. Suppose that A, B are real $n \times n$ matrices, B diagonal with entries b_1, \dots, b_n and $A = (a_{ij})$ such that $a_{ij} \geq 0$ for $i \neq j$. If there exists a nonnegative vector $p = (p_1, \dots, p_n) \geq 0$ such that (i) $\sum_{i=1}^n p_i b_i = 0$ and (ii) $\sum_{i=1}^n p_i a_{ii} \geq 0$, then $\dot{x} = (A + uB)x$ is not controllable in the positive orthant.

Proof. Define $\Phi(x) = \Phi(x_1, \dots, x_n)$ on R^n by $\Phi(x) = \prod_{i=1}^n x_i^{p_i}$, with p_1, \dots, p_n satisfying (i) and (ii). We show that

$$\langle \text{grad } \Phi, (A + uB)x \rangle \geq 0$$

for all $x \neq 0$ such that $x \geq 0$. This implies that Φ is monotone nondecreasing along any solution curve $x(t)$, $t \geq 0$, of $\dot{x}(t) = (A + u(t)B)x(t)$ such that $x(0) > 0$. Thus, we do not have controllability on R_+^n . Computing we obtain

$$\left(\frac{\partial \Phi}{\partial x_1}, \dots, \frac{\partial \Phi}{\partial x_n} \right) (A + uB)(x_1, \dots, x_n)' = \sum_{i,j} \Phi_i a_{ij} x_j + u \sum_i \Phi_i b_i x_i.$$

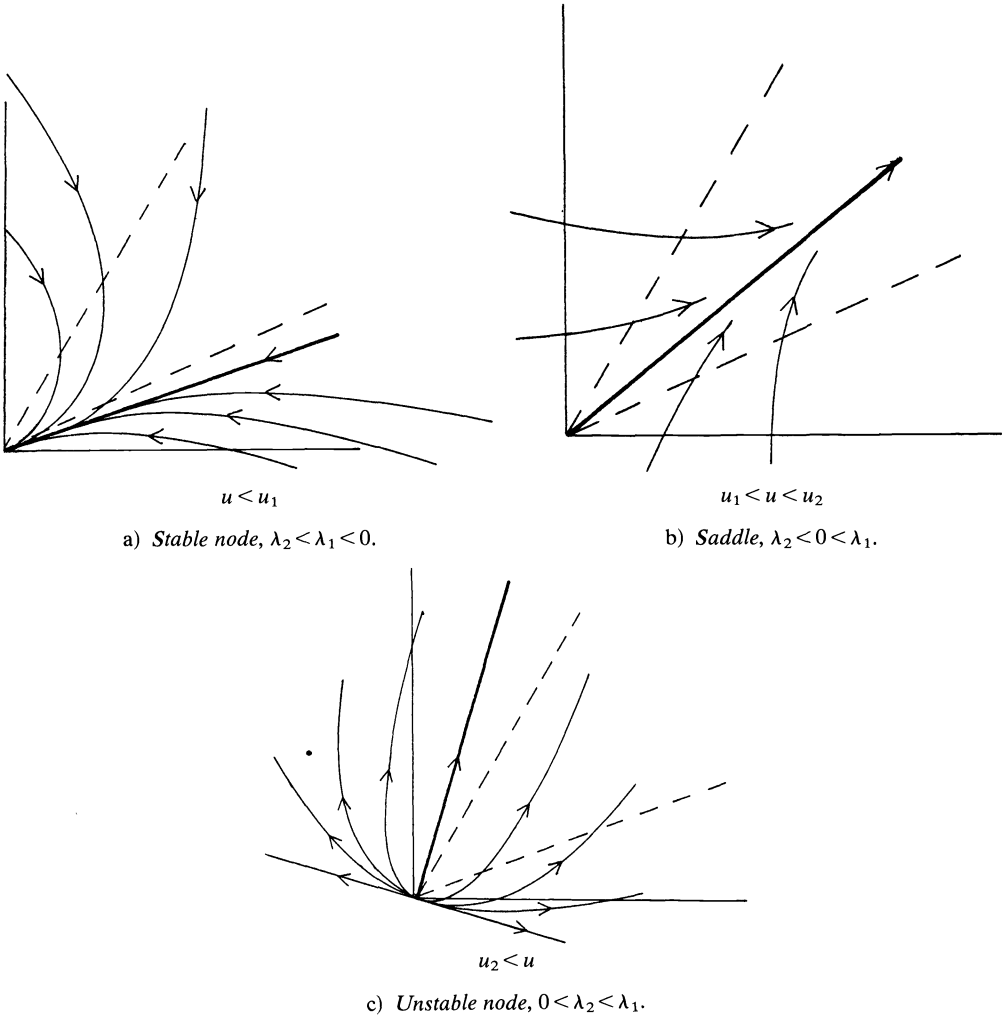


FIG. 5

However, $\Phi_i = \partial\Phi/\partial x_i = p_i x_i^{p_i-1} \prod_{j \neq i} x_j^{p_j} = p_i x_1^{-1} \Phi$ if $x_i > 0$ for all i . Thus

$$\begin{aligned}
 \langle \text{grad } \Phi, (A + uB)x \rangle &= \sum_{i,j} p_i x_i^{-1} \Phi a_{ij} x_j + u \sum_i p_i b_i \Phi \\
 (6.2) \qquad &= \Phi \left[\sum_{i=1}^n p_i a_{ii} + \sum_{i \neq j} (p_i x_i^{-1}) a_{ij} x_j \right] \geq 0
 \end{aligned}$$

for all $x > 0$. This completes the proof.

Remark 6.3. If A, B are as in the hypothesis but no $b_i = 0$ (this is the generic case) and $\sum_i p_i a_{ii} > 0$, then, clearly conditions (i) and (ii) are satisfied for matrices \tilde{A}, \tilde{B} near A and B , i.e., in an open subset of pairs A, B of the same type (A essentially positive and $\text{rank } B = n$).

Remark 6.4. The conditions obviously do not specify $p = (p_1, \dots, p_n)$ uniquely. Also note that we could equally well search for $(p_1, \dots, p_n) \leq 0$ satisfying (i) and (ii') $\sum p_i a_{ii} \leq 0$. Finally, as we see by studying the $n = 2$ case, it is possible to weaken the requirement that all components of p have the same sign.

Consider the case $n = 2$. Given A, B with $a_{12} > 0$ and $a_{21} > 0$, $b_1 \neq 0 \neq b_2$, we let $\Phi(x) = x_1^{p_1} x_2^{p_2}$.

(6.5)

$$\langle \text{grad } \Phi, (A + uB)x \rangle = \left\{ p_1 a_{11} + p_2 a_{22} + \left(\frac{p_1}{x_1} \right) a_{12} x_2 + \left(\frac{p_2}{x_2} \right) a_{21} x_1 + (p_1 b_1 + p_2 b_2) u \right\} \Phi.$$

For $x > 0$, $\Phi > 0$, and hence the sign of the expression is determined by the terms in parenthesis. If we wish the sign to be the same for all $x > 0$, we clearly must require p_1, p_2 to be chosen so that $p_1 b_1 + p_2 b_2 = 0$. The solutions are all of the form $p_1 = k b_2$, $p_2 = -k b_1$. We also note that the sign of the expression is not altered by multiplying by $x_1 x_2$. Thus, the sign of $\langle \text{grad } \Phi, (A + uB)x \rangle$ for $x > 0$ is always the same if and only if that of the quadratic form $Q = -(b_2 a_{11} - b_1 a_{22}) x_1 x_2 - b_2 a_{12} x_2^2 + b_1 a_{21} x_1^2$ or

$$(6.6) \quad Q(x_1, x_2) = a_{21} b_1 x_1^2 + (b_1 a_{22} - b_2 a_{11}) x_1 x_2 - a_{12} b_2 x_2^2$$

stays the same for $x > 0$.

A sufficient condition for this is that the discriminant \mathcal{D} of $Q(x)$ be negative:

$$(6.7) \quad \mathcal{D} = (a_{22} b_1 - a_{11} b_2)^2 + 4 b_1 b_2 a_{12} a_{21} < 0.$$

If the discriminant is positive, we have as the graph of $z = Q(x_1, x_2)$ a saddle surface, and only the case $Q(x_1, x_2) = c x_1 x_2$ would guarantee that the sign of Q does not change in the first quadrant. But we have assumed a_{12}, a_{21}, b_1, b_2 are all nonzero. Hence $\mathcal{D} > 0$ implies that the sign of $Q(x)$ changes in the first quadrant, and we are not able to draw the conclusion of noncontrollability in this case. The case $\mathcal{D} = 0$ is somewhat special and occurs only if

$$(6.8) \quad (a_{22} b_1 - a_{11} b_2)^2 = -4 b_1 b_2 a_{12} a_{21},$$

which, in particular, requires b_1 and b_2 to have opposite signs. We will eliminate it from consideration as "nongeneric".

Finally, note that the expression \mathcal{D} above was considered earlier, it is exactly the discriminant of the quadratic polynomial in u given by $\det(A + uB)$. Hence, $\mathcal{D} < 0$ implies that for all u the characteristic roots λ_1, λ_2 of the $A + uB$ satisfy $\lambda_2 < 0 < \lambda_1$. Thus we have what was conjectured earlier as a corollary to our theorem. (See (4.4) and subsequent discussion.)

COROLLARY 6.9. *If $\mathcal{D} < 0$, or equivalently, $\lambda_2 < 0 < \lambda_1$, for all $u \in \mathbb{R}$, then the system $\dot{x} = (A + uB)x$ is not controllable on \mathbb{R}_+^n .*

Note that p_1 and p_2 might be of opposite sign (but we must have $p_1 b_1 + p_2 b_2 = 0$). This shows that it is possible to weaken the requirement $p \geq 0$ in the theorem.

7. Increasing the number of controls. The following theorem is very easy to prove, but it answers a natural question and is worth stating as indicating some possibilities.

THEOREM. *Suppose $A = (a_{ij})$ is an $n \times n$ real matrix with $a_{ij} \geq 0$ for $i \neq j$ and that B_1, \dots, B_n are n linearly independent diagonal matrices. Consider the system on \mathbb{R}_+^n ,*

$$(**) \quad \dot{x} = (A + u_1 B_1 + \dots + u_n B_n)x,$$

where $u = (u_1, \dots, u_n)$ are piecewise continuous control functions with values in \mathbb{R}^n . If $x^{(0)}$ and $x^{(1)}$ are points of \mathbb{R}_+^n and $x(t)$, $0 \leq t \leq T$ is any piecewise differentiable path in \mathbb{R}_+^n , then there exist controls such that $x(t)$ is a solution of (**).

Proof. It is enough to show that for each vector $y = (y_1, \dots, y_n)$ at $x = (x_1, \dots, x_n) > 0$ there exists a vector $u = (u_1, \dots, u_n)$ such that $y = (A + \sum u_i B_i)x$.

There is no loss of generality in supposing $B_k = (\delta_{ik}\delta_{kj})$ the diagonal matrix with +1 in the k th row and column and zeros elsewhere. Thus, for x given, we have

$$(A + \sum u_i B_i)x = \begin{pmatrix} a_{11} + u_1 & a_{12} & a_{1n} \\ a_{21} & a_{22} + u_2 & a_{2n} \\ a_{n1} & a_{n2} & a_{nn} + u_n \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} u_1 x_1 + \sum a_{1j} x_j \\ \vdots \\ u_n x_n + \sum a_{nj} x_j \end{pmatrix},$$

where x_1, \dots, x_n are all positive and fixed as is $\sum_j a_{ij} x_j$ for $i = 1, \dots, n$. Obviously we need only choose u_1, \dots, u_n according to the rule

$$u_i = \frac{1}{x_i} \left(y_i - \sum_j a_{ij} x_j \right), \quad i = 1, \dots, n,$$

in order to get $(A + \sum u_i B_i)x = y$. This assures that we have total controllability in a very strong sense. In particular, given any x and a neighborhood U of x , there exists $V \subset U$ a neighborhood of x such that we have controllability in V without leaving U .

Acknowledgment. The author wishes to acknowledge with gratitude many helpful discussions of this subject with David Elliott, who suggested the problem, and further help and encouragement from Roger Brockett.

REFERENCES

- [1] R. BELLMAN, *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1960.
- [2] R. W. BROCKETT AND J. M. C. CLARK, *The geometry of the conditional density equation*, Proc. Oxford Conference on Stochastic Control, Oxford, 1968.
- [3] R. W. BROCKETT, *On the reachable set for bilinear systems*, Lecture Notes in Economics and Mathematical Systems, Springer, New York, 1970.
- [4] F. R. GANTMACHER, *The Theory of Matrices*, vol II, Chelsea, New York, 1959.
- [5] V. JURDJEVIC AND I. KUPKA, *Control systems subordinated to a group action: accessibility*, to appear.
- [6] A. J. KRENER, *A generalization of Chow's theorem and the bang-bang theorem to nonlinear control problems*, SIAM J. Control, 11 (1973), pp. 670–676.
- [7] ———, *Bilinear and nonlinear realizations of input-output maps*, SIAM J. Control, 13 (1975) pp. 827–834.
- [8] C. LOBRY, *Contrôlabilité des systèmes nonlineaires*, SIAM J. Control, 8 (1970), pp. 573–605.
- [9] D. G. LUENBERGER, *Introduction to Dynamic Systems*, John Wiley, New York, 1979.
- [10] L. MARKUS, *Catastrophies of economic equilibria*, in *Calculus of Variations and Control Theory*, Academic Press, New York, 1976.
- [11] N. J. PULLMAN, *Matrix Theory and Its Applications*, Marcel Dekker, New York, 1976.
- [12] H. J. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.

QUADRATIC CONTROL OF EVOLUTION EQUATIONS WITH DELAYS IN CONTROL*

AKIRA ICHIKAWA†

Abstract. The quadratic cost problem of evolution equations with delays in control is considered. A semigroup model which involves no explicit delays in control, but contains an unbounded control operator is introduced. With the aid of a family of approximating systems, it is shown that the optimal feedback control and the minimum cost are characterized by the solution of a Riccati equation. Three examples are given to illustrate the theory. The filtering problem of evolution equations with observation delays is also solved through the duality relation.

Introduction. One of the most general forms of a controlled linear evolution equation is as follows [12], [20], [25], [29], [30];

$$(0.1) \quad z(t) = T(t-t_0)z_0 - A \int_{t_0}^t T(t-r)B_0u(r) dr + \int_{t_0}^t T(t-r)B_1u(r) dr,$$

which corresponds to the differential equation

$$(0.1)' \quad \frac{d}{dt} z(t) = A[z(t) - B_0u(t)] + B_1u(t), \quad z(t_0) = z_0,$$

where $T(t)$ is a strongly continuous semigroup on a Hilbert space Z , u is a control with values in a Hilbert space U , $B_0, B_1 \in \mathcal{L}(U, Z)$ (the space of bounded linear operators mapping U into Z), and A is the infinitesimal generator of $T(t)$.

If $z(t)$ is continuous for each $u \in L_2(t_0, t_1; U)$, then the following cost is meaningful:

$$(0.2) \quad J(u) = \langle Gz(t_1), z(t_1) \rangle + \int_{t_0}^{t_1} [\langle Mz(t), z(t) \rangle + \langle Nu(t), u(t) \rangle] dt,$$

where $0 \leq G, M \in \mathcal{L}(Z)$, $0 < N \in \mathcal{L}(U)$ with $N^{-1} \in \mathcal{L}(U)$ and $\langle \cdot, \cdot \rangle$ denotes inner products in U and Z . If $B_0 = 0$, then (0.1) becomes

$$(0.3) \quad z(t) = T(t-t_0)z_0 + \int_{t_0}^t T(t-r)B_1u(r) dr.$$

The theory of the quadratic cost control problem (0.3), (0.2) is complete and the optimal control is characterized by the solution of a Riccati equation [7], [9], [10], [13]. Applications include partial differential equations of parabolic and hyperbolic type and delay differential equations. The theory was extended in [8] to the case where B_1 is unbounded in the sense that B_1 maps U outside Z . The results in [8] can be applied to partial differential equations with boundary or pointwise controls. However, there is no general theory which covers the quadratic control problem for differential delay equations involving control delays. The simplest case of a delay equation with a single delay in state and control was considered by Koivo and Lee [17]. They obtained the optimal feedback law, but failed to give the minimum cost. Kwong [18] then gave an expression for it and developed a further study into the infinite time quadratic problem. But their theory is not a direct extension of that based on Riccati equations as in [7], [8].

* Received by the editors May 15, 1981. This research was supported in part by the Sakkokai Foundation.

† Faculty of Engineering, Shizuoka University, Hamamatsu 432, Japan.

In this paper we consider evolution equations with both point and distributed delays in control. We introduce a dynamics for the segment of the control $u(t+s)$, $-b \leq s \leq 0$, and obtain a system of the form (0.1)'. We then show that the new system is equivalent to another one of the form (0.3) with an unbounded B . Thus, we expect that the quadratic problem is solved in terms of a Riccati equation as in [8]. We shall show this using a family of approximating systems of the form (0.3) with bounded control operators. Zabczyk [29] considered the quadratic problem (0.1), (0.2) with $G=0$ and transformed it to the quadratic problem of the form (0.3), (0.2) with an unbounded M . If, in particular, the unbounded operator has a continuous extension to Z , then the quadratic problem is solved using the standard results in [9]. He showed that the quadratic problem for delay equations with only distributed delays is an example of this type. Recently, Vinter and Kwong [27] also introduced an evolution equation of the form (0.3) with a bounded B_1 to study the infinite time quadratic problem for a delay differential equation. However, their model allows only for a distributed delay in control.

Section 1 is concerned with a model for infinite dimensional systems with delays in control. We introduce a first order differential equation for the segment of the control function and obtain an evolution equation of the form (0.1)'. We then transform the new system into the one of the form (0.3). We also give some preliminary results. In § 2 we introduce a family of approximating systems of the form (0.3) with bounded B_1 's and solve quadratic problems for them. Then we show that our original problem can be solved as a limiting case of these problems. We give three examples to illustrate the theory. It is known [12], [20], [25], [30] that the system (0.1) can describe a class of partial differential equations with boundary controls. In § 3 we briefly discuss that the quadratic problem for such systems can be solved using a family of approximating systems. So certain results in [8] can be obtained easily. In § 4 we consider the filtering problem for an infinite dimensional system with delay in observation. Again this is more complicated than the standard filtering problem without delay [5], [9]. The filtering problem for delay equations with observation delays was studied by Bagchi [2], and the filter stability was considered by Kwong and Willsky [18] and Kwong [19]. The theory of filtering for evolution equations is given in [5], [9], [22] when observation operators are bounded. The theory was extended to the case of unbounded observation in [6] using the results in [8]. A general theory which allows for unbounded observations was also established by Ouvrard [23], [24] using a martingale projection theorem. It can be applied to the smoothing problem of delay equations with observation delays. Here we show that our filtering problem can be solved using the results in § 2 since filtering with observation delay is dual to quadratic control with control delay.

This paper is a revised version of the report [15].

1. Evolution equations with delay. Let X and U be real separable Hilbert spaces. Let $\mathcal{L}(U, X)$ be the space of bounded linear operators mapping U into X . We write $\mathcal{L}(X)$ for $\mathcal{L}(X, X)$. We denote by $|\cdot|$ norms of vectors and operators. Let $L_2(p, r; U)$ be the space of U -valued square integrable functions on $[p, r]$. Consider the controlled system

$$(1.1) \quad \begin{aligned} \frac{dx(t)}{dt} &= Ax(t) + B_0 u(t) + \sum_{i=1}^k C_i u(t-b_i) + \int_{-b}^0 C_{01}(s) u(t+s) ds, \\ x(t_0) &= x_0, \quad u(s) = u_0(s), \quad t_0 - b \leq s < t_0, \quad 0 < b_1 < \cdots < b_k \leq b, \end{aligned}$$

where A is the infinitesimal generator of a strongly continuous semigroup $R(t)$ on X ,

$0 < t_1 < \infty$, $0 \leq t_0 < t_1$ is fixed, but arbitrary, $u(\cdot) \in L_2(t_0, t_1; U)$, $B_0, C_i \in \mathcal{L}(U, X)$, $i = 1, \dots, k$, $C_{01} \in \mathcal{L}(U, X)$ is strongly measurable with $|C_{01}| \leq a < \infty$ and $u_0 \in L_2(t_0 - b, t_0; U)$. We define the solution of (1.1) by

$$(1.2) \quad x(t) = R(t - t_0)x_0 + \int_{t_0}^t R(t - r) \left[B_0 u(r) + \sum_{i=1}^k C_i u(r - b_i) + \int_{-b}^0 C_{01}(s) u(r + s) ds \right] dr,$$

where the integrals are understood in the sense of Bochner. Then $x(t)$ is strongly continuous but not differentiable, in general. However, if (1.1) has an absolutely continuous solution, it is given by (1.2). Our control problem is to minimize the cost functional

$$(1.3) \quad J(u) = \langle Gx(t_1), x(t_1) \rangle + \int_{t_0}^{t_1} [\langle Mx(t), x(t) \rangle + Nu(t), u(t) \rangle] dt,$$

where $G, M \in \mathcal{L}(X)$ are self-adjoint and nonnegative, $N \in \mathcal{L}(U)$ is self-adjoint and positive with $N^{-1} \in \mathcal{L}(U)$ and $\langle \cdot, \cdot \rangle$ denotes inner products in X and U .

Delfour and Mitter [10] solved the quadratic problem for general delay differential equations using a product space formulation. The advantage of their approach is that their infinite dimensional system has a standard linear input-output relation and that the optimal control and the minimum cost are completely characterized by the solution of a Riccati equation. Koivo and Lee [17] solved the quadratic problem for a system with a single delay in state and in control. Delays in control give an additional difficulty in analysis [1], [17] and the Riccati-type theory is not directly applicable. Following the spirit of the product formulation [4], [10], we go one step further and introduce a new state space which includes the segment of the control. Then we obtain a linear input-output relation, and we expect that the Riccati-type theory is valid. In other words, we introduce a new state space which carries the necessary information to determine the future.

Set $y(t, s) = u(t + s)$, then we can formally write a differential equation for it:

$$(1.4) \quad \frac{\partial y(t, s)}{\partial t} = \frac{\partial y(t, s)}{\partial s}, \quad y(t_0, s) = y_0(s) = u_0(t_0 + s), \quad t > t_0, \quad -b \leq s \leq 0$$

with boundary condition

$$(1.5) \quad y(t, 0) = u(t).$$

If $u(\cdot)$ is smooth, then $u(t + \cdot)$ is a solution of (1.4), (1.5). Now let $Y = L_2(-b, 0; U)$ and consider the semigroup of left translation $S(t)$:

$$(1.6) \quad [S(t)y](s) = \begin{cases} y(t + s), & -b \leq s \leq -t \\ 0, & -t < s \leq 0 \\ 0, & -b \leq s \leq 0 \end{cases} \quad \text{if } t \leq b, \quad y \in Y.$$

Its generator is given by

$$(1.7) \quad Dy = \frac{dy}{ds},$$

with domain $\mathcal{D}(D) = \{y \in Y \mid y \text{ is absolutely continuous, } y' \in Y, y(0) = 0\}$.

Then we can write (1.4), (1.5) as an abstract differential equation in Y [28], [29]:

$$(1.8) \quad \frac{dy}{dt} = D(y - Fu), \quad y(t_0) = y_0,$$

where $F \in \mathcal{L}(U, Y)$ is defined by $(Fw)(s) = w$, $-b \leq s \leq 0$, $w \in U$. If (1.8) has an absolutely continuous solution, it is given by

$$(1.9) \quad y(t) = S(t-t_0)y_0 - D \int_{t_0}^t S(t-r)Fu(r) dr.$$

In fact, using the definitions of $S(t)$ and F , we can show that (1.9) is always defined and

$$(1.10) \quad [y(t)](s) = \begin{cases} u(t+s), & -(t-t_0) < s \leq 0 \\ u_0(t+s), & -b \leq s \leq -(t-t_0) \\ u(t+s), & -b \leq s \leq 0 \end{cases} \quad \begin{matrix} \text{if } t-t_0 \leq b, \\ \\ \text{if } t-t_0 > b. \end{matrix}$$

Define an operator $C: Y \rightarrow X$,

$$(1.11) \quad Cy = \sum_{i=1}^k C_i y(-b_i) + \int_{-b}^0 C_{01}(s)y(s) ds.$$

Then C is unbounded on Y but is bounded for example on $W^{1,2}(-b, 0; U) = [y \in Y \mid y \text{ is absolutely continuous, } y' \in Y]$. Note that if $y(t, s) = h(t+s)$ for some $h \in L_p(t_0-b, t_1; U)$, then $Cy \in L_p(0, t_1; X)$, $p \geq 1$. Now we rewrite (1.1) in the form

$$(1.12) \quad \frac{dx}{dt} = Ax + Cy + B_0 u, \quad x(t_0) = x_0, \quad \frac{dy}{dt} = D(y - Fu), \quad y(t_0) = y_0.$$

Then (1.2) is equivalent to

$$(1.13) \quad \begin{aligned} (a) \quad x(t) &= R(t-t_0)x_0 + \int_{t_0}^t R(t-r)[Cy(r) + B_0 u(r)] dr, \\ (b) \quad y(t) &= S(t-t_0)y_0 - D \int_{t_0}^t S(t-r)Fu(r) dr. \end{aligned}$$

It is convenient to introduce the Hilbert space $Z = X \times Y$. Define for each $z = \begin{bmatrix} x \\ y \end{bmatrix} \in Z$ and $t \geq t_0$

$$(1.14) \quad T(t-t_0) \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} R(t-t_0)x + \int_{t_0}^t R(t-r)CS(r-t_0)y dr \\ S(t-t_0)y \end{bmatrix}.$$

Then $T(t)$ is a strongly continuous semigroup on Z and its generator is [14]

$$(1.15) \quad \mathcal{A} = \begin{bmatrix} A & C \\ 0 & D \end{bmatrix}, \quad \mathcal{D}(\mathcal{A}) = \left\{ \begin{bmatrix} x \\ y \end{bmatrix} \in Z \mid x \in \mathcal{D}(A), y \in \mathcal{D}(D) \right\}.$$

We can easily verify that (1.12) can be written

$$(1.16) \quad \frac{d}{dt} z(t) = \mathcal{A} \left[z(t) - \begin{bmatrix} 0 \\ F \end{bmatrix} u(t) \right] + \begin{bmatrix} CF + B_0 \\ 0 \end{bmatrix} u(t), \quad z(t_0) = z_0 = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}.$$

Thus, we have obtained a system of the form (0.1)'. Next we shall show that (1.13) can be transformed into a familiar form

$$(1.17) \quad z(t) = T(t-t_0)z_0 + \int_{t_0}^t T(t-r)Bu(r) dr, \quad z_0 = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix},$$

although $B \notin \mathcal{L}(U, Z)$. For this purpose we consider the extension $\bar{D} \in \mathcal{L}(V, V^*)$ of the generator D given by $\bar{D}v = dv/ds - \Delta v(0)$, where $V = W^{1,2}(-b, 0; U)$, V^* the dual of V and $\Delta \in \mathcal{L}(U, V^*)$ is defined by $\langle \Delta u, w \rangle = \langle u, w(0) \rangle$ for each $w \in V$ and (\cdot, \cdot) denotes

the duality between V^* and V . Since $Fu \in V$ for each u , we have $\bar{D}Fu \in V^*$ and, in fact, $-\bar{D}Fu = \Delta u$. Thus, we can rewrite (1.13b) formally as

$$(1.18) \quad y(t) = S(t-t_0)y_0 - \int_{t_0}^t S(t-r)\bar{D}Fu(r) dr = S(t-t_0)y_0 + \int_{t_0}^t S(t-r)\Delta u(r) dr.$$

Now we give a meaning to (1.18). First note that $S(t)\Delta \in \mathcal{L}(U, V^*)$. So we can interpret the integration in the sense of Bochner in V^* , and $y(t)$ is a continuous V^* -valued function. It turns out that $y(t)$ is a Y -valued function and coincides with $y(t)$ given by (1.13b). Let $L_{2\text{loc}}(t_0, \infty; U)$ be the space of locally square integrable functions in U .

LEMMA 1.1. For each $u \in L_{2\text{loc}}(t_0, \infty; U)$ and $p \geq t_0$,

$$\begin{aligned} -D \int_p^t S(t-r)Fu(r)dr &= \int_p^t S(t-r)\Delta u(r) dr \\ &= \begin{cases} u(t+s), & -b \leq s \leq 0 \\ u(t+s), & -(t-p) < s \leq 0 \\ 0, & -b \leq s \leq -(t-p) \end{cases} \quad \begin{matrix} \text{if } t-p > b, \\ \text{if } t-p \leq b. \end{matrix} \end{aligned}$$

Proof. Appendix 1.

Using this lemma we rewrite (1.13) as

$$z(t) = T(t-t_0)z_0 + \int_{t_0}^t T(t-r) \begin{bmatrix} B_0 \\ \Delta \end{bmatrix} u(r) dr.$$

Setting $B = \begin{bmatrix} B_0 \\ \Delta \end{bmatrix}$, we obtain (1.17). We note that the integration is not defined in Z , although $z(t)$ is a continuous function in Z . The cost functional (1.3) is written as

$$(1.19) \quad J(u; t_0, z_0) = \langle \tilde{G}z(t_1), z(t_1) \rangle + \int_{t_0}^{t_1} [\langle \tilde{M}z(t), z(t) \rangle + Nu(t), u(t)] dt,$$

where

$$\tilde{G} = \begin{bmatrix} G & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \tilde{M} = \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix}.$$

Thus, we have transformed the quadratic problem (1.2), (1.3) into the one (1.17) (or (1.16)), (1.19). We naturally expect that the optimal control is given by the feedback law

$$(1.20) \quad u = -N^{-1}B^*Q(t)z \quad \text{with } J(u; t_0, z_0) = \langle Q(t_0)z_0, z_0 \rangle,$$

where $Q(t)$ is the solution of the Riccati equation

$$(1.21) \quad \frac{d}{dt} \langle Q(t)z, z \rangle + 2\langle Q(t)z, z \rangle + \langle Mz, z \rangle - \langle N^{-1}B^*Q(t)z, B^*Q(t)z \rangle = 0,$$

$$Q(t_1) = G, \quad z \in \mathcal{D}(\mathcal{A}).$$

In the following section, we shall show that this is in fact the case. In the rest of this section, we give some preliminary results. Let $0 < \lambda \in \rho(D)$, the resolvent set of D , and let $R(\lambda, D)$ be the resolvent of D [3]. Consider the family of approximations of (1.13b):

$$y_\lambda(t) = S(t-t_0)y_0 - R_\lambda D \int_{t_0}^t S(t-r)Fu(r) dr,$$

where $R_\lambda = \lambda R(\lambda, D)$. Then we have

$$(1.22) \quad y_\lambda(t) = S(t-t_0)y_0 + \int_{t_0}^t S(t-r)\Delta_\lambda u(r) dr,$$

where $\Delta_\lambda = -\lambda DR(\lambda, D)F \in \mathcal{L}(U, Y)$. Since $R_\lambda \rightarrow I$ (the identity) strongly as $\lambda \rightarrow \infty$ [3], $y_\lambda(t)$ converges to $y(t)$ as $\lambda \rightarrow \infty$. Approximations of this type based on the resolvent are also used in [16] to study stability and quadratic control of stochastic evolution equations.

LEMMA 1.2. Suppose that u_λ converges weakly to u in $L_2(t_0, t_1; U)$. Then

$$\tilde{y}_\lambda(t) = S(t-t_0)y_0 + \int_{t_0}^t S(t-r)\Delta_\lambda u_\lambda(r) dr.$$

converges weakly to $y(t)$ given by (1.13b) for all $t_0 \leq t \leq t_1$.

Proof. Appendix 2.

As we see in (1.20), we need to consider B^* . First note that $\Delta^* \in \mathcal{L}(V, U)$ and $\Delta^* w = w(0)$, $w \in V$. We can define $\Delta^* y$ for $y \in Y$ which is continuous at 0. We define $B^* = [B_0^*, \Delta^*]$, where Δ^* is understood in the generalized sense. Let $B_\lambda = [B_{\Delta_\lambda}^*]$.

LEMMA 1.3. If $y \in Y$ and $y(s)$ is continuous at $s = 0$, then $\Delta_\lambda^* y \rightarrow y(0)$ as $\lambda \rightarrow \infty$.

Proof. Appendix 3.

LEMMA 1.4. Let $V(t, s) \in \mathcal{L}(Z)$ be strongly measurable on $t_0 \leq s \leq t \leq t_1$ with $|V(t, s)| \leq a_1 < \infty$ and let

$$P(t)z = T^*(t_1-t)\tilde{G}V(t_1, t)z + \int_t^{t_1} T^*(r-t)\tilde{M}V(r, t)z dr.$$

Then $B^*P(t) \in \mathcal{L}(Z, U)$ and $|B^*P(t)| \leq a_2 < \infty$ for all t . Moreover, $B_\lambda^*P(t)z \rightarrow B^*P(t)z$ for all t and z as $\lambda \rightarrow \infty$.

Proof. As we see below, it is sufficient to consider the first term. Set

$$V(t_1, t)z = \begin{bmatrix} x(t_1, t) \\ y(t_1, t) \end{bmatrix}, \quad T^*(t_1-t)GV(t_1, t)z = \begin{bmatrix} P_0(t)z \\ P_1(t)z \end{bmatrix},$$

Then from (1.14) we obtain

$$\begin{aligned} P_0(t)z &= R^*(t_1-t)Gx(t_1, t), \\ [P_1(t)z](s) &= \sum_{i=1}^k C_i^* R^*(t_1-t-b_i-s)Gx(t_1, t) \\ &\quad + \int_{-b}^0 C_{01}^*(r)R^*(t_1-t-r-s)Gx(t_1, t) dr, \end{aligned}$$

where C_i^* , C_{01}^* , R^* are the adjoint operators of C_i , C_{01} , R , respectively, and we set $R^*(p) = 0$ if $p < 0$. Thus, $B^*P(t)$ is well defined and $|B^*P(t)| \leq a_2 < \infty$. By Lemma 1.3 we conclude that $B_\lambda^*P(t)z \rightarrow B^*P(t)z$ as $\lambda \rightarrow \infty$.

LEMMA 1.5. Let $V_\lambda(t, s) \in \mathcal{L}(Z)$, $\lambda \geq \lambda_0 > 0$ be a family of operators such that $V_\lambda(t, s) \rightarrow V(t, s)$ strongly as $\lambda \rightarrow \infty$ for all $t_0 \leq s \leq t \leq t_1$. Let

$$P_\lambda(t)z = T^*(t_1-t)\tilde{G}V_\lambda(t_1, t)z + \int_t^{t_1} T^*(r-t)\tilde{M}V_\lambda(r, t)z dr.$$

Then $B_\lambda^*P_\lambda(t)z \rightarrow B^*P(t)z$ for all t .

Proof. This is a consequence of Lemmas 1.3 and 1.4.

2. The optimal control. To solve the quadratic problem (1.17), (1.19) (which we call QP1) we consider a family of approximating systems:

$$(2.1) \quad z_\lambda(t) = T(t-t_0)z_0 + \int_{t_0}^t T(t-r)B_\lambda u(r) dr,$$

where $B_\lambda = [B_\lambda^0]$ and $\lambda \geq \lambda_0$ for some $0 < \lambda_0 \in \rho(D)$. Consider also

$$(2.2) \quad J_\lambda(u; t_0, z_0) = \langle \tilde{G}z_\lambda(t_1), z_\lambda(t_1) \rangle + \int_{t_0}^{t_1} [\langle \tilde{M}z_\lambda(t), z_\lambda(t) \rangle + \langle Nu(t), u(t) \rangle] dt.$$

Since $B_\lambda \in \mathcal{L}(U, Z)$, the optimal control for (2.1), (2.2) is well known [7], [9].

PROPOSITION 2.1. *The feedback control $u_\lambda = -N^{-1}B_\lambda^*Q_\lambda(t)z$ is optimal for (2.1), (2.2) and $J_\lambda(u_\lambda; t_0, z_0) = \langle Q_\lambda(t_0)z_0, z_0 \rangle$, where $Q_\lambda(t)$ is the unique solution of the Riccati equations*

$$(2.3) \quad \frac{d}{dt} \langle Q_\lambda(t)z, z \rangle + 2\langle Q_\lambda(t)z, z \rangle + \langle \tilde{M}z, z \rangle - \langle N^{-1}B_\lambda^*Q_\lambda(t)z, B_\lambda^*Q_\lambda(t)z \rangle = 0, \\ Q_\lambda(t_1) = \tilde{G}, \quad z \in \mathcal{D}(\mathcal{A}),$$

$$(a) \quad Q_\lambda(t)z = T^*(t_1-t)\tilde{G}T(t_1-t)z \\ + \int_t^{t_1} T^*(r-t)[\tilde{M} - Q_\lambda(r)B_\lambda N^{-1}B_\lambda^*Q_\lambda(r)]T(r-t)z dr,$$

$$(2.4) \quad (b) \quad Q_\lambda(t)z = T^*(t_1-t)\tilde{G}U_\lambda(t_1, t)z + \int_t^{t_1} T^*(r-t)\tilde{M}U_\lambda(r, t)z dr,$$

$$(c) \quad Q_\lambda(t)z = U_\lambda^*(t_1, t)\tilde{G}U_\lambda(t_1, t)z \\ + \int_t^{t_1} U_\lambda^*(r, t)[\tilde{M} + Q_\lambda(r)B_\lambda N^{-1}B_\lambda^*Q_\lambda(r)]U_\lambda(r, t)z dr,$$

in the class of strongly continuous nonnegative operators in $\mathcal{L}(Z)$, and $U_\lambda(t, s) \in \mathcal{L}(Z)$ is the unique solution of the integral equation

$$(2.5) \quad U_\lambda(t, s)z = T(t-s)z - \int_s^t T(t-r)B_\lambda N^{-1}B_\lambda^*Q_\lambda(r)U_\lambda(r, s)z dr$$

with properties

$$(2.6) \quad (a) \quad U_\lambda(t, r)U_\lambda(r, s)z = U_\lambda(t, s)z, \quad 0 \leq t_0 \leq s \leq r \leq t \leq t_1, \\ (b) \quad U_\lambda(t, t) = I, \quad \text{the identity on } Z, \\ (c) \quad U_\lambda(t, s) \text{ is jointly continuous on } t_0 \leq s \leq t \leq t_1.$$

Moreover, $Q_\lambda(t) \leq Q_0(t)$, where

$$Q_0(t)z = T^*(t_1-t)\tilde{G}T(t_1-t)z + \int_t^{t_1} T^*(r-t)\tilde{M}T(r-t)z dr.$$

LEMMA 2.1.

$$\sup_{t_0 \leq t \leq t_1} \left| \int_{t_0}^t T(t-r)B_\lambda u(r) dr \right| \leq a|u(\cdot)|_{L_2(t_0, t_1; U)}$$

for some $a > 0$ independent of λ .

Proof. Appendix 4.

When the feedback control $u_\lambda = -N^{-1}B_\lambda^*Q_\lambda(t)z$ is employed, the optimal control as a function of t is given by

$$(2.7) \quad u_\lambda(t) = u_\lambda(t; t_0, z_0) = -N^{-1}B_\lambda^*Q_\lambda(t)U_\lambda(t, t_0)z_0,$$

and as we see below, it is convergent in $L_2(t_0, t_1; U)$ as $\lambda \rightarrow \infty$.

LEMMA 2.2. $u_\lambda(t; t_0, z_0)$ converges in $L_2(t_0, t_1; U)$ as $\lambda \rightarrow \infty$ to a function of the form $u(t; t_0, z_0) = K(t, t_0)z_0$, $K(t, t_0) \in \mathcal{L}(Z, U)$. Moreover, $u(t; t_0, z_0)$ is optimal for the problem QP1.

Proof. Since $J_\lambda(u_\lambda; t_0, z_0) \leq J(0; t_0, z_0) = \langle Q_0(t_0)z_0, z_0 \rangle$, $u_\lambda(\cdot)$ is uniformly bounded in $L_2(t_0, t_1; U)$ for $\lambda \geq \lambda_0 > 0$. Hence, there exists a subsequence denoted again by $u_\lambda(\cdot)$ which is convergent weakly to $\bar{u}(\cdot) \in L_2(t_0, t_1; U)$. Then by virtue of Lemma 1.2, $z_\lambda(t)$ given by (2.1) converges weakly in Z to

$$(2.8) \quad \bar{z}(t) = T(t - t_0)z_0 + \int_{t_0}^t T(t - r)B\bar{u}(r) dr.$$

By lower semicontinuity of the quadratic functional, we have

$$(2.9) \quad J(\bar{u}; t_0, z_0) \leq \liminf_{\lambda \rightarrow \infty} J_\lambda(u_\lambda; t_0, z_0).$$

On the other hand, we have $J_\lambda(u_\lambda; t_0, z_0) \leq J_\lambda(\bar{u}; t_0, z_0)$ by the optimality of u_λ which in turn implies

$$(2.10) \quad \limsup_{\lambda \rightarrow \infty} J_\lambda(u_\lambda; t_0, z_0) \leq \lim_{\lambda \rightarrow \infty} J_\lambda(\bar{u}; t_0, z_0) = J(\bar{u}; t_0, z_0).$$

Thus, from (2.9) and (2.10), we conclude

$$(2.11) \quad J(\bar{u}; t_0, z_0) = \lim_{\lambda \rightarrow \infty} J_\lambda(u_\lambda; t_0, z_0).$$

This implies that the $L_2(t_0, t_1; U)$ norm of u_λ converges to that of \bar{u} . This fact together with the weak convergence of u_λ to \bar{u} implies that $u_\lambda \rightarrow \bar{u}$ strongly in $L_2(t_0, t_1; U)$. Now we show that \bar{u} is optimal for QP1. Let u be arbitrary in $L_2(t_0, t_1; U)$, then $J_\lambda(u_\lambda; t_0, z_0) \leq J_\lambda(u; t_0, z_0)$. Passing to the limit $\lambda \rightarrow \infty$ along the subsequence given above, we conclude that $J(\bar{u}; t_0, z_0) \leq J(u; t_0, z_0)$. Thus, \bar{u} is optimal. Next we show that the original sequence $u_\lambda \rightarrow \bar{u}$ in $L_2(t_0, t_1; U)$. In fact, suppose that there were a sequence which lies outside of a neighborhood of \bar{u} in $L_2(t_0, t_1; U)$. Then we can extract a subsequence which is weakly convergent to some element $\tilde{u} \in L_2(t_0, t_1; U)$. Then by the same reasoning above, we conclude that \tilde{u} is optimal for QP1. By strict convexity of $J(\cdot)$, we have $\bar{u} = \tilde{u}$, which is a contradiction. Thus,

$$u_\lambda(t; t_0, z_0) = -N^{-1}B_\lambda^*Q_\lambda(t)U_\lambda(t, t_0)z_0 \rightarrow \bar{u}(t; t_0, z_0) \quad \text{in } L_2(t_0, t_1; U).$$

Hence, we can write $u(t; t_0, z_0) = K(t, t_0)z_0$ for some $K(t, t_0) \in \mathcal{L}(Z, U)$.

LEMMA 2.3. $U_\lambda(t, s)$ converges strongly uniformly to some operator $U(t, s) \in \mathcal{L}(Z)$ which has the property (2.6) and $|U(t, s)| \leq a < \infty$ for some a independent of t, s .

Proof. Note that

$$(2.12) \quad U_\lambda(t, t_0)z_0 = T(t - t_0)z_0 - \int_{t_0}^t T(t - r)B_\lambda N^{-1}B_\lambda^*Q_\lambda(r)U_\lambda(r)U_\lambda(r, t_0)z_0 dr.$$

Thus, by Lemmas 2.1 and 2.2, $U_\lambda(t, t_0)z_0$ converges to

$$(2.13) \quad U(t, t_0)z_0 = T(t - t_0)z_0 - \int_{t_0}^t T(t - r)BK(r, t_0)z_0 dr.$$

By Proposition 2.1 we have

$$\begin{aligned} \int_{t_0}^{t_1} \langle Nu_\lambda(t; t_0, z_0), u_\lambda(t; t_0, z_0) \rangle dt &\leq \langle Q_\lambda(t_0)z_0, z_0 \rangle \\ &\leq \langle Q_0(t_0)z_0, z_0 \rangle \leq a_0|z_0|^2, \end{aligned}$$

where $a_0 = \sup_{0 \leq t \leq t_1} |Q_0(t)|$. Thus, by Lemma 2.1 we easily see $|U_\lambda(t, t_0)| \leq a < \infty$ for some constant a independent of λ, t, t_0 and the convergence above is uniform in t and t_0 . Hence, $|U(t, t_0)| \leq a$. Since t_0 is fixed but arbitrary, we have the assertion.

LEMMA 2.4. $Q_\lambda(t)$ converges strongly to $Q(t)$:

$$(2.14) \quad Q(t)z = T^*(t_1 - t)\tilde{G}U(t_1, t)z + \int_t^{t_1} T^*(r - t)\tilde{M}U(r, t)z dr,$$

where $U(t, s)$ is defined by (2.13). Moreover, $B^*Q(t) \in \mathcal{L}(Z, U)$ is well defined and $|B^*Q(t)| \leq a < \infty$ for all $0 \leq t \leq t_1$ and $B_\lambda^*Q_\lambda(t)z \rightarrow B^*Q(t)z$ for all t and z .

Proof. The strong convergence of $Q_\lambda(t)$ to $Q(t)$ follows from (2.4b) and Lemma 2.3. The second part is also the immediate consequence of Lemmas 1.4 and 2.3.

From this lemma we obtain the relation $K(t, s) = -N^{-1}B^*Q(t)U(t, s)$, and hence $U(t, s)$ satisfies

$$(2.15) \quad U(t, s)z = T(t - s)z - \int_s^t T(t - r)BN^{-1}B^*Q(r)U(r, s)z dr.$$

For integral equations of this type, we have the following.

PROPOSITION 2.2. Let $K(\cdot): [t_0, t_1] \rightarrow \mathcal{L}(Z, U)$ be strongly measurable and $|K(t)| \leq \tilde{a} < \infty$ for all t . Then the integral equation

$$U_K(t, s)z = T(t - s)z - \int_s^t T(t - r)BN^{-1}K(r)U_K(r, s)z dr$$

has a unique solution with the following properties:

- $$(2.16) \quad \begin{aligned} (a) \quad &U_K(t, s) \text{ is continuous on } t_0 \leq s \leq t \leq t_1, \\ (b) \quad &U_K(t, r)U_K(r, s) = U_K(t, s) \text{ for all } t_0 \leq s \leq r \leq t \leq t_1, \\ (c) \quad &U_K(t, t) = I \text{ for all } t_0 \leq t \leq t_1, \\ (d) \quad &|U_K(t, s)| \leq a < \infty \text{ for all } t_0 \leq s \leq t \leq t_1. \end{aligned}$$

This proposition can be proved by the standard method based on a contraction mapping theorem [14].

Note that $U(t, t_0)z_0$ is the response to the feedback control $\tilde{u} = -N^{-1}B^*Q(t)z$ with initial condition (t_0, z_0) , and so it is the optimal trajectory. Summing up we have the following.

THEOREM 2.1. The operator $Q_\lambda(t)$ converges strongly to a strongly continuous nonnegative operator $Q(t) \in \mathcal{L}(Z)$ and $B_\lambda^*Q_\lambda(t)$ strongly to $B^*Q(t)$ which is piecewise

continuous in $\mathcal{L}(Z, U)$. The operators $Q(t)$ and $B^*Q(t)$ satisfy the Riccati equations

$$(2.17) \quad \frac{d}{dt} \langle Q(t)z, z \rangle + 2 \langle Q(t)z, z \rangle + \mathcal{A}z + \langle \tilde{M}z, z \rangle - \langle N^{-1}B^*Q(t)z, B^*Q(t)z \rangle = 0,$$

$$Q(t_1) = \tilde{G}, \quad z \in \mathcal{D}(\mathcal{A}),$$

$$(a) \quad Q(t)z = T^*(t_1 - t)\tilde{G}T(t_1 - t)z$$

$$+ \int_t^{t_1} T^*(r - t)[\tilde{M} - (B^*Q(r))^*N^{-1}B^*Q(r)]T(r - t)z \, dr,$$

$$(2.18) \quad (b) \quad Q(t)z = T^*(t_1 - t)\tilde{G}U(t_1, t)z$$

$$+ \int_t^{t_1} T^*(r - t)\tilde{M}U(r, t)z \, dr,$$

$$(c) \quad Q(t)z = U^*(t_1, t)\tilde{G}U(t_1, t)z$$

$$+ \int_t^{t_1} U^*(r, t)[\tilde{M} + (B^*Q(r))^*N^{-1}B^*Q(r)]U(r, t)z \, dr,$$

where $U(t, s)$ is the unique solution of the integral equation (2.14). The optimal control for QP1 is given by the feedback law

$$\bar{u} = -N^{-1}B^*Q(t)z,$$

and the minimum cost is $J(\bar{u}; t_0, z_0) = \langle Q(t_0)z_0, z_0 \rangle$.

Let $z = \begin{bmatrix} x \\ y \end{bmatrix}$; then we can write

$$Q(t)z = \begin{bmatrix} Q_{00}(t) & Q_{01}(t) \\ Q_{10}(t) & Q_{11}(t) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix},$$

where $Q_{01}^*(t) = Q_{10}(t)$.

THEOREM 2.2. *The operators $Q_{10}(t)$ and $Q_{11}(t)$ have the following representation :*

$$(2.19) \quad [Q_{10}(t)x](s) = Q_{10}(t, s)x,$$

$$[Q_{11}(t)y](s) = \int_{-b}^0 Q_{11}(t, s, r)y(r) \, dr$$

for some piecewise continuous operators $Q_{10}(t, s) \in \mathcal{L}(X, U)$ and $Q_{11}(t, s, r) \in \mathcal{L}(U)$.

Proof. To simplify the proof, we assume $Cy = C_1y(-b)$. In the general case all the expressions below become lengthy, but the extension is otherwise straightforward. As in Lemma 1.4, we obtain from (2.18b)

$$(2.20) \quad (a) \quad Q_{00}(t)x = R^*(t_1 - t)GU_{00}(t_1, t)x + \int_t^{t_1} R^*(r - t)MU_{00}(r, t)x \, dr,$$

$$(b) \quad [Q_{10}(t)x](s) = C_1^*R^*(t_1 - t - b - s)GU_{00}(t_1, t)x$$

$$+ \int_t^{t_1} C_1^*R^*(r - t - b - s)MU_{00}(r, t)x \, dr,$$

$$(c) \quad [Q_{11}(t)](s) = C_1^*R^*(t_1 - t - b - s)GU_{01}(t_1, t)y$$

$$+ \int_t^{t_1} C_1^*R^*(r - t - b - s)MU_{01}(r, t)y \, dr,$$

where

$$U(t, s) = \begin{bmatrix} U_{00}(t, s) & U_{01}(t, s) \\ U_{10}(t, s) & U_{11}(t, s) \end{bmatrix}$$

and we set $R^*(p) = 0$ if $p < 0$. Then by (2.20b) $Q_{10}(t, s)$ is piecewise continuous. From (2.20c) we see that $Q_{11}(t)$ has a kernel $Q_{11}(t, s, r)$ which is piecewise continuous in s . But $Q_{11}(t, s, r)$ is symmetric in s and r , so it is also piecewise continuous in r . In particular, it is continuous at $s = 0$ except when $t_1 - t = b$. Next we shall show that $Q_{11}(t, s, r) \in \mathcal{L}(U)$ is uniformly bounded. For this purpose consider

$$Q_0(t) = \begin{bmatrix} q_{00}(t) & q_{01}(t) \\ q_{10}(t) & q_{11}(t) \end{bmatrix}.$$

Then

$$(2.21) \quad q_{00}(t)x = R^*(t_1 - t)GR(t_1 - t)x + \int_t^{t_1} R^*(r - t)MR(r - t)x \, dr,$$

and $q_{10}(t)$ and $q_{11}(t)$ have kernels

$$(2.22) \quad \begin{aligned} q_{10}(t, s)x &= C_1^*R^*(t_1 - t - b - s)GR(t_1 - t)x \\ &\quad + \int_t^{t_1} C_1^*R^*(r - t - b - s)MR(r - t)x \, dr, \\ q_{11}(t, s, r)u &= C_1^*R^*(t_1 - t - b - s)GR(t_1 - t - b - r)C_1u \\ &\quad + \int_t^{t_1} C_1^*R^*(p - t - b - s)MR(p - t - b - r)C_1u \, dp, \end{aligned}$$

respectively, where we set $R(p) = 0$ if $p < 0$. Thus, $q_{10}(t, s)$ and $q_{11}(t, s, r)$ are piecewise continuous and uniformly bounded. From $0 \leq Q(t) \leq Q_0(t)$ it follows that $0 \leq Q_{11}(t) \leq q_{11}(t)$. Thus, we have for all t

$$(2.23) \quad \begin{aligned} 0 &\leq \int_{-b}^0 \int_{-b}^0 \langle Q_{11}(t, s, r)y(r), y(s) \rangle \, dr \, ds \\ &\leq \int_{-b}^0 \int_{-b}^0 \langle q_{11}(t, s, r)y(r), y(s) \rangle \, dr \, ds, \quad y \in Y. \end{aligned}$$

Let $Q_{11}(t, s, r)$ and $q_{11}(t, s, r)$ be continuous at $s = r = s_0$. Then for each $u \in U$, we can construct a sequence of continuous functions $y_n(\cdot)$ with supports contained in a neighborhood of s_0 such that

$$\begin{aligned} \lim_{n \rightarrow \infty} \int_{-b}^0 \int_{-b}^0 \langle Q_{11}(t, s, r)y_n(r), y_n(s) \rangle \, dr \, ds &= \langle Q_{11}(t, s_0, s_0)u, u \rangle, \\ \lim_{n \rightarrow \infty} \int_{-b}^0 \int_{-b}^0 \langle q_{11}(t, s, r)y_n(r), y_n(s) \rangle \, dr \, ds &= \langle q_{11}(t, s_0, s_0)u, u \rangle. \end{aligned}$$

Thus, we conclude that $0 \leq \langle Q_{11}(t, s_0, s_0)u, u \rangle \leq \langle q_{11}(t, s_0, s_0)u, u \rangle$. Even if s_0 is not a point of continuity for Q_{11} , q_{11} , similar inequalities hold for left and right limits of Q_{11} and q_{11} at $s = r = s_0$. Hence, we conclude that

$$|Q_{11}(t, s, s)| \leq \sup_s |q_{11}(t, s, s)| \leq \sup_{s, r} |q_{11}(t, s, r)| = a < \infty.$$

Next let s_0 and r_0 be continuity points of Q_{11} and q_{11} . Then for each $u, v \in U$, we can construct a sequence of continuous functions $y_n(\cdot)$ with supports contained in the union of some neighborhoods of s_0 and r_0 with a similar property as above. Then we can show

$$\begin{aligned} & \langle Q_{11}(t, s_0, s_0)u, u \rangle + \langle Q_{11}(t, r_0, r_0)v, v \rangle + 2\langle Q_{11}(t, s_0, r_0)u, v \rangle \\ & \leq q_{11}(t, s_0, s_0)u, u \rangle + \langle q_{11}(t, r_0, r_0)v, v \rangle + 2\langle q_{11}(t, s_0, r_0)u, v \rangle. \end{aligned}$$

From this we obtain the bound $|Q_{11}(t, s_0, r_0)| \leq 3a$. Again, we can extend this inequality for arbitrary points s_0 and r_0 . From (2.18a) we have $Q_{11}(t) = q_{11}(t) - \bar{q}_{11}(t)$, where

$$\begin{bmatrix} \bar{q}_{00}(t) & \bar{q}_{01}(t) \\ \bar{q}_{10}(t) & \bar{q}_{11}(t) \end{bmatrix} z = \int_t^{t_1} T^*(r-t)[B^*Q(r)]^*N^{-1}B^*Q(r)T(r-t)z \, dr.$$

Thus the kernel of $\bar{q}_{11}(t)$ can be written

$$\bar{q}_{11}(t, s, r)u = \int_t^{t_1} q(p, t, s)N^{-1}q(p, t, r)u \, dp,$$

where $q(p, t, r)u = [B_0^*Q_{00}(p) + Q_{10}(p, 0)]R(p-t-r-b)C_1u + Q_{11}(p, 0, r-p+t)u$. Since $q(p, t, r)$ is uniformly bounded, it follows that $Q_{11}(t, s, r)$ is piecewise continuous in all variables and uniformly bounded.

Next we shall consider the uniqueness of a solution of the Riccati equation (2.17) and (2.18).

LEMMA 2.5. *Let $Q(t)$ be a solution of (2.18b) with properties in Theorems 2.1 and 2.2 (denoted property \mathcal{P}). Then*

$$\langle Q(t_0)z_0, z_0 \rangle = J(\tilde{u}; t_0, z_0) \leq J(u; t_0, z_0) \quad \text{for any } u \in L_2(t_0, t_1; U),$$

where $\tilde{u}(t) = -N^{-1}B^*Q(t)U(t, t_0)z_0$ is the control associated with the feedback law $\tilde{u} = -N^{-1}B^*Q(t)z$.

Proof. From (2.18b) we obtain

$$(2.24) \quad \tilde{u}(t) = -N^{-1}B^*[T^*(t_1-t)GU(t_1, t_0)z_0 + \int_t^{t_1} T^*(r-t)MU(r, t_0)z_0 \, dr].$$

Then, by direct calculation, we have

$$\begin{aligned} (2.25) \quad J(\tilde{u} + v; t_0, z_0) &= \langle Q(t_0)z_0, z_0 \rangle + \langle \tilde{G}z(t_1; v), z(t_1; v) \rangle \\ &\quad + \int_{t_0}^{t_1} [\langle \tilde{M}z(t; v), z(t; v) \rangle + \langle Nv(t), v(t) \rangle] \, dt, \end{aligned}$$

where $z(t; v) = \int_{t_0}^t T(t-r)Bv(r) \, dr$. The cancellation of cross terms such as $2 \int_{t_0}^{t_1} \langle N\tilde{u}(t), v(t) \rangle \, dt$ can be proved using the relation

$$\tilde{u}(t) = -\lim_{\lambda \rightarrow \infty} N^{-1}B_\lambda^*[T^*t_1-t)GU(t_1, t_0)z_0 + \int_t^{t_1} T^*(r-t)MU(r, t_0)z_0 \, dr].$$

Thus, $J(\tilde{u}; t_0, z_0) \leq J(u; t_0, z_0)$ for any admissible u .

LEMMA 2.6. *If $Q(t)$ satisfies (2.17), (2.18a) or (2.18b) with property \mathcal{P} , then it satisfies the other two as well as (2.18c).*

Proof. The equivalence of (2.17) and (2.18a) follows from [7]. Now let $Q(t)$ be a solution of (2.18b) with property \mathcal{P} . Define

$$\tilde{Q}_\lambda(t)z = T^*(t_1-t)\tilde{G}\tilde{U}_\lambda(t_1, t)z + \int_t^{t_1} T^*(r-t)\tilde{M}\tilde{U}_\lambda(r, t)z \, dr,$$

where $\tilde{U}_\lambda(t, s)$ is the unique solution of

$$\tilde{U}_\lambda(t, s)z = T(t-s)z - \int_s^t T(t-r)B_\lambda N^{-1}B^*Q(r)\tilde{U}_\lambda(r, s)z \, dr.$$

Then from [7] it follows that

$$\begin{aligned}\tilde{Q}_\lambda(t)z &= T^*(t_1-t)\tilde{G}T(t_1-t)z + \int_t^{t_1} T^*(r-t)[\tilde{M} - Q_\lambda(r)B_\lambda N^{-1}B^*Q(r)]T(r-t)z \, dr, \\ \tilde{Q}_\lambda(t)z &= U_\lambda^*(t_1, t)\tilde{G}\tilde{U}_\lambda(t_1, t)z + \int_t^{t_1} \tilde{U}_\lambda^*(r, t)[\tilde{M} + 2Q_\lambda(r)B_\lambda N^{-1}B_\lambda^*Q_\lambda(r) \\ &\quad - Q_\lambda(r)B_\lambda N^{-1}B^*Q(r)]\tilde{U}_\lambda(r, t)z \, dr.\end{aligned}$$

Since we can easily show that $\tilde{U}_\lambda(t, s) \rightarrow U(t, s)$ strongly, it follows that $\tilde{Q}_\lambda(t) \rightarrow Q(t)$ strongly and $B_\lambda^*\tilde{Q}_\lambda(t) \rightarrow B^*Q(t)$ strongly. Hence, we obtain (2.18a) and (2.18c) from these equations. Similarly, we can show that the solution of (2.18a) satisfies (2.18b) and (2.18c).

THEOREM 2.3. *There exists a unique solution of (2.17), (2.18a) and (2.18b) with property \mathcal{P} . The feedback control $\bar{u} = -N^{-1}B^*Q(t)z$ is optimal and the minimum cost is $J(\bar{u}; t_0, z_0) = \langle Q(t_0)z_0, z_0 \rangle$.*

COROLLARY 2.1. *The optimal control for QP1 is given by the feedback law*

$$(2.26) \quad \bar{u}(t) = -N^{-1} \left[[B_0^*Q_{00}(t) + Q_{10}(t, 0)]x(t) + \int_{-b}^0 Q_{11}(t, 0, s)\bar{u}(t+s) \, ds \right]$$

and the minimum cost by

$$(2.27) \quad \begin{aligned}J(\bar{u}; t_0, z_0) &= \langle Q_{00}(t_0)x_0, x_0 \rangle + 2 \left\langle \int_{-b}^0 Q_{01}(t_0, s)u_0(s) \, ds, x_0 \right\rangle \\ &\quad + \int_{-b}^0 \int_{-b}^0 \langle Q_{11}(t_0, s, r)u_0(r), u_0(s) \rangle \, dr \, ds.\end{aligned}$$

We can decompose (2.17) into the following set of equations:

$$\begin{aligned}(a) \quad & \frac{d}{dt} \langle Q_{00}(t)x, x \rangle + 2 \langle Q_{00}(t)x, Ax \rangle + \langle Mx, x \rangle \\ & - \langle N^{-1}[B_0^*Q_{00}(t) + \Delta^*Q_{10}(t)]x, [B_0^*Q_{00}(t) + \Delta^*Q_{10}(t)]x \rangle = 0, \\ & \quad \quad \quad x \in \mathcal{D}(A), \\ & \quad \quad \quad Q_{00}(t_1) = G, \\ (b) \quad & \frac{d}{dt} \langle Q_{01}(t)y, x \rangle + \langle Q_{01}(t)y, Ax \rangle + \langle Cy, Q_{00}(t)x \rangle + \langle Dy, Q_{10}(t)x \rangle \\ & - \langle N^{-1}[B_0^*Q_{01}(t) + \Delta^*Q_{11}(t)]y, [B_0^*Q_{00}(t) + \Delta^*Q_{10}(t)]x \rangle = 0, \\ & \quad \quad \quad Q_{01}(t_1) = 0, \quad x \in \mathcal{D}(A), \quad y \in \mathcal{D}(D), \\ (c) \quad & \frac{d}{dt} \langle Q_{11}(t)y, y \rangle + 2 \langle Q_{01}(t)y, Cy \rangle + 2 \langle Q_{11}(t)y, Dy \rangle \\ & - \langle N^{-1}[B_0^*Q_{01}(t) + \Delta^*Q_{11}(t)]y, [B_0^*Q_{01}(t) + \Delta^*Q_{11}(t)]y \rangle = 0, \\ & \quad \quad \quad Q_{11}(t_1) = 0, \quad y \in \mathcal{D}(D).\end{aligned}$$

Remark. In order to realize the feedback control $\bar{u} = -N^{-1}B^*Q(t)z$ we only need to know $B^*Q(t)$ or $Q_{00}(t)$, $Q_{10}(t, 0)$ and $Q_{11}(t, 0, s)$. We can derive integral equations for them from (2.18a). Computationally, it may be more convenient to solve them rather than (2.17). Once $B^*Q(t)$ is known, for example, $Q(t)$ is easily obtained by integration.

Now we give three examples to illustrate our theory.

Example 2.1. We take $X = R^n$, $U = R^p$, $A \in R^{n \times n}$, $B_0, C_i, C_{01} \in R^{n \times p}$, $G, M \in R^{n \times n}$ and $N \in R^{p \times p}$. Then (1.1) is an ordinary differential equation and Q_{00} , Q_{01} , and Q_{11} are $R^{n \times n}$ -, $R^{n \times p}$ - and $R^{p \times p}$ -matrix valued functions. We assume $b_k = b$. Then from (2.28) we obtain the following:

$$\begin{aligned}
 & \frac{d}{dt} Q_{00}(t) + A^* Q_{00}(t) + Q_{00}(t) A + M \\
 (a) \quad & -[Q_{00}(t) B_0 + Q_{01}(t, 0)] N^{-1} [B_0^* Q_{00}(t) + Q_{10}(t, 0)] = 0, \\
 & Q_{00}(t_1) = G, \\
 & \left(\frac{\partial}{\partial t} - \frac{\partial}{\partial s} \right) Q_{01}(t, s) + A^* Q_{01}(t, s) + Q_{00}(t) \left[\sum_{i=1}^{k-1} C_i \delta(s + b_i) + C_{01}(s) \right] \\
 (2.29) \quad (b) \quad & -[Q_{00}(t) B_0 + Q_{01}(t, 0)] N^{-1} [B_0^* Q_{01}(t, s) + Q_{11}(t, 0, s)] = 0, \\
 & Q_{01}(t_1, s) = 0, \quad Q_{01}(t, -b) = Q_{00}(t) C_k, \\
 & \left(\frac{\partial}{\partial t} - \frac{\partial}{\partial s} - \frac{\partial}{\partial r} \right) Q_{11}(t, s, r) + \sum_{i=1}^{k-1} [C_i^* \delta(s + b_i) + C_{01}^*(s)] Q_{01}(t, r) \\
 & + Q_{10}(t, s) \left[\sum_{i=1}^{k-1} C_i \delta(r + b_i) + C_{01}(r) \right] \\
 (c) \quad & -[Q_{10}(t, s) B_0 + Q_{11}(t, s, 0)] N^{-1} \\
 & \times [B_0^* Q_{01}(t, r) + Q_{11}(t, 0, r)] = 0, \\
 & Q_{11}(t_1, s, r) = 0, \quad Q_{11}(t, -b, r) = C_k^* Q_{01}(t, r), \quad Q_{11}(t, s, -b) = Q_{10}(t, s) C_k,
 \end{aligned}$$

where $\delta(\cdot)$ is the delta function.

If we set $Q_{01} = 0$, $Q_{11} = 0$, then (2.29a) is a usual Riccati equation.

Example 2.2. Let $X = R^n \times L_2(-a, 0; R^n)$ and $U = R^p$. Let A be the operator on X defined by

$$(2.30) \quad A \begin{bmatrix} x(0) \\ x(\cdot) \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^m A_i x(-a_i) + \int_{-a}^0 A_{01}(\mu) x(\mu) d\mu \\ dx/d\mu \end{bmatrix}$$

with domain

$$\mathcal{D}(A) = \left\{ \begin{bmatrix} x(0) \\ x(\cdot) \end{bmatrix} \middle| x(\cdot) \in W^{1,2}(-a, 0; R^n) \right\}$$

where $0 = a_0 < a_1 < \dots < a_m = a$, $A_i, A_{01} \in R^{n \times n}$, and A_{01} is a bounded measurable function. We take $B_0 u = \begin{pmatrix} B_0^u \\ 0 \end{pmatrix}$, $\hat{B}_0 \in R^{n \times p}$ and

$$C y = \left[\sum_{i=1}^k C_i y(-b_i) + \int_{-b}^0 C_{01}(s) y(s) ds \right],$$

where C_i, C_{01} are given as in Example 2.1.

It is known [9], [10], [11] that A generates a strongly continuous semigroup on X and (1.1) is the evolution equation model for the delay differential equation in R^n ;

$$(2.31) \quad \frac{d\hat{x}(t)}{dt} = \sum_{i=0}^m A_i \hat{x}(t-a_i) + \int_{-a}^0 A_{01}(\mu) \hat{x}(t+\mu) d\mu + \hat{B}_0 u(t) + \sum_{i=1}^k C_i u(t-b_i) + \int_{-b}^0 C_{01}(s) u(t+s) ds,$$

$$\hat{x}(\mu) = \hat{x}_0(\mu), \quad t_0 - a \leq \mu < t_0, \quad \hat{x}(t_0) = \hat{x}_0, \quad u(s) = u_0(s), \quad t_0 - b \leq s < t_0.$$

As a cost functional, we take

$$(2.32) \quad J(u) = \langle \hat{G} \hat{x}(t_1), \hat{x}(t_1) \rangle + \int_{t_0}^{t_1} [\langle \hat{M} \hat{x}(t), \hat{x}(t) \rangle + \langle Nu(t), u(t) \rangle] dt,$$

where $\hat{G}, \hat{M} \in R^{n \times n}$ and $N \in R^{p \times p}$. Thus, in this case $M = \begin{pmatrix} \hat{M} & 0 \\ 0 & 0 \end{pmatrix}$ and $G = \begin{pmatrix} \hat{G} & 0 \\ 0 & 0 \end{pmatrix}$. The quadratic problem in [17] is a special case of this example.

Now we set

$$Q_{00}(t) = \begin{bmatrix} P_{00}(t) & P_{01}(t) \\ P_{10}(t) & P_{11}(t) \end{bmatrix}, \quad Q_{01}(t) = \begin{bmatrix} R_{01}(t) \\ R_{02}(t) \end{bmatrix}.$$

Then

$$\begin{aligned} P_{00} &\in R^{n \times n}, & P_{01} &\in \mathcal{L}(L_2(-a, 0; R^n), R^n), \\ P_{10} &= P_{01}^*, & P_{11} &\in \mathcal{L}(L_2(-a, 0; R^n)), \\ R_{01} &\in \mathcal{L}(L_2(-a, 0; R^n), R^n), & R_{02} &\in \mathcal{L}(L_2(-b, 0; R^p), L_2(-a, 0; R^n)). \end{aligned}$$

We can derive differential equations for P_{00} and the kernels of other operators. From (2.28a) we obtain

$$(2.33) \quad \begin{aligned} &\frac{d}{dt} P_{00}(t) + A_0^* P_{00}(t) + P_{00}(t) A + \hat{M} + P_{01}(t, 0) + P_{10}(t, 0) \\ &- [P_{00}(t) \hat{B}_0 + R_{01}(t, 0)] N^{-1} [\hat{B}_0^* P_{00}(t) + R_{10}(t, 0)] = 0, \quad P_{00}(t_1) = \hat{G}, \end{aligned}$$

$$(2.34) \quad \begin{aligned} &\left(\frac{\partial}{\partial t} - \frac{\partial}{\partial \mu} \right) P_{01}(t, \mu) + A_0^* P_{01}(t, \mu) + P_{00}(t) \left[\sum_{i=1}^{m-1} A_i \delta(\mu + a_i) + A_{01}(\mu) \right] \\ &+ P_{11}(t, 0, \mu) - [P_{00}(t) \hat{B}_0 + R_{10}(t, 0)] N^{-1} [\hat{B}_0^* P_{01}(t, \mu) + R_{20}(t, 0, \mu)] = 0, \\ &P_{01}(t_1, \mu) = 0, \quad P_{01}(t, -a) = P_{00}(t) A_m, \end{aligned}$$

$$(2.35) \quad \begin{aligned} &\left(\frac{\partial}{\partial t} - \frac{\partial}{\partial \mu} - \frac{\partial}{\partial \nu} \right) P_{11}(t, \mu, \nu) + \left[\sum_{i=1}^{m-1} A_i^* \delta(\mu + a_i) + A_{01}^*(\mu) \right] P_{01}(t, \nu) \\ &+ P_{10}(t, \mu) \left[\sum_{i=1}^{m-1} A_i \delta(\nu + a_i) + A_{01}(\nu) \right] \\ &- [P_{10}(t, \mu) \hat{B}_0 + R_{02}(t, \mu, 0)] N^{-1} [\hat{B}_0^* P_{01}(t, \nu) + R_{20}(t, 0, \nu)] = 0, \\ &P_{11}(t_1, \mu, \nu) = 0, \quad P_{11}(t, -a, \nu) = A_m^* P_{01}(t, \nu), \quad P_{11}(t, \mu, -a) = P_{10}(t, \mu) A_m. \end{aligned}$$

Equation (2.28b) is decomposed into

$$(2.36) \quad \left(\frac{\partial}{\partial t} - \frac{\partial}{\partial s} \right) R_{01}(t, s) + A_0^* R_{01}(t, s) + P_{00}(t) \left[\sum_{i=1}^{k-1} C_i \delta(s + b_i) + C_{01}(s) \right] \\ + R_{02}(t, 0, s) - [P_{10}(t) \hat{B}_0 + R_{01}(t, 0)] N^{-1} [\hat{B}_0^* R_{10}(t, s) + Q_{11}(t, 0, s)] = 0, \\ R_{01}(t_1, s) = 0, \quad R_{01}(t, -b) = P_{00}(t) C_k,$$

$$(2.37) \quad \left(\frac{\partial}{\partial t} - \frac{\partial}{\partial \mu} - \frac{\partial}{\partial s} \right) R_{02}(t, \mu, s) + P_{01}(t, \mu) \left[\sum_{i=1}^{k-1} C_i \delta(s + b_i) + C_{01}(s) \right] \\ + \left[\sum_{i=1}^{m-1} A_i \delta(\mu + a_i) + A_{01}^*(\mu) \right] R_{10}(t, s) \\ - [P_{10}(t, \mu) \hat{B}_0 + R_{02}(t, \mu, 0)] N^{-1} [\hat{B}_0^* R_{10}(t, s) + Q_{11}(t, 0, s)] = 0, \\ R_{02}(t_1, \mu, s) = 0, \quad R_{02}(t, \mu, -b) = P_{10}(t, \mu) C_k, \quad R_{02}(t, -a, s) = A_m R_{10}(t, s).$$

Equation (2.28c) is written

$$(2.38) \quad \left(\frac{\partial}{\partial t} - \frac{\partial}{\partial s} - \frac{\partial}{\partial r} \right) Q_{11}(t, s, r) + \left[\sum_{i=1}^{k-1} C_i^* \delta(s + b_i) + C_{01}^*(s) \right] R_{01}(t, r) \\ + R_{10}(t, s) \left[\sum_{i=1}^{k-1} C_i \delta(r + b_i) + C_{01}(r) \right] \\ - [R_{01}(t, s) \hat{B}_0 + Q_{11}(t, s, 0)] N^{-1} [\hat{B}_0^* R_{10}(t, r) + Q_{11}(t, 0, r)] = 0, \\ Q_{11}(t_1, s, r) = 0, \quad Q_{11}(t, -b, r) = C_k^* R_{01}(t, r), \quad Q_{11}(t, s, -b) = R_{10}(t, s) C_k.$$

The optimal control is given by the feedback law

$$(2.39) \quad \bar{u}(t) = -N^{-1} \left\{ [\hat{B}_0^* P_{00}(t) + R_{10}(t, 0)] \hat{x}(t) \right. \\ \left. + \int_{-a}^0 [\hat{B}_0^* P_{01}(t, r) + R_{20}(t, 0, r)] \hat{x}(t+r) dr + \int_{-b}^0 Q_{11}(t, 0, s) \bar{u}(t+s) ds \right\},$$

and the minimum cost is

$$(2.40) \quad J(\bar{u}) = \langle P_{00}(t_0) \hat{x}_0, \hat{x}_0 \rangle + 2 \left\langle \int_{-a}^0 P_{01}(t_0, \mu) \hat{x}_0(\mu) d\mu, \hat{x}_0 \right\rangle \\ + \int_{-a}^0 \int_{-a}^0 \langle P_{11}(t_0, \mu, \nu) \hat{x}_0(\nu), \hat{x}_0(\mu) \rangle d\nu d\mu + 2 \left\langle \int_{-b}^0 R_{01}(t_1, s) u_0(s) ds, \hat{x}_0 \right\rangle \\ + 2 \int_{-a}^0 \int_{-b}^0 \langle R_{02}(t_0, \mu, s) u_0(s), \hat{x}_0(\mu) \rangle ds d\mu \\ + \int_{-b}^0 \int_{-b}^0 \langle Q_{11}(t_0, s, r) u_0(r), u_0(s) \rangle dr ds.$$

If we set $R_{10} = 0$, $R_{20} = 0$ and $Q_{11} = 0$, then (2.33), (2.34) and (2.35) coincide with the Riccati equation for the delay differential equation with no delay in control [9].

Example 2.3. Consider the heat equation on $[0, 1]$

$$(2.41) \quad \frac{\partial x(t, \mu)}{\partial t} = \frac{\partial^2 x(t, \mu)}{\partial \mu^2} + b(\mu) u(t) + c(\mu) u(t-b), \\ x(t_0, \mu) = x_0(\mu), \quad x(t, 0) = x(t, 1) = 0, \quad u(s) = u_0(s), \quad t_0 - b \leq s < t_0.$$

For this example we take $X = L_2(0, 1)$, $U = R^1$, $Y = L_2(-b, 0)$, $b(\cdot)$, $c(\cdot) \in X$ and A is the differential operator

$$A = \frac{d^2}{d\mu^2},$$

$$\mathcal{D}(A) = [x \in X | x, x' \text{ are absolutely continuous, } x', x'' \in X, x(0) = x(1) = 0].$$

Let N be a positive number, and let $G(\mu, \nu)$ and $M(\mu, \nu)$ be the kernels of G , $M \in \mathcal{L}(X)$, respectively. Then we can derive partial differential equations for the kernels of $Q_{ij}(t)$, $i, j = 0, 1$.

$$\begin{aligned} \left(\frac{\partial}{\partial t} + \frac{\partial^2}{\partial \mu^2} + \frac{\partial^2}{\partial \nu^2} \right) Q_{00}(t, \mu, \nu) + M(\mu, \nu) - N^{-1} \left[\int_0^1 Q_{00}(t, \mu, \theta) b(\theta) d\theta + Q_{01}(t, 0, \mu) \right] \\ \times \left[\int_0^1 Q_{11}(t, \theta, \nu) b(\theta) d\theta + Q_{10}(t, 0, \nu) \right] = 0, \end{aligned} \quad (2.42)$$

$$Q_{00}(t_1, \mu, \nu) = G(\mu, \nu),$$

$$Q_{00}(t, 0, \nu) = Q_{00}(t, 1, \nu) = Q_{00}(t, \mu, 0) = Q_{00}(t, \mu, 1) = 0,$$

$$\begin{aligned} \left(\frac{\partial}{\partial t} + \frac{\partial^2}{\partial \mu^2} - \frac{\partial}{\partial s} \right) Q_{01}(t, \mu, s) - N^{-1} \left[\int_0^1 Q_{00}(t, \mu, \nu) b(\nu) d\nu + Q_{01}(t, \mu, 0) \right] \\ \times \left[\int_0^1 Q_{01}(t, \nu, s) b(\nu) d\nu + Q_{11}(t, 0, s) \right] = 0, \end{aligned} \quad (2.43)$$

$$Q_{01}(t_1, \mu, s) = 0, \quad Q_{01}(t, 0, s) = Q_{01}(t, 1, s) = 0,$$

$$Q_{01}(t, \mu, -b) = \int_0^1 Q_{00}(t, \mu, \nu) c(\nu) d\nu,$$

$$\begin{aligned} \left(\frac{\partial}{\partial t} - \frac{\partial}{\partial s} - \frac{\partial}{\partial r} \right) Q_{11}(t, s, r) - N^{-1} \left[\int_0^1 Q_{10}(t, \nu, s) b(\nu) d\nu + Q_{11}(t, s, 0) \right] \\ \times \left[\int_0^1 Q_{01}(t, \nu, r) b(\nu) d\nu + Q_{11}(t, 0, r) \right] = 0, \end{aligned} \quad (2.44)$$

$$Q_{11}(t_1, s, r) = 0, \quad Q_{11}(t, -b, r) = \int_0^1 Q_{01}(t, \mu, r) c(\mu) d\mu,$$

$$Q_{11}(t, s, -b) = \int_0^1 Q_{10}(t, \mu, s) c(\mu) d\mu.$$

3. Boundary control. Let Z and U be real Hilbert spaces. Let A be the infinitesimal generator of a strongly continuous semigroup $T(t)$ on Z , and let $B_0, B_1 \in \mathcal{L}(U, Z)$. It is known [12], [20], [25], [30] that the mathematical model

$$(3.1) \quad z(t) = T(t)z_0 - A \int_{t_0}^t T(t-r)B_0 u(r) dr + \int_{t_0}^t T(t-r)B_1 u(r) dr$$

can describe a class of partial differential equations with both distributed and boundary controls. Under certain conditions, (3.1) is continuous and the following cost is

meaningful;

$$(3.2) \quad J(u) = \langle Gz(t_1), z(t_1) \rangle + \int_{t_0}^{t_1} [\langle Mz(t), z(t) \rangle + \langle Nu(t), u(t) \rangle] dt,$$

where $0 \leq G$, $M \in \mathcal{L}(Z)$ and $0 < N \in \mathcal{L}(U)$.

Let $R(\lambda, A)$ be the resolvent of A then as in (2.1) we can define a family of approximating systems

$$(3.3) \quad z_\lambda(t) = T(t)z_0 + \int_{t_0}^t T(t-r)B_\lambda u(r) dr,$$

where $B_\lambda = -\lambda AR(\lambda, A)B_0 + B_1$. Let $J_\lambda(u)$ be the functional (3.2) with z replaced by z_λ . Then, as in Proposition 2.1, the solution for the quadratic problem (3.2) and $J_\lambda(u)$ is known. We expect that the optimal solutions for these auxiliary problems converge as $\lambda \rightarrow \infty$ to the optimal solution for (3.1), (3.2). In fact, this is the case if we assume for example the following:

There exists a Hilbert space V such that $V \subset Z \subset V^*$ with continuous injections, where V^* is the dual of V . $T(t)$ maps V^* into Z and

$$(3.4) \quad |T(t)w|_Z \leq \frac{a}{t^\alpha} |w|_{V^*}, \quad w \in V^*, \quad 0 \leq \alpha < \frac{1}{2}.$$

Moreover, (3.1) is equivalent to

$$z(t) = T(t)z_0 + \int_{t_0}^t T(t-r)Bu(r) dr,$$

where $B \in \mathcal{L}(U, V^*)$.

Then it can be shown that $B_\lambda Q_\lambda(t)z$, $Q_\lambda(t)z$ converge to $B^*Q(t)z$, $Q(t)z$ in $L_2(t_0, t_1; U)$ and in Z , respectively, where $Q_\lambda(t)$ is the solution of the Riccati equation for the problem (3.3), $J_\lambda(u)$ and $Q(t)$ is the unique solution of the Riccati equation

$$\frac{d}{dt} \langle Q(t)z, z \rangle + 2\langle Q(t), Az \rangle + \langle Mz, z \rangle - \langle Q(t)BN^{-1}B^*Q(t)z, z \rangle = 0,$$

$$Q(t_1) = G,$$

$$Q(t)z = T^*(t_1-t)GT(t_1-t)z + \int_t^{t_1} T^*(r-t)[M - Q(r)BN^{-1}B^*Q(r)]T(r-t)z dr,$$

$$Q(t)z = T^*(t_1-t)GU(t_1, t)z + \int_t^{t_1} T^*(r-t)MU(r, t)z dr,$$

$$Q(t)z = U^*(t_1, t)GU(t_1, t)z + \int_t^{t_1} U^*(r, t)[M + Q(r)BN^{-1}B^*Q(r)]U(r, t)z dr,$$

in the class of strongly continuous nonnegative operator in $\mathcal{L}(Z)$ with $|B^*Q(t)|_{\mathcal{L}(Z, U)}$, $|Q(t)B|_{\mathcal{L}(U, Z)} \in L_2(t_0, t_1)$, and $U(t, s)$ is the unique solution of

$$U(t, s)z = T(t-s)z - \int_s^t T(t-r)BN^{-1}B^*Q(r)U(r, s)z dr.$$

The optimal control is given by the feedback law

$$\bar{u} = -N^{-1}B^*Q(t)z \quad \text{with } J(\bar{u}) = \langle Q(t_0)z_0, z_0 \rangle.$$

We can repeat the proof in § 2, but the analysis is different and, in fact, is much simpler because of the estimate (3.4). The details are omitted since this problem is solved under a more general setting in [8].

We can consider yet a more general model

$$\begin{aligned} x(t) &= R(t)x_0 - A \int_{t_0}^t R(t-r)F_1Cy(r) dr + \int_{t_0}^t R(t-r)B_1y(r) dr + \int_{t_0}^t R(t-r)B_2u(r) dr, \\ y(t) &= S(t)y_0 - D \int_{t_0}^t S(t-r)F_2u(r) dr, \end{aligned}$$

which is the mixture of the models (1.1) and (3.1) and may represent boundary control systems with delay in control. The analysis is then more involved, but it is possible to formulate a quadratic problem for this system.

4. Filtering with delay in observation. Consider the following signal and observation processes

$$(4.1) \quad x(t) = R(t)x_0 + \int_0^t R(t-r)F dw(r),$$

$$(4.2) \quad y(t) = \int_0^t \left[\sum_{i=0}^k C_i x(r-b_i) + \int_{-b}^0 C_{01}(s)x(r+s) ds \right] dr + v(t),$$

where $R(t)$ is a strongly continuous semigroup on a real separable Hilbert space X with generator A , $w(t)$ is a Wiener process in a real separable Hilbert space H with incremental covariance W [9], $F \in \mathcal{L}(H, X)$, $C_i \in \mathcal{L}(X, R^p)$, $C_{01} \in \mathcal{L}(X, R^p)$ is an essentially bounded function, $v(t)$ is a p -dimensional Wiener process with covariance $0 < V \in R^{p \times p}$, x_0 is a square integrable random variable with zero mean and covariance $0 \leq P_0 \in \mathcal{L}(X)$ and x_0 , $w(t)$ and $v(t)$ are independent. Our filtering problem is to find for each $t_1 > 0$ the best linear estimate of the state $x(t_1)$ based on the observation $y(s)$, $0 \leq s \leq t_1$, which is of the form

$$(4.3) \quad \hat{x}(t_1) = \int_0^{t_1} K(t_1, r) dy(r),$$

where $K \in \mathcal{L}(R^p, X)$ with $\int_0^{t_1} |K(t_1, r)|^2 dr < \infty$, or equivalently, to minimize

$$(4.4) \quad E[\langle x(t_1) - \hat{x}(t_1), x \rangle] \quad \text{for each } x \in X,$$

where E denotes the expectation.

This is the dual problem of the quadratic cost control problem in § 2. In fact, consider

$$(4.5) \quad \begin{aligned} \frac{d\xi(t)}{dt} &= -A^* \xi(t) + C_0^* u(t) + \sum_{i=1}^k C_i^* u(t+b_i) + \int_0^b C_{01}^* (-s) u(t+s) ds, \\ \xi(t_1) &= \xi_1, \quad u(t) = 0, \quad t > t_1, \end{aligned}$$

$$(4.6) \quad J(u) = \langle P_0 \xi(0), \xi(0) \rangle + \int_0^{t_1} [\langle FWF^* \xi(t), \xi(t) \rangle + \langle Vu(t), u(t) \rangle] dt,$$

where A^* , C_i^* , C_{01}^* are adjoint operators of A , C_i , C_{01} , respectively. We can prove the following as in Theorem 3.1 [21].

LEMMA 4.1. For each $\xi_1 \in X$ and $u(\cdot) \in L_2(0, t_1; R^p)$,

$$(4.7) \quad E[\langle \xi_1, x(t_1) \rangle - \int_0^{t_1} \langle u(t), dy(t) \rangle]^2 = J(u).$$

Now set $\eta(t) = \xi(t_1 - t)$, $\nu(t) = -u(t_1 - t)$, then the control problem (4.5), (4.6) is equivalent to

$$(4.8) \quad \begin{aligned} \frac{d\eta(t)}{dt} &= A^* \eta(t) + C_0^* \nu(t) + \sum_{i=1}^k C_i^* \nu(t - b_i) + \int_{-b}^0 C_{01}^*(s) \nu(t + s) ds, \\ \eta(0) &= \xi_1, \quad \nu(t) = 0, \quad t < 0, \end{aligned}$$

$$(4.9) \quad J(\nu) = \langle P_0 \eta(t_1), \eta(t_1) \rangle + \int_0^{t_1} [\langle FWF^* \eta(t), \eta(t) \rangle + \langle V \nu(t), \nu(t) \rangle] dt.$$

Thus, we can use the result in § 2 to obtain the optimal control

$$(4.10) \quad \bar{u}(t) = V^{-1} \{ [C_0 P_{00}(t) + P_{10}(t, 0)] \xi(t) + \int_{-b}^0 P_{11}(t, 0, s) \bar{u}(t - s) ds \}$$

and the minimum cost

$$(4.11) \quad J(\bar{u}) = \langle P_{00}(t_1) \xi_1, \xi_1 \rangle,$$

where

$$P(t) = \begin{bmatrix} P_{00}(t) & P_{01}(t) \\ P_{10}(t) & P_{11}(t) \end{bmatrix} \in \mathcal{L}(Z), \quad Z = X \times L_2(-b, 0; R^p)$$

satisfies the Riccati equation

$$(4.12) \quad \frac{d}{dt} \langle P(t)z, z \rangle - 2 \langle P(t)z, \tilde{\mathcal{A}}z \rangle - \langle \widetilde{FWF^*} z, z \rangle + \langle V^{-1} \tilde{B}^* P(t)z, \tilde{B}^* P(t)z \rangle = 0,$$

$$P(0) = \tilde{P}_0 = \begin{bmatrix} P_0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \widetilde{FWF^*} = \begin{bmatrix} FWF^* & 0 \\ 0 & 0 \end{bmatrix}, \quad z \in \mathcal{D}(\tilde{\mathcal{A}}),$$

and $P_{10}(t, s)$, $P_{11}(t, s, r)$ are the kernels of $P_{10}(t)$, $P_{11}(t)$, respectively, and

$$(4.13) \quad \tilde{\mathcal{A}} = \begin{bmatrix} A^* & C^* \\ 0 & D \end{bmatrix}, \quad \mathcal{D}(\tilde{\mathcal{A}}) = \left[\begin{bmatrix} x \\ y \end{bmatrix} \in Z \mid x \in \mathcal{D}(A^*), y \in \mathcal{D}(D) \right],$$

$$(4.14) \quad C^* y = \sum_{i=1}^k C_i^* y(-b_i) + \int_{-b}^0 C_{01}^*(s) y(s) ds,$$

$\tilde{B} = \begin{bmatrix} C_0^* \\ \Delta^0 \end{bmatrix}$ and Δ is defined as in § 1.

The optimal trajectory $\xi(t)$ is given by

$$(4.15) \quad \xi(t) = \bar{U}_{00}(t_1, t) \xi_1,$$

where

$$\bar{U}(t, s) = \begin{bmatrix} \bar{U}_{00}(t, s) & \bar{U}_{01}(t, s) \\ \bar{U}_{10}(t, s) & \bar{U}_{11}(t, s) \end{bmatrix}$$

is the solution of

$$(4.16) \quad \bar{U}(t, s)z = \tilde{T}(t - s)z - \int_s^t \tilde{T}(r - s) \tilde{B} V^{-1} \tilde{B}^* P(r) \bar{U}(r, s)z dr,$$

and $\tilde{T}(t)$ is the semigroup generated by $\tilde{\mathcal{A}}$. The optimal control as a function of t is

$$(4.17) \quad \begin{aligned} u(t) &= V^{-1} B^* P(t) \bar{U}(t_1, t) \begin{bmatrix} \xi_1 \\ 0 \end{bmatrix} \\ &= V^{-1} [P_{10}(t, 0) \bar{U}_{00}(t_1, t) + P_{11}(t, 0, \cdot) \bar{U}_{10}(t_1, t)] \xi, \end{aligned}$$

where $P_{11}(t, 0, \cdot) \in \mathcal{L}(L_2(-b, 0; R^p), R^p)$ is defined by

$$P_{11}(t, 0, \cdot)h = \int_{-b}^0 P_{11}(t, 0, s)h(s) ds, \quad h \in L_2(-b, 0; R^p).$$

Hence, the optimal filter is given by

$$(4.18) \quad \bar{x}(t_1) = \int_0^{t_1} [\bar{U}_{00}^*(t_1, t)P_{01}(t, 0) + \bar{U}_{01}^*(t_1, t)P_{11}(t, \cdot, 0)]V^{-1} dy(t).$$

From (4.11) it follows that the covariance of the error process $x(t_1) - \bar{x}(t_1)$ is $P_{00}(t_1)$. Thus, we have:

THEOREM 4.1. *The optimal filter for (4.1) and (4.2) is given by (4.18), where $P(t)$ and $\bar{U}(t, s)$ satisfy (4.12) and (4.16), respectively, and $P_{00}(t_1)$ is the covariance of the error $x(t_1) - \bar{x}(t_1)$.*

To solve (4.12) we need to know the explicit form of A^* . In the case of delay equations, it is given in [26].

Appendix 1. Proof of Proposition 1.1. We shall show that

$$\begin{aligned} -D \int_p^t S(t-r)Fu(r)dr &= \int_p^t S(t-r)Bu(r)dr \\ &= \begin{cases} u(t+s), & -b \leq s \leq 0 & \text{if } t-p > b, \\ \begin{cases} u(t+s), & -(t-p) < s \leq 0 \\ 0, & -b \leq s \leq -(t-p) \end{cases} & \text{if } t-p \leq b. \end{cases} \end{aligned}$$

Recall that $[Fu(r)](s) = u(r)$, $-b \leq s \leq 0$. So we have

$$[S(t-r)Fu(r)](s) = \begin{cases} \begin{cases} u(r), & -(t-r) \geq s \geq -b \\ 0, & -(t-r) < s \leq 0 \end{cases} & \text{if } t-r \leq b, \\ \begin{cases} 0, & -b \leq s \leq 0 \end{cases} & \text{if } t-r > b, \end{cases}$$

where 0 stands for the null vector in Y . From this we obtain

$$\left[\int_p^t S(t-r)Fu(r)dr \right](s) = \begin{cases} \int_{t+s}^t u(r)dr, & -b \leq s \leq 0 & \text{if } t-p > b, \\ \begin{cases} \int_{t+s}^t u(r)dr, & -(t-p) < s \leq 0 \\ \int_p^t u(r)dr, & -b \leq s \leq -(t-p) \end{cases} & \text{if } t-p \leq b, \end{cases}$$

which in turn implies that

$$\left[-D \int_p^t S(t-r)Fu(r)dr \right](s) = \begin{cases} u(t+s), & -b \leq s \leq 0 & \text{if } t-p > b, \\ \begin{cases} u(t+s), & -(t-p) < s \leq 0 \\ 0, & -b \leq s \leq -(t-p) \end{cases} & \text{if } t-p \leq b. \end{cases}$$

Next let $w \in V$ and consider

$$\begin{aligned} & \left(\int_p^t S(t-r)Bu(r) dr, w \right) \\ &= \int_p^t (S(t-r)Bu(r), w) dr \\ &= \begin{cases} \int_{t-b}^t \langle u(r), w(r-t) \rangle dr = \int_{-b}^0 \langle u(t+s), w(s) \rangle ds & \text{if } t-p > b, \\ \int_p^t \langle u(r), w(r-t) \rangle dr = \int_{p-t}^0 \langle u(t+s), w(s) \rangle ds & \text{if } t-p \leq b. \end{cases} \end{aligned}$$

Thus,

$$\left[\int_p^t S(t-r)Bu(r) dr \right](s) = \begin{cases} u(t+s), & -b \leq s \leq 0 & \text{if } t-p > b, \\ u(t+s), & -(t-p) < s \leq 0 & \\ 0, & -b \leq s \leq -(t-p) & \text{if } t-p \leq b. \end{cases}$$

Appendix 2. Proof of Lemma 1.2. It is sufficient to show that

$$\left\langle y, \int_{t_0}^t S(t-r)\Delta_\lambda u_\lambda(r) dr \right\rangle \rightarrow \left\langle y, \int_{t_0}^t S(t-r)\Delta u(r) dr \right\rangle.$$

We assume that $t - t_0 > b$. Then by Lemma 1.1, we have

$$\begin{aligned} \left\langle y, \int_{t_0}^t S(t-r)\Delta_\lambda u_\lambda(r) dr \right\rangle &= \langle \lambda R^*(\lambda, D)y, u_\lambda(t+\cdot) \rangle \\ &= \langle \lambda R^*(\lambda, D)y - y, u_\lambda(t+\cdot) \rangle + \langle y, u_\lambda(t+\cdot) \rangle. \end{aligned}$$

Now

$$|\langle \lambda R^*(\lambda, D)y - y, u_\lambda(t+\cdot) \rangle| \leq |\lambda R^*(\lambda, D)y - y|_Y |u_\lambda(t+\cdot)|_Y \rightarrow 0$$

as $\lambda \rightarrow \infty$ since $\lambda R^*(\lambda, D) \rightarrow I$ strongly as $\lambda \rightarrow \infty$ and $|u_\lambda(t+\cdot)|_Y$ is bounded, and

$$\begin{aligned} \lim_{\lambda \rightarrow \infty} \int_{-b}^0 \langle y(s), u_\lambda(t+s) \rangle ds &= \lim_{\lambda \rightarrow \infty} \int_{t-b}^t \langle y(r-t), u_\lambda(r) \rangle dr \\ &= \int_{t-b}^t \langle y(r-t), u(r) \rangle dr \\ &= \left\langle y, \int_{t_0}^t S(t-r)\Delta u(r) dr \right\rangle \end{aligned}$$

by the weak convergence of u_λ to u .

Similarly, we can prove the assertion with $t - t_0 \leq b$.

Appendix 3. Proof of Lemma 1.3. We note that

$$[R(\lambda, D)v](s) = \int_s^0 e^{\lambda(s-r)} v(r) dr, \quad v \in Y$$

and

$$[\Delta_\lambda u](s) = [-\lambda DR(\lambda, D)Fu](s) = \lambda e^{\lambda s} u \quad \text{for each } u \in U.$$

Thus,

$$\Delta_{\lambda}^* y = \int_{-b}^0 \lambda e^{\lambda s} y(s) ds.$$

Since $y(s)$ is continuous at $s=0$, for each $\varepsilon > 0$ there exists a $0 < \delta < b$ such that $|y(s) - y(0)| < \varepsilon$ for $-\delta < s < 0$. Now

$$\Delta_{\lambda}^* y - y(0) = \int_{-\delta}^0 \lambda e^{\lambda s} y(s) ds - y(0) + \int_{-b}^{-\delta} \lambda e^{\lambda s} y(s) ds.$$

The second term goes to zero as $\lambda \rightarrow \infty$ by the dominated convergence theorem and

$$\begin{aligned} \left| \int_{-\delta}^0 \lambda e^{\lambda s} y(s) ds - y(0) \right| &= \left| \int_{-\delta}^0 \lambda e^{\lambda s} [y(s) - y(0)] ds - e^{-\lambda \delta} y(0) \right| \\ &\leq \varepsilon + e^{-\lambda \delta} |y(0)|. \end{aligned}$$

Hence, $\lim_{\lambda \rightarrow \infty} \Delta_{\lambda}^* y = y(0)$.

Appendix 4. Proof of Lemma 2.1. It is sufficient to consider

$$\begin{aligned} \left| \int_{t_0}^t S(t-r) \Delta_{\lambda} u(r) dr \right| &= \left| -\lambda R(\lambda, D) D \int_{t_0}^t S(t-r) F u(r) dr \right| \\ &\leq 2 \left| D \int_{t_0}^t S(t-r) F u(r) dr \right| \quad \text{for large } \lambda \\ &\leq 2 \sqrt{\int_{-b}^0 |u(t+s)|^2 ds} \leq 2 \sqrt{\int_{t_0}^{t_1} |u(t)|^2 dt}. \end{aligned}$$

Acknowledgment. I would like to thank R. F. Curtain for some helpful discussions.

REFERENCES

- [1] Y. ALEKAL, P. BRUNOVSKY, D. H. CHYUNG AND E. B. LEE, *The quadratic problem for systems with time delays*, IEEE Trans. Automat. Contr., AC-16 (1971), pp. 673-687.
- [2] A. BAGCHI, *A martingale approach to state estimation in delay-differential systems*, J. Math. Anal. Appl., 56 (1976), pp. 195-210.
- [3] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer-Verlag, New York, 1976.
- [4] J. G. BORISOVIC AND A. S. TURBABIN, *On the Cauchy problem for linear nonhomogeneous differential equation with retarded arguments*, Soviet Math. Dokl., 10 (1969), pp. 401-405.
- [5] R. F. CURTAIN, *Infinite dimensional estimation theory for linear systems*, Control Theory Centre Report 38, University of Warwick, England 1976.
- [6] ———, *Stochastic distributed systems with point observations and boundary control: An abstract theory*, Stochastics, 3 (1979), pp. 85-104.
- [7] R. F. CURTAIN AND A. J. PRITCHARD, *The infinite dimensional Riccati equation for systems described by evolution operators*, this Journal, 14 (1976), pp. 951-983.
- [8] ———, *An abstract theory for unbounded control action for distributed parameter systems*, this Journal, 15 (1977), pp. 566-611.
- [9] ———, *Infinite Dimensional Linear System Theory*, Lecture Notes in Control and Information Sciences 8, Springer-Verlag, New York, 1978.
- [10] M. C. DELFOUR AND S. K. MITTER, *Controllability, observability and optimal feedback control of affine hereditary differential systems*, this Journal, 10 (1972), pp. 298-327.
- [11] M. C. DELFOUR, C. MCCALLA AND S. K. MITTER, *Stability and infinite-time quadratic cost problem for linear hereditary differential systems*, this Journal, 13 (1975), pp. 48-88.
- [12] H. O. FATTORINI, *Boundary control systems*, this Journal, 6 (1968), pp. 349-385.

- [13] J. S. GIBSON, *The Riccati integral equations for optimal control problems on Hilbert spaces*, this Journal, 17 (1979), pp. 537–565.
- [14] A. ICHIKAWA, *Evolution equations with delay*, Control Theory Centre Report 52, University of Warwick, England, 1977.
- [15] ———, *Optimal control and filtering of evolution equations with delay in control and observations*, Control Theory Centre Report No. 53, University of Warwick, England, 1977.
- [16] ———, *Dynamic programming approach to stochastic evolution equations*, this Journal, 17 (1979) pp. 152–174.
- [17] H. N. KOIVO AND E. B. LEE, *Controller synthesis for linear systems with retarded state and control variables and quadratic cost*, Automatica, 8 (1972), pp. 203–208.
- [18] R. H. KWONG AND A. S. WILLSKY, *Estimation and filter stability of stochastic delay systems*, this Journal, 16 (1978), pp. 660–681.
- [19] R. H. KWONG, *A stability theory for the linear-quadratic-Gaussian problem for systems with delays in the state, control and observations*, this Journal, 18 (1980), pp. 49–75.
- [20] I. LASIECKA, *Unified theory for abstract parabolic boundary problems: a semigroup approach*, Appl. Math. Optim., 6 (1980), pp. 287–333.
- [21] A. LINDQUIST, *A theorem on duality between estimation and control for linear stochastic systems with time delay*, J. Math. Anal. Appl., 37 (1972), pp. 516–536.
- [22] S. K. MITTER AND R. B. VINTER, *Filtering for linear stochastic hereditary differential systems*, Lecture Notes in Economics and Mathematical Systems 107, Springer-Verlag, New York, 1975, pp. 1–21.
- [23] J. Y. OUVARD, *Martingale projection and linear filtering in Hilbert spaces. II: The theory*, this Journal, 16 (1978), pp. 912–937.
- [24] ———, *Linear filtering in Hilbert spaces. II: An application to the smoothing theory for hereditary systems with observation delays*, this Journal, 16 (1978), pp. 938–952.
- [25] R. TRIGGIANI, *Boundary feedback stabilizability of parabolic equations*, Appl. Math. Optim., 6 (1980), pp. 201–220.
- [26] 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.
- [27] R. B. VINTER AND R. H. KWONG, *The infinite time quadratic control problem for linear systems with state and control delays: An evolution equation approach*, this Journal 19 (1981), pp. 139–153.
- [28] J. ZABCZYK, *On decomposition of generators*, this Journal, 16 (1978), pp. 523–534.
- [29] ———, *On systems with delays in control*, Report CRM-777, CRM, Université de Montreal, Canada, 1978.
- [30] ———, *On stabilizability of boundary control systems*, Report CRM-785, CRM, Université de Montreal, Canada, 1978.

CAUSALITY AND STABILITY OF LINEAR SYSTEMS DESCRIBED BY PARTIAL DIFFERENTIAL OPERATORS*

YAKAR KANNAI†

Abstract. Definitions are suggested for the concepts of causality and stability of linear systems described by partial differential operators with constant coefficients. These definitions are modelled after the one-dimensional case. The operators having causal and stable fundamental solutions are characterized algebraically, and Nyquist-type criteria are established.

1. Introduction and statement of results. In the theory of linear systems [1], [5], it is said that the linear system described by the ordinary differential operator (with constant coefficients)

$$(1.1) \quad Lu = \sum_{j=0}^m a_j \frac{d^j u}{dt^j}$$

is *stable* if either one of the following three equivalent conditions is satisfied:

- (i) All the roots of the polynomial $\sum_{j=0}^m a_j s^j$ have negative real parts.
- (ii) The impulse response $F(t)$, defined by $LF(t) = \delta(t)$ (where $\delta(t)$ is the Dirac measure) and $F(t) = 0$ for $t < 0$, is exponentially decreasing as t tends to $+\infty$.
- (iii) B.I.B.O.: if $f(t)$ is a continuous bounded function in $[0, \infty)$, then the solution $u(t)$ of $Lu = f$ with the initial condition $u(0) = \cdots = u^{(m-1)}(0) = 0$ is also bounded in $[0, \infty)$.

Note that (ii) can be weakened in our case to (ii'): $F(t)$ tends to zero as t tends to $+\infty$, and that (iii) is equivalent—by the closed graph theorem or otherwise—to (iii'): There exists a constant C such that if $|f(t)| \leq 1$ for $t \in [0, \infty)$, then $|u(t)| \leq C$ for $t \in [0, \infty)$. Note also that the argument principle can be applied for determination of stability from (i); one determines the change of $\arg \sum a_j (iw)^j$ as w goes from $-\infty$ to $+\infty$. The resulting criterion for stability is called the Nyquist criterion [5, p. 417].

There are several conceivable possibilities for developing an analogous theory for systems governed by partial differential operators (with constant coefficients). Guided by the desire to use transform techniques (and thus to obtain conditions as similar as possible to (i), say), we are led to consider problems in cones (rather than in domains with more complicated boundaries). Thus, let V be a closed, convex cone in R^n with a nonempty interior (this includes the case in which both the time variable t and the space variables x are constrained to be nonnegative, as well as the case in which only t is restricted to be nonnegative; in the sequel we will make no distinction between space and time variables). Let $P(D)$ be a nonzero partial differential operator with constant coefficients. We will use the standard notation [3] for such operators. Note in particular that $D = ((1/i)\partial/\partial x_1, \dots, (1/i)\partial/\partial x_n)$ (thus we use Fourier—rather than Laplace—transforms). We are looking for a temperate fundamental solution F of $P(D)$ with special properties. This fundamental solution F should play the role of the impulse response. While it is always possible in the o.d.e. case to find F such that $\text{supp } F \subset [0, \infty)$, the analogous condition has to be postulated explicitly in the p.d.e. case:

Causality. We say that F is *causal* if $\text{supp } F \subset V$.

* Received by the editors July 24, 1981.

† Department of Theoretical Mathematics, Weizmann Institute of Science, Rehovot, Israel. The work of this author represented in this paper was carried out at the Weizmann Institute of Science, Rehovot, Israel, and at the University of Minnesota, Minneapolis, Minnesota.

Note that F is causal if and only if $\text{supp}(F * \varphi) \subset V$ whenever $\text{supp } \varphi \subset V$, and the convolution $F * \varphi$ makes sense.

The space of continuous and bounded functions does not appear to be very natural in the study of partial differential equations. It seems that the following definition is the appropriate one (for $n > 1$).

Stability. We say that F is *stable* if $F * \varphi \in \mathcal{S}$ for every $\varphi \in \mathcal{S}$, where \mathcal{S} is the Schwartz space of infinitely differentiable functions in R^n which, together with all their derivatives, decay at infinity faster than any (negative) power of $|x|$, and $F * \varphi$ denotes the convolution of F and φ .

We can now state our main result.

THEOREM 1. *The operator $P(D)$ has a causal and stable temperate fundamental solution if and only if $P(\zeta)$ never vanishes for any complex vector ζ with $\text{Im } \zeta \in -V^*$, where V^* is the dual cone of V , i.e., $V^* = \{y \in R^n : \langle y, x \rangle \geq 0 \text{ for all } x \in V\}$.*

Theorem 1 seems to establish the equivalence of the analogues of (i), (ii') and (iii). What about (ii)? Since F is only a distribution and not a function (and thus F does not necessarily possess a value at a given point), we have to be precise about what we mean by exponential decrease of F . What has to decrease is not F itself but its averages, $F * \varphi$, for $\varphi \in C_0^\infty(R^n)$. ($F * \varphi$ would not, in general, decrease exponentially if we take φ to be only in \mathcal{S} .) We will say that F is exponentially decreasing if there exists a positive constant C such that for all $\varphi \in C_0^\infty(R^n)$, $|(F * \varphi)(x)| = O(\exp(-C \text{dist}(x, V \cap (-V))))$. We have the following result.

THEOREM 2. *The operator $P(D)$ has a causal and stable, exponentially decreasing fundamental solution F if and only if there exists a positive constant C such that $P(\zeta) \neq 0$ if $\langle \text{Im } \zeta, x \rangle \leq C \text{dist}(x, V \cap (-V))$ for all $x \in V$.*

For example, if $V = \{x : x_i \geq 0 \text{ for all } 1 \leq i \leq n\}$, then $V^* = V$ and the algebraic condition of Theorem 1 states that $P(\zeta) \neq 0$ if $\text{Im } \zeta_j \leq 0$ for all $1 \leq j \leq n$, whereas in Theorem 2, it is required that there exists an $\varepsilon > 0$ such that $P(\zeta) \neq 0$ if $\text{Im } \zeta_j \leq \varepsilon$ for all $1 \leq j \leq n$. Note that in the one-dimensional case, the roots of a polynomial form a finite set, so that if their real parts are negative, they are uniformly bounded by a negative number. If $n > 1$, then the variety $\{\zeta : P(\zeta) = 0\}$ might approach asymptotically a cone, even if the variety is disjoint from the cone. Consider, for example, the polynomial (in two variables)

$$P(\zeta_1, \zeta_2) = (\zeta_1 - i)\zeta_2 - 1.$$

Then $P(\zeta_1, \zeta_2) \neq 0$ if $\text{Im } \zeta_j \leq 0$, $j = 1, 2$, but P vanishes on the curve $\zeta_1 = -s$, $\zeta_2 = (i - s)/(1 + s^2)$, and this curve is asymptotically approaching the cone as s tends to $+\infty$. Hence (ii) and (ii') are no longer equivalent. It can be debated whether exponential decrease is really important in the applications or if Theorem 1 contains all the relevant information.

Temperate fundamental solutions supported in cones have been investigated earlier (see [2], [8] and the references quoted there). Our interest, however, lies in temperature fundamental solutions which are both causal and stable; this actually simplifies a number of the arguments. Nevertheless, we will quote in § 2 the appropriate theorem from [2] and then apply it to the proof of Theorem 1.

It is desirable to have criteria, similar to the Nyquist criterion, which will enable one to determine whether indeed the assumptions of Theorem 1 are satisfied. It turns out that if we know already that P has no real roots (i.e., $P(\xi) \neq 0$ if $\xi \in R^n$) then $P(\zeta) \neq 0$ for $\text{Im } \zeta \in -V^*$ if $P(\zeta)$ does not vanish for $\text{Im } \zeta$ in any extremal ray of $-V^*$. (This fact follows from a local version of Bochner's tube theorem [6]; we will derive it directly from Theorem 1.) If $-V^*$ is finitely generated (as in the special case

$V = \{x : x_i \geq 0 \text{ for all } 1 \leq i \leq n\}$, then it suffices to apply Nyquist criterion finitely many times. Those matters will be discussed in § 3.

2. Proofs of Theorems 1 and 2. Recall that a closed, convex cone $V \subset R^n$ is said to be salient (or proper) if $V \cap (-V) = \{0\}$. In general, $V \cap (-V)$ is a linear subspace of R^n ; we shall use the notations $n' = \dim(V \cap (-V))$, $n'' = n - n'$, and coordinates $x = (x', x'')$ such that $V \cap (-V)$ is defined by $x'' = 0$. There exists a closed, convex, salient cone W in $R^{n''}$ such that $V = R^{n'} \oplus W$. If V has a nonempty interior in R^n (as we assume throughout), then W has a nonempty interior (relative to $R^{n''}$).

The following result from [2] (Part of Theorem 1 of [2]) characterizes completely differential operators having temperate fundamental solutions supported in cones, i.e., causal operators.

THEOREM A. *The following conditions on $V = W \oplus R^{n'}$ and the differential operator $P(D)$ are equivalent.*

(2.1) $P(D)$ has a temperate fundamental solution with support in V .

(2.2) For every $\xi' \in R^{n'}$, either $P(\xi', \zeta'') \neq 0$, if $\text{Im } \zeta'' \in -\text{int } W^*$, or $P(\xi', \zeta'') = 0$ for all $\zeta'' \in C^{n''}$.

If V is salient, then $n' = 0$, $n'' = n$ and $W = V$. In that case (2.2) reads (recall that P is assumed to be nonzero):

(2.3) $P(\zeta) \neq 0$ if $\text{Im } \zeta \in -\text{int } V^*$.

Suppose that $P(\zeta) \neq 0$ if $\text{Im } \zeta \in -V^*$. Then $P(\xi) \neq 0$ for all real vectors ξ . It follows easily from a corollary of the Seidenberg–Tarski theorem [3, p. 276] that there exist real constants C and α with $C > 0$ such that $|P(\xi)| \geq C(1 + |\xi|)^\alpha$ for all $\xi \in R^n$. Hence, $1/P(\xi)$ is a temperate distribution and as such its inverse Fourier transform $F(x)$ exists too as a temperate distribution. Moreover, $1/P(\xi) \in C^\infty(R^n)$ and the derivatives of $1/P$ have at most a polynomial growth in $|\xi|$. Hence if $\varphi \in \mathcal{S}$, so that $\hat{\varphi} \in \mathcal{S}$, we have $\hat{\varphi}/P \in \mathcal{S}$ too. It follows that $F * \varphi \in \mathcal{S}$. That $\text{supp } F \subset V$ can either be proved directly, applying the Paley–Wiener theorem, or else by applying Theorem A, according to which $P(D)$ has a temperate fundamental solution F_1 with support in V . By construction, F is a temperate fundamental solution of $P(D)$. Hence the temperate distribution $u = F - F_1$ is a solution of $P(D)u = 0$. Then $P(\xi)\hat{u}(\xi) = 0$. But $P(\xi)$ has no real zeros. Hence $\hat{u}(\xi) = 0$, and $F = F_1$. Thus the operator $P(D)$ has a causal and stable temperate fundamental solution.

Suppose, conversely, that the operator $P(D)$ has a causal and stable fundamental solution F . Then $\hat{F} \in \mathcal{S}'$ and $P(\xi)\hat{F}(\xi) = 1$. Hence $\hat{F}(\xi)$ is a smooth function in an open dense subset of R^n , namely, in the set $\{\xi \in R^n; P(\xi) \neq 0\}$. For any $\xi_0 \in R^n$, choose $\hat{\varphi} \in \mathcal{S}$ such that $\hat{\varphi}(\xi_0) \neq 0$. Then

$$(\widehat{F\hat{\varphi}})(x) = (F * \varphi)(-x) \in \mathcal{S}.$$

Hence the temperate distribution $\widehat{F\hat{\varphi}}$ is actually C^∞ smooth and $\xi_0 \notin \text{sing supp } \hat{F}$ (this argument is modelled after wave-front set considerations; compare [4]). Hence $\hat{F} \in C^\infty(R^n)$ and $P(\xi) \neq 0$ for all $\xi \in R^n$. In particular, if V is not salient, then the polynomial $P(\xi', \zeta'')$ can vanish identically in $\zeta'' \in C^{n''}$ for no real vector $\xi' \in R^{n'}$. Hence, the first alternative in (2.2) must hold, i.e., $P(\xi', \zeta'') \neq 0$ if $\xi' \in R^{n'}$ and $\text{Im } \zeta'' \in -\text{int } W^*$. (There is no ξ' if V is salient.) Assume now that $P(\xi_0 + i\eta_0) = 0$ for some $\xi_0 \in R^n$ and $\eta_0'' \in -\partial W^*$. Then necessarily $\eta_0 \neq 0$. Set (for $z \in C$)

$$P(\xi + z\eta) = \sum_{j=0}^m a_j(\xi, \eta)z^j.$$

Then $a_0(\xi_0, \eta_0) = P(\xi_0) \neq 0$. Hence, the polynomial in z , $P(\xi_0 + z\eta_0)$, is not identically zero. This polynomial vanishes at $z = i$. The coefficients $a_j(\xi, \eta)$ depend continuously (in fact polynomially) on η . By Rouché's theorem, the polynomial $P(\xi_0 + z\eta)$ has a zero near $z = i$ if η is sufficiently close to η_0 . Let η_1 , with $\eta'_1 = 0$, $n'_1 \in -\text{int } W^*$, be near η_0 . Then there exists a complex number $\varepsilon_1 + i\varepsilon_2$ with $|\varepsilon_i| < 1$ such that $P(\xi_0 + (i - \varepsilon_1 - i\varepsilon_2)\eta_1) = 0$. But

$$\text{Im} [\xi_0 + (i - \varepsilon_1 - i\varepsilon_2)\eta_1]'' = -(1 - \varepsilon_2)\eta''_1 \in -\text{int } W^*,$$

a contradiction. It follows that $P(\xi + i\eta) \neq 0$ for all $\xi \in R^n$ and η such that $\eta' = 0$, $\eta'' \in -W^*$, not just $\eta'' \in -\text{int } W^*$. If V is salient, then $W^* = V^*$. If V is not salient, then $V^* = \{0\} \oplus W^*(0 \in R^n)$. Hence in all cases $P(\xi + i\eta) \neq 0$ for all $\xi \in R^n$ and $\eta \in -V^*$, and the proof of Theorem 1 is completed.

Suppose now that $P(\xi) \neq 0$ if $\langle \text{Im } \zeta, x \rangle \leq C|x''|$ all $x \in V$ and some positive constant C and choose $a \in V^*$ so that $a'' \in \text{int } W^*$ and $|a| < C$. By Theorem 1, $P(D)$ has a causal and stable temperate fundamental solution F with $\hat{F}(\xi) = 1/P(\xi)$. Let $\varphi \in C_0^\infty(R^n)$ be arbitrary. By the Paley–Wiener theorem (in the formulation given in [3]), there exists a constant A and for every N there exists a constant C_N such that $\hat{\varphi}$ is an entire function and

$$(2.4) \quad |\hat{\varphi}(\xi + i\eta)| \leq C_N e^{A|\eta|} (1 + |\xi|)^{-N} \quad \text{for } \xi, \eta \in R^n.$$

An application of the Seidenberg–Tarski theorem implies the existence of real constants D and α with $D > 0$ such that

$$(2.5) \quad |P(\xi + i\lambda a)| \geq D(1 + |\xi|)^\alpha$$

for all $\xi \in R^n$ and $0 \leq \lambda \leq 1$ (note that $\langle \lambda a, x \rangle \leq C|x''|$ if $x \in V$ and $0 \leq \lambda \leq 1$). Hence the integration used for computing $(F * \varphi)(x)$ can be pushed to the complex domain:

$$(2.6) \quad (F * \varphi)(x) = (2\pi)^{-n/2} \int_{R^n} \frac{e^{i\langle x, \xi \rangle} \hat{\varphi}(\xi)}{P(\xi)} d\xi = (2\pi)^{-n/2} \int_{R^n} \frac{e^{i\langle x, \xi + ia \rangle} \hat{\varphi}(\xi + ia)}{P(\xi + ia)} d\xi.$$

Applying (2.4) and (2.5) to (2.6), we infer that there exists a constant E , independent of x , such that

$$(2.7) \quad |(F * \varphi)(x)| \leq E e^{-\langle x, a \rangle}.$$

On the compact set $W \cap \{x : |x| = 1\}$, both $\langle x, a \rangle$ and $\text{dist}(x, R^n) = |x''| = \text{dist}(x, V \cap (-V))$ are positive, and both are homogeneous of degree 1 on V . Hence there exists a constant C_2 such that $\text{dist}(x, V \cap (-V)) \leq C_2 \langle x, a \rangle$ on V . The estimate (2.7) yields the desired exponential decrease on V . Let R be such that $x \in \text{supp } \varphi$ implies that $\|x\| \leq R$. If $\text{dist}(x, V) > R$ then $(F * \varphi)(x) = 0$. If $x \notin V$ but $\text{dist}(x, V) \leq R$ then $x = y + z$ with $y \in V$ and $\|z\| \leq R$, and

$$(2.8) \quad i\langle x, \xi + ia \rangle = i\langle x, \xi \rangle - \langle y, a \rangle - \langle z, a \rangle.$$

If $\text{dist}(x, V \cap (-V)) \geq 2R$ (the other case does not matter since $(F * \varphi)(x)$ is bounded) then $2 \text{dist}(y, V \cap (-V)) \geq \text{dist}(x, V \cap (-V))$, and there exists a constant C_3 such that $\langle y, a \rangle \geq C_3 \text{dist}(x, V \cap (-V))$. Applying (2.8) to (2.6), we thus obtain the exponential decrease in this case too. It follows that F is exponentially decreasing in the sense given to that term in the introduction.

Suppose, conversely, that $P(D)$ has a causal and stable temperate fundamental solution F such that F is exponentially decreasing, i.e., there exists a constant C such

that for all $\varphi \in C_0^\infty(R^n)$ there exists a constant C_φ with

$$(2.9) \quad |(F * \varphi)(x)| \leq C_\varphi \exp[-C \operatorname{dist}(x, R'^n)] = C_\varphi \exp[-C|x''|].$$

Then for every $\eta \in R^n$ with $\langle \eta, x \rangle < C|x''|$ for $x \in V$, the function $e^{\langle x, \eta \rangle} (F * \varphi)(x)$ is integrable in R^n for all $\varphi \in C_0^\infty(V)$ (note that by causality $\operatorname{supp}(F * \varphi) \subset V$ if $\operatorname{supp} \varphi \subset V$), and the Fourier transform

$$(2.10) \quad (\widehat{F * \varphi})(\xi + i\eta) = (2\pi)^{-n/2} \int e^{-i\langle x, \xi + i\eta \rangle} (F * \varphi)(x) dx$$

can be continued analytically (from the set $\eta \in -V^*$, where known to exist by Theorem 1) to the set $U = \{\eta : \langle \eta, x \rangle < C|x''| \text{ for all } x \in V\}$. For any vector $\zeta \in C^n$, there exists $\varphi \in C_0^\infty(V)$ with $\hat{\varphi}(\zeta) \neq 0$. (In fact, let $\Psi \in C_0^\infty(V)$ be such that $\hat{\Psi}(0) \neq 0$ and choose $\varphi(x) = \Psi(x)C^{i\langle x, \zeta \rangle}$.) Hence $\hat{F}(\zeta)$ must also admit analytic continuation to the set U . But $\hat{F}(\zeta) = 1/P(\zeta)$ if $\operatorname{Im} \zeta \in -V^*$. Hence $1/P(\zeta)$ is also analytic in U , i.e., $P(\zeta) \neq 0$ if $\operatorname{Im} \zeta \in U$, and Theorem 2 is proved.

3. Sufficient conditions for stability. Let $V \subset R^n$ be a convex cone with a nonempty interior and let $P(D)$ be a nonzero differential polynomial. By Theorem 1 one has to find out whether $P(\zeta)$ does not vanish on the "tube domain" $R^n - iV^*$ in order to determine whether $P(D)$ has a causal and stable fundamental solution. Note that V^* is salient if V has a nonempty interior. Thus the algebraic problem is the following: Given a nonzero polynomial P and a salient closed convex cone $K (= -V^*)$, is it true that $P(\zeta) \neq 0$ for all ζ such that $\operatorname{Im} \zeta \in K$? It turns out that it suffices to restrict the attention to a lower dimensional subset of $R^n + iK$. Let $E(K)$ denote the union of the set of extremal rays of the cone K and the origin. Note that the salient cone K is the convex hull of $E(K)$.

THEOREM 3. *The polynomial P does not vanish on $R^n + iK$ if and only if P does not vanish on $R^n + iE(K)$.*

Proof. The only nontrivial statement is that if P does not vanish on $R^n + iE(K)$ then P does not vanish on $R^n + iK$. The nonvanishing of P on R^n implies that $P(D)$ has a unique temperate fundamental solution $F(x)$. Let $K_a = \{\lambda a\}$, $a \in K$, $\lambda > 0$, be an extremal ray of K . By assumption, $P(\zeta) \neq 0$ if $\operatorname{Im} \zeta \in K_a \cup \{0\}$ ($0 \in R^n$). By Theorem 1, $\operatorname{supp} F \subset \{x : \langle x, a \rangle \leq 0\}$. But the rays K_a generate K . Hence the intersection of the closed half-spaces $\{x : \langle x, a \rangle \leq 0\}$ coincides with $-K^*$ and $\operatorname{supp} F \subset -K^*$. By Theorem 1, $P(\zeta) \neq 0$ if $\operatorname{Im} \zeta \in -(-K^*)^* = K$.

A different proof of Theorem 3 can be obtained by applying a local version of Bochner's tube theorem [6] to the holomorphic function $1/P$.

It follows from Theorem 3 that we need a method for checking whether or not $P(\xi + i\lambda a) \neq 0$ for all $\xi \in R^n$ and $\lambda > 0$, for a fixed vector $a \in R^n$ (given that $P(\xi) \neq 0$ for all $\xi \in R^n$). Choose coordinates so that $a = (0, \dots, 0, 1)$, and set $\xi = (\xi', \xi_n)$. We thus need a criterion to determine whether or not $P(\xi', \xi_n) \neq 0$ for $\operatorname{Im} \xi_n > 0$. Set

$$(3.1) \quad P(\xi', \xi_n) = \sum_{j=0}^m a_j(\xi') \xi_n^j.$$

The set $\{\xi' : a_m(\xi') \neq 0\}$ consists of finitely many connected open subsets $\Omega_1, \dots, \Omega_k$ of R^{n-1} [7, pp. 11, 108]. Let $N(\xi')$ denote the number (including multiplicities) of zeros of $P(\xi', \xi_n)$ with positive imaginary part. Then $N(\xi')$ is constant in each Ω_j , $j = 1, \dots, k$, since Rouché's theorem and the continuity of the coefficient a_j imply that $N(\xi')$ is locally constant in a neighborhood of a point where $a_m(\xi') \neq 0$. Hence, it suffices to compute $N(\xi')$ for k points, one in each set Ω_j . Thus choose $\xi'_j \in \Omega_j$ and

let $R > 0$ be sufficiently large so that for no $|\zeta_n| > R - 1$ one has a zero of $P(\xi'_j, \zeta_n)$. Let Δ denote the increments of $\arg P(\xi'_j, \zeta_n)$ as ζ_n changes from $-R$ to $+R$ (along the real axis). Then by the Nyquist criterion [5, p. 417], we see that $N(\xi'_j) = 0$ if Δ is close to $m\pi$.

If K is finitely generated, it follows that finitely many Nyquist-like determinations suffices to check whether or not $P(D)$ has a causal and stable fundamental solution.

Acknowledgments. I am very much indebted to I. Horowitz for suggesting the problem, to E. Bedford for explaining to me how the argument principle extends to polydiscs, and to W. Littman and A. Tannenbaum for helpful conversations.

Note added in proof. Fundamental solutions supported in cones (for systems) were constructed by M. Eskin and E. Shamir, *Elliptic solvability of augmented differential complexes on piecewise smooth manifolds* (to appear).

REFERENCES

- [1] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.
- [2] A. ENQVIST, *On temperate fundamental solutions supported by a convex cone*, Ark. Mat., 14 (1976), pp. 35–41.
- [3] L. HÖRMANDER, *Linear Partial Differential Operators*, Springer, Berlin, 1963.
- [4] ———, *Fourier integral operators*, I, Acta. Math., 127 (1971), pp. 79–183.
- [5] W. KAPLAN, *Operational Methods for Linear Systems*, Addison-Wesley, Reading, MA, 1962.
- [6] H. KOMATSU, *A local version of Bochner's tube theorem*, J. Fac. Sci. Univ. Tokyo, Sect. 1A, 19 (1972), pp. 201–214.
- [7] J. MILNOR, *Singular points of complex hypersurfaces*. Ann. Math. Studies 61, Princeton University Press, Princeton, NJ, 1968.
- [8] B. E. PETERSON, *Holomorphic functions with bounds*, Complex Analysis and its Applications, vol. III, International Atomic Energy Agency, Vienna, 1976, pp. 107–120.

STABLE AND REGULAR REACHABILITY FOR RELAXED HEREDITARY DIFFERENTIAL SYSTEMS*

FRITZ COLONIUS†

Abstract. We investigate reachability properties of relaxed nonlinear hereditary differential systems in the state space $W^{n,\infty}[-r, 0]$. Reachability of the relaxed system is equivalent to approximate reachability of the ordinary system. Local reachability for the relaxed system can be obtained under much weaker conditions than for the ordinary system.

The notion of stable reachability allows us to relate reachability of points in \mathbb{R}^n and states in $W^{n,\infty}[-r, 0]$. Regular reachability is of primary importance in optimal control problems with function space end condition, because in case of regularity Lagrange multipliers in the dual of $W^{n,\infty}[-r, 0]$ can be identified with functions in $W^{n,\infty}[-r, 0]$. Regularly reachable final states and regular trajectories are characterized and it is shown that regularity is a generic property of trajectories. Examples are given where all trajectories reaching a certain final state are regular.

Introduction. In this paper, we study reachability properties of the following nonlinear hereditary differential system Σ :

$$(0.1) \quad \dot{x}(t) = f(x_t, u(t), t) \quad (t \in T := [t_0, t_1]),$$

$$(0.2) \quad x_{t_0} = \varphi_0,$$

$$(0.3) \quad u(t) \in \Omega(t) \quad (t \in T),$$

where $0 \leq r < \infty$, $t_1 - r > t_0$, $f: C^n[-r, 0] \times \mathbb{R}^m \times T \rightarrow \mathbb{R}^n$, $\Omega(t) \subset \mathbb{R}^m$ and $\varphi_0 \in C^n[-r, 0]$ are fixed and

$$x_t(s) := x(t+s), \quad s \in [-r, 0].$$

r denotes the length of the system memory. The state of this system is given by the function segment x_t and the reachability theory depends essentially on the choice of the infinite dimensional state space Z .

Jacobs/Kao [15], Banks/Jacobs/Langenhop [2], [3], Colonius/Hinrichsen [11] dealt with complete exact reachability in the Sobolev space $Z = W^{n,2}[-r, 0]$ for unconstrained linear systems. This work was primarily motivated by fixed final state optimal control problems with unconstrained (resp. energy constrained) controls $u \in L_2^m(T)$. A natural choice of Z for problems with pointwise constrained controls in L_2^m —as considered here—is $Z = W^{n,\infty}[-r, 0]$. (This compatibility requirement on the state and control function spaces is exposed very clearly by Kurczyk/Olbrot [17].)

However, as is well known, there is a severe drawback of reachability theory in $W^{n,p}[-r, 0]$, $1 \leq p \leq \infty$: complete exact reachability for linear systems can only be guaranteed if the dimension m of the control space is not less than the dimension n of the phase space. Hence it appears reasonable to replace exact reachability by approximate reachability. This concept, in fact, has been studied for unconstrained linear systems by many authors (see, e.g., Delfour/Mitter [12], Manitius/Triggiani [20], Manitius [19] in $Z = \mathbb{R}^n \times L_2^n[-r, 0]$, Olbrot [23] in various Banach spaces). While this turned out successful for certain purposes (e.g., feedback stabilizability and the infinite time linear quadratic optimal control problem), it is inadequate for fixed final state optimal control problems. For these problems exact reachability of the linearized system is necessary in order to guarantee the existence of Lagrange multipliers (see

* Received by the editors November 11, 1980, and in final revised form October 17, 1981.

† Forschungsschwerpunkt Dynamische Systeme, Universität Bremen, Bibliothekstraße, Postfach 330 440, 2800 Bremen 33, West Germany.

Kurcyusz [16]). Furthermore, with respect to the phase space $Z = W^{n,\infty}[-r, 0]$, Olbrot [23, Thm. 5.3] has shown that for linear unconstrained systems approximate reachability is equivalent to exact reachability; i.e., nothing is gained by allowing approximate reachability. In this paper we propose the following way out of this apparent dilemma: we study instead of the system (0.1), (0.2), $(\widehat{0.3})$ the *relaxed system* in the sense of Warga [27]. That is, we consider instead of measurable control functions u satisfying $(\widehat{0.3})$ relaxed (measure-valued) control functions

$$(0.3) \quad v \in \mathcal{S}^{\#}$$

and replace u in (0.1) by v .

Exact reachability with relaxed controls is equivalent to approximate reachability with ordinary *pointwise constrained controls* (see Theorem 2.1 and Remark 2.1 below).

It is well known that for linear systems final states which are not exactly reachable can only be approximated by applying an *unbounded* sequence of control functions (see Kurcyusz [16]). We show, however, that neither this nor Olbrot's result on the equivalence of exact and approximate reachability in $W^{n,\infty}[-r, 0]$ remains true if the *control appears nonlinearly*. Here local reachability of the relaxed system is a much weaker property than local reachability of the ordinary system. In particular, the condition $m \geq n$ on the dimensions m and n of the control and phase space, respectively, is not necessary for local reachability of the relaxed system. The consequences for optimal control problems are studied in the companion paper [10].

Since the controls are bounded we study local reachability. Now suppose we can reach a certain final state φ_1 with a trajectory x^0 corresponding to a control $v^0 \in \mathcal{S}^{\#}$. Then in order to reach φ in a neighborhood of φ_1 in $W^{n,\infty}[-r, 0]$, we first have to reach $\varphi(-r)$ at time $t_1 - r$ and then trace exactly the velocity function $\dot{\varphi}(t - t_1)$ on $[t_1 - r, t_1]$.

In order to achieve this we introduce the concept of *stable reachability*. This means that a complete neighborhood of $\varphi_1(-r) \in \mathbb{R}^n$ can be reached with arbitrarily small deviations from x^0 . Then the hereditary effects influencing the velocity on $[t_1 - r, t_1]$ can be kept small, and in order to reach a complete neighborhood of φ_1 in $W^{n,\infty}[-r, 0]$ we only have to compensate small deviations from $\varphi_1 = x_{t_1}^0$.

If a complete neighborhood of $\dot{\varphi}_1(t - t_1) = \dot{x}^0(t) = f(x_t^0, v^0(t), t)$, $t \in [t_1 - r, t_1]$ in $L_{\infty}^n[t_1 - r, t_1]$ can be covered by altering the control v^0 only on $[t_1 - r, t_1]$, we say that φ_1 is reached *regularly* with x^0 (for exact definitions see Definitions 2.1, 2.2, 3.1). The notion of regular reachability is of primary importance for optimal control problems with function space end condition, because in case of regularity Lagrange multipliers in the dual of $W^{n,\infty}[-r, 0]$ can be identified with functions in $W^{n,\infty}[-r, 0]$ (see Colonius [9], [10] (this issue, pp. 695–712)). The analysis of the somewhat delicate relations among local, stable and regular reachability constitutes the main content of this paper. The results extend and sharpen those of Colonius [9]. After the preliminary § 1, we discuss in § 2 the relation between stable and local reachability. In particular, it turns out that both notions are equivalent for linear relaxed systems, while for nonlinear systems there may be final states which are locally but not stably reachable. In § 3, we show that regular reachability plus a stable reachability property in \mathbb{R}^n implies stable reachability in $W^{n,\infty}[-r, 0]$. This, conversely, does not imply regular reachability, but a weaker property. For linear systems of the form (see (3.4) below)

$$\dot{x}(t) = L(t)x_t + B(t)u(t)$$

this property specializes to the well-known rank condition for $B(t)$ on $[t_1 - r, t_1]$ plus essential boundedness of the generalized inverse $B(t)^+$ on $[t_1 - r, t_1]$. This proves, in

the state space $W^{n,\infty}[-r, 0]$, a conjecture by Banks/Jacobs/Langenhov [3]. In § 4, attention is focused on properties of the whole reachable set \mathcal{R} and of the set \mathcal{T} of trajectories. The subsets of regularly reachable states in \mathcal{R} and of regular trajectories in \mathcal{T} are described. It is shown that regularity is a generic property of trajectories. Examples are given where *all* trajectories reaching a certain final state are regular.

Notation and conventions. The Banach space of continuous functions on the compact set $A \subset \mathbb{R}^m$ with values in \mathbb{R}^n is denoted by $C^n(A)$. For $1 \leq p \leq \infty$, $W^{n,p}[a, b]$ denotes the Sobolev space of absolutely continuous functions $x: [a, b] \rightarrow \mathbb{R}^n$ with derivative $\dot{x} \in L_p^n[a, b]$, that is, with p -integrable, respectively essentially bounded derivative. The norm in the Banach space $W^{n,p}[a, b]$ is given by $\|x\| := (|x(a)|, \|\dot{x}\|_{L_p})$, where $|\cdot|$ denotes the Euclidean norm in finite dimensional space. $W^{n,p}[a, b]$ is identified in the canonical way with $\mathbb{R}^n \times L_p^n[a, b]$. The topological dual of a Banach space X is denoted by X^* . The interior of a set A in a Banach space is denoted by $\text{int } A$ and $\text{co } A$ is its convex hull. For $\delta > 0$, $\text{int}_\delta A$ is the set of all points in $\text{int } A$ having at least distance δ to the boundary ∂A of A .

$\mathcal{L}(X_1, X_2)$ is the space of bounded linear operators from a Banach space X_1 into a Banach space X_2 . For a measurable subset $S \subset \mathbb{R}$, we write $(s \in S)$ instead of for Lebesgues almost all $s \in S$. For the sets $\Omega(t)$ of admissible control values we assume that $\Omega(t) \subset \Omega_0 (t \in T)$, where $\Omega_0 \subset \mathbb{R}^m$ is compact, $t \mapsto \Omega(t)$ is measurable, and $\Omega(t)$ is closed for almost all $t \in T$. The set of Radon probability measures on Ω_0 is denoted by $\text{rpm}(\Omega_0)$. The set of relaxed controls v defined on the fixed time interval $T := [t_0, t_1]$ with values in the set of Radon probability measures on Ω_0 having support contained in $\Omega(t)$ is denoted by \mathcal{S}^* . Then \mathcal{S}^* can be identified with a subset of the dual space $\mathcal{N} := L_1(T, C(\Omega_0))^*$ and is considered in the weak* topology (for all this see Warga [27]).

Relaxed controls satisfy the following weak measurability requirement:

$$t \mapsto c(v(t)) := \int_{\Omega_0} c(\omega) v(t)(d\omega)$$

is measurable for each $c \in C(\Omega_0)$.

For f as in Lemma 1.1 below we define

$$f(x_t, v(t), t) := \int_{\Omega_0} f(x_t, \omega, t) v(t)(d\omega),$$

and identify an ordinary control u with the relaxed control $\delta_{u(\cdot)} \in \mathcal{S}^*$, where $\delta_{u(t)}$ denotes the point measure concentrated at $u(t) \in \Omega(t)$. T_1 denotes the final interval $[t_1 - r, t_1]$. Since the initial state φ_0 of the relaxed system Σ described by (0.1), (0.3) remains fixed, we tacitly assume that x satisfies $x_{t_0} = \varphi_0$, when we speak of “a trajectory x of Σ ”. A pair (x, v) is called a solution of Σ , if $v \in \mathcal{S}^*$ and x satisfies (0.1), (0.2).

1. Preliminaries. In this section we formulate conditions on f which will be needed in the sequel, and we investigate how the right-hand side of (0.1) depends on trajectories and controls. Then it is shown that the relaxation of the problem is equivalent to the convexification of the set of velocity vectors. Proofs are very condensed or omitted (see Colonius [8]).

LEMMA 1.1. Assume that the following conditions on f are satisfied:

(1.1) The function $f: C^n[-r, 0] \times \mathbb{R}^m \times T \rightarrow \mathbb{R}^n$ is continuous in $(\varphi, \omega) \in C^n[-r, 0] \times \mathbb{R}^m$ and measurable in $t \in T$;

(1.2) There is $p: \mathbb{R}_+ \times T \rightarrow \mathbb{R}_+$ such that for all $x \in C^n[t_0 - r, t_1]$ and $\omega \in \Omega_0$

$$|f(x_t, \omega, t)| \leq p(\|x_t\|_\infty, t) \quad (t \in T),$$

where $p(s, \cdot) \in L_\infty^1(T)$ for all $s \in \mathbb{R}_+$ and $p(\cdot, t)$ is monotonically increasing for almost all $t \in T$.

Let $((x^k, v^k)) \subset C^n[t_0 - r, t_1] \times \mathcal{S}^*$ be a sequence with $x^k \rightarrow x^0$ in $C^n[t_0 - r, t_1]$ and $v^k \rightarrow v^0$ weakly* in \mathcal{S}^* . Then $(f(x_t^k, v^k(t), t), t \in T) \rightarrow (f(x_t^0, v^0(t), t), t \in T)$ weakly* in $L_\infty^n(T)$, hence weakly in $L_2^n(T)$.

If $p(s, \cdot)$ in (1.2) is only known to be a L_2 -function, the convergence property still holds in the weak L_2 -topology.

Variants of this lemma implying weak L_2 -convergence are well known, even for functions f allowing lags in the controls (Berkovitz [5], Warga [26], Bates [4]). Weak* convergence in L_∞ follows similarly.

LEMMA 1.2. Assume that (1.1) and (1.2) hold and moreover the following condition is satisfied:

(1.3) The function f is continuously Fréchet differentiable in the first argument, the corresponding derivative $D_1 f(\varphi, \omega, t)$ is continuous in (φ, ω, t) and for all $\omega \in \Omega_0$

$$\|D_1 f(\varphi, \omega, t)\| \leq p(\|\varphi\|_\infty, t) \quad (t \in T)$$

where p is as in (1.2).

Let $((x^k, v^k)) \subset C^n[t_0 - r, t_1] \times \mathcal{S}^*$ be a sequence with $x^k \rightarrow x^0$ in $C^n[t_0 - r, t_1]$ and $v^k \rightarrow v^0$ in the strong norm topology on \mathcal{S}^* (see Warga [27, Thm. IV.1.9]). Then

$$\operatorname{ess\,sup}_t |f(x_t^k, v^k(t), t) - f(x_t^0, v^0(t), t)| \rightarrow 0$$

and the Fréchet derivatives $D_1 f(\varphi, v^k(t), t)$ exist and have the form $(\psi \in C^n[-r, 0])$

$$D_1 f(\varphi, v^k(t), t)\psi = \int_{\Omega_0} D_1 f(\varphi, \omega, t) v^k(t)(d\omega);$$

furthermore

$$\operatorname{ess\,sup}_t \|D_1 f(x_t^k, v^k(t), t) - D_1 f(x_t^0, v^0(t), t)\| \rightarrow 0.$$

Proof. Using (1.1)–(1.3) and the mean value theorem, we find

$$\begin{aligned} & \|f(x_t^k, v^k(t), t) - f(x_t^0, v^0(t), t)\|_\infty \\ & \leq \operatorname{ess\,sup}_{t, \omega} |f(x_t^k, \omega, t)| \|v^k - v^0\| + \operatorname{ess\,sup}_{t, \omega} |f(x_t^k, \omega, t) - f(x_t^0, \omega, t)| \\ & \leq \operatorname{ess\,sup}_t p(c_0, t) \|v^k - v^0\| + \operatorname{ess\,sup}_t p(c_0, t) \|x^k - x^0\|_\infty \end{aligned}$$

for a constant $c_0 > 0$. The right-hand side converges to 0 by assumption. The existence and the form of the derivatives $D_1 f(\varphi, v^k(t), t)$ follow by Warga [27, Thm. II.3.10]. For $\psi \in C^n[-r, 0]$ with $\|\psi\|_\infty \leq 1$

$$\begin{aligned} & \|D_1 f(x_t^k, v^k(t), t)\psi - D_1 f(x_t^0, v^0(t), t)\psi\| \\ & \leq \sup_{t, \omega} \|D_1 f(x_t^k, \omega, t) - D_1 f(x_t^0, \omega, t)\| \|v^k\| + \sup_{t, \omega} \|D_1 f(x_t^0, \omega, t)\| \|v^k - v^0\|. \end{aligned}$$

This converges to 0, since $\Omega \times T$ is compact and $D_1 f$ is continuous. \square

LEMMA 1.3. Let the conditions (1.1) and (1.2) be satisfied and assume that the trajectories of Σ are uniformly bounded. Then the set \mathcal{T} of trajectories x of Σ is compact and sequentially compact in the weak* topology of $W^{n, \infty}(T)$ and the uniform norm topology.

Proof. Sequential weak* compactness of \mathcal{T} follows by sequential weak* compactness of \mathcal{S}^* , Warga [27, Thm. IV.3.11, and Lemma 1.1]. Since $W^{n, \infty}$ is the dual of the separable space $W^{n, 1}$, weak* compactness of \mathcal{T} is equivalent to weak* sequential

compactness (see Dunford/Schwartz [13, p. 437]. Finally, compactness in the topology of uniform convergence follows, because the embedding of $W^{n,\infty}(T)$ into $C^n(T)$ is compact. \square

Remark 1.1. The trajectories of Σ are uniformly bounded, if (1.3) is satisfied. The following lemma allows us to characterize relaxed velocity vectors.

LEMMA 1.4. *For a measurable subset $S \subset T$, consider a measurable function $z : S \rightarrow \mathbb{R}^n$ and a function $\Phi : \Omega_0 \times S \rightarrow \mathbb{R}^n$ with $\Phi(\omega, \cdot)$ measurable for all $\omega \in \Omega_0$ and $\Phi(\cdot, t)$ continuous for a.a. $t \in S$. Then the following three conditions are equivalent:*

- (i) $z(t) \in \text{co } \Phi(\Omega(t), t) \quad (t \in S);$
- (ii) $z(t) = \Phi(v(t), t) \quad (t \in S)$

for an element $v \in \mathcal{S}^*$;

$$(iii) \quad z(t) = \sum_{i=0}^n \alpha_i(t) \Phi(u_i(t), t) \quad (t \in S)$$

for some measurable $\alpha_i : T \rightarrow \mathbb{R}_+$ and $u_i : T \rightarrow \Omega_0$ with $\sum_{i=0}^n \alpha_i(t) = 1$ and $u_i(t) \in \Omega(t)$.

Proof. The proof follows by Warga [27, Thms. I.6.13, IV 3.13; compare also Thm. VI.3.2]. \square

Remark 1.2. Consider

$$\{(f(x, v(t), t), t \in T) : v \in \mathcal{S}^*\}.$$

Then the lemma above implies that this set coincides with

$$\{z \in L_\infty^n(T) : z(t) \in \text{co } f(x, \Omega(t), t) (t \in T)\}.$$

This shows that along a fixed trajectory x the set of relaxed velocity vectors coincides with the convex hull of the set of ordinary velocity vectors. Hence the relaxed system Σ is equivalent to the relaxed system considered by Oguztöreli [22, § 8.9].

LEMMA 1.5. *Assume that (1.1)–(1.3) hold and, additionally, that the following condition is satisfied:*

(1.4) *For each relaxed control $v \in \mathcal{S}^*$ there is a unique trajectory $x = S(v) \in C^n[t_0 - r, t_1]$ satisfying (0.1) and (0.2).*

Then $S(v)$ depends in a continuously Fréchet differentiable manner on $v \in \mathcal{S}^$; for $v, v^0 \in \mathcal{S}^*$ and $x^0 := S(v^0)$, the trajectory $x(v) := DS(v^0)(v - v^0) \in C^n[t_0 - r, t_1]$ satisfies*

$$(1.5) \quad \dot{x}(t) = D_1 f(x_t^0, v^0(t), t) x_t + f(x_t^0, v(t) - v^0(t), t) \quad (t \in T),$$

$$(1.6) \quad x_{t_0} = 0.$$

Proof. The proof follows by the implicit function theorem and computation of derivatives according to Lemma 1.2. \square

Remark 1.3. It is convenient to write the linearized system in the form (1.5), because (1.5) involves the relaxed control $v \in \mathcal{S}^*$; $v - v^0$ is *not* in the set \mathcal{S}^* of relaxed controls.

Remark 1.4. Assumption (1.4) requires existence and uniqueness for trajectories corresponding to *relaxed* controls. This problem can be reduced to existence and uniqueness theory of functional differential equations using the representation of relaxed velocities introduced by Gamkrelidze: by Lemma 1.4 for each relaxed trajectory x there exist measurable functions $\alpha_0, \alpha_1, \dots, \alpha_n : T \rightarrow \mathbb{R}_+$ with $\sum \alpha_i(t) = 1$ and ordinary controls u_0, u_1, \dots, u_n with values in $\Omega(t)$ such that

$$\dot{x}(t) = \sum_{i=0}^n \alpha_i(t) f(x_t, u_i(t), t) \quad (t \in T),$$

and conversely. Hence relaxed trajectories satisfy a functional differential equation; for results on the solution of these equations see Hale [14].

Remark 1.5. The set of trajectories of (1.5) coincides with the set of trajectories of the system

$$(1.6) \quad \dot{x}(t) = D_1 f(x_t^0, v^0(t), t)x_t + u(t) \quad (t \in T)$$

where $u : T \rightarrow \mathbb{R}^n$ are measurable functions with values $u(t)$ in

$$\text{co } f(x_t^0, \Omega(t), t) - f(x_t^0, v^0(t), t) = \text{co } f(x_t^0, \Omega(t), t) - \dot{x}^0(t).$$

This is a consequence of Lemma 1.4. Observe that (1.5) (and (1.6)) again is uniquely solvable by Hale [14, Chap. 6, Thm. 2.1].

2. Local and stable reachability. In this section we formally state the relation between the ordinary and the relaxed problem indicated in the introduction. Then we prove sufficiency of a criterion for stable reachability and show that local and stable reachability are equivalent for linear relaxed systems.

In the rest of this paper we assume that the conditions (1.1), (1.2) and (1.4) are satisfied.

THEOREM 2.1. *If the trajectories of Σ are uniformly bounded, the following assertions hold:*

(i) *Suppose (x^0, v^0) is a solution of Σ . Then there is a sequence $((x^k, v^k))$ of solutions of Σ , where v^k is an ordinary control, such that $v^k \rightarrow v^0$ weakly* in \mathcal{S}^* and $x^k \rightarrow x^0$ in $C^n[t_0 - r, t_1]$.*

(ii) *If, conversely, $((x^k, v^k))$ is a sequence of solutions of Σ (where v^k is not necessarily an ordinary control) with $x_{t_1}^k \rightarrow \varphi_1$ in $C^n[-r, 0]$, then there exists a solution (x^0, v^0) of Σ such that for a subsequence $v^k \rightarrow v^0$ weakly* in \mathcal{S}^* and $x^k \rightarrow x^0$ in $C^n[t_0 - r, t_1]$; in particular $x_{t_1}^0 = \varphi_1$.*

Proof. (i) Since the ordinary controls are dense in \mathcal{S}^* (Warga [27, Thm. IV.3.10]) there is a sequence (v^k) of ordinary controls converging to v^0 . Then a subsequence of the corresponding sequence of trajectories $x^k := S(v^k)$ converges by Lemma 1.3, and by Lemma 1.2 and assumption (1.4) its limit is x^0 .

(ii) follows similarly. \square

Remark 2.1. The assertions of the theorem remain true if uniform convergence of the trajectories (resp. final states) is replaced by weak* convergence in $W^{n,\infty}$.

The following definition presents an *abus de langage*; however, it is quite convenient.

DEFINITION 2.1. A function $\varphi_1 \in W^{n,\infty}[-r, 0]$ is called *locally reachable* at time t_1 , if and only if there is a neighborhood N of φ_1 in $W^{n,\infty}[-r, 0]$ such that for each $\varphi \in N$ there is a trajectory x of Σ with $\varphi = x_{t_1}$. Analogously, we define local reachability of a point α in \mathbb{R}^n .

DEFINITION 2.2. A function $\varphi_1 \in W^{n,\infty}[-r, 0]$ is called *stably reachable* at time t_1 with a trajectory x^0 of Σ , if and only if for each neighborhood V of x^0 in $C^n[t_0 - r, t_1]$ there is a neighborhood N of φ_1 in $W^{n,\infty}[-r, 0]$ such that for each $\varphi \in N$ there is a trajectory $x \in V$ of Σ with $\varphi = x_{t_1}$. Analogously, we define stable reachability of a point α in \mathbb{R}^n .

Stable reachability of φ (resp., α) with x^0 means that one can reach a complete neighborhood of φ_1 (resp., α) with arbitrarily small deviations from x^0 . Obviously this implies local reachability. However, the following example which is taken from Lee/Markus [18, p. 257] illustrates that in general the converse does not hold. It is a nonlinear system with $r = 0$, i.e., an ordinary differential system.

Example 2.1. Consider

$$\begin{aligned}
 \dot{x}^1(t) &= x^2(t)u^1(t) - x^1(t)u^2(t), \\
 \dot{x}^2(t) &= -x^1(t)u^1(t) - x^2(t)u^2(t), \quad t \in [0, \pi], \\
 x^1(0) &= 1, \quad x^2(0) = 0, \\
 (u^1(t), u^2(t)) &\in \Omega := \{(\omega_1, \omega_2) \in \mathbb{R}^2 : |\omega_1| \leq 1, |\omega_2| \leq 1\}.
 \end{aligned}
 \tag{2.1}$$

Since for fixed $x = (x^1, x^2)$ the velocity sets are convex, it suffices to consider ordinary controls (see Remark 1.2). In polar coordinates (r, φ) , the system is described by

$$\begin{aligned}
 \dot{r}(t) &= -r(t)u^2(t), \quad \dot{\varphi}(t) = -u^1(t), \\
 r(0) &= 1, \quad \varphi(0) = 0.
 \end{aligned}$$

The reachable set at time $t_1 = \pi$ is an annulus around 0 with inner radius $e^{-\pi}$ and outer radius e^{π} . It is not difficult to see that all pairs (r, π) with $e^{-\pi} < r < e^{\pi}$ are locally reachable, but not stably reachable at time $t_1 = \pi$.

Remark 2.2. It seems that this is a general phenomenon: We conjecture that any point being for the first time in the interior of the reachable set without having—before that time—been at the boundary of the reachable set is not stably reachable.

Condition (2.2), below, is related to regular reachability (see § 3) and will play an important role in the rest of this paper. The following *stability lemma* shows that it remains valid under small perturbations of the trajectory x^0 .

LEMMA 2.1. Assume that (1.3) and the following condition are satisfied:

(2.2) For a measurable subset $S \subset T$, for $x^0 \in C^n[t_0 - r, t_1]$, and $z \in L^n_\infty(S)$, there is a neighborhood V of $0 \in \mathbb{R}^n$ such that

$$V \subset z(t) + \text{co } f(x_t^0, \Omega(t), t) \quad (t \in S).$$

Then there are $\delta > 0$ and a neighborhood \tilde{V} of $0 \in \mathbb{R}^n$ such that for all x with $\|x - x^0\|_\infty < \delta$

$$\tilde{V} \subset z(t) + \text{co } f(x_t, \Omega(t), t) \quad (t \in S).$$

Proof. Take V as an n -simplex with vertices e_0, e_1, \dots, e_n . Then by Lemma 1.4 there are $v_i \in \mathcal{S}^{\#}$ with

$$e_i = z(t) + f(x_t^0, v_i(t), t) \quad (t \in S).$$

By Lemma 1.2, $f(x_t, v_i(t), t)$ is essentially uniformly close to $f(x_t^0, v_i(t), t)$ for small $\|x - x^0\|_\infty$. Thus there is $\delta > 0$ such that for all x with $\|x - x^0\|_\infty < \delta$ and for a.a. $t \in S$ the points $f(x_t, v_i(t), t)$, $i = 0, 1, \dots, n$, are also vertices of n -simplices containing a fixed neighborhood \tilde{V} of $0 \in \mathbb{R}^n$. This proves the lemma, since convex combinations of $f(x_t, v_i(t), t)$, $i = 0, 1, \dots, n$ remain in $\text{co } f(x_t, \Omega(t), t)$. \square

Remark 2.3. Assumption (1.3) may be replaced by continuity of $f(\varphi, \omega, t)$ in (φ, ω, t) (compare the second part of Lemma 1.2).

THEOREM 2.2. Suppose the trajectory x^0 of Σ reaches $\varphi_1 \in W^{n,\infty}[-r, 0]$ at time t_1 , and there are $\varepsilon > 0$ and a neighborhood V of $0 \in \mathbb{R}^n$ such that

$$(2.3) \quad V \subset -x^0(t) + \text{co } f(x_t^0, \Omega(t), t) \quad (t \in [t_1 - r - \varepsilon, t_1 - r]).$$

Then $\varphi_1(-r)$ is stably reachable with x^0 at time $t_1 - r$.

If the inclusion in (2.3) holds also for a.a. $t \in T_1$, the function φ_1 is stably reachable with x^0 at time t_1 .

Proof. By the stability lemma, Lemma 2.1, there are $\delta_0 > 0$ and a neighborhood U of $\dot{x}^0|_{[t_1-r-\varepsilon, t_1-r]} \in L_\infty^n$ such that for all x with $\|x - x^0\|_\infty < \delta_0$ and all $z \in U$

$$(2.4) \quad z(t) \in \text{co } f(x_t, \Omega(t), t) \quad (t \in [t_1-r-\varepsilon, t_1-r]).$$

For all $\delta > 0$, the set N_δ defined by

$$\begin{aligned} N_\delta := \{ \alpha \in \mathbb{R}^n : \alpha = x(t_1-r) \text{ for some } x \in C^n[t_0-r, t_1] \\ \text{with } x(t) = x^0(t) \text{ for } t \in [t_0-r, t_1-r-\varepsilon], \\ \|x - x^0\|_\infty < \delta \text{ and } \dot{x}|_{[t_1-r-\varepsilon, t_1-r]} \in U \} \end{aligned}$$

forms a neighborhood of $\varphi_1(-r)$ in \mathbb{R}^n . For $0 < \delta < \delta_0$ take $\alpha \in N_\delta$. Then there is $x \in C^n[t_0-r, t_1]$ as in the definition of N_δ with

$$(2.5) \quad \dot{x}(t) \in \text{co } f(x_t, \Omega(t), t) \quad (t \in [t_1-r-\varepsilon, t_1-r]).$$

By Lemma 1.4, x is a trajectory of Σ , reaching α by definition. Thus x^0 reaches $\varphi_1(-r)$ stably at time t_1-r .

Stable reachability of φ_1 follows similarly, since for each $\delta > 0$ and each neighborhood U of $\dot{x}^0|_{[t_1-r-\varepsilon, t_1]}$ in L_∞^n the set

$$\begin{aligned} \{ \varphi \in W^{n,\infty}[-r, 0] : \varphi = x_{t_1} \text{ for some } x \in C^n[t_0-r, t_1] \\ \text{with } x(t) = x^0(t) \text{ for } t \in [t_0-r, t_1-r-\varepsilon], \\ \|x - x^0\|_\infty < \delta, \text{ and } \dot{x}|_{[t_1-r-\varepsilon, t_1]} \in U \} \end{aligned}$$

forms a neighborhood of φ_1 in $W^{n,\infty}[-r, 0]$. \square

Next we consider a class of systems Λ which are linear in x . As a *relaxed* system, Λ is linear.

$$\begin{aligned} \Lambda \quad \dot{x}(t) &= L(t)x_t + b(v(t), t) \quad (t \in T), \\ x_{t_0} &= \varphi_0, \end{aligned}$$

where $v \in \mathcal{S}^*$ and

(2.6) for $L : T \rightarrow \mathcal{L}(C^n[-r, 0], \mathbb{R}^n)$ the map $t \mapsto L(t)\varphi$ is measurable for all $\varphi \in C^n[-r, 0]$ and $\text{ess sup } \|L(t)\| < \infty$; the function $b : T \times \Omega_0 \rightarrow \mathbb{R}^n$ is measurable in the first, continuous in the second argument, and satisfies

$$\text{ess sup}_{t \in T} \sup_{\omega \in \Omega_0} |b(t, \omega)| < \infty;$$

furthermore, $\varphi_0 \in C^n[-r, 0]$ and $t_1-r > t_0$.

For each $v \in \mathcal{S}^*$, the system Λ has a unique trajectory x with $x_t = \varphi_0$, Hale [14, p. 142]. The system Λ is a special case of Σ satisfying conditions (1.1)–(1.4).

Remark 2.4. Consider a function $b : T \times \Omega_0 \rightarrow \mathbb{R}^n$ with the same properties as in (2.6). Then the map $t \mapsto b(t, \cdot)$ may be considered as a map on T with values in $C^n(\Omega_0)$. Hence, since $C^n(\Omega_0)$ is separable, b may be identified with an element of $L_\infty(T, C^n(\Omega_0))$ where $\|b\|_\infty := \text{ess sup}_{t \in T} \sup_{\omega \in \Omega_0} |b(t, \omega)|$ (cf. Warga [27, p. 122 and Theorem I.5.26]).

Remark 2.5. The set of trajectories of Λ coincides with the set of trajectories of the following system with ordinary controls:

$$\dot{x}(t) = L(t)x_t + u(t) \quad (t \in T),$$

where the controls u are measurable functions defined on T with values $u(t)$ in $\text{co } b(\Omega(t), t)$. Hence the reachability theories for these two classes of systems coincide. However, we write down the proofs for the system Λ , since the corresponding optimal control theories are different and the extra expense of writing is minimal.

THEOREM 2.3. *For the system Λ a function φ_1 in $W^{n,\infty}[-r, 0]$, (respectively a point α in \mathbb{R}^n), is stably reachable iff it is locally reachable.*

Proof. One direction is trivial. Let N be a neighborhood of φ_1 in $W^{n,\infty}[-r, 0]$ such that each element of N is reachable at time t_1 . Suppose that φ_1 is reached with the trajectory x^0 of Λ corresponding to $v^0 \in \mathcal{S}^*$. For $\varepsilon > 0$, define a neighborhood N^ε of φ_1 by

$$N^\varepsilon := \varphi_1 + \varepsilon(N - \varphi_1).$$

Let $\varphi^\varepsilon \in N^\varepsilon$. Then there is $\varphi \in N$ with $\varphi^\varepsilon = \varphi_1 + \varepsilon(\varphi - \varphi_1)$. The function φ is reachable, say with x corresponding to $v \in \mathcal{S}^*$. Then by linearity, φ^ε is reached with $x^\varepsilon := x^0 + \varepsilon(x - x^0)$ corresponding to

$$v^\varepsilon := v^0 + \varepsilon(v - v^0) \in \mathcal{S}^*.$$

For all $\delta > 0$ there is $\varepsilon > 0$ such that for all $\varphi^\varepsilon \in N^\varepsilon$ the trajectory x^ε satisfies $\|x^\varepsilon - x^0\|_\infty < \delta$, since the trajectories of Λ are uniformly bounded. Thus φ_1 is reached stably with x^0 . The assertion for $\alpha \in \mathbb{R}^n$ follows similarly. \square

Remark 2.6. Suppose that φ_1 is reached stably with x^0 and x^1 is any trajectory of Λ reaching φ_1 . Then the theorem above shows that φ_1 is reached stably with x^1 . Hence for linear systems it is not necessary to specify the trajectory reaching φ_1 stably, if φ_1 is stably reachable with some trajectory.

Remark 2.7. For scalar systems (i.e. $n = 1$) stable reachability of points in \mathbb{R} follows under very mild assumptions. Suppose the trajectories of Σ are uniformly bounded and there are $v^0, \bar{v}, \underline{v} \in \mathcal{S}^*$ such that the map $v \mapsto x(v, t_1)$ has Gateaux derivatives at v^0 in the directions $\bar{v} - v^0$ and $\underline{v} - v^0$ satisfying

$$D(x(t_1, v^0); \bar{v} - v^0) > 0 \text{ and } D(x(t_1, v^0); \underline{v} - v^0) < 0.$$

Then $x(t_1, v^0)$ is stably reachable with $x(v^0)$, since the map $v \mapsto x(t_1, v)$ is continuous and $n = 1$.

3. Regular and stable reachability. In this section, we introduce the concept of regular reachability and investigate its relation to local and stable reachability.

DEFINITION 3.1. A function φ_1 in $W^{n,\infty}[-r, 0]$ is called *regularly reachable at time t_1* with a trajectory x^0 of Σ , iff $\varphi_1(-r) = x^0(t_1 - r)$ and there exists a neighborhood V of $0 \in \mathbb{R}^n$ such that

$$(3.1) \quad V \subset -\dot{\varphi}_1(t - t_1) + \text{co } f(x_t^0, \Omega(t), t) \quad (t \in T_1).$$

A trajectory x^0 is called *regular*, iff $x_{t_1}^0$ is reached regularly at time t_1 with x^0 ; otherwise x^0 is called *irregular*. By Lemma 1.4 it follows—which is to be expected—that $\varphi_1 = x_{t_1}^0$, if x^0 reaches φ_1 regularly. A function $\varphi_1 \in W^{n,\infty}[-r, 0]$ is reached regularly with x^0 iff $\varphi_1(-r) = x^0(t_1 - r)$ and in L_∞^n

$$\begin{aligned} \dot{\varphi}_1 &\in \text{int}\{z \in L_\infty^n[-r, 0]: z(t - t_1) \in \text{co } f(x_t^0, \Omega(t), t) (t \in T_1)\} \\ &= \text{int}\{z \in L_\infty^n[-r, 0]: z(t - t_1) = f(x_t^0, v(t), t) (t \in T_1) \text{ for a } v \in \mathcal{S}^*\}. \end{aligned}$$

Regular reachability means that not only $\dot{\varphi}_1(t - t_1)$, but a whole uniform neighborhood of $\dot{\varphi}_1(t - t_1)$ is contained in the set of relaxed velocity vectors if the system at time t is in the state x_t^0 .

If the optimal trajectory x^0 is regular, the maximum principle constitutes a necessary optimality criterion (see Colonius [9], [10]).

The following theorem relates stable reachability of points in \mathbb{R}^n and final states in $W^{n,\infty}[-r, 0]$ via the notion of regular reachability.

THEOREM 3.1. *Assume that Σ satisfies (1.3) and that for $\varphi_1 \in W^{n,\infty}[-r, 0]$ the following assumptions hold:*

(a) *the point $\varphi_1(-r)$ is reached stably with x^0 at $t_1 - r$;*

(b) *the function φ_1 is reached regularly with x^0 at t_1 .*

Then φ_1 is reached stably with x^0 at t_1 .

Proof. By assumption (b) and the stability lemma 2.1, there are $\delta > 0$ and a neighborhood N of $\varphi_1 \in W^{n,\infty}[-r, 0]$ such that for all $\varphi \in N$ and all x with $\|x - x^0\|_\infty < \delta$

$$\dot{\varphi}(t - t_1) \in \text{co } f(x, \Omega(t), t) \quad (t \in T_1).$$

By assumption (a), there is a neighborhood N_1 of $\varphi_1 \in W^{n,\infty}[-r, 0]$, such that for all $\varphi \in N_1$ there is $x^\varphi \in C^n[t_0 - r, t_1 - r]$ with $\|x^\varphi - x^0\|_\infty < \delta$ and

$$\dot{x}^\varphi(t) = f(x_t^\varphi, v^\varphi(t), t) \quad (t \in [t_0, t_1 - r])$$

for a relaxed control $v^\varphi \in \mathcal{S}^*$, and

$$x_{t_0}^\varphi = \varphi_0, \quad x^\varphi(t_1 - r) = \varphi(-r).$$

Without loss of generality, we may assume that $N_1 \subset N$. The function x^φ can be extended to an absolutely continuous function on $[t_0 - r, t_1]$ by

$$x_{t_1}^\varphi := \varphi.$$

It only remains to prove that x^φ is a trajectory of Σ . We have to show that there is $v \in \mathcal{S}^*$ such that

$$\dot{x}^\varphi(t) = \dot{\varphi}(t - t_1) = f(x_t^\varphi, v(t), t) \quad (t \in T_1)$$

or, equivalently,

$$\dot{\varphi}(t - t_1) \in \text{co } f(x_t^\varphi, \Omega(t), t) \quad (t \in T_1).$$

This follows by $N_1 \subset N$ and $\|x^\varphi - x^0\|_\infty < \delta$. \square

COROLLARY 3.1. *Assume that $\varphi_1 \in W^{n,\infty}[-r, 0]$ is regularly reachable with x^0 . Then for all y in a neighborhood of $\dot{\varphi}_1 \in L^n[-r, 0]$ there is a trajectory x of Σ with*

$$\dot{x}_{t_1} = y \text{ and } x(t) = x^0(t) \text{ for } t \in [t_0, t_1 - r].$$

Proof. First reach $\varphi_1(-r)$ with x^0 . Then use regularity as above in order to steer the system to

$$(\varphi_1(-r) + \int_{-r}^s y(\tau) d\tau, s \in [-r, 0]) \in W^{n,\infty}[-r, 0].$$

\square

Theorem 3.1 shows that, up to a finite dimensional reachability condition at $t_1 - r$, regular reachability implies stable reachability. We shall prove a partial converse of this for a special class of systems.

Define the class of systems Σ_a , where the control term appears additively, by:

$$\dot{x}(t) = a(x_t, t) + b(v(t), t) \quad (t \in T),$$

$$x_{t_0} = \varphi_0,$$

where $v \in \mathcal{S}^*$ and, in addition to the requirements of (2.6), (3.2) the map $a : C^n[-r, 0] \times T \rightarrow \mathbb{R}^n$ is continuous in the first and measurable in the second argument, and $|a(\varphi, t)|$ is essentially bounded for bounded arguments.

Σ_a is a special case of Σ satisfying (1.1) and (1.2). Furthermore, we assume unique solvability, i.e., (1.4).

Remark 3.1. As in Remark 2.5, the set of trajectories of Σ_a coincides with the set of trajectories of

$$\dot{x}(t) = a(x, t) + u(t),$$

where the controls u take values $u(t)$ in $\text{co } b(\Omega(t), t)$.

We shall make use of the following generalization of a classical result due to A. Denjoy.

THEOREM 3.2. *Let X be a complete separable metric space, S a compact metric space, $h: [a, b] \times S \rightarrow X$ such that $h(\cdot, s)$ is measurable and $h(t, \cdot)$ is continuous for all $(t, s) \in [a, b] \times S$. Then for a subset T' of $[a, b]$ with $\lambda(T') = \lambda([a, b])$ and for all $t \in T'$ there is a set $E \subset [a, b]$ such that t is a point of density of E and $h|_E \times S$ is continuous in t .*

Proof. By a strong version of Lusin's theorem (see Warga [27, Thm. I.5.26(2)]), there is for all $\varepsilon > 0$ a closed subset F_ε of $[a, b]$ such that $\lambda([a, b] \setminus F_\varepsilon) \leq \varepsilon$ and $h|_{F_\varepsilon} \times S$ is continuous. This implies the theorem above in exactly the same way as the usual Lusin's theorem implies Natanson [21, Satz 2, p. 296]. \square

THEOREM 3.3. *Assume that the trajectories of Σ_a are uniformly bounded, and that there is a neighborhood U of $z^0 \in L_\infty^n(T_1)$ such that for all $z \in U$ there is a trajectory x with $z = (\dot{x})_{t_1}$. Then*

$$\text{int}\{z \in L_\infty^n(T_1): z(t) \in \text{co } b(\Omega(t), t) (t \in T_1)\} \neq \emptyset.$$

Proof. There is a trajectory x^0 with $z^0 = (\dot{x}^0)_{t_1}$. Define for $t \in T_1$

$$V_t := -z^0(t) + a(x_t^0, t) + \text{co } b(\Omega(t), t).$$

Then V_t is convex and $0 \in V_t$. Due to the assumptions on Σ_a , all sets V_t are contained in a compact set $C \subset \mathbb{R}^n$. Consider $\Phi: t \mapsto V_t$ as a map from T_1 into the set of closed subsets of C . Then Φ is well defined and measurable by Warga [27, Thm. I.7.6]. Thus by Warga [27, Thm. I.7.8], there is a countable set $\{\xi_i: i \in \mathbb{N}\}$ of measurable selections of Φ such that

$$\{\xi_i(t): i \in \mathbb{N}\} \text{ is dense in } \Phi(t) = V_t \text{ for a.a. } t \in T_1.$$

For $y \in \mathcal{E} := \{y \in \mathbb{R}^n: |y| = 1\}$ and $t \in T_1$ let

$$\bar{h}(y, t) := \sup\{y\xi_i(t): i \in \mathbb{N}\}, \quad \underline{h}(y, t) := \inf\{y\xi_i(t): i \in \mathbb{N}\}.$$

Then \bar{h} and \underline{h} are continuous in y , measurable in t and satisfy

$$\bar{h}(y, t) = \max\{y\xi: \xi \in V_t\}, \quad \underline{h}(y, t) = \min\{y\xi: \xi \in V_t\}.$$

We shall prove that there are a constant $\alpha > 0$ and a set $T' \subset T_1$ with $\lambda(T') = \lambda(T_1)$ such that for all $t \in T'$ and all $y \in \mathcal{E}$

$$(3.3) \quad \bar{h}(y, t) \geq \underline{h}(y, t) + \alpha.$$

Then, by Warga [27, Thm. I.7.10], there exists a measurable function $\xi: T_1 \rightarrow C$ with $\xi(t) \in V_t$, $t \in T'$ such that for all $y \in \mathcal{E}$ and all $t \in T'$

$$\min\{y\xi: \xi \in V_t\} + \frac{\alpha}{2} \leq y\xi(t) \leq \max\{y\xi: \xi \in V_t\} - \frac{\alpha}{2}.$$

Then $\xi \in \text{int}\{z \in L_\infty^n(T_1): z(t) \in V_t (t \in T_1)\}$, and the theorem follows.

The set \mathcal{E} and the set \mathcal{T} of trajectories of Σ_a are compact metric spaces, since \mathcal{T} is compact in $C^n[t_0-r, t_1]$ by Lemma 1.3. Consider a as a map: $\mathcal{T} \times T_1 \rightarrow \mathbb{R}^n$, $(x, t) \mapsto a(x, t)$. Then a is continuous in x and measurable in t .

Now Theorem 3.2 implies that there is $T' \subset T_1$ with $\lambda(T') = \lambda(T)$ such that for all $t' \in T'$, all $y \in \mathcal{E}$, and all $x \in \mathcal{T}$ $\bar{h}(y, t)$, $\underline{h}(y, t)$, and $a(x, t)$ are approximately continuous in $t = t'$. There exists a constant $\alpha > 0$ with the following property: For any $t' \in T'$ and any $y \in \mathcal{E}$, there is a step function z such that $z^0 + z \in U$ and that $yz(t)$ has a jump of length greater than 2α in $t = t'$.

For any such z there exists a solution (x, v) of Σ_a with

$$z(t) = -z^0(t) + a(x, t) + b(v(t), t) \quad (t \in T_1).$$

Hence

$$yz(t) = y[a(x, t) - a(x_t^0, t)] + y[-z^0(t) + a(x_t^0, t) + b(v(t), t)] \quad (t \in T_1).$$

The left-hand side has a jump of length greater than 2α in $t = t'$ and $y[a(x, t) - a(x_t^0, t)]$ is approximately continuous in $t = t'$. Hence (3.3) follows. \square

Remark 3.2. The proof above owes a lot to the proof of Schwarzkopf [24, Thm. 1].

Remark 3.3. Let $\varphi_1 \in W^{n,\infty}[-r, 0]$. Then the assumptions of Theorem 3.3 are satisfied for $z^0(t) = \dot{\varphi}_1(t - t_1)$, $t \in T_1$, if φ_1 is reached locally (resp. stably) with x^0 at t_1 , and the trajectories of Σ_a are uniformly bounded. Then it also follows that $\varphi_1(-r)$ is reached locally (resp. stably) with x^0 at $t_1 - r$. If one could prove that V_t contains a neighborhood of $0 \in \mathbb{R}^n$ which is independent of $t \in T_1$, the converse of Theorem 3.1 would hold. However, in the following we give an example of a linear system where a certain final state φ_1 is locally reachable (i.e., by Theorem 2.3 also stably reachable), but reached irregularly with a certain trajectory x^0 . Hence in general, we cannot show more than Theorem 3.3. Intuitively, this may be explained as follows:

In the proof of Theorem 3.1, we reach φ_1 with x^0 . In order to reach all φ in a neighborhood of $\varphi_1 \in W^{n,\infty}[-r, 0]$, we reach first $\varphi(-r)$; by stable reachability of $\varphi_1(-r)$, we can achieve this with small deviations from $x^0|_{[t_0, t_1-r]}$. The hereditary effects influencing the possible velocity vectors of the system on T_1 are compensated using regular reachability of φ_1 : By the stability lemma 2.1, regularity allows to choose controls in such a way that *arbitrary* small deviations from x^0 are compensated and all velocity vectors $\dot{\varphi}(t - t_1)$ in a uniform neighborhood of $\dot{\varphi}_1(t - t_1)$ are reached.

Now one can try to reach $\varphi(-r)$ in such a way that the hereditary effects influencing the system behavior on T_1 prepare the reaching of φ . Then it is not necessary to compensate *arbitrary* small deviations from x^0 on $[t_0, t_1 - r]$ and we may get local reachability without regularity. This is performed in the following scalar example. Theorem 3.3 shows the limitations of this manipulation: One can achieve a shift of the set of velocity vectors, but V_t must contain interior points.

Example 3.1. Consider

$$\dot{x}(t) = x(t-2) + u(t) \quad (t \in T := [0, 6]),$$

$$x_0 = 0,$$

$$\Omega := [-3, 1].$$

We claim that

$$\varphi_1(t) := \begin{cases} t+2, & t \in [-2, -1], \\ 1, & t \in [-1, 2] \end{cases}$$

is locally and irregularly reachable. Define an (ordinary) control function u^1 by

$$u^1(t) := \begin{cases} 0 & \text{for } t \in [0, 4], \\ 1 & \text{for } t \in (4, 5], \\ 0 & \text{for } t \in (5, 6]. \end{cases}$$

The corresponding trajectory x^1 is given by

$$x^1(t) := \begin{cases} 0 & \text{for } t \in [-2, 4], \\ t-4 & \text{for } t \in (4, 5], \\ 1 & \text{for } t \in [5, 6]. \end{cases}$$

Clearly x^1 is irregular and satisfies $x_6^1 = \varphi_1$. Now we shall show that the assumptions of Theorem 2.2 are satisfied for a certain trajectory x^0 . This will imply that φ_1 is locally reachable.

Define a control function u^0 by

$$u^0(t) = \begin{cases} 1 & \text{for } t \in [0, 2), \\ -2 & \text{for } t \in [2, 4), \\ -\frac{1}{2}(t-4)^2 + 2(t-2) - 5 & \text{for } t \in [4, 5), \\ -\frac{1}{2}(t-4)^2 + 2(t-2) - 6 & \text{for } t \in [5, 6]. \end{cases}$$

The corresponding trajectory x^0 is given by

$$x^0(t) = \begin{cases} 0 & \text{for } t \in [-2, 0], \\ t & \text{for } t \in [0, 2], \\ \frac{1}{2}(t-2)^2 - 2t + 6 & \text{for } t \in [2, 4], \\ t-4 & \text{for } t \in [4, 5], \\ 1 & \text{for } t \in [5, 6]. \end{cases}$$

Clearly $x_6^0 = \varphi_1$.

An elementary analysis shows that $u^0(t)$ is uniformly in the interior of Ω on $[2, 6]$; i.e., the assumptions of Theorem 2.2 are satisfied. \square

Assume that (1.3) holds. Then the system Σ can be linearized, and the linearized system Σ_{lin} has the form (1.5), (1.6) as indicated in Lemma 1.5.

Regular reachability has the remarkable property, that it is invariant under linearization. More precisely, we have

PROPOSITION 3.1. *Suppose that (1.3) holds, and $\varphi_1 = x_{t_1}^0$, where x^0 is the trajectory of Σ corresponding to $v^0 \in \mathcal{S}^*$. Then x^0 reaches φ_1 regularly iff the zero trajectory of Σ_{lin} reaches $0 \in W^{n,\infty}[-r, 0]$ regularly.*

Proof. Write down the definitions! \square

This property motivates interest in regular reachability of $0 \in W^{n,\infty}[-r, 0]$ with the zero trajectory. Consider the following linear system:

$$(3.4) \quad \dot{x}(t) = L(t)x_t + B(t)v(t) \quad (t \in T),$$

$$(3.5) \quad x_{t_0} = 0,$$

where we assume that the assumptions in (2.6) are satisfied and $B \in L_\infty(T, \mathbb{R}^{n \times m})$.

COROLLARY 3.2. *Let $0 \in \text{int}_\delta \text{co } \Omega(t)(t \in T_1)$ for a $\delta > 0$. Then the following two conditions on the system (3.4), (3.5) are equivalent:*

- (i) *The trajectory $x^0 = 0$ reaches $0 \in W^{n,\infty}[-r, 0]$ regularly at time t_1 ;*
- (ii) *Rank $B(t) = n$ ($t \in T_1$) and $B(t)^+$ is essentially bounded on T_1 .*

Furthermore, $0 \in W^{n,\infty}[-r, 0]$ is locally reachable at time t_1 , iff (ii) holds and $0 \in \mathbb{R}^n$ is locally reachable at time $t_1 - r$.

Proof. Define the multiplication operator $\tilde{B}: L_\infty^m(T_1) \rightarrow L_\infty^n(T_1)$ by

$$(\tilde{B}u)(t) = B(t)u(t) \quad (t \in T_1).$$

By Kurcyusz/Olbrot [17, Lemma 3], \tilde{B} has a closed image iff $B(t)^+$ is essentially bounded on T_1 . Hence $B(t)^+$ is essentially bounded and $\text{rank } B(t) = n$ ($t \in T_1$) iff the linear map B is surjective, hence open. Because of $0 \in \text{int}_\delta \text{co } \Omega(t)$, this in turn is equivalent to the existence of a neighborhood V of $0 \in \mathbb{R}^n$ such that

$$V \subset V_t = \text{co } B(t)\Omega(t) = B(t) \text{co } \Omega(t) \quad (t \in T_1).$$

This proves the equivalence of (i) and (ii). If (i) holds and $0 \in \mathbb{R}^n$ is locally reachable at $t_1 - r$, then $0 \in W^{n,\infty}[-r, 0]$ is locally reachable by Theorem 2.3 and Theorem 3.1.

The converse follows by Theorem 3.3 and the assumption

$$0 \in \text{int}_\delta \text{co } \Omega(t) (t \in T_1). \quad \square$$

Remark 3.4. The last equivalence in the corollary above is the analogue of results in the theory of *unconstrained* linear hereditary systems. Here the following is known (for controls $u \in L_p^m(T)$, $1 \leq p \leq \infty$):

Suppose that $B(t)^+$ is essentially bounded on T_1 . Then each element $\varphi \in W^{n,p}[-r, 0]$ is reachable at time t_1 , iff each element $\alpha \in \mathbb{R}^n$ is reachable at time $t_1 - r$ and $\text{rank } B(t) = n$ ($t \in T_1$) (see Jacobs/Kao [15], Banks/Jacobs/Langenhof [2], [3], Kurcyusz/Olbrot [17], Colonius/Hinrichsen [11]). In fact, by a category argument, Corollary 3.2 implies the result above for $p = \infty$. Furthermore, complete reachability in $W^{n,\infty}$ implies even that $B(t)^+$ is essentially bounded on T_1 . This proves a conjecture by Banks/Jacobs/Langenhof [3] (see also Kurcyusz/Olbrot [17, p. 48] and Colonius/Hinrichsen [11, p. 878]). In turn the cited result implies the last assertion of Corollary 3.2, provided that $B(t)^+$ is essentially bounded on T_1 as shown in Colonius [9, Remark 10].

Remark 3.5. If $B(t)$ depends continuously on t and has a constant rank, then by Kurcyusz/Olbrot [17, Lemma 4] the generalized inverse $B(t)^+$ is bounded.

Remark 3.6. Suppose there is $\varepsilon > 0$ such that for a.a. $t \in [t_1 - r - \varepsilon, t_1 - r]$ one has $0 \in \text{int}_\delta \text{co } \Omega(t)$ for a $\delta > 0$, and the generalized inverse $B(t)^+$ is essentially bounded and has rank n . Then by Theorem 2.2 $0 \in \mathbb{R}^n$ is locally reachable at time $t_1 - r$.

4. Analysis of the reachable set \mathcal{R} of final states and of the set \mathcal{T} of trajectories. In § 3, we analyzed reachability properties of a specific final state φ_1 and of a specific trajectory x^0 . Now we change our point of view by analyzing properties of the whole set \mathcal{T} of trajectories:

$$\mathcal{T} := \{x \in C^n[t_0 - r, t_1]: x \text{ is a trajectory of } \Sigma\}$$

and of the whole reachable set \mathcal{R} of final states

$$\mathcal{R} := \{\varphi \in W^{n,\infty}[-r, 0]: \text{there is } x \in \mathcal{T} \text{ with } \varphi = x_{t_1}\}.$$

Since the initial state φ_0 remains fixed throughout, we may identify $x \in \mathcal{T}$ with $x|T$ and consider \mathcal{T} as a subset of $W^{n,\infty}(T)$. The set \mathcal{R} will always be considered in $W^{n,\infty}$ -topology.

Observe that by definition $\varphi \in \text{int } \mathcal{R}$ iff φ is locally reachable.

LEMMA 4.1. Consider the system Σ_a of § 3, and assume that $x^0, x^1 \in \mathcal{T}$ correspond to $v^0, v^1 \in \mathcal{S}^*$ and that x^1 is regular. Then the trajectories $x^\gamma \in \mathcal{T}$ corresponding to $v^\gamma := (1 - \gamma)v^0 + \gamma v^1 \in \mathcal{S}^*$ where $0 < \gamma \leq 1$ are regular.

Proof. By regularity of x^1 there is $\delta > 0$ such that

$$\dot{x}^1(t) - a(x_t^1, t) = b(v^1(t), t) \in \text{int}_\delta \text{co } b(\Omega(t), t) \quad (t \in T_1).$$

Since $b(v^0(t), t) \in \text{co } b(\Omega(t), t)$ ($t \in T_1$) and $\text{co } b(\Omega(t), t)$ is convex, this implies for $0 < \gamma \leq 1$

$$\begin{aligned} \dot{x}^\gamma(t) - a(x_t^\gamma, t) &= b(v^\gamma(t), t) = (1 - \gamma)b(v^0(t), t) + \gamma b(v^1(t), t) \\ &\in \text{int}_{\gamma\delta} \text{co } b(\Omega(t), t) \quad (t \in T_1). \end{aligned}$$

This shows regularity of x^γ . \square

Remark 4.1. Using the same construction as in the proof above, one can see that for Σ_a the set \mathcal{T} is path connected, i.e., any two trajectories $x^0, x^1 \in \mathcal{T}$ can be connected by a continuous path $(x^\gamma, \gamma \in [0, 1])$. One only has to prove that the map $v^\gamma \rightarrow x^\gamma: \mathcal{S}^* \rightarrow W^{n,\infty}(T)$ is continuous with respect to norm topology on $W^{n,\infty}(T)$ and strong norm topology on \mathcal{S}^* . This follows by Lemma 1.3 and Lemma 1.2.

Define the finite dimensional reachable set \mathcal{R}^f at time $t_1 - r$ as

$$\mathcal{R}^f := \{\alpha \in \mathbb{R}^n : \text{there is } \varphi \in \mathcal{R} \text{ with } \alpha = \varphi(-r)\},$$

and let

$$\mathcal{R}_\alpha := \{\varphi \in \mathcal{R} : \varphi(-r) = \alpha\} \quad \text{for } \alpha \in \mathcal{R}^f.$$

Then

$$\mathcal{R} = \bigcup_{\alpha \in \mathcal{R}^f} \mathcal{R}_\alpha$$

is a decomposition of the reachable set parametrized by the elements of the finite dimensional reachable set at time $t_1 - r$. We identify \mathcal{R}_α with a subset of $L_\infty^n[-r, 0]$, and consider \mathcal{R}_α in the induced topology. Then

$$\text{int } \mathcal{R}_\alpha = \{\varphi_1 \in \mathcal{R}_\alpha : \text{there is a } \delta > 0 \text{ such that all } \varphi \in W^{n,\infty}[-r, 0]$$

$$\text{with } \varphi(-r) = \alpha \text{ and } \|\dot{\varphi}_1 - \dot{\varphi}_\infty\| < \delta \text{ are in } \mathcal{R}_\alpha\}.$$

Corollary 3.1 shows that each regularly reachable φ_1 lies in $\text{int } \mathcal{R}_{\varphi_1(-r)}$.

THEOREM 4.1. Assume that Σ_a satisfies (1.3). Then

(i) there exists a regularly reachable final state φ_1 in \mathcal{R}_α for each $\alpha \in \mathcal{R}^f$ iff

$$(4.1) \quad \text{int}_\delta \text{co } b(\Omega(t), t) \neq \emptyset \quad (t \in T_1) \quad \text{for a } \delta > 0.$$

(ii) Suppose (4.1) holds. Then the set of regularly reachable final states is open and dense in \mathcal{R}_α for each $\alpha \in \mathcal{R}^f$.

(iii) Suppose that φ_1 is reached regularly and stably with a certain trajectory x^0 . Then there is a neighborhood N of φ_1 in $W^{n,\infty}[-r, 0]$ such that all $\varphi \in N$ are regularly reachable.

Proof. (i) One direction is trivial. Suppose (4.1) holds and let $z^0 \in L_\infty^n(T_1)$ with

$$z^0(t) \in \text{int}_\delta \text{co } b(\Omega(t), t) \quad (t \in T_1).$$

For any trajectory x^0 define x^1 by

$$x^1(t) := x^0(t), \quad t \in [t_0 - r, t_1 - r], \quad \dot{x}^1(t) = a(x_t^1, t) + z^0(t) \quad (t \in T_1).$$

Then x^1 is a regular trajectory and $\varphi_1 := x_{t_1}^1 \in \mathcal{R}_{x^0(t_1-r)}$.

(ii) Let $\varphi_1 \in \mathcal{R}$ be reached with the trajectory x^0 corresponding to $v^0 \in \mathcal{S}^*$. By

(i) there exists $\varphi_2 \in \mathcal{R}_{\varphi_1(-r)}$ reached with a regular trajectory x^2 corresponding to

$v^2 \in \mathcal{S}^*$. Define x^1 as the trajectory corresponding to the control $v^1 \in \mathcal{S}^*$ which coincides with v^0 on $[t_0, t_1 - r)$ and with v^2 on T_1 . Then $x^1(t_1 - r) = \varphi_1(-r)$ and Lemma 4.1 together with Lemma 1.2 implies that in each neighborhood of φ_1 in $\mathcal{R}_{\varphi_1(-r)}$ there is a regularly reachable final state.

The set of regularly reachable final states is open in \mathcal{R}_α for $\alpha \in \mathcal{R}^f$ by the stability lemma 2.1.

(iii) The assumptions imply that there is $\delta > 0$ such that

$$\dot{\varphi}_1(t - t_1) - a(x_t^0, t) \in \text{int}_{3\delta} \text{co } b(\Omega(t), t) \quad (t \in T_1).$$

Furthermore there is a neighborhood N of φ_1 in $W^{n,\infty}[-r, 0]$, such that all $\varphi \in N$ are reached with trajectories x^φ satisfying

$$\text{ess sup}_{t \in T_1} |a(x_t^\varphi, t) - a(x_t^0, t)| < \delta.$$

If we choose N so small that $\|\varphi - \varphi_1\| < \delta$ for all $\varphi \in N$, we find

$$\dot{\varphi}(t - t_1) - a(x_t^\varphi, t) \in \text{int}_\delta \text{co } b(\Omega(t), t) \quad (t \in T_1). \quad \square$$

COROLLARY 4.1. Suppose that Σ_a satisfies (1.3). Then in $L_\infty^\text{n} \text{int}\{\dot{\varphi} \in L_\infty^\text{n}(T_1): \varphi \in \mathcal{R}\} \neq \emptyset$ iff (4.1) holds.

Proof. If condition (4.1) is satisfied, then by Theorem 4.1 there exists a regularly reachable final state φ_1 . By Corollary 3.1 we find $\dot{\varphi}_1 \in \text{int}\{\dot{\varphi} \in L_\infty^\text{n}(T_1): \varphi \in \mathcal{R}\}$. The converse follows by Theorem 3.3. \square

Remark 4.2. If (4.1) holds a.e. on $[t_1 - r - \varepsilon, t_1]$ for an $\varepsilon > 0$, then by Theorem 2.2 it follows that $\text{int } \mathcal{R} \neq \emptyset$.

Remark 4.3. If the function b does not depend explicitly on t and $\Omega(t) = \Omega_0$ ($t \in T$), then the condition (4.1) (nonempty interior in L_∞ -norm) reduces to $\text{int co } b(\Omega_0) \neq \emptyset$. Furthermore, by the remark above and Corollary 4.1, this condition is equivalent to $\text{int } \mathcal{R} \neq \emptyset$.

Remark 4.4. By Remark 2.4, functions b satisfying the requirements in (2.6) may be considered as elements in $L_\infty(T, C^n(\Omega_0))$. Now suppose that for a.a. $t \in T$ the set $\Omega(t)$ contains at least $n + 1$ points. Then the set of functions b satisfying condition (4.1) is open and dense in $L_\infty(T, C^n(\Omega_0))$, and in this sense condition (4.1) is *generic*. Openness is easily seen. Density can be proved using a construction similar to that in the proof of Lemma 4.1: There exist $n + 1$ measurable functions $\omega_i: T \rightarrow \Omega_0$, $i = 0, 1, \dots, n$ with $\omega_i(t) \in \Omega(t)$ and $\omega_i(t) \neq \omega_j(t)$ for a.a. $t \in T$ and $i \neq j$. Then for a.a. $t \in T$ there exists $b_1(t, \cdot) \in C^n(\Omega_0)$ mapping $\omega_0(t), \dots, \omega_n(t)$ into $n + 1$ points in \mathbb{R}^n being in general position, i.e., having the property $\text{int}_\delta \text{co}\{b_1(t, \omega_0(t)), \dots, b_1(t, \omega_n(t))\} \neq \emptyset$ for some $\delta > 0$. We may assume that δ is independent of t and that $b_1 \in L_\infty(T, C^n(\Omega_0))$. Let b_0 be an arbitrary element of $L_\infty(T, C^n(\Omega_0))$, and define

$$b^\gamma(t, \omega) := \gamma b_1(t, \omega) + (1 - \gamma)b_0(t, \omega) \quad \text{for } t \in T, \quad \omega \in \Omega_0.$$

Then $b^\gamma \rightarrow b_0$ in $L_\infty(T, C^n(\Omega_0))$ and

$$\text{int}_{\gamma\delta} \text{co}\{b\gamma(t, \omega_0(t)), \dots, b^\gamma(t, \omega_n(t))\} \neq \emptyset,$$

i.e., the functions $b\gamma$ satisfy condition (4.1).

Remark 4.5. Theorem 4.1 (ii) shows that regular reachability is *generic* in \mathcal{R}_α . Part (iii) shows regular reachability is preserved under small perturbations of φ_1 , provided that it is also stable.

For linear systems, we can completely characterize regular reachability by strengthening the properties (ii) and (iii) in Theorem 4.1.

COROLLARY 4.2. *Suppose that for the linear system Λ condition (4.1) is satisfied. Then $\varphi_1 \in W^{n,\infty}[-r, 0]$ is regularly reachable iff $\varphi_1 \in \text{int } \mathcal{R}_{\varphi_1(-r)}$. Furthermore, each element of $\text{int } \mathcal{R}$ is regularly reachable.*

Proof. In view of Corollary 3.1 we have to show that each $\varphi_1 \in \text{int } \mathcal{R}_{\varphi_1(-r)}$ is regularly reachable. By Theorem 4.1(ii) there is $\varphi = W^{n,\infty}[-r, 0]$ such that $\varphi_1 + \varphi \in \mathcal{R}_{\varphi_1(-r)}$ is regularly reachable, say with x^0 , and $\varphi_1 - \varphi \in \mathcal{R}_{\varphi_1(-r)}$ is reachable, say with x^1 . By linearity and Lemma 4.1,

$$\varphi_1 = \frac{1}{2}(\varphi_1 + \varphi) + \frac{1}{2}(\varphi_1 - \varphi)$$

is reached regularly with $\frac{1}{2}x^0 + \frac{1}{2}x^1$. Furthermore for each $\varphi \in \text{int } \mathcal{R}$ we have $\varphi \in \text{int } \mathcal{R}_{\varphi(-r)}$. This implies the last assertion. \square

Remark 4.6. It follows from Corollary 4.2 that each final state, which is reachable only with irregular trajectories lies in $\partial \mathcal{R}$ and clearly

$$\bigcup_{\alpha \in \mathcal{R}^f} \partial \mathcal{R}_\alpha \subset \partial \mathcal{R}.$$

However, there are regularly reachable final states in $\partial \mathcal{R} \setminus \bigcup_{\alpha \in \mathcal{R}^f} \partial \mathcal{R}_\alpha$: For instance, reach $\alpha \in \partial \mathcal{R}^f$ with a trajectory x^0 satisfying $\dot{x}^0(t) - L(t)x_t^0 \in \text{int}_\delta \text{co } b(\Omega(t), t) (t \in T_1)$ for a $\delta > 0$. Then $x_{t_1}^0 \in \partial \mathcal{R} \setminus \partial \mathcal{R}_\alpha$.

Regular reachability of a final function φ_1 means that there exists a certain trajectory reaching φ_1 regularly. Now in an optimization problem, we are interested in a specific *unknown* trajectory, the optimal one, out of all trajectories reaching φ_1 . Thus regular reachability of φ_1 is only *the minimal degree of well-posedness* we have to require in an optimal control problem. In the following, we go a step further by asking: What can be said about regularity of *all* trajectories reaching a given final function φ_1 ?

Consider the system Σ_a . Here the regularity condition for x^0 is satisfied iff there is a neighborhood V of $0 \in \mathbb{R}^n$ such that

$$V \subset -\dot{\varphi}_1(t - t_1) + a(x_t^0, t) + \text{co } b(\Omega(t), t) \quad (t \in T_1).$$

Now suppose that we can specify a uniform bound c_0 for $|a(x_t^0, t)|$, $t \in T_1$. If $\text{co } b(\Omega(t), t)$ can be made big enough compared to c_0 , all trajectories reaching φ_1 are regular. We give two examples where this idea applies.

Example 4.1.

$$\dot{x}(t) = a(x_t, t) + v(t) \quad (t \in T),$$

$$x_{t_0} = \varphi_0, \quad x_{t_1} = \varphi_1,$$

where a is as in the definition of Σ_a , and additionally, satisfies

$$|a(x_t, t)| \leq 1 \quad (t \in T_1)$$

for all trajectories x reaching φ_1 . x is regular, iff there is a neighborhood V of $0 \in \mathbb{R}^n$ such that

$$V + \dot{\varphi}_1(t - t_1) - a(x_t, t) \subset \text{int}_\delta \text{co } \Omega(t) \quad (t \in T_1).$$

for a $\delta > 0$.

If, e.g., $\varphi_1 = 0$, take Ω as a n -simplex with vertices $\omega_0, \omega_1, \dots, \omega_n$ containing the unit ball of \mathbb{R}^n in its interior. Then, obviously, all trajectories reaching φ_1 are regular. Observe that by Corollary 4.2 $\mathcal{R}_{\varphi_1(-r)}$ is open.

Example 4.2.

$$\dot{x}(t) = A_1 x(t-1) + v(t) \quad (t \in T := [0, k+1]),$$

$$x_0 = \varphi_0, \quad x_{k+1} = \varphi_1,$$

where $\varphi_0, \varphi_1 \in C^n[-1, 0]$, $k \in \mathbb{N}$, $A_1 \in M_{nn}$, $\Omega := \{\omega \in \mathbb{R}^n : |\omega| \leq q\}$ and $q > 0$. Then for $t \in [k, k+1]$

$$\begin{aligned} |A_1 x(t-1)| &\leq \|A_1\| \|x_k\|_\infty \leq \|A_1\| \{(1 + \|A_1\|) \|x_{k-1}\|_\infty + q\} \\ &\leq \|A_1\| \left\{ (1 + \|A_1\|)^k \|\varphi_0\|_\infty + q \sum_{j=0}^k (1 + \|A_1\|)^j \right\}. \end{aligned}$$

All trajectories reaching φ_1 are regular, if

$$(4.2) \quad \|\dot{\varphi}_1\|_\infty + \|A_1\| \left\{ (1 + \|A_1\|)^k \|\varphi_0\|_\infty + q \sum_{j=0}^k (1 + \|A_1\|)^j \right\} < q.$$

If, e.g., $\varphi_0 = \varphi_1 = 0$ and

$$(4.3) \quad \|A_1\| \sum_{j=0}^k (1 + \|A_1\|)^j < 1,$$

condition (4.2) is satisfied for all $q > 0$.

This simple example illustrates that the regularity requirement for *all* trajectories reaching a given final function is very restrictive: Observe that (4.3) implies $\|A_1\| < 1/(k+1)$. However, if Ω may change in time, regularity can be guaranteed if $\Omega(t)$ is big enough for $t \in T_1$.

The following theorem characterizes regular trajectories by a reachability property, and claims that regularity is a *generic* property of trajectories.

THEOREM 4.2. $x^0 \in \mathcal{T}$ is a regular trajectory of Σ_a iff in $L_\infty^n(T_1)$

$$(4.4) \quad (\dot{x}^0)_{t_1} \in \text{int}\{(\dot{x})_{t_1} : x \in \mathcal{T} \text{ and } x(t) = x^0(t) \text{ for } t \in [t_0 - r, t_1 - r]\}.$$

If φ_1 is a regularly reachable final state of Λ the set of regular trajectories is open and dense in $\{x^0 \in \mathcal{T} : x_{t_1}^0 = \varphi_1\}$.

Proof. One direction follows by Corollary 3.1. Conversely, assume that $x^0 \in \mathcal{T}$ satisfies (4.4). By Corollary 4.1 and the construction in the proof of Theorem 4.1(ii), there is in each neighborhood of $x^0 \in W^{n,\infty}(T)$ a regular trajectory x^1 coinciding with x^0 on $[t_0 - r, t_1 - r]$. By (4.4), x^1 may be chosen such that $x^2 := x^0 - (x^1 - x^0) \in \mathcal{T}$. Then by Lemma 4.1

$$x^0 = \frac{x^1}{2} + \frac{x^2}{2}$$

is regular.

The final assertion of the theorem follows again by the construction in the proof of Theorem 4.1(ii) and linearity. \square

Remark 4.7. The results above illustrate that regularity occurs also for higher dimensional systems with scalar control. In Example 4.1, take the vertices $\omega_0, \omega_1, \dots, \omega_n$ as images of the points i/n , and consider $\tilde{\Omega} := \{0, 1/n, \dots, 1\} \subset [0, 1]$ as new control set with control actions ω_i . More generally, condition (4.1) is, e.g.,

satisfied for the scalar control

$$b(\omega) = \begin{pmatrix} \omega \\ \omega^2 \\ \vdots \\ \omega^n \end{pmatrix}, \quad \text{where } \Omega(t) = \Omega_0 := [0, 1] \quad (t \in T).$$

Then $\text{int co } b(\Omega_0) \neq \emptyset$, and hence it follows for the linear system Λ that $\text{int } \mathcal{R} \neq \emptyset$; each element of $\text{int } \mathcal{R}$ is regularly reachable and the regular trajectories reaching φ_1 form an open, dense subset of the set of all trajectories reaching φ_1 .

Remark 4.8. Let $B \subset C^n[t_0 - r, t_1]$ and assume that there is a neighborhood V of $0 \in \mathbb{R}^n$ such that for all $x^0 \in B$

$$V \subset -\dot{\varphi}_1(t - t_1) + \text{co } f(x_t^0, \Omega, t) \quad (t \in T_1)$$

(compare Definition 3.1). Then it appears natural to apply a fixed point theorem in order to prove local reachability of φ_1 . In fact, if B is defined by certain growth conditions on f , and a finite dimensional condition for the reachability of $\varphi_1(-r)$ at time $t_1 - r$ is added, one can prove such a result using Kakutani's fixed point theorem (Colonius [8, Thms. 3.1, 3.3]). For Example 4.1, this yields that $\varphi_1 = 0$ is locally reachable [8, Beispiel 4.1]. In the context of control theory, this classical argument was used, e.g., by Tarnove [25]. Angell [1] and Chukwu [6], [7] applied it to hereditary differential systems in order to achieve reachability of a fixed final state.

Acknowledgment. I thank Diederich Hinrichsen for many fruitful discussions. This paper originated from my doctoral thesis which was guided by him.

REFERENCES

- [1] T. S. ANGELL, *Existence theorems for a class of optimal control problems with delay*, Doctoral thesis, University of Michigan, Ann Arbor, 1969.
- [2] H. T. BANKS, M. Q. JACOBS, AND C. E. LANGENHOP, *Function space controllability for linear functional differential equations*, Differential Games and Control Theory, E. O. Roxin, P. Liu and R. L. Sternberg, eds., Marcel Dekker, New York-Basel, 1974, pp. 81-98.
- [3] ———, *Characterization of the controlled states in $W_2^{(1)}$ of linear hereditary systems*, SIAM J. Control, 13 (1975), pp. 611-649.
- [4] G. R. BATES, *Hereditary optimal control problems*, Ph.D. thesis, Purdue University, W. Lafayette, IN, 1977.
- [5] L. D. BERKOVITZ, *A penalty function proof of the maximum principle*, Appl. Math. Optim., 2 (1975/6), pp. 291-303.
- [6] E. N. CHUKWU, *Functional inclusion and controllability of nonlinear neutral functional differential systems*, J. Opt. Theory Appl., 29 (1979), pp. 291-300.
- [7] ———, *Controllability of delay systems with restrained controls*, J. Opt. Theory Appl., 29 (1979), pp. 301-320.
- [8] F. COLONIUS, *Hereditäre differenzierbare Systeme mit Funktionenraum-Endbedingung und punktwise Steuerbeschränkungen: Notwendige Optimalitätsbedingungen und Erreichbarkeit*, Dissertation, Universität Bremen, Bremen, 1979.
- [9] ———, *Regularization of Lagrange multipliers for time delay systems with fixed final state*, in Optimization and Optimal Control, A. Auslender, W. Oettli and J. Stoer, eds., Springer-Verlag, Berlin-Heidelberg-New York, 1981, pp. 163-177.
- [10] ———, *The maximum principle for relaxed hereditary differential systems with function space end condition*, this Journal, this issue, pp. 695-712.
- [11] F. COLONIUS AND D. HINRICHSEN, *Optimal control of functional differential systems*, this Journal, 16 (1978), pp. 861-879.
- [12] M. C. DELFOUR AND S. K. MITTER, *Controllability, observability, and optimal feedback control of affine hereditary differential systems*, SIAM J. Control, 10 (1972), pp. 298-328.

- [13] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part I: General Theory*, Wiley-Interscience, New York, 1967.
- [14] J. HALE, *Theory of Functional Differential Equations*, second ed., Springer-Verlag, Berlin-Heidelberg-New York, 1977.
- [15] M. Q. JACOBS AND T. J. KAO, *An optimum settling problem for time lag systems*, J. Math. Anal. Appl., 40 (1972), pp. 687–707.
- [16] S. KURCYUSZ, *On the existence and nonexistence of Lagrange multipliers in Banach spaces*, J. Opt. Theory Appl., 20 (1976), pp. 81–110.
- [17] S. KURCYUSZ AND A. W. OLBROT, *On the closure in $W^{1,q}$ of the attainable subspace of linear time lag systems*, J. Differential Equations, 24 (1977), pp. 29–50.
- [18] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
- [19] A. MANITIUS, *Necessary and sufficient conditions of approximate controllability for general linear retarded systems*, this Journal, 19 (1981), pp. 516–532.
- [20] A. MANITIUS AND R. TRIGGIANI, *Function space controllability of linear retarded systems, a derivation from abstract operator conditions*, this Journal, 16 (1978), pp. 599–643.
- [21] I. P. NATANSON, *Theorie der Funktionen einer reellen Veränderlichen*, Akademie Verlag, Berlin, 1961.
- [22] M. N. OGUZTÖRELI, *Time-Lag Control Systems*, Academic Press, New York, 1966.
- [23] A. W. OLBROT, *Control of retarded systems with function space constraints, part 2: approximate controllability*, Control and Cybernetics, 6 (1977), pp. 17–71.
- [24] A. B. SCHWARZKOPF, *Relaxed control problems with state equality constraints*, SIAM J. Control, 13 (1975), pp. 677–694.
- [25] I. TARNOVE, *A controllability problem for nonlinear systems*, in Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967.
- [26] J. WARGA, *Optimal controls with pseudo-delays*, SIAM J. Control, 12 (1974), pp. 286–299.
- [27] ———, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

THE MAXIMUM PRINCIPLE FOR RELAXED HEREDITARY DIFFERENTIAL SYSTEMS WITH FUNCTION SPACE END CONDITION*

FRITZ COLONIUS†

Abstract. This paper contains a proof of the global pointwise maximum principle for relaxed hereditary differential systems with general function space end condition. First a multiplier theorem establishes the existence of Lagrange multipliers (l_0, l) , where $l_0 \in \mathbb{R}_+$ and l is in the dual of the Sobolev space $W^{n,\infty}[-r, 0]$. Then l can be identified with an element of $W^{n,\infty}[-r, 0]$ provided that the optimal trajectory satisfies a certain regularity condition. This yields two equivalent forms of the maximum principle. Using the results on regular reachability obtained in a companion paper, the maximum principle is shown to be—in a certain sense—generically valid.

Introduction. This paper deals with necessary optimality conditions in the following control problem for hereditary differential systems:

$$(0.1) \quad \text{Minimize} \int_{t_0}^{t_1} g(x(t), u(t), t) dt$$

subject to

$$(0.2) \quad \dot{x}(t) = f(x_t, u(t), t) \quad (t \in T := [t_0, t_1]),$$

$$(0.3) \quad x_{t_0} = \varphi_0,$$

$$(0.4) \quad h(x(t), t) = 0 \quad \text{all } t \in T_1 := [t_1 - r, t_1],$$

$$(0.5) \quad u(t) \in \Omega(t) \quad (t \in T),$$

where $x_t(s) := x(t+s) \in \mathbb{R}^n$, $s \in [-r, 0]$ and $t_1 - r > t_0$, $0 \leq r < \infty$, $g: \mathbb{R}^n \times \mathbb{R}^m \times T \rightarrow \mathbb{R}$, $f: C^n[-r, 0] \times \mathbb{R}^m \times T \rightarrow \mathbb{R}^n$, $h: \mathbb{R}^n \times T_1 \rightarrow \mathbb{R}^k$, $\varphi_0 \in C^n[-r, 0]$ and $\Omega(t) \subset \Omega_0 \subset \mathbb{R}^m$, Ω_0 compact, are given and $(t \in T)$ means for Lebesgue almost all $t \in T$. r denotes the length of the time delay. The state at time t of the hereditary system (0.2) is given by the function segment x_t . Hence, (0.4) is a condition for the final state x_{t_1} . This infinite-dimensional or “function space” end condition appears appropriate if the behavior of the system after t_1 is of any interest. We call h the output function of the system (0.2). The function space end condition (0.4)—opposed to a finite dimensional one $h(x(t_1), t_1) = 0$ —presents particular difficulties which were dealt with in a series of papers: Banks/Jacobs [2], Banks/Kent [3], Barbu [5], Barbu/Precupanu [6, Chap. 4, § 3], Bien [9], Bien/Chyung [10], Buehler [12], Colonius [13], [14], [16], Colonius/Hinrichsen [17], [18], Das [19], Jacobs [25], Jacobs/Kao [26], Kamenskii [27], [28], Kent [33], [34], Kurcyusz [35], Olbrot [39], Utthoff [45] (see also the surveys given in Banks [1], Banks/Manitius [4], Kamenskii/Skubachevskii [32]). Apparently, the first formulation of such a problem—in the context of the calculus of variations—was given by Elsgolts (see Zverkin et al. [54]).

It is appropriate to discuss the relation of the present paper to Banks/Kent [3], Barbu [5], Bien/Chyung [10], Olbrot [39] and Colonius [16] also dealing with constraints like (0.5).

Banks and Kent split the end condition into two conflicting inequality constraints and use methods by Neustadt in order to prove that the maximum principle is a necessary and, in case of normality and convexity, also sufficient optimality condition.

* Received by the editors November 11, 1980, and in final revised form October 17, 1981.

† Forschungsschwerpunkt Dynamische Systeme, Universität Bremen, Bibliothekstraße, Postfach 330 440, 2800 Bremen 33, West Germany.

However, due to their problem formulation, the Lagrange multipliers corresponding to the two conflicting inequality constraints may eliminate each other. Hence, in the necessity part, nontriviality cannot be guaranteed.

Barbu [5] uses the methods of convex analysis and the existence theory of differential equations associated with nonlinear monotone operators in Hilbert space. Problems with fixed final states and pointwise control constraints are formally included [5, Problem (2.4)–(2.6) and § 5]. However, the required assumption (local reachability in $W^{2,n}$ -norm) cannot be satisfied for bounded sets of admissible control values (see the discussion in Colonius [16, § 4]).

Colonius [16] gives a special approach to linear time invariant single delay systems with fixed final states. In this case stronger results than in the present paper are obtainable.

Bien and Chyung [10] transform the pure phase constraint into a mixed control/phase variable constraint using a classical device (cf. Pontryagin, et al. [41]). For a trajectory x , condition (0.4) is—under sufficient smoothness conditions—equivalent to

$$(0.6) \quad 0 = h(x(t_1 - r), t_1 - r),$$

$$(0.7) \quad 0 = \frac{d}{dt} h(x(t), t) = \frac{\partial}{\partial x} h(x(t), t) \dot{x}(t) + \frac{\partial}{\partial t} h(x(t), t) \\ = \frac{\partial}{\partial x} h(x(t), t) f(x(t), u(t), t) + \frac{\partial}{\partial t} h(x(t), t) \quad \text{a.a. } t \in T_1.$$

They generalize the theory of Makowski and Neustadt [37] to hereditary systems with a single constant delay (Olbro [39, Remark (6.3c)] proposes a similar procedure having reduced the retarded system to an unretarded one). The obtained maximum principle has only a local (in the sense of Girsanov [20]) form on the final interval T_1 . That is, the maximum condition has a differentiated form. It has to be assumed that the optimal solution satisfies a certain a priori condition. This regularity condition also has a local form and involves derivatives along the optimal solution as well as cone approximations to the set of admissible control values. As the authors remark, it appears very difficult to assure the validity of the regularity condition *before* the computation of the optimal solution. Furthermore, the regularity condition requires implicitly that the number m of independent control variables is not less than the dimension k of the output space. This restrictive condition appears also in the optimal control of nonlinear systems with *energy constrained* controls (see Kurcyusz [35]).

The problems connected with the end condition led Olbro [39] to another problem formulation. He required instead of (0.4) (for the fixed final state problem, where $h(x(t), t) = x(t) - \varphi_1(t - t_1)$, $\varphi_1: [-r, 0] \rightarrow \mathbb{R}^n$) that

$$\|x_{t_1} - \varphi_1\| \leq \varepsilon \quad \text{for a constant } \varepsilon > 0.$$

Here the norm may be taken in various Banach spaces. Then he proved necessary optimality conditions for these much simpler problems. In an engineering interpretation, the number ε specifies the accuracy required in reaching the final state, and the norm to measure the distance between the desired and the reached final state can be chosen on the basis of technological requirements. However, it is not clear what happens for $\varepsilon \rightarrow 0$; in particular, the problem might become ill behaved for small ε .

It appears more satisfactory to require that the end condition be fulfilled with *arbitrary accuracy*, i.e., we minimize over sequences of controls and corresponding

trajectories satisfying approximately the end condition (0.4). This can be achieved by a “relaxation” of the problem (see Young [52], Warga [48]). Following the approach by Warga [48], we consider the set $\mathcal{S}^\#$ of relaxed controls and in the “preproblem” (0.1)–(0.4), (0.5) replace the condition (0.5) by

$$(0.5) \quad v \in \mathcal{S}^\#$$

and insert v in (0.1), (0.2) instead of u . We denote (0.1)–(0.5) as Problem 1 and study necessary optimality conditions for an optimal solution (x^0, v^0) of this problem, which is well defined under the assumptions stated in § 1. This extends the approach in Colonius [13], [14], where the fixed final state problem was treated, to the general function space end condition (0.4).

The relaxation of the problem will allow us to weaken the regularity assumption needed in Bien/Chyung [10]. In particular, the condition $m \geq k$ on the dimensions of the control and output spaces, respectively, is no longer necessary.

We obtain a global, pointwise maximum principle, provided that the optimal trajectory satisfies a certain regularity condition referring to the infinite dimensional part (0.7) of the end condition. This regularity condition has a similar form to those given—in the theory of ordinary differential systems—by Warga [46], [47], Schwarzkopf [42] for inequality and by Schwarzkopf [43], [44] for equality constraints on control and phase variables.

Two equivalent forms of the maximum principle are stated corresponding to the formulations (0.4) and (0.6), (0.7) of the end condition. We exploit the results in Colonius [15] (this issue, pp. 675–694) in order to show that the maximum principle is—in a certain sense—generically valid.

This paper is built up as follows. In § 1, the assumptions are formulated. Section 2 establishes the existence of Lagrange multipliers $(l_0, l_1, l_2) \in \mathbb{R}_+ \times \mathbb{R}^k \times (L_\infty^k(T_1))^*$, provided that the infinite dimensional part of the attainable set of the linearized system has a nonempty interior. In § 3, the regularity assumption on the optimal trajectory is used to regularize l_2 . That is, l_2 can be identified with an element of $L_\infty^k(T_1)$. The global pointwise maximum principle is obtained as a straightforward consequence. Section 4 discusses the range of validity of the maximum principle.

We retain the notation and conventions of Colonius [15].

1. Assumptions. The following assumptions will be imposed throughout this paper:

(1.1) The functions $f: C^n[-r, 0] \times \mathbb{R}^m \times T \rightarrow \mathbb{R}^n$ and $g: \mathbb{R}^n \times \mathbb{R}^m \times T \rightarrow \mathbb{R}$ are continuous in $(\varphi, \omega) \in C^n[-r, 0] \times \mathbb{R}^m$ and $(y, \omega) \in \mathbb{R}^n \times \mathbb{R}^m$, respectively, and measurable in $t \in T$.

(1.2) There are $p, q: \mathbb{R}_+ \times T \rightarrow \mathbb{R}_+$ such that for all $\varphi \in C^n[-r, 0]$, $y \in \mathbb{R}^m$ and $\omega \in \Omega_0$,

$$|f(\varphi, \omega, t)| \leq p(\|\varphi\|_\infty, t) \quad (t \in T),$$

$$|g(y, \omega, t)| \leq q(|y|, t) \quad (t \in T),$$

and $p(s, \cdot) \in L_\infty^1(T)$, $q(s, \cdot) \in L_2^1(T)$ and $p(\cdot, t)$, $q(\cdot, t)$ are monotonically increasing for all $(s, t) \in \mathbb{R}_+ \times T$.

(1.3) The functions f and g are continuously Fréchet differentiable in the first argument; the corresponding derivatives $D_1 f(\varphi, \omega, t)$ and $D_1 g(y, \omega, t)$ are continuous in (φ, ω, t) and (y, ω) , respectively; $h = (h_1, \dots, h_k): \mathbb{R}^n \times T_1 \rightarrow \mathbb{R}^k$ is continuously Fréchet differentiable and the derivative is continuously Fréchet differentiable with

respect to $y \in \mathbb{R}^n$;

$$|D_1 f(\varphi, \omega, t)| \leq p(\|\varphi\|_\infty, t) \quad (t \in T),$$

$$|D_1 g(y, \omega, t)| \leq q(|y|, t) \quad (t \in T),$$

where p, q are as in (1.2).

(1.4) For each relaxed control $v \in \mathcal{S}^*$, there is a unique trajectory $x \in C^n[t_0 - r, t_1]$ satisfying (0.2) and (0.3).

The conditions (1.1)–(1.3) may be required only in a neighborhood of the optimal solution similarly as in Berkovitz [8], Bates [7]. The assumptions on f coincide with Colonius [15, (1.1)–(1.4)]. Hence, [15, Lemmas 1.1, 1.2, 1.3 and 1.5] are valid here. The continuity (instead of integrability) assumption on $D_1 f(\varphi, \omega, t)$ in t is stronger than desirable. However, it is needed for [15, Lemma 1.2] used in the proof of Lemma 2.1 below. By [15, Thm. 2.1], the relaxed problem admits the interpretation given in the introduction. Furthermore, existence of an optimal solution is guaranteed:

THEOREM 1.1. *If the trajectories x , satisfying (0.2) and (0.3), are uniformly bounded, there exists an optimal solution (x^0, v^0) of Problem 1.*

The proof of this theorem follows from compactness of \mathcal{S}^* and Colonius [15, Lemma 1.3].

In the rest of this paper we assume that (x^0, v^0) is an optimal solution of Problem 1. Furthermore, for the sake of simplicity, we let $\varphi_0(0) = 0$.

2. The abstract maximum principle. In this section, we reformulate Problem 1 in an abstract setting of operators on Banach spaces. Then we obtain the existence of Lagrange multipliers $(l_0, l_1, l_2) \in \mathbb{R}_+ \times \mathbb{R}^k \times (L_\infty^k(T_1))^*$ provided that the output set of the linearized system has a nonempty interior. This yields two different forms of the abstract maximum principle reflecting the two formulations (0.4) and (0.6), (0.7) of the end condition. These two formulations are equivalent under the assumptions stated in § 1.

For the ease of notation, we introduce

$$\mathbf{C}^n(T) := \{x \in C^n(T) : x(t_0) = 0\},$$

$$\mathbf{W}^{n,\infty}(T) := \{x \in W^{n,\infty}(T) : x(t_0) = 0\}.$$

Observe that the natural embedding of $\mathbf{W}^{n,\infty}(T)$ into $\mathbf{C}^n(T)$ is compact. $\mathbf{W}^{n,\infty}(T)$ is in a natural way isomorphic to $L_\infty^n(T)$. We extend each element x of $\mathbf{W}^{n,\infty}(T)$ (resp. $\mathbf{C}^n(T)$) to a continuous function $x : [t_0 - r, t_1] \rightarrow \mathbb{R}^n$ by $x_{t_0} := \varphi_0$.

Define

$$G : \mathbf{C}^n(T) \times \mathcal{N} \rightarrow \mathbb{R}$$

$$\text{by } G(x, v) := \int_T g(x(t), v(t), t) dt;$$

$$F : \mathbf{C}^n(T) \times \mathcal{N} \rightarrow \mathbf{W}^{n,\infty}(T)$$

$$\text{by } F(x, v)(t) := \int_{t_0}^t f(x_s, v(s), s) ds, \quad t \in T;$$

$$B : \mathbf{C}^n(T) \rightarrow \mathbb{R}^k$$

$$\text{by } B(x) := h(x(t_1 - r), t_1 - r);$$

$$C: \mathbf{C}^n(T) \times \mathcal{N} \rightarrow L_\infty^k(T_1)$$

$$\text{by } C(x, v)(t) := \frac{\partial}{\partial x} h(x(t), t) f(x, v(t), t) + \frac{\partial}{\partial t} h(x(t), t), \quad t \in T_1;$$

$$K: \mathbf{W}^{n,\infty}(T) \rightarrow L_\infty^k(T_1)$$

$$\text{by } (Kx)(t) := \frac{\partial}{\partial x} h(x(t), t) \dot{x}(t) + \frac{\partial}{\partial t} h(x(t), t), \quad t \in T_1.$$

Then (B, K) and (B, C) may be considered as maps with values in

$$W^{k,\infty}(T_1) = \mathbb{R}^k \times L_\infty^k(T_1).$$

Define the solution map $S: \mathcal{S}^* \rightarrow W^{n,\infty}(T)$ by $S(v) = x$, where x is the response of system (0.2) with initial condition (0.3) to the relaxed control v .

Problem 1 is equivalent to the abstract one, Problem 2.

Problem 2.

$$\underset{v \in \mathcal{S}^*}{\text{Minimize}} \ G(S(v), v)$$

subject to

$$(B, K)(S(v)) = 0.$$

Observe that

$$\begin{aligned} (2.1) \quad (B, K)(S(v)) &= (B(S(v)), K(S(v))) = (B(S(v)), C(S(v), v)) \\ &= (B, C)(S(v), v). \end{aligned}$$

Remark 2.1. The problem formulation with the operator (B, K) will yield the form of the maximum principle proposed by Banks/Kent [3], the formulation with (B, C) will yield the form given by Bien/Chyung [10].

In the following, we assure differentiability of the involved operators and give concrete representations for the derivatives. First, by the Riesz representation theorem, there is a $(n \times n)$ -matrix function η defined on $T \times [t_0 - r, t_1]$ such that

$$(2.2) \quad D_1 f(x_s^0, v^0(s), s) x_s = \int_{t_0-r}^s d_t \eta(s, t) x(t) \quad (s \in T)$$

for all $x \in C^n[t_0 - r, t_1]$; here $\eta(\cdot, t)$ is measurable and $\eta(s, \cdot)$ is of bounded variation for $(s, t) \in T \times [t_0 - r, t_1]$. The integral is meant in the Riemann–Stieltjes sense.

For all $s \in T$, we require that $\eta(s, \cdot)$ is normalized, i.e., left continuous on $(t_0 - r, t_1)$ and $\eta(s, t) = \eta(s, s) = 0$, $t_0 \leq s \leq t \leq t_1$. This determines $\eta(s, \cdot)$ uniquely. One can show (using Bourbaki [11, § 8, exc. 6]) that η is even measurable on the rectangle $T \times [t_0 - r, t_1]$. Define $h(x, t) = 0$ for $x \in \mathbb{R}^n$, $t \in T \setminus T_1$, and let for $t \in T$:

$$(2.3) \quad H_x(t) := \left(\frac{\partial}{\partial x_j} h_i(x^0(t), t) \right)_{\substack{i=1, \dots, k, \\ j=1, \dots, n,}}$$

$$(2.4) \quad H_{xx_j}(t) := \left(\frac{\partial^2}{\partial x_l \partial x_j} h_i(x^0(t), t) \right)_{\substack{i=1, \dots, k, \\ l=1, \dots, n,}}$$

$$(2.5) \quad H_{xt}(t) := \left(\frac{\partial^2}{\partial x_j \partial t} h_i(x^0(t), t) \right)_{\substack{i=1, \dots, k, \\ j=1, \dots, n,}}$$

- LEMMA 2.1. (i) *The maps F and G are linear in $v \in \mathcal{N}$, C is affine in $v \in \mathcal{N}$.*
(ii) *The restrictions of F , G , (B, C) and (B, K) to $\mathbf{W}^{n,\infty}(T) \times \mathcal{S}^*$ are weakly* continuous, S is weakly* continuous.*
(iii) *The maps B , C , F , G , K and S are continuously Fréchet differentiable.*
(iv) *The derivatives have the following form:*

$$[D_1 F(x_0, v_0)x](t) = \int_{t_0}^t D_1 f(x_s^0, v^0(s), s)x_s ds, \quad t \in T,$$

$$D_1 G(x^0, v^0)x = \int_T \frac{\partial}{\partial x} g(x^0(t), v^0(t), t)x(t) dt,$$

$$DB(x^0)x = H_x(t_1 - r)x(t_1 - r),$$

$$\begin{aligned} [D_1 C(x^0, v^0)x](t) &= \sum_{j=1}^n f_j(x_t^0, v^0(t), t)H_{xx_j}(t)x(t) \\ &\quad + H_x(t)D_1 f(x_t^0, v^0(t), t)x_t + H_{xt}(t)x(t), \quad t \in T_1, \text{ where } x_{t_0} := 0, \end{aligned}$$

$$[DK(x^0)x](t) = \sum_{j=1}^n \dot{x}_j^0(t)H_{xx_j}(t)x(t) + H_x(t)\dot{x}(t) + H_{xt}(t)x(t), \quad t \in T_1.$$

For $v \in \mathcal{S}^*$, the function $x(v) := DS(v^0)(v - v^0)$ satisfies

$$\begin{aligned} \dot{x}(t) &= D_1 f(x_t^0, v^0(t), t)x_t + f(x_t^0, v(t) - v^0(t), t) \quad (t \in T), \\ (2.6) \quad x_{t_0} &= 0. \end{aligned}$$

Proof. Assertion (i) follows directly from the definitions. Assertion (ii) follows from Colonius [15, Lemma 1.1]. The operators B , C , F , G , K are continuously Fréchet differentiable by [15, Lemma 1.2] and the chain rule. This yields also the form of the derivatives as indicated in (iv). Finally, the existence and the form of the derivative $DS(v^0)$ follows from [15, Lemma 1.5]. \square

Consider the linearized system (2.6) in its abstract form

$$(2.7) \quad x = D_1 F(x^0, v^0)x + F(x^0, v - v^0), \quad v \in \mathcal{S}^*,$$

with output

$$(2.8) \quad (DB(x^0), DK(x^0))(x) \in W^{k,\infty}(T_1).$$

Then the chain rule and (2.1) imply

$$(2.9) \quad DK(x^0)x(v) = D_1 C(x^0, v^0)x(v) + C(x^0, v) - C(x^0, v^0), \quad v \in \mathcal{S}^*.$$

Define the set \mathcal{A} of attainable output values by

$$\mathcal{A} := \{\varphi \in W^{k,\infty}(T_1): \text{there is a } v \in \mathcal{S}^* \text{ such that } \varphi = (DB(x^0), DK(x^0))x(v)\},$$

and let

$$\pi_\infty \mathcal{A} := \{z \in L_\infty^k(T_1): \text{there is } \varphi \in \mathcal{A} \text{ with } z = \dot{\varphi}\}.$$

The set \mathcal{A} will always be considered in $W^{k,\infty}$ -norm and the set $\pi_\infty \mathcal{A}$ in L_∞^k -norm. We have

$$\begin{aligned} (2.10) \quad \pi_\infty \mathcal{A} &= \{z \in L_\infty^k(T_1): \text{there is a } v \in \mathcal{S}^* \text{ such that } z = D_1 C(x^0, v^0)x(v) + C(x^0, v)\}, \end{aligned}$$

since

$$\begin{aligned} 0 &= \frac{d}{dt} h(x^0(t), t) = \frac{\partial}{\partial x} h(x^0(t), t) f(x_t^0, v^0(t), t) + \frac{\partial}{\partial t} h(x^0(t), t) \\ &= [C(x^0, v^0)](t) \quad (t \in T_1). \end{aligned}$$

THEOREM 2.1. *Let $v^0 \in \mathcal{S}^*$, $x^0 = S(v^0)$ be an optimal solution of Problem 2 and assume that $\text{int } \pi_\infty \mathcal{A} \neq \emptyset$. Then there are nontrivial Lagrange multipliers $(l_0, l) := (l_0, l_1, l_2) \in \mathbb{R}_+ \times \mathbb{R}^k \times (L_\infty^k(T_1))^*$ such that*

$$(2.11) \quad \begin{aligned} &l_0 D_1 G(x^0, v^0) x(v) + l_0 G(x^0, v - v^0) + l_1 \cdot DB(x^0) x(v) \\ &+ l_2 \circ DK(x^0) x(v) \geq 0 \quad \text{for all } v \in \mathcal{S}^*. \end{aligned}$$

If $0 \in \text{int } \pi_\infty \mathcal{A}$, then $(l_0, l_1) \neq (0, 0)$, and if $0 \in \text{int } \mathcal{A}$, then $l_0 \neq 0$.

Proof. If $0 \in \partial \pi_\infty \mathcal{A}$, there is a support functional l_2 at 0 to $\pi_\infty \mathcal{A}$, and (2.11) is satisfied with $l_0 = 0$, $l_1 = 0$. Now let $0 \in \text{int } \pi_\infty \mathcal{A}$. We verify the assumptions in Colonius [16, Thm. 1.3].

By Lemma 2.1, $v \mapsto G(S(v), v)$ and $v \mapsto (B, K)S(v)$ are continuously Fréchet differentiable and weakly* continuous. By Warga [48, Thm. IV.3.11], \mathcal{S}^* is weakly* compact and convex. The space $L_\infty^k(T_1)$ can weakly* continuously be embedded into the Hilbert space $L_2^k(T_1)$. Hence there are Lagrange multipliers $(l_0, l_1, l_2) \in \mathbb{R}_+ \times \mathbb{R}^k \times (L_\infty^k(T_1))^*$ satisfying (2.11) and $(l_0, l_1) \neq (0, 0)$. By the same theorem, $l_0 \neq 0$ if $0 \in \text{int } \mathcal{A}$. \square

Remark 2.2. In the case where $0 \in \text{int } \mathcal{A}$, Theorem 2.1 follows easily from the multiplier theorem by Zowe/Kurcyusz [53, Thm. 2.1], which does not presuppose properties with respect to weak* topology.

Remark 2.3. By weak* continuity and compactness of \mathcal{S}^* , $\pi_\infty \mathcal{A}$ is a weakly* closed convex set in $L_\infty^k(T_1)$. Hence, Phelps [40, Thm. 1] shows that the set M of points in $\pi_\infty \mathcal{A}$ admitting a weakly* continuous support functional $l_2 \neq 0$ is norm dense in the norm boundary of $\pi_\infty \mathcal{A}$. If $0 \in M$, then $(0, 0, l_2) \in \mathbb{R}_+ \times \mathbb{R}^k \times L_1^k(T_1) \subset \mathbb{R}_+ \times \mathbb{R}^k \times (L_\infty^k(T_1))^*$ are Lagrange multipliers satisfying (2.11). This shows that also in the case where $\text{int } \pi_\infty \mathcal{A} = \emptyset$ there are “many” points in $\pi_\infty \mathcal{A}$ admitting nontrivial Lagrange multipliers, even with $l_2 \in L_1^k(T_1)$. The same argument applies where $\text{int } \mathcal{A} = \emptyset$. Then one obtains Lagrange multipliers $(0, l) \in \mathbb{R}_+ \times W^{k,1}(T_1)$.

COROLLARY 2.1. *Let the assumptions of Theorem 2.1 be satisfied. Then*

(i) *the adjoint variable $y^K \in (\mathbf{W}^{n,\infty}(T))^*$ defined by*

$$y^K = D_1 F(x^0, v^0)^* y^K + l_0 D_1 G(x^0, v^0) + DB(x^0)^* l_1 + DK(x^0)^* l_2$$

satisfies

$$l_0 G(x^0, v - v^0) + y^K \circ F(x^0, v - v^0) \geq 0 \quad \text{for all } v \in \mathcal{S}^*.$$

(ii) *The adjoint variable $y^C \in (\mathbf{C}^n(T))^*$ defined by*

$$y^C = D_1 F(x^0, v^0)^* y^C + l_0 D_1 G(x^0, v^0) + DB(x^0)^* l_1 + D_1 C(x^0, v^0)^* l_2$$

satisfies

$$l_0 G(x^0, v - v^0) + y^C \circ F(x^0, v - v^0) + l_2 \circ C(x^0, v) \geq 0 \quad \text{for all } v \in \mathcal{S}^*.$$

(iii) *The difference $y^K - y^C \in (\mathbf{W}^{n,\infty}(T))^*$ satisfies*

$$y^K - y^C = D_1 F(x^0, v^0)^* (y^K - y^C) + [DK(x^0)^* - D_1 C(x^0, v^0)^*] l_2.$$

Proof. Observe that both adjoint equations are uniquely solvable since $\text{Id}_{\mathbf{C}(T)^*} - D_1F(x^0, v^0)^*$ and $\text{Id}_{W^{n,\infty}(T)^*} - D_1F(x^0, v^0)^*$ are isomorphisms, where Id denotes the identity map on the respective spaces. By definition of y^K , (2.11) and (2.7), we find

$$\begin{aligned} & l_0 G(x^0, v - v^0) + y^K \cdot F(x^0, v - v^0) \\ &= l_0 G(x^0, v - v^0) + [l_0 D_1 G(x^0, v^0) + l_1 \cdot DB(x^0) + l_2 \cdot DK(x^0)] \\ & \quad \cdot (\text{Id} - D_1 F(x^0, v^0))^{-1} F(x^0, v - v^0) \\ &= l_0 G(x^0, v - v^0) + l_0 D_1 G(x^0, v^0) x(v) + l_1 \cdot DB(x^0) x(v) + l_2 \cdot DK(x^0) x(v) \\ &\cong 0 \quad \text{for all } v \in \mathcal{S}^*. \end{aligned}$$

Assertions (ii) and (iii) follow similarly, taking into account (2.9). \square

3. The global pointwise maximum principle. The abstract Theorem 2.1 is only a first step. The optimality conditions involve the multiplier $l_2 \in (L_\infty^k(T_1))^*$ which may not be identifiable with a real function. In order to regularize l_2 , we make use of the following notion.

DEFINITION 3.1. A trajectory x^0 satisfying (0.2), (0.3) is called *regular* (with respect to (0.4)) if and only if there exists a neighborhood V of $0 \in \mathbb{R}^k$ such that

$$(3.1) \quad V \subset \frac{\partial}{\partial x} h(x^0(t), t) [\text{co } f(x_t^0, \Omega(t), t) - \dot{x}^0(t)] \quad (t \in T_1).$$

Observe that 0 is always in the set at the right-hand side. By Colonius [15, Lemma 1.4] x^0 is regular if and only if there is a neighborhood V_∞ of $0 \in L_\infty^k(T_1)$ such that

$$V_\infty \subset \{C(x^0, v) : v \in \mathcal{S}^*\}.$$

The proof of the global pointwise maximum principle is prepared by the following two lemmas, which contain the crux of the proof.

LEMMA 3.1. *If the optimal trajectory x^0 is regular, then $0 \in \text{int } \pi_\infty \mathcal{A}$.*

Proof. First we construct an inverse of the map $v \mapsto C(x^0, v)$ from \mathcal{S}^* into $L_\infty^k(T_1)$. Take V in (3.1) as a k -simplex with vertices e_1, e_2, \dots, e_{k+1} . Let $e_0 := 0$. Since $0 \in \text{int } V$, we can divide V into $k+1$ subsimplices V^j , $j = 1, \dots, k+1$ with vertices $e_0, \dots, e_{j-1}, e_{j+1}, \dots, e_{k+1}$. Then there are $v^i \in \mathcal{S}^*$ such that

$$e_i = C(x^0, v^i), \quad i = 1, \dots, k+1,$$

and we know that

$$e_0 = 0 = C(x^0, v^0).$$

We may assume that for $i = 1, \dots, k+1$,

$$v^i(t) = v^0(t), \quad t \in [t_0, t_1 - r].$$

By definition of C ,

$$e_i = H_x(t) f(x_t^0, v^i(t) - v^0(t), t) \quad (t \in T_1).$$

For $t \in T$ we define continuous piecewise affine maps

$$A_t : V \rightarrow \text{rpm}(\Omega_0)$$

in the following way:

$$A_t(e^i) := v^i(t), \quad i = 0, 1, \dots, k+1.$$

Each $z \in V$ lies in some V^j . Consider the barycentric coordinates α_i with respect to V^j , i.e.,

$$z = \sum_{\substack{i=0 \\ i \neq j}}^{k+1} \alpha_i e_i,$$

where $\alpha_i \geq 0$, $\sum_{i=0, i \neq j}^{k+1} \alpha_i = 1$, and define

$$A_t z := \sum_{\substack{i=0 \\ i \neq j}}^{k+1} \alpha_i A_t(e_i).$$

We have to show that this is well defined. Clearly, $A_t z \in \text{rpm}(\Omega_0)$. Suppose that $z \in V^j \cap V^l$. Then only those coordinates α_i of z in V^j (resp. V^l) are nonzero for which $e_i \in V^j \cap V^l$. These coordinates have the same value in V^j and V^l . Now let

$$V_\infty := \{z \in L_\infty^k(T_1) : z(t) \in V(t \in T_1)\}$$

and define $A : V_\infty \rightarrow \mathcal{S}^\#$ by

$$(Az)(t) := A_t z(t) \quad (t \in T).$$

Then $Az \in \mathcal{S}^\#$, since the coordinate functions of z may be chosen measurable and the support of $(Az)(t)$ is contained in $\Omega(t)$. We have

$$(3.2) \quad C(x^0, Az) = z$$

for all $z \in V_\infty$, since C and A satisfy this equality pointwise.

A is extended to a map $A : L_\infty^k(T_1) \rightarrow \mathcal{N}$ by an affine continuation of A_t in the $k+1$ sectors of \mathbb{R}^n corresponding to V^j , $j=0, 1, \dots, k$. Then A satisfies a Lipschitz condition for a constant L , 0 (see Warga [48, p. 268] for the definition of $\|\cdot\|_{\mathcal{N}}$):

$$(3.3) \quad \|Az_1 - Az_2\|_{\mathcal{N}} \leq L \|z_1 - z_2\|_\infty.$$

Now consider the equation

$$(3.4) \quad x = D_1 F(x^0, v^0)x + F(x^0, A(z - D_1 C(x^0, v^0)x)) - F(x^0, v^0).$$

Since for fixed $z \in L_\infty^k(T_1)$ the value of the right-hand side depends only on $x_t \in C^n[-r, 0]$, this equation can be written as

$$\dot{x}(t) = f^z(x_t, t) \quad (t \in T), \quad x_{t_0} = 0.$$

Using (3.3) and the assumptions (1.1)–(1.3) one finds that f^z is continuous in the first argument and measurable in the second argument; furthermore, f^z satisfies a global Lipschitz condition uniformly with respect to z in the first argument. Hence for each $z \in L_\infty^k(T_1)$ there exists a unique solution of this functional differential equation (cf. Hale [22, Thm. 2.3 and p. 55], which by definition is equivalent to (3.4).

Take $z = 0$. Then $x = 0$ solves (3.4) since

$$F(x^0, A(z - D_1 C(x^0, v^0)x)) = F(x^0, A(0)) = F(x^0, v^0).$$

The uniform global Lipschitz condition and Gronwall's inequality imply that x depends continuously on the right-hand side of (3.4). Hence there is $\delta > 0$ such that for all z with $\|z\|_\infty < \delta$,

$$z - D_1 C(x^0, v^0)x \in V_\infty.$$

Thus by construction of A there is $v \in \mathcal{S}^\#$ with

$$(3.5) \quad v = A(z - D_1 C(x^0, v^0)x).$$

Then

$$(3.6) \quad x = D_1 F(x^0, v^0)x + F(x^0, v - v^0),$$

and by (3.2)

$$(3.7) \quad C(x^0, v) = z - D_1 C(x^0, v^0)x.$$

Since this holds for all z with $\|z\|_\infty < \delta$, we find by (2.10) that $0 \in \text{int } \pi_\infty \mathcal{A}$. \square

LEMMA 3.2. *If the optimal trajectory x^0 is regular and (l_0, l_1, l_2) are Lagrange multipliers satisfying (2.11), then l_2 can be identified with a function $\rho \in L_\infty^k(T_1) \subset (L_\infty^k(T_1))^*$.*

Proof. Consider the subspace S of simple functions in $L_\infty^k(T_1)$. S is dense in L_∞ (see, e.g., Hewitt/Stromberg [23, Thm. 11.35]). We shall prove that $l_2|_S$ is continuous with respect to L_1 -norm on S . Then $l_2|_S$ can be extended to a continuous linear functional on $L_1^k(T_1)$ which by duality of L_1 and L_∞ can be identified with an element ρ of $L_\infty^k(T_1)$. Then l_2 and the functional defined by ρ coincide on S , hence, on $L_\infty^k(T_1)$.

The general element $s \in S$ has the form

$$s(t) = \sum_{i=1}^l \sum_{j=1}^k s_{ij} \chi_{E_i}(t) a_j \quad (t \in T_1),$$

where $s_{ij} \in \mathbb{R}$, $\{a_j\}$ is a base of \mathbb{R}^n and $\{E_i\}$ is a measurable decomposition of T_1 .

Since x^0 is regular, we may assume that $\pm \chi_{E_i}(t) a_j \in V$, where V satisfies (3.1). Then there are $v_{ij}^\pm \in \mathcal{S}^\#$ such that

$$\pm \chi_{E_i} a_j = C(x^0, v_{ij}^\pm)$$

and

$$(3.8) \quad \int_T \|v_{ij}^\pm(t) - v^0(t)\| dt < 2\lambda(E_i).$$

Decompose s into its positive and negative parts:

$$s = \sum_{i=1}^l \sum_{j=1}^k (s_{ij}^+ - s_{ij}^-) \chi_{E_i} a_j,$$

where $s_{ij}^\pm := \max(0, \pm s_{ij})$.

Apply the maximum condition in Corollary 2.1(ii) $2kl$ times in order to obtain

$$\begin{aligned} l_2(s) &= l_2\left(\sum_{i,j} (s_{ij}^+ - s_{ij}^-) \chi_{E_i} a_j\right) \\ &= \sum_{i,j} \{s_{ij}^+ l_2 \circ C(x^0, v_{ij}^+) + s_{ij}^- l_2 \circ C(x^0, v_{ij}^-)\} \\ &\geq -\sum_{i,j} \{s_{ij}^+ [l_0 G(x^0, v_{ij}^+ - v^0) + y^C \circ F(x^0, v_{ij}^+ - v^0)] \\ &\quad + s_{ij}^- [l_0 G(x^0, v_{ij}^- - v^0) + y^C \circ F(x^0, v_{ij}^- - v^0)]\} \\ &\geq -c_0 \sum_{i,j} (s_{ij}^+ + s_{ij}^-) \lambda(E_i) \\ &= -c_0 \|s\|_{L_1} \end{aligned}$$

for a constant $c_0 > 0$; here we used (3.8) and the assumptions (1.1) and (1.2).

Hence, for $\|s\|_{L_1} \rightarrow 0$,

$$\liminf l_2(s) \geq 0;$$

the same argument for $-s$ proves that $l_2(s) \rightarrow 0$ for $\|s\|_{L_1} \rightarrow 0$. Thus, l_2 is continuous on S in L_1 -norm and the lemma is proven. \square

The following maximum principle has the form proposed by Banks/Kent [3] for a more general class of systems (including neutral equations) with fixed final state.

THEOREM 3.1 (BK-form of the maximum principle). *Let x^0, v^0 be an optimal solution of Problem 1 and assume that x^0 is a regular trajectory.*

Then there exist Lagrange multipliers $(l_0, l_1, \rho) \in \mathbb{R}_+ \times \mathbb{R}^k \times L_\infty^k(T_1)$ such that $(l_0, l_1) \neq (0, 0)$ and the adjoint variable $\psi^K \in L_\infty^k(T_1)$ defined by

$$\begin{aligned} \text{(i)} \quad \psi^K(t) = & - \int_t^{t_1} \eta(s, t)^* \psi^K(s) ds + l_0 \int_t^{t_1} \frac{\partial}{\partial x} g(x^0(s), v^0(s), s) ds \\ & + H_x(t_1 - r)^* l_1 + \int_{T_1} \left[\sum_{j=1}^n \dot{x}_j^0(s) H_{xx_j}(s)^* + H_{xt}(s)^* \right] \rho(s) ds \end{aligned}$$

for $t \in [t_0, t_1 - r]$,

$$\begin{aligned} \psi^K(t) = & - \int_t^{t_1} \eta(s, t)^* \psi^K(s) ds + l_0 \int_t^{t_1} \frac{\partial}{\partial x} g(x^0(s), v^0(s), s) ds \\ & + H_x(t)^* \rho(t) + \int_t^{t_1} \left[\sum_{j=1}^n \dot{x}_j^0(s) H_{xx_j}(s)^* + H_{xt}(s)^* \right] \rho(s) ds \end{aligned}$$

for $t \in (t_1 - r, t_1]$, satisfies

$$\text{(ii)} \quad l_0 g(x^0(t), v^0(t), t) + \psi^K(t) f(x_t^0, v^0(t), t) \geq l_0 g(x^0(t), \omega, t) + \psi^K(t) f(x_t^0, \omega, t)$$

for all $\omega \in \Omega(t)$, a.a. $t \in T$.

Proof. In view of Corollary 2.1(i), Lemma 3.1 and Lemma 3.2, we have to compute adjoint operators. By partial integration, $DK(x^0)^* l_2 \in (W^{n, \infty}(T))^*$ can be identified with the following element of $W^{n, \infty}(T)$:

$$DK(x^0)^* l_2 = \begin{cases} \int_{T_1} \left[\sum_{j=1}^n \dot{x}_j^0(t) H_{xx_j}(t)^* + H_{xt}(t)^* \right] \rho(t) dt, & t \in [t_0, t_1 - r], \\ \int_t^{t_1} \left[\sum_{j=1}^n \dot{x}_j^0(s) H_{xx_j}(s)^* + H_{xt}(s)^* \right] \rho(s) ds + H_x(t)^* \rho(t), & t \in (t_1 - r, t_1]. \end{cases}$$

Similarly,

$$\begin{aligned} [l_0 D_1 G(x^0, v^0)](t) &= -l_0 \int_t^{t_1} \frac{\partial}{\partial x} g(x^0(s), v^0(s), s) ds, \quad t \in T, \\ [DB(x^0)^* l_1](t) &= \begin{cases} H_x(t_1 - r)^* l_1, & t \in [t_0, t_1 - r], \\ 0, & t \in (t_1 - r, t_1]. \end{cases} \end{aligned}$$

Then y^K can be identified with an element $\psi^K \in L_\infty^k(T)$ and

$$[D_1 F(x^0, v^0)^* y^K](t) = - \int_t^{t_1} \eta(s, t)^* \psi^K(s) ds, \quad t \in T.$$

This yields the adjoint equation (i). Furthermore,

$$y^K \circ F(x^0, v - v^0) = \int_T \psi^K(t) f(x_t^0, v(t) - v^0(t), t) dt,$$

and the maximum condition in integral form follows. This implies (ii) by standard arguments (see, e.g., Warga [48, Thm. VI.2.3]). \square

Remark 3.1. Since $\eta(s, \cdot)$ is of bounded variation, one can identify $\psi^K|_{[t_0, t_1-r]}$ with a function of bounded variation. For special systems, e.g., systems with a single constant delay, $\psi^K|_{[t_0, t_1-r]}$ is even absolutely continuous (see Banks/Kent [3, p. 583]).

Remark 3.2. One can give the following nontriviality condition in terms of the adjoint variable ψ^K :

$$(0, 0, 0) \neq (l_0, \psi^K(t_1-r), \psi^K|_{T_1}) \in \mathbb{R}_+ \times \mathbb{R}^n \times L_\infty^n(T_1).$$

Assume that $(l_0, \psi^K(t_1-r), \psi^K|_{T_1})$ is trivial. Regularity implies that the multiplication operator

$$\tilde{H}: L_\infty^n(T_1) \rightarrow L_\infty^k(T_1), \quad (\tilde{H}z)(t) := H_x(t)z(t) \quad (t \in T_1)$$

is surjective. Hence Kurcyusz/Olbrot [36, Lemmas 3 and 4] imply that

$$\text{rank } H_x(t)^* = \text{rank } H_x(t) = k$$

for all $t \in T_1$ and that the generalized inverse $[H_x(t)^*]^+$ of $H_x(t)^*$ is bounded on T_1 .

Using the adjoint equation (i), we find that ρ satisfies the homogeneous Volterra equation

$$\rho(t) = - \int_t^{t_1} [H_x(t)^*]^+ \left[\sum_{j=1}^n \dot{x}_j^0(s) H_{xx_j}(s)^* + H_{xt}(s)^* \right] \rho(s) ds \quad (t \in T_1).$$

By unique solvability it follows that $\rho(t) = 0$, $t \in T_1$. Then

$$0 = \psi^K(t_1-r) = -H_x(t_1-r)^* l_1$$

implies that $l_1 = 0$. This contradicts the nontriviality condition $(l_0, l_1) \neq (0, 0)$.

Remark 3.3. Consider the maximum principle for the case of a fixed final state. Then $l_0 = 0$ implies $\psi^K(t) = 0$ ($t \in T_1$). This follows from regularity, since the maximum condition has the form

$$\psi^K(t) f(x_t^0, v^0(t), t) \leq \psi^K(t) y$$

for all $y \in \text{co } f(x_t^0, \Omega(t), t)$.

Remark 3.4. If there exists $\varepsilon > 0$ such that

$$x^0(t) \in \text{int co } f(x_t^0, \Omega(t), t) \quad (t \in [t_1-r-\varepsilon, t_1]),$$

then $l_0 \neq 0$.

Suppose $l_0 = 0$. Then as in Remark 3.3 it follows that $\psi^K(t) = 0$ ($t \in [t_1-r-\varepsilon, t_1]$). This used in the adjoint equation (i) shows that $l_1 = 0$, contradicting the nontriviality condition. Observe that we do not require the existence of a *uniform* neighborhood of $\dot{x}^0(t)$ contained in $\text{co } f(x_t^0, \Omega(t), t)$ as it is required for stable reachability in Colonius [15, Theorem 2.2].

The following theorem gives a second form of the maximum principle proposed by Bien/Chyung [10]. Recall that the dual space of $\mathbf{C}^n(T)$ can be identified with $\mathbf{NBV}^n(T)$, the space of normalized functions of bounded variation on T with values in \mathbb{R}^n being right continuous in t_0 (cf. the definition (2.2) of η).

THEOREM 3.2 (BC-form of the maximum principle). *Let x^0, v^0 be an optimal solution of Problem 1 and assume that x^0 is regular.*

Then there exist Lagrange multipliers $(l_0, l_1, \rho) \in \mathbb{R}_+ \times \mathbb{R}^k \times L_\infty^k(T)$ such that $(l_0, l_1) \neq (0, 0)$, $\rho|_{[t_0, t_1-r]} = 0$, and the adjoint variable $\psi^C \in \mathbf{NBV}^n(T)$ defined by

$$\begin{aligned} \text{(i)} \quad \psi^C(t) = & - \int_t^{t_1} \eta(s, t)^* (\psi^C(s) + H_x(s)^* \rho(s)) ds \\ & - l_0 \int_t^{t_1} \frac{\partial}{\partial x} g(x^0(s), v^0(s), s) ds \\ & - \int_t^{t_1} \left[\sum_{j=1}^n f_j(x_s^0, v^0(s), s) H_{xx_j}(s)^* + H_{xt}(s)^* \right] \rho(s) ds \\ & - \begin{cases} H_x(t_1-r)^* l_1 & t \in [t_0, t_1-r], \\ 0, & t \in (t_1-r, t_1], \end{cases} \end{aligned}$$

is right continuous in t_0 and satisfies

$$\begin{aligned} \text{(ii)} \quad & -l_0 g(x^0(t), v^0(t), t) + (\psi^C(t) - H_x(t)^* \rho(t)) f(x_t^0, v^0(t), t) \\ & \geq -l_0 g(x^0(t), \omega, t) + (\psi^C(t) - H_x(t)^* \rho(t)) f(x_t^0, \omega, t) \end{aligned}$$

for all $\omega \in \Omega(t)$, a.a. $t \in T$.

Proof. Theorem 3.2 follows from Corollary 2.1(ii) in the same way as Theorem 3.1 follows from Corollary 2.1(i). Observe that $D_1 C(x^0, v^0)l_2$ can be identified with an element of $\mathbf{NBV}^n(T)$:

$$\begin{aligned} [D_1 C(x^0, v^0)^* l_2](t) = & - \int_t^{t_1} \eta(t, s)^* H_x(s)^* \rho(s) ds \\ & - \int_t^{t_1} \left[\sum_{j=1}^n f_j(x_s^0, v^0(s), s) H_{xx_j}(s)^* + H_{xt}(s)^* \right] \rho(s) ds. \end{aligned}$$

y^C can be identified with $\psi^C \in \mathbf{NBV}^n(T)$. Furthermore,

$$\begin{aligned} y^C \circ F(x^0, v - v^0) &= \int_T \int_{t_0}^t f(x_s^0, v(s) - v^0(s), s) ds d\psi^C(t) \\ &= - \int_T f(x_t^0, v(t) - v^0(t), t) \psi^C(t) dt \end{aligned}$$

and

$$l_2 \circ C(x^0, v) = \int_T \rho(t) H_x(t) f(x_t^0, v(t) - v^0(t), t) dt.$$

This yields the maximum condition (ii). \square

Remark 3.5. On subintervals of T , the adjoint equation and the maximum condition have a simpler form.

Since $\rho(t) = 0$ on $[t_0, t_1-r]$, the maximum condition does not involve ρ on this interval. Furthermore,

$$\int_t^{t_1} \eta(s, t)^* H_x(s)^* \rho(s) ds = \int_{T_1} \eta(s, t)^* H_x(s)^* \rho(s) ds$$

for $t \in [t_0, t_1-r]$ and for $t \in [t_0, t_1-2r]$ and $s \in [t_1-r, t_1]$,

$$\eta(s, t) = \eta(s, s-r);$$

this follows from the definition (2.2) of η . In special cases, e.g., systems with a single

constant delay, the adjoint variable ψ^C is even absolutely continuous on T with the possible exception of a jump at $t_1 - r$.

Remark 3.6. Again the nontriviality condition can be written in terms of the adjoint variable

$$(0, 0) \neq (l_0, \psi^C|_{T_1}) \in \mathbb{R}_+ \times \mathbf{NBV}^n(T_1).$$

Suppose that $(l_0, \psi^C|_{T_1})$ is trivial. Then $\rho(t) = 0$ ($t \in T_1$), since regularity is assumed and the maximum condition on T_1 has the form

$$\rho(t)H_x(t)f(x_t^0, v^0(t), t) \leq \rho(t)y$$

for all $y \in H_x(t) \text{ co } f(x_t^0, \Omega(t), t)$. Hence,

$$0 = \psi^C(t_1 - r) = -H_x(t_1 - r)^* l_1.$$

Since by regularity, $\text{rank } H_x(t_1 - r)^* = k$, it follows that $l_1 = 0$ contradicting the condition $(l_0, l_1) \neq (0, 0)$.

Remark 3.7. Compare the two forms of the maximum principle: The adjoint variables ψ^K and ψ^C are related by an integral equation which in its abstract form is given in Corollary 2.1(iii).

The BK-form appears natural because the end condition is originally a pure phase equality constraint. Hence, the Lagrange multiplier corresponding to the end condition should appear only in the adjoint equation, not in the maximum condition as is the case in the BC-form.

Remark 3.8. Bien/Chyung [10] require, instead of (0.6), that $h(x(t_1), t_1) = 0$. Then the adjoint variable has a corresponding jump in t_1 , while (0.6) induces a jump of ψ^C in $t_1 - r$. Bien and Chyung redefine their adjoint variable in t_1 such that the Lagrange multiplier corresponding to the finite dimensional part of the end condition does not appear explicitly in the adjoint equation (see Bien/Chyung [10, Thm. 3.1(ii) and (iii)]). Apart from this minor variation, their adjoint equation coincides with the adjoint equation above specialized to the case of a single constant delay. However, the maximum condition given above has a global form on the whole interval T , not only on $[t_0, t_1 - r]$ as in Bien/Chyung [10].

Remark 3.9. Sufficiency of the maximum principle has been analyzed by Banks/Kent [3] and Bien/Chyung [10]. Restricting their analysis to the fixed final state problem, Banks and Kent establish sufficiency of the BK-form under the usual normality and convexity assumptions. Bien and Chyung show sufficiency of the BC-form under similar assumptions for general function space and conditions. Taking into account the equivalence of the two forms, their results extend those of Banks and Kent.

Remark 3.10. We have treated the phase equality constraint (0.4) by using its equivalent formulation as a mixed control phase variable constraint. Now it is clear from the proof of Theorem 3.2 that we can deal with *any* constraint of the form

$$b(x, v(t), t) = 0 \quad (t \in T_1)$$

where $b: C^n[-r, 0] \times \mathbb{R}^m \times T_1 \rightarrow \mathbb{R}^k$ satisfies the same assumptions as f in (1.1)–(1.3). In particular, for $r = 0$ the results of Makowski/Neustadt [37] are extended to relaxed—instead of ordinary—optimal solutions of problems with ordinary differential equations. The maximum principle obtained in this way is equivalent to that in Schwarzkopf [43] (where $r = 0$, $T_1 = T$). An advantage of our approach is that we can make use of results in general optimization theory (cf. Theorem 2.1)). Furthermore,

a relation between regularity and structural properties of the system is obtained (Lemma 3.1).

Remark 3.11. The regularization procedure performed in the proof of Lemma 3.2 can easily be extended in order to deal with inequality constraints of the form

$$b(x, v(t), t) \leq 0 \quad (t \in T_1),$$

where b is as in the remark above.

Suppose that there exist Lagrange multipliers $(l_0, l_1, l_2) \in \mathbb{R}_+ \times \mathbb{R}^k \times (L_\infty^k(T_1))^*$ satisfying the analogue of (2.11) and additionally

$$\langle l_2, z \rangle \geq 0 \quad \text{for all } z \in L_\infty^k(T_1) \quad \text{with } z \geq 0.$$

Then we impose the following regularity condition: There exists a neighborhood V of $0 \in \mathbb{R}^k$ such that

$$V \cap \mathbb{R}_-^k \subset \text{co } b(x_t^0, \Omega(t), t) \cap \mathbb{R}_-^k \quad (t \in T_1),$$

where $\mathbb{R}_-^k := \{x = (x_1, \dots, x_k) \in \mathbb{R}^k : x_i \leq 0\}$.

In order to prove that l_2 can be identified with an element of $L_\infty^k(T_1)$, consider first nonpositive simpler functions s . Here everything goes through due to the above regularity assumption. Now suppose s is a *general* simple function. Decompose s into its positive and negative parts:

$$s = s^+ - s^-, \quad \text{where } s^+, s^- \geq 0.$$

Due to the positivity of l_2 , we find

$$\langle l_2, s \rangle = \langle l_2, s^+ \rangle + \langle l_2, -s^- \rangle \geq \langle l_2, -s^- \rangle$$

and

$$\langle l_2, -s \rangle \geq \langle l_2, -s^+ \rangle.$$

Since $\langle l_2, -s^- \rangle$ and $\langle l_2, -s^+ \rangle$ converge to 0 for $\|s\|_{L_1} \rightarrow 0$, also $\langle l_2, s \rangle$ converges to 0. Hence, l_2 can be identified with an element of $L_\infty^k(T_1)$.

Remark 3.12. The maximum principle reduces the optimal control problem (0.1)–(0.5) to an “infinite defect boundary value problem” (as defined by Kamenskii/Myshkis [31]) consisting of a retarded and an advanced equation coupled by the maximum condition. Solution of such a system is very difficult (cf. also Grimm/Schmitt [21], Kamenskii [27], Hutson [24], Kamenskii/Kamenskii/Myshkis [30]).

However, the existence of Lagrange Multipliers is also important in order to prove the convergence of computational procedures (cf. Wierzbicki/Hatko [49], Wierzbicki/Kurcyusz [50], who use shifted penalty methods in order to compute solutions of problems with function space end condition, and Williamson/Polak [51]).

Remark 3.13. Conversely, one obtains results on infinite defect boundary value problems by considering a corresponding optimal control problem. Existence of an optimal solution and validity of the maximum principle imply that a certain boundary value problem has a solution (see Kamenskii [29]).

Remark 3.14. Colonius [16] gave an example of an optimal control problem where the assumptions of Theorem 2.1 are met while the optimal trajectory is not regular. It is shown that the maximum principle is *not* satisfied (for a certain performance index), i.e., there are no nontrivial Lagrange multipliers $(l_0, l) \in \mathbb{R}_+ \times W^{n,\infty}[-r, 0]$. This shows that the regularity assumption is crucial for the validity of the maximum principle.

4. Discussion. The proof of the maximum principle relies on two main assumptions: (i) the existence of Lagrange multipliers $(l_0, l_1) \in \mathbb{R}_+ \times (W^{k,\infty}(T_1))^*$ can be established if $\pi_\infty \mathcal{A}$ has a nonempty interior in $L_\infty^k(T_1)$, and (ii) l can be identified with an element of $W^{k,\infty}(T_1)$ if the optimal trajectory is regular. We shall discuss, for the fixed final state problem, how restrictive these assumptions are. For simplicity, we assume that $\Omega(t) = \Omega_0$ on T and that the function f defining the right-hand side of the system equation is independent of t and an element of the following Banach space \mathcal{F} :

$$\mathcal{F} := \left\{ f: C^n[-r, 0] \times \mathbb{R}^m \rightarrow \mathbb{R}^n: \|f\| \right. \\ \left. := \max \left\{ \sup_{\varphi \in C^n, \omega \in \mathbb{R}^m} |f(\varphi, \omega)|, \sup_{\varphi \in C^n, \omega \in \mathbb{R}^m} \|D_1 f(\varphi, \omega)\| \right\} < \infty \right\}.$$

We have for $f \in \mathcal{F}$:

$$\pi_\infty \mathcal{A} = \{z \in L_\infty^n[-r, 0]: \text{there is } v \in \mathcal{S}^\# \text{ such that } z = (\dot{x})_{t_1} \text{ for } x \text{ satisfying } x_{t_0} = 0 \\ \text{and } \dot{x}(t) = D_1 f(x_t^0, v^0(t))x_t + f(x_t^0, v(t) - v^0(t))(t \in T)\}.$$

The linearized system satisfies the assumptions in Colonius [15, Corollary 4.1]. Hence, $\text{int } \pi_\infty \mathcal{A} \neq \emptyset$ if and only if

$$(4.1) \quad \text{int } \{z \in L_\infty^n(T_1): z(t) \in \text{co } f(x_t^0, \Omega_0)(t \in T_1)\} \neq \emptyset.$$

By the same arguments as in [15, Remark 4.4] one can see that the set of elements $f \in \mathcal{F}$ satisfying condition (4.1) for all $x^0 \in C^n[t_0 - r, t_1]$ is open and dense in \mathcal{F} provided that Ω_0 contains at least $n + 1$ points. In this sense the condition $\text{int } \pi_\infty \mathcal{A} \neq \emptyset$ is generically satisfied for hereditary differential systems defined by $f \in \mathcal{F}$.

Linear systems Λ (see Colonius [15, § 2]) are not included in the class of systems defined by \mathcal{F} . However, on the basis of [15, Remark 4.4], a similar genericity statement can easily be proven.

Observe, however, that the situation is quite different if we restrict ourselves to the class of functions f where $\omega \in \mathbb{R}^m$ appears *affinely*. Then the condition $n \leq m$ on the dimensions of the state space (=output space) and the control space is necessary for $\text{int } \pi_\infty \mathcal{A} \neq \emptyset$.

Now consider the second assumption concerning regularity. Colonius [15, Examples 4.1, 4.2] specifies classes of systems, where *all* trajectories reaching a certain final function are regular. In general the situation is much more complicated, and we have to look at the linearized system.

Colonius [15, Prop. 3.1] states that a trajectory x^0 is regular if and only if the zero trajectory of the corresponding linearized system is regular. Since $0 \in \mathcal{A}$, we have either that $0 \in \partial \mathcal{A}$ or $0 \in \text{int } \mathcal{A}$. The first situation is a degenerate one: if $0 \in \partial \mathcal{A}$ and $\text{int } \mathcal{A} \neq \emptyset$ there are Lagrange multipliers $(0, l) \in \mathbb{R}_+ \times (W^{n,\infty}[-r, 0])^*$ and the optimality condition (2.11) is independent of the performance index (see Remark 2.3 for the case where $\text{int } \mathcal{A} = \emptyset$).

Now suppose $0 \in \text{int } \mathcal{A}$. Then by [15, Corollaries 4.1, 4.2], 0 is regularly reachable, and by [15, Thm. 4.2], the set of regular trajectories is open and dense in the set of all trajectories reaching 0 . Thus, irregularity, in particular, irregularity of the zero trajectory, is “exceptional” (compare, e.g., the discussion by Maurin [38, p. 29]).

In this sense, the maximum principle is generically valid and its use as a necessary optimality condition appears to be justified.

Acknowledgment. I thank Diederich Hinrichsen for many fruitful discussions. This paper originated from my doctoral thesis which was guided by him.

REFERENCES

- [1] H. T. BANKS, *Control of functional differential equations with function space boundary conditions*, Delay and Functional Differential Equations and their Applications, K. Schmitt, ed., Academic Press, New York, 1972, pp. 1–16.
- [2] H. T. BANKS AND M. Q. JACOBS, *An attainable sets approach to optimal control of functional differential equations with function space boundary conditions*, J. Differential Equations, 13 (1973), pp. 127–149.
- [3] H. T. BANKS AND G. A. KENT, *Control of functional differential equations of retarded and neutral type with target sets in function space*, SIAM J. Control, 10 (1972), pp. 567–593.
- [4] H. T. BANKS AND A. MANITIUS, *Application of abstract variational theory to hereditary systems—A survey*, IEEE Trans. Automat. Control, AC-19 (1974), pp. 524–533.
- [5] V. BARBU, *Convex control problems for linear differential systems of retarded type*, Ricerche Math., XXVI (1977), pp. 3–26.
- [6] V. BARBU AND TH. PRECUPANU, *Convexity and Optimization in Banach Spaces*, Editura Academiei, Bucuresti/Sijthoff and Noordhoff, Groningen, 1978.
- [7] G. R. BATES, *Hereditary optimal control problems*, Ph.D. thesis, Purdue University, W. Lafayette, IN, 1977.
- [8] L. D. BERKOVITZ, *A penalty function proof of the maximum principle*, Appl. Math. Optim., 2 (1975/76), pp. 291–303.
- [9] Z. BIEN, *Optimal control of delay systems*, Ph.D. thesis, University of Iowa, Ames, Iowa, 1975.
- [10] Z. BIEN AND D. H. CHYUNG, *Optimal control of delay systems with a final function condition*, Internat. J. Control, 32 (1980), pp. 539–560.
- [11] N. BOURBAKI, *Intégration*, Livre V, Chap. V (2^{me} éd.), Hermann, Paris, 1968.
- [12] H. H. BUEHLER, *Application of Neustadt's theory of extremals to an optimal control problem with a functional differential equation and a functional inequality constraint*, Appl. Math. Optim., 2 (1975/76), pp. 34–74.
- [13] F. COLONIUS, *Necessary optimality conditions for nonlinear hereditary differential systems with function space end constraints*, Functional Differential Systems and Related Topics, M. Kisielewicz, ed., The Higher College of Engineering in Zielona Gora, Zielona Gora, Poland, 1980, pp. 62–71.
- [14] ———, *Regularization of Lagrange multipliers for time delay systems with fixed final state*, in Optimization and Optimal Control, A. Auslender, W. Oettli and J. Stoer, eds., Springer-Verlag, Berlin–Heidelberg–New York, 1981, pp. 163–177.
- [15] ———, *Stable and regular reachability for relaxed hereditary differential systems*, this Journal, this issue, pp. 675–694.
- [16] ———, *A penalty function proof of a Lagrange multiplier theorem with application to linear delay systems*, Appl. Math. Optim., 7 (1981), pp. 309–334.
- [17] F. COLONIUS AND D. HINRICHSSEN, *Optimal control of functional differential systems*, this Journal, 16 (1978), pp. 861–879.
- [18] ———, *Optimal control of hereditary differential systems*, in Recent Theoretical Developments in Control, M. J. Gregson, ed., Academic Press, London, 1978, pp. 215–240.
- [19] P. C. DAS, *Application of Dubovitskii/Milyutin formalism to optimal settling problems with constraints*, Optimization and Optimal Control (Oberwolfach 1974), Springer-Verlag, Berlin–Heidelberg–New York, 1975.
- [20] I. V. GIRSANOV, *Lectures on Mathematical Theory of Extremum Problems*, Springer-Verlag, Berlin–Heidelberg–New York, 1967.
- [21] L. J. GRIMM AND K. SCHMITT, *Boundary value problems for differential equations with deviating arguments*, Aequationes Math., 4 (1970), pp. 176–190.
- [22] J. HALE, *Theory of Functional Differential Equations*, second ed., Springer-Verlag, Berlin–Heidelberg–New York, 1977.
- [23] E. HEWITT AND K. STROMBERG, *Real and Abstract Analysis*, Springer-Verlag, Berlin–Heidelberg–New York, 1969.
- [24] V. HUTSON, *A note on a boundary value problem for linear differential difference equations of mixed type*, J. Math. Anal. Appl., 61 (1977), pp. 416–425.
- [25] M. Q. JACOBS, *An optimization problem for an n^{th} order scalar neutral functional differential equation with functional side conditions*, in Delay and Functional Differential Equations and Their Applications, K. Schmitt, ed., Academic Press, New York, 1972, pp. 345–372.
- [26] M. Q. JACOBS AND T. J. KAO, *An optimum settling problem for time lag systems*, J. Math. Anal. Appl., 40 (1972), pp. 687–707.
- [27] G. A. KAMENSKII, *Variational and boundary value problems with deviating argument*, Differential Equations, 6 (1970), pp. 1026–1032.

- [28] ———, *On a conditional extremum of a functional with deviating argument*, Soviet Math. Dokl., 18 (1977), pp. 921–924.
- [29] ———, *A variational method for solving boundary value problems for certain linear differential equations with deviating arguments*, Differential Equations, 13 (1977), pp. 820–825.
- [30] A. G. KAMENSKII, G. A. KAMENSKII AND A. D. MYSHKIS, *On the convergence of the finite difference method of numerically solving boundary value problems for linear differential-difference equations*, Soviet Math. Dokl., 18 (1977), pp. 321–324.
- [31] G. A. KAMENSKII AND A. D. MYSHKIS, *Boundary value problems with infinite defect*, Differential Equations, 7 (1971), pp. 1612–1618.
- [32] G. A. KAMENSKII AND A. L. SKUBACHEVSKII, *Extremals of functionals with deviating argument*, Moskovskii Aviazionnii Institut, Moscow, 1979 (in Russian).
- [33] G. A. KENT, *Optimal control of functional differential equations of neutral type*, Doctoral thesis, Brown Univ., Providence, RI, 1971.
- [34] ———, *A maximum principle for optimal control problems with neutral functional differential systems*, Bull. Amer. Math. Soc., 77 (1971), pp. 565–570.
- [35] S. KURCYUSZ, *A local maximum principle for operator constraints and its application to systems with time lag*, Control and Cyber., 2 (1973), pp. 99–125.
- [36] S. KURCYUSZ AND A. W. OLBROT, *On the closure in $W^{1,q}$ of the attainable subspace of linear time lag systems*, J. Differential Equations, 24 (1977), pp. 29–50.
- [37] K. MAKOWSKI AND L. W. NEUSTADT, *Optimal control problems with mixed control phase variable equality and inequality constraints*, SIAM J. Control, 12 (1974), pp. 184–228.
- [38] K. MAURIN, *Analysis, Part II*, PWN-Polish Scientific Publishers, Warszawa, and D. Reidel Publishing Company, London, 1980.
- [39] A. W. OLBROT, *Control of retarded systems with function space constraints: necessary optimality conditions*, Control and Cyber., 5 (1976), pp. 5–31.
- [40] R. R. PHELPS, *Weak* support points of convex sets in E^** , Israel J. Math., 2 (1964), pp. 177–182.
- [41] L. S. PONTRYAGIN, V. G. BOLTJANSKII, R. V. GAMKRELIDZE AND E. E. MISCHENKO, *The Mathematical Theory of Optimal Control Processes*, John Wiley, New York, 1962.
- [42] A. B. SCHWARZKOPF, *Optimal controls for problems with a restricted state space*, SIAM J. Control, 10 (1972), pp. 487–511.
- [43] ———, *Relaxed control problems with state equality constraints*, SIAM J. Control, 13 (1975), pp. 677–694.
- [44] ———, *Optimal controls with equality state constraints*, J. Opt. Theory Appl., 19 (1976), pp. 455–468.
- [45] J. UTTHOFF, *Optimale Kontrolle neutraler Funktional-differentialgleichungen*, Diplomarbeit, Freie Universität Berlin, Berlin, 1979.
- [46] J. WARGA, *Minimizing variational curves restricted to a preassigned set*, Trans. Amer. Math. Soc., 12 (1964), pp. 449–455.
- [47] ———, *Unilateral variational problems defined by integral equations*, Michigan Math. J., 12 (1965), pp. 449–480.
- [48] ———, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
- [49] A. P. WIERZBICKI AND A. HATKO, *Computational methods in Hilbert space for optimal control problems with delays*, 5th IFIP Conference on optimization Techniques, R. Conti, A. Ruberti, eds., Springer-Verlag, Berlin-Heidelberg-New York, 1973.
- [50] A. P. WIERZBICKI AND S. KURCYUSZ, *Projection on a cone, penalty functionals and duality theory for problems with inequality constraints in a Hilbert space*, this Journal, 15 (1977), pp. 25–56.
- [51] L. J. WILLIAMSON AND E. POLAK, *Relaxed controls and the convergence of optimal control algorithms*, this Journal, 14 (1976), pp. 737–756.
- [52] L. C. YOUNG, *Lectures on the Calculus of Variations and Optimal Control Theory*, Saunders, Philadelphia, 1969.
- [53] J. ZOWE AND S. KURCYUSZ, *Regularity and stability for the mathematical programming problem in Banach spaces*, Appl. Math. Opt., 5 (1979), pp. 49–62.
- [54] A. M. ZVERKIN, G. A. KAMENSKII, S. B. NORKIN AND L. E. ELSGOLTS, *Differential equations with a perturbed argument*, Russian Math. Surveys, 17 (1962), pp. 61–146.

OPTIMAL INPUT DESIGN FOR DYNAMIC SYSTEM PARAMETER ESTIMATION: THE D_s -OPTIMALITY CASE*

Z. H. QURESHI AND T. S. NG†

Abstract. The problem of optimal input design for estimating part of the parameters of a dynamic system is considered. We show that two design criteria, G_s -optimality and D_s -optimality are equivalent. Furthermore, a proven sequential design algorithm is extended to cover this case.

1. Introduction. The problem of optimal input design is fundamental to many identification experiments and has been investigated by many authors. Mehra [1] provided a comprehensive survey of the literature on optimal input synthesis. In the same paper, he extended the Kiefer–Wolfowitz equivalence theorem [2] to dynamic systems and proposed a sequential design procedure based on that of Fedorov [3]. An extensive study on the subject by Zarrop [4] showed that certain sequential design procedures, which incorporate a number of algorithms previously proposed by a number of authors [1], [3], [5], converge globally to a D -optimal design. Optimal input design with constraints or in the presence of feedback has also been examined quite extensively; see, for example, Goodwin and Payne [6], Goodwin, Murdoch and Payne [7], Ng, Goodwin and Söderström [8], Söderström [9] and others.

It is evident that in designing identification experiments for different purposes, different design criteria should be used. For example, if the design is to minimize the covariance of the parameter estimator, a scalar function of the parameter covariance matrix (or the inverse of the information matrix M^{-1} , see (2.1) for a definition) is a good choice. On the other hand, if the purpose of an identification experiment is to accurately predict the output sequence, then it is more reasonable to cost the output variance. A number of input design criteria are suggested and discussed in Mehra [1]. For our purpose, we consider only two of them:

- (a) A design ξ^\dagger is called D -optimal if it maximizes $\det M(\xi)$.
- (b) A design ξ^\dagger is called G -optimal if it minimizes $\max_\omega d(\omega, \xi)$.

For a definition of $d(\omega, \xi)$ see (2.3).

Excluding trivial cases, it is obvious that an optimal design which satisfies all criteria does not exist. However, the two criteria D -optimality and G -optimality, which compare the results of an experiment in two different spaces, are strictly related to each other [1]–[5]. The equivalence of these two design criteria has proved to be extremely important. It is fundamental to most D -optimal design algorithms and many computational improvements [1]–[5] in both linear regression problems and dynamic system design problems.

For dynamic systems, the studies so far have concentrated on estimating all system parameters. The problem of estimating part of the system parameters has not been addressed. In this paper, the optimal input design problem for estimating S out of p ($S \leq p$) parameters, where p is the total number of system parameters, will be considered. We shall prove the equivalence theorem [1]–[5], [10] and extend a sequential design algorithm [1], [4] to the D_s -optimality case.

The paper is organized as follows: in § 2, some definitions and terminology are established. In § 3, we extend the proof of the equivalence of D_s -optimality and

* Received by the editors December 8, 1980, and in revised form September 15, 1981.

† Department of Electrical Engineering, The University of Wollongong, Wollongong, N.S.W., Australia.

G_s -optimality and in § 4, the convergence of a sequential design algorithm is presented and illustrated by two simple examples.

2. Problem statement. We now introduce our terminology and notation, which is essentially that of [4]. Consider a linear, time invariant, single-input-single-output discrete time dynamic system:

$$y_k = \frac{B(z^{-1})}{A(z^{-1})} u_{k-d} + \frac{D(z^{-1})}{C(z^{-1})} e_k,$$

where

$$\begin{aligned} A(z^{-1}) &= 1 + a_1 z^{-1} + \cdots + a_n z^{-n}, \\ B(z^{-1}) &= b_0 + b_1 z^{-1} + \cdots + b_m z^{-m}, \\ C(z^{-1}) &= 1 + c_1 z^{-1} + \cdots + c_q z^{-q}, \\ D(z^{-1}) &= d_0 + d_1 z^{-1} + \cdots + d_r z^{-r}, \end{aligned}$$

e_k is Gaussian i.i.d. with unit variance.

Following the derivation of [4] and [8], the average information matrix for system parameters a_1, \dots, a_n and b_0, \dots, b_m is given by:

$$(2.1) \quad M(\xi) = \int_{-\pi}^{\pi} h(e^{j\omega}) h^*(e^{j\omega}) d\xi'(\omega) = \text{Re} \int_0^{\pi} h(e^{j\omega}) h^*(e^{j\omega}) d\xi(\omega),$$

where $h^T = [h_1, h_2, \dots, h_p]$ and

$$\begin{aligned} h_i(z^{-1}) &= \begin{cases} \frac{B(z^{-1})C(z^{-1})}{D(z^{-1})A^2(z^{-1})} \cdot z^{-(d+i)}, & i = 1, \dots, n, \\ -\frac{C(z^{-1})}{D(z^{-1})A(z^{-1})} \cdot z^{-(d+i-n-1)}, & i = n+1, \dots, p, \end{cases} \\ p &= n + m + 1, \\ \xi(\omega) &= \begin{cases} 2\xi'(\omega), & \omega \in (0, \pi), \\ \xi'(\omega), & \omega = 0, \pi, \end{cases} \end{aligned}$$

and h^* is the complex conjugate transpose of h . Any such ξ is called a normalized design and the set of all ξ will be denoted by Ξ , i.e., $\Xi = \{\xi / \int_0^{\pi} d\xi = 1\}$. The information matrix for parameters c_1, \dots, c_q and d_0, \dots, d_r is independent of any design ξ (Goodwin and Payne [11]) and hence will not be considered any further here. In what follows, we also assume that $M(\xi)$ is nonsingular; i.e., that the design ξ is persistently exciting [11].

Since we are interested in estimating part of the system parameters, we partition the information matrix and its inverse into

$$M(\xi) = \begin{pmatrix} M_1(\xi) & M_2(\xi) \\ M_2^T(\xi) & M_3(\xi) \end{pmatrix} \quad \text{and} \quad M^{-1}(\xi) = \begin{pmatrix} M^{(1)}(\xi) & M^{(2)}(\xi) \\ M^{(2)T}(\xi) & M^{(3)}(\xi) \end{pmatrix},$$

where $M_1(\xi)$ and $M^{(1)}(\xi)$ are $S \times S$ matrices and $M_3(\xi)$ and $M^{(3)}(\xi)$ are $(p-s) \times (p-s)$ matrices.

Now define 1)

$$(2.2) \quad \bar{M}(\xi) = [M^{(1)}(\xi)]^{-1} = M_1(\xi) - M_2(\xi)M_3^{-1}(\xi)M_2^T(\xi)$$

and it follows that

$$(2.2a) \quad \det \bar{M}(\xi) = \det [M_1(\xi) - M_2(\xi)M_3^{-1}(\xi)M_2^T(\xi)] = \frac{\det M(\xi)}{\det M_3(\xi)}.$$

2) ξ^\dagger yields a global D_s -optimal design if

$$\det \bar{M}(\xi^\dagger) = \max_{\xi} \det \bar{M}(\xi).$$

3) ξ^\dagger yields a local D_s -optimal design if

$$\frac{\partial}{\partial \alpha} \log \det \bar{M}[(1-\alpha)\xi^\dagger + \alpha\xi] \big|_{\alpha=0} \leq 0 \quad \text{for all } \xi.$$

4) The generalized variance $d(\omega, \xi)$, which in physical terms, can be interpreted as the ratio of the variance of the system frequency response to the noise power at frequency ω [11], in the case $S \leq p$ to be

$$(2.3) \quad d_s(\omega, \xi) = h^*(e^{j\omega})M^{-1}(\xi)h(e^{j\omega}) - h^{(2)*}(e^{j\omega})M_3^{-1}(\xi)h^{(2)}(e^{j\omega}),$$

where

$$\begin{aligned} h^T(e^{j\omega}) &= [h_1(e^{j\omega}) \cdots h_s(e^{j\omega}) \mid h_{s+1}(e^{j\omega}) \cdots h_p(e^{j\omega})] \\ &= [h^{(1)T}(e^{j\omega}) \mid h^{(2)T}(e^{j\omega})]. \end{aligned}$$

We now proceed to prove the equivalence theorem for the case where S out of p parameters are estimated.

3. Equivalence theorem. Before we proceed to state the main result, we first show that the local minimum of optimizing $\det \bar{M}(\xi)$ is also the global minimum.

THEOREM 3.1. Consider $\xi^0 \in \Xi$ and let $\xi = (1-\alpha)\xi^\dagger + \alpha\xi^0 \in \Xi$, $\alpha \in [0, 1]$. Then

- 1) ξ^\dagger maximizes $\det \bar{M}(\xi) \forall \xi \in \Xi$,
- 2) $\partial/\partial \alpha \log \det \bar{M}[(1-\alpha)\xi^\dagger + \alpha\xi^0] \leq 0$

are equivalent.

Proof. Clearly (1) \rightarrow (2). To show that (2) \rightarrow (1), we first state two matrix lemmas without proof.

LEMMA 3.1 (Kiefer [12]). For $i = 1, 2, \dots, r$, let C_i be $S \times (p-S)$, let D_i be positive definite symmetric $S \times S$ and suppose $\alpha_i > 0$, $\sum_{i=1}^r \alpha_i = 1$. Then

$$\left[\sum_{i=1}^r \alpha_i C_i \right] \left[\sum_{i=1}^r \alpha_i D_i \right]^{-1} \left[\sum_{i=1}^r \alpha_i C_i^T \right] \leq \sum_{i=1}^r \alpha_i C_i D_i^{-1} C_i^T,$$

with equality if and only if the matrix C_i, D_i^{-1} is the same for all i .

LEMMA 3.2 (Fedorov [3]). If A and B are nonnegative symmetric definite $S \times S$ matrices, then $\log \det (\alpha A + (1-\alpha)B)$ is concave for $\alpha \in [0, 1]$ and is strictly concave unless $A = B$ or $A + B$ is singular.

Now assume that ξ^\dagger does not maximize $\det \bar{M}(\xi) \forall \xi \in \Xi$. Then there exists $\xi^0 \in \Xi$ such that

$$\log \det \bar{M}(\xi^0) - \log \det \bar{M}(\xi^\dagger) \geq 0.$$

Let $\xi = (1 - \alpha)\xi^\dagger + \alpha\xi^0 \in \Xi$. Then

$$\begin{aligned}\bar{M}(\xi) &= [M^{(1)}(\xi)]^{-1} \\ &= M_1(\xi) - M_2(\xi)M_3^{-1}(\xi)M_2^T(\xi) \\ &= [(1 - \alpha)M_1(\xi^\dagger) + \alpha M_1(\xi^0)] \\ &\quad - [(1 - \alpha)M_2(\xi^\dagger) + \alpha M_2(\xi^0)] \cdot [(1 - \alpha)M_3^{-1}(\xi^\dagger) + \alpha M_3^T(\xi^0)]^{-1} \\ &\quad \cdot [(1 - \alpha)M_2(\xi^\dagger) + \alpha M_2(\xi^0)]^T.\end{aligned}$$

Applying Lemma 3.1 gives

$$\begin{aligned}\bar{M}(\xi) &\geq (1 - \alpha)M_1(\xi^\dagger) + \alpha M_1(\xi^0) - (1 - \alpha)M_2(\xi^\dagger)M_3^{-1}(\xi^\dagger)M_2^T(\xi^\dagger) \\ &\quad - \alpha M_2(\xi^0)M_3^{-1}(\xi^0)M_2^T(\xi^0) \\ &= (1 - \alpha)\bar{M}(\xi^\dagger) + \alpha\bar{M}(\xi^0).\end{aligned}$$

Hence, using Lemma 3.2, we get

$$\begin{aligned}\log \det \bar{M}(\xi) &\geq \log \det [(1 - \alpha)\bar{M}(\xi^\dagger) + \alpha\bar{M}(\xi^0)] \\ &\geq (1 - \alpha) \log \det \bar{M}(\xi^\dagger) + \alpha \log \det \bar{M}(\xi^0), \\ \frac{\partial}{\partial \alpha} \log \det \bar{M}(\xi)|_{\alpha=0} &\geq \log \det \bar{M}(\xi^0) - \log \det \bar{M}(\xi^\dagger) \geq 0\end{aligned}$$

which violates 2). This completes the proof of Theorem 3.1.

We now extend the equivalence theorem to the case of estimating S out of p parameters.

THEOREM 3.2 (equivalence theorem, $S \leq p$ case). *The following statements are equivalent:*

- 1) ξ^\dagger maximizes $\det \bar{M}(\xi)$,
- 2) ξ^\dagger minimizes $\max_\omega d_s(\omega, \xi)$,
- 3) $\max_\omega d_s(\omega, \xi) = S$,

where $d_s(\omega, \xi)$ is defined by (2.3).

Proof. We proceed (3) \rightarrow (2), (1) \rightarrow (3), (3) \rightarrow (1) and (2) \rightarrow (3).

By definition of $d_s(\omega, \xi)$, it follows that

$$\begin{aligned}\int_0^\pi d_s(\omega, \xi) d\xi(\omega) &= \int_0^\pi [h^*(e^{j\omega})M^{-1}(\xi)h(e^{j\omega}) - h^{*(2)}(e^{j\omega})M_3^{-1}(\xi)h^{(2)}(e^{j\omega})] \cdot d\xi(\omega) \\ &= t_r \left[M^{-1}(\xi) \operatorname{Re} \int_0^\pi h(e^{j\omega})h^*(e^{j\omega}) d\xi(\omega) \right] \\ &\quad - t_r \left[M_3^{-1}(\xi) \operatorname{Re} \int_0^\pi h^{(2)}(e^{j\omega})h^{*(2)}(e^{j\omega}) d\xi(\omega) \right] \\ &= t_r[I_p] - t_r[I_{(p-s)}] = S.\end{aligned}$$

Thus

$$(3.1) \quad \max_\omega d_s(\omega, \xi) \geq S.$$

Now consider

$$(3.2) \quad \max_\omega d_s(\omega, \xi^\dagger) = \min_\xi \max_\omega d_s(\omega, \xi).$$

It follows from (3.1) that a sufficient condition for ξ^\dagger to satisfy (3.2) is that $\max_{\omega} d_s(\omega, \xi^\dagger) = S$. Thus (3) \rightarrow (2).

To show (1) \rightarrow (3), let $\xi^0 \in \Xi$ be any design and consider the design $\xi = (1 - \alpha)\xi^\dagger + \alpha\xi^0 \in \Xi$. Using Theorem 3.1 and (2.2a), (1) can be written as:

$$(3.3) \quad \frac{\partial}{\partial \alpha} \log \det \bar{M}[(1 - \alpha)\xi^\dagger + \alpha\xi^0]_{\alpha=0} = \frac{\partial}{\partial \alpha} \log \det M[(1 - \alpha)\xi^\dagger + \alpha\xi^0]_{\alpha=0} - \frac{\partial}{\partial \alpha} \log \det M_3[(1 - \alpha)\xi^\dagger + \alpha\xi^0]_{\alpha=0} \leq 0.$$

Now

$$(3.4) \quad \begin{aligned} \frac{\partial}{\partial \alpha} \log \det M[(1 - \alpha)\xi^\dagger + \alpha\xi^0]_{\alpha=0} &= t_r M^{-1}(\xi^\dagger) [M(\xi^0) - M(\xi^\dagger)] \\ &= t_r M^{-1}(\xi^\dagger) M(\xi^0) - p. \end{aligned}$$

Similarly

$$(3.5) \quad \frac{\partial}{\partial \alpha} \log \det M_3[(1 - \alpha)\xi^\dagger + \alpha\xi^0]_{\alpha=0} = t_r M_3^{-1}(\xi^\dagger) M_3(\xi^0) - (p - s).$$

Combining (3.3), (3.4) and (3.5) gives:

$$(3.6) \quad \frac{\partial}{\partial \alpha} \log \det \bar{M}[(1 - \alpha)\xi^\dagger + \alpha\xi^0]_{\alpha=0} = t_r M^{-1}(\xi^\dagger) M(\xi^0) - t_r M_3^{-1}(\xi^\dagger) M_3(\xi^0) - S \leq 0.$$

Let ξ^0 be a single frequency design at ω , (3.6) and the definition of M and M_3 for a single frequency design then imply that

$$h^*(e^{j\omega}) M^{-1}(\xi^\dagger) h(e^{j\omega}) - h^{(2)*}(e^{j\omega}) M_3^{-1}(\xi^\dagger) h^{(2)}(e^{j\omega}) \leq S$$

or

$$(3.7) \quad d(\omega, \xi^\dagger) - d_{(p-s)}(\omega, \xi^\dagger) = d_s(\omega, \xi^\dagger) \leq S.$$

Comparing (3.7) with (3.1) shows that (1) \rightarrow (3).

Conversely if $d_s(\omega, \xi^\dagger) = S$ holds, we know that (3.6) holds for all single frequency design ξ^0 , hence (3) \rightarrow (1).

Finally (3.1) and (3.7) indicate that ξ^\dagger minimizes $\max_{\omega} d_s(\omega, \xi)$; thus (2) \rightarrow (3). This completes the proof of the theorem.

In the next section, we investigate a sequential design algorithm based on the equivalence theorem.

4. A sequential design algorithm. Sequential design algorithms with proven convergence to a D -optimal design were proposed by Mehra [1] and extensively studied by Zarrop [4]. We shall study, in this section, an algorithm which is essentially due to Mehra [1] and Zarrop [4] but extends to the D_s -optimality case.

ALGORITHM.

- (1) Choose a design $\xi_0 \in \Xi$ such that $M(\xi_0)$ is nonsingular.
- (2) Set $k = 1$.
- (3) Choose a frequency ω_k such that $d_s(\omega_k, \xi_k) = \max_{\omega \in [0, \pi]} d_s(\omega, \xi_k)$.
- (4) If $d_s(\omega_k, \xi_k) = S$, stop.
- (5) Update design to $\xi_{k+1} = (1 - \alpha_k)\xi_k + \alpha_k \xi_{\omega_k} \in \Xi$.
- (6) Set $k = k + 1$; go to (3).

THEOREM 4.1. *If the sequence $\{\alpha_k\}$ in the algorithm is chosen so that*

$$\alpha_k \in [0, 1], \quad \lim_{k \rightarrow \infty} \alpha_k = 0, \quad \sum_{k=0}^{\infty} \alpha_k = \infty,$$

then $\lim_{k \rightarrow \infty} \xi_k = \xi^\dagger \in \Xi$ a D_s -optimal design.

Proof. By (2.2a)

$$(4.1) \quad \frac{\det \bar{M}(\xi_{k+1})}{\det \bar{M}(\xi_k)} = \frac{\det M(\xi_{k+1})/\det M(\xi_k)}{\det M_3(\xi_{k+1})/\det M_3(\xi_k)}.$$

Following Zarrop [4], we have:

$$(4.2) \quad M(\xi_{k+1}) = (1 - \alpha_k)M(\xi_k) + \alpha_k \operatorname{Re} \{h(e^{j\omega}k)h^*(e^{j\omega}k)\},$$

$$(4.3) \quad \frac{\det M(\xi_{k+1})}{\det M(\xi_k)} = (1 - \alpha_k)^p \{1 + \beta_k d(\omega_k, \xi_k) + \beta_k^2 g(\omega_k, \xi_k)\},$$

where $\beta_k = \alpha_k/(1 - \alpha_k)$

$$d(\omega, \xi) = h^*(e^{j\omega})M^{-1}(\xi)h(e^{j\omega}),$$

$$g(\omega, \xi) = \frac{1}{4}\{d^2(\omega, \xi) - |d_1(\omega, \xi)|^2\},$$

$$d_1(\omega, \xi) = h^T(e^{j\omega})M^{-1}(\xi)h(e^{j\omega}).$$

Similarly

$$(4.4) \quad M_3(\xi_{k+1}) = (1 - \alpha_k)M_3(\xi_k) + \alpha_k \operatorname{Re} \{h^{(2)}(e^{j\omega_k})h^{(2)*}(e^{j\omega_k})\}$$

and

$$(4.5) \quad \frac{\det M_3(\xi_{k+1})}{\det M_3(\xi_k)} = (1 - \alpha_k)^r \{1 + \beta_k d_r(\omega_k, \xi_k) + \beta_k^2 g_r(\omega_k, \xi_k)\},$$

where $r = p - s$,

$$d_r(\omega, \xi) = h^{(2)*}(e^{j\omega})M_3^{-1}(\xi)h^{(2)}(e^{j\omega}),$$

$$g_r(\omega, \xi) = \frac{1}{4}\{d^2(\omega, \xi) - |d_{1r}(\omega, \xi)|^2\},$$

$$d_{1r}(\omega, \xi) = h^{(2)T}(e^{j\omega})M_3^{-1}(\xi)h^{(2)}(e^{j\omega}).$$

Combining (4.1), (4.3) and (4.5) gives

$$(4.6) \quad \frac{\det \bar{M}(\xi_{k+1})}{\det \bar{M}(\xi_k)} = (1 - \alpha)^s \cdot \frac{1 + \beta_k d(\omega_k, \xi_k) + \beta_k^2 g(\omega_k, \xi_k)}{1 + \beta_k d_r(\omega_k, \xi_k) + \beta_k^2 g_r(\omega_k, \xi_k)}.$$

Since $\lim_{k \rightarrow \infty} \alpha_k = 0$, the first order expansion of (4.6) in α gives

$$\begin{aligned} \log \det \bar{M}(\xi_{k+1}) &\simeq \log \det \bar{M}(\xi_k) - \alpha_k S + \alpha_k d - \alpha_k d_r \\ &= \log \det \bar{M}(\xi_k) + \alpha_k (d_s - S). \end{aligned}$$

Now $\lim_{k \rightarrow \infty} \alpha_k = 0$, therefore there exists k_0 , η dependent on k_0 such that $0 < \eta \leq 1$ and for all $k \geq k_0$,

$$(4.7) \quad \log \det \bar{M}(\xi_{k+1}) > \log \det \bar{M}(\xi_k) + \eta \alpha_k \{d_s(\omega_k, \xi_k) - S\}.$$

Step (3) of the algorithm ensures that $d_s(\omega_k, \xi_k) \geq S$ (cf. Theorem 3.2) and therefore $\{\det \bar{M}(\xi_k), k \geq k_0\}$ is a monotonically increasing sequence bounded above by

$\det M(\xi^\dagger)$, therefore

$$(4.8) \quad \lim_{k \rightarrow \infty} \det \bar{M}(\xi_k) = \det \bar{M}(\xi') \leq \det \bar{M}(\xi^\dagger)$$

for some $\xi' \in \Xi$.

To show that $\det \bar{M}(\xi') = \det \bar{M}(\xi^\dagger)$, assume the contrary; then there exists $\varepsilon > 0$ such that $d_s(\omega_k, \xi_k) - S \geq \varepsilon$, for all $k > k_0$ and from (4.7)

$$\log \det \bar{M}(\xi_{k_0}) + \varepsilon \eta \sum_{k=k_0}^{\infty} \alpha_k < \log \det \bar{M}(\xi').$$

By assumption, $\sum_{k=0}^{\infty} \alpha_k = \infty$ and the sequence $\{\det \bar{M}(\xi_k)\}$ is unbounded and this contradicts (4) and thus completes the proof.

It is noted that many of the variations to improve the rate of convergence discussed in Zarrop [4], Wynn [13] and Atwood [5] do not necessarily apply in this case. In particular, the iterative design procedures introduced in Fedorov [3], Wynn [13] and Atwood [5] may not converge.

Example 4.1. Consider the system (Zarrop [4, p. 121])

$$y_t = (b_0 + b_1 z^{-1})u_{t-1} + \frac{d_0 + d_1 z^{-1}}{1 + C_1 z^{-1}} e_t.$$

Following § 2, we have

$$\theta^T = [b_0 b_1],$$

$$h_1(z^{-1}) = -\frac{1 + c_1 z^{-1}}{d_0 + d_1 z^{-1}} z^{-1},$$

$$h_2(z^{-1}) = -\frac{1 + c_1 z^{-1}}{d_0 + d_1 z^{-1}} z^{-2},$$

$$M = \frac{1 + c_1^2 + 2c_1 \cos \omega}{d_0^2 + d_1^2 + 2d_0 d_1 \cos \omega} \cdot \begin{bmatrix} 1 & \cos \omega \\ \cos \omega & 1 \end{bmatrix}.$$

Estimating b_0 using the sequential design algorithm given in § 4 where $\alpha_k = 1/(k+1)$ is used, Table 1 gives the final design for several values of c_1 , d_0 and d_1 with initial design $\omega = 0.5$, $\lambda = 1$.

TABLE 1
Final design for Example 4.1

c_1	d_0	d_1	ω	λ
0.5	1	0.3	1.46	1
0.1	-0.8	0.3	1.11	1
-0.5	-1	-0.3	2.10	1

In all cases, step 3 chooses the optimal frequency after several iterations. The convergence characteristics for three different initial designs for the case $c_1 = .5$, $d_0 = 1$, $d_1 = .3$ are shown in Fig. 4.1.

It is interesting to compare the D_s -optimal and the D -optimal designs. For this system with $c_1 = .5$, $d_0 = 1$, and $d_1 = 0.3$ the D -optimal design gives $\omega = 1.3804$ and

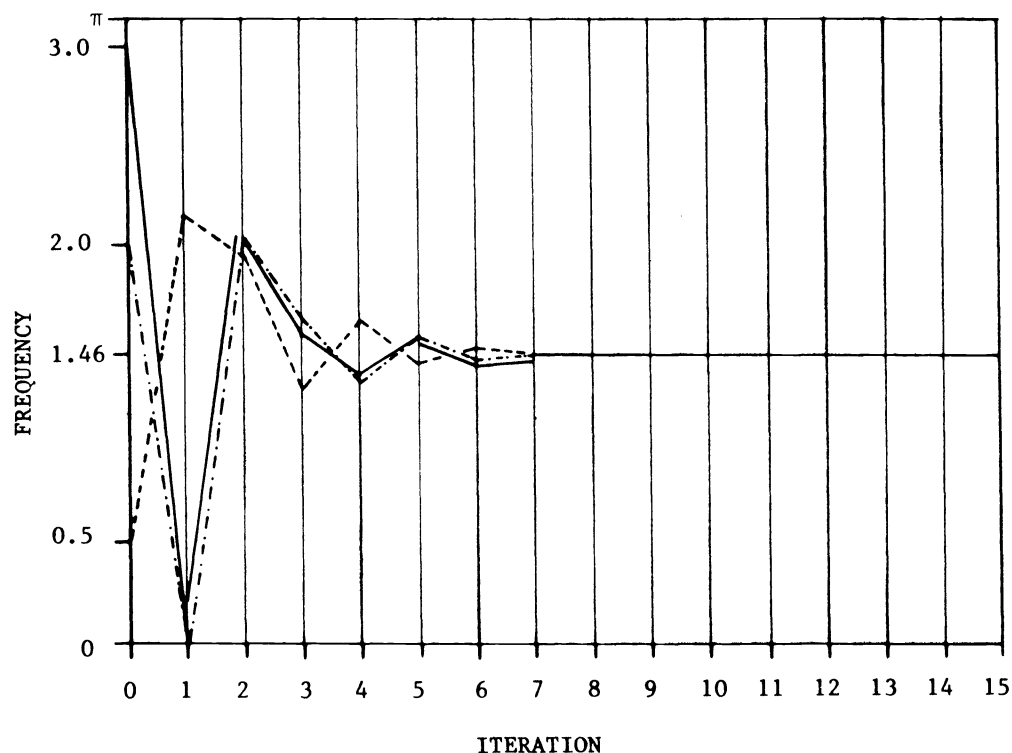


FIG. 4.1. Convergence characteristics for different initial design frequencies.

variance $\hat{b}_0 = 0.8673$ (Zarrop [4, p. 121]). The D_s -optimal design gives $\omega = 1.46$ and variance $\hat{b}_0 = 0.8604$. It can be seen from these results that the D_s -optimal design achieves a lower value on the variance of the parameter estimate.

Example 4.2. Consider the system

$$y_t = \frac{b_0 u_{t-1}}{1 + a_1 z^{-1}} + \frac{d_0 + d_1 z^{-1}}{1 + c_1 z^{-1}} e_t.$$

Following § 2, we have:

$$\theta^T = [a_1 b_0],$$

$$h_1(z^{-1}) = b_0 \cdot \frac{1 + c_1 z^{-1}}{d_0 + d_1 z^{-1}} \cdot \frac{z^{-1}}{1 + a_1 z^{-1}},$$

$$h_2(z^{-1}) = -\frac{1 + c_1 z^{-1}}{d_0 + d_1 z^{-1}} \cdot \frac{z^{-1}}{1 + a_1 z^{-1}}$$

and

$$M = b_0^2 \cdot \frac{1 + c_1^2 + 2c_1 \cos \omega}{d_0^2 + d_1^2 + 2d_0 d_1 \cos \omega} \cdot \frac{1}{(1 + a_1^2 + 2a_1 \cos \omega)}$$

$$\cdot \begin{bmatrix} 1 & -\frac{1}{b_0}(a_1 + \cos \omega) \\ -\frac{1}{b_0}(a_1 + \cos \omega) & \frac{1}{b_0^2}(1 + a_1^2 + 2a_1 \cos \omega) \end{bmatrix}.$$

Estimating a_1 using the sequential design algorithm in § 4 with $\alpha_k = 1/(k+1)$, the final design for several values of a_1 , b_0 , c_1 , d_0 and d_1 are shown in Table 2.

TABLE 2
Final design for Example 4.2

a_1	b_0	c_1	d_0	d_1	ω_1	λ_1	ω_2	λ_2
-0.1	1.0	-0.1	1.0	0.0	0	0.5	π	0.5
-0.3	0.5	0.5	1.0	0.7	0	0.5	π	0.5

In all cases, the initial design is $\omega = 0.5$, $\lambda = 1$. The same final design is also obtained with different initial frequencies. In all the cases tried, the final design always converges to two frequencies.

5. Concluding remarks. Optimal input design for identification experiments based on two different criteria has been shown to be equivalent for the case of estimating part of the system parameters. A proven sequential design algorithm which converges to a D -optimal design has also been extended to the case of estimating part of the system parameters and illustrated by two simple examples.

In this paper, we have assumed that $M(\xi)$ is nonsingular. Methods used to overcome singular $M(\xi)$ in linear regression problems, e.g., transformations used in Kiefer [10] and Atwood [14] cannot be extended to the dynamic case. Further work is being carried out on the singular case and will be reported at a later stage.

REFERENCES

- [1] R. K. MEHRA, *Optimal input signals for parameter estimation in dynamic systems—survey and new results*, IEEE Trans. Automat. Control, AC-19 (1974), pp. 753–763.
- [2] J. KIEFER AND J. WOLFOWITZ, *The equivalence of two extremum problems*, Canad. J. Math., 12 (1960), pp. 363–366.
- [3] V. V. FEDOROV, *Theory of Optimal Experiments*, Academic Press, New York, 1972.
- [4] M. B. ZARROP, *Optimal Experiment Design for Dynamic System Identification*, Springer-Verlag, New York, 1979.
- [5] C. L. ATWOOD, *Sequences converging to D-optimal design of experiments*, Ann. Statist., 1 (1973), pp. 342–352.
- [6] G. C. GOODWIN AND R. L. PAYNE, *Design and characterization of optimal test signals for SISO parameter estimation*, Paper TT-1, 3rd IFAC Symposium, the Hague, 1973.
- [7] G. C. GOODWIN, J. C. MURDOCH AND R. L. PAYNE, *Optimal test signal design for linear SISO system identification*, Int. J. Control, 17 (1973), pp. 45–55.
- [8] T. S. NG, G. C. GOODWIN AND T. SÖDERSTRÖM, *Optimal experiment design for linear systems with input-output constraints*, Automatica, 13 (1977), pp. 571–577.
- [9] T. SÖDERSTRÖM, *Notes on the design of optimal identification experiments*, Rep. UPTEC 7563R, Inst. of Technology, Uppsala University, Sweden, September 1975.
- [10] J. KIEFER, *Optimum designs in regression problems II*, Annal. Math. Statist., 32 (1961), pp. 298–325.
- [11] G. C. GOODWIN AND R. L. PAYNE, *Dynamic System Identification: Experiment Design and Data Analysis*, Academic Press, New York, 1977.
- [12] J. KIEFER, *Optimum experiment designs*, J.R.S.S. (Series B), 21 (1959), pp. 272–319.
- [13] H. P. WYNN, *The sequential generation of D-optimum experimental design*, Ann. Math. Statist., 41 (1970), pp. 1655–1664.
- [14] C. L. ATWOOD, *Optimal and efficient designs of experiments*, Ann. Math. Statist., 40 (1969), pp. 1570–1602.

ON THE CONVERGENCE OF THE DISCRETE TIME DYNAMIC PROGRAMMING EQUATION FOR GENERAL SEMIGROUPS*

A. BENSOUSSAN† AND M. ROBIN‡

Abstract. We consider several classes of control problems for Markov processes (continuous control, optimal stopping, impulse control). The formulation we use is valid for general Markov semigroups. We study the discrete time approximation of the dynamic programming equation, using mainly an analytical approach. Probabilistic interpretation is given for some of the results.

Introduction. Let $\phi(t)$ be a Markov semigroup, and more generally, $\phi^v(t)$ a family of semigroups depending on a parameter. We formulate several problems corresponding to different stochastic control situations. The “continuous control” case is the following: to find u (Borel bounded or uniformly continuous bounded) which is the *maximum element* of the set

$$(*) \quad u \leq \int_0^t \phi^v(s) L_v e^{-\alpha s} ds + \phi^v(t) u e^{-\alpha t} \quad \forall t \geq 0, \quad \forall v,$$

where L_v is given and $\alpha > 0$.

The stopping time problem is the following: to find $u \in B$ or C which is the maximum element of the set

$$(**) \quad u \leq \int_0^t \phi(s) L e^{-\alpha s} dt + e^{-\alpha t} \phi(t) u \quad \forall t \geq 0,$$

$$u \leq \psi,$$

where ψ is a given function (called the obstacle).

The impulse control problem corresponds to (**) when the obstacle function ψ is not given a priori but depends on the solution u , $\psi = M(u)$, where M is a nonlinear operator.

We consider discretized versions of the above problems. Problem (*) is approximated by

$$(*)_h \quad u_h = \min_v \left[\int_0^h e^{-\alpha s} \phi^v(s) L_v ds + e^{-\alpha h} \phi^v(h) u_h \right].$$

Problem (**) is approximated by

$$(**)_h \quad u_h = \min \left[\psi, \int_0^h e^{-\alpha s} \phi(s) L ds + e^{-\alpha h} \phi(h) u_h \right].$$

The impulse control problem is approximated by

$$(***)_h \quad u_h = \min \left[M u_h, \int_0^h e^{-\alpha s} \phi(s) L ds + e^{-\alpha h} \phi(h) u_h \right].$$

The same kind of discretization process was used in the pioneering work of Nisio (cf. [7a], [7b], [7c]) on nonlinear semigroups associated to optimal stochastic control

* Received by the editors March 2, 1981, and in revised form October 19, 1981.

† University of Paris-Dauphine and Institut National de Recherche en Informatique et en Automatique, Domaine de Voluceau, Rocquencourt, B.P. 105, 78150 Le Chesnay, France.

‡ Institut National de Recherche en Informatique et en Automatique, Domaine de Voluceau-Rocquencourt, B.P. 105, 78150 Le Chesnay, France.

problems. Here we consider *directly* the stationary dynamic programming equation in an abstract state space and for general semigroups. The results concerning the approximation of the stopping time problem and the impulse control problem are new, to the best of our knowledge (although Shiryaev [9] has also considered discretization to study optimal stopping time problems, in a probabilistic framework).

After having established all the analytical results, we give the probabilistic interpretation of (*) and (**), but not of (***)_n, which is too technical and would extend the paper beyond reasonable limits.

1. The problem of semigroup envelopes.

1.1. Notation and assumptions. Let E be a Polish space provided with the Borel σ -algebra \mathcal{E} . We note that B is the space of Borel bounded functions on E and that C is the space of bounded uniformly continuous functions on E . We consider a family $\phi^v(t)$, $v \in V$ of operators such that

$$(1.1) \quad V \text{ is a finite set,}$$

$$\phi^v(t) \in \mathcal{L}(B; B), \quad \phi^v(0) = I,$$

$$(1.2) \quad \|\phi^v(t)\| \leq 1,$$

$$\phi^v(t+s) = \phi^v(t)\phi^v(s), \quad \phi^v(t)\varphi \geq 0 \text{ if } \varphi \geq 0.$$

A semigroup of operators on B satisfying (1.2) is called a *Markov semigroup*.

We will also assume that

$$(1.3) \quad \phi^v(t): C \rightarrow C,$$

$$(1.4) \quad t \rightarrow \phi^v(t)\varphi(x) \text{ is continuous from } (0, \infty) \rightarrow \mathbb{R} \quad \forall x \text{ fixed, } \quad \forall \varphi \in C.$$

Let next $L(x, v)$ be a function such that

$$(1.5) \quad L_v(x) \equiv L(x, v) \in B, \quad \int_0^\infty e^{-\alpha t} \phi^v(t) L_v dt \in C,$$

where α is a positive number.

We formulate the following problem: to find u maximum solution of

$$(1.6) \quad \begin{aligned} u &\leq \int_0^t e^{-\alpha s} \phi^v(s) L_v ds + e^{-\alpha t} \phi^v(t) u \quad \forall t \geq 0, \quad \forall v, \\ u &\in B. \end{aligned}$$

We will study (1.6) by the following discretization procedure:

$$(1.7) \quad u_h = \min_v \left[\int_0^h e^{-\alpha s} \phi^v(s) L_v ds + e^{-\alpha h} \phi^v(h) u_h \right], \quad u_h \in C.$$

1.2. Convergence result. Our main result in this section will be the following.

THEOREM 1.1. *We assume (1.1), (1.2), (1.3), (1.4), (1.5). Then $u_{1/2^q} \downarrow u$, the maximum solution of (1.6), as $q \uparrow +\infty$. \square*

We will need several lemmas. We first recall a property of Markov semigroups which is given in Dynkin [4].

$$(1.8) \quad \begin{aligned} &\text{Let } \varphi_n \in B \text{ such that } \varphi_n(x) \rightarrow \varphi(x) \quad \forall x \text{ and } \|\varphi_n\| \leq K. \text{ Then} \\ &\phi(t)\varphi_n(x) \rightarrow \phi(t)\varphi(x) \quad \forall x. \end{aligned}$$

LEMMA 1.1. *There exists one and only one solution of (1.7).*

Proof. Let $z \in B$, set $T_h z = \min_v \left[\int_0^h e^{-\alpha t} \phi^v(t) L_v dt + e^{-\alpha h} \phi^v(h) z \right]$.

Since the set V is finite, $T_h z \in B$. Note also that when $z \in C$, $T_h z \in C$. Moreover,

$$\|T_h z_1 - T_h z_2\| \leq e^{-\alpha h} \|z_1 - z_2\|,$$

hence T_h is a contraction and therefore has one and only one fixed point u_h . \square

LEMMA 1.2. *Let $z \in B$ such that $z \leq T_h z$. Then $z \leq u_h$.*

Proof. T_h is increasing in z , hence $z \leq T_h z \leq T_h^2 z, \dots, \leq T_h^n z \leq u_h$ by the contraction property, hence the result. \square

LEMMA 1.3. *We have*

$$(1.9) \quad u_h \leq u_{2h}.$$

Proof. For any v , we have

$$u_h \leq \int_0^h e^{-\alpha s} \phi^v(s) L_v ds + e^{-\alpha h} \phi^v(h) u_h;$$

hence

$$e^{-\alpha h} \phi^v(h) u_h \leq \int_0^h e^{-\alpha(s+h)} \phi^v(s+h) L_v ds + e^{-2\alpha h} \phi^v(2h) u_h,$$

which implies

$$u_h \leq \int_0^{2h} e^{-\alpha s} \phi^v(s) L_v ds + e^{-2\alpha h} \phi^v(2h) u_h.$$

Hence $u_h \leq T_{2h} u_h$, which with Lemma 1.2 implies (1.9).

LEMMA 1.4. *We have*

$$(1.10) \quad u_h \geq -K, \quad K = \max_v \frac{\|L_v^-\|}{\alpha}.$$

Proof. Assume $z \geq -K$. Then

$$\begin{aligned} \int_0^h e^{-\alpha s} \phi^v(s) L_v ds + e^{-\alpha h} \phi^v(h) z &\geq -e^{-\alpha h} + \int_0^h e^{-\alpha s} \phi^v(s) L_v^- ds \\ &\geq -K e^{-\alpha h} - \max_v \|L_v^-\| \int_0^h e^{-\alpha s} ds \geq -K, \end{aligned}$$

hence $T_h z \geq -K$. Starting with $z = 0$, we deduce $T_h^n 0 \geq -K$, hence $u_h \geq -K$. \square

Proof of Theorem 1.1. Let us set $u_q = u_{1/2^q}$. Then from Lemmas 1.3 and 1.4 it follows that

$$(1.11) \quad u_q \downarrow u \quad \text{as } q \uparrow \infty.$$

Note that u is u.s.c.

Furthermore, we have

$$u_h \leq \int_0^{mh} e^{-\alpha s} \phi^v(s) L_v ds + e^{-\alpha mh} \phi^v(mh) u_h \quad \forall m \text{ integer}.$$

Taking $h = 1/2^q$, $m = j^{2^{q-l}}$ with $l \leq q$, we obtain

$$(1.12) \quad u_q \leq \int_0^{j/2^l} e^{-\alpha s} \phi^v(s) L_v ds + e^{-\alpha j/2^l} \phi^v\left(\frac{j}{2^l}\right) u_q, \quad l \leq q, \quad j \text{ integer}.$$

We can let $q \uparrow +\infty$ in (1.12), j, l fixed; from (1.8) and (1.11) we deduce

$$(1.13) \quad \begin{aligned} u &\leq \int_0^{1/2^l} e^{-\alpha s} \phi^v(s) L_v ds + e^{-\alpha j/2^l} \phi^v\left(\frac{j}{2^l}\right) u \\ &\leq \int_0^{j/2^l} e^{-\alpha s} \phi^v(s) L_v ds + e^{-\alpha j/2^l} \phi^v\left(\frac{j}{2^l}\right) u_q, \end{aligned}$$

where now j, l, q are arbitrary integers. We take $j = [t2^l] + 1$ and let $l \rightarrow \infty$. Since $u_q \in C$ we may use assumption (1.4) to assert that

$$u \leq \int_0^t e^{-\alpha s} \phi^v(s) L_v ds + e^{-\alpha t} \phi^v(t) u_q,$$

in which we may let again q tend to $+\infty$. This proves that u satisfies (1.6). It is certainly the maximum element. Indeed, let \tilde{u} be a solution of (1.6). Then clearly $\tilde{u} \leq T_h \tilde{u}$, which with Lemma 1.2 implies $\tilde{u} \leq u_h$, and hence, $\tilde{u} \leq u$. \square

1.3. Additional regularity results. We assume here that

$$(1.14) \quad |L(x, v) - L(y, v)| \leq K|x - y|^\delta, \quad 0 \leq \delta \leq 1,$$

$$(1.15) \quad \begin{aligned} \forall g \in C^{0,\delta}(E) \quad (\text{i.e., } |g(x) - g(y)| \leq \|g\|_\delta |x - y|^\delta), \\ |\phi^v(t)g(x) - \phi^v(t)g(y)| \leq e^{\lambda t} \|g\|_\delta |x - y|^\delta, \quad \lambda \geq 0. \end{aligned}$$

THEOREM 1.2. *We make the assumptions of Theorem 1.1 and (1.14), (1.15). Then if $\alpha > \lambda$, $u \in C^{0,\delta}(E)$.*

Proof. Let $z \in C^{0,\delta}(E)$. Let us fix x_0 in E . There exists v_0 (depending on x_0) such that

$$T_h z(x_0) = \int_0^h e^{-\alpha s} \phi^{v_0}(s) L_{v_0}(x_0) ds + e^{-\alpha h} \phi^{v_0}(h) z(x_0).$$

Now, letting x be arbitrary, we have

$$T_h z(x) \leq \int_0^h e^{-\alpha s} \phi^{v_0}(s) L_{v_0}(x) ds + e^{-\alpha h} \phi^{v_0}(h) z(x).$$

Hence by the difference

$$\begin{aligned} T_h z(x) - T_h z(x_0) &\leq \int_0^h e^{-\alpha s} (\phi^{v_0}(s) L_{v_0}(x) - \phi^{v_0}(s) L_{v_0}(x_0)) ds \\ &\quad + e^{-\alpha h} (\phi^{v_0}(h) z(x) - \phi^{v_0}(h) z(x_0)), \end{aligned}$$

and from assumptions (1.14), (1.15), it follows that

$$\leq \int_0^h e^{-\alpha s} e^{\lambda s} K |x - x_0| ds + e^{-\alpha h} e^{\lambda h} \|z\|_\delta |x - x_0|^\delta.$$

Hence, if $x \neq x_0$,

$$\frac{T_h z(x) - T_h z(x_0)}{|x - x_0|^\delta} \leq K \frac{1 - e^{-(\alpha - \lambda)h}}{\alpha - \lambda} + e^{-(\alpha - \lambda)h} \|z\|_\delta,$$

and, since x_0, x are arbitrary, we deduce

$$(1.16) \quad \|T_h z\|_\delta \leq K \frac{1 - e^{-(\alpha - \lambda)h}}{\alpha - \lambda} + e^{-(\alpha - \lambda)h} \|z\|_\delta.$$

Iterating we obtain

$$\|T_{hz}^k\|_\delta \leq K \frac{1 - e^{-(\alpha-\lambda)h}}{\alpha - \lambda} (1 + e^{-(\alpha-\lambda)h} + \dots + e^{-(\alpha-\lambda)(k-1)h}) + e^{-k(\alpha-\lambda)} \|z\|_\delta,$$

hence, letting $k \rightarrow \infty$,

$$(1.17) \quad \|u_h\|_\delta \leq \frac{K}{\alpha - \lambda}.$$

Taking $h = 1/2^q$ and letting $q \uparrow +\infty$, we deduce $\|u\|_\delta \leq K/(\alpha - \lambda)$. \square

Remark 1.1. We now show that we can get rid of the assumption $\alpha > \lambda$ provided that we look for a solution in $C^0(E)$. This result seems new.

We assume

$$(1.18) \quad t \rightarrow \phi^v(t)\varphi(x) \text{ is (Lebesgue) measurable from } (0, \infty) \text{ in } R \ \forall \varphi \in B, \forall x \text{ fixed.}$$

We first state a technical result.

LEMMA 1.5. *Let $\phi(t)$ be a Markov semigroup on B , i.e., satisfying (1.2), which verifies (1.18). Let $w \in B$ such that*

$$(1.19) \quad w \leq \int_0^t e^{-\alpha s} \phi(s)g \, ds + e^{-\alpha t} \phi(t)w \quad \forall t \geq 0,$$

where $g \in B$. Then for any $\beta > 0$ one has

$$(1.20) \quad w \leq \int_0^t e^{-\beta s} \phi(s)(g + (\beta - \alpha)w) \, ds + e^{-\beta t} \phi(t)w \quad \forall t \geq 0.$$

Proof. We set

$$H(t) = w - \int_0^t e^{-\alpha s} \phi(s)g \, ds - e^{-\alpha t} \phi(t)w;$$

we have

$$H(0) = 0, \quad H(t) \leq 0 \quad \forall t.$$

In fact, we have the additional property

$$(1.21) \quad H(t) \leq H(s) \quad \text{for } t \geq s.$$

Indeed, (1.21) amounts to proving that

$$(1.22) \quad e^{-\alpha s} \phi(s)w \leq e^{-\alpha t} \phi(t)w + \int_s^t e^{-\alpha \lambda} \phi(\lambda)g \, d\lambda, \quad s \leq t.$$

But from (1.19),

$$w \leq \int_0^{t-s} e^{-\alpha r} \phi(r)g \, dr + e^{-\alpha(t-s)} \phi(t-s)w;$$

hence

$$e^{-\alpha s} \phi(s)w \leq \int_0^{t-s} e^{-\alpha(r+s)} \phi(r+s)g \, dr + e^{-\alpha t} \phi(t)w,$$

i.e. (1.22). We next have

$$e^{-(\beta-\alpha)t}w = e^{-\beta t}\phi(t)w + e^{-(\beta-\alpha)t} \int_0^t e^{-\alpha s}\phi(s)g \, ds + e^{-(\beta-\alpha)t}H(t),$$

and integrating between 0 and T , we deduce

$$\begin{aligned} [1 - e^{-(\beta-\alpha)T}]w &= \int_0^T (\beta - \alpha) e^{-\beta t}\phi(t)w \, dt + \int_0^T (\beta - \alpha) e^{-(\beta-\alpha)t}H(t) \, dt \\ &\quad + \int_0^T (\beta - \alpha) e^{-(\beta-\alpha)t} \left(\int_0^t e^{-\alpha s}\phi(s)g \, ds \right) dt \\ &= \int_0^T (\beta - \alpha) e^{-\beta t}\phi(t)w \, dt \\ &\quad + \int_0^T (\beta - \alpha) e^{-(\beta-\alpha)t}H(t) \, dt - e^{-(\beta-\alpha)T} \int_0^T e^{-\alpha t}\phi(t)g \, dt \\ &\quad + \int_0^T e^{-\beta t}\phi(t)g \, dt; \end{aligned}$$

hence

$$\begin{aligned} w &= \int_0^T e^{-\beta t}\phi(t)(g + (\beta - \alpha)w) \, dt + e^{-\beta T}\phi(T)w + e^{-(\beta-\alpha)T}H(T) \\ &\quad + \int_0^T (\beta - \alpha)^{-(\beta-\alpha)t}H(t) \, dt. \end{aligned}$$

If $\beta \geq \alpha$, since $H(t) \leq 0$, we clearly have (1.20) with $t = T$. If $\beta < \alpha$, then using (1.21) we have

$$H(t) \geq H(T), \quad (\beta - \alpha)H(t) \leq (\beta - \alpha)H(T);$$

hence,

$$\begin{aligned} e^{-(\beta-\alpha)T} + \int_0^T (\beta - \alpha) e^{-(\beta-\alpha)t}H(t) \, dt \\ \leq e^{-(\beta-\alpha)T}H(T) + (\beta - \alpha)H(T) \int_0^T e^{-(\beta-\alpha)t} \, dt = H(T) \leq 0. \end{aligned}$$

Therefore, (1.20) holds in all cases for $t \leq T$. Since T is arbitrary, the desired result is proved. \square

We then can prove:

LEMMA 1.6. *Under the assumptions of Theorem 1.1 and (1.18), the maximum solution of (1.6) is also the maximum solution of*

$$\begin{aligned} (1.23) \quad &u \in B, \\ &u \leq \int_0^t e^{-\beta s}\phi^v(s)(L_v + (\beta - \alpha)u) \, ds + e^{-\beta t}\phi^v(t)u \quad \forall v, \quad \forall t \geq 0, \end{aligned}$$

where $\beta > 0$ is arbitrary.

Proof. Let u be the maximum solution of (1.6). In view of Lemma 1.5, u is also a solution of (1.23) for all $\beta > 0$. If w is another solution of (1.23), by Lemma 1.5,

w also satisfies (1.6). Therefore $w \leq u$, and this proves that u is the maximum solution of (1.23). \square

We then state:

THEOREM 1.3. *We make the assumptions of Theorem 1.1, and (1.14), (1.15), (1.18). Then u , the maximum solution of (1.6), belongs to C and $u_{1/2^q}$ (defined in Theorem 1.1) converges to u , uniformly on every compact of E .*

Proof. Let $z \in C$ and let ξ be the maximum element of

$$(1.24) \quad \begin{aligned} &\xi \in B, \\ &\xi \leq \int_0^t e^{-\beta s} \phi^v(s)(L_v + (\beta - \alpha)z) ds + e^{-\beta t} \phi^v(t)\xi \quad \forall v, \quad \forall t. \end{aligned}$$

We thus have defined a map $S: C \rightarrow B$.

According to Theorem 1.2, provided $\beta > \lambda$ (which we suppose), $S: C^{0,\delta} \rightarrow C^{0,\delta}$.

Consider now $S_h: B \rightarrow B$ defined by $\xi_h = S_h(z)$, where

$$(1.25) \quad \xi_h = \min_v \left[\int_0^h e^{-\beta s} \phi^v(s)(L_v + (\beta - \alpha)z) ds + e^{-\beta h} \phi^v(h)\xi_h \right].$$

Note that $S_h: C \rightarrow C$. One easily checks the estimates

$$(1.26) \quad \|S_h(z_1) - S_h(z_2)\| \leq \frac{\beta - \alpha}{\beta} \|z_1 - z_2\| \quad \forall z_1, z_2 \in B.$$

Since

$$(1.27) \quad S(z) = \lim_{q \rightarrow +\infty} S_{1/2^q}(z) \quad \forall z \in C,$$

we deduce from (1.26) that

$$(1.28) \quad \|S(z_1) - S(z_2)\| \leq \frac{\beta - \alpha}{\beta} \|z_1 - z_2\| \quad \forall z_1, z_2 \in C.$$

We also note the relation, which follows from Lemma 1.5,

$$(1.29) \quad u \leq S_h(u).$$

Define now $u^n = S^n(0)$, $u_h^n = S_h^n(0)$. Since S maps $C^{0,\delta}$ into itself, then $u^n \in C^{0,\delta}$. From (1.28) it follows $\|u^{n+1} - u^n\| \leq (\beta - \alpha)/\beta^n \|u^1\|$, and thus $u^n \rightarrow w$ in C .

We are going to prove that

$$(1.30) \quad u = w,$$

which will prove the desired result. We also have from (1.26)

$$(1.31) \quad S_h^n(0) \rightarrow w_h \text{ in } C \text{ and } w_h \text{ is the fixed point of } S_h \text{ (in } B).$$

From (1.29) it follows that

$$(1.32) \quad u \leq w_h.$$

Now from (1.26), (1.28) we can assert that

$$(1.33) \quad \|u^n - w\| \leq \left(\frac{\beta - \alpha}{\beta} \right)^n \frac{\beta}{\alpha} \|u^1\|,$$

$$(1.34) \quad \|u_h^n - w_h\| \leq \left(\frac{\beta - \alpha}{\beta} \right)^n \frac{\beta}{\alpha} \max_v \frac{\|L_v\|}{\beta}.$$

But we have

$$(1.35) \quad u_{1/2^n}^n \downarrow u^n \quad \text{as } q \uparrow \infty \quad \forall n.$$

This is proved by induction on n , as in Theorem 1.1. From (1.33), (1.34), (1.35) we conclude

$$(1.36) \quad w_{1/2^n}(x) \downarrow w(x) \quad \forall x \quad \text{as } q \uparrow \infty.$$

But (1.32) and (1.36) imply

$$(1.37) \quad u \leq w.$$

But since w is a fixed point of S , we have

$$w \leq \int_0^t e^{-\beta s} \phi^v(s)(L_v + (\beta - \alpha)w) ds + e^{-\beta t} \phi^v(t)w$$

and, from Lemma 1.5,

$$w \leq \int_0^t e^{-\alpha s} \phi^v(s)L_v ds + e^{-\alpha t} \phi^v(t)w;$$

therefore, $w \leq u$, which with (1.37) implies (1.30). \square

1.4. Probabilistic interpretation. We assume here, in addition to (1.2), that $\phi(t)1 = 1$. Let us define $\Omega = E^{[0, \infty)}$, $x(t, \omega)$ the canonical process, $\mu_t^s = \sigma(x(\lambda); t \leq \lambda \leq s)$, $\mu_t = \mu_t^\infty$. For simplicity we take (without loss of generality) $V = \{1, 2, \dots, m\}$. To $i \in V$, we associate a probability P_i^{xt} on (Ω, μ_t) such that

$$(1.38) \quad E_i^{xt} \phi(x(s)) = \phi^i(s-t)\phi(x) \quad \forall s \geq t.$$

We denote by W the class of the step processes adapted to μ_0^t with values in V . More precisely, if $V \in W$, then there exists a sequence $\tau_0 = 0 < \tau_1 < \dots < \tau_n < \dots$ which is *deterministic*, increasing and convergent to $+\infty$ and

$$(1.39) \quad V = v(\cdot), \quad v(t; \omega) = v_n(\omega), \quad t \in [\tau_n, \tau_{n+1}),$$

where v_n is $\mu_0^{\tau_n}$ -measurable with values in V .

We can define a family $Q_{\omega, t}^V$ of probabilities on (Ω, μ_0) indexed by the pair $\omega, t(\omega \in \Omega, t \geq 0)$, such that for $\Gamma \in \mu_0^{t-} = \sigma(\mu_0^s, 0 \leq s < t)$ and $\Delta \in \mu_n$, then one has

$$(1.40) \quad Q_{\omega, t}^V(\Gamma \cap \Delta) = \chi_\Gamma(\omega) P_{v(t; \omega)}^{x(t; \omega)^t}(\Delta).$$

Let us next define a sequence \tilde{P}_V^x of probabilities on (Ω, μ_0) as follows:

$$(1.41) \quad \tilde{P}_V^x = P_{v_0}^{x, 0},$$

where $v_0 = v_0(x)$, v_0 being a Borel function from $E \rightarrow V$, by (1.39). We then define by induction

$$(1.42) \quad \tilde{P}_V^{x, n+1}(\Gamma) = E^{\tilde{P}_V^x} Q_{\omega, \tau_{n+1}}^V(\Gamma).$$

We then note that

$$(1.43) \quad \tilde{P}_V^x(\Gamma) = \tilde{P}_V^{x, n+1}(\Gamma) = \dots \quad \forall \Gamma \in \mu_0^{\tau_{n+1}}.$$

Indeed, take $\Gamma \in \mu_0^{\tau_{n+1}}$, then

$$\Gamma = \Gamma_1 \cap \Gamma_2 \quad \text{with } \Gamma_1 \in \mu_0^{\tau_{n+1}-0}, \quad \Gamma_2 \in \sigma(x(\tau_{n+1})).$$

Therefore, by formula (1.40),

$$Q_{\omega, \tau_{n+1}}^V(\Gamma) = \chi_{\Gamma_1}(\omega) P_{v(\tau_{n+1}; \omega)}^{x(\tau_{n+1}; \omega), \tau_{n+1}}(\Gamma_2) = \chi_{\Gamma_1}(\omega) \chi_{\Gamma_2}(\omega) = \chi_{\Gamma_1 \cap \Gamma_2}(\omega) = \chi_{\Gamma}(\omega);$$

therefore, by (1.42) we obtain (1.43).

The family $P_V^n(\Gamma)$ on $(\Omega, \mu_0^{\tau_{n+1}})$ forms a system of compatible probabilities, and since E is a Polish space, then (cf. Neveu [6]) there exists one and only one probability P_V^x on (Ω, μ_0) such that

$$(1.44) \quad P_V^x = \bar{P}_V^x \text{ on } \mu_0^{\tau_{n+1}}.$$

LEMMA 1.7. *Let $\varphi \in B$ and $\tau_n \leq t < \tau_{n+1}$. Then we have*

$$(1.45) \quad E_V^x[\varphi(x(t)) | \mu_0^{\tau_n}] = \phi^{v_n}(t - \tau_n) \varphi(x(\tau_n)).$$

Proof. Let ξ_n be $\mu_0^{\tau_n}$ -measurable and bounded. Since $t < \tau_{n+1}$, we have

$$\begin{aligned} E_V^x \xi_n \varphi(x(t)) &= \bar{E}_V^x \xi_n \varphi(x(t)) = \bar{E}_V^x E^{Q_{\omega, \tau_n}^V} \xi_n(\omega') \varphi(x(t; \omega')) \\ &= \bar{E}_V^x \xi_n(\omega) E^{Q_{\omega, \tau_n}^V} \varphi(x(t; \omega')). \end{aligned}$$

But from (1.40) and (1.38)

$$E^{Q_{\omega, \tau_n}^V} \varphi(x(t; \omega')) = \phi^{v(\tau_n)}(t - \tau_n) \varphi(x(\tau_n))$$

and

$$\bar{E}_V^x \xi_n(\omega) \phi^{v(\tau_n)}(t - \tau_n) \varphi(x(\tau_n)) = E_V^x \xi_n \phi^{v(\tau_n)}(t - \tau_n) \varphi(x(\tau_n)),$$

from which (1.45) follows. \square

We next define, for $V \in W$,

$$(1.46) \quad J^x(V) = E_V^x \int_0^\infty e^{-\alpha t} L(x(t), v(t)) dt.$$

We have

$$J^x(V) = E_V^x \sum_{n=0}^\infty e^{-\alpha \tau_n} \int_{\tau_n}^{\tau_{n+1}} e^{-\alpha(t-\tau_n)} L(x(t), v_n) dt.$$

But from Lemma 1.7, for $\tau_n \leq t < \tau_{n+1}$,

$$E_V^x L(x(t), v_n) = E_V^x \phi^{V_n}(t - \tau_n) L_{v_n}(x(\tau_n)).$$

Therefore

$$(1.47) \quad E_V^x \int_{\tau_n}^{\tau_{n+1}} e^{-\alpha(t-\tau_n)} L(x(t), v_n) dt = E_V^x \int_0^{\tau_{n+1}-\tau_n} e^{-\alpha t} \phi^{v_n}(t) L_{v_n}(x(\tau_n)) dt.$$

Let us consider the function

$$(1.48) \quad L_h(x, v) = \int_0^h e^{-\alpha t} \phi^v(t) L_v(x) dt.$$

Then from (1.47)

$$E_V^x \int_{\tau_n}^{\tau_{n+1}} e^{-\alpha(t-\tau_n)} L(x(t), v_n) dt = L_{\tau_{n+1}-\tau_n}(x(\tau_n), v_n),$$

and defining

$$(1.49) \quad x_n(\omega) = x(\tau_n, \omega)$$

we obtain the following formula

$$(1.50) \quad J^x(V) = E_V^x \sum_{n=0}^{\infty} e^{-\alpha \tau_n} L_{\tau_{n+1}-\tau_n}(x_n(\omega), v_n(\omega)).$$

Let us set

$$(1.51) \quad W_h = \{V \in W \mid \tau_n = nh\}.$$

We have the following.

THEOREM 1.4. *We make the assumptions of Theorem 1.1 and (1.2). Then we have*

$$(1.52) \quad u_h(x) = \min_{V \in W_h} J^x(V).$$

Proof. For $V \in W_h$ we have

$$(1.53) \quad J_x(V) = E_V^x \sum_{n=0}^{\infty} e^{-\alpha nh} L_h(x_n, v_n),$$

and v_n is μ_0^{nh} -measurable. Note that v_0 is not random. Now from (1.7) we have

$$(1.54) \quad u_h(x) \leq \int_0^h e^{-\alpha s} \phi^v(s) L_v(x) ds + e^{-\alpha h} \phi^v(h) u_h(x) \quad \forall x, v.$$

Take V to be an arbitrary control. In (1.54) replace x by $x_n(\omega)$, and v by $v_n(\omega)$. We get

$$(1.55) \quad u_h(x_n) \leq \int_0^h e^{-\alpha s} \phi^{v_n}(s) L_{v_n}(x_n) ds + e^{-\alpha h} \phi^{v_n}(h) u_h(x_n).$$

But from Lemma 1.7 we can assert that

$$(1.56) \quad \phi^{v_n}(h) u_h(x_n) = E_V^x [u_h(x_{n+1}) | \mu_0^{nh}];$$

also from (1.48)

$$L_h(x_n, v_n) = \int_0^h e^{-\alpha s} \phi^{v_n}(s) L_{v_n}(x_n) ds.$$

Collecting results we deduce from (1.55), after taking the mathematical expectation and multiplying by $e^{-\alpha nh}$,

$$E_V^x e^{-\alpha nh} u_h(x_n) \leq E_V^x e^{-\alpha nh} L_h(x_n, v_n) + E_V^x e^{-\alpha(n+1)h} u_h(x_{n+1}).$$

Summing from $n = 0$ to $N - 1$, we obtain

$$u_h(x) \leq E_V^x \sum_{n=0}^{N-1} e^{-\alpha nh} L_h(x_n, v_n) + E_V^x e^{-\alpha Nh} u_h(x_N).$$

Letting $N \rightarrow \infty$ and remembering that u_h is bounded, we obtain

$$(1.57) \quad u_h(x) \leq J^x(V).$$

Let now $\hat{v}(x)$ be such that

$$u_h(x) = \int_0^h e^{-\alpha s} \phi^{\hat{v}(x)}(s) L_{\hat{v}(x)}(x) ds + e^{-\alpha h} \phi^{\hat{v}(x)}(h) u_h(x) \quad \forall x.$$

We can find $\hat{v}(x)$ Borel. To \hat{v} we associate \hat{V} in W_h as follows:

$$\hat{V} = (\hat{v}_0(x), \dots, \hat{v}_n(\omega), \dots),$$

with $\hat{v}_0(x) = \hat{v}(x)$, $\hat{v}_n(\omega) = \hat{v}(nh; \omega)$.

A calculation similar to that made above shows that $u_h(x) = J^x(\hat{V})$, which completes the proof of (1.52). \square

We can now give the interpretation of the maximum solution of (1.6). Let us define

$$(1.58) \quad W_q = W_{1/2^q}.$$

Then $W_{q+1} \supset W_q$, and hence

$$\min_{V \in W_q} J^x(V) \geq \min_{V \in W_{q+1}} J^x(V),$$

and we recover the fact that $u_q \geq u_{q+1}$.

THEOREM 1.5. *We make the assumptions of Theorem 1.1 and (1.2). Then the maximum solution of (1.6) can be interpreted as*

$$(1.59) \quad u(x) = \inf_{V \in \bigcup_q W_q} J^x(V).$$

Proof. Let \tilde{u} be the right-hand side of (1.59). Since $u_q \geq \tilde{u}$ we have $u \geq \tilde{u}$.

On the other hand, we know that

$$u \leq u_q \leq J^x(V) \quad \forall V \in W_q;$$

hence

$$u(x) \leq J^x(V) \quad \forall V \in \bigcup_q W_q,$$

which implies $u \leq \tilde{u}$, hence the result. \square

If E is compact, it follows from Theorem 1.3 that $u_h \rightarrow u$ in C ; then it is easy to check that we have

COROLLARY 1.5. *Under the assumptions of Theorem 1.5, if in addition E is compact, then $\forall \varepsilon > 0$. There exists an ε -optimal control in $\bigcup_q W_q$, i.e., $\forall \varepsilon > 0, \exists V \in \bigcup_q W_q$ such that $J^x(V) \leq u(x) + \varepsilon \quad \forall x$.*

2. Stopping time problem.

2.1. Assumptions—notation. Let (E, \mathcal{E}) and B, C be as in § 1.1. We consider a Markov semigroup on B , $\phi(t)$. We will assume

$$(2.1) \quad t \rightarrow \phi(t)\varphi(x) \text{ is continuous from } (0, \infty) \rightarrow R \quad \forall x \text{ fixed, } \forall \varphi \in B.$$

Let

$$(2.2) \quad \psi \in B,$$

$$(2.3) \quad L \in B \text{ such that } t \rightarrow \phi(t)L(x) \text{ is (Lebesgue) measurable } \forall x \text{ fixed.}$$

If we assume more regularity on ψ, L , namely

$$(2.4) \quad \psi \in C, \quad \int_0^\infty e^{-\alpha t} \phi(t)L \, dt \in C,$$

then we will use a weaker form of (2.1) (i.e., (1.4)),

$$(2.5) \quad t \rightarrow \phi(t)\varphi(x) \text{ is continuous from } (0, \infty) \rightarrow R \quad \forall x \text{ fixed, } \forall \varphi \in C,$$

$$(2.6) \quad \phi(t): C \rightarrow C \quad \forall t > 0.$$

We consider the following problem: to find a maximum solution of the set of inequalities

$$(2.7) \quad \begin{aligned} u &\in B, & u &\leq \psi, \\ u &\leq \int_0^t e^{-\alpha s} \phi(s) L \, ds + e^{-\alpha t} \phi(t) u \quad \forall t \geq 0. \end{aligned}$$

We will study (2.7) by the following discretization procedure:

$$(2.8) \quad u_h = \min \left[\psi, \int_0^h e^{-\alpha t} \phi(t) L \, dt + e^{-\alpha h} \phi(h) u_h \right], \quad u_h \in B.$$

2.2. Convergence result.

THEOREM 2.1. *We assume (1.2), (2.1), (2.2), (2.3) or (1.2), (2.4), (2.5), (2.6). Then the set of functions satisfying (2.7) is not empty and has a maximum element.*

Proof. Let, for $z \in B$,

$$T_h(z) = \min \left[\psi, \int_0^h e^{-\alpha t} \phi(t) L \, dt + e^{-\alpha h} \phi(h) z \right].$$

Then $T_h(z) \in B$. Now when (2.4) and (2.6) are satisfied, $T_h: C \rightarrow C$. Moreover T_h is a contraction in B ; hence (2.8) has one and only one solution. Then we proceed as in the proof of Theorem 1.1. We check that

$$(2.9) \quad z \in B \text{ and } z \leq T_h z \text{ implies } z \leq u_h, \quad u_h \leq u_{2h},$$

$$(2.10) \quad u_h \geq -K, \quad K = \max \left[\|\psi^-\|, \frac{\|L^-\|}{\alpha} \right].$$

Indeed,

$$-K \leq \psi, \quad -K \leq \int_0^\infty e^{-\alpha t} \phi(t) L \, dt - e^{-\alpha h} \phi(h) K;$$

hence $-K \leq u_h$ by (2.9).

We next set $u_q = u_{1/2^q}$ and

$$(2.11) \quad u_q \downarrow u, \quad u \in B, \quad u \geq -K.$$

Clearly $u \leq \psi$. As in the proof of Theorem 1.1, we have

$$(2.12) \quad u \leq \int_0^{j/2^l} e^{-\alpha s} \phi(s) L \, ds + e^{-\alpha j/2^l} \phi\left(\frac{j}{2}\right) u \quad \forall j, l \text{ integer}.$$

Let $t > 0$, take $j = [t2^l] + 1$ and let $l \rightarrow \infty$. By assumption (2.1) we can assert that

$$(2.13) \quad u \leq \int_0^t e^{-\alpha s} \phi(s) L \, ds + e^{-\alpha t} \phi(t) u.$$

Therefore u is an element of (2.7). It is the maximum element since, if \tilde{u} is another element, then clearly $\tilde{u} \leq T_h \tilde{u}$; hence $\tilde{u} \leq u_h$, which implies $\tilde{u} \leq u$.

If we assume (2.4), (2.5), (2.6) instead of (2.1), we cannot pass to the limit as $l \rightarrow \infty$ in (2.12). We proceed as in Theorem 1.1.

We first majorize u by u^q in the right-hand side of (2.12), and use the fact that $u^q \in C$ to pass to the limit as $l \rightarrow \infty$. We obtain (2.13) with u replaced by u^q in the right-hand side, and we can let $q \rightarrow \infty$ to obtain the desired result. \square

2.3. Regularity. In this section we assume that $\phi(t)$ satisfies (1.2), (2.6) and a stronger property than (2.5), namely

$$(2.14) \quad t \rightarrow \phi(t)\varphi \text{ is continuous from } [0, \infty) \text{ into } C \quad \forall \varphi \in C^1.$$

We also assume (2.4) for the data ψ, L . Then our objective is to prove the following.

THEOREM 2.2. *We assume (1.2), (2.6), (2.14) and (2.4). Then the maximum solution of (2.7) belongs to C . Moreover, $u_h \rightarrow u$ in C .*

Although we will not use explicitly the infinitesimal generator $-A$ of $\phi(t)$, we will use the following property of its domain $D(A)(\phi(t))$ considered as an operator from $C \rightarrow C$:

$$(2.15) \quad D(A) = \left\{ g \in C \mid g = \int_0^\infty e^{-\alpha t} \phi(t) G dt, G \in C \right\}, \quad D(A) \text{ is dense in } C.$$

We consider the penalized problem

$$(2.16) \quad u_\varepsilon = \int_0^\infty e^{-\alpha t} \phi(t) \left(L - \frac{1}{\varepsilon} (u_\varepsilon - \psi)^+ \right) dt, \quad u_\varepsilon \in C.$$

Then (2.16) has one and only one solution. Indeed, from Lemma (1.5), (2.16) is equivalent to

$$(2.17) \quad \begin{aligned} u_\varepsilon &= \int_0^\infty e^{-(\alpha+1/\varepsilon)t} \phi(t) \left(L + \frac{1}{\varepsilon} u_\varepsilon - \frac{1}{\varepsilon} (u_\varepsilon - \psi)^+ \right) dt \\ &= \int_0^\infty e^{-(\alpha+1/\varepsilon)t} \phi(t) \left(L + \frac{1}{\varepsilon} z_\varepsilon \wedge \psi \right) dt; \end{aligned}$$

then it is easy to check that T_ε is a contraction and, thus, its unique fixed point is the unique solution of (2.16). To prove that

$$(2.18) \quad u_h \rightarrow u \quad \text{in } C,$$

we will rely on a discretized penalized problem. In the mean time, we will recover convergence properties of u_ε to u as a side effect.

We now introduce the penalized discretized problem

$$(2.19) \quad u_h^\varepsilon = h L_h - \frac{h}{\varepsilon} (u_h^\varepsilon - \psi)^+ + e^{-\alpha h} \phi(h) u_h^\varepsilon, \quad u_h^\varepsilon \in C,$$

where

$$(2.20) \quad L_h = \frac{1}{h} \int_0^h e^{-\alpha t} \phi(t) L dt.$$

LEMMA 2.1. *Problem (2.19) has one and only one solution.*

Proof. We define on C the map

$$(2.21) \quad T_{hz}^\varepsilon = \sum_{n=0}^\infty \left(\frac{\varepsilon}{h+\varepsilon} \right)^{n+1} e^{-n\alpha h} \phi(nh) \left[h L_h + h \frac{\psi \wedge z}{\varepsilon} \right].$$

Let us check that it is a contraction. Indeed,

$$\|T_{hz_1}^\varepsilon - T_{hz_2}^\varepsilon\| \leq \sum_{n=0}^\infty \left(\frac{\varepsilon}{h+\varepsilon} \right)^{n+1} e^{-n\alpha h} \frac{h}{\varepsilon} \|z_1 - z_2\| \leq \frac{\|z_1 - z_2\|}{1 + \varepsilon \frac{1 - e^{-\alpha h}}{h}}.$$

¹ This is equivalent to $\phi(t)\varphi \rightarrow \varphi$ in C as $t \downarrow 0 \quad \forall \varphi \in C$ (cf. Dynkin [4]). Moreover, (2.14) is equivalent to (2.5) if E is compact.

Let u_h^ε be the unique fixed point. Then

$$\begin{aligned} u_h^\varepsilon &= \frac{\varepsilon}{h+\varepsilon} \left(hL_h + h \frac{\psi \wedge u_h^\varepsilon}{\varepsilon} \right) + e^{-\alpha h} \frac{\varepsilon \phi(h) u_h^\varepsilon}{h+\varepsilon} \\ &= \frac{\varepsilon h L_h}{h+\varepsilon} + \frac{h}{h+\varepsilon} u_h^\varepsilon - \frac{h}{h+\varepsilon} (u_h^\varepsilon - \psi)^+ + \frac{\varepsilon}{h+\varepsilon} e^{-\alpha h} \phi(h) u_h^\varepsilon, \end{aligned}$$

from which (2.19) follows. \square

LEMMA 2.2. *We have*

$$(2.22) \quad u_h^\varepsilon \leq u_h^{\varepsilon'} \quad \text{if } \varepsilon \leq \varepsilon'$$

and

$$(2.23) \quad \|u_h^\varepsilon\| \leq K.$$

Proof. We have

$$(2.24) \quad T_h^\varepsilon u_h^{\varepsilon'} = \sum_{n=0}^{\infty} \left(\frac{\varepsilon}{h+\varepsilon} \right)^{n+1} e^{-n\alpha h} \left[hL_h + \frac{h\psi \wedge u_h^{\varepsilon'}}{\varepsilon} \right].$$

But from (2.19)

$$u_h^{\varepsilon'} \frac{\varepsilon}{h+\varepsilon} = \frac{\varepsilon h L_h}{h+\varepsilon} - \frac{h}{\varepsilon'} \frac{\varepsilon}{h+\varepsilon} (u_h^{\varepsilon'} - \psi)^+ + \frac{\varepsilon}{h+\varepsilon} e^{-\alpha h} \phi(h) u_h^{\varepsilon'}$$

or

$$\begin{aligned} u_h^{\varepsilon'} &= \frac{\varepsilon h L_h}{h+\varepsilon} + \frac{h}{h+\varepsilon} u_h^{\varepsilon'} - \frac{h}{\varepsilon'} \frac{\varepsilon}{h+\varepsilon} (u_h^{\varepsilon'} - \psi)^+ + \frac{\varepsilon}{h+\varepsilon} e^{-\alpha h} \phi(h) u_h^{\varepsilon'} \\ &= \frac{\varepsilon h L_h}{h+\varepsilon} + \frac{\varepsilon}{h+\varepsilon} \psi \wedge u_h^{\varepsilon'} + \frac{h}{h+\varepsilon} \left(1 - \frac{\varepsilon}{\varepsilon'} \right) (u_h^{\varepsilon'} - \psi)^+ + \frac{\varepsilon}{h+\varepsilon} e^{-\alpha h} \phi(h) u_h^{\varepsilon'}, \end{aligned}$$

and, thus

$$(2.25) \quad u_h^{\varepsilon'} = \sum_{n=0}^{\infty} \left(\frac{\varepsilon}{h+\varepsilon} \right)^{n+1} e^{-n\alpha h} \phi(nh) \left[hL_h + \frac{h}{\varepsilon} u_h^{\varepsilon'} + h \left(\frac{1}{\varepsilon} - \frac{1}{\varepsilon'} \right) (u_h^{\varepsilon'} - \psi)^+ \right].$$

Hence, comparing (2.24) and (2.25), we see that

$$T_h^\varepsilon u_h^{\varepsilon'} \leq u_h^{\varepsilon'} \quad \text{if } \varepsilon \leq \varepsilon'.$$

Iterating it follows that $(T_h^\varepsilon)^n u_h^{\varepsilon'} \leq u_h^{\varepsilon'}$; hence (2.22).

Let us now notice that $u^0 = \int_0^\infty e^{-\alpha t} \phi(t) L dt$ satisfies

$$(2.26) \quad u^0 = hL_h + e^{-\alpha h} \phi(h) u^0.$$

Then using (2.19) we deduce

$$u^0 - u_h^\varepsilon \geq e^{-\alpha h} \phi(h) (u^0 - u_h^\varepsilon)$$

and iterating

$$u^0 - u_h^\varepsilon \geq e^{-n\alpha h} \phi(nh) (u^0 - u_h^\varepsilon) \rightarrow 0,$$

and therefore

$$(2.27) \quad u^0 \geq u_h^\varepsilon.$$

But

$$(2.28) \quad u^0 \leq \frac{\|L^+\|}{\alpha},$$

hence

$$(2.29) \quad u_h^\varepsilon \leq \frac{\|L^+\|}{\alpha}.$$

On the other hand, we have

$$\begin{aligned} T_h^\varepsilon 0 &= \sum_{n=0}^{\infty} \left(\frac{\varepsilon}{h+\varepsilon} \right)^{n+1} e^{-n\alpha h} \phi(nh) \left(hL_h - \frac{h}{\varepsilon} \psi^- \right) \\ &\geq - \sum_{n=0}^{\infty} \left(\frac{\varepsilon}{h+\varepsilon} \right)^{n+1} e^{-n\alpha h} \left(\|L^-\| \frac{1-e^{-\alpha h}}{\alpha} + \frac{h}{\varepsilon} \|\psi^-\| \right) \geq - \frac{\|L^-\|}{\alpha} - \|\psi^-\| = -K. \end{aligned}$$

Assume now that $z \geq -K$. Then from (2.21)

$$\begin{aligned} T_h^\varepsilon z &\geq \sum_{n=0}^{\infty} \left(\frac{\varepsilon}{h+\varepsilon} \right)^{n+1} e^{-n\alpha h} \phi(nh) \left[-\|L^-\| \frac{1-e^{-\alpha h}}{\alpha} - \frac{h}{\varepsilon} K \right] \\ &\geq - \frac{\varepsilon}{h+\varepsilon(1-e^{-\alpha h})} \left[\|L^-\| \frac{1-e^{-\alpha h}}{\alpha} + \frac{h}{\varepsilon} K \right] \geq -K. \end{aligned}$$

Therefore it follows that $u_h^\varepsilon \geq -K$, which proves (2.23), changing the definition of K . \square

LEMMA 2.3. *We have*

$$(2.30) \quad u_h^\varepsilon \geq u_h.$$

Proof. We have

$$(2.31) \quad T_h^\varepsilon u_h = \sum_{n=0}^{\infty} \left(\frac{\varepsilon}{h+\varepsilon} \right)^{n+1} e^{-n\alpha h} \phi(nh) \left[hL_h + \frac{h\psi \wedge u_h}{\varepsilon} \right]$$

and

$$u_h^\varepsilon \leq hL_h + e^{-\alpha h} \phi(h) u_h$$

or

$$u_h \leq \frac{\varepsilon h L_h}{h+\varepsilon} + \frac{h}{h+\varepsilon} u_h + \frac{\varepsilon}{h+\varepsilon} e^{-\alpha h} \phi(h) u_h;$$

hence

$$(2.32) \quad u_h \leq \sum_{n=0}^{\infty} \left(\frac{\varepsilon}{h+\varepsilon} \right)^{n+1} e^{-n\alpha h} \phi(nh) \left[hL_h + \frac{h}{\varepsilon} u_h \right].$$

Noting that $u_h \leq \psi$, hence, $u_h = \psi \wedge u_h$, it follows from (2.31), (2.32) that $T_h^\varepsilon u_h \geq u_h$, from which we deduce (2.30).

LEMMA 2.4. *Let $\tilde{\psi} \in C$ and \tilde{u}_h^ε be the solution of (2.19) with $\tilde{\psi}$ replacing ψ . Then we have*

$$(2.33) \quad \|\tilde{u}_h^\varepsilon - u_h^\varepsilon\| \leq \|\psi - \tilde{\psi}\|.$$

Proof. Let z, \tilde{z} be in C such that $\|z - \tilde{z}\| \leq \|\psi - \tilde{\psi}\|$. Then from (2.21) it follows that

$$\|T_{h\tilde{z}}^\varepsilon - T_{h\tilde{z}}^\varepsilon\| \leq \sum_{n=0}^{\infty} \left(\frac{\varepsilon}{h+\varepsilon}\right)^{n+1} e^{-n\alpha h} \frac{h}{\varepsilon} \|\psi - \tilde{\psi}\| \leq \|\psi - \tilde{\psi}\|,$$

and hence (2.33). \square

We consider now the case when $\psi \in D(A)$. Therefore we can write

$$\begin{aligned} \psi &= \int_0^\infty e^{-\alpha t} \phi(t) \Lambda \, dt, \quad \Lambda \in C \\ (2.34) \quad &= \int_0^h e^{-\alpha t} \phi(t) \Lambda \, dt + e^{-\alpha h} \phi(h) \psi = h \Lambda_h + e^{-\alpha h} \phi(h) \psi. \end{aligned}$$

We have the following.

LEMMA 2.5. Assume (2.34). Then

$$(2.35) \quad \|u_h^\varepsilon - u_h\| \leq \varepsilon (\|\Lambda\| + \|L\|).$$

Proof. Let us define

$$\tilde{L}_h = L_h - \Lambda_h, \quad \tilde{u}_h^\varepsilon = u_h^\varepsilon - \psi.$$

From (2.19) and (2.34), it follows that

$$\tilde{u}_h^\varepsilon = h \tilde{L}_h - \frac{h}{\varepsilon} (\tilde{u}_h^\varepsilon)^+ + e^{-\alpha h} \phi(h) \tilde{u}_h^\varepsilon.$$

Hence, as already seen,

$$\begin{aligned} \tilde{u}_h^\varepsilon &= \sum_{n=0}^{\infty} \left(\frac{\varepsilon}{h+\varepsilon}\right)^{n+1} e^{-n\alpha h} \phi(nh) \left[h \tilde{L}_h - \frac{h}{\varepsilon} (\tilde{u}_h^\varepsilon)^- \right] \\ &\leq \sum_{n=0}^{\infty} \left(\frac{\varepsilon}{h+\varepsilon}\right)^{n+1} e^{-n\alpha h} \phi(nh) h \tilde{L}_h \leq \frac{h\varepsilon}{h+\varepsilon-\varepsilon e^{-\alpha h}} \|\tilde{L}_h^+\| \leq \varepsilon \|\tilde{L}_h^+\|, \end{aligned}$$

and therefore

$$(2.36) \quad \frac{(\tilde{u}_h^\varepsilon)^+}{\varepsilon} \leq \|\tilde{L}_h^+\|.$$

Now (cf. Lemma 2.2)

$$(2.37) \quad T_h^\varepsilon u_h^{\varepsilon'} - u_h^{\varepsilon'} = - \sum_{n=0}^{\infty} \left(\frac{\varepsilon}{h+\varepsilon}\right)^{n+1} e^{-n\alpha h} \phi(nh) h \left(\frac{1}{\varepsilon} - \frac{1}{\varepsilon'}\right) (u_h^{\varepsilon'} - \psi)^+,$$

and from (2.36) $T_h^\varepsilon u_h^{\varepsilon'} - u_h^{\varepsilon'} \geq -\varepsilon' \|\tilde{L}_h^+\|$ for $\varepsilon < \varepsilon'$.

Now from (2.34) we deduce

$$(2.38) \quad \psi = \sum_{n=0}^{\infty} \left(\frac{\varepsilon}{h+\varepsilon}\right)^{n+1} e^{-n\alpha h} \left[h \Lambda_h + \psi \frac{h}{\varepsilon} \right];$$

thus

$$(T_h^\varepsilon)^2 u_h^{\varepsilon'} - \psi = \sum_{n=0}^{\infty} \left(\frac{\varepsilon}{h+\varepsilon}\right)^{n+1} e^{-n\alpha h} \left[h \tilde{L}_h - \frac{h}{\varepsilon} (T_h^\varepsilon u_h^{\varepsilon'} - \psi)^- \right],$$

and using (2.37) we have

$$-(T_h^\varepsilon u_h^{\varepsilon'} - \psi)^- \cong \tilde{u}_h^{\varepsilon'} - \varepsilon' \|\tilde{L}_h^+\|.$$

Therefore

$$(2.39) \quad (T_h^\varepsilon)^2 u_h^{\varepsilon'} - \psi \cong \sum_{n=0}^{\infty} \left(\frac{\varepsilon}{h + \varepsilon} \right)^{n+1} e^{-n\alpha h} \left[h \tilde{L}_h + h \frac{\tilde{u}_h^{\varepsilon'}}{\varepsilon} - h \frac{\varepsilon'}{\varepsilon} \|\tilde{L}_h^+\| \right].$$

Now from the expression (2.25) of $\tilde{u}_h^{\varepsilon'}$ and (2.34) we can write

$$(2.40) \quad \tilde{u}_h^{\varepsilon'} = \sum_{n=0}^{\infty} \left(\frac{\varepsilon}{h + \varepsilon} \right)^{n+1} e^{-n\alpha h} \left[h \tilde{L}_h + \frac{h}{\varepsilon} \tilde{u}_h^{\varepsilon'} - \frac{h}{\varepsilon} (\tilde{u}_h^{\varepsilon'})^+ \right],$$

which together with (2.39) implies

$$(2.41) \quad (T_h^\varepsilon)^2 u_h^{\varepsilon'} - \psi \cong \tilde{u}_h^{\varepsilon'} - \frac{h\varepsilon'}{\varepsilon} \sum_{n=0}^{\infty} \left(\frac{\varepsilon}{h + \varepsilon} \right)^{n+1} e^{-n\alpha h} \|\tilde{L}_h^+\| \cong \tilde{u}_h^{\varepsilon'} - \varepsilon' \|\tilde{L}_h^+\|.$$

Therefore

$$(T_h^\varepsilon)^2 u_h^{\varepsilon'} - u_h^{\varepsilon'} \cong -\varepsilon' \|\tilde{L}_h^+\|,$$

and by iteration, we obtain

$$(2.42) \quad u_h^\varepsilon - u_h^{\varepsilon'} \cong -\varepsilon' \|\tilde{L}_h^+\| \quad \text{for } \varepsilon < \varepsilon'.$$

Let now u_h^* be the decreasing limit of the sequence u_h^ε , as $\varepsilon \downarrow 0$. Then from (2.42) we obtain $0 \cong u_h^* - u_h^{\varepsilon'} \cong -\varepsilon' \|\tilde{L}_h^+\|$. Therefore $u_h^\varepsilon \rightarrow u_h^*$ in C . Then from (2.19) one easily checks that $u_h^* = u_h$; hence finally,

$$(2.43) \quad \|u_h^\varepsilon - u_h\| \cong \varepsilon \|\tilde{L}_h^+\| \cong \varepsilon (\|\Lambda\| + \|L\|),$$

i.e., (2.35). \square

LEMMA 2.6. *We have*

$$(2.44) \quad u_h^\varepsilon \rightarrow u^\varepsilon \quad \text{in } C \text{ as } h \rightarrow 0.$$

Proof. We have

$$\begin{aligned} u^\varepsilon - u_h^\varepsilon &= \sum_{n=0}^{\infty} e^{-n\alpha h} \phi(nh) \left[\int_0^h e^{-\alpha s} \phi(s) \left(L + \frac{1}{\varepsilon} \psi \wedge u^\varepsilon \right) e^{-nh/\varepsilon} ds \right. \\ &\quad \left. - \left(\frac{\varepsilon}{\varepsilon + h} \right)^{n+1} h \left(L_h + \frac{1}{\varepsilon} \psi \wedge u_h^\varepsilon \right) \right] \\ &= \sum_{n=0}^{\infty} e^{-n\alpha h} \phi(nh) \left[\int_0^h e^{-\alpha s} \phi(s) \left(L + \frac{1}{\varepsilon} \psi \wedge u^\varepsilon \right) e^{-nh/\varepsilon} ds \right. \\ &\quad \left. - \left(\frac{\varepsilon}{\varepsilon + h} \right)^{n+1} h \left(L_h + \frac{1}{\varepsilon} \psi \wedge u_h^\varepsilon \right) \right] \\ &\quad + \sum_{n=0}^{\infty} e^{-n\alpha h} \phi(nh) \left(\frac{\varepsilon}{\varepsilon + h} \right)^{n+1} \frac{h}{\varepsilon} (\psi \wedge u^\varepsilon - \psi \wedge u_h^\varepsilon) \end{aligned}$$

$$= \text{I} + \text{II}.$$

But

$$\begin{aligned}
 |\text{II}| &\leq \frac{\|u^\varepsilon - u_h^\varepsilon\|}{1 + \varepsilon \frac{1 - e^{-\alpha h}}{h}} \\
 |\text{I}| &\leq \sum_{n=0}^{\infty} e^{-n\alpha h} \left\| L_h + \frac{1}{\varepsilon} \psi \wedge u^\varepsilon \right\| \left| e^{-nh/\varepsilon} - \left(\frac{1}{1 + h/\varepsilon} \right)^{n+1} \right| \\
 &\quad + \sum_{n=0}^{\infty} h e^{-n\alpha h} \frac{\left\| \int_0^h (e^{-\alpha s} \phi(s) - I) \frac{\psi \wedge u^\varepsilon}{\phi} \right\|}{h} \\
 &\leq \frac{h}{\varepsilon} \left\| L_h + \frac{1}{\varepsilon} \psi \wedge u^\varepsilon \right\| \sum_{n=0}^{\infty} e^{-n\alpha h} (n+2) + \sum_{n=0}^{\infty} h e^{-n\alpha h} \frac{\left\| \int_0^h (e^{-\alpha s} \phi(s) - I) \frac{\psi \wedge u^\varepsilon}{\varepsilon} \right\|}{h} \\
 &= O_\varepsilon(h) \rightarrow 0 \quad \text{as } h \rightarrow 0 \text{ for fixed } \varepsilon.
 \end{aligned}$$

But then we get

$$\|u^\varepsilon - u_h^\varepsilon\| \leq \left(1 + \frac{h}{\varepsilon(1 - e^{-\alpha h})} \right) O_\varepsilon(h) \leq \left(1 + \frac{e^\alpha}{\varepsilon} \right) O_\varepsilon(h),$$

which implies (2.44). \square

Proof of Theorem 2.2. We first notice that from (2.33) and (2.44) we can assert that

$$(2.45) \quad \|u^\varepsilon - \tilde{u}^\varepsilon\| \leq \|\psi - \tilde{\psi}\|.$$

Now take $h = 1/2^q$, $u_q = u_{1/2^q}$. We know from Theorem 2.1 that $u_q \downarrow u$. Assume that $\psi \in D(A)$; then from (2.35)

$$|u_q^\varepsilon(x) - u_q(x)| \leq \varepsilon(\|\Lambda\| + \|L\|) \quad \forall x.$$

Letting $q \uparrow +\infty$, we deduce

$$(2.46) \quad \|u^\varepsilon - u\| \leq \varepsilon(\|\Lambda\| + \|L\|) \quad \text{if } \psi \in D(A).$$

Similarly, from (2.22) we deduce

$$(2.47) \quad u^\varepsilon \leq u^{\varepsilon'} \quad \text{if } \varepsilon \leq \varepsilon'$$

and from (2.23)

$$(2.48) \quad \|u^\varepsilon\| \leq K.$$

Therefore $u^\varepsilon \downarrow u^*$ as $\varepsilon \downarrow 0$. From (2.46) it clearly follows that $u^* = u$ when $\psi \in D(A)$. In fact, this is true in general. Indeed, from (2.30) we deduce $u^\varepsilon \geq u$. Hence,

$$(2.49) \quad u^* \geq u.$$

But let us show that

$$(2.50) \quad u^\varepsilon \rightarrow u^* \quad \text{in } C.$$

Indeed, take $\psi_N \in D(A)$. $\psi_N \rightarrow \psi$ in C . All quantities depending on ψ_N instead of ψ will be indexed by N . From (2.45) we deduce $\|u^* - u_N^*\| \leq \|\psi - \psi_N\|$, and from (2.46)

$$\|u_N^\varepsilon - u_N^*\| \leq \varepsilon(\|\Lambda_N\| + \|L\|).$$

Therefore

$$\begin{aligned}\|u^\varepsilon - u^*\| &\leq \|u^\varepsilon - u_N^\varepsilon\| + \|u_N^\varepsilon - u_N^*\| + \|u_N^* - u^*\| \\ &\leq 2\|\psi - \psi_N\| + \varepsilon(\|\Lambda_N\| + \|L\|),\end{aligned}$$

which proves (2.50). But then from (2.16) we deduce

$$\int_0^t e^{-\alpha s} \phi(s) (u^* - \psi)^+ ds = 0 \quad \forall t.$$

Dividing by t and letting $t \rightarrow 0$, it follows from assumption (2.14) and the fact that $(u^* - \psi)^+ \in C$ that $(u^* - \psi)^+ = 0$; hence,

$$(2.51) \quad u^* \leq \psi.$$

Since we also have

$$u^\varepsilon \leq \int_0^t e^{-\alpha s} \phi(s) L ds + e^{-\alpha t} \phi(t) u^\varepsilon,$$

we deduce

$$u^* \leq \int_0^t e^{-\alpha s} \phi(s) L ds + e^{-\alpha t} \phi(t) u^*,$$

which combined with (2.51) implies $u^* \leq u$ by definition of u . This and (2.49) prove that

$$(2.52) \quad u^\varepsilon \rightarrow u \quad \text{in } C.$$

Therefore $u \in C$. It remains to prove that (2.18) holds. We have

$$(2.53) \quad u - u_h = u - u_N + u_N - u_{N,\varepsilon} + u_{N,\varepsilon} - u_{N,\varepsilon,h} + u_{N,\varepsilon,h} - u_{N,\varepsilon,h} + u_{N,\varepsilon,h} - u_h.$$

But we have

$$(2.54) \quad \|u - u_N\| \leq \|\psi - \psi_N\|$$

from (2.45) and (2.52),

$$(2.55) \quad \|u_N - u_{N,\varepsilon}\| \leq C_N \varepsilon,$$

from (2.46),

$$(2.56) \quad \|u_{N,\varepsilon} - u_{N,\varepsilon,h}\| \leq O_{N,\varepsilon}(h)$$

by Lemma 2.6, where $O_{N,\varepsilon}(h) \rightarrow 0$ as $h \rightarrow 0$, N, ε being fixed,

$$(2.57) \quad \|u_{N,\varepsilon,h} - u_{N,h}\| \leq C_N \varepsilon,$$

from (2.38), and

$$(2.58) \quad \|u_{N,h} - u_h\| \leq \|\psi - \psi_N\|$$

from (2.33) (after letting $\varepsilon \downarrow 0$ and using the fact that $u_h^\varepsilon \downarrow u_h$ as $\varepsilon \rightarrow 0$). Indeed, from Lemma 2.2 $u_h^\varepsilon \downarrow u_h^*$, and from (2.30) $u_h^* \geq u_h$. But from (2.19), after multiplying by ε and letting $\varepsilon \rightarrow 0$, $u_h^* \leq \psi$, we have

$$u_h^* \leq \int_0^h e^{-\alpha s} \phi(s) L ds + e^{-\alpha h} \phi(h) u_h^*,$$

which implies $u_h^* \leq u_h$ (by 2.9). Hence $u_h^* = u_h$. Collecting results, we may in (2.53) let first h tend to 0, then ε to 0, then N to $+\infty$. We obtain $\|u - u_h\| \rightarrow 0$, which completes the proof of Theorem 2.2. \square

2.4. Probabilistic interpretation. Returning to (2.7), (2.8) and we will give a probabilistic interpretation of both functions u and u_h . We construct on (Ω, μ_0) the unique probability P^x such that $(\Omega, \mu_0, P^x, \mu_0^t, x(t))$ is a Markov process verifying

$$(2.59) \quad E^x \varphi(x(t)) = \phi(t) \varphi(x) \quad \forall \varphi \in B,$$

and by the Markov property,

$$(2.60) \quad E^x[\varphi(x(t)) | \mu^s] = \phi(t-s) \varphi(x(s)) \quad s \leq t \quad \forall \varphi \in B \quad \text{a.s. } P^x.$$

Let θ be a μ^t stopping time; we define

$$(2.61) \quad J^x(\theta) = E^x \left[\int_0^\theta e^{-\alpha t} L(x(t)) dt + e^{-\alpha \theta} \psi(x(\theta)) \right].$$

We consider stopping times of the form $\theta = \nu h$, where ν is a random integer satisfying

$$(2.62) \quad \{\nu = n\} \in \mu^{nh} \quad \forall n.$$

Note that θ is indeed a μ^t stopping time, since

$$\{\theta \leq t\} = \left\{ \nu \leq \frac{t}{h} \right\} = \left\{ \nu \leq \left[\frac{t}{h} \right] \right\} \in \mu[t/h] \cdot h \subset \mu^t.$$

THEOREM 2.3. *We make the assumptions of Theorem 2.1 and (1.2)'. Then the solution of discretized problem (2.8) is given explicitly by*

$$(2.63) \quad u_h(x) = \inf_{\theta = \nu h} J_x(\theta).$$

Moreover, there exists an optimal stopping time $\theta_h = \hat{\nu}_h h$, where

$$(2.64) \quad \hat{\nu}_h = \min_n [u_h(x(nh)) = \psi(x(nh))].$$

Proof. From (2.8) it follows that for all n integer

$$(2.65) \quad u_h \leq \int_0^{nh} e^{-\alpha t} \phi(t) L dt + e^{-\alpha nh} \phi(nh) u_h.$$

By the Markov property for $m \geq n$,

$$\begin{aligned} E^x \left[u_h(x(mh)) e^{-\alpha mh} + \int_{nh}^{mh} e^{-\alpha s} L(x(s)) ds | \mu^{nh} \right] \\ = \phi((m-n)h) u_h(x(nh)) e^{-\alpha mh} + \int_{nh}^{mh} e^{-\alpha s} \phi(s-nh) L(x(nh)) ds, \end{aligned}$$

and from (2.65)

$$\geq e^{-\alpha nh} u_h(x(nh)).$$

By Doob's theorem we can replace m, n by random integers $\nu_1 \leq \nu_2$ satisfying (2.62), namely,

$$e^{-\alpha \nu_1 h} u_h(x(\nu_1 h)) \leq E^x \left[u_h(x(\nu_2 h)) e^{-\alpha \nu_2 h} + \int_{\nu_1 h}^{\nu_2 h} e^{-\alpha s} L(x(s)) ds | \mu^{\nu_1 h} \right],$$

from which follows (with $\nu_1 = 0$, $\nu_2 = \nu$),

$$(2.66) \quad u_h(x) \leq J^x(\nu h).$$

Now from (2.8) we have

$$(2.67) \quad u_h(x(nh)) = \min \left[\psi(x(nh)), \int_0^h e^{-\alpha t} \phi(t) L(x(nh)) dt + e^{-\alpha h} \phi(h) u_h(x(nh)) \right].$$

Now by the Markov property (2.60)

$$\phi(t) L(x(nh)) = E^x[L(x(nh+t)) | \mu^{nh}];$$

hence

$$\int_0^h e^{-\alpha t} \phi(t) L(x(nh)) dt = E^x \left[e^{\alpha nh} \int_{nh}^{(n+1)h} e^{-\alpha s} L(x(s)) ds | \mu^{nh} \right].$$

Therefore (2.67) reads

$$e^{-\alpha nh} u_h(x(nh)) \min \left[e^{-\alpha nh} \psi(x(nh)), E^x \left(\int_{nh}^{(n+1)h} e^{-\alpha s} L(x(s)) ds + e^{-\alpha(n+1)h} u_h(x(n+1)h) \right) | \mu^{nh} \right].$$

Multiplying both sides by $\chi_{n < \hat{\nu}_h}$, since for $n < \hat{\nu}_h$, $u_h(x(nh)) < \psi(x(nh))$, we have

$$(2.68) \quad e^{-\alpha nh} u_h(x(nh)) \chi_{n < \hat{\nu}_h} = \chi_{n < \hat{\nu}_h} E^x \left(\int_{nh}^{(n+1)h} e^{-\alpha s} L(x(s)) ds + e^{-\alpha(n+1)h} u_h(x(n+1)h) \right) | \mu^{nh}.$$

Since $\chi_{n < \hat{\nu}_h}$ is μ^{nh} measurable, we can enter it into the conditional expectation. Taking then expectations on both sides, we deduce

$$E^x e^{-\alpha nh} u_h(x(nh)) \chi_{n < \hat{\nu}_h} = E^x \chi_{n < \hat{\nu}_h} \left(\int_{nh}^{(n+1)h} e^{-\alpha s} L(x(s)) ds + e^{-\alpha(n+1)h} u_h(x(n+1)h) \right).$$

By summing over n , we obtain

$$(2.69) \quad E^x \sum_{n=0}^{\hat{\nu}_h-1} e^{-\alpha nh} u_h(x(nh)) = E^x \left[\int_0^{\hat{\nu}_h h} e^{-\alpha s} L(x(s)) ds + \sum_{n=1}^{\hat{\nu}_h} e^{-\alpha nh} u_h(x(nh)) \right].$$

In fact, in writing (2.69) we have implicitly admitted that $\nu_h \geq 1$. If $\nu_h = 0$, from (2.64), then necessarily, $u_h(x) = \psi(x)$. Therefore we can assert that (2.69) holds for any x such that $u_h(x) < \psi(x)$. It clearly follows from (2.69) that

$$u_h(x) = E^x \left[\int_0^{\hat{\nu}_h h} e^{-\alpha s} L(x(s)) ds + e^{-\alpha \hat{\nu}_h h} u_h(x(\hat{\nu}_h h)) \right].$$

But when $\hat{\nu}_h < \infty$, $u_h(x(\hat{\nu}_h h)) = \psi(x(\hat{\nu}_h h))$, hence

$$(2.70) \quad u_h(x) = J^x(\hat{\nu}_h h),$$

provided again that $u_h(x) < \psi(x)$. When $u_h(x) = \psi(x)$, then P^x a.s. $\nu_h = 0$, hence (2.70) still holds. This completes the proof of the desired results. \square

We write now

$$(2.71) \quad \Theta_q = \left\{ \theta = \frac{\nu}{2^q}, \{ \nu = n \} \in \mu^{n/2^q} \forall n \right\}.$$

Then $\Theta_q \subset \Theta_{q-1}$ and

$$u_q = u_{1/2^q} = \inf_{\Theta_q} J_x(\theta).$$

We recover that the sequence u_q decreases. We can next state:

THEOREM 2.4. *We make the assumptions of Theorem 2.1, and (1.2)'. Then the maximum solution of (2.7) is given explicitly by*

$$(2.72) \quad u(x) = \inf_{\theta \in \bigcup_q \Theta_q} J_x(\theta).$$

Proof. Similar to that of Theorem 1.5. \square

Remark 2.1. In the case when the obstacle ψ and the solution $u \in C$, then in (2.72) the inf can be taken on all stopping times, and it is in fact achieved. This situation is more classical (cf. A. Bensoussan, J. L. Lions [2], M. Robin [8]); therefore we do not describe it. \square

3. Case of implicit obstacles. We consider now problem (2.7) with an implicit obstacle, which is the situation which arises in impulse control, although we will not develop the probabilistic interpretation (cf. A. Bensoussan, J. L. Lions [3], M. Robin [8]). We will assume (1.2), (2.6), (2.14) and

$$(3.1) \quad L \in B \int_0^\infty e^{-\alpha t} \phi(t) L dt \in C, \quad L \geq 0.$$

Let also M be an operator such that

$$(3.2) \quad \begin{aligned} M: C \rightarrow C \text{ is Lipschitz, concave and monotone increasing} \\ (\text{i.e., } M\varphi_1 \leq M\varphi_2 \text{ if } \varphi_1 \leq \varphi_2), M(0) \geq k > 0. \end{aligned}$$

We consider the set of functions satisfying

$$(3.3) \quad \begin{aligned} u \in C, \quad u \leq Mu, \\ u \leq \int_0^t e^{-\alpha s} \phi(s) L ds + e^{-\alpha t} \phi(t) u, \end{aligned}$$

and the corresponding discretized version

$$(3.4) \quad u_h = \min \left[Mu_h, \int_0^h e^{-\alpha t} \phi(t) L dt + e^{-\alpha h} \phi(h) u_h \right], \quad u_h \in C.$$

We then have:

THEOREM 3.1. *We assume (1.2), (2.6), (2.14) and (3.1), (3.2). Then (3.4) has one and only one solution. Moreover, $u_h \rightarrow u$ in C , where u is the maximum element of (3.3).*

Proof. Denote by $\sigma_h(\psi)$ the solution of (2.8) and by $\sigma(\psi)$ the maximum element of (2.7). With our assumptions, Theorem 2.2 applies; hence σ_h, σ are maps from C into C .

From (2.9) one easily checks that

$$(3.5) \quad \sigma_h \text{ is monotone increasing and concave.}$$

Similarly,

$$(3.6) \quad \sigma \text{ is monotone increasing and concave.}$$

We next define maps

$$(3.7) \quad A_h = \sigma_h \circ M,$$

$$(3.8) \quad A = \sigma \circ M.$$

By the assumptions on M , A_h and A are also monotone increasing and concave. Let $\mu \in (0, 1)$ such that $\mu \|u_0\| \leq k$, where u^0 has been defined in (2.26). One then checks that

$$(3.9) \quad A_h(0) \geq \mu u^0,$$

$$(3.10) \quad A(0) \geq \mu u^0.$$

From this and the fact that A, A_h are monotone increasing and concave, it follows that A, A_h have a unique fixed point. This is a general property of increasing, concave operators on bounded functions, due to B. Hanouzet, J. L. Joly [5] (cf., also A. Bensoussan [1], A. Bensoussan, J. L. Lions [3]). Moreover, considering the iterative sequences

$$u_h^{n+1} = A_h(u_h^n), \quad u_h^0 = u^0, \quad u^{n+1} = A(u^n), \quad u^0 \text{ defined before.}$$

Then one has

$$(3.11) \quad \|u_h^n - u_h\| \leq (1 - \mu)^n \frac{\|L\|}{\alpha},$$

$$(3.12) \quad \|u^n - u\| \leq (1 - \mu)^n \frac{\|L\|}{\alpha},$$

where u_h and u are the fixed points of A_h, A . Clearly, u_h is the solution of (3.4). Using the fact that $u^n \downarrow u$, one checks that u is the maximum solution of (2.7).

Now for n fixed by Theorem 2.2 $\|u_h^n - u^n\| \rightarrow 0$ as $h \rightarrow 0$.

This and the uniform estimates (3.11), (3.12) imply that $u_h \rightarrow u$ in C .

Remark 3.1. The new result in Theorem 3.1 is the fact that $u_h \rightarrow u$ in C . Of course, this relies on Theorem 2.2 which is new. \square

4. Examples. We give here some examples of semigroups.

4.1. Diffusions. Consider the diffusion with values in \mathbb{R}^n ,

$$(4.1) \quad dy = g(y) dt + \sigma(y) dw(t), \quad y(0) = x,$$

where

$$(4.2) \quad g, \sigma \text{ are Lipschitz functions and bounded.}$$

We define the second order differential operator

$$(4.3) \quad A = -g \cdot \nabla - \text{tr } a D^2,$$

where $a \equiv \frac{1}{2} \sigma \sigma^*$.

We take $E = \mathbb{R}^n$ and define

$$(4.4) \quad \phi(t)\varphi(x) = E\varphi(y_x(t)).$$

Then (1.2) is satisfied. Let us check (1.3). Indeed, we have

$$(4.5) \quad E|y_x(t) - y_{x'}(t)|^2 \leq |x - x'|^2 e^{C_1 t}.$$

We set

$$z(x, t) = \phi(t)\varphi(x);$$

then if

$$\rho(\delta) = \sup_{|x-x'|\leq\delta} |\varphi(x) - \varphi(x')|,$$

we have

$$|z(x, t) - z(x', t)| \leq \rho(\delta) + 2 \frac{\|\varphi\|}{\delta^2} C_0 |x - x'| e^{C_1 t},$$

which proves that $\phi(t)\varphi$ belongs to C .

We next have $E|y_x(t) - x|^2 \leq Ct$; hence $|z(x, t) - z(x, 0)| \leq \rho(\delta) + 2(\|\varphi\|/\delta^2)Ct$ which proves (1.4).

Let us now prove (1.15). We have, assuming $\varphi \in C^{0,\delta}$,

$$|z(x, t) - z(x', t)| \leq \|\varphi\|_{C^{0,\delta}} E|y_x(t) - y_{x'}(t)|^\delta \leq \|\varphi\|_{C^{0,\delta}} |x - x'| e^{C_1 \delta t/2},$$

which proves the desired result.

If $\varphi \in B$, then for $t > 0$, $z(x, t)$ is a regular function of x, t ; hence, clearly (1.18) is satisfied, as well as the restrictive assumption (2.1).

4.2. Diffusion stopped at the exit of a domain. Let \mathcal{O} be a bounded smooth domain of R^n , and let τ_x be the 1st exit time of the diffusion $y_x(t)$ from \mathcal{O} . We set

$$\phi(t)\varphi(x) = E\varphi(y_x(t \wedge \tau_x)).$$

Then (1.2) is satisfied. Writing $z(z, t) = \phi(t)\varphi(x)$, then z is a solution of the nonhomogeneous Dirichlet problem

$$\begin{aligned} \frac{\partial z}{\partial t} + A_z &= 0, \\ (4.6) \quad z|_{\Sigma} &= \varphi, \quad \Sigma = \partial\mathcal{O} \times (0, T), \\ z(x, 0) &= \varphi. \end{aligned}$$

From results on P.D.E., one can assert that (1.3), (1.4) are verified. Concerning the Hölder property (1.15), although we can assert that $z(x, t)$ is Hölder in x , the precise estimate (1.15) does not seem to be correct. It is convenient for this type of problem to introduce subspaces of B, C , namely

$$B_0 = \{\varphi \in B, \varphi|_{\Gamma} = 0\}, \quad \Gamma = \partial\mathcal{O}, \quad C_0 = \{\varphi \in C, \varphi|_{\Gamma} = 0\}.$$

Then clearly $\phi(t): B_0 \rightarrow B_0$, and $z(x, t)$ is a solution of

$$(4.7) \quad \frac{\partial z}{\partial t} + Az = 0, \quad z|_{\Sigma} = 0, \quad z(x, 0) = \varphi(x).$$

In that case z is a regular function of x, t for $t > 0$. Hence (1.18) and (2.1) are verified for $\varphi \in B_0$. Now consider $L_v \in B_0, \psi \in B$ or $C, \psi|_{\Gamma} \geq 0$, then the maximum solution of (1.6) belongs to B_0 . So does the maximum element of (2.7). This follows from the fact that this property is verified for the discretized problem. \square

One can also replace the operator A by an operator $A - L$, where

$$L\varphi(x) = \int_{R^n} (\varphi(x+z) - \varphi(x) - \nabla\varphi(x) \cdot z\chi_{|z|\leq 1}) b(x, z) m(dz)$$

or

$$\int_{R^n} (\varphi(x + \gamma(x, z)) - \varphi(x) - \nabla \varphi \cdot \gamma(x, z)) m(dz),$$

provided that b, γ are Lipschitz in x , uniformly with respect to z and b bounded, γ linear growth in z . By results on the Cauchy problem for $A - L$, and on stochastic differential equations with jumps (cf. Skorokhod [10], Bensoussan, Lions [3]) one can obtain analogous results for the corresponding Markov semigroup.

REFERENCES

- [1] A. BENSOUSSAN, *On the semigroup approach to variational and quasi-variational inequalities*, Proc. of the 1st Franco-South East Asian Conference on Mathematical Sciences, 1978.
- [2] A. BENSOUSSAN AND J.-L. LIONS, *Applications des inéquations variationnelles en contrôle stochastique*, Dunod, Paris, 1978.
- [3] A. BENSOUSSAN AND J.-L. LIONS, *Contrôle impulsif et inéquations quasi-variationnelles*, to be published.
- [4] E. DYNKIN, *Theory of Markov Processes*, Prentice-Hall, Englewood Cliffs, NJ., 1961.
- [5] B. HANOUEZ AND J. L. JOLY, *Convergence uniforme des itérées définissant la solution d'une inéquation quasi-variationnelle*, Note C.R.A.S., Paris, 1978.
- [6] J. NEVEU, *Bases mathématiques de calcul des probabilités*, Masson, Paris, 1970.
- [7a] M. NISIO, *Some remarks on stochastic optimal controls*, 3rd USSR-Japan symposium on Probability Theory, 1975, Lecture Notes in Mathematics, 550, Springer-Verlag, New York, 1976.
- [7b] ———, *On a non linear semi-group attached to stochastic optimal control*, Publ. Res. Inst. Math. Sci. (RIMS), 12 (1976), pp. 513–537.
- [7c] ———, *On stochastic optimal controls and envelopes of Markovian semi-groups*, Proc. International Symposium on Stochastic Differential Equations, Kyoto, 1976, pp. 297–325.
- [8] M. ROBIN, *Contrôle impulsif des processus de Markov*, Thèse, Université de Paris IX, 1978.
- [9] M. SHIRYAEV, *Optimal Stopping Rules*, Springer-Verlag, Berlin, 1978.
- [10] A. V. SKOROKHOD, *Studies in the Theory of Random Processes*, Addison Wesley, Reading, MA, 1965.

EXISTENCE AND UNIQUENESS OF OPTIMAL CONTROLS FOR A QUASILINEAR PARABOLIC EQUATION

THOMAS I. SEIDMAN[†] AND HONG-XING ZHOU[‡]

Abstract. We consider the quasilinear parabolic equation $\dot{y} + \mathbf{A}y + \mathbf{F}(y) = \varphi$ on $Q := (0, T) \times \Omega$. Viewing φ as a control, we seek to minimize $J(\varphi) := \|\varphi - \hat{\varphi}\|^2 + \lambda \|y - \hat{y}\|^2 + \mu \|y(T) - \hat{\eta}\|^2$. Under suitable hypotheses it is shown that one has existence of an optimal control φ_* and that this satisfies the appropriate optimality system. Further, for small data (so $J_* := \min J$ is small enough) one has global uniqueness and continuous dependence on the data.

1. Introduction. We will be considering quasilinear parabolic partial differential equations of the form

$$(1.1) \quad \dot{y} + \mathbf{A}y + \mathbf{F}(y) = \varphi \quad \text{on } Q := (0, T) \times \Omega.$$

Here Ω is a region in \mathbb{R}^m , \mathbf{A} is a suitable elliptic operator and \mathbf{F} is a Nemytsky operator on $L^2(\mathcal{Q})$:

$$\mathbf{F}: y \mapsto F(\cdot, y(\cdot)).$$

(It should cause no confusion to use the same symbol \mathbf{F} for the function $\mathbf{F}: \mathcal{Q} \times \mathbb{R} \rightarrow \mathbb{R}$ and for the induced Nemytsky operator.) We state the following hypotheses for F :

$$(1.2) \quad F(\cdot, 0) = 0;$$

F satisfies Carathéodory conditions and a one-sided Lipschitz condition

$$-2[F(\cdot, r) - F(\cdot, r')] \leq \gamma[r - r'] \quad \text{for } r \leq r';$$

\mathbf{F} maps $\mathcal{H} := L^2(\mathcal{Q})$ into itself.

$$(1.3) \quad F(x, \cdot) \text{ is differentiable (a.e. } x \in \Omega) \text{ with } G := \partial F / \partial r \text{ Lipschitz continuous:}$$

$$|G(\cdot, r) - G(\cdot, r')| \leq K|r - r'|.$$

We also introduce the cost functional:

$$(1.4) \quad \mathcal{J} = \mathcal{J}(\varphi) := \|\varphi - \hat{\varphi}\|^2 + \lambda \|y - \hat{y}\|^2 + \mu \|y(T) - \hat{\eta}\|^2,$$

with the norms taken in \mathcal{H} , \mathcal{H} and $\mathfrak{H} := L^2(\Omega)$.

We have not yet mentioned boundary conditions or initial data for (1.1). Suppose \mathbf{B} is a boundary operator suitably related to \mathbf{A} for consideration of

$$(1.5) \quad \begin{aligned} \dot{y} + \mathbf{A}y + \mathbf{F}(y) &= \varphi && \text{on } \mathcal{Q}, \\ \mathbf{B}y &= y^* \text{ (given)} && \text{on } \Sigma := (0, T) \times \partial\Omega, \\ y(0) &= \eta \text{ (given)} && \text{on } \Omega \text{ at } t = 0. \end{aligned}$$

Then we let y_0 be the solution of the *linear* problem

$$(1.6) \quad \begin{aligned} \dot{y}_0 + \mathbf{A}y &= \hat{\varphi} && \text{on } \mathcal{Q}, \\ \mathbf{B}y_0 &= y^* && \text{on } \Sigma, \\ y_0 &= \eta && \text{on } \Omega \text{ at } t = 0. \end{aligned}$$

* Received by the editors April 2, 1981, and in revised form December 14, 1981.

[†] Université de Nice, France, and Department of Mathematics, University of Maryland Baltimore County, Baltimore, Maryland 22128.

[‡] Department of Mathematics, University of Maryland Baltimore County, Baltimore, Maryland 22128, and Shandong University, People's Republic of China.

If we set

$$(1.7) \quad \begin{aligned} \bar{F}(\cdot, r) &:= F(\cdot, r + y_0(\cdot)) - F(\cdot, y_0(\cdot)), & \tilde{\varphi} &:= -F(\cdot, y_0(\cdot)), \\ \tilde{y} &:= \hat{y} - y_0, & \tilde{\eta} &:= \hat{\eta} - y_0(T), & \bar{y} &:= y - y_0, & \bar{\varphi} &:= \varphi - \hat{\varphi}, \end{aligned}$$

then (1.4), (1.5) reduce to

$$(1.8) \quad \mathcal{J} = \mathcal{J}(\varphi) := \|\varphi\|^2 + \lambda \|y - \tilde{y}\|^2 + \mu \|y(T) - \tilde{\eta}\|^2,$$

$$\dot{y} + \mathbf{A}y + \mathbf{F}(y) = \tilde{\varphi} + \varphi \quad \text{on } \mathcal{Q},$$

$$(1.9) \quad \mathbf{B}y = 0 \quad \text{on } \Sigma,$$

$$y(0) = 0 \quad \text{on } \Omega \text{ at } t = 0.$$

(We have omitted the bars on \bar{F} , \bar{y} , $\bar{\varphi}$ in (1.8), (1.9) for simplicity of future reference.) Note that \bar{F} satisfies $\bar{F}(\cdot, 0) = 0$ even if the original F did not; clearly, \bar{F} satisfies (1.2), (1.3) if F satisfied the other conditions.

We will be keeping Ω , T , \mathbf{A} , \mathbf{B} , λ , μ fixed so that the *data* of the problem,

minimize \mathcal{J} given by (1.8) with y given by (1.9),

consists of specification of

$$(1.10) \quad \begin{aligned} &\mathbf{F} \text{ satisfying (1.2)–and possibly also (1.3);} \\ &\tilde{y} \text{ in } \mathcal{H}, \tilde{\eta} \text{ in } \mathfrak{H} := L^2(\Omega); \tilde{\varphi} \text{ in } \mathcal{H}. \end{aligned}$$

When we speak of “continuous dependence on the data”, etc., we mean the dependence for the new problem on *this* data. Clearly, however, *existence, uniqueness and continuous dependence on data* for the original problem given by (1.4), (1.5) will be equivalent to this provided y^* , η are to be topologized so that $y_0, y_0(T)$ depend continuously in $\mathcal{H}, \mathfrak{H}$ on these.

From (1.9) we will henceforth always be working with homogeneous boundary conditions, and we absorb this in the specification of (the domain of) \mathbf{A} . We assume that we can introduce a Banach space \mathfrak{V} of functions on Ω satisfying the boundary conditions such that

$$\mathfrak{V} \hookrightarrow \mathfrak{H} \hookrightarrow \mathfrak{V}^*, \quad \mathbf{A} : \mathfrak{V} \hookrightarrow \mathfrak{V}^*$$

with

$$(1.11) \quad \|v\|_{\mathfrak{V}}^2 := \langle \mathbf{A}v, v \rangle \quad (v \in \mathfrak{V}).$$

Remark. Note that if one has

$$\langle \mathbf{A}v, v \rangle + \alpha \|v\|_{\mathfrak{H}}^2 \geq c \|v\|_{\mathfrak{V}}^2 \quad (v \in \mathfrak{V}),$$

then one can shift a term αv from F to $\mathbf{A}v$ (this may change the value of γ in (1.2) but otherwise leaves the hypotheses unchanged) and *then* introduce (1.11) as an equivalent norm for \mathfrak{V} .

We mention that there will be occasion later to impose the condition that \mathfrak{V} embeds in $L^4(\Omega)$:

$$(1.12) \quad \mathfrak{V} \subset L^4(\Omega) \quad \text{so } \|v\|_4 \leq c_* \|v\|_{\mathfrak{V}} \quad \text{for } v \in \mathfrak{V}.$$

Our basic assumption on \mathbf{A} is that:

The solution map of the linear problem

$$(1.13) \quad \dot{y} + \mathbf{A}y = \varphi \quad \text{on } \mathcal{Q} \quad y(0) = 0 \quad \text{on } \Omega \text{ at } t = 0,$$

is a compact linear map: $\varphi \mapsto y: \mathcal{H} \rightarrow \mathcal{H}$.

For convenience we also assume \mathbf{A} is selfadjoint with respect to \mathfrak{H} .

It will be convenient to introduce the spaces $\mathcal{X} := L^\infty((0, T) \rightarrow \mathfrak{H})$, $\mathcal{V} := L^2((0, T) \rightarrow \mathfrak{H})$, $\mathcal{W} := \mathcal{X} \cap \mathcal{V}$ and to set

$$\|y\|_{\mathcal{W}} := \max \{\|y\|_{\mathcal{X}}, \|y\|_{\mathcal{V}}\}.$$

In the next section we will consider the map $\mathbf{Y} := \varphi \mapsto y$ defined by (1.9) (with $\tilde{\varphi} = 0$) and will show it is well defined and continuously Fréchet differentiable from \mathcal{H} to \mathcal{W} . The form of the derivative \mathbf{Y}' can be anticipated by the obvious formal calculation. Thus, $\mathbf{Y}'(\varphi): h \mapsto z$, with z determined by the linear parabolic problem

$$(1.14) \quad \begin{aligned} \dot{z} + \mathbf{A}z + \mathbf{G}(y)z &= h \quad \text{on } \mathcal{Q}, \\ z(0) &= 0 \quad \text{on } \Omega \text{ at } t = 0, \end{aligned}$$

where $\mathbf{G}(y)$ is given by the Nemytsky operator induced by $G := \partial F / \partial r$, as in (1.3), for $y = \mathbf{Y}(\varphi)$.

It is not difficult to show the existence of optimal controls, minimizing \mathcal{J} . (Indeed, an argument of similar form [4] shows this in the technically more delicate case of boundary control.) The differentiability of \mathbf{Y} permits the characterization of the optimal control φ_* by the *necessary conditions* (corresponding to $\mathcal{J}'(\varphi_*) = 0$):

$$(1.15) \quad \begin{aligned} -\dot{\varphi} + \mathbf{A}\varphi + \mathbf{G}(y)\varphi &= \lambda(\tilde{y} - y) \quad \text{on } \mathcal{Q}, \\ \varphi(T) &= \mu[\hat{\eta} - y(T)] \quad \text{on } \Omega \text{ at } t = T, \end{aligned}$$

where now $y = \mathbf{Y}(\varphi + \tilde{\varphi})$. The coupled system (1.9), (1.15) is called an *optimality system* for \mathcal{J} . This will be the basis of the argument in the third section for uniqueness and continuous dependence of φ_* .

2. The control problem. Before proceeding with investigation of the solution map \mathbf{Y} , we introduce a technical lemma generalizing the Gronwall inequality.

LEMMA 2.1. *Let $\alpha, \beta, \gamma + 1$ be nonnegative constants and suppose $\nu(\cdot)$ is integrable with*

$$(2.1) \quad 0 \leq \nu(t) \leq \alpha^2 + 2\beta \left[\int_0^t \nu(s) ds \right]^{1/2} + \gamma \int_0^t \nu(s) ds$$

for $t \geq 0$. Then setting $C_\gamma := \exp[(\gamma + 1)T/2]$, one has

$$(2.2) \quad \sup_{[0, T]} \nu \leq C_\gamma^2 (\alpha^2 + \beta^2).$$

Proof. Since

$$2\beta \left[\int_0^t \nu \right]^{1/2} \leq \beta^2 + \int_0^t \nu,$$

(2.1) implies

$$0 \leq \nu(t) \leq (\alpha^2 + \beta^2) + (\gamma + 1) \int_0^t \nu(s) ds.$$

Applying the usual Gronwall inequality to this gives

$$\nu(t) \leq (\alpha^2 + \beta^2) e^{(\gamma+1)t} \quad \text{for } t \geq 0,$$

and (2.2) follows immediately from this. \square

Suppose, now, we are given \mathbf{F} satisfying (1.2). We wish to define $\mathbf{Y}: \varphi \mapsto y$ as a map: $\mathcal{H} \rightarrow \mathcal{W}$ by

$$(2.3) \quad \dot{y} + \mathbf{A}y + \mathbf{F}(y) = \varphi \quad \text{on } \mathcal{Q}, \quad y(0) = 0 \quad \text{on } \Omega \text{ at } t = 0.$$

LEMMA 2.2. *For \mathbf{F} satisfying (1.2) and φ in \mathcal{H} , the problem (2.3) has a solution y in \mathcal{W} with*

$$(2.4) \quad \|y\|_{\mathcal{W}} \leq C_{\gamma} \|\varphi\|_{\mathcal{H}}.$$

Proof. It is known [1] that the hypothesis $\mathbf{F}: \mathcal{H} \rightarrow \mathcal{H}$ in (1.2) implies that the Nemytsky operator is bounded and continuous. For z in \mathcal{H} define a map $\mathbf{T}: z \mapsto \mathbf{F}(z) \mapsto v$ by

$$\dot{v} + \mathbf{A}v = \varphi - \mathbf{F}(z) \quad \text{on } \mathcal{Q}, \quad z(0) = 0.$$

Clearly, as $z \mapsto \mathbf{F}(z)$ is continuous, the assumption (1.13) ensures that \mathbf{T} is continuous and compact. By the Leray–Schauder theorem, \mathbf{T} will have a fixed point—obviously a solution of (2.3)—in $\mathcal{B}_r := \{z \in \mathcal{H}: \|z\|_{\mathcal{H}} \leq r\}$ provided $\theta \mathbf{T}$ has no fixed point on $\partial \mathcal{B}_r$ for $0 \leq \theta < 1$, i.e., provided $\|y\|_{\mathcal{H}} \neq r$ whenever y satisfies

$$\dot{y} + \mathbf{A}y = \theta \varphi - \theta \mathbf{F}(y) \quad \text{on } \mathcal{Q}, \quad y(0) = 0.$$

Multiplying by $2y$ and integrating over $(0, t) \times \Omega$, this gives

$$\begin{aligned} \nu(t) &:= \|y(t)\|_{\mathfrak{H}}^2 + 2 \int_0^t \|y\|_{\mathfrak{H}}^2 = 2\theta \int_0^t \langle y, \varphi \rangle_{\mathfrak{H}} + 2\theta \int_0^t \langle y - 0, \mathbf{F}(y) - \mathbf{F}(0) \rangle_{\mathfrak{H}} \\ &\leq 2\theta \|\varphi\|_{\mathcal{H}} \left[\int_0^t \|y\|_{\mathfrak{H}}^2 \right]^{1/2} + \theta \gamma \int_0^t \|y\|_{\mathfrak{H}}^2, \end{aligned}$$

using (1.2). This has the form of (2.1) with $\alpha = 0$, $\beta = \theta \|\varphi\|_{\mathcal{H}}$, γ as in (1.2). Thus,

$$\|y\|_{\mathcal{W}} \leq \theta C_{\gamma} \|\varphi\|_{\mathcal{H}}$$

for any fixed point y of $\theta \mathbf{T}$ ($0 \leq \theta \leq 1$). Since $\|y\|_{\mathcal{H}}^2 \leq T \|y\|_{\mathcal{H}}^2$ one sees that one cannot have $\|y\|_{\mathcal{H}} = r$ if $r > \sqrt{T} C_{\gamma} \|\varphi\|_{\mathcal{H}}$. This gives both existence and the estimate (2.4). \square

LEMMA 2.3. *The problem (2.3) defines a uniformly Lipschitz continuous map $\mathbf{Y}: \varphi \mapsto y: \mathcal{H} \mapsto \mathcal{W}$ with*

$$(2.5) \quad \|\mathbf{Y}(\varphi_1) - \mathbf{Y}(\varphi_2)\|_{\mathcal{W}} \leq C_{\gamma} \|\varphi_1 - \varphi_2\|_{\mathcal{H}}.$$

Proof. Let φ_1, φ_2 be two controls and y_1, y_2 solutions of the corresponding problems (2.3); set $\psi := \varphi_1 - \varphi_2$, $z := y_1 - y_2$. Subtracting the equations one obtains

$$(2.6) \quad \begin{aligned} \dot{z} + \mathbf{A}z &= \psi - [\mathbf{F}(y_1) - \mathbf{F}(y_2)] \quad \text{on } \mathcal{Q}, \\ z(0) &= 0 \quad \text{on } \Omega \text{ at } t = 0. \end{aligned}$$

Then, multiplying by $2z$ and integrating over $(0, t) \times \Omega$,

$$\nu(t) := \|z(t)\|_{\mathfrak{H}}^2 + 2 \int_0^t \|z\|_{\mathfrak{H}}^2 \leq 2\|\psi\|_{\mathcal{H}} \left[\int_0^t \|z\|_{\mathfrak{H}}^2 \right]^{1/2} + 2\gamma \int_0^t \|z\|_{\mathfrak{H}}^2,$$

much as in the previous proof. As there, this implies by (2.2) that $\|z\|_{\mathcal{W}} \leq C_\gamma \|\psi\|_{\mathcal{X}}$. In particular, of course, if $\psi = 0$ one has $z = 0$, so \mathbf{Y} is single-valued. Note that the uniform Lipschitz constant C_γ depends only on T, γ and not otherwise on Ω or \mathbf{A} (although the form of the norm in \mathcal{V} does depend on \mathbf{A} through the modulus of ellipticity and on Ω through the constant of the Poincaré inequality). \square

We next wish to prove the differentiability of $\mathbf{Y}: \mathcal{H} \rightarrow \mathcal{W}$. In the course of the argument, besides needing (1.3) for the first time, we are faced with estimating the integral of a triple product involving solutions of equations like (2.3) and (1.14). These solutions will be in \mathcal{W} , and we are led to impose the condition (1.12). (While we will work directly with the $L^2(\Omega)$ and $L^4(\Omega)$ norms ($\|\cdot\|_{\mathfrak{L}}$ and $\|\cdot\|_4$), note that interpolation between $\mathcal{X}_0 = L^\infty((0, T) \rightarrow \mathfrak{L})$ and $\mathcal{X}_1 = L^2((0, T) \rightarrow L^4(\Omega))$ gives embedding

$$\mathcal{W} \subset \mathcal{X}_\theta := L^{2/\theta}((0, T) \rightarrow L^{4/(2-\theta)}(\Omega)) \quad (0 \leq \theta \leq 1).$$

In particular, $\theta = 2/3$ gives $\mathcal{X}_\theta = L^3(\mathcal{Q})$, explaining the integrability of such triple products. More generally, one can show that the triple scalar product is a continuous trilinear functional on $\mathcal{X}_\theta \times \mathcal{X}_\rho \times \mathcal{X}_\sigma$ for $0 \leq \theta, \rho, \sigma \leq 1$ with $\theta + \rho + \sigma = 2$.)

THEOREM 2.4. *Let \mathbf{F} satisfy (1.2), (1.3) and assume (1.12). Then the mapping $\mathbf{Y}: \varphi \mapsto y$ defined by (2.3) is continuously Fréchet differentiable from \mathcal{H} to \mathcal{W} with the derivative $\mathbf{Y}'(\varphi)$ given by (1.14).*

Proof. Given φ, h in \mathcal{H} , let $y := \mathbf{Y}(\varphi)$ and $y_h := \mathbf{Y}(\varphi + h) =: y + z$ (i.e., $z := y_h - y$). Now let z_0 be the solution of (1.14) using this y, h .

Remark. The estimate (2.7) below shows that the linear operator $z \mapsto [\dot{z} + \mathbf{A}z + \mathbf{G}(y)z]$ has closed range in \mathcal{H} , and, as the nullspace is trivial (same estimation!), this shows solvability of (1.14) for arbitrary h in \mathcal{H} .

Multiplying (1.14) by $2z$ and integrating over $(0, t) \times \Omega$ gives

$$\nu(t) := \|z(t)\|_{\mathfrak{L}}^2 + 2 \int_0^t \|z\|_{\mathfrak{L}}^2 \leq 2\|h\|_{\mathcal{H}} \left[\int_0^t \|z\|_{\mathfrak{L}}^2 \right]^{1/2} + \gamma \int_0^t \|z\|_{\mathfrak{L}}^2,$$

observing that (1.2), (1.3) ensure the bound: $-2\mathbf{G}(y) \leq \gamma$. From this, using Lemma 2.1, and from (2.6), we see that

$$(2.7) \quad \|z_0\|_{\mathcal{W}}, \|z\|_{\mathcal{W}} \leq C_\gamma \|h\|_{\mathcal{H}}.$$

This shows that the (obviously linear) map: $h \mapsto z_0$ defined by (1.14) is continuous.

Denoting the map, in anticipation, by $\mathbf{Y}'(\varphi)$, we have shown that $\|\mathbf{Y}'(\varphi)\|_{\mathcal{H} \rightarrow \mathcal{W}} \leq C_\gamma$. Now consider the remainder

$$w := z - z_0 = \mathbf{Y}(\varphi + h) - \mathbf{Y}(\varphi) - \mathbf{Y}'(\varphi)h.$$

To show Fréchet differentiability (and to justify the notation $\mathbf{Y}'(\varphi)$ for the map: $h \mapsto z_0$ defined by (1.14)) we must show that $\|w\|_{\mathcal{W}} = o(\|h\|_{\mathcal{H}})$. A simple manipulation of the equations for y, y_h, z_0 shows that w satisfies

$$(2.8) \quad \dot{w} + \mathbf{A}w + \mathbf{G}(y)w = \psi \quad \text{on } \mathcal{Q}, \quad w(0) = 0 \quad \text{on } \Omega \text{ at } t = 0,$$

where

$$\begin{aligned} \psi(\cdot) &= G(\cdot, y)z - [F(\cdot, y + z) - F(\cdot, y)] \\ &= G(\cdot, y)z - \int_{y(\cdot)}^{y(\cdot) + z(\cdot)} G(\cdot, r) dr \\ &= z \int_0^1 [G(\cdot, y) - G(\cdot, y + sz)] ds. \end{aligned}$$

Using the Lipschitz continuity (1.3) of G , this gives

$$|\psi| \leq |z| \int_0^1 K |sz| ds = \frac{1}{2} K |z|^2.$$

Invoking (1.12) and noting (2.7), we see that $w\psi$ is integrable. We have then

$$(2.9) \quad \begin{aligned} 2 \int_0^t \int_{\Omega} w\psi &\leq K \int_0^t \|w\|_{\mathfrak{S}} \|z\|_4^2 \leq K \left(\sup_s \|w(s)\|_{\mathfrak{S}} \right) \|z\|_{L^2((0,T) \rightarrow L^4(\Omega))}^2 \\ &\leq KC_*^2 \|w\|_{\mathcal{W}} \|z\|_{\mathcal{W}}^2. \end{aligned}$$

Now, multiplying (2.8) by $2w$ and integrating over $(0, t) \times \Omega$,

$$\nu(t) := \|w(t)\|_{\mathfrak{S}}^2 + 2 \int_0^t \|w\|_{\mathfrak{S}}^2 \leq 2 \int_0^t \int_{\Omega} w\psi + \gamma \int_0^t \|w\|_{\mathfrak{S}}^2.$$

Using Lemma 2.1 with $\alpha^2 = 2 \iint |w\psi|$ and $\beta = 0$ (equivalently, using the Gronwall inequality) and noting (2.7), one obtains

$$\|w\|_{\mathcal{W}}^2 \leq C_{\gamma}^2 Kc_*^2 \|w\|_{\mathcal{X}} \|z\|_{\mathcal{W}}^2 \leq C_{\gamma}^4 Kc_*^2 \|w\|_{\mathcal{W}} \|h\|_{\mathfrak{S}}^2,$$

which shows $\|w\|_{\mathcal{W}} = O(\|h\|_{\mathcal{X}}^2) = o(\|h\|_{\mathcal{X}})$, as desired. Thus, $\mathbf{Y}'(\varphi)$, defined by (1.14), is indeed the Fréchet derivative of \mathbf{Y} at φ .

To show continuity we estimate $\|\mathbf{Y}'(\varphi_1)h - \mathbf{Y}'(\varphi_2)h\|_{\mathcal{W}}$. Let $z_j := \mathbf{Y}'(\varphi_j)h$, given by (1.14) using $\mathbf{G}(y_j)$ where $y_j := \mathbf{Y}(\varphi_j)$ for $j = 1, 2$. Now set $w := z_1 - z_2 = [\mathbf{Y}'(\varphi_1) - \mathbf{Y}'(\varphi_2)]h$ so

$$\begin{aligned} \dot{w} + \mathbf{A}w + \mathbf{G}(y_1)w &= \psi \quad \text{on } \mathcal{Q}, \\ w(0) &= 0 \quad \text{on } \Omega \text{ at } t = 0, \end{aligned}$$

where

$$\psi := [\mathbf{G}(y_2) - \mathbf{G}(y_1)]z_2.$$

As in deriving (2.9) this gives $|\psi| \leq K|y_1 - y_2||z_2|$ and so

$$(2.10) \quad 2 \int_0^t \int_{\Omega} w\psi \leq Kc_*^2 \|w\|_{\mathcal{X}} \|y_1 - y_2\|_{\mathcal{W}} \|z_2\|_{\mathcal{W}}.$$

From this it follows, as above, that

$$\|w\|_{\mathcal{W}}^2 \leq C_{\gamma}^2 Kc_*^2 \|w\|_{\mathcal{X}} \|y_1 - y_2\|_{\mathcal{W}} \|z_2\|_{\mathcal{W}} \leq C_{\gamma}^4 Kc_*^2 \|w\|_{\mathcal{W}} \|\varphi_1 - \varphi_2\|_{\mathcal{X}} \|h\|_{\mathcal{X}}$$

using (2.6), (2.7). We, thus, obtain the uniform Lipschitz estimate

$$(2.11) \quad \|\mathbf{Y}'(\varphi_1) - \mathbf{Y}'(\varphi_2)\|_{\mathcal{X} \rightarrow \mathcal{W}} \leq C_{\gamma}^4 Kc_*^2 \|\varphi_1 - \varphi_2\|_{\mathcal{X}}. \quad \square$$

We will also, in the next section, have occasion to consider the map $\varphi \mapsto y(T) = \mathbf{Y}(\varphi)(T)$ defined by (2.3). Since the map: $y \mapsto y(t)$ is clearly continuous from \mathcal{X} to \mathfrak{S} , it is clear that as an immediate corollary of Lemma 2.3 and Theorem 2.4 one has:

COROLLARY 2.5. *For each t in $[0, T]$ (in particular at T), the map $\mathbf{Y}_t: \varphi \mapsto y(t)$ defined by (2.3) is—subject to the hypotheses above—well-defined, uniformly Lipschitz continuous from \mathcal{X} to \mathfrak{S} and with a Lipschitz continuous Fréchet derivative $\mathbf{Y}'_t(\varphi)$ given by $\mathbf{Y}'_t(\varphi)h = z(t)$, defined by (1.14). One has the same estimates (2.5), (2.7), (2.11) for $\mathbf{Y}_t, \mathbf{Y}'_t$ as for \mathbf{Y}, \mathbf{Y}' . \square*

The final result of this section will be used later to prove existence of an optimal control.

LEMMA 2.6. *Let \mathbf{F} satisfy (1.2). Then the maps \mathbf{Y}, \mathbf{Y}_t defined by (2.3) are continuous from the weak topology of \mathcal{H} to the strong topology of \mathcal{H} and the weak topology of \mathfrak{S} , respectively.*

Proof. Given φ in \mathcal{H} , suppose $\varphi_k \rightharpoonup \varphi$ (weak convergence in \mathcal{H}). Let $y := \mathbf{Y}(\varphi)$, $y_k = \mathbf{Y}(\varphi_k)$, $f = \mathbf{F}(y)$, $f_k = \mathbf{F}(y_k)$. Note that $\{\varphi_k\}$ is bounded in \mathcal{H} so by Lemma 2.3 $\{y_k\}$ is bounded in \mathcal{W} and so in \mathcal{H} and by (1.2) $\{f_k\}$ is also bounded in \mathcal{H} . Extracting a subsequence if necessary, we may assume $[\varphi_k - f_k] \rightharpoonup [\varphi - \tilde{f}]$ for some \tilde{f} in \mathcal{H} . Now view y_k as the solution of the linear problem

$$(2.12) \quad \begin{aligned} \dot{y}_k + \mathbf{A}y_k &= [\varphi_k - f_k] \quad \text{on } \mathcal{Q}, \\ y_k(0) &= 0 \quad \text{on } \Omega \text{ at } t = 0. \end{aligned}$$

By (1.13) this shows that $\{y_k\}$ converges strongly in \mathcal{H} to the solution \tilde{y} of

$$(2.12') \quad \begin{aligned} \tilde{y} + \mathbf{A}\tilde{y} &= [\varphi - \tilde{f}] \quad \text{on } \mathcal{Q}, \\ \tilde{y}(0) &= 0 \quad \text{on } \Omega \text{ at } t = 0. \end{aligned}$$

Since $y_k \rightarrow \tilde{y}$ in \mathcal{H} the continuity of the Nemytsky operator $\mathbf{F}: \mathcal{H} \rightarrow \mathcal{H}$ gives $f_k = \mathbf{F}(y_k) \rightarrow \mathbf{F}(\tilde{y})$. Thus, $\tilde{f} = \mathbf{F}(\tilde{y})$ and (2.12') becomes exactly (2.3). As we have demonstrated uniqueness for (2.3), this shows $\tilde{y} = \mathbf{Y}(\varphi) =: y$. For the subsequence, then, $y_k \rightarrow y$. As the limit $y := \mathbf{Y}(\varphi)$ is unique, a standard argument gives the same result for the (full) original sequence $\{\varphi_k\}$. This proves the continuity of \mathbf{Y} from \mathcal{H}_w (\mathcal{H} with its weak topology) to \mathcal{H} .

Let \mathbf{Y}^0 be the solution map for (2.3) corresponding to the special case: $F \equiv 0$. By Corollary 2.5 the map \mathbf{Y}_t^0 is continuous from \mathcal{H} to \mathfrak{S} . Since in this special case the map is linear, this implies continuity from \mathcal{H}_w to \mathfrak{S}_w as well. Thus, since we have shown above that $[\varphi_k - f_k] \rightharpoonup [\varphi - \mathbf{F}(y)]$, we have

$$\mathbf{Y}_t(\varphi_k) = y_k(t) = \mathbf{Y}_t^0(\varphi_k - f_k) \rightharpoonup \mathbf{Y}_t^0(\varphi - \mathbf{F}(y)) = y(t) = \mathbf{Y}_t(\varphi).$$

This shows the continuity from \mathcal{H}_w to \mathfrak{S}_w of the (nonlinear) map \mathbf{Y}_t . \square

3. Optimal controls. We suppose now that (with Ω, T, \mathbf{A} as above and $\lambda, \mu \geq 0$ fixed) we are given a Nemytsky operator \mathbf{F} , satisfying (1.2), and functions $\tilde{y}, \tilde{\eta}, \tilde{\varphi}$ in $\mathcal{H}, \mathfrak{S}, \mathcal{H}$, respectively. Let $\tilde{\mathbf{Y}}: \varphi \mapsto y = \mathbf{Y}(\varphi + \tilde{\varphi})$ be the map: $\mathcal{H} \rightarrow \mathcal{W}$ defined by (1.9) and introduce as in (1.8) the functional \mathcal{J} defined by

$$(3.1) \quad \mathcal{J}(\varphi) := \|\varphi\|_{\mathcal{H}}^2 + \lambda \|y - \tilde{y}\|_{\mathcal{H}}^2 + \mu \|y(T) - \tilde{\eta}\|_{\mathfrak{S}}^2, \quad y = \tilde{\mathbf{Y}}(\varphi) = \mathbf{Y}(\varphi + \tilde{\varphi}).$$

We seek to minimize \mathcal{J} , i.e., to find a control φ_* (necessarily in \mathcal{H} by the form of (3.1)) such that

$$(3.2) \quad \mathcal{J}(\varphi_*) = \mathcal{J}_* := \inf \{J(\varphi): \varphi \in \mathcal{H}\}.$$

We refer to such a φ_* as an *optimal control* and call $y_* := \tilde{\mathbf{Y}}(\varphi_*)$ the corresponding *optimal trajectory*.

We begin with the trivial but useful observation that

$$(3.3) \quad \begin{aligned} \mathcal{J}_* &\leq \mathcal{J}(0) = \lambda \|\mathbf{Y}(\tilde{\varphi}) - \tilde{y}\|_{\mathcal{H}}^2 + \mu \|\mathbf{Y}_T(\tilde{\varphi}) - \tilde{\eta}\|_{\mathfrak{S}}^2 \\ &\leq \lambda [\|\tilde{y}\|_{\mathcal{H}} + \sqrt{T} C_\gamma \|\tilde{\varphi}\|_{\mathcal{H}}]^2 + \mu [\|\tilde{\eta}\|_{\mathfrak{S}} + C_\gamma \|\tilde{\varphi}\|_{\mathcal{H}}]^2, \end{aligned}$$

using (3.1) and Lemma 2.2. Thus, we have some a priori bounds; we note that

$$(3.4) \quad \|\varphi_*\|_{\mathcal{H}}^2 \leq \mathcal{J}_*, \quad \lambda \|y_* - \tilde{y}\|_{\mathcal{H}}^2 \leq \mathcal{J}_*, \quad \mu \|y_*(T) - \tilde{\eta}\|_{\mathfrak{S}}^2 \leq \mathcal{J}_*.$$

Using Lemma 2.2 we also have

$$(3.5) \quad \|y_*\|_{\mathcal{W}}, \quad \|y_*(T)\|_{\mathfrak{H}} \leq C_\gamma \|\varphi_* + \tilde{\varphi}\|_{\mathcal{H}} \leq C_\gamma [\|\tilde{\varphi}\|_{\mathcal{H}} + \mathcal{J}_*^{1/2}].$$

THEOREM 3.1. *Let \mathbf{F} satisfy (1.2) and let \tilde{y} , $\tilde{\eta}$, $\tilde{\varphi}$ be in \mathcal{H} , \mathfrak{H} , \mathcal{H} , respectively. Then the functional \mathcal{J} attains its minimum so there exists an “optimal pair” $[\varphi_*, y_*]$ with (3.2) and $y_* = \tilde{\mathbf{Y}}(\varphi_*)$. If in addition \mathbf{F} satisfies (1.3) and \mathbf{Y} is Gâteaux differentiable at φ_* (with differential $\mathbf{Y}'(\varphi_*)$ given by (1.14)), then the optimal pair satisfies the “optimality system” for (3.1):*

$$(3.6) \quad \begin{aligned} \dot{y} + \mathbf{A}y + \mathbf{F}(y) &= \tilde{\varphi} + \varphi && \text{on } \mathcal{Q}, \\ y(0) &= 0 && \text{on } \Omega \text{ at } t = 0, \end{aligned}$$

$$(3.7) \quad \begin{aligned} -\dot{\varphi} + \mathbf{A}\varphi + \mathbf{G}(y)\varphi &= \lambda[\tilde{y} - y] && \text{on } \mathcal{Q}, \\ \varphi(T) &= \mu[\tilde{\eta} - y(T)] && \text{on } \Omega \text{ at } t = T. \end{aligned}$$

Proof. Let $\{\varphi_k\}$ be a minimizing sequence for \mathcal{J} so $\mathcal{J}(\varphi_k) \rightarrow \mathcal{J}_*$. Clearly, $\{\mathcal{J}(\varphi_k)\}$ bounded implies $\{\varphi_k\}$ bounded in \mathcal{H} , and we may assume $\varphi_k \rightarrow \varphi_*$ for some φ_* . By Lemma 2.6 we have $y_k = \tilde{\mathbf{Y}}(\varphi_k) \rightarrow y_* = \tilde{\mathbf{Y}}(\varphi_*)$ in \mathcal{H} and $y_k(T) \rightarrow \mathbf{Y}_*(T)$ in \mathfrak{H}_w . Thus, as the norms for \mathcal{H} , \mathfrak{H} are lower semicontinuous from \mathcal{H}_w , \mathfrak{H}_w to \mathbb{R} , one has

$$\mathcal{J}(\varphi_*) \leq \liminf \mathcal{J}(\varphi_k) = \mathcal{J}_*,$$

so, by the definition of \mathcal{J}_* , one has (3.2).

Now, where \mathbf{Y} is (Gâteaux) differentiable, \mathcal{J} is a composition of differentiable maps. One has

$$\frac{1}{2} \mathcal{J}'(\varphi) : h \mapsto \langle h, \varphi \rangle_{\mathcal{H}} + \lambda \langle \tilde{\mathbf{Y}}'(\varphi)h, y - \hat{y} \rangle_{\mathcal{H}} + \mu \langle \tilde{\mathbf{Y}}'_T(\varphi)h, y(T) - \hat{\eta} \rangle_{\mathfrak{H}}$$

with $y = \tilde{\mathbf{Y}}(\varphi)$. (This uses the obvious embedding $\mathcal{W} \hookrightarrow \mathcal{H}$ for $\mathbf{Y}'(\varphi)h$ and for y .) Taking adjoints one then has

$$(3.8) \quad \frac{1}{2} \mathcal{J}'(\varphi) = \varphi - ([\mathbf{Y}'(\varphi)]^* \lambda [\tilde{y} - y] + [\mathbf{Y}'_T(\varphi)]^* \mu [\tilde{\eta} - y(T)]).$$

To compute the adjoints multiply (1.14) by a function w and integrate over \mathcal{Q} to obtain

$$\begin{aligned} \langle w, h \rangle_{\mathcal{H}} &= \langle w, \dot{z} + \mathbf{A}z + gz \rangle_{\mathcal{H}}, \quad g := \mathbf{G}(y) \\ &= \langle w, z \rangle_{\mathfrak{H}} \Big|_0^T - \langle \dot{w}, z \rangle_{\mathcal{H}} + \int_0^T \langle w, \mathbf{A}z \rangle_{\mathfrak{H}} + \langle gw, z \rangle_{\mathcal{H}} \\ &= \langle w(T), z(T) \rangle_{\mathfrak{H}} + \langle -\dot{w} + \mathbf{A}w + gw, z \rangle_{\mathcal{H}} \\ &= \langle w(T), \mathbf{Y}'_T(\varphi)h \rangle_{\mathfrak{H}} + \langle -\dot{w} + \mathbf{A}w + gw, \mathbf{Y}'(\varphi)h \rangle_{\mathcal{H}}. \end{aligned}$$

Thus, the parenthesized expression on the right-hand side of (3.8) is just the solution φ of (3.7). At the minimizer φ_* of \mathcal{J} , one must, of course, have $\mathcal{J}'(\varphi_*) = 0$, so (3.8) asserts that φ_* satisfies (3.7) with $y = \mathbf{Y}(\varphi_*)$, which is just (3.6). \square

We now note that we may use the system (3.6), (3.7) to obtain estimates for pairs $[\varphi, y]$ by much the same methods as in the previous section.

LEMMA 3.2. *Let \mathbf{F} satisfy (1.2), (1.3). Then any solution pair $[\varphi, y]$ in $\mathcal{H} \times \mathcal{H}$ for the system (3.6), (3.7) must be in $\mathcal{W} \times \mathcal{W}$.*

Proof. Since (3.6) just means $y = \mathbf{Y}(\varphi + \tilde{\varphi})$, Lemma 2.2 gives y in \mathcal{W} . On the other hand, multiplying (3.7) by 2φ and integrating over $(t, T) \times \Omega$ gives, using the

one-sided bound in (1.2) to see that $-2\mathbf{G}(y) \leq \gamma$,

$$\begin{aligned} -\|\varphi\|_{\mathfrak{S}}^2|_t^T + 2 \int_t^T \|\varphi\|_{\mathfrak{B}}^2 &= 2\lambda \iint \varphi(\tilde{y} - y) - 2 \iint \mathbf{G}(y)\varphi^2 \\ &\leq 2\lambda \|\tilde{y} - y\|_{\mathfrak{H}} \left[\int_t^T \|\varphi\|_{\mathfrak{S}}^2 \right]^{1/2} + \gamma \int_t^T \|\varphi\|_{\mathfrak{S}}^2. \end{aligned}$$

Applying Lemma 2.1 with

$$\begin{aligned} \nu(T-t) &:= \|\varphi(t)\|_{\mathfrak{S}}^2 + 2 \int_t^T \|\varphi\|_{\mathfrak{B}}^2, \quad \alpha := \|\varphi(T)\|_{\mathfrak{S}} = \mu \|\tilde{\eta} - y(T)\|_{\mathfrak{S}}, \\ \beta &:= \lambda \|\tilde{y} - y\|_{\mathfrak{H}}, \quad \gamma \text{ as in (1.2)}, \end{aligned}$$

one obtains

$$(3.9) \quad \|\varphi\|_{\mathcal{W}}^2 \leq C_{\gamma}^2 [\lambda^2 \|\tilde{y} - y\|_{\mathfrak{H}}^2 + \mu^2 \|\tilde{\eta} - y(T)\|_{\mathfrak{S}}^2].$$

Since we already know y is in \mathcal{W} , we know $y(T)$ is in \mathfrak{S} , so the right-hand side of (3.9) is finite. Thus, φ is also in \mathcal{W} . \square

Note that (1.3) was used here only to know that $G = \partial F / \partial r$ is defined but not for the Lipschitz bound.

For the final results on uniqueness and continuous dependence on data of optimal controls, we will be forced to impose the requirement that $\lambda \geq \mu^2$. The form of this condition seems somewhat artificial. However, we note that the only inhomogeneity in the relation between λ and μ in \mathcal{J} lies in the possibility of rescaling time by a substitution: $t = \sigma s$; if this is done, the effect is to replace λ by $\lambda_{\sigma} := \sigma^2 \lambda$ and μ by $\mu_{\sigma} := \sigma \mu$ so the condition $\lambda \geq \mu^2$ is precisely equivalent to $\lambda_{\sigma} \geq \mu_{\sigma}^2$. Since the case: $\lambda = \mu = 0$ is of no interest—one obviously minimizes \mathcal{J} uniquely simply by taking $\varphi = 0$ —this analysis shows that, once the case: $\lambda = 0, \mu > 0$ is eliminated, there is no (further) loss of generality in normalizing so $\lambda = 1$.

We will also be forced to impose a further restriction on \mathbf{F} . In terms of the constant γ of (1.2), this requirement is that $\gamma < 2\|\mathbf{E}\|^{-2}$, where \mathbf{E} is the embedding map: $\mathfrak{B} \rightarrow \mathfrak{S}$. A more precise condition (for which the condition on γ is sufficient) is

$$(3.10) \quad -\int_{\Omega} \mathbf{G}(v)w^2 \leq \theta \|w\|_{\mathfrak{B}}^2 \quad \text{for } v, w \in \mathfrak{B}, \quad \theta < 1.$$

It is not difficult to see that (3.10) is equivalent to the strict monotonicity of $[\mathbf{A} + \mathbf{F}]$:

$$\langle [\mathbf{A} + \mathbf{F}](v_1) - [\mathbf{A} + \mathbf{F}](v_2), v_1 - v_2 \rangle \geq (1 - \theta) \|v_1 - v_2\|_{\mathfrak{B}}^2 \quad \text{for } v_1, v_2 \in \mathfrak{B}$$

with $(1 - \theta) > 0$.

THEOREM 3.3. *Suppose (1.12) holds and $\mu < 1$, where \mathcal{J} is normalized so $\lambda = 1$. Suppose also that \mathbf{F} satisfies (1.2), (1.3) and (3.10). Then if*

$$(3.11) \quad \mathcal{J}_* < \rho^2 := [2(1 - \theta) / KC_{\gamma} c_*^2]^2,$$

the optimal control φ_ is unique (indeed, \mathcal{J} then has no other stationary points). This holds, in particular, for sufficiently small data $\tilde{\varphi}, \tilde{y}, \tilde{\eta}$ in $\mathcal{H}, \mathcal{H}, \mathfrak{S}$.*

Proof. Let $[\varphi_*, y_*]$ be an optimal pair. By Theorems 3.1 and 2.4, one knows that such a pair exists and necessarily satisfies (3.6), (3.7) [N.B.: with $\lambda = 1$]. Suppose φ were another stationary point of \mathcal{J} and $y := \mathbf{Y}(\tilde{\varphi} + \varphi)$. Then the pair $[\varphi, y]$ would also

satisfy (3.6), (3.7). Letting $\psi := \varphi - \varphi_*$ and $z = y - y_*$, we then have

$$(3.12) \quad \begin{aligned} \dot{z} + \mathbf{A}z + [\mathbf{F}(y) - \mathbf{F}(y_*)] &= \psi \quad \text{on } \mathcal{Q}, \\ z(0) &= 0 \quad \text{on } \Omega \text{ at } t = 0, \end{aligned}$$

$$(3.13) \quad \begin{aligned} -\dot{\psi} + \mathbf{A}\psi + \mathbf{G}(y)\psi &= -z + [\mathbf{G}(y_*) - \mathbf{G}(y)]\varphi_* \quad \text{on } \mathcal{Q}, \\ \psi(T) &= -\mu z(T) \quad \text{on } \Omega \text{ at } t = T. \end{aligned}$$

Observe that (3.10) gives

$$\begin{aligned} \langle \mathbf{F}(y) - \mathbf{F}(y_*), z \rangle &= \int_{\Omega} \int_0^1 \mathbf{G}(y_* + sz) z \, ds \, z \\ &= \int_0^1 \left(\int_{\Omega} \mathbf{G}(y_* + sz) z^2 \right) ds \cong \int_0^1 [-\theta \|z\|_{\mathfrak{B}}^2] \, ds, \end{aligned}$$

so multiplying (3.12) by z and integrating over Ω gives

$$(3.14) \quad \left(\frac{1}{2} \|z\|_{\mathcal{H}}^2 \right)' + (1 - \theta) \|z\|_{\mathfrak{B}}^2 \leq \langle z, \psi \rangle.$$

Multiplying (3.13) by ψ , integrating over Ω and using (3.10), (1.3) gives

$$(3.15) \quad \left(-\frac{1}{2} \|\psi\|_{\mathfrak{F}}^2 \right)' + (1 - \theta) \|\psi\|_{\mathfrak{B}}^2 \leq -\langle z, \psi \rangle + \int_{\Omega} K |z \varphi_*| |\psi|.$$

Now note that (3.9) gives

$$(3.16) \quad \|\varphi_*(t)\|_{\mathfrak{F}}^2 \leq C_{\gamma}^2 [\|\tilde{y} - y_*\|_{\mathcal{H}}^2 + \mu^2 \|\tilde{\eta} - y_*(T)\|_{\mathfrak{F}}^2] \leq C_{\gamma}^2 \mathcal{F}_*$$

since $\mu < 1$ and $\mathcal{F}_* = \mathcal{F}(\varphi_*)$. Thus, using (1.12),

$$\begin{aligned} \int_{\Omega} |z \varphi_* \psi| &\leq \|z\|_4 \|\psi\|_4 \|\varphi_*\|_{\mathfrak{F}} \leq C_*^2 \|z\|_{\mathfrak{B}} \|\psi\|_{\mathfrak{B}} C_{\gamma} \mathcal{F}_*^{1/2} \\ &\leq [C_*^2 C_{\gamma} \mathcal{F}_*^{1/2} / 2] [\|z\|_{\mathfrak{B}}^2 + \|\psi\|_{\mathfrak{B}}^2]. \end{aligned}$$

Adding (3.14) to (3.15) and using this, one has

$$\frac{1}{2} (\|z\|_{\mathfrak{B}}^2 - \|\psi\|_{\mathfrak{F}}^2)' + (1 - \theta) [\|z\|_{\mathfrak{B}}^2 + \|\psi\|_{\mathfrak{B}}^2] \leq [K c_* C_{\gamma} \mathcal{F}_*^{1/2} / 2] [\|z\|_{\mathfrak{B}}^2 + \|\psi\|_{\mathfrak{B}}^2].$$

We now integrate this over $[0, T]$ noting that (as $\mu^2 < 1$)

$$[\|z\|_{\mathfrak{F}}^2 - \|\psi\|_{\mathfrak{F}}^2]_0^T = (1 - \mu^2) \|z(T)\|_{\mathfrak{F}}^2 + \|\psi(0)\|_{\mathfrak{F}}^2 \geq 0.$$

We obtain

$$(1 - \theta) [\|z\|_{\mathcal{V}}^2 + \|\psi\|_{\mathcal{V}}^2] \leq (\mathcal{F}_*^{1/2} / \rho) (1 - \theta) [\|z\|_{\mathcal{V}}^2 + \|\psi\|_{\mathcal{V}}^2]$$

or, equivalently, $\Gamma [\|z\|_{\mathcal{V}}^2 + \|\psi\|_{\mathcal{V}}^2] \leq 0$ with, using (3.9),

$$(3.17) \quad \Gamma := (1 - \theta) [1 - \mathcal{F}_*^{1/2} / \rho] > 0.$$

This forces $\|z\|_{\mathcal{V}} = \|\psi\|_{\mathcal{V}} = 0$, so $\varphi = \varphi_*$, $y = y_*$. \square

We now wish to consider the dependence of this unique optimal control on the data: $\tilde{\varphi}$, \tilde{y} , $\tilde{\eta}$ and \mathbf{F} as well. We topologize $\tilde{\varphi}$, \tilde{y} , $\tilde{\eta}$ in \mathcal{H} , \mathcal{H} , \mathfrak{H} , of course and will measure perturbations of \mathbf{F} by conditions of the form

$$(3.18) \quad \begin{aligned} |F_1(\cdot, r) - F(\cdot, r)| &\leq \delta(\cdot) + \varepsilon(t)|r| \quad \text{with } \delta \in \mathcal{H} \text{ and } \varepsilon \in L^1(0, T), \\ |G_1(\cdot, r) - G(\cdot, r)| &\leq \bar{\delta}(\cdot) + \bar{\varepsilon}|r| \quad \text{with } \bar{\delta} \in \mathcal{H} \text{ and } \bar{\varepsilon} \in \mathbb{R}. \end{aligned}$$

It does not follow from this that \mathbf{F}_1 will satisfy (1.3), (3.10) when \mathbf{F} does. However, in saying that $\mathbf{F}_1 \rightarrow \mathbf{F}$, we intend that each \mathbf{F}_1 *does* satisfy (1.2), (1.3), (3.10) with θ taken as uniform (but not necessarily K) as $\delta, \varepsilon, \tilde{\delta}, \bar{\varepsilon}$ go to 0 in the appropriate senses.

THEOREM 3.4. *Under the same hypotheses as for Theorem 3.3, including (3.11), the unique optimal pair depends continuously on the data: $\mathbf{F}, \tilde{\varphi}, \tilde{y}, \tilde{\eta}$. That is if $[\varphi, y]$ is the (unique) optimal pair corresponding to $\mathbf{F}, \tilde{\varphi}, \tilde{y}, \tilde{\eta}$ as in Theorem 3.3 and if $\mathbf{F}_k \rightarrow \mathbf{F}$ in the sense described above with $\tilde{\varphi}_k \rightarrow \tilde{\varphi}, \tilde{y}_k \rightarrow \tilde{y}, \tilde{\eta}_k \rightarrow \tilde{\eta}$ in $\mathcal{H}, \mathcal{H}, \mathcal{S}$, respectively, then each corresponding problem with perturbed data has an optimal pair $[\varphi_k, y_k]$ and $[\varphi_k, y_k] \rightarrow [\varphi, y]$ in $\mathcal{W} \times \mathcal{W}$.*

Proof. Under the given assumptions, Theorem 3.3 gives uniqueness for the optimal pair $[\varphi, y]$ which by Theorem 3.1 satisfies the optimality system (3.6), (3.7). By Theorem 3.1 we know that $[\varphi_k, y_k]$ exists and satisfies the corresponding optimality system (3.6)_k, (3.7)_k. Also, we will show next that \mathcal{J}_{k*} (the minimum cost for the perturbed problem) cannot be much greater than \mathcal{J}_* and so satisfies (3.11) for large k . Nevertheless, since we need not have a uniform bound for K (which appears in the definition of ρ), Theorem 3.3 need not apply and we cannot conclude that $[\varphi_k, y_k]$ is unique.

Clearly, $\mathcal{J}_{k*} = \mathcal{J}_k(\varphi_k) \leq \mathcal{J}_k(\varphi)$, where \mathcal{J}_k is the perturbed cost functional (defined, also, through the perturbed equation). Since $\tilde{y}_k, \tilde{\eta}_k$ are close to $\tilde{y}, \tilde{\eta}$, it suffices to show that \bar{y} is close to y where \bar{y} is the solution of the perturbed equation using the unperturbed control φ . Setting $\bar{z} := \bar{y} - y$ we have

$$\begin{aligned} \bar{z} + \mathbf{A}\bar{z} + [\mathbf{F}(\bar{y}) - \mathbf{F}(y)] &= (\tilde{\varphi}_k - \tilde{\varphi}) + [\mathbf{F}(\bar{y}) - \mathbf{F}_k(\bar{y})] \quad \text{on } \mathcal{Q}, \\ \bar{z}(0) &= 0 \quad \text{on } \Omega \text{ at } t = 0. \end{aligned}$$

Multiplying by \bar{z} and integrating over $(0, t) \times \Omega$, using (3.10) and (3.18), one obtains

$$\frac{1}{2} \|\bar{z}(t)\|_{\mathcal{S}}^2 + (1 - \theta) \int_0^t \|\bar{z}\|_{\mathcal{S}}^2 \leq [\|\tilde{\varphi}_k - \tilde{\varphi}\|_{\mathcal{H}} + \|\delta\|_{\mathcal{H}}] \left[\int_0^t \|\bar{z}\|_{\mathcal{S}}^2 \right]^{1/2} + \|\varepsilon\|_1 \|\bar{z}\|_{\mathcal{X}}^2.$$

Applying Lemma 2.1 as earlier (note that $\gamma = 0$ here), this gives

$$\|\bar{z}\|_{\mathcal{X}} \leq (\|\tilde{\varphi}_k - \tilde{\varphi}\|_{\mathcal{H}} + \|\delta\|_{\mathcal{H}})^2 + 2\|\varepsilon\|_1 \|\bar{z}\|_{\mathcal{X}}^2,$$

which bounds \bar{z} when $2\|\varepsilon\|_1 < 1$. This suffices to show that $\limsup \mathcal{J}_{k*} \leq \mathcal{J}_*$. (The result of this theorem, of course, shows $\mathcal{J}_{k*} \rightarrow \mathcal{J}_*$.) Since $\|\varphi_k\|_{\mathcal{H}}^2 \leq \mathcal{J}_k(\varphi_k) = \mathcal{J}_{k*}$, this bounds $\{\varphi_k\}$ uniformly in \mathcal{H} . As in Lemma 2.2, one, thus, has a uniform bound on $\{y_k\}$ in \mathcal{W} . At the same time, as in (3.9), (3.16), the bound on $\{\mathcal{J}_{k*}\}$ gives a uniform bound on $\{\varphi_k\}$ in \mathcal{W} . (Note that these use the uniformity of θ but not of K .)

The argument now is along much the same lines as for Theorem 3.3. Letting $\psi := \varphi_k - \varphi$ and $z := y_k - y$, we have

$$\begin{aligned} (3.19) \quad \dot{z} + \mathbf{A}z + [\mathbf{F}_k(y_k) - \mathbf{F}_k(y)] &= \psi + (\tilde{\varphi}_k - \tilde{\varphi}) + [\mathbf{F}(y) - \mathbf{F}_k(y)] \quad \text{on } \mathcal{Q}, \\ z(0) &= 0 \quad \text{on } \Omega \text{ at } t = 0, \end{aligned}$$

$$\begin{aligned} -\dot{\psi} + \mathbf{A}\psi + \mathbf{G}(y_k)\psi \\ (3.20) \quad &= -z + (\tilde{y}_k - \tilde{y}) + [\mathbf{G}(y_k) - \mathbf{G}_k(y_k)]\varphi + [\mathbf{G}(y) - \mathbf{G}(y_k)]\varphi \quad \text{on } \mathcal{Q}, \\ \psi(T) &= -\mu z(T) + \mu(\tilde{\eta}_k - \tilde{\eta}) \quad \text{on } \Omega \text{ at } t = T. \end{aligned}$$

Multiplying (3.19) by z , integrating over Ω and using (3.10), (3.17),

$$(3.21) \quad \frac{1}{2} (\|z\|_{\mathcal{S}}^2)' + (1 - \theta) \|z\|_{\mathcal{S}}^2 \leq \langle \psi, z \rangle_{\mathcal{S}} + \|\tilde{\varphi}_k - \tilde{\varphi}\|_{\mathcal{S}} \|z\|_{\mathcal{S}} + \|\delta\|_{\mathcal{S}} \|z\|_{\mathcal{S}} + \varepsilon \|z\|_{\mathcal{X}}^2.$$

Similarly, multiplying (3.20) by ψ , integrating over Ω and using (1.3), (3.10), (3.16), (3.19),

$$(3.22) \quad \begin{aligned} & -\frac{1}{2}(\|\psi\|_{\mathfrak{H}}^2)' + (1-\theta)\|\psi\|_{\mathfrak{B}}^2 \\ & \leq -\langle \psi, z \rangle_{\mathfrak{H}} + \|\tilde{y}_k - \tilde{y}\|_{\mathfrak{H}}\|\psi\|_{\mathfrak{H}} + (\|\tilde{\delta}\|_{\mathcal{X}} + \varepsilon\|y_k\|_{\mathcal{X}})\|\varphi\|_{\mathfrak{B}}\|\psi\|_{\mathfrak{B}} + KC_{\gamma}\mathcal{J}_*^{1/2}\|z\|_{\mathfrak{B}}\|\psi\|_{\mathfrak{B}}. \end{aligned}$$

Adding these and integrating over $[0, T]$ gives, as in the proof of Theorem 3.3,

$$\Gamma(\|z\|_{\mathcal{V}}^2 + \|\psi\|_{\mathcal{V}}^2) \text{ small,}$$

where Γ is as in (3.17) and, since we have a priori bounds on φ , φ_k , y , y_k in \mathcal{W} , (small) means something bounded linearly in the perturbations. This gives (locally Lipschitz-ian) continuity in $\mathcal{V} \times \mathcal{V}$.

Note, however, that one can use the estimate for $\|\psi\|_{\mathcal{V}}$ to estimate the term $\int \langle \psi, z \rangle$ in (3.21) after integrating over $[0, t]$. Then application of Lemma 2.1 gives a suitable estimate for $\|z\|_{\mathcal{X}}$. Similarly, one can estimate $-\langle \psi, z \rangle$ in (3.22), after integrating over $[t, T]$, using the estimate obtained for $\|z\|_{\mathcal{V}}$, and apply Lemma 2.1 to obtain an estimate for $\|\psi\|_{\mathcal{X}}$. Thus, one has continuity in $\mathcal{W} \times \mathcal{W}$ as desired. \square

4. Remarks. For these remarks we restrict ourselves—for simplicity and specificity—to the canonical case of the heat equation: $\mathbf{A} = -\Delta$ with Dirichlet boundary conditions, assuming Ω bounded in \mathbb{R}^m with $\partial\Omega$ smooth.

Remark 4.1. We note first (cf. [2]) that the basic assumptions (1.11), (1.13) hold for this problem

$$(4.1) \quad \begin{aligned} \dot{y} - \Delta y + \mathbf{F}(y) &= \tilde{\varphi} + \varphi && \text{on } \mathcal{Q}, \\ y(0) &= 0 && \text{on } \Omega \text{ at } t = 0, \end{aligned}$$

if we take $\mathfrak{B} := H_0^1(\Omega)$, $\mathfrak{H} := L^2(\Omega)$, $\mathfrak{B}^* = H^{-1}(\Omega)$, while the assumption that \mathfrak{B} embed in \mathfrak{H} (with the indicated norm $\langle \mathbf{A}v, v \rangle^{1/2}$ for \mathfrak{B}) follows from the Poincaré inequality.

Note also that the Sobolev embedding theorem gives the condition (1.12) $H_0^1(\Omega) \subset L^4(\Omega)$ provided $m \leq 4$. While this restriction on the dimension does not affect consideration of (4.1) for physical examples ($m = 3$), we will see that it can be relaxed somewhat. \square

Remark 4.2. The estimates have been taken throughout in terms of the norms of \mathcal{X} , \mathcal{V} . For the linear case ($\mathbf{F}(u) = b(\cdot)u$), it is known [2] that with forcing terms in $\mathcal{H} = L^2(\mathcal{Q})$ one can expect to obtain a solution y in

$$(4.2) \quad \mathcal{Y} := H^{2,1}(\mathcal{Q}) := L^2((0, T) \rightarrow H^2(\Omega)) \cap H^1((0, T) \rightarrow L^2(\Omega)).$$

The same regularity can be obtained for the nonlinear problem (4.1) by consideration of the composed map:

$$(4.3) \quad \begin{aligned} \varphi \mapsto y & \quad \mapsto [\varphi - \mathbf{F}(y)] \mapsto y \\ L^2(\mathcal{Q}) =: \mathcal{H} \rightarrow \mathcal{W} \supset \mathcal{H} \rightarrow \mathcal{H} & \quad \rightarrow \mathcal{Y} := H^{2,1}(\mathcal{Q}). \end{aligned}$$

Here, the first map is just \mathbf{Y} as treated above (or from $\mathcal{H}_{\mathcal{W}}$ to \mathcal{H} as in Lemma 2.6), the second is essentially determined by the Nemytsky operator \mathbf{F} , and finally, the third is just the solution map for the linear problem of (1.13). Strengthening the one-sided Lipschitz condition of (1.2) to a “regular” uniform Lipschitz condition, it follows from (4.3) and the present Lemma 2.3 that the conclusion of that lemma can be strengthened to assert uniform Lipschitz continuity for $\mathbf{Y}: \mathcal{H} \rightarrow \mathcal{Y}$. \square

Remark 4.3. The key to the proof of Theorem 2.4 is the estimation of $\int_{\Omega} |wz|^2$ via (2.9). From the final comment of Remark 4.2, one has $\|z\|_{\mathcal{Y}} \leq (\text{const}) \|h\|_{\mathcal{X}}$. Also,

the strengthening of (1.2) adopted there implies $G(\cdot)$ uniformly bounded, so (1.14) gives z_0 in \mathcal{Y} and gives w in \mathcal{Y} . Supposing we had $\mathcal{Y} \subset L^4(\mathcal{Q})$, one might bound the right-hand side of (2.8) in $L^2(\mathcal{Q})$ by $O(\|h\|_{\mathcal{H}}^2)$ which for this linear equation gives $w = O(\|h\|_{\mathcal{H}}^2)$ in \mathcal{Y} and so Fréchet differentiability of $\mathbf{Y}: \mathcal{H} \rightarrow \mathcal{Y}$. One can show $\mathcal{Y} \subset L^4(\mathcal{Q})$ for $m \leq 5$. On the other hand, if one has only $\mathcal{Y} \subset L^3(\mathcal{Q})$, then essentially the same argument as was used originally to prove Theorem 3.4 gives Fréchet differentiability of $\mathbf{Y}: \mathcal{H} \rightarrow \mathcal{W}$, as there. (In either case the derivative $\mathbf{Y}'(\varphi)$ is given by (1.14).) We now observe that for $m \leq 9$ interpolation theory does give $\mathcal{Y} \subset L^3(\mathcal{Q})$. Thus, the conclusion of Theorem 2.4 actually can be shown to hold not just for $m \leq 4$ but up through 9. \square

Remark 4.4. Given the optimality system (3.6), (3.7), one can hope to obtain better regularity for φ than that it has in \mathcal{W} . Under the hypothesis here adopted that G is bounded, the linear theory [2] gives also that φ as the solution of (3.7) will be in $\mathcal{Y} = H^{2,1}(\mathcal{Q})$ also—provided one has the data for $\varphi(T)$ suitably regular as well in $H_0^1(\Omega)$. Since it follows from (4.3) that $y(T)$ will be in $H_0^1(\Omega)$, we will have this under the assumption that one restricts the data $\tilde{\eta}$ to $H_0^1(\Omega)$. \square

To summarize the discussion of this section, we have demonstrated the following:

THEOREM 4.5. *Let $\Omega \subset \mathbb{R}^m$ with $m \leq 9$ and let \mathbf{F} satisfy (1.2), (1.3)—strengthened to assume G bounded. Let $\tilde{\varphi}, \tilde{y}, \tilde{\eta}$ be given in $L^2(\mathcal{Q}), L^2(\mathcal{Q}), H_0^1(\Omega)$, respectively. Then the control problem posed by (4.1), (3.1) has an optimal pair $[\varphi_*, y_*]$. This satisfies (3.6), (3.7) and is in $\mathcal{Y} \times \mathcal{Y}$. \square*

Remark 4.6. Clearly, the increased regularity of φ_* (beyond the a priori knowledge that $\varphi_* \in \mathcal{H}$ to keep \mathcal{J} finite) should be usable in (3.6) to get still greater regularity for y_* . In turn, greater regularity of y_* should be usable in (3.7). To make effective use of this would require sufficiently strong regularity assumptions on the data: $\mathbf{F}, \tilde{\varphi}, \tilde{y}, \tilde{\eta}$. Given such regularity, familiar “bootstrapping” arguments permit the use of (3.6), (3.7) in recursive alternation to obtain correspondingly strong regularity for φ_* and y_* . \square

Remark 4.7. We conclude with an indication as to how essentially the same arguments can be made to apply to boundary control problems—up to a point.

Suppose then (3.6), defining the map: $\varphi \mapsto y$, is replaced by

$$(4.4) \quad \begin{aligned} \dot{y} - \Delta y + \mathbf{F}(y) &= \tilde{\varphi} && \text{on } \mathcal{Q}, \\ y_\nu + \alpha y &= \varphi && \text{on } \Sigma, \\ y(0) &= 0 && \text{on } \Omega \text{ at } t = 0. \end{aligned}$$

Here, $\tilde{\varphi}$ is in \mathcal{H} as before and \mathbf{F} is to satisfy (1.2). We assume α is in $L^\infty(\Sigma)$ and nonnegative; y_ν denotes the (exterior) normal derivative of y at $\partial\Omega$ and the control φ is now a function on Σ . We now take the functional \mathcal{J} to be

$$\mathcal{J}(\varphi) := \|\varphi\|^2 + \lambda \|y - \tilde{y}\|^2 + \mu \|y(T) - \tilde{\eta}\|^2$$

with, of course, y determined by (4.4) and the norms taken in $\mathcal{H}' = L^2(\Sigma)$, \mathcal{H} and \mathcal{H} .

The basic technique of estimation will be much as earlier. We will write $\|v\|_{\mathfrak{B}}^2$ and $\|v\|_{\mathfrak{V}}^2$ for $\int_\Omega |\nabla v|^2$ and $\int_0^T \|v\|_{\mathfrak{B}}^2$, respectively, even though we now take \mathfrak{B} to be $H^1(\Omega)$ so that these are only seminorms. On the other hand, $\|\cdot\|_{\mathcal{W}}$ (as defined earlier) is a norm on

$$\mathcal{W} := L^\infty((0, T) \rightarrow L^2(\Omega)) \cap L^2((0, T) \rightarrow H^1(\Omega)).$$

We note that trace theory gives

$$(4.5) \quad \|w\|_{\mathcal{H}'} \leq \hat{c} \|w\|_{\mathcal{W}} \quad \text{for } w \in \mathcal{W},$$

where on the left $w \in \mathcal{H}'$ is the (Dirichlet) trace on Σ of $w \in \mathcal{W}$. Multiplying (4.4) by $2y$ and integrating over $(0, t) \times \partial\Omega$ now gives

$$\begin{aligned} & \|y(t)\|_{\mathfrak{S}}^2 + 2 \int_0^t \|y\|_{\mathfrak{S}}^2 + 2 \int_0^t \int_{\partial\Omega} \alpha y^2 \\ &= 2 \int_0^t \int_{\Omega} \tilde{\varphi} y - 2 \int_0^t \int_{\Omega} [\mathbf{F}(y) - \mathbf{F}(0)][y - 0] + 2 \int_0^t \int_{\partial\Omega} \varphi y \\ &\leq 2\|\varphi\|_{\mathcal{H}'} \hat{c} \|y\|_{\mathcal{W}} + 2\|\tilde{\varphi}\|_{\mathcal{H}} \left[\int_0^t \|y\|_{\mathfrak{S}}^2 \right]^{1/2} + \gamma \int_0^t \|y\|_{\mathfrak{S}}^2. \end{aligned}$$

Applying Lemma 2.1 and reasoning as earlier, this gives

$$\|y\|_{\mathcal{W}}^2 \leq C_{\gamma}^2 [\|\tilde{\varphi}\|_{\mathcal{H}}^2 + 2\hat{c}\|\varphi\|_{\mathcal{H}'} \|y\|_{\mathcal{W}}],$$

which bounds y in \mathcal{W} . One similarly gets uniform Lipschitz continuity of the map $\mathbf{Y}: \varphi \rightarrow y: \mathcal{H}' \rightarrow \mathcal{W}$. An argument along the lines of (4.3) then gives (uniform) Lipschitz continuity of $\mathbf{Y}: \mathcal{H}' \rightarrow \mathcal{Y}$ with

$$\mathcal{Y} = H^{3/2, 3/4}(\mathcal{Q}) := L^2((0, T) \rightarrow H^{3/2}(\Omega)) \cap H^{3/4}((0, T) \rightarrow L^2(\Omega))$$

now replacing (4.2). As in Remark 4.3, we note that interpolation theory gives $\mathcal{Y} \subset L^3(\mathcal{Q})$ provided now $m \leq 6$. With that restriction on m and with \mathbf{F} now also required to satisfy (1.3), one argues as for Theorem 2.4 to get differentiability of $\mathbf{Y}: \mathcal{H}' \rightarrow \mathcal{W}$ with $\mathbf{Y}'(\varphi)h = z$ given by:

$$(4.6) \quad \begin{aligned} \dot{z} - \mathbf{A}z + \mathbf{G}(y)z &= 0 && \text{on } \mathcal{Q}, \\ z_{\nu} + \alpha z &= h && \text{on } \Sigma, \\ z(0) &= 0 && \text{on } \Omega \text{ at } t = 0. \end{aligned}$$

One then sees that there exists an optimal control minimizing \mathcal{J} . Computing the adjoints of the maps: $\varphi \mapsto z, z(T)$, defined by (4.6), one shows that such an optimal control satisfies

$$(4.7) \quad \varphi = v|_{\Sigma},$$

where

$$(4.8) \quad \begin{aligned} -\dot{v} - \mathbf{A}v + \mathbf{G}(y)v &= \lambda[y - \tilde{y}] && \text{on } \mathcal{Q}, \\ v_{\nu} + \alpha v &= 0 && \text{on } \Sigma, \\ v(T) &= \mu[y(T) - \tilde{\eta}] && \text{on } \Omega \text{ at } t = T. \end{aligned}$$

Thus this, together with (4.4), is the optimality system for this boundary control problem. As earlier, there is no difficulty in using the system to obtain further regularity results for the optimal pair $[\varphi, y]$; see [3].

At this point, however, the program of the earlier analysis for (3.6) can no longer be continued here. The key to the uniqueness and continuity results (Theorems 3.3, 3.4) is the cancellation of the terms $\langle z, \psi \rangle_{\mathfrak{S}}$ on adding (3.14), (3.15). In attempting to proceed analogously here, however, one obtains $\langle z, w \rangle_{\mathfrak{S}}$ for one equation (taking traces on $\partial\Omega$; here $z = y_1 - y_2$, $w = v_1 - v_2$) and $\langle z, w \rangle_{\mathfrak{S}}$ for the other. No such cancellation is now possible, and the analogous argument breaks down.

One can, however, prove *local* uniqueness (still for small data) using an entirely different argument. Consider the map Φ defined by

$$(4.9) \quad \Phi: (\varphi, \tilde{\varphi}, \tilde{y}, \tilde{\eta}) \mapsto y \mapsto v \mapsto (\varphi - v|_{\Sigma}, \tilde{\varphi}, \tilde{y}, \tilde{\eta})$$

in which y is obtained by solving (4.4) and then using this y , v is obtained by solving (4.8). With φ in \mathcal{H}' and suitable spaces for $\tilde{\varphi}, \tilde{y}, \tilde{\eta}$ (and suitable regularity assumed for F) one sees that Φ is continuously differentiable. Note that $[\varphi, \mathbf{Y}(\varphi)]$ is a solution pair for the optimality system (4.4), (4.8) (necessary for φ to be an optimal control) if and only if

$$(4.10) \quad \Phi(\varphi, \tilde{\varphi}, \tilde{y}, \tilde{\eta}) = (0, \tilde{\varphi}, \tilde{y}, \tilde{\eta}).$$

Thus, if we can show $\Phi'(0)$ is invertible, then the inverse function theorem gives local invertibility of Φ near 0 and so local uniqueness of the optimal control for small data). One easily calculates Φ' to have the form

$$\Phi' = \begin{pmatrix} \mathbf{M} & * & * & * \\ & \mathbf{1} & & \\ & & \mathbf{1} & \\ & & & \mathbf{1} \end{pmatrix}$$

with \mathbf{M} given at 0 by $\mathbf{M}\psi := \psi - w|_{\Sigma}$, where now,

$$(4.11) \quad \begin{aligned} \dot{z} - \Delta z + \mathbf{G}(0)z &= 0 && \text{on } \mathcal{Q}, \\ z_{\nu} + \alpha z &= \psi && \text{on } \Sigma, \\ z(0) &= 0 && \text{on } \Omega \text{ at } t = 0, \end{aligned}$$

$$(4.12) \quad \begin{aligned} -\dot{w} - \Delta w + \mathbf{G}(0)w &= \lambda z && \text{on } \mathcal{Q}, \\ w_{\nu} + \alpha w &= 0 && \text{on } \Sigma, \\ w(T) &= \mu z(T) && \text{on } \Omega \text{ at } t = T. \end{aligned}$$

Clearly, Φ' is invertible if and only if \mathbf{M} is, and as $\mathbf{M}(0)$ is a compact perturbation of the identity, we need only verify that $\mathbf{M}(0)\psi = 0$ implies $\psi = 0$. We note however that [(4.11), (4.12) and $\psi - w|_{\Sigma} = 0$] is just the optimality system for the problem of minimizing

$$\tilde{\mathcal{J}}(\psi) := \|\psi\|^2 + \lambda \|z\|^2 + \mu \|z(T)\|^2 \quad (\text{with (4.11)}).$$

(This is just the special case of (4.4), (4.7), (4.8) with $\mathbf{F}(y) := \mathbf{G}(0)y$.) As (4.11) is *linear*, $\tilde{\mathcal{J}}$ is strictly convex and so has the unique stationary point $\psi = 0$. This completes the proof of local uniqueness of the optimal control. This argument does not, of course, give an explicit estimate (such as (3.11) in combination with (3.3)) for the size of the neighborhood of 0 giving this uniqueness. \square

5. Added remarks. A somewhat unsatisfying feature of the arguments presented above is that the necessary conditions (1.15) have only been justified subject to a dimension restriction due to “embedding theorem” considerations. This does not affect applicability to physical problems ($m = 3$) but is unaesthetic. One notes that, if (1.15) would be known to hold, then one could obtain additional regularity and—with this—may be able to justify the differentiability argument of Theorem 2.4 and so verify the “necessary conditions” (1.15). This argument is circular. Subsequent to the original submission of this paper and motivated by these considerations, an abstract

argument [5] was developed which rigorously justifies such (apparently) circular arguments. In particular, this gives Gâteaux differentiability of \mathbf{Y} at the optimal φ_* . Subsequent to the above, following a remark of G. Chavent, a new argument for differentiability of \mathbf{Y} (everywhere on \mathcal{H}) was constructed which imposes no dimension restriction.

Our task is to show that $z_\tau \rightarrow z$ as $\tau \rightarrow 0$, where $z_\tau := (y_\tau - y)/\tau$ for $\tau \neq 0$ with $y = \mathbf{Y}(\varphi)$, $y_\tau = \mathbf{Y}(\varphi + \tau h)$ and z is the solution of (1.14). By Lemma 2.3 we note that z_τ is bounded in \mathcal{H} independently of τ . Further, z_τ satisfies

$$(5.1) \quad \dot{z}_\tau + \mathbf{A}z_\tau = \zeta_\tau := h - \mathbf{G}(\hat{y}_\tau)z_\tau,$$

where $\hat{y}_\tau := y + \theta_\tau[y_\tau - y]$ with $0 < \theta_\tau < 1$, noting that pointwise

$$\mathbf{F}(y_\tau) - \mathbf{F}(y) = \mathbf{G}(\hat{y}_\tau)[y_\tau - y]$$

by the mean value theorem. Clearly,

$$\|\hat{y}_\tau - y\| \leq \|y_\tau - y\| = O(\tau) \rightarrow 0$$

so $g_\tau := \mathbf{G}(\hat{y}_\tau) \rightarrow g_0 := \mathbf{G}(y)$ in \mathcal{H} as $\tau \rightarrow 0$. With $G = \partial F / \partial r$ assumed uniformly bounded, it is clear that \mathbf{G} is a continuous Nemytsky operator on \mathfrak{S} and that $g_\tau z_\tau$ is bounded (uniformly in τ) in \mathfrak{S} so ζ_τ is uniformly bounded and one can extract a subsequence such that $\zeta_\tau \rightarrow \zeta_0$. Using (1.13) for (5.1), one then has $z_\tau \rightarrow z_0$ in \mathcal{H} , where z_0 is the solution of

$$\dot{z}_0 + \mathbf{A}z_0 = \zeta_0 := w - \lim_{(\text{subseq.})} [h - g_\tau z_\tau].$$

Again extracting a subsequence, one has $g_\tau \rightarrow g_0$ and $z_\tau \rightarrow z_0$ pointwise almost everywhere on \mathcal{Q} as $\tau \rightarrow 0$. It follows that $\zeta_0 = h - g_0 z_0$ so z_0 is the (unique) solution z of (1.14). By uniqueness also, the same convergence holds without extracting subsequences so z is, indeed, the Gâteaux differential of \mathbf{Y} at φ in the direction h . As earlier, analysis of (1.14) then shows that this gives a Fréchet derivative, and all the desired results follow. \square

REFERENCES

- [1] M. A. KRASNOSELSKI, *Topological Methods in the Theory of Nonlinear Integral Equations*, Pergamon, Oxford, 1965.
- [2] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, vol. 2, Springer-Verlag, New York, 1972.
- [3] T. I. SEIDMAN, *Regularity of boundary controls for parabolic equations*, pp. 536–550, *Analysis and Optimization of Systems* (A. Bensoussan and J. L. Lions, eds.), Lecture Notes in Control and Information Science 28, Springer-Verlag, New York, 1980.
- [4] ———, *L'existence de contrôles optimaux pour un problème distribué nonlinéaire*, to appear.
- [5] ———, *Existence and regularity of extrema*, J. Math. Anal. Appl., to appear.

AN ALGORITHM FOR SOLVING S -GAMES AND DIFFERENTIAL S -GAMES*

JERZY A. FILAR[†] AND T. E. S. RAGHAVAN[‡]

Abstract. We present an algorithm for solving S -games. Our algorithm can be used to compute approximately the value of the game as well as ε -optimal strategies of the two players. For games with similar structure to S -games which do not necessarily possess a value, the algorithm can sometimes be used as a heuristic procedure for determining the existence of a minimax solution. Further, it is shown that a certain simple class of differential games (we call them “differential S -games”) can be viewed as static games and solved by the above procedure.

Key words. S -game, control theory, differential S -game, compactness, pre-randomized controls

1. Introduction. We shall consider a two-person, zero-sum game played as follows: Player P_2 chooses a point s in a compact set S in R^n while player P_1 chooses a coordinate i of R^n . The payoff to P_1 is the i th coordinate s_i of s . Such games were first studied by Blackwell and Girshick [2, pp. 47–51] and are often referred to as S -games.

In this note we present an algorithm (§ 2) for approximately computing the value and ε -optimal strategies, for both players in an S -game. Our algorithm reduces the problem to a sequence of minimization problems. Thus the computational strength of the procedure is determined by the strength of the existing techniques for solving the associated minimum problems. If the set S is not known to be compact, our algorithm can lead to a heuristic procedure for testing the existence of the value.

We expect that one fruitful area of application for our algorithm will be that of finding solutions for what we call differential S -games. These are differential games in which only one player (the minimizer) controls the law of motion, while the other player can decide only which out of a fixed number of functionals will be used to compute the payoff. Thus a differential S -game is an essentially control theoretic problem¹ complicated by the uncertainty as to which performance index is being used during the process. In § 3 we show that under suitable assumptions these games can be analyzed as S -games, and consequently that our algorithm can be applied. Of course, the effectiveness of the algorithm will now depend on our ability to solve the associated sequence of optimal control problems. A simple example is given in § 4.

2. An algorithm for S -games. Let Γ be an S -game as described in the introduction. For such a game, Blackwell and Girshick [2] established the following result.

THEOREM 2.1. *The game has a value, $v(\Gamma)$. Both players have optimal randomized strategies and further, player P_2 possesses an optimal strategy which randomizes on at most n points of S .*

We shall now outline a procedure which shows that within some $\varepsilon > 0$, an approximation to the value and to ε -optimal strategies for both players can be obtained by solving a sequence of approximate matrix games, and a corresponding sequence of minimization problems. Our algorithm is a generalization of Troutt's (see [9, pp. 343–345]) algorithm for nonconvex S_n -games.

* Received by the editors April 15, 1981, and in final revised form January 8, 1982. This work was supported in part by the U.S. Air Force Office of Scientific Research under grant 78-3495B.

[†] Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, Maryland 21218.

[‡] Department of Mathematics, University of Illinois at Chicago Circle, Chicago, Illinois 60680.

¹ For an introduction to the theories of differential games and optimal control we refer the reader to Isaacs [4] and Berkowitz [1], respectively.

Let $\lambda = (\lambda_1, \dots, \lambda_n)$ be an arbitrary probability vector, and define for each $s \in S$

$$(1) \quad H(\lambda, s) = \sum_{i=1}^n \lambda_i s_i.$$

Here s_i = the i th coordinate of s . Our algorithm will be as powerful as our ability to solve the following minimization problem:

$$(2) \quad \underset{s \in S}{\text{minimize}} H(\lambda, s).$$

In the next two sections it will be seen that this algorithm is, potentially at least, applicable to a large class of problems.

Notation. If P_1 uses a strategy $\lambda = (\lambda_1 \dots \lambda_n)$ and P_2 uses a strategy $\eta = (\eta_1 \dots \eta_k)$ on the points s^1, s^2, \dots, s^k of S (i.e., chooses s^j with probability η_j), then the expected payoff is simply

$$(3) \quad \Phi(\lambda, \eta) = \sum_{i=1}^n \sum_{j=1}^k \lambda_i \eta_j s_i^j,$$

where s_i^j = i th coordinate of s^j . Further, let $A(s^1, s^2, \dots, s^k)$ be an $n \times k$ matrix whose j th column is the point $s^j \in S$. Also, let $v(A)$ stand for the value of a matrix game A .

THE ALGORITHM

Step 1. [Initialization. See Remark 2.5.] For each $i = 1, 2, \dots, n$ solve the minimization problem (2) with $\lambda = e_i$, the i th vector of the standard basis of R^n . This yields a set of n points in S , say, $\xi^1, \xi^2, \dots, \xi^n$. Now construct the matrix $A^1 = A(\xi^1, \xi^2, \dots, \xi^n)$, and compute its value, $v(A^1)$, and a pair of optimal strategy vectors λ^1 and η^1 for players P_1 and P_2 in the matrix game A^1 .

Step 2. Solve the minimization problem (2) with $\lambda = \lambda^1$. This yields a point s^2 such that $H(\lambda^1, s^2) = \min_s H(\lambda^1, s)$. If $H(\lambda^1, s^2) = v(A^1)$, stop. Otherwise, construct the matrix game $A^2 = A^1(\xi^1, \xi^2, \dots, \xi^n, s^2)$ and solve it.

Step 3. Repeat Step 2 with $\lambda = \lambda^2$ (λ^2 is P_1 's optimal strategy in A^2), and so on.

The above procedure generates a sequence of solutions $v(A^m), \lambda^m, \eta^m$ to the matrix games A^m as well as a sequence of points $\xi^1, \dots, \xi^n, s^2, s^3, s^4, \dots$ in S . The relationship between these sequences and the solution of the original game is summarized in the following result.

THEOREM 2.2. (i) *If the algorithm terminates in k iterations, then $v(\Gamma) = v(A^k)$, λ^k is optimal for player P_1 and η^k is optimal for P_2 , where η^k is to be regarded as a randomization on the points of S corresponding to the columns of A^k .*

(ii) $v(A^k) \rightarrow v(\Gamma)$ as $k \rightarrow \infty$.

(iii) *Given $\varepsilon > 0$, there exists a positive integer M such that λ^m, η^m are ε -optimal for players P_1 and P_2 for all $m \geq M$.*

Proof. (i) The algorithm terminating in k steps means that $H(\lambda^k, s) \geq v(A^k)$ for all $s \in S$. But $\Phi(\lambda, \eta^k) \leq v(A^k)$ for all probability vectors λ since η^k is optimal in the matrix game A^k .

(ii) It follows that

$$H(\lambda^k, s^{k+1}) = \min_s H(\lambda^k, s) \leq v(\Gamma).$$

Since λ^{k+1} is optimal in the matrix game A^{k+1} and since $v(\Gamma) \leq v(A^{k+1})$ we have that for every k

$$(4) \quad \begin{aligned} H(\lambda^k, s^{k+1}) &\leq v(\Gamma) \leq v(A^{k+1}) \leq H(\lambda^{k+1}, \xi) \\ \text{whenever } \xi &\in \{\xi^1, \dots, \xi^n, s^2, s^3, \dots, s^{k+1}\}. \end{aligned}$$

Since the vectors λ^k are probability vectors, $\|\lambda^{k_\nu} - \lambda^{k_{\nu+1}}\|_\infty \rightarrow 0$ for some subsequence $\{\lambda^{k_\nu}\}_{\nu=1}^\infty$. Thus from (4) we obtain

$$(5) \quad \begin{aligned} 0 \leq v(A^{k_{\nu+1}}) - v(\Gamma) &\leq H(\lambda^{k_{\nu+1}}, s^{k_{\nu+1}}) - H(\lambda^{k_\nu}, s^{k_{\nu+1}}) \\ &\leq \|\lambda^{k_{\nu+1}} - \lambda^{k_\nu}\|_\infty \sum_{i=1}^n |s_i^{k_{\nu+1}}|. \end{aligned}$$

Hence $v(A^{k_\nu}) \rightarrow v(\Gamma)$ as $\nu \rightarrow \infty$. Since $v(A^k) \downarrow$ limit, this limit must be $v(\Gamma)$.

(iii) The fact that η^m becomes an ε -optimal randomized strategy for P_2 for m sufficiently large follows immediately from (ii), since P_1 has no restrictions placed on him in the matrix games A^m .

For player P_1 , assume that there does not exist a positive integer M such that (iii) is satisfied for all $m \geq M$. This implies that there exists a subsequence $\{m_r\}_{r=1}^\infty$ of positive integers such that for every r ,

$$(6) \quad H(\lambda^{m_r}, s^{m_r+1}) < v(\Gamma) - \varepsilon.$$

Further, the subsequence of probability vectors $\{\lambda^{m_r}\}_{r=1}^\infty$ must have a convergent subsequence. Without loss of generality assume that $\|\lambda^{m_r} - \lambda^0\|_\infty \rightarrow 0$ as $r \rightarrow \infty$. Now it is possible to find r and l such that $m_r < m_l$ and

$$(7) \quad -\frac{\varepsilon}{2} < H(\lambda^{m_r}, s^{m_r+1}) - H(\lambda^{m_l}, s^{m_r+1}) < \frac{\varepsilon}{2}.$$

Now using the fact that

$$v(\Gamma) \leq v(A^{m_l}) \leq H(\lambda^{m_l}, s^{m_r+1}),$$

we obtain a contradiction to (6), namely

$$(8) \quad v(\Gamma) - \frac{\varepsilon}{2} \leq H(\lambda^{m_r}, s^{m_r+1}).$$

Remark 2.3. Note that we are not claiming that the sequence $\{\lambda^m\}_{m=1}^\infty$ converges but merely that for sufficiently large m , λ^m is ε -optimal for P_1 .

Remark 2.4. Unlike Troutt's algorithm for nonconvex S_n -games, in our construction the matrix A^{k+1} contains all the columns of A^k . It is this feature of the construction which leads to ε -optimality of P_1 's strategies. Further, Troutt's proof of the monotone convergence of the values of his truncated matrices to the value of the S_n -game seems to be incomplete.

Remark 2.5. While the construction of the first matrix game A^1 is not used in the proof of Theorem 2.1, it is a natural starting point. For instance, if A^1 happens to have a saddle-point we would know that the pair of strategies corresponding to the saddle-point are pure optimal strategies for players P_1 and P_2 .

We shall now briefly discuss the situation where the set S is not known to be compact. It could happen that the game still possesses a minimax solution, and our algorithm could sometimes still be useful as a heuristic procedure for "guessing" the solution. This is a consequence of the following result which can be very easily proved.

COROLLARY 2.6. Consider an S -game in which S is not necessarily compact. Let $\bar{\theta}_k = v(A^k)$ and suppose that $\theta_k = \min_s H(\lambda^k, s)$ exists for every k . If the sequences $\{\theta_k\}_{k=1}^\infty$ and $\{\bar{\theta}^k\}_{k=1}^\infty$ converge to a common limit θ , then the value exists and equals θ . Thus the algorithm of Theorem 2.2 still yields an approximate solution.

Remark 2.7. Note that as soon as our algorithm produces an interval $[\theta_k, \bar{\theta}^k]$ which is small and θ_{k+1} which is near θ_k , we are in possession of a pair (λ^{k+1}, η^k) of reasonably good strategies for P_1 and P_2 in the following sense: The expected gain of P_1 will be at least θ_{k+1} when he uses λ^{k+1} while the expected loss of P_2 will be no more than $\bar{\theta}^k$ when he uses η^k . With such a pair of strategies at hand, the question of whether the value actually exists may be of little practical interest.

3. A class of differential S -games. In this section we introduce a special class of differential games which can be analyzed as static S -games and hence we shall call them *differential S -games*. Such a game can also be viewed as a game between a controller in the usual control theoretic sense and an antagonistic opponent who can decide which of the finitely many possible payoff functionals will be used to compute the controller's loss. That choice of the payoff functional remains unknown until the end of the game. More precisely, we consider a differential game Γ with R^m as state space played by two players P_1 and P_2 over the fixed time interval $I = [0, 1]$. At each time $t \in I$, P_2 picks an element $v(t)$ from a compact subset V of R^l in such a way that $v(t)$ is measurable. Player P_1 , on the other hand, picks out a number i from the set $N = \{1, 2, \dots, n\}$ at the beginning of the game.

The "law of motion" of the game is specified by the differential equation

$$(9) \quad \frac{dx}{dt} = f(t, x, v(t)), \quad x(0) = x_0.$$

Here $x \in R^m$ and $f: I \times R^m \times V \rightarrow R^m$ is a continuous function satisfying a Lipschitz condition

$$(10) \quad \|f(t, x_1, v) - f(t, x_2, v)\| \leq K \|x_1 - x_2\|$$

whenever $x_1, x_2 \in R^m$, $t \in I$, $v \in V$; K is some fixed positive number.

The equation (9) now has a unique solution $x(t)$ corresponding to any given initial condition $x(0) = x_0$; the resulting solution shall be called the *trajectory* corresponding to $v(\cdot)$. We may now compute the *payoff* $H(i, v)$ which P_2 will pay to P_1 at the end of the game, by

$$(11) \quad H(i, v) = \int_0^1 h(t, x(t), i, v(t)) dt.$$

Here $h: I \times R^m \times N \times V \rightarrow R$ is a continuous uniformly bounded function.

In order to obtain our results we need to restrict P_2 's set of admissible controls further. We shall assume that the *pure controls* $v(t) = (v_1(t), \dots, v_l(t))$ of P_2 constitute a closed subset $\mathcal{V} \subset (L_1)^l$ and satisfy the conditions (for every j)

$$(i) \quad v_j(t) \in [a_j, b_j],$$

$$(ii) \quad \lim_{\Delta \rightarrow 0} \sup_{\gamma \in \mathcal{V}} \int_0^1 |v_j(t + \Delta) - v_j(t)| dt = 0.$$

These conditions guarantee that \mathcal{V} is compact in the space $(L_1)^l$ of integrable functions. To check this we refer the reader to Lusternik and Sobolev [5, p. 69]. We can now define a *pre-randomized control* of P_2 as a probability distribution on \mathcal{V} , and denote the set of all such controls by $\bar{\mathcal{V}}$. Similarly the set of all probability vectors $\lambda =$

$(\lambda_1, \dots, \lambda_n)$, denoted by \tilde{N} , will represent the space of all prerandomized controls of player P_1 . Thus the payoff to P_1 resulting from a pair of prerandomized controls (λ, ν) is given by

$$(12) \quad \Phi(\lambda, \nu) = \sum_{i=1}^n \int_{\mathcal{V}} \int_0^1 \lambda_i h(t, x(t), i, v(t)) dt d\nu(v).$$

We shall say that Γ has a *value* $v(\Gamma)$ in prerandomized strategies if there exists a pair (λ^0, ν^0) such that for all $\lambda \in \tilde{N}$ and $\nu \in \tilde{\mathcal{V}}$

$$(13) \quad \Phi(\lambda, \nu^0) \leq \Phi(\lambda^0, \nu^0) \leq \Phi(\lambda^0, \nu).$$

Note that the conditional compactness conditions (i) and (ii) can be easily verified for certain natural classes of controls. For instance, the set of all “bang-bang” controls with at most k discontinuities satisfies these conditions and so its closure will be compact.

We can now associate with every pure control $v \in \mathcal{V}$ the n -component vector $H(v) = (H(1, v), H(2, v), \dots, H(n, v))$. The set of such vectors will be denoted by S .

LEMMA 3.1. *S is a compact set in R^n .*

Proof. It is clear that S is bounded. Consider a sequence of points $s_p, p = 1, 2, \dots$, of S . Suppose that this sequence converges to s_∞ . We shall show that $s_\infty \in S$. With each s_p we can associate a pure control v_p and a trajectory x_p satisfying (9). Since the set of all trajectories can be shown to be relatively compact in the Banach space $[C(I)]^m$ of continuous functions $x: I \rightarrow R^m$ (by arguments such as Elliott and Kalton's [3, pp. 14–17]), and since \mathcal{V} is compact, we conclude that a subsequence of $\{x_p\}$ and the same subsequence of $\{v_p\}$ both converge. Since L^1 convergence implies almost everywhere convergence for a subsequence, without loss of generality $v_p \rightarrow v_\infty$ almost everywhere and $x_p \rightarrow x_\infty$ in norm. We shall first prove

$$(14) \quad \dot{x}_\infty = f(t, x_\infty, v_\infty) \quad \text{a.e.}$$

Since for $\tau \in I$, $x_p(\tau) = x_p(0) + \int_0^\tau f(t, x_p(t), v_p(t)) dt$ for every $p = 1, 2, \dots$, and since f is continuous, it follows that in the limit $x_\infty(\tau) = x_\infty(0) + \int_0^\tau f(t, x_\infty(t), v_\infty(t)) dt$. Thus (14) must hold.

Now, by continuity and boundedness of h we can conclude that, for each i ,

$$H(i, v_p) = \int_0^1 h(t, x_p(t), i, v_p(t)) dt \rightarrow H(i, v_\infty).$$

Hence $(s_\infty)_i = H(i, v_\infty)$ for each i , so the lemma is proved.

It is now clear that our game is equivalent to an S -game as formulated in § 1. Namely, player P_2 chooses a point s of a compact set S in R^n while player P_1 chooses a coordinate i of R^n . The payoff to P_1 is the i th coordinate s_i of s .

Let $\lambda = (\lambda_1 \dots \lambda_n)$ be a probability vector, and define for every $v \in \mathcal{V}$

$$(15) \quad H(\lambda, v) = \int_0^1 \left(\sum_{i=1}^n \lambda_i h(t, x(t), i, v(t)) \right) dt.$$

The algorithm of § 2 is now applicable to the differential S -game and will be as powerful as our ability to solve the following optimal control problem in pure controls:

$$(16) \quad \begin{aligned} &\text{minimize } H(\lambda, v) \quad \text{subject to} \\ &\quad \quad \quad v \in \mathcal{V} \\ &\frac{dx}{dt} = f(t, x(t), v(t)), \quad x(0) = x_0. \end{aligned}$$

Remark 3.2. Under additional restrictions it may turn out that a solution can be found in the class of pure controls for at least one player. For instance, if

$$\frac{dx}{dt} = A(t)x + Bv, \quad x(0) = x_0$$

and

$$H(i, v) = \mu_i(x) + \int_0^1 h(i, v(t)) dt,$$

where $A(t)$ is a nonnegative matrix, $\mu_i(x)$ a nonnegative linear functional and h_i is a convex function, then player P_2 has an optimal pure control (see [6]).

4. A simple example. We shall now illustrate the working of our algorithm. The example solved below shows that this algorithm can sometimes yield approximate optimal solutions even in games which do not precisely fit the theory of §§ 2 and 3, provided that the conceptual ingredients of an S -game are present.

The law of motion is given by

$$(17) \quad \frac{dx}{dt} = v, \quad x(0) = 1.$$

Player P_1 has only two choices, 1 or 2, and

$$h(t, x, i, v) = \begin{cases} 10v^2 & \text{if } i = 1, \\ x^2 + v^2 & \text{if } i = 2. \end{cases}$$

The expression (15) of § 3 now becomes

$$(18) \quad H(\lambda, v) = \int_0^1 [\lambda_1 10v^2(t) + \lambda_2(x^2(t) + v^2(t))] dt.$$

Since in this example $H(\lambda, v)$ is convex in v and linear in λ , it follows from Sion's Theorem [8] that we can take \mathcal{V} to be the whole of $L_2[0, 1]$ and be assured that the value exists. In order to apply our algorithm the optimal control problem (16) has to be solvable. If the minimization in (16) is taken over the piecewise continuous functions instead of \mathcal{V} , then this problem can be solved by a standard application of the maximum principle (see Rozonoer [7, pp. 1291]). However, since piecewise continuous functions are dense in $L_2[0, 1]$, the latter solution is also optimal for (16). That solution is of the following form:

For any $\lambda = (\lambda_1, \lambda_2)$ define the constants $\gamma = (\lambda_2/(1 + 9\lambda_1))^{1/2}$, $A_1 = 1/(e^{2\gamma} + 1)$, $A_2 = 1 - A_1$. The optimal trajectory for a particular λ (with $\lambda_2 > 0$) is

$$(19) \quad \bar{x}(t) = A_1 e^{\gamma t} + A_2 e^{-\gamma t},$$

while the optimal control is given by

$$\bar{v}(t) = \gamma(A_1 e^{\gamma t} - A_2 e^{-\gamma t}).$$

If $\lambda_2 = 0$, quite clearly $\bar{v}(t) \equiv 0$.

We shall now summarize the numerical results of the first five iterations:

$$A^5 = \begin{bmatrix} 0 & 1.7080 & .4268 & .9597 & .6659 & .8055 \\ 1 & .7616 & .8218 & .7765 & .7952 & .7850 \end{bmatrix}.$$

The optimal strategies for players P_1 and P_2 in this matrix game are

$$\lambda^5 = (.0681, .9319), \quad \eta^5 = (0, 0, 0, 0, .1368, .8632).$$

Let $\theta_k = H(\lambda^k, v^{k+1}) = \min_v H(\lambda^k, v)$ and $\bar{\theta}_k = v(A^k)$ for $k = 1, 2, \dots$. In this example the lower value θ_k and the upper value $\bar{\theta}_k$ are given by

$$\begin{array}{rcccccc} \bar{\theta}_k: & .8775 & .8037 & .7907 & .7875 & .7864, \\ \theta_k: & .7616 & .7729 & .7847 & .7852 & .7862. \end{array}$$

Thus after five iterations we know that $v(\Gamma)$ lies in the interval $[\bar{\theta}_k, \theta_k]$ (also see Remark 2.7).

Remark 5.1. While the $\bar{\theta}_k$'s are decreasing monotonically, the θ_k 's need not increase monotonically. In fact, $\theta_6 = .7862$ and $\theta_7 = .7860$. Further, columns which are not necessarily for optimal play in the matrix game A^k may become necessary for player P_2 at a later stage.

Acknowledgment. We wish to thank A. J. Goldman from The Johns Hopkins University and D. J. Wilson from the University of Melbourne for reading an earlier version of this manuscript and for their helpful comments. The referee suggested the present organization of the paper.

REFERENCES

- [1] L. D. BERKOVITZ, *Optimal Control Theory*, Springer-Verlag, New York, 1974.
- [2] D. BLACKWELL AND M. A. GIRSHICK, *Theory of Games and Statistical Decisions*, John Wiley, New York, 1954.
- [3] R. J. ELLIOTT AND N. J. KALTON, *The existence of value in differential games*, Mem. Amer. Math. Soc., 126 (1972-73), pp. 1-67.
- [4] R. ISAACS, *Differential Games*, John Wiley, New York, London, 1965.
- [5] L. A. LUSTERNIK AND V. J. SOBOLEV, *Elements of Functional Analysis*, Gordon and Breach, New York, 1961.
- [6] T. PARTHASARATHY AND T. E. S. RAGHAVAN, *Existence of saddle points and Nash equilibrium points for differential games*, this Journal, 13 (1975), pp. 977-980.
- [7] L. I. ROZONOER AND L. S. PONTRYAGIN, *Maximum principle in the theory of optimum systems I*, Automat. Remote Control, 20 (1959), pp. 1288-1302.
- [8] M. SION, *On a general minimax theorem*, Pacific J. Math., 8 (1958), pp. 171-176.
- [9] M. D. TROUTT, *Algorithms for non-convex S_n -games*, Math. Programming, 14 (1978), pp. 332-348.

ADMISSIBLE CONTROLLABILITY OF VIBRATING SYSTEMS WITH CONSTRAINED CONTROLS*

KIMIYAKI NARUKAWA†

Abstract. We study the control of a vibrating string when the controls $f(x, t)$ are applied only on a limited subset of the string and constrained by continuous functions $\alpha(x)$, $\beta(x)$ as $\alpha(x) \leq f(x, t) \leq \beta(x)$. We find that it is possible to transfer almost all states of vibration into rest by a constrained control. The higher space dimensional vibrating systems are also considered.

1. Introduction. In this paper we shall be concerned with the control problems of a vibrating string. The forced motion of a string with density $\rho(x)$ and modulus of elasticity $c(x)$ is described by the equation

$$(1.1) \quad \rho(x) \left[\frac{\partial^2 u}{\partial t^2} \right] - \frac{\partial}{\partial x} \left[c(x) \frac{\partial u}{\partial x} \right] = f(x, t), \quad 0 < x < l, \quad t > 0.$$

Let the both endpoints be fixed, that is,

$$(1.2) \quad u(0, t) = u(l, t) = 0, \quad t > 0.$$

Here $\rho(x)$ and $c(x)$ are sufficiently smooth, bounded and strictly positive functions. Let controls be exercised by means of external forces $f(x, t)$. When continuous functions $\alpha(x)$ and $\beta(x)$ ($\alpha(x) \leq \beta(x)$, $\alpha(x) \neq \beta(x)$) are given, we are concerned with the question whether it is possible to transfer a given initial state of vibration into rest by a control with the constraint $\alpha(x) \leq f(x, t) \leq \beta(x)$. The functions $\alpha(x)$ and $\beta(x)$ are called constraint functions. When the controls are applied only on a limited open subset E , we have only to choose $\alpha(x)$ and $\beta(x)$ such that $\alpha(x) \leq \beta(x)$ on E and $\alpha(x) = \beta(x) = 0$ on $[0, l] - E$. If $\alpha(x) \equiv 0$ or $\beta(x) \equiv 0$, then the control system is considered as the case where the controls are applied only to one direction, that is, only pushing or pulling forces are applied. If a time-independent external force $g(x)$ is applied on the system (for example, a gravity, a magnetic force, etc.), then we have only to take constraint functions as $\tilde{\alpha}(x) = g(x) + \alpha(x)$ and $\tilde{\beta}(x) = g(x) + \beta(x)$.

In general, the control problems where the control $f(t)$ and the output $u(t)$ are related by the differential equation

$$(1.3) \quad \frac{du(t)}{dt} = Au(t) + Bf(t)$$

have been considered by many authors. Here A is the infinitesimal generator of a C_0 semigroup of bounded linear operators $U(t)$, $t \geq 0$, on a Banach space X and B is a bounded linear operator from a Banach space Y to X .

In case the dimensions of X and Y are finite and controls are constrained as $f(t) \in W$ almost all t , where W is a compact set in Y , some necessary and sufficient conditions for the controllability have been obtained by Lee and Markus [6], Saperstone [12], Saperstone and Yorke [13], Brammer [1], etc.

In case the dimensions of X and Y are infinite and W is a unit ball in Y or some other conditions, under some assumptions, admissible null controllability was obtained by Fattorini [4] and Narukawa [7].

* Received by the editors February 24, 1981.

† Department of Mathematics, Faculty of Integrated Sciences, Hiroshima University, Hiroshima, Japan 730.

In this paper we consider the control system (1.1) with (1.2) in the case when zero is not contained in the interior of W . The control problems described by (1.1) without constraints, with the controls of the form $g(x)f(t)$, where $g(x)$ is a given function, was considered by Russell [8]. The boundary control of the vibrating string without constraints was also considered by Russell [9] and Krabs [5]. The work [8] of Russell has the closest relationship to the results of this paper.

In the last section we obtain the results for the higher space dimensional case when $\alpha(x) < \beta(x)$. The controllability of this system without constraints was obtained by Fattorini [4] and Chen [2].

2. Notation and definitions. For an integer $m (\geq 0)$, $H^m(0, l)$ denotes the usual Sobolev space and $H_0^m(0, l)$ denotes the closure of $C_0^\infty(0, l)$ in $H^m(0, l)$. If we set $V(t) = [u(t), (\partial u / \partial t)(t)]$, then (1.1) is reduced to the first order equation

$$(2.1) \quad \frac{dV(t)}{dt} = AV(t) + Bf(t),$$

where

$$A = \begin{pmatrix} 0 & I \\ L & 0 \end{pmatrix}, \quad Lu = \rho^{-1}(x) \frac{\partial}{\partial x} \left[c(x) \frac{\partial u}{\partial x} \right]$$

and

$$Bf(x, t) = \begin{pmatrix} 0 \\ \rho^{-1}(x)f(x, t) \end{pmatrix}.$$

Here, and from now on, we put $[u, v] = \langle u, v \rangle$. We introduce the inner products

$$((u, v)) = \int_0^l c(x) \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} dx \quad \text{for } u, v \in H_0^1(0, l)$$

and

$$(u, v) = \int_0^l \rho(x) u(x) v(x) dx \quad \text{for } u, v \in L^2(0, l).$$

Then the norms defined by these inner products are equivalent to those of $H_0^1(0, l)$ and $L^2(0, l)$, respectively. Let \mathcal{H} be the Hilbert space $H_0^1(0, l) \times L^2(0, l)$ endowed with the inner product

$$([u_1, u_2], [v_1, v_2])_{\mathcal{H}} = ((u_1, v_1)) + (u_2, v_2)$$

for $[u_1, u_2], [v_1, v_2] \in \mathcal{H}$.

If the domain of A is defined as $(H^2(0, l) \cap H_0^1(0, l)) \times H_0^1(0, l)$, then A generates a unitary group $U(t)$ on \mathcal{H} . For any $V_0 = [u_0, u_1] \in \mathcal{H}$ and $f(t) \in L_{\text{loc}}^1(0, T; L^2(0, l))$, we define the mild solution $V(t) = [u(t), (\partial u / \partial t)(t)]$ of (1.1) or (2.1) with the initial state $V(0) = V_0$ by the equation

$$(2.2) \quad V(t) = U(t)V_0 + \int_0^t U(t-s)Bf(s) ds.$$

For a Banach space Y and $1 \leq p \leq \infty$, $L^p(0, T; Y)$ denotes the space of all strongly measurable, Y -valued functions $f(t)$ defined in $0 \leq t \leq T$ with

$$\|f\|_p = \left(\int_0^T \|f(t)\|^p dt \right)^{1/p} < \infty$$

endowed with the norm $\|\cdot\|_p$ (the definition is modified in the usual way when $p = \infty$).

To define admissible controllability, we recall some definitions for the general control system

$$(1.3) \quad \frac{du(t)}{dt} = Au(t) + Bf(t).$$

As is stated in the introduction, A is the infinitesimal generator of a C_0 semigroup of bounded linear operators $U(t)$, $t \geq 0$, on a Banach space X and B is a bounded linear operator from a Banach space Y to X .

DEFINITION 2.1. 1) A subspace D of X is said to be controllable in $L^p(0, T; Y)$ if D is contained in the attainable set

$$\left\{ \int_0^T U(T-s)Bf(s) ds \mid f \in L^p(0, T; Y) \right\}.$$

2) A subspace D of X is said to be null controllable in $L^p(0, T; Y)$ if for each $u_0 \in D$ there exists $f \in L^p(0, T; Y)$ such that

$$U(T)u_0 + \int_0^T U(T-s)Bf(s) ds = 0.$$

3) The control system (1.3) is said to be exactly controllable in $L^p(0, T; Y)$ if the whole space X is controllable in $L^p(0, T; Y)$.

DEFINITION 2.2. Let \mathcal{F} be a subset in $L^1_{loc}([0, \infty); Y)$.

1) A subspace D of X is said to be admissibly controllable in the constraint set \mathcal{F} if for each u_0 and u_1 in D , there exists $f \in \mathcal{F}$ satisfying

$$u_1 = U(T)u_0 + \int_0^T U(T-s)Bf(s) ds.$$

2) A subspace D of X is said to be admissibly null controllable in the constraint set \mathcal{F} if for each $u_0 \in D$ there exists $f \in \mathcal{F}$ satisfying

$$U(T)u_0 + \int_0^T U(T-s)Bf(s) ds = 0.$$

DEFINITION 2.3. The control system (1.3) is said to be admissibly approximately (null) controllable in the constraint set \mathcal{F} if there exists a dense subset D of X which is admissibly (null) controllable in \mathcal{F} .

3. Controllability without constraints. In this section we show the controllability theorem without constraints by the method of Russell [8] and by the stabilizability and controllability theorem by Slemrod [14] and Russell [10].

For any nonnegative and smooth function $\gamma(x)$, let us consider the control system

$$(3.1) \quad \rho(x) \left[\frac{\partial^2 u}{\partial t^2} \right] - \left(\frac{\partial}{\partial x} \right) \left[c(x) \left(\frac{\partial u}{\partial x} \right) \right] = \gamma(x)f(x, t), \quad 0 < x < l, \quad t > 0,$$

with Dirichlet boundary condition (1.2). As are introduced in § 2,

$$(3.2) \quad A = \begin{pmatrix} 0 & I \\ L & 0 \end{pmatrix}, \quad Lu = \rho^{-1}(x) \frac{d}{dx} \left[c(x) \frac{du}{dx} \right],$$

and the domains of A and L are $(H^2(0, l) \cap H_0^1(0, l)) \times H_0^1(0, l)$ and $H^2(0, l) \cap H_0^1(0, l)$ respectively. Further let us define the norms of $D(A^m)$ and $D((-L)^{m/2})$ for a positive

integer m in the usual way as

$$\begin{aligned}\| [u, v] \|_{D(A^m)}^2 &= \| [u, v] \|_{\mathcal{H}}^2 + \left\| A^m \begin{pmatrix} u \\ v \end{pmatrix} \right\|_{\mathcal{H}}^2 \quad \text{for } [u, v] \in D(A^m), \\ \| w \|_{D((-L)^{m/2})}^2 &= \| w \|_{L^2(0, l)}^2 + \| (-L)^{m/2} w \|_{L^2(0, l)}^2 \quad \text{for } w \in D((-L)^{m/2})\end{aligned}$$

respectively. Then it is well known that $D(A^m)$ and $D((-L)^{m/2})$ are closed subspaces of $(H^{m+1}(0, l) \cap H_0^1(0, l)) \times (H^m(0, l) \cap H_0^1(0, l))$ and $H^m(0, l) \times H_0^1(0, l)$ respectively, and further, the norms of these spaces are equivalent respectively.

First we investigate the controllability by the square integrable control functions in t .

THEOREM 3.1. *Let m be a nonnegative integer. If $\gamma(x)$ is not identically zero, then there exists a positive time T such that $D(A^m)$ (= the domain of A^m) is controllable in $L^2(0, T; D((-L)^{m/2}))$ for the control system (3.1) with (1.2).*

Proof. We show the results along the lines of Russell [8]. For details see [8]. By Liouville's transformation, that is,

$$u^* = [c(x)\rho(x)]^{1/4}u, \quad x^* = \int_0^x \left[\frac{\rho(\xi)}{c(\xi)} \right]^{1/2} d\xi,$$

we obtain a new equation in u^* involving derivatives with respect to x^* and t . Reverting to the use of u and x rather than u^* and x^* , this equation is

$$(3.3) \quad \frac{\partial^2 u}{\partial t^2} - r(x)u - \frac{\partial^2 u}{\partial x^2} = \tilde{\gamma}(x)\tilde{f}(x, t), \quad 0 < x < K,$$

where $K = \int_0^l [\rho(x)/c(x)]^{1/2} dx$, $r(x)$ is a continuous function on $[0, K]$ and $\gamma(x)$ is a nonnegative and smooth function.

Put

$$\begin{aligned}\tilde{A} &= \begin{pmatrix} 0 & 1 \\ P & 0 \end{pmatrix}, \quad P = \frac{d^2}{dx^2} + r(x), \\ D(\tilde{A}) &= (H^2(0, K) \cap H_0^1(0, K)) \times H_0^1(0, K) \subset H_0^1(0, K) \times L^2(0, K), \\ D(P) &= H^2(0, K) \cap H_0^1(0, K) \subset L^2(0, K).\end{aligned}$$

Let $\{-\lambda_k\}$ and $\{\varphi_k\}$ be the eigenvalues and eigenfunctions respectively of P and let $\{\varphi_k\}$ form an orthonormal basis in $L^2(0, K)$. Let the initial state $[u_0, u_1] \in D(\tilde{A}^m)$ be expanded as

$$u_0(x) = \sum \mu_k \varphi_k(x), \quad u_1(x) = \sum \nu_k \varphi_k(x).$$

Then, by simple computation, a necessary and sufficient condition that $\tilde{f}(x, t)$ steers $[u_0, u_1]$ to the zero state at T is

$$(3.4) \quad \nu_k = - \int_0^T \int_0^K \cos(\omega_k t) \varphi_k(x) \tilde{\gamma}(x) \tilde{f}(x, t) dx dt,$$

$$(3.5) \quad \mu_k = \int_0^T \int_0^K \omega_k^{-1} \sin(\omega_k t) \varphi_k(x) \tilde{\gamma}(x) \tilde{f}(x, t) dx dt, \quad k = 1, 2, \dots,$$

where $\{\omega_k\}$ are square roots of $\{\lambda_k\}$.

For $T \geq 2K$ there exists a biorthogonal system $\{p_k(t), q_k(t)\}$ for $\{\sin(\omega_k t), \cos(\omega_k t)\}$ in $L^2(0, T)$, that is,

$$\begin{aligned} \int_0^T \sin(\omega_k t) p_l(t) dt &= \delta_{kl}, & \int_0^T \cos(\omega_k t) p_l(t) dt &= 0, \\ \int_0^T \cos(\omega_k t) q_l(t) dt &= \delta_{kl}, & \int_0^T \sin(\omega_k t) q_l(t) dt &= 0. \end{aligned}$$

Put

$$\begin{aligned} \tilde{f}(x, t) &= -\sum \nu_k q_k(t) \varphi_k(x) \left[\int_0^K \tilde{\gamma}(x) |\varphi_k(x)|^2 dx \right]^{-1} \\ &\quad + \sum \mu_k \omega_k p_k(t) \varphi_k(x) \left[\int_0^K \tilde{\gamma}(x) |\varphi_k(x)|^2 dx \right]^{-1} \\ (3.6) \quad &= \sum q_k(t) v_k(x) + \sum p_k(t) w_k(x). \end{aligned}$$

Here

$$v_k(x) = -\nu_k \varphi_k(x) \left[\int_0^K \tilde{\gamma}(x) |\varphi_k(x)|^2 dx \right]^{-1}$$

and

$$w_k(x) = \mu_k \omega_k \varphi_k(x) \left[\int_0^K \tilde{\gamma}(x) |\varphi_k(x)|^2 dx \right]^{-1}.$$

By Lemma 3.1, mentioned later, there exists a positive δ independent of k such that

$$\int_0^K \tilde{\gamma}(x) |\varphi_k(x)|^2 dx \geq \delta > 0.$$

Thus we have

$$\|v_k\|_{D((-P)^{m/2})}^2 + \|w_k\|_{D((-P)^{m/2})}^2 \leq \delta^{-2} (\nu_k^2 \omega_k^{2m} + \omega_k^{2(m+1)} \mu_k^2).$$

Hence

$$\begin{aligned} \sum \|v_k\|_{D((-P)^{m/2})}^2 + \sum \|w_k\|_{D((-P)^{m/2})}^2 \\ (3.7) \quad &\leq \delta^{-2} \left(\sum \nu_k^2 \omega_k^{2m} + \sum \mu_k^2 \omega_k^{2(m+1)} \right) \\ &\leq \delta^{-2} \left\| \begin{pmatrix} u_0 \\ u_1 \end{pmatrix} \right\|_{D(\tilde{A}^m)}. \end{aligned}$$

The convergence of (3.7) is equivalent to the fact that $\tilde{f}(x, t)$ belongs to $L^2(0, T; D((-P)^{m/2}))$. (For the proof see Russell [11, p. 669–670]. The proof in [11] is given for scalar valued functions, but is similar for Hilbert valued functions.) Thus we have $\tilde{f}(x, t) \in L^2(0, T; D((-P)^{m/2}))$, and it is easy to see that for the control system (3.3) with (1.2) $D(\tilde{A}^m)$ is controllable in $L^2(0, T; D((-P)^{m/2}))$. Reversing the variable x , we have the controllability for the control system (3.1) with (1.2).

Next we show the lemma used in the proof of Theorem 3.1.

LEMMA 3.1. *Let P be the differential operator obtained by Liouville's transformation and $\{\varphi_k(x)\}$ be the eigenfunctions of P with Dirichlet boundary condition which form an orthonormal basis in $L^2(0, K)$. Then for any $0 \leq a < b \leq K$, there exists a positive δ*

which is independent of k such that

$$\int_a^b |\varphi_k(x)|^2 dx \geq \delta > 0.$$

Proof. It is well known that there exists an eigenfunction $u_k(x)$ for each eigenvalue λ_k such that

$$u_k(x) = \sin(\omega_k x) + O(\omega_k^{-1}),$$

where $O(\omega_k^{-1})$ denotes the large order of ω_k^{-1} as $k \rightarrow \infty$. For the proof see, e.g., Yosida [15, p. 112]. Thus

$$\begin{aligned} \int_0^K |u_k(x)|^2 dx &\leq \int_0^K \sin^2(\omega_k x) dx + O(\omega_k^{-1}) \\ &= \frac{K}{2} + O(\omega_k^{-1}). \end{aligned}$$

Let $\varphi_k(x) = c_k u_k(x)$. Then

$$1 = \int_0^K |\varphi_k(x)|^2 dx = |c_k|^2 \int_0^K |u_k(x)|^2 dx.$$

Hence there exists some positive μ such that

$$1 \leq \mu |c_k|^2.$$

Thus $|c_k|^2 \geq \mu^{-1}$. Therefore,

$$\begin{aligned} \int_a^b |\varphi_k(x)|^2 dx &\geq \mu^{-1} \int_a^b |u_k(x)|^2 dx \\ &= \mu^{-1} \left[\int_a^b \sin^2(\omega_k x) dx + O(\omega_k^{-1}) \right] = \frac{b-a}{2\mu} + O(\omega_k^{-1}). \end{aligned}$$

Thus there exists an integer n such that

$$\frac{b-a}{2\mu} + O(\omega_k^{-1}) \geq \frac{b-a}{4\mu} \quad \text{for } k \geq n.$$

Since

$$\int_a^b |\varphi_k(x)|^2 dx \neq 0,$$

there exists $\delta (>0)$ such that

$$\int_a^b |\varphi_k(x)|^2 dx \geq \delta > 0.$$

In order to show the controllability theorem by the bounded measurable control functions in t , we recall some theorems by Slemrod and Russell.

THEOREM 3.2 (Dolecki–Russell [3]). *Let X and Y be Hilbert spaces and the general control system (1.3) be exactly controllable in $L^2(0, T; Y)$. Then there exists a positive δ such that*

$$\int_0^T \|B^* U^*(T-t)x\|^2 dt \geq \delta \|x\|^2$$

for any $x \in X$. Here $*$ denotes the adjoint operators.

THEOREM 3.3 (Slemrod [14]). *Let X and Y be Hilbert spaces and $U(t)$ be a group. If there exist positive ε and δ satisfying*

$$\int_0^\varepsilon \|B^*U^*(-t)x\|^2 dt \geq \delta \|x\|^2 \quad \text{for any } x \in X,$$

then the control system (3.1) is stabilizable; that is, there exists a bounded linear operator K from X to Y such that $A+BK$ generates a uniformly, asymptotically stable C_0 semigroup.

THEOREM 3.4 (Russell [10]). *Let the control system (1.3) be stabilizable both forward and backward; that is, there exist bounded linear operators K^+ and K^- from X to Y such that $S^+(t) \rightarrow 0$ as $t \rightarrow \infty$ and $S^-(t) \rightarrow 0$ as $t \rightarrow -\infty$, where $S^\pm(t)$ are the C_0 semigroups generated by $A+BK^\pm$ respectively. Then the control system (1.3) is exactly null controllable in $L^\infty(0, T; Y)$ for large T .*

Now we have:

THEOREM 3.5. *Let the hypotheses of Theorem 3.1 be satisfied. Then there exists a positive time T such that $D(A^m)$ is controllable in $L^\infty(0, T; D((-L)^{m/2}))$ for the control system (3.1) with (1.2).*

Proof. Let X and Y be $D(A^m)$ and $D((-L)^{m/2})$ respectively. Then A generates a unitary group $U(t)$ on $D(A^m)$ and B is a bounded linear operator from Y to X . Since the control system is exactly controllable in $L^2(0, T; Y)$ by Theorem 3.1, by Theorem 3.2 there exists a positive δ satisfying

$$\int_0^T \|B^*U^*(T-t)x\|^2 dt \geq \delta \|x\|^2 \quad \text{for any } x \in X.$$

Since (1.3) is invariant under time reversal, the hypotheses of Theorem 3.3 hold. Thus the hypotheses of Theorem 3.4 are satisfied. Hence, by Theorem 3.4, $X = D(A^m)$ is null controllable in $L^\infty(0, T; D((-L)^{m/2}))$. Since the control system is invariant under time reversal, controllability follows from null controllability.

4. Controllability with constraints. Let us consider the control problems when the control functions are constrained in some limited set. For a positive integer m and continuous functions $\alpha(x)$ and $\beta(x)$ on $[0, l]$ ($\alpha(x) \leq \beta(x)$, $\alpha(x) \not\equiv \beta(x)$), we take the constraint sets of controls as

$$(4.1) \quad \mathcal{F}_{\alpha, \beta}^m = \bigcup_{T>0} \{f(x, t) \in L^\infty(0, T; H^m(0, l)) \mid \alpha(x) \leq f(x, t) \leq \beta(x) \text{ for almost all } t, x\}.$$

Under these constraints we consider the control problems of the control system (1.1) with (1.2).

LEMMA 4.1. *Let m be a nonnegative integer and $g(x)$ be a function in $D((-L)^{m/2})$. Then the mild solution $u(x, t)$ of (1.1) with $f(x, t) = g(x)$ and initial data $u(x, 0) = 0$ and $(\partial u / \partial t)(x, 0) = 0$ satisfies the following inequalities:*

$$(4.2) \quad \left\| \left[u(t), \frac{\partial u}{\partial t}(t) \right] \right\|_{D(A^{m+1})} \leq c_0,$$

where c_0 is a constant not depending on t , A and L are the closed operators defined by (3.2).

Proof. Let $\{\varphi_k(x)\}$ be the eigenfunctions of $-L$ which form an orthonormal basis and $\{\omega_k\}$ be the square roots of the eigenvalues $\{\lambda_k\}$. Then $u(t)$ and $(\partial u / \partial t)(t)$ are

represented as

$$u(t) = \sum \mu_k(t) \varphi_k(x), \quad \frac{\partial u}{\partial t}(t) = \sum \nu_k(t) \varphi_k(x),$$

where

$$(4.3) \quad \mu_k(t) = \frac{g_k}{\omega_k} \int_0^t \sin [\omega_k(t-s)] ds = g_k \omega_k^{-2} [1 - \cos (\omega_k t)],$$

$$(4.4) \quad \nu_k(t) = g_k \int_0^t \cos [\omega_k(t-s)] ds = g_k \omega_k^{-1} \sin (\omega_k t)$$

and

$$g_k = \int_0^l g(x) \varphi_k(x) dx.$$

Since

$$\begin{aligned} \left\| \left[u(t), \frac{\partial u}{\partial t}(t) \right] \right\|_{D(A^{m+1})}^2 &= \left\| \begin{pmatrix} u(t) \\ (\partial u / \partial t)(t) \end{pmatrix} \right\|_{\mathcal{H}}^2 + \left\| A^{m+1} \begin{pmatrix} u(t) \\ (\partial u / \partial t)(t) \end{pmatrix} \right\|_{\mathcal{H}}^2 \\ &\leq \text{const.} \left(\|(-L)^{(m+2)/2} u(t)\|^2 + \left\| (-L)^{(m+1)/2} \frac{\partial u}{\partial t}(t) \right\|^2 \right) \end{aligned}$$

and

$$\begin{aligned} &\|(-L)^{(m+2)/2} u(t)\|^2 + \left\| (-L)^{(m+1)/2} \frac{\partial u}{\partial t}(t) \right\|^2 \\ &= \sum \omega_k^{2(m+2)} |\mu_k(t)|^2 + \sum \omega_k^{2(m+1)} |\nu_k(t)|^2, \end{aligned}$$

we have, by (4.3) and (4.4),

$$\left\| \left[u(t), \frac{\partial u}{\partial t}(t) \right] \right\|_{D(A^{m+1})}^2 \leq \text{const.} \left(\sum \omega_k^{2m} |g_k|^2 \right).$$

Since $g(x)$ belongs to $D((-L)^{m/2})$, the series converges and

$$(4.5) \quad \sum \omega_k^{2m} |g_k|^2 \leq c_0,$$

where c_0 is a constant not depending on t . Thus we have the result.

In order to show the admissible controllability, let us recall the following theorem and corollary for the general control system (1.3) which was obtained by Narukawa [7].

THEOREM 4.1. *Let $1 < p \leq \infty$ and assume the following:*

- 1) *The controlled space D is a Banach space endowed with a norm $\|\cdot\|_D$ stronger than the norm $\|\cdot\|_X$ of X ; that is, there exists some positive constant γ such that $\|u\|_X \leq \gamma \|u\|_D$ for any $u \in D$;*
 - 2) *the controlled space D is invariant under $U(t)$ for each $t \geq 0$, that is, $U(t)D \subset D$ for all $t \geq 0$;*
 - 3) *$U(t)$ is contractive on D , that is, $\|U(t)u\|_D \leq \|u\|_D$ for all $t \geq 0$;*
 - 4) *the controlled space D is null controllable in $L^p(0, T_0; Y)$ for some $T_0 > 0$.*
- Then the controlled space D is admissibly null controllable in the constraint set*

$$F_\eta^p = \bigcup_{T>0} \{f \in L^p(0, T; Y) \mid \|f(t)\|_p \leq \eta\}$$

for any $\eta > 0$.

COROLLARY 4.1. *Let $1 < p \leq \infty$ and A be the infinitesimal generator of a contraction C_0 semigroup. For some positive α if $D((-A)^\alpha)$ is null controllable in $L^p(0, T; Y)$, then $D((-A)^\alpha)$ is admissibly null controllable in F_η^p for any $\eta > 0$.*

Let us define the null controllable set N_T at T in F_η^∞ defined as

$$N_T = \left\{ u_0 \in X \mid \text{there exists } f \in F_\eta^\infty \text{ such that } U(T)u_0 + \int_0^T U(T-s)Bf(s) dx = 0 \right\}.$$

Then, by Corollary 4.1 and the proof of Theorem 4.1, we have:

LEMMA 4.2. *Let the hypotheses in Corollary 4.1 be satisfied. Then, for any positive number r , there exists a positive time T such that*

$$\{u \in X \mid \|u\| \leq r\} \subset N_T.$$

Now let us go back to the control system (3.1) with (1.2). Noting that the control system is invariant under time reversal and the closed operator defined by (3.2) generates a unitary group in \mathcal{H} , by Theorem 3.5 and Lemma 4.2, we have:

LEMMA 4.3. *Let the hypotheses in Theorem 3.1 be satisfied. Then for any positive number r , there exists a positive time T at which the zero state can be steered by the control f belonging in $L^\infty(0, T; D((-L)^{m/2}))$ and satisfying the inequality*

$$\|f(t)\|_{D((-L)^{m/2})} \leq \eta \quad \text{almost all } t,$$

to any state in $\{[u, v] \in D(A^m) \mid \|[u, v]\|_{D(A^m)} \leq r\}$.

Now, for the control system (1.1) with (1.2), we have:

THEOREM 4.2. *Let m be a positive integer. If the constraint functions $\alpha(x)$ and $\beta(x)$ belong to $D((-L)^{m/2})$, then $D(A^m)$ is admissibly controllable in the constraint set $\mathcal{F}_{\alpha, \beta}^m$.*

Proof. Put $g(x) = [\alpha(x) + \beta(x)]/2$. Since $g(x) \in D((-L)^{m/2})$, by Lemma 4.1 we have

$$\left\| \left[w(t), \frac{\partial w}{\partial t}(t) \right] \right\|_{D(A^m)} \leq c_0$$

for the mild solution $w(x, t)$ of (1.1) with $f(x, t) = g(x)$ and initial state $w(x, 0) = (\partial w / \partial t)(x, 0) = 0$.

Put $E = \{x \in [0, l] \mid \alpha(x) \neq \beta(x)\}$. Then $E \neq \emptyset$. Thus we can choose a nonnegative and nonzero function $\gamma(x)$ in $C_0^\infty(0, l)$ so that $\text{supp } \gamma(x) \subset E$. Let us take $\eta > 0$ so that

$$(4.6) \quad \eta \gamma(x) \leq [\beta(x) - \alpha(x)]/2.$$

By Lemma 4.3 for any given $\tilde{\eta} > 0$ and $r > 0$, there exists $T > 0$ such that the set

$$R_{\tilde{\eta}}(T) = \left\{ \int_0^T U(T-s) \begin{pmatrix} 0 \\ \gamma f(s) \end{pmatrix} ds \mid f(t) \in L^\infty(0, T; D((-L)^{m/2})) \text{ and } \|f(t)\|_{D((-L)^{m/2})} \leq \tilde{\eta} \text{ almost all } t \right\}$$

contains the ball $B_r = \{[u, v] \in D(A^m) \mid \|[u, v]\|_{D(A^m)} \leq r\}$, where $U(t)$ is the group generated by A . Since the norms of $D((-L)^{m/2})$ are equivalent to the one of $H^m(0, l)$, by Sobolev's inequality,

$$(4.7) \quad \sup_{x \in (0, l)} |f(x, t)| \leq c_1 \|f(t)\|_{D((-L)^{1/2})} \leq c_1 \|f(t)\|_{D((-L)^{m/2})} \leq c_1 \tilde{\eta}$$

holds for almost all t . Let us take $\tilde{\eta}$ so small that $c_1 \tilde{\eta} < \eta$. Then, by (4.6) and (4.7),

we have

$$\alpha(x) \leq g(x) + \gamma(x)f(x, t) \leq \beta(x) \quad \text{almost all } t$$

for all functions $f(x, t) \in L^\infty(0, T; D((-L)^{m/2}))$ satisfying

$$\|f(t)\|_{D((-L)^{m/2})} \leq \tilde{\eta} \quad \text{almost all } t.$$

By Sobolev's imbedding theorem, $f(x, t)$ is continuous in x for almost all t . Hence the constraint set contains the set

$$\bigcup_{T>0} \{g(x) + \gamma(x)f(x, t) \mid f(x, t) \in L^\infty(0, T; D((-L)^{m/2})) \text{ satisfying } \|f(t)\|_{D((-L)^{m/2})} \leq \tilde{\eta} \text{ almost all } t\}.$$

Thus, by (4.2) and the equality

$$\int_0^T U(T-s) \begin{pmatrix} 0 \\ g + \gamma f(s) \end{pmatrix} ds = \int_0^T U(T-s) \begin{pmatrix} 0 \\ g \end{pmatrix} ds + \int_0^T U(T-s) \begin{pmatrix} 0 \\ \gamma f(s) \end{pmatrix} ds,$$

the set

$$\left\{ \int_0^T U(T-s) \begin{pmatrix} 0 \\ h(s) \end{pmatrix} ds \mid h \in \mathcal{F}_{\alpha, \beta}^m \right\}$$

contains the set $R_{\tilde{\eta}}(T) + [w(T), (\partial w / \partial t)(T)]$. Since for any $r > 0$ there exists $T > 0$ such that $R_{\tilde{\eta}}(T) \supset B_r$, and the inequality $\| [w(T), (\partial w / \partial t)(T)] \|_{D(A^m)} \leq c_0$, we have

$$\bigcup_{T>0} \left\{ R_{\tilde{\eta}}(T) + \left[w(T), \frac{\partial w}{\partial t}(T) \right] \right\} = D(A^m).$$

Thus, for any $[u_1, v_1] \in D(A^m)$, there exist a positive time T and a control $h(x, t)$ in $\mathcal{F}_{\alpha, \beta}^m$ which steers the zero state to the final state $[u_1, v_1]$. Noting that the control system (1.1) with (1.2) is invariant under time reversal, we have the admissible controllability of $D(A^m)$ in the constraint set $\mathcal{F}_{\alpha, \beta}^m$.

Let us define the the constraint set with smooth controls as

$$(4.8) \quad C_{\alpha, \beta}^\infty = \bigcup_{T>0} \{f(x, t) \in L^\infty(0, T; C^\infty[0, l]) \mid \alpha(x) \leq f(x, t) \leq \beta(x) \text{ holds for almost all } t\}.$$

Since $\bigcap_{m \geq 1} D(A^m)$ is dense in \mathcal{H} and $\bigcap_{m \geq 1} \mathcal{F}_{\alpha, \beta}^m = C_{\alpha, \beta}^\infty$, it is easy to see the following corollary.

COROLLARY 4.2. *Let $\alpha(x)$ and $\beta(x)$ be smooth functions such that $\alpha(x) \leq \beta(x)$ and $\alpha(x) \not\equiv \beta(x)$. Then the control system (1.1) with (1.2) is admissibly approximately controllable in the constraint set $C_{\alpha, \beta}^\infty$.*

Remark 4.1. The set $R_{\tilde{\eta}}(T)$ in the proof of Theorem 4.3 is increasing in T , that is, $R_{\tilde{\eta}}(T_1) \subset R_{\tilde{\eta}}(T_2)$ for $T_1 \leq T_2$. Hence, we have the following more precise statements: For any $[u_0, v_0]$ and $[u_1, v_1]$ in $D(A^m)$ ($m \geq 1$), there exists a positive time T_0 such that the initial state $[u_0, v_0]$ is steered to the final state $[u_1, v_1]$ at any time T greater than T_0 by the control function in $\mathcal{F}_{\alpha, \beta}^m$.

Remark 4.2. In Lemma 4.1 we showed the mild solution with exterior force $g(x)$ is bounded. Further, it is shown that under the assumptions of Lemma 4.1 for any $\varepsilon > 0$ there exists a sequence $\{T_n\}$ such that $0 < T_1 < T_2 < \cdots < T_n < \cdots \rightarrow \infty$ and $\| [u(T_n), (\partial u / \partial t)(T_n)] \|_{D(A^{m+1})} \leq \varepsilon$ for each T_n .

In fact, in the proof of Lemma 4.1, there exists an integer N such that

$$\sum_{k \geq N+1} \omega_k^{2m} |g_k|^2 \leq \frac{\varepsilon}{2},$$

because the series $\sum \omega_k^{2m} |g_k|^2$ converges. By the following Lemma 4.4, for any $\eta > 0$ there exist a sequence $T_1 < T_2 < \dots < T_n < \dots \rightarrow \infty$ and integers $\{q_{n_k}\}_{n=1,2,\dots, k=1,2,\dots, N}$ such that

$$|\omega_k T_n - 2\pi q_{n_k}| < \eta, \quad k = 1, 2, \dots, N \quad \text{for each } T_n.$$

Let us take $\eta > 0$ so small that

$$\sum_{k=1}^N \omega_k^{2m} |g_k|^2 [(1 - \cos(\omega_k T_n))^2 + \sin^2(\omega_k T_n)] < \frac{\varepsilon}{2}, \quad n = 1, 2, \dots.$$

Then we have

$$\begin{aligned} \left\| \left[u(T_n), \frac{\partial u}{\partial t}(T_n) \right] \right\|_{D(A^{m+1})} &\leq \sum_{k=1}^N \omega_k^{2m} |g_k|^2 [(1 - \cos(\omega_k T_n))^2 + \sin^2(\omega_k T_n)] + \sum_{k=N+1}^{\infty} \omega_k^{2m} |g_k|^2 \\ &< \varepsilon. \end{aligned}$$

LEMMA 4.4. *Let real numbers $\{\omega_k\}_{1 \leq k \leq N}$ be given. Then for any $\varepsilon > 0$ there exist a sequence $\{T_n\}$ and integers $\{p_{n_k}\}_{n=1,2,\dots, k=1,2,\dots, N}$ such that $T_n \rightarrow \infty$ as $n \rightarrow \infty$ and*

$$|\omega_k T_n - p_{n_k}| < \varepsilon, \quad k = 1, 2, \dots, N, \quad \text{for each } T_n.$$

This lemma is proved by Kronecker's approximation theorem and by induction for N .

Remark 4.3. Krabs [5] considered the general one-dimensional vibrating systems with boundary controls. We can also obtain the similar results for these systems with distributed controls.

Remark 4.4. In the proof of Theorem 4.2, we showed that $D(A^m)$ is controllable by the controls $g(x) + \gamma(x)f(x, t)$, where $g(x) \in D((-L)^{m/2})$ and $f(x, t) \in L^\infty(0, T; D((-L)^{m/2}))$. But by Lemma 4.1, $g(x)$ need not belong to $D((-L)^{m/2})$ and has only to belong to $D((-L)^{(m-1)/2})$. Thus, especially by taking $m = 1$, $\alpha(x)$ and $\beta(x)$ in $H^1(0, l)$, we have the admissible controllability of $D(A)$ in $\mathcal{F}_{\alpha, \beta}^1$. Hence, in this case, the control functions $\alpha(x)$ and $\beta(x)$ need not vanish at $x = 0$ or l .

5. The higher dimensional case. In this section we consider the case where the space dimension is greater than 1.

Let the control system be

$$(5.1) \quad \rho(x) \left[\frac{\partial^2 u}{\partial t^2} \right] - \sum \frac{\partial}{\partial x_i} \left[c_{ij}(x) \frac{\partial u}{\partial x_j} \right] = f(x, t)$$

in $\Omega \times (0, \infty)$ with Dirichlet boundary condition

$$(5.2) \quad u(x, t) = 0 \quad \text{on } \partial\Omega \times (0, \infty).$$

Here Ω is a bounded domain in \mathbf{R}^n with smooth boundary $\partial\Omega$, $\rho(x)$ and $c_{ij}(x)$ are smooth and bounded functions in Ω which satisfy

$$0 < \rho_0 \leq \rho(x) \leq \rho_1, \quad c_{ij}(x) = c_{ji}(x) \quad \text{for } x \in \Omega$$

and

$$\sum c_{ij}(x) \xi_i \xi_j \geq \delta \sum |\xi_i|^2 \quad \text{for any } \xi = (\xi_i) \in \mathbf{R}^n$$

with positive constants ρ_0, ρ_1 and δ .

As is similar to the one-dimensional case, we define the closed operators A and L and the Hilbert space \mathcal{H} . Then A generates a unitary group, and thus, we can define the mild solutions and control problems in the similar way to the one-dimensional case. For a positive integer m and continuous functions $\alpha(x)$ and $\beta(x)$ on Ω satisfying $\alpha(x) < \beta(x)$ in Ω , we take the constraint set of controls as

$$(5.3) \quad \mathcal{F}_{\alpha,\beta}^m = \bigcup_{T>0} \{f(x, t) \in L^\infty(0, T; H^m(\Omega)) \mid \alpha(x) \leq f(x, t) \leq \beta(x) \text{ almost all } t \text{ and } x\}.$$

We have Lemma 4.1 in the same way for $n \geq 2$. For the higher dimensional case, we do not know whether Theorems 3.1 and 3.5 hold or not. When $\gamma(x) = 1$, Fattorini [4] and Chen [2] proved the controllability of the control system (5.1) with (5.2). Similarly it is easy to see that Theorem 3.5 and Lemma 4.3 hold when $\gamma(x) = 1$.

Now we have:

THEOREM 5.1. *Let m be an integer greater than $[n/2] + 1$. If $\alpha(x)$ and $\beta(x)$ belong to $D((-L)^{m/2})$ and satisfy*

$$(5.4) \quad \frac{\partial}{\partial n} \alpha(x) < \frac{\partial}{\partial n} \beta(x) \quad \text{on } \partial\Omega,$$

where $\partial/\partial n$ is an inward unit normal derivative on $\partial\Omega$, then $D(A^m)$ is admissibly controllable in the constraint set $\mathcal{F}_{\alpha,\beta}^m$.

Proof. Since $D((-L)^{m/2})$ is contained in $H^m(\Omega) \cap H_0^1(\Omega)$ and the norms of these spaces are equivalent, by Sobolev's imbedding theorem we have $D((-L)^{m/2}) \subset B_0^1(\Omega)$ topologically, where

$$B_0^1(\Omega) = \{u \in C^1(\bar{\Omega}) \mid u = 0 \text{ on } \partial\Omega\}$$

with the norm

$$\|u\|_{B_0^1(\Omega)} = \sup_{x \in \Omega} |u(x)| + \sum_{i=1}^n \sup_{x \in \Omega} \left| \frac{\partial u}{\partial x_i}(x) \right|.$$

Thus, for any $\varphi \in D((-L)^{m/2})$, φ belongs to $B_0^1(\Omega)$ and satisfies

$$(5.5) \quad \|\varphi\|_{B_0^1(\Omega)} \leq \text{const.} \cdot \|\varphi\|_{D((-L)^{m/2})}.$$

Put $g(x) = [\alpha(x) + \beta(x)]/2$. Then $g(x) \in B_0^1(\Omega)$. Since the inequality (5.4) holds, it is easy to see that the inequality

$$(5.6) \quad \alpha(x) \leq g(x) + \varphi(x) \leq \beta(x)$$

holds if $\varphi \in B_0^1(\Omega)$ and $\|\varphi\|_{B_0^1(\Omega)}$ is sufficiently small. Thus, by the inequality (5.5), we can choose $\eta > 0$ so small that the inequality (5.6) holds for any $\varphi \in D((-L)^{m/2})$ satisfying $\|\varphi\|_{D((-L)^{m/2})} \leq \eta$. Hence, it is shown that the constraint set contains the set

$$\bigcup_{T>0} \{g(x) + f(x, t) \mid f(x, t) \in L^\infty(0, T; D((-L)^{m/2}) \text{ satisfying } \|f(t)\|_{D((-L)^{m/2})} \leq \eta \text{ almost all } t\}.$$

Since Lemma 4.1 and 4.3 hold, the rest of proof is similar to the one of Theorem 4.2.

Similarly in the one-dimensional case, let us define the constraint set with smooth controls as

$$C_{\alpha,\beta}^\infty = \bigcup_{T>0} \{f(x, t) \in L^\infty(0, T; C^\infty(\Omega)) \mid \alpha(x) \leq f(x, t) \leq \beta(x) \text{ holds for almost all } t\}.$$

Then we have:

COROLLARY 5.1. *Let $\alpha(x)$ and $\beta(x)$ be in $C^1(\bar{\Omega})$ satisfying $\alpha(x) < \beta(x)$ in Ω and the inequality (5.4). Then the control system (5.1) with (5.2) is admissibly approximately controllable in the constraint set $C_{\alpha,\beta}^\infty$.*

Remark 5.1. The conclusions in Remark 4.1 also hold for the control system (5.1) with (5.2).

Remark 5.2. For simplicity we consider the Dirichlet boundary condition. But we have the similar results for the control system with more general boundary conditions,

$$\alpha u(x) + \beta \frac{\partial u}{\partial n}(x) = 0 \quad \text{on } \partial\Omega,$$

where α and β are constants satisfying $\alpha^2 + \beta^2 \neq 0$.

REFERENCES

- [1] R. BRAMMER, *Controllability in linear autonomous systems with positive controllers*, this Journal, 10 (1972), pp. 339–353.
- [2] G. CHEN, *Control and stabilization for the wave equation in a bounded domain*, this Journal, 17 (1979), pp. 66–81.
- [3] S. DOLECKI AND D. L. RUSSELL, *A general theory of observation and control*, this Journal, 15 (1977), pp. 185–220.
- [4] H. O. FATTORINI, *The time optimal problem for distributed control of systems described by the wave equation*, *Control Theory of Systems Governed by Partial Differential Equations*, A. K. Aziz, J. W. Wingate, M. J. Balas, eds., Academic Press, New York, 1977.
- [5] W. KRABS, *On boundary controllability of one dimensional vibrating systems*, *Math. Meth. in the Appl. Sci.*, 1 (1979), pp. 322–345.
- [6] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
- [7] K. NARUKAWA, *Admissible null controllability and optimal time control*, preprint.
- [8] D. L. RUSSELL, *Nonharmonic Fourier series in the control theory of distributed parameter systems*, *J. Math. Anal. Appl.*, 18 (1967), pp. 542–560.
- [9] ———, *Control theory of hyperbolic equations related to certain questions in harmonic analysis and spectral theory*, *J. Math. Anal. Appl.*, 40 (1972), pp. 336–368.
- [10] ———, *Exact boundary value controllability theorems for wave and heat processes in star-complemented regions*, *Differential Games and Control Theory*, Roxin, Liu, Sternberg, eds., Marcel Dekker, New York, 1974.
- [11] ———, *Controllability and stabilizability theory for linear partial differential equations: Recent progress and open questions*, *SIAM Rev.*, 20 (1978), pp. 639–739.
- [12] S. H. SAPERSTONE, *Global controllability of linear systems with positive controls*, this Journal, 11 (1973), pp. 417–423.
- [13] S. H. SAPERSTONE AND J. YORKE, *Controllability of linear oscillatory systems using positive controls*, this Journal, 9 (1971), pp. 253–262.
- [14] M. SLEMROD, *A note on complete controllability and stabilizability for linear control systems in Hilbert space*, this Journal, 12 (1974), pp. 500–508.
- [15] K. YOSIDA, *Lectures on Differential and Integral Equations*, Interscience, New York, 1960.

APPROXIMATING A SECOND-ORDER DIRECTIONAL DERIVATIVE FOR NONSMOOTH CONVEX FUNCTIONS*

J. B. HIRIART-URRUTY†

Abstract. For a lower-semicontinuous convex function f , the approximate second-order directional derivative $(d, \delta) \mapsto f''_\varepsilon(x_0; d, \delta)$ is defined through the ε -directional derivative $f'_\varepsilon(x; d)$. The function $v_d: x \mapsto v_d(x) = f'_\varepsilon(x; d)$ is, for all $\varepsilon > 0$, locally Lipschitz on $\text{int}(\text{dom} f)$ and, at those points where it is not differentiable, v_d admits a directional derivative $v'_d(x_0; \delta)$ for all δ , which we precisely denote by $f''_\varepsilon(x_0; d, \delta)$. The objective of the present work is two-fold: to classify all the possible differentiability properties of v_d according to the behavior of the function $\lambda \mapsto f(x_0 + \lambda d)$ on \mathbb{R}_+ , and to study the existence or nonexistence of the limit of $f''_\varepsilon(x_0; d, \delta)$ when $\varepsilon \rightarrow 0^+$.

Introduction. Defining a generalized second-order directional derivative $(d, \delta) \mapsto f''(x_0; d, \delta)$ for an arbitrary convex function f is of main concern in the current research in convex optimization. Introducing such an object, tractable from the computational viewpoint, is highly desirable to design second-order methods. Recently, some works have been devoted to an approximate second-order directional derivative $(d, \delta) \mapsto f''_\varepsilon(x_0; d, \delta)$ which is defined through the approximate directional derivative $d \mapsto f'_\varepsilon(x; d)$ [2], [7], [9]. The starting point was that $f'_\varepsilon(x; d)$ enjoys for $\varepsilon > 0$ some noteworthy properties different as for their nature from those of the "exact" directional derivative $f'(x; d)$. As a function of the state variable x , $f'(x; d)$ is upper-semicontinuous at all points x_0 of $\text{int}(\text{dom} f)$ and continuous at those points x_0 where f is differentiable. On the one hand, having available the $f'(x; d)$, $x \in \text{int}(\text{dom} f)$ and $d \in \mathbb{R}^n$, allows us to recover f from them since

$$(0.1) \quad f(x) = f(x_0) + \int_0^1 f'(x_0 + t(x - x_0); x - x_0) dt.$$

On the other hand, $f'(\cdot; d)$ is differentiable only at some privileged points which we shall consider later. The use of $f'_\varepsilon(x_0; d)$ as a substitute for $f'(x_0; d)$ turns out to be advantageous in many respects. The main property of the function

$$v_d: x \mapsto f'_\varepsilon(x; d)$$

is that it is *locally Lipschitz* on $\text{int}(\text{dom} f)$ [6], [11]. So v_d is differentiable almost everywhere on $\text{int}(\text{dom} f)$, which allows us to define a generalized gradient $\partial v_d(x_0)$ in Clarke's sense [3] at all $x_0 \in \text{int}(\text{dom} f)$. Moreover, at those points x_0 where it is not differentiable, v_d admits a directional derivative $v'_d(x_0; \delta)$ for all δ , which we shall also denote by $f''_\varepsilon(x_0; d, \delta)$. Since $v_d(x)$ can be written as the marginal function of a mathematical program,

$$(P) \quad v_d(x) = \max \{ \langle x^*, d \rangle \mid f^*(x^*) + f(x_0) - \langle x_0, x^* \rangle - \varepsilon \leq 0 \},$$

the *existence* of the directional derivatives of v_d was implicit in results like those of Gol'shtein [5] (the so-called mixed case) or Hogan [8]. However, the *exact expression* of $v'_d(x_0; \delta)$ had to be worked out in the particular case involved, and such a project has been initiated by Lemaréchal and Nurminskii [9]. The formula giving $v'_d(x_0; \delta)$ was proved by them under the assumption that f is finite and coercive (i.e.,

* Received by the editors March 12, 1981.

† Université de Clermont-Ferrand II, Complexe Scientifique des Cézéaux, Département de Mathématiques Appliquées, B.P. 45, 63170 Aubière, France. Current address, Université Paul Sabatier, 118, route de Narbonne, 31062 Toulouse Cédex, France.

$\lim_{\|x\| \rightarrow +\infty} f(x)/\|x\| = +\infty$); it involves mainly two sets:

$$\Lambda_d(x_0) = \{\lambda > 0 | [f(x_0 + \lambda d) - f(x_0) + \varepsilon] \lambda^{-1} = f'_\varepsilon(x_0; d)\}$$

and

$$\partial_\varepsilon f(x_0)_d = \{x^* \in \partial_\varepsilon f(x_0) | \langle x^*, d \rangle = f'_\varepsilon(x_0; d)\}.$$

The desire to get rid of the coercivity assumption and thereby cover the cases where $\Lambda_d(x_0)$ might be empty led Auslender [2] to rely on

$$M_d(x_0) = \{\mu \geq 0 | r_f(\mu) = f'_\varepsilon(x_0; d)\}$$

rather than $\Lambda_d(x_0)$. Here r_f is a convex function defined by

$$r_f(\mu) = \mu \left[f\left(x_0 + \frac{d}{\mu}\right) - f(x_0) + \varepsilon \right] \quad \text{if } \mu > 0, \\ r_f(0) = f_\infty(d) \quad \text{if } f_\infty(d) < +\infty.$$

The expression of $v'_d(x_0; \delta)$, involving $M_d(x_0)$ and $\partial_\varepsilon f(x_0)_d$, was proved by Auslender [2, § I] under the assumption that f is a finite function. Actually, his approach, which rests on duality results on the program (P), works for more general f , provided that the considered x_0 lies in $\text{int}(\text{dom } f)$. Our aim in the present study is two-fold: *to get a better insight into the differentiability properties of v_d and to study the behavior of $f''_\varepsilon(x_0; d, \delta)$ when ε converges to 0^+* . For the purposes of studying the differentiability properties of v_d (§ 2), our starting point will be the general expression of $v'_d(x_0; \delta)$. As already mentioned, this expression involves the set $\partial_\varepsilon f(x_0)_d$ whose calculation is usually out of reach. So we address ourselves to questions like: How to detect if v_d is differentiable or not, without explicitly calculating it? How to decide if $\partial v_d(x_0)$ contains 0 or not, having only $M_d(x_0)$ at our disposal? and so on. In answering these questions, we shall be led to a classification of all possible situations into three cases accordingly as $\Lambda_d(x_0)$ is nonempty and bounded, unbounded and empty. Each of these three fundamental situations yields peculiar properties as to the differentiability of v_d . We also conclude from the results in this section that there are only two cases as regards to the behavior of $\Lambda_d(x_0)$ when ε goes to 0^+ ; either $\Lambda_d(x_0)$ is nonempty and bounded for $\varepsilon < \bar{\varepsilon}$ or $\Lambda_d(x_0)$ is empty for all ε . It turns out that v_d is strictly differentiable at x_0 (i.e., $\partial v_d(x_0) = \{\nabla v_d(x_0)\}$) whenever it is differentiable at x_0 . The differentiability results of § 2 allow us to give a full description of the generalized gradient of v_d at those points where it is not differentiable (§ 3). Earlier results like those of Auslender [2, § II] and the author [7, § VII] are covered and improved. Moreover, general inner and outer estimates of $\partial v_d(x_0)$ are provided.

A nice feature of $f''_\varepsilon(x_0; d, \delta)$ is that it is defined at all $x_0 \in \text{int}(\text{dom } f)$. However, a fundamental question which has to be answered is the following: to what extent is $f''_\varepsilon(x_0; d, \delta)$ an approximation of the second derivative of f at x_0 ? Section 4 is devoted to this question through the analysis of the behavior of $f''_\varepsilon(x_0; d, \delta)$ when ε goes to 0^+ . In regard to the above question, one of the main results in this section is that $f''_\varepsilon(x_0; d; \delta)$ does have a limit when $\varepsilon \rightarrow 0^+$ at those points x_0 where f is twice differentiable in the Alexandroff's sense [1]. Even if the set of such points is of full measure in $\text{int}(\text{dom } f)$, the expansion

$$(0.2) \quad f'_\varepsilon(x_0 + \delta; d) = f'_\varepsilon(x_0; d) + \int_0^1 f''_\varepsilon(x_0 + t\delta; d, \delta) dt,$$

which is valid for all $\varepsilon > 0$, is not preserved when $\varepsilon \rightarrow 0^+$. This is in the line of difficulties

which arise when using the second derivative (defined almost everywhere) to identify a convex function. For the points x_0 where f is not twice differentiable, we shall sketch what to expect and what not to expect with regard to a definition of $f''(x_0; d, \delta)$.

1. Preliminary definitions and properties. Throughout the sequel, f will be a *lower-semicontinuous convex* function from \mathbb{R}^n into $(-\infty, +\infty]$. Given such a function, the ε -subdifferential of f at $x_0 \in \text{dom } f$ ($\text{dom } f$ is the set where f is finite) is defined for each $\varepsilon \geq 0$ as the set of vectors $x^* \in \mathbb{R}^n$ satisfying

$$(1.1) \quad f(x) \geq f(x_0) + \langle x^*, x - x_0 \rangle - \varepsilon$$

for all $x \in \mathbb{R}^n$. The set of such vectors, denoted by $\partial_\varepsilon f(x_0)$, is a closed convex set which reduces to the subdifferential $\partial f(x_0)$ when $\varepsilon = 0$. Moreover, $\partial_\varepsilon f(x_0)$ is known to be *nonempty and compact* whenever x_0 lies in $\text{int}(\text{dom } f)$. Since only these points will be of interest in the present study, we, henceforth, assume that $x_0 \in \text{int}(\text{dom } f)$.

There are two fundamental ways of characterizing $\partial_\varepsilon f(x_0)$: through the conjugate function f^* of f or in terms of approximate difference quotients of f . We recall these characterizations which both will be used throughout.

PROPOSITION 1.1. [14].

(a) $x^* \in \partial_\varepsilon f(x_0)$ if and only if

$$(1.2) \quad f(x_0) + f^*(x^*) - \langle x_0, x^* \rangle \leq \varepsilon;$$

(b) the support function of $\partial_\varepsilon f(x_0)$ is given as

$$(1.3) \quad d \mapsto f'_\varepsilon(x_0; d) = \inf_{\lambda > 0} \{ [f(x_0 + \lambda d) - f(x_0) + \varepsilon] \lambda^{-1} \}.$$

For fixed $\varepsilon > 0$ and $d \neq 0$, we denote by $\Lambda_d(x_0)$ the set of $\lambda_0 > 0$ for which

$$[f(x_0 + \lambda_0 d) - f(x_0) + \varepsilon] \lambda_0^{-1} = f'_\varepsilon(x_0; d).$$

The behavior of the approximate difference quotient

$$q_f : \lambda \mapsto [f(x_0 + \lambda d) - f(x_0) + \varepsilon] \lambda^{-1}$$

on \mathbb{R}_+^* is of particular importance with regard to deriving properties of $\Lambda_d(x_0)$. The behavior of q_f near 0^+ and $+\infty$ is known since

$$\lim_{\lambda \rightarrow 0^+} q_f(\lambda) = +\infty$$

and

$$\lim_{\lambda \rightarrow +\infty} q_f(\lambda) = \sup_{\lambda > 0} \{ [f(x_0 + \lambda d) - f(x_0)] \lambda^{-1} \} = f_\infty(d).$$

Here f_∞ is what is known as the *recession function* of f (or the *asymptotic function* of f). Since $x_0 \in \text{int}(\text{dom } f)$, there exists $\lambda^\uparrow \in]0, +\infty]$ such that q_f is finite on $]0, \lambda^\uparrow[$. Clearly, q_f is a lower-semicontinuous quasiconvex function on \mathbb{R}_+^* . Even more, q_f enjoys a *pseudoconvexity* property in the sense that the stationary points of q_f in \mathbb{R}_+^* are also the global minima of q_f on \mathbb{R}_+^* . Let us make that more precise. By stationary points of q_f on \mathbb{R}_+^* , we mean the $\lambda_0 \in \mathbb{R}_+^*$ for which the necessary condition for optimality

$$(1.4) \quad 0 \in \partial q_f(\lambda_0)$$

holds. Here, $\partial q_f(\lambda_0)$ stands for Clarke's *generalized gradient* of q_f at λ_0 [3], [16]. The definition and basic properties of the generalized gradient for *locally Lipschitz* functions will be recalled later. However, q_f is not necessarily locally Lipschitz around all

λ_0 satisfying (1.4) so that one has to rely on a general definition of ∂q_f such as given, for example, in [16]. Anyway, with this general definition and existing chain rules [16], it can be proved that

$$\partial q_f(\lambda_0) = [\langle \partial f(x_0 + \lambda d), d \rangle - q_f(\lambda_0)] \lambda_0^{-1}$$

for all $\lambda_0 > 0$. Therefore, the optimality condition (1.4) reduces to

$$(1.5) \quad q_f(\lambda_0) \in \langle \partial f(x_0 + \lambda d), d \rangle.$$

This condition turns out to be sufficient for $\lambda_0 \in \Lambda_d(x_0)$, as it can be directly verified from the definitions. The reason why the necessary condition for optimality is also sufficient can be explained by the following: the function r_f defined on \mathbb{R}_+^* by

$$r_f(\mu) = q_f\left(\frac{1}{\mu}\right)$$

is convex. Clearly, $\lambda_0 \in \mathbb{R}_+^*$ is a minimum of q_f on \mathbb{R}_+^* if and only if $\mu_0 = 1/\lambda_0$ is a minimum of r_f on \mathbb{R}_+^* . Now, the necessary and sufficient condition for optimality of $\mu_0 > 0$, $0 \in \partial r_f(\mu_0)$, is made equivalent to (1.4) by a chain rule (on generalized gradients of lower-semicontinuous functions [16]) which states that

$$\partial r_f(\mu_0) = -\partial q_f(\lambda_0) \lambda_0^2, \quad \lambda_0 \mu_0 = 1.$$

It might be more convenient to deal with r_f rather than q_f . From the computational viewpoint, it is easier to minimize the convex function r_f on \mathbb{R}_+ especially as

$$\lim_{\mu \rightarrow \infty} r_f(\mu) = +\infty, \quad \lim_{\mu \rightarrow 0^+} r_f(\mu) = f_\infty(d) > -\infty.$$

We agree on posing $r_f(0) = f_\infty(d)$ whenever $f_\infty(d) < +\infty$. Consequently, r_f always achieves its minimum value ($= f'_\varepsilon(x_0; d)$) on \mathbb{R}_+ . If we denote by $M_d(x_0)$ the set of $\mu_0 \in \mathbb{R}_+$ for which

$$r_f(\mu_0) = f'_\varepsilon(x_0; d),$$

$M_d(x_0)$ is a nonempty compact interval of \mathbb{R}_+ and

$$(1.6) \quad \Lambda_d(x_0) = \left\{ \frac{1}{\mu_0} \mid \mu_0 \in M_d(x_0), \mu_0 > 0 \right\}.$$

The definition itself of M_d and the continuity properties of f make that the multifunction $x \rightrightarrows M_d(x)$ is upper-semicontinuous at x_0 and bounded in a neighborhood of x_0 .

The case where q_f does not achieve the infimum value $f'_\varepsilon(x_0; d)$ on \mathbb{R}_+^* corresponds to the situation where $q_f(\lambda) > f_\infty(d)$ for all $\lambda > 0$; in terms of the function r_f this means that $\mu_0 = 0$ is the unique element of $M_d(x_0)$.

The main result concerning the behavior of the function

$$x \mapsto v_d(x) = f'_\varepsilon(x; d)$$

is that it is locally Lipschitz on $\text{int}(\text{dom } f)$ [6], [11]. So v_d has a generalized gradient ∂v_d in Clarke's sense [3] at all $x_0 \in \text{int}(\text{dom } f)$. In that respect, we recall the definition and basic properties of $\partial v_d(x_0)$. By definition, $\partial v_d(x_0)$ is the convex hull of the set

$$\{x^* \mid \exists x_k \rightarrow x_0 \text{ in } E_1 \text{ with } \nabla v_d(x_k) \rightarrow x^*\},$$

where E_1 is the set of full measure in $\text{int}(\text{dom } f)$ where v_d is differentiable. Actually, as observed by Clarke [3, Proposition 1.11], $\partial v_d(x_0)$ is blind to sets of null measure in the sense that the calculation of $\partial v_d(x_0)$ (as above) requires knowledge of $\nabla v_d(x)$

only for x on a set of full measure around x_0 . The support function of $\partial v_d(x_0)$ is the so-called *generalized directional derivative* of v_d defined as

$$\delta \mapsto v_d^0(x_0; \delta) = \limsup_{\substack{x \rightarrow x_0 \\ \lambda \rightarrow 0^+}} [v_d(x + \lambda \delta) - v_d(x)] \lambda^{-1}.$$

In terms of limits of linear mappings $\langle \nabla v_d(x), \delta \rangle$, we also have that

$$v_d^0(x_0; \delta) = \limsup_{\substack{x \rightarrow x_0 \\ x \in \Delta}} \langle \nabla v_d(x), \delta \rangle$$

for a set Δ of full measure around x_0 .

At those points x_0 where v_d is not differentiable, v_d admits however a directional derivative $v_d'(x_0; \delta)$ for all δ . The formula giving $v_d'(x_0; \delta)$ was proved by Lemaréchal and Nurminskii [9] under the assumption that f is finite and has an everywhere finite conjugate function f^* . For such a function (also called cofinite or coercive), $\Lambda_d(x_0)$ is nonempty and bounded for all x_0 and d , so that $v_d'(x_0; \delta)$ can be expressed in terms of $\Lambda_d(x_0)$. Indeed,

$$(1.7) \quad v_d'(x_0; \delta) = \min_{\lambda \in \Lambda_d(x_0)} \max_{x^* \in \partial_\epsilon f(x_0)_d} \{\lambda^{-1} [\langle x^*, \delta \rangle - f'(x_0; \delta)]\},$$

where

$$\partial_\epsilon f(x_0)_d = \{x^* \in \partial_\epsilon f(x_0) \mid \langle x^*, d \rangle = f'_\epsilon(x_0; d)\}.$$

The desire to get rid of the coercivity assumption and thereby cover the cases where $\Lambda_d(x_0)$ might be empty led Auslender [2] to rely on $M_d(x_0)$ rather than $\Lambda_d(x_0)$; the formula below giving $v_d'(x_0; \delta)$ was proved by him under the assumption that f is an everywhere finite function. Actually, this formula holds true for more general f than we are dealing with here provided that $x_0 \in \text{int}(\text{dom } f)$.

THEOREM 1.2. *The directional derivative of v_d at x_0 in the δ direction is given as*

$$(1.8) \quad v_d'(x_0; \delta) = \min_{\mu \in M_d(x_0)} \max_{x^* \in \partial_\epsilon f(x_0)_d} \{\mu [\langle x^*, \delta \rangle - f'(x_0; \delta)]\}.$$

In view of (1.6), one sees that the above formula is a generalization of the earlier one (1.7). $v_d'(x_0; \delta)$ is also noted as $f''_\epsilon(x_0; d, \delta)$; both notations will be used throughout.

The set $\partial_\epsilon f(x_0)_d$ is a certain face of $\partial_\epsilon f(x_0)$, the one consisting of all the points x^* at which d is a normal vector. In an alternate formulation, it is known that¹

$$\partial_\epsilon f(x_0)_d = \partial \psi_{\partial_\epsilon f(x_0)}^*(d).$$

Hence, the $\psi_{\partial_\epsilon f(x_0)_d}^*(\delta)$ involved in the formulation of $v_d'(x_0; \delta)$ can be rephrased as

$$\psi_{\partial_\epsilon f(x_0)_d}^*(\delta) = \lim_{\sigma \rightarrow 0^+} [f'_\epsilon(x_0; d + \sigma \delta) - f'_\epsilon(x_0; d)] \sigma^{-1}$$

The expression of $v_d'(x_0; \delta)$ can be simplified for some directions δ . Let L be the linear space associated with the affine hull of $\partial_\epsilon f(x_0)_d$ and let $\mathcal{A}(d)$ be the orthogonal complement of L ($1 \leq \dim \mathcal{A}(d) \leq n$). For a $\delta \in \mathcal{A}(d)$, the linear form $x^* \mapsto \langle x^*, \delta \rangle$ is constant on $\partial_\epsilon f(x_0)_d$ or, in other words, the breadth of $\partial_\epsilon f(x_0)_d$ in the δ direction is null. Thus, for such a δ , we have that

$$v_d'(x_0; \delta) = \min_{\mu \in M_d(x_0)} \{\mu [\alpha(\delta) - f'(x_0; \delta)]\},$$

¹ ψ_A^* denotes the support function of A .

where $\alpha(\delta)$ is the constant $\langle \partial_\varepsilon f(x_0)_d, \delta \rangle$. In particular, $\mathcal{A}(d)$ is the whole space \mathbb{R}^n whenever $\partial_\varepsilon f(x_0)_d$ is reduced to a single element $x_d^*(x_0)$.

The ordinary directional derivative $v'_d(x_0; \delta)$ and the generalized directional derivative $v_d^0(x_0; \delta)$ are both defined at x_0 and

$$v'_d(x_0; \delta) \leq v_d^0(x_0; \delta).$$

v_d will be called *strictly tangentially convex* at x_0 if

$$v'_d(x_0; \delta) = v_d^0(x_0; \delta)$$

for all $\delta \in \mathbb{R}^n$. This property is stronger than requiring the convexity of $\delta \mapsto v'_d(x_0; \delta)$ (*tangential convexity* of v_d at x_0)². Similarly, v_d will be said to be *strictly tangentially concave* at x_0 if $-v_d$ is strictly tangentially convex at x_0 .

To end these preliminaries, we recall the framework we work in: $x_0 \in \text{int}(\text{dom } f)$, $d \neq 0$ and $\varepsilon > 0$. Strictly speaking, all the objects defined, $\Lambda_d(x_0)$, $M_d(x_0)$, v_d , ... also depend on ε . We shall drop the ε for a while and put it back when the behavior in ε will be considered (§ 4).

2. Differentiability points of v_d .

THEOREM 2.1. *Let x be a point around x_0 where v_d is differentiable. Then $M_d(x)$ is single-valued ($M_d(x) = \{\mu_d(x)\}$) and we have that*

$$(2.1) \quad \mu_d(x) \partial_\varepsilon f(x)_d = \nabla v_d(x) + \mu_d(x) \partial f(x).$$

Moreover, at such an x , $\nabla v_d(x)$ is nonnull if and only if $\mu_d(x) > 0$.

As a consequence, we note that $\partial_\varepsilon f(x)_d$ is singlevalued ($\partial_\varepsilon f(x)_d = \{x_d^*(x)\}$) whenever both v_d and f are differentiable at x and $\mu_d(x) > 0$. In such a case³,

$$(2.2) \quad x_d^*(x) = \mu_d(x)^{-1} \nabla v_d(x) + \nabla f(x).$$

Before proving Theorem 2.1, we shall prove a technical lemma which will be useful in the sequel.

LEMMA 2.2. *Suppose $\Lambda_d(x)$ is nonempty; then $f'_\varepsilon(x; d) > f'(x; d)$.*

Proof. According to the definitions themselves,

$$[f(x + \lambda_0 d) - f(x) + \varepsilon] \lambda_0^{-1} = f'_\varepsilon(x; d)$$

for all $\lambda_0 \in \Lambda_d(x)$, while

$$\inf_{\lambda > 0} \{[f(x + \lambda d) - f(x)] \lambda^{-1}\} = f'(x; d).$$

Hence, we have the desired strict inequality since $\varepsilon > 0$. \square

Proof of Theorem 2.1. Let x be a point (around x_0) where v_d is differentiable. For all $\delta \in \mathbb{R}^n$, we have that

$$(2.3) \quad \min_{\mu} \max_{x^*} \{\mu [\langle x^*, \delta \rangle - f'(x; \delta)]\} = \langle \nabla v_d(x), \delta \rangle.$$

By writing $\langle \nabla v_d(x), \delta \rangle = -\langle \nabla v_d(x), -\delta \rangle$, the above relation yields

$$(2.4) \quad \min_{\mu} \max_{x^*} \{\mu [\langle x^*, \delta \rangle - f'(x; \delta)]\} = \max_{\mu} \min_{x^*} \{\mu [\langle x^*, \delta \rangle + f'(x; -\delta)]\} = \langle \nabla v_d(x), \delta \rangle$$

for all δ .

² We feel that these appellations are better than the earlier ones: *quasidifferentiable* for tangential convex and (*subdifferentially*) *regular* for strictly tangentially convex. More particularly, the tangentially linear case is the differentiable case while the strictly linear case reduces to the strictly differentiable case.

³ $\mu_d(x)$ is defined as the unique element of $M_d(x)$ whenever M_d is single-valued at x .

Let us make, for example, $\delta = d$ in (2.4); we get that

$$(2.5) \quad \min_{\mu} \{\mu[f'_\varepsilon(x; d) - f'(x; d)]\} = \max_{\mu} \{\mu[f'_\varepsilon(x; d) + f'(x; -d)]\}.$$

Since $-f'(x; d) \leq f'(x; -d)$, the above implies that

$$(2.6) \quad \max_{\mu} \{\mu[f'_\varepsilon(x; d) - f'(x; d)]\} = \min_{\mu} \{\mu[f'_\varepsilon(x; d) - f'(x; d)]\}.$$

That will be our key relation.

If $f'_\varepsilon(x; d) > f'(x; d)$, we derive from (2.6) that $M_d(x) = \{\mu_d(x)\}$. Now if $f'_\varepsilon(x; d) = f'(x; d)$, we know by Lemma 2.2 that $\Lambda_d(x)$ is necessarily empty so that $M_d(x) = \{0\}$. So, in any case, $M_d(x)$ is single-valued at the considered x .

Now (2.3) can be rewritten as

$$\max_{x^* \in \partial_\varepsilon f(x)_d} [\mu_d(x) \langle x^*, \delta \rangle] = \langle \nabla v_d(x), \delta \rangle + \mu_d(x) f'(x; \delta)$$

for all δ . Whence the expression of $\mu_d(x) \partial_\varepsilon f(x)_d$.

Clearly, $\nabla v_d(x)$ is null whenever $\mu_d(x)$ is null. Conversely, suppose that $\nabla v_d(x) = 0$. According to the relation (2.1) just proved, we have that

$$\mu_d(x) \partial_\varepsilon f(x)_d = \mu_d(x) \partial f(x).$$

If $\mu_d(x)$ was not null, the above would imply that $\partial_\varepsilon f(x)_d = \partial f(x)$ and thus $f'_\varepsilon(x; d) = f'(x; d)$, which is in contradiction with the nonnullity of $\mu_d(x)$ (Lemma 2.2). Hence, the result is proved. \square

The results of Theorem 2.1 can be made more precise if more is known on the behavior of the function q_f on \mathbb{R}_+^* . Actually, there are three possible situations we are going to study in detail now.

In the first place, let d satisfy the following at x_0 :

$$(S_1) \quad \exists \lambda_*, \quad [f(x_0 + \lambda_* d) - f(x_0) + \varepsilon] \lambda_*^{-1} < f_\infty(d).$$

This condition is necessary and sufficient for $\Lambda_d(x_0)$ to be *nonempty and bounded*. In alternate ways, (S_1) is necessary and sufficient for having $0 \notin M_d(x_0)$ or for securing $f'_\varepsilon(x_0; d) < f_\infty(d)$. (S_1) is certainly satisfied by any d for which $f_\infty(d) = +\infty$.

THEOREM 2.3. *Let d satisfy (S_1) . There then exists a neighborhood V of x_0 such that $\mu_d(x) > 0$ and*

$$(2.7) \quad \partial_\varepsilon f(x)_d = \mu_d(x)^{-1} \nabla v_d(x) + \partial f(x)$$

whenever v_d is differentiable at $x \in V$.

An immediate corollary to the above result is as follows:

COROLLARY 2.4. *Let d satisfy (S_1) . Then the following assertions are equivalent in a neighborhood of x_0 :*

- (a) v_d and f are differentiable at x ;
- (b) $M_d(x) = \{\mu_d(x)\}$, $\partial_\varepsilon f(x)_d = \{x_d^*(x)\}$, $\partial f(x) = \{\nabla f(x)\}$.

At such an x , $\nabla v_d(x)$ is expressed as

$$(2.8) \quad \nabla v_d(x) = \mu_d(x) [x_d^*(x) - \nabla f(x)].$$

Proof of Theorem 2.3. Due to the upper-semicontinuity of the set-valued mapping M_d , $M_d(x) \subset \mathbb{R}_+^*$ for all x in a neighborhood V of x_0 . Therefore, it is a consequence of Theorem 2.1 that $\mu_d(x) > 0$ whenever v_d is differentiable at $x \in V$. \square

Remark 1. Note that $\nabla v_d(x)$ is nonnull whenever it exists in V . Along the same lines, we observe that $f'_\varepsilon(x; d) > f'(x; d)$ for all x in V ; consequently,

$$v'_d(x; d) = f''_\varepsilon(x; d, d) > 0$$

for all x in V .

Remark 2. In the situation we are concerned with, the differentiability of v_d at $x \in V$ does not secure the differentiability of f at x . Nevertheless, the following partial differentiability result is easily derived from (2.7):

$$f'(x; \delta) = -f'(x; -\delta) \text{ for all } \delta \in \mathcal{A}(d),$$

whenever v_d is differentiable at $x \in V$.

We now turn our attention to the cases where $f'_\varepsilon(x_0; d) = f_\infty(d)$. To begin with, let us consider d such that

$$(S_2) \quad \min_{\lambda > 0} \{ [f(x_0 + \lambda d) - f(x_0) + \varepsilon] \lambda^{-1} \} = f_\infty(d).$$

This condition is necessary and sufficient for $\Lambda_d(x_0)$ to be *nonempty and unbounded*. In other words, (S_2) corresponds to the situation where $M_d(x_0)$ contains 0 but is not reduced to it. Actually, the case we deal with here is somewhat peculiar. If $\underline{\lambda} = \min \{ \lambda \mid \lambda \in \Lambda_d(x_0) \}$, all the $\lambda \geq \underline{\lambda}$ are in $\Lambda_d(x_0)$, and one easily checks that

$$f(x_0 + \lambda d) = f(x_0) - \varepsilon + \lambda f_\infty(d)$$

for all $\lambda \geq \underline{\lambda}$. Moreover, we note that if d satisfies (S_2) for a certain $\bar{\varepsilon}$, then d satisfies (S_1) for all $\varepsilon < \bar{\varepsilon}$.

Contrary to (S_1) , the condition (S_2) is not necessarily secured in a neighborhood of x_0 whenever it is satisfied at x_0 and that will be the cause of the differences with the previous situation.

THEOREM 2.5. *Let d satisfy (S_2) at x_0 . Given an open neighborhood V of x_0 , let V_0 denote the set of $x \in V$ where v_d is differentiable. Then $\nabla v_d(x)$ is nonnull (or equivalently $\mu_d(x) > 0$) on a subset of positive measure W of V_0 .*

Proof. We have to prove that ∇v_d does not vanish almost everywhere in V . If that was the case, we would have $\partial v_d(x_0) = \{0\}$ and thus v_d differentiable at x_0 . But, since $M_d(x_0)$ is not reduced to $\{0\}$, that contradicts the first statement in Theorem 2.1. \square

Remark. As in the previous situation, $\partial_\varepsilon f(x)_d$ is single-valued at those points $x \in W$ where f is differentiable. Also note that $f''_\varepsilon(x; d, d) > 0$ almost everywhere in W but certainly not at x_0 . As for v'_d at x_0 , the following can be observed:

$$v'_d(x_0; d) = f''_\varepsilon(x_0; d, d) = 0, \quad v'_d(x_0; \delta) \leq 0 \quad \text{for all } \delta \in \mathbb{R}^n$$

and $v'_d(x_0; \delta) < 0$ for a certain δ .

The third possible situation for d is as follows:

$$(S_3) \quad [f(x_0 + \lambda d) - f(x_0) + \varepsilon] \lambda^{-1} > f_\infty(d) \text{ for all } \lambda > 0.$$

This condition is necessary and sufficient for $\Lambda_d(x_0)$ to be *empty* and corresponds to the case where $M_d(x_0) = \{0\}$. It is also equivalent to the statement: “ v_d is differentiable at x_0 with $\nabla v_d(x_0) = 0$ ” (Theorem 2.1). Here, contrary to what holds in the situation (S_2) , $\mu_d(x)$ can be null almost everywhere in a neighborhood of x_0 .

It is worthwhile to break the present case into two, accordingly as $f'_\varepsilon(x_0; d) - f'(x_0; d)$ is strictly positive or not.

Let us examine the first case, namely when

$$(S_{3.a}) \quad (S_3) \text{ holds and } f'_\varepsilon(x_0; d) > f'(x_0; d).$$

The first thing to be observed is that $(S_{3.a})$ is a transient stage in the sense that if $(S_{3.a})$ holds for a certain ε one then ends by getting the final stage (S_1) (and remaining in it) when ε goes to 0.

The second observation concerns all the situations (including (S_1) and (S_2)) where $f'_\varepsilon(x_0; d) > f'(x_0; d)$. Due to the definition of $\partial_\varepsilon f(x_0)_d$, one easily checks that

$$(2.9) \quad f'_\varepsilon(x_0; d) > f'(x_0; d) \Leftrightarrow \partial f(x_0) \cap \partial_\varepsilon f(x_0)_d = \emptyset.$$

Now since the function $x \mapsto f'_\varepsilon(x; d) - f'(x; d)$ is lower-semicontinuous at x_0 , there exists a neighborhood V of x_0 such that

$$f'_\varepsilon(x; d) > f'(x; d)$$

for all $x \in V$. As a consequence, $\partial f(x) \cap \partial_\varepsilon f(x)_d = \emptyset$ whenever x lies in V . This result must be put together with the relationship between $\partial f(x_0)$ and $\partial_\varepsilon f(x_0)_d$ stated when v_d is differentiable at x and $\mu_d(x) > 0$ (Theorem 2.1).

The second case to be considered is the one where

$$(S_{3.b}) \quad f'_\varepsilon(x_0; d) = f'(x_0; d) (= f_\infty(d))$$

and thus

$$f(x_0 + \lambda d) = f(x_0) + \lambda f'(x_0; d)$$

for all $\lambda > 0$. Here $(S_{3.b})$ is a permanent stage in the sense that if $(S_{3.b})$ holds for some ε_0 it then holds for all ε . Contrary to the previous case, one may have $\partial f(x) \cap \partial_\varepsilon f(x)_d \neq \emptyset$ for some x around x_0 (actually this holds for all x , including x_0 , where $f'_\varepsilon(x; d) = f'(x; d)$). Even more, it might happen that $f'_\varepsilon(x; d) = f'(x; d)$ for all x in a neighborhood of x_0 ; an example of this situation is $f(x) = |x|$ at $x_0 = 0$.

As a general rule, the assumption (S_3) does not guarantee that $\partial_\varepsilon f(x_0)_d$ is single-valued, even when f is differentiable at x_0 . The next result gives an expression of $\partial_\varepsilon f(x_0)_d$ peculiar to the situations (S_2) and (S_3) .

THEOREM 2.6. *Let d satisfy (S_2) or (S_3) at x_0 . We then have*

$$(2.10) \quad \partial_\varepsilon f(x_0)_d = \partial_\varepsilon f(x_0) \cap \partial f_\infty(d).$$

Proof. A necessary and sufficient condition for d to satisfy (S_2) or (S_3) at x_0 is that $f'_\varepsilon(x_0; d) = f_\infty(d)$. Now, f_∞ is the support function of $\text{dom } f^*$ [14, Thm 13.3], whence $x^* \in \partial_\varepsilon f(x_0)_d$ if and only if

$$x^* \in \partial_\varepsilon f(x_0)$$

and

$$\langle x^*, d \rangle = \psi^*(d | \text{dom } f^*).$$

The latter relation is equivalent to $x^* \in \partial \psi^*(d | \text{dom } f^*)$ since the considered x^* lies in $\text{dom } f^*$. Hence the result (2.10) is proved. \square

Remark. If not within the situation (S_1) , it is not generally true that $\partial_\varepsilon f(x)_d$ is single-valued almost everywhere around x_0 ; see Example 3.2. More will be said about $\partial_\varepsilon f(x)_d$ later on.

3. Generalized gradient and generalized directional derivative of v_d .

3.1. At a point x (around x_0) where both v_d and f are differentiable, we have that

$$(3.1) \quad \nabla v_d(x) = \mu_d(x)[x_d^*(x) - \nabla f(x)],$$

where $x_d^*(x)$ is the *unique* element of $\partial_\epsilon f(x)_d$ when $\mu_d(x) > 0$ and *any* element of $\partial_\epsilon f(x)_d$ whenever $\mu_d(x) = 0$ (Theorem 2.1). With (3.1) and the formulation of $v_d'(x_0; \delta)$, we are in a position to derive general differentiability results on v_d .

It is convenient to pose Δ as the set of full measure around x_0 where both v_d and f are differentiable.

THEOREM 3.1. $\partial v_d(x_0)$ is the convex hull of the compact set

$$(3.2) \quad \{x^* | \exists x_k \rightarrow x_0 \text{ in } \Delta \text{ with } \mu_d(x_k) [x_d^*(x_k) - \nabla f(x_k)] \rightarrow x^*\},$$

and $v_d^0(x_0; \delta)$ is given for all δ as

$$(3.3) \quad v_d^0(x_0; \delta) = \limsup_{x \rightarrow x_0} v_d'(x; \delta).$$

Proof. Following what has been recalled in § 1, the calculation of $\partial v_d(x_0)$ demands knowledge of $\nabla v_d(x)$ for all x in a set of full measure around x_0 , what is done in Δ by (3.1).

Concerning the generalized directional derivative, we have that

$$v_d^0(x_0; \delta) = \limsup_{\substack{x \rightarrow x_0 \\ \Delta}} \langle \nabla v_d(x), \delta \rangle$$

for all δ , whence

$$v_d^0(x_0; \delta) \leq \limsup_{x \rightarrow x_0} v_d'(x; \delta).$$

For the converse inequality, we observe that

$$v_d'(x; \delta) \leq v_d^0(x; \delta)$$

and that the mapping $x \mapsto v_d^0(x; \delta)$ is upper-semicontinuous at x_0 . Hence, the announced result is proved. \square

As a general rule, $\partial v_d(x_0)$ contains $\nabla v_d(x_0)$ when the latter exists. The next result states that actually $\partial v_d(x_0)$ is reduced to $\{\nabla v_d(x_0)\}$ whenever v_d is differentiable at x_0 .

PROPOSITION 3.2. *Suppose v_d is differentiable at x_0 ; then v_d is strictly differentiable at x_0 .*

Proof. Suppose firstly that $\nabla v_d(x_0) = 0$. $M_d(x_0)$ is then reduced to the 0 element (Theorem 2.1.) Due to the upper-semicontinuity of M_d , all the $\mu_d(x_k)$ involved in the calculation of $\partial v_d(x_0)$ [formula (3.2)] converge to 0. Thus, $\partial v_d(x_0) = \{0\}$.

The general case boils down to the above one by setting

$$g(x) = f(x) - \langle \nabla v_d(x_0), x \rangle.$$

For such a g ,

$$w_d(x) = v_d(x) - \langle \nabla v_d(x_0), d \rangle,$$

where $w_d(x)$ stands for $g'_\epsilon(x; d)$. Hence, the desired result is achieved since

$$\partial w_d(x_0) = \partial v_d(x_0) - \{\nabla v_d(x_0)\}$$

and

$$\partial w_d(x_0) = \{\nabla w_d(x_0)\} = \{0\}. \quad \square$$

In § 2, equivalent assumptions to (S_1) , (S_2) and (S_3) were displayed in terms of $\Lambda_d(x_0)$ or $M_d(x_0)$. The next statement gives a further characterization of the (S_i) , expressed this time in terms of $\partial v_d(x_0)$.

PROPOSITION 3.3. *The following equivalences hold:*

- (i) *d satisfies (S_1) at x_0 if and only if $0 \notin \partial v_d(x_0)$;*
- (ii) *d satisfies (S_2) at x_0 if and only if $\partial v_d(x_0)$ contains 0 but is not reduced to it;*
- (iii) *d satisfies (S_3) at x_0 if and only if $\partial v_d(x_0) = \{0\}$.*

Proof. (iii) is a mere consequence of Proposition 3.2 and an earlier characterization of (S_3) .

For (ii) let d satisfy (S_2) at x_0 . As noticed in the remark following the proof of Theorem 2.5, v_d is not differentiable at x_0 and $v'_d(x_0; \delta) \leq 0$ for all δ . Consequently, ∂v_d is not single-valued at x_0 and

$$v_d^0(x_0; -\delta) \geq -v'_d(x_0; \delta) \geq 0$$

for all δ , whence $0 \in \partial v_d(x_0)$. For the converse, the only thing to prove is that $0 \in M_d(x_0)$ whenever $0 \in \partial v_d(x_0)$. Since the three set-valued mappings M_d , $\partial_\epsilon f(\cdot)_d$ and ∂f are upper-semicontinuous at x_0 , the formula (3.2) yields the following estimate of $\partial v_d(x_0)$,

$$(3.4) \quad \partial v_d(x_0) \subset M_d(x_0)[\partial_\epsilon f(x_0)_d - \partial f(x_0)].$$

If $\partial f(x_0) \cap \partial_\epsilon f(x_0)_d$ was nonempty, one would have $f'_\epsilon(x_0; d) = f'(x; d)$ (relation (2.9)) and thus $M_d(x_0) = \{0\}$ (Lemma 2.2), which is the situation (S_3) . Therefore, $0 \notin \partial_\epsilon f(x_0)_d - \partial f(x_0)$ and (3.4) then implies that $0 \in M_d(x_0)$.

As for the equivalence (i), it is simply derived from the two other ones. \square

Generally speaking, the directional derivative (function)

$$v'_d : \delta \mapsto v'_d(\delta) = v'_d(x_0; \delta)$$

is neither convex nor concave. So, it is worthwhile to exhibit the general expressions of the biconjugate functions $(v'_d)^{**}$ and $(-v'_d)^{**}$. Since v'_d is positively homogeneous with $v'_d(0) = 0$, $(v'_d)^{**}$ and $(-v'_d)^{**}$ are support functions of convex sets we propose to determine now. For that, let us recall the definition of the $*$ -difference of sets, such as introduced by Pontryagin [12]. Given two sets, A and B , the $*$ -difference of A and B , denoted as A^*B , consists of vectors x satisfying the condition $x + B \subset A$. Clearly, A^*B is convex whenever A and B are convex. As for example, the result (2.1) in Theorem 2.1 can be rewritten as

$$(3.5) \quad \nabla v_d(x) = \mu_d(x)[\partial_\epsilon f(x_d)_d^* \partial f(x)],$$

with the convention that $0 \cdot \emptyset = 0$.

The next statement is a particular case of a more general result by Pshenichnyi [13, p. 179–182] on the conjugate of the difference of two convex functions.

LEMMA 3.4. *Let A and B be two nonempty compact convex sets. Then*

$$(3.6) \quad \psi_{A^*B}^* = (\psi_A^* - \psi_B^*)^{**}.$$

In view of the above, v'_d can be reformulated as

$$(3.7) \quad v'_d = \min_{\mu \in M_d(x_0)} [\psi_{\mu \partial_\epsilon f(x_0)_d}^* - \psi_{\mu \partial f(x_0)}^*].$$

We, therefore, are in a position to express $(v'_d)^{**}$ and $(-v'_d)^{**}$ as support functions of convex sets involving $*$ -differences.

THEOREM 3.5. *$(v'_d)^{**}$ is the support function of*

$$\Gamma_1(x_0) = \bigcap_{\mu \in M_d(x_0)} \{\mu[\partial_\epsilon f(x_0)_d^* \partial f(x_0)]\},$$

while $(-v'_d)^{**}$ is the support function of

$$\Gamma_2(x_0) = M_d(x_0)[\partial f(x_0) \stackrel{*}{\circ} \partial_{\varepsilon} f(x_0)_d].$$

Proof. We have that

$$(v'_d)^* = \sup_{\mu} [\psi_{\mu \partial_{\varepsilon} f(x_0)_d}^* - \psi_{\mu \partial f(x_0)}^*]^*$$

[14, p. 149]. Since $\mu \partial_{\varepsilon} f(x_0)_d - \mu \partial f(x_0)$ is closed and convex, (3.6) gives

$$[\psi_{\mu \partial_{\varepsilon} f(x_0)_d}^* - \psi_{\mu \partial f(x_0)}^*]^* = \psi_{[\mu \partial_{\varepsilon} f(x_0)_d \stackrel{*}{\circ} \partial f(x_0)]},$$

whence the expression of $(v'_d)^{**}$ is immediately derived.

$(-v'_d)^{**}$ is the closed convex hull of the collection

$$\{\psi_{\mu \partial f(x_0)}^* - \psi_{\mu \partial_{\varepsilon} f(x_0)_d}^*\}^* = \psi_{\mu \partial f(x_0) \stackrel{*}{\circ} \mu \partial_{\varepsilon} f(x_0)_d} \mid \mu \in M_d(x_0).$$

Hence, $(-v'_d)^{**}$ is the support function of

$$\text{co} \left\{ \bigcup_{\mu \in M_d(x_0)} \mu [\partial f(x_0) \stackrel{*}{\circ} \partial_{\varepsilon} f(x_0)_d] \right\}.$$

But since $M_d(x_0) \subset \mathbb{R}_+$, the above set is nothing else than $\Gamma_2(x_0)$. \square

Since we have

$$\max \{(v'_d)^{**}(\delta), (-v'_d)^{**}(-\delta)\} \leq v_d^0(x_0; \delta)$$

for all δ , the theorem just proved induces that both $\Gamma_1(x_0)$ and $-\Gamma_2(x_0)$ are included in $\partial v_d(x_0)$, a comparison result which must be put together with the inclusion (3.4). However, except in the case where $\partial_{\varepsilon} f(x_0)_d$ is a shifted copy of $\partial f(x_0)$ (like in (3.5) when $\mu_d(x_0) > 0$), both $\partial_{\varepsilon} f(x_0)_d \stackrel{*}{\circ} \partial f(x_0)$ and $\partial f(x_0) \stackrel{*}{\circ} \partial_{\varepsilon} f(x_0)_d$ cannot be nonempty at the same time. So, we are led to examine situations where one or more among M_d , $\partial_{\varepsilon} f(\cdot)_d$, ∂f are single-valued at x_0 .

The first case we consider is when $M_d(x_0)$ is reduced to one element, namely $\mu_d(x_0)$. The statement below improves an earlier result by Auslender [2, Thm. 2] where the expressions of $\partial v_d(x_0)$ are derived under the stronger assumption that M_d is single-valued in a neighborhood of x_0 .

COROLLARY 3.6. Assume $M_d(x_0) = \{\mu_d(x_0)\}$. Then

$$(3.8) \quad \partial v_d(x_0) = \mu_d(x_0) \text{ co } \{u^* \mid \exists x_k \rightarrow x_0 \text{ in } \Delta \text{ with } x_k^* - \nabla f(x_k) \rightarrow u^*\}.$$

Proof. Since M_d is upper-semicontinuous at x_0 and $M_d(x_0) = \{\mu_d(x_0)\}$, all the $\mu_d(x_k)$ involved in the calculation of $\partial v_d(x_0)$ [formula (3.2)] converge to $\mu_d(x_0)$. Hence, the expression (3.8) of $\partial v_d(x_0)$. \square

In the situation we are concerned with, earlier estimates of $\partial v_d(x_0)$ become

$$(3.9) \quad \mu_d(x_0)[\partial_{\varepsilon} f(x_0)_d \stackrel{*}{\circ} \partial f(x_0)] \subset \partial v_d(x_0),$$

$$(3.10) \quad -\mu_d(x_0)[\partial f(x_0) \stackrel{*}{\circ} \partial_{\varepsilon} f(x_0)_d] \subset \partial v_d(x_0),$$

$$(3.11) \quad \partial v_d(x_0) \subset \mu_d(x_0)[\partial_{\varepsilon} f(x_0)_d - \partial f(x_0)].$$

All these inclusions may be strict. Equality holds in (3.9) if and only if v_d is strictly tangentially convex at x_0 , i.e., $v'_d(x_0; \delta) = v_d^0(x_0; \delta)$ for all δ . Both inclusions (3.9) and (3.11) become equalities if and only if $\mu_d(x_0) = 0$ or f is differentiable at x_0 .

Similarly, equality holds in (3.10) if and only if v_d is strictly tangentially concave at x_0 , i.e., $v'_d(x_0; \delta) = -v_d^0(x_0; -\delta)$ for all δ . Inclusions (3.10) and (3.11) are equalities if and only if $\mu_d(x_0) = 0$ or $\partial_{\varepsilon} f(\cdot)_d$ is single-valued at x_0 .

The second case to be considered is when $\partial_\varepsilon f(x_0)_d$ is reduced to the single element $x_d^*(x_0)$. Among others, this situation holds at all $x_0 \in \text{int}(\text{dom } f)$ for all f defined on the real line. Combining the earlier estimate (3.4) and the fact that $\partial v_d(x_0)$ always contains $-\Gamma_2(x_0)$, one readily gets the following result.

COROLLARY 3.7. *Assume $\partial_\varepsilon f(x_0)_d = \{x_d^*(x_0)\}$. Then v_d is strictly tangentially concave at x_0 and*

$$(3.12) \quad \partial v_d(x_0) = M_d(x_0)[x_d^*(x_0) - \partial f(x_0)].$$

The above improves an earlier result by the author [7, Corollary 7.6] where the expression (3.12) was stated under the stronger assumption that $\partial_\varepsilon f(\cdot)_d$ is single-valued in a neighborhood of x_0 .

3.2 Examples. The next two examples serve as illustrations of the foregoing. They are simple enough to make all the calculations easy and versatile enough to illustrate various situations.

Example 3.1. Let $f: \mathbb{R} \rightarrow \mathbb{R}$ be defined as

$$f(x) = \max(-2x, -x, x-2).$$

We set $d = 1$ and $\varepsilon = 2$. The domain \mathbb{R} of f is partitioned into three subsets accordingly as (S_1) , (S_2) or (S_3) is satisfied. For the afore-defined f , we have the following:

x_0	$x_0 < -2$	$x_0 = -2$	$-2 < x_0 < 0$	$x_0 = 0$	$0 < x_0 < 1$	$x_0 \geq 1$
Situation	(S_1)	(S_1)	(S_1)	(S_2)	$(S_{3,a})$	$(S_{3,b})$
$M_d(x_0)$	$\left\{-\frac{1}{x_0}\right\}$	$\left[\frac{1}{3}, \frac{1}{2}\right]$	$\left\{\frac{1}{1-x_0}\right\}$	$[0, 1]$	$\{0\}$	$\{0\}$
$x_d^*(x_0)$	$\left\{\frac{-2(x_0+1)}{x_0}\right\}$	$\{-1\}$	$\left\{\frac{2x_0+1}{1-x_0}\right\}$	$\{1\}$		

v_d is known to be strictly tangentially concave on \mathbb{R} (Corollary 3.7) and the only points where v_d is not differentiable are -2 and 0 . Since $v_d(x)$ is merely $x_d^*(x)$ here, one can check all the results and formulas encountered in § 2 and § 3.1.

Example 3.2. Let $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ be defined as

$$f(\xi_1, \xi_2) = |\xi_1| + \xi_2^+.$$

We set $d = (0, 1)$ and ε an arbitrary positive number. There clearly exists a neighborhood V_ε of $x_0 = (0, 0)$ such that

$$\partial_\varepsilon f(x) = \partial f(x_0) = \text{co} \{(1, 0), (-1, 0), (1, 1), (-1, 1)\},$$

and

$$M_d(x) = \{0\}$$

for all x in V_ε . Since f_∞ equals f here, this example can serve as an illustration of Theorem 2.6 with

$$\partial_\varepsilon f(x)_d = \text{co} \{(1, 1), (-1, 1)\}$$

for all $x \in V_\varepsilon$.

Preserving the same f and d as above, let now $x_0 = (0, -2)$ and $\varepsilon = 1$. There exists a neighborhood V of x_0 such that, whenever $x = (\xi_1, \xi_2)$ lies in V ,

$$M_d(x) = \{\mu_d(x)\} = \left\{ -\frac{1}{\xi_2} \right\},$$

$$\partial_\varepsilon f(x)_d = \begin{cases} \left\{ \left(1, -\frac{1}{\xi_2} \right) \right\} & \text{if } \xi_1 > 0, \\ \left\{ \left(-1, -\frac{1}{\xi_2} \right) \right\} & \text{if } \xi_1 < 0. \end{cases}$$

Consequently v_d is differentiable at x_0 with $\nabla v_d(x_0) = (0, 1/4)$. At this x_0 we have

$$\partial_\varepsilon f(x_0)_d = \text{co} \left\{ \left(1, \frac{1}{2} \right), \left(-1, \frac{1}{2} \right) \right\}$$

and

$$\partial f(x_0) = \text{co} \{ (1, 0), (-1, 0) \},$$

so that

$$\nabla v_d(x_0) = \mu_d(x_0) [\partial_\varepsilon f(x_0)_d \overset{*}{\partial} f(x_0)],$$

as indicated by formula (2.5).

We now turn our attention to examples of assumptions securing that M_d or $\partial_\varepsilon f(\cdot)_d$ are single-valued so that results of Corollary 3.6 or Corollary 3.7 could apply.

As verified in [2, Thm. 3], a sufficient condition for M_d to be single-valued at x_0 is that $f_d : \lambda \mapsto f(x_0 + \lambda d)$ be strictly convex on \mathbb{R}_+^* . Consequently, $M_d(x) = \{\mu_d(x)\}$ for all $x \in \text{int}(\text{dom } f) \cap (x_0 + \mathbb{R}d)$ whenever f is strictly convex on $\text{int}(\text{dom } f) \cap (x_0 + \mathbb{R}d)$. We, therefore, have the following global statement.

PROPOSITION 3.8. *M_d is single-valued on $\text{int}(\text{dom } f)$ for all d whenever f is strictly convex on $\text{int}(\text{dom } f)$.*

Conditions ensuring that $\partial_\varepsilon f(\cdot)_d$ is single-valued at x_0 are, in a certain sense, dual to the above-described ones. In that respect, we begin by giving a further formulation of $\partial_\varepsilon f(x_0)_d$.

LEMMA 3.9. *Assume $\Lambda_d(x_0)$ is nonempty. Then*

$$(3.13) \quad \partial_\varepsilon f(x_0)_d = \{x^* \in \partial f(x_0 + \lambda d) \mid \langle x^*, d \rangle = f'_\varepsilon(x_0; d)\}$$

for all $\lambda \in \Lambda_d(x_0)$.

Proof. The x^* in $\partial_\varepsilon f(x_0)_d$ are those which satisfy

$$(3.14) \quad f^*(x^*) + f(x_0) - \langle x_0, x^* \rangle \leq \varepsilon$$

or, equivalently,

$$(3.15) \quad [f^*(x^*) + f(x_0 + \lambda d) - \langle x^*, x_0 + \lambda d \rangle] \lambda^{-1} \leq [f(x_0 + \lambda d) - f(x_0) + \varepsilon] \lambda^{-1} - \langle x^*, d \rangle$$

for all $\lambda > 0$ for which $f(x_0 + \lambda d)$ is finite.

If now $x^* \in \partial_\varepsilon f(x_0)_d$, we have that

$$\langle x^*, d \rangle = [f(x_0 + \lambda d) - f(x_0) + \varepsilon] \lambda^{-1}$$

for all $\lambda \in \Lambda_d(x_0)$. Hence, the left-hand side of (3.15) is null, which means that $x^* \in \partial f(x_0 + \lambda d)$.

Conversely, for all $\lambda \in \Lambda_d(x_0)$, we have that

$$f^*(x^*) + f(x_0) - \langle x_0, x^* \rangle = \varepsilon$$

whenever $x^* \in \partial f(x_0 + \lambda d)$ and $\langle x^*, d \rangle = f'_\varepsilon(x_0; d)$. Hence, the result is proved. \square

The above improves an earlier result by Lemaréchal and Nurminskii [9, Lemma 1] where the inclusion $\partial_\varepsilon f(x_0)_d \subset \partial f(x_0 + \lambda d)$ was stated for finite coercive functions. In the case where $\Lambda_d(x_0)$ is empty, we simply observe that

$$(3.16) \quad \partial_\varepsilon f(x_0)_d = \{x^* \in \partial_{\varepsilon(\lambda)} f(x_0 + \lambda d) | \langle x^*, d \rangle = f'_\varepsilon(x_0; d)\}$$

for all $\lambda > 0$ with $\varepsilon(\lambda)$ standing for $f(x_0 + \lambda d) - f(x_0) + \varepsilon - \lambda \langle x^*, d \rangle$. In that respect, we note that $\lim_{\lambda \rightarrow \infty} \varepsilon(\lambda)$ is $r_f^+(0)$, i.e., the right-hand derivative at 0 of the (convex) function r_f .

In consequence of Lemma 3.9, we have that v_d is strictly tangentially concave at all $x \in \text{int}(\text{dom } f) \cap (x_0 + \mathbb{R}d)$ whenever f is differentiable on $\text{dom } \partial f \cap (x_0 + \mathbb{R}d)$. Therefore, the counterpart situation to the one tackled in Proposition 3.8 concerns the so-called *essentially smooth* functions [14, p. 251]. f is called essentially smooth if it satisfies the following three conditions for $\Omega = \text{int}(\text{dom } f)$:

(a) Ω is nonempty;

(b) f is differentiable throughout Ω ;

(c) $\lim_{k \rightarrow \infty} \|\nabla f(x_k)\| = +\infty$ whenever $\{x_k\}$ is a sequence in Ω converging to a boundary point x_∞ of Ω .

Since an f which is finite and differentiable throughout \mathbb{R}^n is in particular essentially smooth, the next statement is just an extension of [7, Corollary 7.7].

PROPOSITION 3.10. *Assume f is essentially smooth. Then v_d is strictly tangentially concave on $\text{int}(\text{dom } f)$ for all d . Moreover,*

$$(3.17) \quad \partial v_d(x_0) = \left\{ \frac{\nabla f(x_0 + \lambda d) - \nabla f(x_0)}{\lambda} \mid \lambda \in \Lambda_d(x_0) \right\}$$

for all $x_0 \in \text{int}(\text{dom } f)$ and d for which $\partial v_d(x_0)$ is not reduced to $\{0\}$.

Proof. We hold apart the case where $\partial v_d(x_0) = \{0\}$, which is tantamount to the emptiness of $\Lambda_d(x_0)$ (Proposition 3.3). As for whether $\Lambda_d(x_0)$ is nonempty, Lemma 3.9 asserts that

$$x_0 + \lambda d \in \text{dom } \partial f$$

whenever $\lambda \in \Lambda_d(x_0)$. Now, $\text{dom } \partial f$ is $\text{int}(\text{dom } f)$ and $\partial f(x) = \{\nabla f(x)\}$ for all $x \in \text{int}(\text{dom } f)$ [14, Thm. 26.1]. Hence, the results are derived from Corollary 3.7. \square

When f combines the properties of being strictly convex on $\Omega = \text{int}(\text{dom } f)$ and essentially smooth, the pair (Ω, f) is called a *convex function of Legendre type* [14, p. 258]. For functions v_d associated with such an f , Propositions 3.8 and 3.10 yield a global differentiability property which generalizes and renders clearer an earlier result stated for everywhere finite functions [7, Corollary 7.8].

PROPOSITION 3.11. *Let (Ω, f) be of Legendre type. Then v_d is continuously differentiable on Ω for all d , and for all $x_0 \in \Omega$ and d for which $\nabla v_d(x_0)$ is nonnull, we have that*

$$(3.18) \quad \nabla v_d(x_0) = [\nabla f(x_0 + \lambda_d(x_0)d) - \nabla f(x_0)]\lambda_d(x_0)^{-1},$$

where $\lambda_d(x_0)$ stands for the unique element of $\Lambda_d(x_0)$.

This result is interesting by itself since it allows to calculate $\nabla v_d(x_0)$ without knowing the expression of v_d , provided one is able to determine $\lambda_d(x_0)$. The definition of $\lambda_d(x_0)$, translated in terms of necessary conditions for optimality [cf. relation (1.5)], yields that $\lambda_d(x_0)$ is the (unique) solution of the following nonlinear equation (in λ):

$$(3.19) \quad f(x_0 + \lambda d) - f(x_0) - \lambda \langle \nabla f(x_0 + \lambda d), d \rangle + \varepsilon = 0.$$

Solving this equation explicitly is not an easy matter in most examples. Nevertheless, since the equation in question is one-dimensional, an approximate solution can be obtained by some computational procedure. Also note that the knowledge of $\lambda_d(x_0)$ furnishes by (3.18) and (3.19) a sort of second order expansion for $f(x_0 + \lambda d)$ since

$$(3.20) \quad f(x_0 + \lambda d) = f(x_0) + \lambda \langle \nabla f(x_0), d \rangle + \lambda^2 f''_\varepsilon(x_0; d, d) - \varepsilon$$

for $\lambda = \lambda_d(x_0)$. This kind of development will be considered for an arbitrary f in the next section.

Since (Ω, f) is of Legendre type, so is (Ω^*, f^*) , where $\Omega^* = \text{int}(\text{dom } f^*)$ [14, Thm. 26.5]. So, given a direction d^* , the corresponding function w_{d^*} (associated with f^* and ε) is continuously differentiable on Ω^* and $\nabla w_{d^*}(x_0^*)$ is formulated as in (3.18) mutatis mutandis. It is known that Ω^* can be parameterized through the change of variables $x^* = \nabla f(x)$, $x \in \Omega$, and that

$$f^*(x^*) = \langle (\nabla f)^{-1}(x^*), x^* \rangle - f((\nabla f)^{-1}(x^*))$$

for all $x^* \in \Omega^*$. However, the relationship between $\lambda_d(x_0)$ and $\lambda_{d^*}(x_0^*)$ (where $x_0^* = \nabla f(x_0)$) as well as between $\nabla v_d(x_0)$ and $\nabla w_{d^*}(x_0^*)$ heavily depend on how easily the inverse mapping $(\nabla f)^{-1}$ can be made explicit, a drawback which is inherent to all situations dealing with the Legendre transform.

4. Behavior of $f''_\varepsilon(x_0; d, \delta)$ when ε converges to 0^+ . We have noted in the previous sections that there are two fundamental situations to pay regard to in consideration of the behavior of $f''_\varepsilon(x_0; d, \delta)$ when $\varepsilon \rightarrow 0^+$. The situation where $\Lambda_{d,\varepsilon}(x_0)$ is empty for all ε (situation (S_{3,b})) does not offer much interest since $f''_\varepsilon(x_0; d, \delta) = 0$ for all ε in such a case. The other situation, which will be of main concern in this section, deals with x_0 and d for which $\Lambda_{d,\varepsilon}(x_0)$ is nonempty for $\varepsilon < \bar{\varepsilon}$. In such a case, we have exhibited various formulations of $\partial v_{d,\varepsilon}(x_0)$ and $f''_\varepsilon(x_0; d, \delta)$ which clearly suggest that $f''_\varepsilon(x_0; d, \delta)$ is an approximation of $\langle \nabla^2 f(x_0) d, \delta \rangle$ when the latter makes sense. For example, when $v_{d,\varepsilon}$ is differentiable at x_0 , we observed that

$$\nabla v_{d,\varepsilon}(x_0) = [\partial_\varepsilon f(x_0) \circledast \partial f(x_0)] \lambda_\varepsilon^{-1} \subset [\partial f(x_0 + \lambda_\varepsilon d) \circledast \partial f(x_0)] \lambda_\varepsilon^{-1},$$

where λ_ε stands for the unique element of $\Lambda_{d,\varepsilon}(x_0)$. Of course, the regularity properties of $v_{d,\varepsilon}$, like the differentiability, are not preserved when ε moves to 0. It also clearly appears that the key point for what we are concerned with will be the study of the behavior of $\Lambda_{d,\varepsilon}(x_0)$ when $\varepsilon \rightarrow 0^+$. The first part of the present section is addressed to this question.

In the same way as $f'_\varepsilon(x_0; d)$ is an approximation of $f'(x_0; d)$, $f''_\varepsilon(x_0; d, \delta)$ plays the role of an approximation of the “second derivative of f at x_0 ”, the latter being defined in a suitable way. The second part of this section will show what to expect and what not to expect in that respect.

4.1. Due to the mere definition of $f'_\varepsilon(x_0; d)$, we have the following first order expansion:

$$(4.1) \quad f(x_0 + \lambda d) = f(x_0) + \lambda f'_\varepsilon(x_0; d) - \varepsilon$$

which is valid for all $\lambda \in \Lambda_{d,\varepsilon}(x_0)$. The function $\varepsilon \mapsto f'_\varepsilon(x_0; d)$ is known to be locally Lipschitz on \mathbb{R}_+^* and $f'_\varepsilon(x_0; d)$ decreases to $f'(x_0; d)$ as $\varepsilon \downarrow 0$ (see [6] and references therein). As noted in [7, § 1], the difference $f'_\varepsilon(x_0; d) - f'(x_0; d)$ may decrease slowly towards 0, as slowly as $\varepsilon^{1/2}$.

Let us consider x_0 and d such that $\Lambda_{d,\varepsilon}(x_0)$ is nonempty for a certain $\bar{\varepsilon}$; this simply means that we are not in the situation (S_{3,b}) such as described in § 2. Under this

assumption, we know that $\Lambda_{d,\varepsilon}(x_0)$ is nonempty (and bounded) for all $\varepsilon < \bar{\varepsilon}$. As is clear from the definitions, the properties of the multifunction $\varepsilon \Rightarrow \Lambda_{d,\varepsilon}(x_0)$ will be derived from those of $f'_\varepsilon(x_0; d)$. For that purpose, we set

$$(4.2) \quad \bar{\lambda}_\varepsilon = \max \{ \lambda \mid \lambda \in \Lambda_{d,\varepsilon}(x_0) \}, \quad \underline{\lambda}_\varepsilon = \min \{ \lambda \mid \lambda \in \Lambda_{d,\varepsilon}(x_0) \}$$

for all $\varepsilon \in]0, \bar{\varepsilon}[$. At first glance, one might expect that $\bar{\lambda}_\varepsilon$ (or $\underline{\lambda}_\varepsilon$) converges to 0 when ε goes to 0; this is not generally true, and we shall exhibit a necessary and sufficient condition for that to hold. The next propositions summarize the main results on the behavior of $\Lambda_{d,\varepsilon}(x_0)$ as a multifunction of ε . In particular, it will be shown that even if $\underline{\lambda}_\varepsilon$ converges to 0 it converges slower than ε .

PROPOSITION 4.1. *Let $\underline{\lambda}_\varepsilon$ and $\bar{\lambda}_\varepsilon$ be such as defined in (4.2) for all $\varepsilon < \bar{\varepsilon}$. Then*

$$(4.3) \quad (a) \quad f'_\varepsilon(x_0; d) - f'(x_0; d) \geq \frac{\varepsilon}{\underline{\lambda}_\varepsilon}$$

so that $\varepsilon/\underline{\lambda}_\varepsilon$ converges to 0 when ε converges to 0;

$$(4.4) \quad (b) \quad \bar{\lambda}_{\varepsilon'} \leq \underline{\lambda}_\varepsilon \text{ for all } \varepsilon' < \varepsilon.$$

Proof. According to the definition of $f'_{\varepsilon'}(x_0; d)$, we have that

$$\begin{aligned} [f(x_0 + \bar{\lambda}_{\varepsilon'} d) - f(x_0) + \varepsilon'] / \bar{\lambda}_{\varepsilon'} &\leq [f(x_0 + \underline{\lambda}_\varepsilon d) - f(x_0) + \varepsilon'] / \underline{\lambda}_\varepsilon \\ &\leq [f(x_0 + \underline{\lambda}_\varepsilon d) - f(x_0) + \varepsilon] / \underline{\lambda}_\varepsilon + (\varepsilon' - \varepsilon) / \underline{\lambda}_\varepsilon. \end{aligned}$$

Hence,

$$f'_\varepsilon(x_0; d) \leq f'_{\varepsilon'}(x_0; d) + (\varepsilon' - \varepsilon) / \underline{\lambda}_\varepsilon.$$

Since we may obtain a similar inequality with ε and ε' switched,

$$f'_{\varepsilon'}(x_0; d) \leq f'_\varepsilon(x_0; d) + (\varepsilon - \varepsilon') / \bar{\lambda}_{\varepsilon'},$$

we finally have

$$(4.5) \quad (\varepsilon - \varepsilon') / \underline{\lambda}_\varepsilon \leq f'_\varepsilon(x_0; d) - f'_{\varepsilon'}(x_0; d) \leq (\varepsilon - \varepsilon') / \bar{\lambda}_{\varepsilon'}.$$

By letting ε' converge to 0 in the first inequality of (4.5), we get the inequality (4.3). As for (4.4) it readily comes from the inequalities of (4.5). \square

PROPOSITION 4.2. *Let $\{\varepsilon_n\}$ be a sequence converging to $\varepsilon \in]0, \bar{\varepsilon}[$. Then one of the following holds:*

- (a) $\lim_{n \rightarrow \infty} \underline{\lambda}_{\varepsilon_n} = \lim_{n \rightarrow \infty} \bar{\lambda}_{\varepsilon_n} = \underline{\lambda}_\varepsilon$;
- (b) $\lim_{n \rightarrow \infty} \underline{\lambda}_{\varepsilon_n} = \lim_{n \rightarrow \infty} \bar{\lambda}_{\varepsilon_n} = \bar{\lambda}_\varepsilon$;
- (c) $[\liminf_{n \rightarrow \infty} \underline{\lambda}_{\varepsilon_n}, \limsup_{n \rightarrow \infty} \bar{\lambda}_{\varepsilon_n}] = \Lambda_{d,\varepsilon}(x_0)$.

Proof. Let $\{\lambda_{\varepsilon_n}\}$ satisfy $\lambda_{\varepsilon_n} \in \Lambda_{d,\varepsilon_n}(x_0)$ for all n . We first have to prove that $\{\lambda_{\varepsilon_n}\}$ is bounded. According to the first order expansion (4.1), we have that

$$(4.6) \quad [f(x_0 + \lambda_{\varepsilon_n} d) - f(x_0)] \lambda_{\varepsilon_n}^{-1} = f'_{\varepsilon_n}(x_0; d) - \varepsilon_n \lambda_{\varepsilon_n}^{-1}$$

for all n . If $\limsup_{n \rightarrow \infty} \lambda_{\varepsilon_n} = +\infty$, passing to the limit in the above equality would imply that $f_\infty(d) = f'_\varepsilon(x_0; d)$, which is inconsistent with the hypothesis that $\Lambda_{d,\varepsilon}(x_0)$ is nonempty and bounded (cf. § 2). Consider therefore a subsequence of $\{\lambda_{\varepsilon_n}\}$ converging to λ_∞ . Again passing to the limit in (4.6) clearly shows that $\lambda_\infty \in \Lambda_{d,\varepsilon}(x_0)$. This result combined with the relation (4.4) yields the desired results. \square

We have observed that $\varepsilon/\underline{\lambda}_\varepsilon$ converges to 0 with ε . Without further information on f around x_0 , this result cannot be sharpened, as the next example shows.

Example 4.1. Let f be defined on \mathbb{R} by $f(x) = |x|^\alpha$ with $\alpha \in]1, 2[$. For $x_0 = 0$ and $d = 1$, we have that $\Lambda_{d,\varepsilon}(x_0) = \{(\varepsilon/(\alpha - 1))^{1/\alpha}\}$. Here

$$\lim_{\varepsilon \rightarrow 0} \frac{\varepsilon}{(\lambda_\varepsilon)^\beta} = +\infty,$$

whenever $\beta > \alpha$.

We now turn our attention to the behavior of $\Lambda_{d,\varepsilon}(x_0)$ when ε converges to 0. The next example makes it clear that $\lambda_\varepsilon \in \Lambda_{d,\varepsilon}(x_0)$ does not necessarily converge to 0 even in fairly simple examples.

Example 4.2. Let $f: \mathbb{R} \rightarrow \mathbb{R}$ be defined as

$$f(x) = \max(-x, x - 2).$$

We set $x_0 = 0$, $d = 1$ and $\varepsilon < \bar{\varepsilon} = 2$. In such a case, we clearly have that $\Lambda_{d,\varepsilon}(x_0) = \{1\}$ for all $\varepsilon < 2$. We see in this example that f is affine on $[0, 1]$; that is to say, the difference quotient $[f(x_0 + \lambda d) - f(x_0)]\lambda^{-1}$ is constant ($= f'(x_0; d)$) on $]0, 1]$. This is the kind of situation we would like to avoid in order to secure the convergence of $\bar{\lambda}_\varepsilon$ toward 0.

To take into account the situations where $f_d: \lambda \mapsto f(x_0 + \lambda d)$ might be affine in a neighborhood of 0^+ , we pose the following definition: $a_d^*(x_0)$ is the supremum of all $a \in \mathbb{R}_+$ for which

$$f(x_0 + \lambda d) = f(x_0) + \lambda f'(x_0; d)$$

for all $\lambda \in [0, a]$.

$a_d^*(x_0) = +\infty$ corresponds to the situation where f_d is affine on \mathbb{R}_+ , i.e., the so-called situation (S_{3,b}) in § 2. The case where $a_d^*(x_0) = 0$ is of particular importance. Since the difference quotient $[f(x_0 + \lambda d) - f(x_0)]\lambda^{-1}$ is a nondecreasing function of $\lambda > 0$, the requirement that $a_d^*(x_0) = 0$ is actually equivalent to the following assumption:

(H) *there exists $a > 0$ such that $[f(x_0 + \lambda d) - f(x_0)]\lambda^{-1} > f'(x_0; d)$ for all $\lambda \in]0, a]$.*

We are now ready for stating the main convergence results on $\Lambda_{d,\varepsilon}(x_0)$ when ε goes to 0.

THEOREM 4.3. *Assume that $a_d^*(x_0) < +\infty$ at x_0 for the direction d . Then $\Lambda_{d,\varepsilon}(x_0)$ is nonempty for ε in a neighborhood of 0^+ and*

$$(4.7) \quad \lim_{\varepsilon \rightarrow 0^+} \Lambda_{d,\varepsilon}(x_0) = \{a_d^*(x_0)\}.$$

By (4.7) stated in terms of limits of intervals, we mean that

$$\lim_{\varepsilon \rightarrow 0^+} \lambda_\varepsilon = \lim_{\varepsilon \rightarrow 0^+} \bar{\lambda}_\varepsilon = a_d^*(x_0).$$

It is worth particularizing the statement for the case where $a_d^*(x_0) = 0$.

COROLLARY 4.4. *The following are equivalent:*

- (a) *f satisfies (H) at x_0 for the d direction;*
- (b) *$\Lambda_{d,\varepsilon}(x_0)$ is nonempty for ε in a neighborhood of 0^+ and $\lim_{\varepsilon \rightarrow 0^+} \Lambda_{d,\varepsilon}(x_0) = \{0\}$.*

Proof. Let $a_d^*(x_0) < +\infty$ at x_0 for the direction d . From the definition of $a_d^*(x_0)$ itself, we infer that we are not in the situation (S_{3,b}) so that $\Lambda_{d,\varepsilon}(x_0)$ is nonempty for ε in a certain neighborhood of 0^+ , say $]0, \bar{\varepsilon}]$. Let $\{\varepsilon_n\}$ be a sequence converging to 0^+ and let $\{\lambda_{\varepsilon_n}\}$ satisfy $\lambda_{\varepsilon_n} \in \Lambda_{d,\varepsilon_n}(x_0)$ for all n . As indicated in (4.1), we have that

$$(4.8) \quad f(x_0 + \lambda_{\varepsilon_n} d) = f(x_0) + \lambda_{\varepsilon_n} f'_{\varepsilon_n}(x_0; d) - \varepsilon_n$$

for all n . The sequence $\{\lambda_{\varepsilon_n}\}$ is bounded, as it can readily be seen from the inequality

(4.4). Let λ_∞ be the limit of a subsequence of $\{\lambda_{\varepsilon_n}\}$; by passing to the limit in (4.4), we have

$$f(x_0 + \lambda_\infty d) = f(x_0) + \lambda_\infty f'(x_0; d),$$

so that $\lambda_\infty \leq a_d^*(x_0)$. We pursue our argument for $a_d(x_0) > 0$.

The convex function $r_f: \mu \mapsto \mu[f(x_0 + d/\mu) - f(x_0) + \varepsilon]$ is such that

$$r_f(\mu) = f'(x_0; d) + \varepsilon\mu$$

for $\mu \geq 1/a_d^*(x_0)$. Consequently, $M_{d,\varepsilon}(x_0) \subset]0, 1/a_d^*(x_0)]$ for all ε in $]0, \bar{\varepsilon}[$ and, thus,

$$\lambda_\infty \geq a_d^*(x_0).$$

Hence the result (4.7) is proved. \square

Remark. The result in Theorem 4.3 emphasizes that *all* the λ_ε satisfying $\lambda_\varepsilon \in \Lambda_{d,\varepsilon}(x_0)$ do converge towards a *common* limit as $\varepsilon \rightarrow 0^+$. The fact that this limit might be nonnull is somewhat baffling, and we feel it is a drawback of the present approach. As for simple examples where this case typically occurs we have polyhedral functions. For such a function, we clearly have that $a_d^*(x_0) > 0$ for all d and all $x_0 \in \text{int}(\text{dom } f)$.

4.2. When $\Lambda_{d,\varepsilon}(x_0)$ turns out to be nonempty and bounded, $f''_\varepsilon(x_0; d, \delta) = v'_{d,\varepsilon}(x_0; \delta)$ is expressed as

$$(4.9) \quad f''_\varepsilon(x_0; d, \delta) = \min_{\lambda \in \Lambda_{d,\varepsilon}(x_0)} \{[\psi_{\partial_\varepsilon f(x_0)d}^*(\delta) - f'(x_0; \delta)]/\lambda\}.$$

According as $\psi_{\partial_\varepsilon f(x_0)d}^*(\delta) - f'(x_0; \delta)$ is positive or negative, the minimum over $\Lambda_{d,\varepsilon}(x_0)$ is achieved for $\bar{\lambda}_\varepsilon$ or $\underline{\lambda}_\varepsilon$ in the above formula. For example, we have for $\delta = d$:

$$(4.10) \quad f''_\varepsilon(x_0; d, d) = [f'_\varepsilon(x_0; d) - f'(x_0; d)]/\bar{\lambda}_\varepsilon.$$

This expression of $f''_\varepsilon(x_0; d, d)$, combined with the first order expansion (4.1), yields the following second order expansion:

$$(4.11) \quad f(x_0 + \bar{\lambda}_\varepsilon d) = f(x_0) + \bar{\lambda}_\varepsilon f'(x_0; d) + \bar{\lambda}_\varepsilon^2 f''_\varepsilon(x_0; d, d) - \varepsilon.$$

Here again we emphasize the crucial role played by $\Lambda_{d,\varepsilon}(x_0)$. As indicated in § 3.2 for a particular case, $\Lambda_{d,\varepsilon}(x_0)$, which is known to be the solution set of the equation

$$0 \in \lambda \langle \partial f(x_0 + \lambda d), d \rangle - f(x_0 + \lambda d) + f(x_0) - \varepsilon,$$

is the cornerstone for calculating *both* $f'_\varepsilon(x_0; d)$ and $f''_\varepsilon(x_0; d, d)$.

For the purposes of studying the behavior of $f''_\varepsilon(x_0; d, \delta)$ when ε goes to 0, we shall distinguish three cases according as $a_d^*(x_0)$ is $+\infty$, a positive number or null.

First let $a_d^*(x_0) = +\infty$. In such a case, we know that $f''_\varepsilon(x_0; d, \delta)$ is null for all δ and all ε . Consequently, it is natural to define $f'''(x_0, d, \delta)$ by posing

$$(4.12) \quad f'''(x_0; d, \delta) = 0$$

for all δ .

The case where $0 < a_d^*(x_0) < +\infty$ is somewhat similar. According to (4.9),

$$f''_\varepsilon(x_0; d, \delta) = [\psi_{\partial_\varepsilon f(x_0)d}^*(\delta) - f'(x_0; \delta)]/\lambda_\varepsilon,$$

where λ_ε stands for $\bar{\lambda}_\varepsilon$ or $\underline{\lambda}_\varepsilon$. Anyway λ_ε converges to $a_d^*(x_0)$ (Theorem 4.3). Moreover, we clearly have that

$$\limsup_{\varepsilon \rightarrow 0^+} \psi_{\partial_\varepsilon f(x_0)d}^*(\delta) \leq \psi_{\partial f(x_0)d}^*(\delta)$$

and

$$\liminf_{\varepsilon \rightarrow 0^+} \psi_{\partial_\varepsilon f(x_0)_d}^*(\delta) \geq -\psi_{\partial f(x_0)_d}^*(-\delta)$$

for all δ . Thus, $\lim_{\varepsilon \rightarrow 0^+} f''_\varepsilon(x_0; d, \delta)$ exists and is negative whenever the breadth of $\partial f(x_0)_d$ in the δ direction is null. In particular, this limit is null for $\delta = d$.

So, we retain that

$$(4.13) \quad f''(x_0; d, d) = \lim_{\varepsilon \rightarrow 0^+} f''_\varepsilon(x_0; d, d)$$

is null for the d directions for which $0 < a_d^*(x_0) \leq +\infty$. Getting $f''(x_0; d, d) = 0$ for these directions d is quite natural since

$$f(x_0 + \lambda d) = f(x_0) + \lambda f'(x_0; d)$$

for all λ in a neighborhood of 0^+ .

We now consider the case where $a_d^*(x_0) = 0$. Here $f''_\varepsilon(x_0; d, \delta)$ is not necessarily bounded when $\varepsilon \rightarrow 0^+$, even for $\delta = d$. From the inequalities (4.3) and (4.10), we infer that

$$(4.14) \quad f''_\varepsilon(x_0; d, d) \geq \varepsilon / (\bar{\lambda}_\varepsilon)^2,$$

$$(4.15) \quad \lim_{\varepsilon \rightarrow 0^+} \bar{\lambda}_\varepsilon f''_\varepsilon(x_0; d, d) = 0.$$

So, $\lim_{\varepsilon \rightarrow 0^+} f''_\varepsilon(x_0; d, d) = +\infty$ whenever $(\bar{\lambda}_\varepsilon)^2$ converges to 0 faster than ε ; see Example 4.1 for an illustration of such a behavior. However, we shall prove that $f''_\varepsilon(x_0; d, \delta)$ has a limit when $\varepsilon \rightarrow 0^+$ if f is “twice differentiable at x_0 ”. For that purpose, let us recall some results on the twice differentiability of convex functions.

The subdifferential $\partial f(x)$ of f at x is nonempty for all $x \in \text{int}(\text{dom } f)$ and reduces to $\{\nabla f(x_0)\}$ whenever f is differentiable at x_0 . Defining a second derivative at x_0 in the usual sense of the differential calculus would require that f be differentiable in a neighborhood of x_0 . But simple examples show that f can be nondifferentiable on a dense subset of $\text{int}(\text{dom } f)$ so that the definition of a second derivative has to be made suitable for such a situation. Following Rockafellar [15, p. 887] or Mignot [10, § 1.2], we say that ∂f is differentiable at x_0 if $\partial f(x_0) = \{\nabla f(x_0)\}$ and there is a linear transformation denoted by $A^2 f(x_0)$ such that

$$(4.16) \quad \|\partial f(x) - \nabla f(x_0) - A^2 f(x_0) \cdot (x - x_0)\| = o(\|x - x_0\|),$$

or in other words,

$$\forall \varepsilon > 0, \quad \exists \delta > 0, \quad \forall x \text{ with } \|x - x_0\| \leq \delta, \quad \forall x^* \in \partial f(x), \\ \|x^* - \nabla f(x_0) - A^2 f(x_0) \cdot (x - x_0)\| \leq \varepsilon \|x - x_0\|.$$

This appears as a strong requirement on the behavior of ∂f around x_0 . Nevertheless, we have the following result:

THEOREM 4.3. *∂f is differentiable almost everywhere on $\text{int}(\text{dom } f)$.*

The above is actually a corollary to Mignot’s differentiability theorem on maximal monotone multifunctions [10, Thm. 1.3]. Rephrased in terms of a quadratic approximation of f around a point, Theorem 4.3 implies that at almost all points x_0 of $\text{int}(\text{dom } f)$ there exists a linear transformation $A^2 f(x_0)$ such that

$$(4.17) \quad f(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{1}{2} \langle A^2 f(x_0) \cdot (x - x_0), x - x_0 \rangle + o(\|x - x_0\|^2),$$

a result which dates back to Alexandroff [1].

We denote by E_2 the set of $x_0 \in \text{int}(\text{dom } f)$ where ∂f is differentiable. If we denote by $\hat{\nabla}f$ any mapping satisfying

$$\hat{\nabla}f(x) \in \partial f(x)$$

for all $x \in \text{int}(\text{dom } f)$, the development (4.16) shows that $\hat{\nabla}f$ is differentiable on E_2 and that its derivative is A^2f . Further differentiability properties are also satisfied at the privileged points of E_2 . For example, the *second Schwarz derivative* in the h direction

$$\lim_{\lambda \rightarrow 0} \frac{f(x_0 + \lambda h) - 2f(x_0) + f(x_0 - \lambda h)}{2\lambda^2}$$

as well as the *second right de la Vallée-Poussin derivative* in the h direction

$$\lim_{\lambda \rightarrow 0^+} \frac{1}{\lambda} \left[\frac{f(x_0 + \lambda h) - f(x_0)}{\lambda} - \langle \nabla f(x_0), h \rangle \right]$$

exist at $x_0 \in E_2$ and equal $\frac{1}{2} \langle A^2 f(x_0) h, h \rangle$.

THEOREM 4.4. *Let $x_0 \in E_2$ and let $a_d^*(x_0) = 0$. Then*

$$(4.18) \quad \lim_{\varepsilon \rightarrow 0^+} f''_\varepsilon(x_0; d, \delta) = \langle A^2 f(x_0) d, \delta \rangle$$

for all δ .

Proof. Let $x_\varepsilon^* \in \partial_\varepsilon f(x_0)_d$ satisfy

$$\langle x_\varepsilon^*, \delta \rangle = \psi_{\partial_\varepsilon f(x_0)_d}^*(\delta).$$

$f''_\varepsilon(x_0; d, \delta)$ can be written as

$$[\langle x_\varepsilon^*, \delta \rangle - \langle \nabla f(x_0), \delta \rangle] / \lambda_\varepsilon,$$

with λ_ε converging to 0. Now according to Lemma 3.9, x_ε^* lies in $\partial f(x_0 + \lambda_\varepsilon d)$. Hence, the convergence result (4.18) derives from the definition (4.16) of $A^2 f(x_0)$. \square

COROLLARY 4.5. *Let $x_0 \in E_2$ and let $a_d^*(x_0) = 0$. Then*

$$(4.18)^0 \quad \lim_{\varepsilon \rightarrow 0^+} [\limsup_{x \rightarrow x_0} f''_\varepsilon(x; d, \delta)] = \langle A^2 f(x_0) d, \delta \rangle$$

for all δ and

$$(4.19) \quad \lim_{\varepsilon \rightarrow 0^+} \partial v_{d,\varepsilon}(x_0) = A^2 f(x_0) d.$$

In (4.19), we mean that $\lim_{\varepsilon \rightarrow 0^+} v_\varepsilon^* = A^2 f(x_0) d$ for any v_ε^* in $\partial v_{d,\varepsilon}(x_0)$, $0 < \varepsilon < \bar{\varepsilon}$.

Proof. The outer estimates of $\partial v_{d,\varepsilon}(x_0)$, such as those given in § 3, yield that

$$\limsup_{x \rightarrow x_0} f''_\varepsilon(x; d, \delta) = v_{d,\varepsilon}^0(x_0; \delta) \leq [f'(x_0 + \tilde{\lambda}_\varepsilon d; \delta) - \langle \nabla f(x_0), \delta \rangle] / \tilde{\lambda}_\varepsilon$$

for a certain $\tilde{\lambda}_\varepsilon \in \Lambda_{d,\varepsilon}(x_0)$ and

$$\partial v_{d,\varepsilon}(x_0) \subset \bigcup_{\lambda \in \Lambda_{d,\varepsilon}(x_0)} \left\{ \frac{\partial f(x_0 + \lambda d) - \nabla f(x_0)}{\lambda} \right\}.$$

Thus, (4.18)⁰ and (4.19) are proved by arguing as in the theorem above. \square

Remark. If $x_0 \in E_2$ and $a_d^*(x_0) > 0$, one easily checks that

$$\langle A^2 f(x_0) d, d \rangle = 0$$

and

$$\lim_{\varepsilon \rightarrow 0^+} f''_{\varepsilon}(x_0; d, d) = \lim_{\varepsilon \rightarrow 0^+} [\limsup_{x \rightarrow x_0} f''_{\varepsilon}(x; d, d)] = 0.$$

So the relation

$$\lim_{\varepsilon \rightarrow 0^+} f''_{\varepsilon}(x_0; d, d) = \lim_{\varepsilon \rightarrow 0^+} [\limsup_{x \rightarrow x_0} f''_{\varepsilon}(x; d, d)] = \langle A^2 f(x_0) d, d \rangle$$

holds true for all directions d whenever $x_0 \in E_2$.

In view of that and of the second-order expansion (4.11), we can say that

$$(4.20) \quad \lim_{\varepsilon \rightarrow 0^+} \frac{\varepsilon}{(\bar{\lambda}_{\varepsilon})^2} = \frac{1}{2} \langle A^2 f(x_0) d, d \rangle$$

when $x_0 \in E_2$.

So far, the existence of the limit of $f''_{\varepsilon}(x_0; d, d)$ when $\varepsilon \rightarrow 0^+$ is secured when either $a_d^*(x_0) > 0$ or $x_0 \in E_2$. What can be expected when $a_d^*(x_0) = 0$ and $x_0 \notin E_2$?

As stated in the question, we only consider the “diagonal case”, i.e., $f''_{\varepsilon}(x_0; d, \delta)$ for $\delta = d$. We previously have set an example where $\lim_{\varepsilon \rightarrow 0^+} f''_{\varepsilon}(x_0; d, d)$ was $+\infty$ (Example 4.1). Moreover, it is not hard to find an example for which $\{f''_{\varepsilon}(x_0; d, d) | 0 < \varepsilon < \bar{\varepsilon}\}$ is bounded but has no limit when $\varepsilon \rightarrow 0^+$. According to what has been proved earlier,

$$f''_{\varepsilon}(x_0; d, d) = [f'_{\varepsilon}(x_0; d) - f'(x_0 d)] / \bar{\lambda}_{\varepsilon}$$

is the quotient of two decreasing functions of ε and

$$(4.21) \quad f''_{\varepsilon}(x_0; d, d) \leq [f'(x_0 + \bar{\lambda}_{\varepsilon} d; d) - f'(x_0; d)] / \bar{\lambda}_{\varepsilon}.$$

In default of having necessarily a limit, we focus our attention on conditions ensuring that $f''_{\varepsilon}(x_0; d, d)$ remains bounded when $\varepsilon \rightarrow 0^+$. Only the behavior of $f_d : \lambda \mapsto f(x_0 + \lambda d)$ on \mathbb{R}_+ is relevant for the study of the behavior of $f''_{\varepsilon}(x_0; d, d)$. We shall say that the directional derivative of f is *point-Lipschitz at x_0 in the d direction* if

$$(L_1) \quad \limsup_{\lambda \rightarrow 0^+} [f'(x_0 + \lambda d; d) - f'(x_0; d)] / \lambda < +\infty.$$

In view of (4.21), this condition is sufficient for securing

$$\limsup_{\varepsilon \rightarrow 0^+} f''_{\varepsilon}(x_0; d, d) < +\infty.$$

It also can be rephrased in terms of the behavior of f_d itself by bounding a sort of (upper) second right de la Vallée-Poussin derivative in the d direction.

PROPOSITION 4.6. *The directional derivative of f is point-Lipschitz at x_0 in the d direction if and only if*

$$(L_2) \quad \limsup_{\lambda \rightarrow 0^+} \frac{1}{\lambda} \left[\frac{f(x_0 + \lambda d) - f(x_0)}{\lambda} - f'(x_0; d) \right] < +\infty.$$

Proof. Assume that $x \mapsto f'(x; d)$ is point-Lipschitz at x_0 in the d direction. It is a consequence of the mean-value theorem that there exists $\sigma \in]0, \lambda[$ such that

$$[f(x_0 + \lambda d) - f(x_0)] / \lambda \leq f'(x_0 + \sigma d; d).$$

Since $f'(x_0 + \sigma d; d) - f'(x_0; d)$ is nonnegative, the above inequality yields

$$\frac{1}{\lambda} \left[\frac{f(x_0 + \lambda d) - f(x_0)}{\lambda} - f'(x_0; d) \right] \leq \frac{1}{\sigma} [f'(x_0 + \sigma d; d) - f'(x_0; d)],$$

whence (L_2) is secured.

For proving the converse implication, we draw inspiration from the proof of [15, Prop. 7]. Assuming that (L_2) holds, there exist $k > 0$ and $\lambda_0 > 0$ such that

$$(4.22) \quad f(x_0 + \lambda d) \leq f(x_0) + \lambda f'(x_0; d) + k\lambda^2$$

for all $\lambda \in [0, \lambda_0]$. Let

$$\theta(\lambda) = \begin{cases} k\lambda^2 & \text{if } \lambda \in [0, \lambda_0], \\ +\infty & \text{if not.} \end{cases}$$

Then (4.22) can be expressed as

$$(4.23) \quad f(x_0 + \lambda d) \leq f(x_0) + \lambda f'(x_0; d) + \theta(\lambda)$$

for all λ . Let now $\sigma^* \in \partial f(x_0 + \sigma d)$ satisfy

$$\langle \sigma^*, d \rangle = f'(x_0 + \sigma d; d).$$

We have

$$\begin{aligned} f(x_0 + \lambda d) &\geq f(x_0 + \sigma d) + (\lambda - \sigma) \langle \sigma^*, d \rangle \\ &\geq f(x_0) + \sigma f'(x_0; d) + (\lambda - \sigma) f'(x_0 + \sigma d; d) \end{aligned}$$

for all λ . Combined with (4.23) this yields

$$\lambda [f'(x_0 + \sigma d; d) - f'(x_0; d)] - \theta(\lambda) \leq \sigma [f'(x_0 + \sigma d; d) - f'(x_0; d)]$$

for all λ . Consequently,

$$(4.24) \quad \theta^* \{f'(x_0 + \sigma d; d) - f'(x_0; d)\} \leq \sigma [f'(x_0 + \sigma d; d) - f'(x_0; d)].$$

But $\theta^*(\lambda^*) = (\lambda^*)^2/4k$ for $\lambda^* \in [0, 2\lambda_0 k]$ and $f'(x_0 + \sigma d; d) - f'(x_0; d)$ converges to 0^+ when $\sigma \rightarrow 0^+$. Thus, it comes from (4.24) that

$$f'(x_0 + \sigma d; d) - f'(x_0; d) \leq (4k)\sigma$$

for σ in a certain interval $[0, \sigma_0]$. Hence, (L_1) is secured. \square

To require a condition like (L_1) is actually a mild assumption on the behavior of f_d on \mathbb{R}_+ . We note that (L_1) is satisfied at all $x_0 \in \text{int}(\text{dom } f)$ and for all d when f is a $C^{1,1}$ function, i.e., a function differentiable on $\text{int}(\text{dom } f)$ and whose derivative is locally Lipschitz on $\text{int}(\text{dom } f)$. Also observe that (L_1) is stable under usual operations on convex functions; in particular, (L_1) is secured at x_0 for $f = \max_{i=1, \dots, m} f_i$ whenever it is for the f_i for which $f_i(x_0) = f(x_0)$. So, for a large class of convex functions, it is secured that

$$f''(x_0; d, d) = \limsup_{\varepsilon \rightarrow 0^+} f''_\varepsilon(x_0; d, d) < +\infty$$

at all $x_0 \in \text{int}(\text{dom } f)$ and for all d .

Final remark. Although $A^2 f$ is defined almost everywhere in $\text{int dom } f$, difficulties arise when using it to recover the directional derivatives $f'(x; d)$. Indeed, it is easy to construct a strictly convex differentiable function for which $A^2 f(x)$ is null whenever it exists.

Nevertheless, since $v_{d,\varepsilon}: x \mapsto f'_\varepsilon(x; d)$ is locally Lipschitz and admits directional derivatives at all points, it can be expressed as the integral of these directional derivatives [4, p. 53], namely,

$$f'_\varepsilon(x_0 + \delta; d) = f'_\varepsilon(x_0; d) + \int_0^1 f''_\varepsilon(x_0 + t\delta; d, \delta) dt.$$

This expansion, which is valid for all $\varepsilon > 0$, is not preserved when $\varepsilon \rightarrow 0^+$, even if $\{x \in [x_0, x_0 + \delta] | x \notin E_2\}$ is of one-dimensional null-measure.

5. Conclusion. All the properties of f''_ε we have displayed in the present study emphasize that it indeed plays the role of an approximate second-order directional derivative of f . The charm of f''_ε lies in the fact that it is defined at all x_0 of $\text{int}(\text{dom } f)$ and therefore could serve as a substitute for the second-order derivative in designing second-order minimization methods. However, the definition and properties of such procedures would require us to know the behavior of $f''_\varepsilon(x_0; d, d)$ as a function of d . This question has not been tackled in the present paper and deserves to be done. As it was conspicuous in the study of the behavior of f''_ε when $\varepsilon \mapsto 0^+$, the role of the sets $\Lambda_{d,\varepsilon}(x_0)$ is of utmost importance. Evidently, the properties of $\Lambda_{d,\varepsilon}(x_0)$ as a multifunction of d would give the clues for all questions on the variations of $f''_\varepsilon(x_0; d, d)$ as a function of d . In that respect, we mention the following open problem: is the function $d \mapsto f''_\varepsilon(x_0; d, d)$ convex? The one-dimensional case as well as calculations for particular f adduced plausible evidence to support the conjecture that the answer is positive. Defining an approximate second-order directional derivative for broad classes of nonconvex functions is of main concern in the current research in nonsmooth optimization. In view of what has been carried out in the present work, there is a class of functions for which the desired objective could be achieved, namely the lower- C^2 functions such as defined by Rockafellar [17]. According to Rockafellar, f is lower- C^2 if the following holds for each x_0 : there is for some open neighborhood $V(x_0)$ of x_0 a representation

$$f(x) = \max_{y \in Y} \varphi(x, y) \quad \text{for all } x \in V(x_0),$$

where Y is a compact set and $\varphi: V(x_0) \times Y \rightarrow \mathbb{R}$ is a function which has first and second partial derivatives with respect to x and which along with all these derivatives is continuous jointly in $(x, y) \in V(x_0) \times Y$. Lower- C^2 functions include convex functions and share many properties of them, like the existence almost everywhere of a second-order Taylor expansion [17, Corollary 2]. The close correspondence of lower- C^2 functions with convex ones is also demonstrated by the following result [17, Thm. 6]: around every x_0 a locally Lipschitz lower- C^2 f has a representation

$$f = g - h,$$

with g convex, h quadratic convex. Now since

$$h''_\varepsilon(x_0; d, \delta) = \langle \text{Ad}, \delta \rangle$$

for all ε , with A the second derivative of h , it is natural to pose

$$f''_\varepsilon(x_0; d, \delta) = g''_\varepsilon(x_0; d, \delta) - \langle \text{Ad}, \delta \rangle$$

as a possible definition of an approximate second-order directional derivative for f at x_0 . However, the consistency of this definition as well as its usefulness remain to be proved and require further study.

REFERENCES

- [1] A. D. ALEXANDROFF, *The existence almost everywhere of the second differential of a convex function and some associated properties of convex surfaces*, Uchenye Zapiski Leningr. Gos. Univ. Ser. Mat., 37, No. 6 (1939), pp 3–35. (In Russian.)
- [2] A. AUSLENDER, *Differential properties of the support function of the ϵ -subdifferential of a convex function*, Math. Programming, to appear.
- [3] F. H. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.
- [4] T. M. FLETT, *Differential Analysis*, Cambridge University Press, London, 1980.
- [5] E. G. GOL'SHTEIN, *Theory of Convex Programming*, Transl. Math. Monographs, 36, American Mathematical Society, Providence, R.L., 1972, Chap. 7.
- [6] J.-B. HIRIART-URRUTY, *Lipschitz r -continuity of the approximate subdifferential of a convex function*, Math. Scand., 47 (1980), pp. 123–134.
- [7] ———, *ϵ -subdifferential calculus*, Proc. of the Colloquium Convex Analysis and Optimization, Imperial College, London (28–29 February, 1980), to appear.
- [8] W. W. HOGAN, *Directional derivatives for extremal-value functions with applications to the completely convex case*, Open. Res., 21 (1973), pp. 188–209.
- [9] C. LEMARÉCHAL AND E. NURMINSKII, *Sur la différentiabilité de la fonction d'appui du sous-différentiel approché*, Note aux C. R. Acad. Sci. Paris, 290, Sér. A (1980), pp. 855–858.
- [10] F. MIGNOT, *Contrôle dans les inéquations variationnelles elliptiques*, J. Functional Anal., 22 (1976), pp. 130–185.
- [11] E. A. NURMINSKII, *On ϵ -differential mappings and their applications in nondifferentiable optimization*, Working paper 78–58, I.I.A.S.A., December 1978.
- [12] L. S. PONTRYAGIN, *Linear differential games 2*, Soviet Math. Doklady, 8 (1967), pp. 910–912.
- [13] B. N. PSHENICHNYI, *Leçons sur les jeux différentiels*, Contrôle Optimal et Jeux Différentiels, cahier No. 4, I.R.I.A., 1971.
- [14] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [15] ———, *Monotone operators and the proximal point algorithm*, this Journal, 14 (1976), pp. 877–898.
- [16] ———, *La théorie des sous-gradients et ses applications: fonctions convexes et non convexes*, Presses de l'Université de Montréal, Montreal, 1979.
- [17] ———, *Favorable classes of Lipschitz continuous functions in subgradient optimization*, Working paper, I.I.A.S.A., November, 1980.
- [18] J. STOER AND C. WITZGALL, *Convexity and Optimization in Finite Dimensions*, Springer-Verlag, New York, 1970.

AN OPENNESS CONDITION FOR THE CONTROLLABILITY OF NONLINEAR SYSTEMS*

J. P. GAUTHIER† AND G. BORNARD‡

Abstract. In this paper we study the stability of the controllability property of a system under a general class of state space and system deformations. The main result is a sufficient condition of openness. A corollary shows how this result can be compared to that given by Sussmann [J. Differential Equations, 20 (1976), pp. 292–315]. Some illustrative examples are given.

1. Introduction. Let M be a C^{r+1} connected manifold of dimension m and Γ a system of C^r vector fields on M ($1 \leq r \leq \infty$). This paper deals with the openness of the controllability property in a sense which will be made more precise later.

In the case of a finite system Γ , Sussmann [8] established openness results with respect to a deformation of the system.

The principal result which will be shown here deals with a nontrivial deformation of both the system and the state space.

Let us consider the following situation, which will be met throughout the paper: Γ is not necessarily finite; the orbits of the symmetric system $\{\Gamma, -\Gamma\}$ are submanifolds of M of dimension n , constant over some open saturated neighborhood \tilde{M} of an orbit F in M , $0 < n < m$. It is well known (from Stefan [7]) that this partition defines a C^r foliation \mathcal{F} of \tilde{M} of codimension $p = m - n$ whose leaves are the orbits of Γ . Here the saturation of a set $\subset \tilde{M}$ is the set $\pi^{-1} \circ \pi(\cdot)$, where π is the canonical map from \tilde{M} to the space of the leaves. A set S is said to be saturated if it is equal to its saturation.

The main result to be established here is then:

THEOREM 1 (main result). *Consider \tilde{M} , Γ , \mathcal{F} as just described, and let F be a leaf having the following properties:*

- i. F is compact.
- ii. *Either*
 - ii-a *the holonomy group $H(F)$ of F is finite*
 - or ii-b *there exists a saturated neighborhood of F which is the union of compact leaves, each of them having a fundamental system of neighborhoods which are saturated for \mathcal{F} ,*
 - or ii-c *there exists a saturated open neighborhood N of F which is the union of compact leaves, such that the quotient space N/\mathcal{F} is Hausdorff with the ordinary quotient topology.*
- iii. F is controllable.

Then there exists a saturated neighborhood of F which is the union of controllable leaves. (Here by a controllable leaf F , we mean a leaf F such that the system Γ_F induced by Γ on F is controllable on F .)

Notation is specified in §2, some lemmas are established in §3 and the proof of Theorem 1 is exhibited in §4. Corollaries are given in §5, and in §6 a few examples illustrate some typical situations which can be encountered. In the Appendix, as suggested by the referee, we include a brief, intuitive discussion of the notion of holonomy.

* Received by the editors July 20, 1981, and in revised form December 1, 1981.

† Chargé de Recherche au CNRS; Laboratoire d'Automatique de Grenoble, B.P. 46-38402 St. Martin D'Hères, France.

‡ Attaché de Recherche au CNRS; Ecole Polytechnique d'Alger, 10, Avenue Pasteur, El Harrach, Algeria.

2. Notations and definitions. Let \tilde{M} , Γ , \mathcal{F} be defined as in the introduction. Consider a proper leaf F . (A leaf is proper if its topology as a leaf is the same as its topology as a subspace of \tilde{M} .) Let (V, π, F) , $\pi: V \rightarrow F$, be a locally trivial fibration of a tubular neighborhood V of F in M (see, for instance, Gobillon [3, chap. II]).

One can find, for each point $x \in F$, an (arbitrarily small) open connected neighborhood U having the following properties:

a) $\pi \times f$ is a C^{r+1} diffeomorphism of U onto $Y \times Z$, where Y and Z are open connected subsets in F and $\pi^{-1}(x)$ respectively, and $x = Y \cap Z$.

b) f is a \mathcal{F} -compatible mapping (i.e., $f^{-1}(z)$ is a U -slice of \mathcal{F} for every $z \in Z$) such that the restriction of f to Z is the identity mapping.

c) $\bar{U} \subset U'$, U' being some neighborhood of x meeting the conditions a, b.

Such a neighborhood U will from now on be referred as a regular distinguished neighborhood (R.D.N.). (For the properties of foliations, see, for instance, Epstein [2], Haefliger [4], Lawson [5].)

Consider $\xi = \{X^1, \dots, X^l\}$ a finite set of elements of Γ , $t = (t_1, \dots, t_l) \in R^l$ and $\psi_\xi: \tilde{M} \times R^l \rightarrow M$, the application defined by

$$\psi_\xi(x, t) = X_{t_l}^l \circ \dots \circ X_{t_1}^1(x),$$

where X_t^i denotes the one-parameter subgroup generated by the vector field X^i . ψ_ξ is of rank m (dimension of \tilde{M}) with respect to x .

In the proofs the concept of normal accessibility (Sussmann [8]) will be used extensively. We shall say here that x_1 is normally accessible from x_0 in a leaf F , ($x_0, x_1 \in F$), if there exists a triple (l, ξ, t_0) , $t_0 \in (R_*^+)^l$, such that $\psi_\xi(x_0, t_0) = x_1$ and ψ_ξ is of rank n (dimension of F) with respect to t at (x_0, t_0) .

The normal auto-accessibility of x_0 is the special use of the preceding concept when $x_1 = x_0$.

3. Lemmas. In this section \tilde{M} , Γ , \mathcal{F} are considered as defined in § 1. We do not assume they satisfy the conditions of Theorem 1.

LEMMA 1 (local). Consider F_{x_0} the leaf containing $x_0 \in \tilde{M}$. Assume that F_{x_0} is proper and x_0 is normally auto-accessible in F_{x_0} with the triple (l, ξ_0, t_0) . Then there exists a R.D.N. $U_0 = Y_0 \times Z_0$ of x_0 , a neighborhood $T_0 \subset (R_*^+)^l$ of t_0 and a mapping $\sigma: U_0 \rightarrow T_0$ such that denoting $(y', z') = \psi_{\xi_0}(y, z, \sigma(y, z))$ one has $y' = y$ for every $(y, z) \in Y_0 \times Z_0$.

Moreover, the point $x' = (y', z')$ is normally accessible from $x = (y, z)$ in F_x with the triple $(l, \xi_0, \sigma(x))$.

Proof. Consider a R.D.N. U'_0 of x_0 , a R.D.N. $U''_0 = Y''_0 \times Z''_0 \subset U'_0$ and a neighborhood $T''_0 \subset (R_*^+)^l$ such that $\psi_{\xi_0}(U''_0, T''_0) \subset U'_0$.

Due to the invariance of \mathcal{F} under Γ , ψ_{ξ_0} takes the following form:

$$\psi_{\xi_0}(y, z, t) = (\psi'_{\xi_0}(y, z, t), \psi''_{\xi_0}(z))$$

for every $(y, z) \in Y''_0 \times Z''_0$, $t \in T''_0$, and ψ'_{ξ_0} is a mapping of $U''_0 \times T''_0$ into F_{x_0} of rank n (dimension of F_{x_0}) with respect to t at (x_0, t_0) and such that $\psi'_{\xi_0}(y_0, z_0, t_0) = y_0 = \pi(x_0)$ (normal auto-accessibility).

The result of Lemma 1 is then obtained by applying the implicit function theorem (see Auslander and Mackenzie [1]) to the system $\psi'_{\xi_0}(x, t) = \pi(x)$, choosing U_0 sufficiently small so that the rank of $\psi_{\xi_0}(x, t)$ with respect to t remains n at $(x, \sigma(x))$ for every $x \in U_0$.

LEMMA 2 (openness of the auto-accessibility). Consider F_{x_0} the leaf containing $x_0 \in \tilde{M}$ and assume that F_{x_0} is proper and has a finite holonomy group. Then if x_0 is normally auto-accessible in F_{x_0} , there exists a R.D.N. $U_0 = Y_0 \times Z_0$ of x_0 such that each $x \in U_0$ is normally auto-accessible in F_x .

Proof. Apply Lemma 1. There exists $U'_0, U''_0 = Y''_0 \times Z''_0, T''_0 \subset (R^+_\ast)^l, \sigma: U''_0 \rightarrow T''_0$ with the following properties:

$$\begin{aligned}\psi_{\xi_0}(U''_0, T''_0) &\subset U'_0, \\ \psi_{\xi_0}(y, z, \sigma(y, z)) &= (y, \psi''_{\xi_0}(z)) \text{ for every } (y, z) \in Y''_0 \times Z''_0, \\ \psi_{\xi_0}(y, z, \sigma(y, z)) &\text{ is normally accessible in } F_{(y,z)} \text{ from } (y, z) \\ &\text{ for every } (y, z) \in Y''_0 \times Z''_0,\end{aligned}$$

and we shall denote $\omega(x) = \psi_{\xi_0}(x, \sigma(x))$ for $x \in U''_0$. Clearly, the germ of ψ''_{ξ_0} at z_0 is the element of the holonomy group of F_{x_0} corresponding to the homotopy class of the loop defined in F_{x_0} by the triple $(l, \xi_0, t_0 = \sigma(x_0))$.

Let k be the order of the holonomy group $H(F_{x_0})$. There exists $U_0 = Y_0 \times Z_0 \subset U''_0$ such that $(\psi''_{\xi_0})^i(Z_0) \subset Z''_0$ for $i = 1, \dots, k-1$. Then $(\psi''_{\xi_0})^{k'}$ is the identity mapping of Z_0 onto itself for some divisor k' of k , and consequently, $(\omega)^{k'}$ is the identity mapping of U_0 onto itself.

Consider, for some $x \in U_0$, the sequence of $x^i: x^0 = x, x^{i+1} = \omega(x^i), i = 0, \dots, k'-1$. By construction, one has $x^{k'} = x^0, x^i \in U''_0$ for $i = 0, \dots, k'$. As a consequence each x^{i+1} is normally accessible from x^i , in F_x , and $x^{k'}$ is then normally accessible from x^0 (transitivity of the normal accessibility, Sussmann [8, Property a, p. 296]).

Then for every $x \in U_0, x$ is normally auto-accessible.

4. Proof of Theorem 1. The fact that conditions ii-a, ii-b and ii-c are equivalent is a direct consequence of Epstein [2, Theorems 4.1, 4.2, 4.3]. These conditions can then be used equally in the proof.

The leaf F is controllable (condition iii). Then each point $x \in F$ is normally auto-accessible in F (Sussmann [8, Thm. 4.3]). Apply Lemma 2 to the leaf F . Since it is compact (condition i), there exists a finite set of R.D.N. $U_i = Y_i \times Z_i, i = 0, \dots, r$, such that, if $W = \bigcup_{i=0, \dots, r} U_i, F \subset W$ and every $x \in W$ is normally auto-accessible.

W is a neighborhood of F . By condition ii-b, there exists a saturated neighborhood $V \subset W$ of F .

By construction every $x \in V$ is normally auto-accessible and every leaf meeting V is included in V . Then using again Sussmann [8, Thm. 4.3], every leaf $F' \subset V$, having all its points normally auto-accessible in F' , is controllable.

Remark. In the preceding proofs, the conditions ii-a, b, c were not used completely. The only properties which were needed are:

ii-d the orbits of the holonomy group of F are finite,

ii-e there exists a fundamental system of saturated neighborhood of F ,

and one could think that the conditions ii-a, b, c are too restrictive. In fact, from Epstein [2, Thm. 7.3], the condition ii-d implies ii-a, which in turn is equivalent to ii-b and ii-c. This last equivalence requires the compactness of the leaf, which is implied by ii-e. However, the compactness condition could perhaps be removed by replacing i and ii by ii-a and some regularity conditions on Γ at ∞ .

5. Corollaries. The first corollary gives two cases, more restrictive, implying that the holonomy group of F is finite.

COROLLARY 1. *In Theorem 1, the conditions ii-a, b, c are implied by:*

ii' $H_1(F, \mathbb{R}) = 0, H_1(F, \mathbb{R})$ being the first homology group of F with real coefficient; or by (more restrictive):

ii'' $\pi_1(F)$, the first homotopy group of F is finite.

Proof. These conditions obviously imply condition ii-a. This is sufficient, but one can also remark that from the stability theorems of Thurston [9] and Reeb [6] they also imply directly ii-b.

The second corollary deals with the particular case where the state-space is fixed and only the system is disturbed. It is to be compared with the Sussmann openness theorems (here the manifold is compact and the system is not necessarily finite).

COROLLARY 2. *Let N be a C^{r+1} compact manifold of dimension n , A an open subset of R^p and $\Lambda(\alpha)$ a system of C^r vector fields on N , depending C^r -differentiably on a parameter $\alpha \in A$. If $\Lambda(\alpha_0)$, $\alpha_0 \in A$, is controllable on N , then there exists a neighborhood $U \subset A$ of α_0 such that $\Lambda(\alpha)$ is controllable for every $\alpha \in U$.*

Proof. Consider the C^{r+1} product manifold $M = N \times A$, of dimension $m = n + p$ and the system Γ of C^r complete vector fields on M defined by

$$\Gamma(x, \alpha) = (\Lambda(\alpha), 0) \in T_{N_x} \times T_{A_\alpha}$$

for every $x \in N$, $\alpha \in A$.

Then apply Theorem 1 to M , Γ : the leaves are of the form (M, α) , the quotient topology is obviously Hausdorff and the leaf (M, α_0) is controllable.

Remark. As pointed out by a referee, a direct proof of Corollary 2 can be derived from the results of Sussmann [8]. Moreover, it is interesting to remark that in the case of a compact state-space if Γ (infinite) is controllable, then it exists as a finite collection of vector fields $\Sigma \subset \Gamma$ such that Σ is controllable.

6. Illustrative examples.

Example 1. $M = R^2 - \{0\}$, $\Gamma = \{X\}$, where X is defined by

$$\begin{aligned} \dot{x}_1 &= -x_2 + (1 - \rho)x_1, \\ \dot{x}_2 &= x_1 + (1 - \rho)x_2, \end{aligned} \quad \text{with } \rho = \sqrt{x_1^2 + x_2^2}.$$

The orbits of Γ behave as shown in Fig. 1.

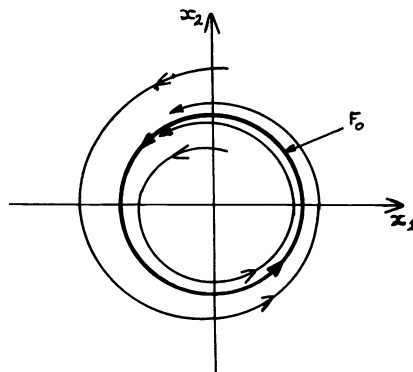


FIG. 1

The leaves are proper, the leaf F_0 is compact, controllable. But (\tilde{M}, Γ) does not satisfy the condition ii, and the leaves other than F_0 are not controllable.

Example 2. M is the open Möbius band of dimension 2, and $\Gamma = \{X\}$, where the leaves, which are the integral curves of X , are as shown in Fig. 2 (two leaves are represented).

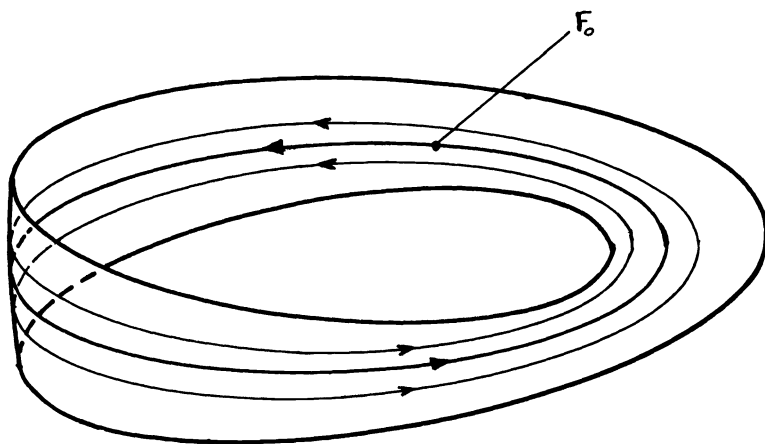


FIG. 2

F_0 is a particular leaf whose homotopy class (one revolution) is different from all the others (two revolutions). The conditions i, ii, iii are satisfied for F_0 , and all the leaves are controllable.

We are here in the case where ψ''_{ξ_0} is not reduced to the identity mapping: ψ''_{ξ_0} is the symmetric application of the transversal fiber into itself.

Remark. For codimension 1 the identity mapping and the symmetric mapping are the only two finite situations which can be met.

Example 3.

$$\tilde{M} = M = \mathbb{R}^3 - \{x \mid x_1 = x_2 = 0\}, \quad \Gamma = \{X_1, X_2, X_3\},$$

with

$$X_1: \begin{cases} \dot{x}_1 = x_1, \\ \dot{x}_2 = x_2, \\ \dot{x}_3 = 0, \end{cases} \quad X_2: \begin{cases} \dot{x}_1 = -x_2, \\ \dot{x}_2 = x_1, \\ \dot{x}_3 = 0, \end{cases} \quad X_3: \begin{cases} \dot{x}_1 = -\phi(x)x_1, \\ \dot{x}_2 = -\phi(x)x_2, \\ \dot{x}_3 = 0, \end{cases}$$

$$\phi(x) = \theta((x_1^2 + x_2^2)x_3^2),$$

where θ is a C^r function such that $\theta(t) > 0$ for $-1 < t < 1$ and $\theta(t) = 0$ elsewhere.

Γ defines a foliation of “horizontal” planes (Fig. 3). The leaf $F_0 = \{x \mid x_3 = 0\}$ is obviously controllable, its holonomy group is finite and the quotient topology is Hausdorff. But the leaves are not compact, and it can be seen that any leaf other than F_0 is not controllable. Let A be the domain where $X_3 \neq 0$, and let B be its complement in M . Clearly, it is not possible with Γ to go back from B to A , and F_0 is the only leaf which does not cross B . The problem arises from the fact that the leaves being not compact, one cannot find a fundamental system of saturated neighborhoods, and moreover, X_3 behaves “badly” at ∞ near F_0 .

7. Appendix: an intuitive discussion of the notion of holonomy. Let us call \mathcal{G}_p^{r+1} the group of germs at $x = 0$ of C^{r+1} local diffeomorphisms f from \mathbb{R}^p to \mathbb{R}^p , preserving the point $x = 0$ ($f(0) = 0$).

The holonomy mapping of a leaf F of \mathcal{F} is a representation ϕ of $\pi_1(F)$ (the first homotopy group of F) in the group \mathcal{G}_p^{r+1} . ϕ will be precisely defined later. The

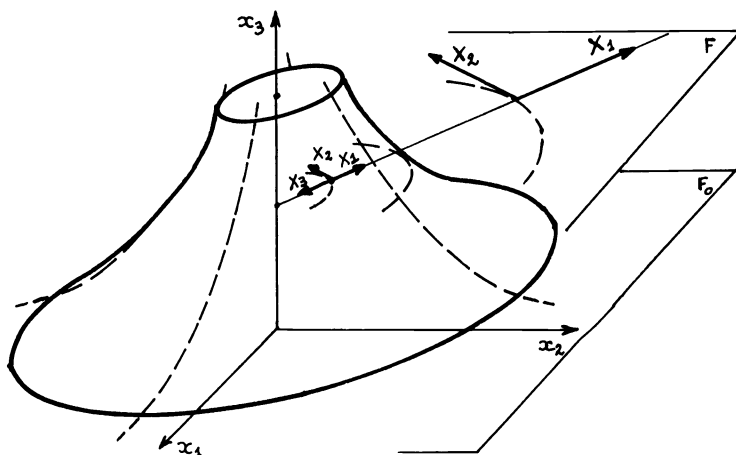


FIG. 3

holonomy group of F is the image H of $\pi_1(F)$ through the holonomy mapping ϕ ;

$$H = \phi(\pi_1(F)).$$

We now construct the holonomy mapping ϕ :

Let $z_0 \in F$. Let U , π be defined as in § 2. Let W be a neighborhood of z_0 in $\pi^{-1}(z_0)$, and ψ be a C^{r+1} diffeomorphism from W to \mathbb{R}^p such that $\psi(z_0) = 0$.

Let $\{\gamma(t) | 0 \leq t \leq 1, \gamma(0) = z_0, \gamma(1) = z_0\}$ be a loop in F with initial point z_0 and endpoint z_0 . For any $(\omega, t) \in W \times [0, 1]$, let F_ω be the leaf through ω and consider the set $H_t(\omega) = \pi^{-1}(\gamma(t)) \cap F_\omega$. This set is discrete, and one can choose $\tilde{\gamma}_\omega(t)$ (in a unique way) in $H_t(\omega)$ such that $\{\tilde{\gamma}_\omega(t) | 0 \leq t \leq 1\}$ is a continuous path in F_ω with $\tilde{\gamma}_\omega(0) = \omega$.

Denote θ for the mapping $z \rightarrow \tilde{\gamma}_z(1)$ from W to $\pi^{-1}(z_0)$. Clearly $\theta(z_0) = z_0$. Consider $\tilde{\theta} = \psi \circ \theta \circ \psi^{-1}$, from \mathbb{R}^p to \mathbb{R}^p . One has $\tilde{\theta}(0) = 0$. The germ of $\tilde{\theta}$ at $x = 0$ is the image of the loop $\{\gamma(t) | 0 \leq t \leq 1\}$ through the holonomy mapping ϕ :

$$\phi(\{\gamma(t) | 0 \leq t \leq 1\}) = \text{germ at } x = 0 \text{ of } \tilde{\theta}.$$

It can be shown that, if $\gamma(t)$ and $\gamma'(t)$ are two loops through z_0 in the same homotopy class, $\phi(\gamma(t)) = \phi(\gamma'(t))$, and that ϕ is a homomorphism.

Acknowledgments. The authors thank Prof. I. Kupka and Prof. G. Sallet for their contribution to this work. Their suggestions resulted in a better formalism and in more elegant proofs than the original ones.

The authors also thank an anonymous referee for his/her very interesting remarks and suggestions.

REFERENCES

- [1] L. AUSLANDER AND R. E. MACKENZIE, *Introduction to Differentiable Manifolds*, Dover, New York, 1977.
- [2] D. B. A. EPSTEIN, *Foliations with all leaves compact*, Annales Inst. Fourier, 26 (1976), pp. 265–282.
- [3] C. GODBILLON, *Géométrie différentielle et mécanique analytique*, Hermann, Paris, 1969.
- [4] A. HAEFLIGER, *Variétés feuilletées*, Anali Scuola Normale Superiore Pisa, 16 (1962), pp. 367–397.
- [5] H. B. LAWSON, *Foliations*, Bull. Amer. Math. Soc., 80 (1974), pp. 369–418.
- [6] G. REEB, *Sur certaines propriétés topologiques des variétés feuilletées*, Actualités Scient. et Industrielles, 1183, 1952.

- [7] P. STEFAN, *Accessible sets, orbits, and foliations with singularities*, Proc. London Math. Soc. 29 (1974), pp. 699–713.
- [8] H. J. SUSSMANN, *Some properties of vector fields systems that are not altered by small perturbations*, J. Differential Equations, 20 (1976), pp. 292–315.
- [9] W. P. THURSTON, *A generalization of the Reeb stability theorem*, Topology, 13 (1974), pp. 347–352.

AN APPROXIMATION THEORY FOR NONLINEAR PARTIAL DIFFERENTIAL EQUATIONS WITH APPLICATIONS TO IDENTIFICATION AND CONTROL*

H. T. BANKS[†] AND K. KUNISCH[‡]

Abstract. Approximation results from linear semigroup theory are used to develop a general framework for convergence of approximation schemes in parameter estimation and optimal control problems for nonlinear partial differential equations. These ideas are used to establish theoretical convergence results for parameter identification using modal (eigenfunction) approximation techniques. Results from numerical investigations of these schemes for both hyperbolic and parabolic systems are given.

1. Introduction. When modeling real-world phenomena one often encounters a situation where a priori knowledge leads one to conjecture a certain type of model equation containing parameters which are unknown. In this paper we are primarily concerned with techniques for recovery of these unknown quantities from given data. In §§ 2 and 3 we present a quite general framework for approximation schemes for abstract nonlinear Cauchy problems. These approximation results are subsequently applied to modal techniques for identification and control problems in §§ 4 and 5, respectively. A summary of some of our numerical experience with parameter estimation problems using these techniques is given in § 6. The examples here were chosen so as to illustrate the feasibility and effectiveness of the method and to investigate possible inherent difficulties. We are quite confident that the ideas outlined here will be applicable in a variety of research areas where mathematical models for the phenomena under study are used. In a forthcoming monograph we shall discuss in more detail identification problems that arise in several areas of applications [35] including seismology [3], [10], [18], reservoir engineering [11], [17], [38], glaciology [16], physics [37], biology [4], [5], [29], [34] and large space structures technology. While our treatment here is restricted to constant unknown parameters, the theoretical ideas extend in large part to problems with unknown function parameters. Indeed, we are currently applying some of our techniques to specific problems from the areas mentioned above; in some cases these efforts involve identification of functions.

In this paper the general approximation results are used to carefully discuss modal approximation schemes for certain classes of partial differential equations (see §§ 4 and 5). Such schemes for specific identification and control problems are, of course, not new. Many discussions in the literature, however, are in the context of very specific examples and frequently no convergence proofs or evidence of numerical studies are supplied. Modal approximations have many advantages, including: they are readily discussed and understood in terms of classical spectral results; they are familiar to

* Received by the editors May 1, 1981, and in revised form January 11, 1982. This work was supported in part by the Air Force Office of Scientific Research under contract AFOSR 76-3092D, in part by the National Science Foundation under grant NSF-MCS 7905774-02, and in part by the U.S. Army Research Office under contract ARO-DAAG 29-79-C-0161.

[†] Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. Part of this research was carried out while this author was a visitor at the Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, VA, which is operated under NASA contracts NAS1-15810 and NAS1-16394.

[‡] Institut für Mathematik, Technische Universität Graz, Kopernikusgasse 24, A-8010 Graz, Austria. Part of this research was carried out while this author was a visitor at the Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University. This author also acknowledges support from the Steierm. Wissenschafts- und Forschungsfonds and the Fulbright Commission.

and readily implemented by practicing engineers, and they give rise to a simple algebraic structure for the approximating ordinary differential equations. However, modal approximations do have some shortcomings, of which we mention several. First, in many practical problems it is very difficult to calculate the true natural modes. Secondly, for certain parabolic partial differential equations modal approximations by their very nature lead to stiff systems of approximating ordinary differential equations. Finally, one can encounter lack of “numerical identifiability” (i.e., the identification problems for the approximating ordinary differential equations yield parameter estimates that converge to different values for different sets of initial estimates) regardless of the well-posedness of the parameter estimation problem for the original partial differential equation model. With respect to the first difficulty pointed out here, we refer to Example 4.4, below, where we explain a “modal” approximation scheme for an identification problem which does not employ the natural modes of the system. For one solution of the latter problems, our experience indicates that for certain classes of parabolic problems spline-based approximation schemes can be more efficient. Details on this aspect of spline methods, along with a number of other features of these techniques, will be given in a separate manuscript currently in preparation.

The parameter identification and estimation problem has received a great amount of attention in the engineering literature and we refer to [1], [23], [31], [32] for review articles. In the future monograph alluded to above, we shall survey the research efforts from the engineering as well as from the mathematical literature. Much of the mathematical literature is concerned with the problem of identifiability, which, loosely speaking, is defined as the problem of injectivity of the map from the set of parameters to the set of outputs. Although this is a very important theoretical and practical question, it will not be a part of the discussion of the present paper.

We point out one important technical aspect that will become clearer in Examples 4.1 and 4.4 below. In general, the eigenfunctions of the model equation will depend on the parameters that are to be identified. For modal approximation schemes this is an extremely undesirable feature from the point of view of implementation, since in practical examples the representation of the operators in the approximating equations will involve a matrix of inner products of the eigenfunctions. It is, of course, desirable to have this matrix independent of the unknown parameters to avoid excessive numerical integrations when performing iterative searches on these parameters.

Our focus in this paper is on the development of semidiscrete approximation schemes for parameter identification and control problems which result in approximating problems governed by ordinary differential equations. Of course, full discretization methods (discretization in time as well as spatial coordinates, resulting in problems governed by difference equations) are of great importance and our investigations of a related theoretical framework, as well as detailed schemes for such an approach, will be reported elsewhere.

In summary, the emphasis of our presentation is twofold. First, we present a general theoretical framework, with unknown parameter-dependent spaces, which can be used to treat many types of problems (including estimation of function space parameters) and approximation schemes (see the remarks in § 7 below). As a concrete example of the use of this framework, we give a detailed treatment of “modal” approximation schemes, thereby putting on a sound theoretical foundation a class of methods that have been used in an ad hoc way by scientists and engineers for some time.

The notation used throughout the paper is quite standard. We employ the usual notation H^l for Sobolev spaces with “functions” and “derivatives” in L_2 , and $|\cdot|$ to

denote norms of elements, as well as those of operators. Only in cases where confusion may arise will we use subscripts to distinguish norms in various spaces.

2. The abstract identification problem and its approximation. We consider the abstract semilinear Cauchy problem

$$(2.1) \quad \begin{aligned} \dot{u}(t) &= A(q)u(t) + F(q, t, u(t)), & t > 0, \\ u(0) &= u_0(q), \end{aligned}$$

where for each $q \in Q \subset R^k$, $A(q)$ is the infinitesimal generator of a linear C_0 -semigroup $\{T(t; q)\}_{t \geq 0}$ on a real Hilbert space $X(q)$ with inner product $\langle \cdot, \cdot \rangle_q$ and norm $|\cdot|_q$ (denoted sometimes below by X , $\langle \cdot, \cdot \rangle$ and $|\cdot|$, respectively, when no loss of clarity results from suppression of the q). We shall, throughout our discussions, employ the concept of *mild solutions*, so that $t \rightarrow u(t; q)$ is called a *solution* of (2.1) if it satisfies

$$(2.2) \quad u(t; q) = T(t; q)u_0(q) + \int_0^t T(t-s; q)F(q, s, u(s; q)) ds.$$

We note that for solutions u we have $t \rightarrow u(t; q)$ continuous. The conditions that we impose on F below will guarantee existence and uniqueness of mild solutions u of (2.1) on any given finite interval $[0, T]$. We shall in certain specific instances below, be required to discuss briefly the relationship between mild and strong (in a classical almost-everywhere sense) solutions of (2.1), but for more general results we refer the reader to [28].

Throughout our presentation we shall assume that $X(q)$ is a function space of R^n -valued “functions” (possibly one of the usual Lebesgue spaces of equivalence classes of functions) defined on the fixed interval $[0, 1]$; consequently, we shall also use the notation $u(t, x; q)$ or $u(t, \cdot; q)$ when discussing solutions of (2.1).

While we shall also discuss control-theoretic applications, much of our attention will be directed towards the problem of identifying the parameter q in (2.1) from observations of the system. Specifically, we assume that (2.1) models some physical, biological, economic, etc., system for which output measurements \hat{y} are available. These measurements may be available in the form of continuous data $\hat{y}(t)$, $0 \leq t \leq T$, or discrete data $\hat{y}(t_i)$, $0 \leq t_1 < \dots < t_r \leq T$. We then seek to find a “best” value for q in Q by minimizing an appropriately defined fit-to-data criterion. To be specific in our formulation here, we shall assume discrete time observations with values $\hat{y}(t_i)$ in an observation space \mathcal{Y} . All of the results of this paper are easily extended to the case of identification problems where one has continuous time data, but we shall not pursue such problems here. Assuming, then, that a criterion function $J: Q \times C(0, T; X(q)) \times \prod_{i=1}^r \mathcal{Y} \rightarrow R^1$ is defined, we formally state the identification problem:

(ID) Given observations $\hat{y} = \{\hat{y}(t_i)\}_{i=1}^r$, minimize $J(q, u(\cdot; q), \hat{y})$ over $q \in Q$ subject to $u(\cdot; q)$ satisfying (2.2).

Several traditional choices of fit-to-data criteria are included in our formulation; namely, we may consider either integral or pointwise (in a spatial sense) evaluation least-squares sums in the above formulation. In the case of integral evaluation we are given measurements $\hat{y}(t_i) \in L_2(0, 1; R^\nu)$ where $\nu \leq n$ and an output map $Y(t, x, q): R^n \rightarrow R^\nu$ on the “state” $u(t, x; q)$. The observation space is given by $\mathcal{Y} = L_2(0, 1; R^\nu)$ and the criterion is defined by

$$(2.3) \quad J(q, u(\cdot; q), \hat{y}) = \sum_{i=1}^r \int_0^1 |\hat{y}(t_i, x) - Y(t_i, x, q)u(t_i, x; q)|^2 dx.$$

We assume that Y is continuous in q and sufficiently regular in x so that $x \rightarrow Y(t_i, x, q)u(t_i, x; q)$ is in $L_2(0, 1; R^\nu)$. For the choice of pointwise or spatially discrete measurements, we assume that we have observations $\hat{y}(t_i) \in \mathcal{Y} \equiv \prod_{j=1}^l R^\nu$, corresponding to measurements of the output at points $\{x_j\}_{j=1}^l$ in $[0, 1]$ at time t_i . These observations represent measurements for $C(t_i, q)\xi(t_i, q)$ where $\xi(t_i, q) = \text{col}(u(t_i, x_1; q), \dots, u(t_i, x_l; q))$ and $C(t_i, q)$ is an $(\nu l) \times (nl)$ -matrix depending continuously on q for each fixed t_i . The associated fit-to-data criterion is then defined by

$$(2.4) \quad J(q, u(\cdot; q), \hat{y}) = \sum_{i=1}^r |\hat{y}(t_i) - C(t_i, q)\xi(t_i, q)|^2.$$

The output maps Y and C introduced in (2.3) and (2.4) are necessitated by the fact that often in practice one can observe only some components (say ν) of the n -dimensional vectors $u(t, x; q)$, and that these observations may depend on the time at which they are made. We further note that the point evaluations at x_j used in defining $\xi(t_i, q)$ above may be meaningless without additional restrictions on the state space, the initial data and/or the right side of the equation in (2.1). A more detailed discussion of the problems arising from use of criteria such as (2.4) when dealing with mild solutions will be given in the context of Example 4.3 below.

We turn next to formulating a sequence (ID^N) of approximating problems on Hilbert spaces $X^N(q)$ for our original identification problem (ID). These problems involve "states" governed by ordinary differential equations and are (in the specific instances we shall propose) tractable using standard numerical procedures. We state first a series of hypotheses and definitions that will be needed at various points in the sequel.

- (H1) For each $N = 1, 2, \dots$, $X^N(q)$ is a closed linear subspace of $X(q)$, endowed with the $X(q)$ topology.
- (H2) The spaces $X(q)$, $q \in Q \subset R^k$, are set-theoretically equal and uniformly topologically isomorphic so that there exists a constant $\mathcal{K} \geq 1$ such that $|v|_{\tilde{q}} \leq \mathcal{K}|v|_q$ for all $v \in X = X(q)$ and $q, \tilde{q} \in Q$.
- (H3) For each $q \in Q$, $A(q)$ generates a linear C_0 -semigroup $T(t; q)$ on $X(q)$.
- (H4) The set Q is a compact subset of R^k .
- (H5) (i) For each $q \in Q$, let $P^N(q): X(q) \rightarrow X^N(q)$ denote the canonical orthogonal projections along $X^N(q)^\perp$ and let $A^N(q): X(q) \rightarrow X^N(q)$ be defined by $A^N(q) = P^N(q)A(q)P^N(q)$. For each N , let $A^N(q)$ generate a linear C_0 -semigroup on $X(q)$ denoted by $T^N(t; q)$.
 (ii) For each N , there exist constants $\hat{\mathcal{M}} = \hat{\mathcal{M}}(N)$ and $\hat{\omega} = \hat{\omega}(N)$, independent of q , such that $|T^N(t; q)| \leq \hat{\mathcal{M}}e^{\hat{\omega}t}$.
- (H6) (i) For each continuous function $u: [0, T] \rightarrow X = X(q)$ (see (H2)), the map $t \rightarrow F(q, t, u(t))$ is measurable.
 (ii) For each constant $M > 0$, there exists a function $k_1 = k_1(M)$ in $L_2(0, T)$ such that for any $q, \tilde{q} \in Q$ we have

$$|F(q, t, u_1) - F(q, t, u_2)|_{\tilde{q}} \leq k_1(t)|u_1 - u_2|_{\tilde{q}}$$

for all $u_1, u_2 \in X$ with $|u_i|_{\tilde{q}} \leq M$.

- (iii) There exists a function k_2 in $L_2(0, T)$ such that

$$|F(q, t, v)|_{\tilde{q}} \leq k_2(t)\{|v|_{\tilde{q}} + 1\}$$

for all $v \in X, q, \tilde{q} \in Q$.

- (iv) For each $(t, v) \in [0, T] \times X$, the map $q \rightarrow F(q, t, v)$ is continuous. (Again, under (H2) recall that all the sets $X = X(q)$ are the same.)

- (H7) The projections $P^N(q): X(q) \rightarrow X^N(q)$ are such that for any sequence $\{q^N\}$ in Q satisfying $q^N \rightarrow \bar{q} \in Q$, one has $|P^N(q^N)z - z|_{q^N} \rightarrow 0$ as $N \rightarrow \infty$ for all $z \in X(\bar{q})$.
- (H8) For each convergent sequence $q^N \rightarrow \bar{q}$ in Q , there are constants \mathcal{M} , ω such that $|T^N(t; q^N)|_{q^N} \leq \mathcal{M} e^{\omega t}$, $|T(t; \bar{q})|_{q^N} \leq \mathcal{M} e^{\omega t}$ uniformly in $N = 1, 2, \dots$.
- (H9) For each convergent sequence $q^N \rightarrow \bar{q}$ in Q , one has for $z \in X(\bar{q})$, $|T^N(t; q^N)z - T(t; \bar{q})z|_{q^N} \rightarrow 0$ as $N \rightarrow \infty$, uniformly in $t \in [0, T]$.

The assumption (H2) will be taken as a standing hypothesis for the remainder of our discussions. For the approximating schemes we develop below, consistency will follow from (H7) while (H8) is a statement of stability. As we shall see, convergence of the schemes (which is (H9)) will follow from (H7), (H8) and the Trotter–Kato theorem.

Remark 2.1. It suffices, under the standing assumption (H2), that the following condition hold in place of (H6)(ii): For some fixed $q^* \in Q$ we have that for each $M > 0$ there is a function k_1 such that for all $q \in Q$ the relation $|F(q, t, u_1) - F(q, t, u_2)|_{q^*} \leq k_1(t)|u_1 - u_2|_{q^*}$ for all $u_1, u_2 \in X$ with $|u_i|_{q^*} \leq M$. Indeed, it is easily seen that this condition, along with (H2), implies (H6)(ii). Similarly, we can in the presence of (H2) equivalently postulate in place of (H6)(iii) the conditions: For some fixed $q^* \in Q$ there exists a function k_2 such that $|F(q, t, v)|_{q^*} \leq k_2(t)\{|v|_{q^*} + 1\}$ for all $v \in X$, $q \in Q$. We further note that existence of a function $k_3 \in L_2(0, T)$ such that $|F(q, t, 0)|_{\bar{q}} \leq k_3(t)$ for $q, \bar{q} \in Q$, a statement of the inequality of (H6)(iii) holding only for $|v|_{\bar{q}}$ sufficiently large (i.e., affine growth at ∞), along with (H6)(ii), are sufficient to imply (H6)(iii).

While the complete role played by the various hypotheses in our development will be clearer after our presentation, a few explanatory comments here might be helpful to readers. First, the desirability of the generality of allowing the underlying Hilbert space X for (2.1) to depend on q in such a way that (H2) obtains will not be apparent from the examples discussed here. (Rather, one must for this consider certain parabolic problems—see the comments in § 7.) However, in light of (H2) as a standing assumption, we are justified in suppressing the canonical isomorphism $\mathcal{J}^N: X(\bar{q}) \rightarrow X(q^N)$ in writing $|P^N(q^N)z - z|_{q^N} \rightarrow 0$ in (H7) rather than the technically correct statement $|P^N(q^N)\mathcal{J}^N z - \mathcal{J}^N z|_{q^N} \rightarrow 0$. Similar observations are pertinent for the statement of (H9) as well as in numerous other places in our presentation where we suppress the \mathcal{J}^N notation.

Condition (H4), while seemingly stringent, is an assumption often valid in practical problems where our theory might be useful. Since under (H3) $A(q)$ is closed, it follows from the closed graph theorem that $A^N(q)$ of (H5)(i) is, in fact, bounded and hence (H5)(i) follows immediately from (H3). It should be recognized that the form of the approximating operators defined in (H5) is a classical one (e.g., see [33, p. 369]) which has also recently been employed in the development of spline approximation techniques for delay differential equations [6]. The definition of $A^N(q)$ involves the implicit assumption that $X^N(q) \subset \text{Dom}(A(q))$; since our goal here is the rigorous formulation of modal approximation schemes for (ID), this restriction poses no difficulties. However, it does prevent a straightforward inclusion of low-order finite-element methods for higher-order partial differential equations in our approximation framework.

Hypothesis (H7) is a common requirement (e.g., see [24], [30]) in approximation theory, demanding that the sequence X^N of subspaces actually approximate the original state space X . Finally, (H6) is comprised of conditions on the nonlinearities in (2.1) that are sufficiently general to include many interesting problems of practical importance but yet are strong enough to guarantee global existence of solutions of (2.1) on

fixed finite intervals. As the knowledgeable reader might expect, these conditions can be replaced by alternate and/or weaker, hypotheses, but only, in general, at the cost of additional tedium in the proofs below. We have tried to compromise between strong conditions that are easily stated and employed in the proofs and ones that are as general (and weak) as possible. Further comments on this matter will be made in § 3.

Before defining the approximating equations for (2.1), we define the projection of the nonlinearity F onto X^N by $F^N(q, t, v) \equiv P^N(q)F(q, t, v)$ for each $(q, t, v) \in Q \times [0, T] \times X$. The approximating family of equations is then given by

$$(2.5) \quad \begin{aligned} \dot{v}(t) &= A^N(q)v(t) + F^N(q, t, v(t)), & t > 0, \\ v(0) &= P^N(q)u_0(q). \end{aligned}$$

Assuming existence of (mild) solutions to (2.5) (this will be established below), we denote (for a given q) these solutions by $u^N(t)$ or alternatively $u^N(t; q)$ or $u^N(t, x; q)$, depending on the context. We then define the approximate identification problems (ID^N) by:

(ID^N) Given observations $\hat{y} = \{\hat{y}(t_i)\}_{i=1}^r$ and a fit-to-data criterion J , minimize $J^N(q) \equiv J(q, u^N(\cdot; q), \hat{y})$ over $q \in Q$ subject to $u^N(\cdot; q)$ satisfying (2.5).

We note that if (in addition to (H1)) $X^N(q)$ is finite-dimensional, then (2.5) can be equivalently interpreted in the strong sense and (ID^N) then becomes an optimization problem constrained by finite-dimensional ordinary differential equations.

In our discussions below, we shall denote by \hat{q}^N any solution of (ID^N) so that it follows by definition that $J^N(\hat{q}^N) \leq J^N(q)$ for all $q \in Q$.

PROPOSITION 2.1. *Assume that (H2), (H3) and (H6) obtain. Then for each $q \in Q$ there exists a unique (mild) solution $u(\cdot; q) \in C(0, T; X(q))$ of (2.1). If, in addition, (H5) holds, there exists, for each $N = 1, 2, \dots$, a unique (mild) solution $u^N(\cdot; q) \in C(0, T; X(q))$ of (2.5).*

Proof. The proofs are completely standard and we only sketch the ideas for (2.1). Uniqueness follows immediately from (H6) and an application of Gronwall's inequality. Existence is established through the usual Picard iterate techniques. Define $v^0(t) = T(t; q)u_0(q)$ and for $j = 1, 2, \dots$,

$$(2.6) \quad v^j(t) = T(t; q)u_0(q) + \int_0^t T(t-s; q)F(q, s, v^{j-1}(s)) ds$$

for $t \in [0, T]$. From (H3) and (H6) it is easily seen that the iterates v^j are all well defined and $v^j: [0, T] \rightarrow X(q)$ is continuous. Moreover, $\{v^j\}_{j=0}^\infty$ is a bounded subset of $C(0, T; X)$. Employing (H6)(ii) and simple inductive arguments, one can establish that $\{v^j\}$ is Cauchy in $C(0, T; X)$. Passing to the limit in (2.6), one obtains the desired results. Existence of unique solutions of (2.5) is argued in an analogous manner by appealing to (H5) for appropriate boundedness.

THEOREM 2.1. *Assume hypotheses (H1)–(H6) hold and let $J(\cdot, \cdot, \hat{y}): Q \times C(0, T; X) \rightarrow \mathbb{R}^1$ be continuous. Moreover, suppose $q \rightarrow u_0(q)$, $q \rightarrow P^N(q)z$ and $q \rightarrow T^N(t; q)z$, $z \in X$, are continuous, with the latter uniformly in $t \in [0, T]$. Then: (i) There exists for each N a solution \hat{q}^N of (ID^N) and the sequence $\{\hat{q}^N\}$ possesses a convergent subsequence $\hat{q}^{N_k} \rightarrow \hat{q}$. (ii) If we further assume that, for any sequence $\{q^j\}$ in Q with $q^j \rightarrow q^0$, we have $|u^j(t; q^j) - u(t; q^0)|_{q^j} \rightarrow 0$ as $j \rightarrow \infty$ uniformly in $t \in [0, T]$, then \hat{q} is a solution of (ID).*

Proof. We show for fixed N that $q \rightarrow J^N(q) \equiv J(q, u^N(\cdot; q), \hat{y})$ is continuous on Q which, in the light of (H4), yields (i). First note that u^N satisfies

$$(2.7) \quad u^N(t; q) = T^N(t; q)P^N(q)u_0(q) + \int_0^t T^N(t-s; q)P^N(q)F(q, s, u^N(s; q)) ds$$

on $[0, T]$. In view of (H5) and (H2) we find that there exist constants M and M_1 such that $|T^N(t; q)|_{\tilde{q}} \leq M$, $|P^N(q)|_{\tilde{q}} \leq M_1$ uniformly for $t \in [0, T]$, $q, \tilde{q} \in Q$. It follows from (H6)(iii) and (2.7) that

$$\begin{aligned} |u^N(t; q)|_q &\leq M|P^N(q)u_0(q)|_q + M \int_0^t |P^N(q)F(q, s, u^N(s; q))|_q ds \\ &\leq M|u_0(q)| + M \int_0^t k_2(s)\{|u^N(s; q)|_q + 1\} ds. \end{aligned}$$

Since $q \rightarrow u_0(q)$ is continuous, this implies (via Gronwall's inequality) that $|u^N(t; q)|_q$ is uniformly bounded for $(t, q) \in [0, T] \times Q$. This in turn implies (by (H6)(iii)) that the mapping $s \rightarrow T^N(t-s; q)P^N(\tilde{q})F(\tilde{q}, s, u^N(s; \tilde{q}))$ from $[0, T]$ to X is dominated by an integrable function uniformly in $q, \tilde{q}, \tilde{q} \in Q$. (This will permit us to invoke, below, the usual dominated convergence theorem.) Assuming that $q^i \rightarrow \tilde{q}$, $q^i, \tilde{q} \in Q$ are arbitrary, we obtain the following estimates:

$$\begin{aligned} &|u^N(t; \tilde{q}) - u^N(t; q^i)| \\ &\leq |T^N(t; \tilde{q})P^N(\tilde{q})u_0(\tilde{q}) - T^N(t; q^i)P^N(\tilde{q})u_0(\tilde{q})| \\ &\quad + |T^N(t; q^i)P^N(\tilde{q})u_0(\tilde{q}) - T^N(t; q^i)P^N(q^i)u_0(\tilde{q})| \\ &\quad + |T^N(t; q^i)P^N(q^i)u_0(\tilde{q}) - T^N(t; q^i)P^N(q^i)u_0(q^i)| \\ &\quad + \int_0^t |T^N(t-s; \tilde{q}) - T^N(t-s; q^i)|P^N(\tilde{q})F(\tilde{q}, s, u^N(s; \tilde{q}))| ds \\ &\quad + \int_0^t |T^N(t-s; q^i)(P^N(\tilde{q}) - P^N(q^i))F(\tilde{q}, s, u^N(s; \tilde{q}))| ds \\ &\quad + \int_0^t |T^N(t-s; q^i)P^N(q^i)\{F(\tilde{q}, s, u^N(s; \tilde{q})) - F(q^i, s, u^N(s; q^i))\}| ds \\ &= \rho_1(j) + \rho_2(j) + \rho_3(j) + \rho_4(j) + \rho_5(j) + \rho_6(j), \end{aligned}$$

where the ρ_i 's are defined as indicated (ρ_i the i th term), $i = 1, \dots, 6$, and all norms are $|\cdot|_{\tilde{q}}$. We then have by hypothesis

$$\rho_1(j) = |\{T^N(t; \tilde{q}) - T^N(t; q^i)\}P^N(\tilde{q})u_0(\tilde{q})| \rightarrow 0 \quad \text{as } j \rightarrow \infty,$$

uniformly in $t \in [0, T]$. Also ρ_2 and $\rho_3 \rightarrow 0$ by the continuity assumptions on P^N and u_0 and the boundedness of $T^N(t; q^i)$. Dominated convergence implies that $\rho_4 \rightarrow 0$ and $\rho_5 \rightarrow 0$ as $j \rightarrow \infty$. Finally,

$$\begin{aligned} \rho_6(j) &\leq MM_1 \left\{ \int_0^t |F(\tilde{q}, s, u^N(s; \tilde{q})) - F(q^i, s, u^N(s; \tilde{q}))| ds \right. \\ &\quad \left. + \int_0^t |F(q^i, s, u^N(s; \tilde{q})) - F(q^i, s, u^N(s; q^i))| ds \right\} \\ &\leq MM_1 \rho_7(j) + \int_0^t k_1(s)|u^N(s; \tilde{q}) - u^N(s; q^i)| ds, \end{aligned}$$

where $\rho_7(j) \rightarrow 0$ as $j \rightarrow \infty$ and k_1 depends on the uniform bounds for $u^N(t; \tilde{q})$, $u^N(t; q^j)$ (see (H6)(ii)). Thus we find

$$|u^N(t; \tilde{q}) - u^N(t; q^j)| \leq \varepsilon_j(t) + \int_0^t k_1(s) |u^N(s; \tilde{q}) - u^N(s; q^j)| ds,$$

where $\varepsilon_j(t) \rightarrow 0$, uniformly in $t \in [0, T]$, as $j \rightarrow \infty$. Applying Gronwall's inequality, we have uniform (in t) continuity of $q \rightarrow u^N(t; q)$ on Q which implies the desired continuity of J^N .

Turning to (ii) and letting $\{\hat{q}^{N_i}\}$ be a convergent subsequence with $\hat{q}^{N_i} \rightarrow \hat{q}$, we first observe that $J^{N_i}(\hat{q}^{N_i}) \leq J^{N_i}(q)$ for all $q \in Q$. By hypothesis, $u^{N_i}(t; \hat{q}^{N_i}) \rightarrow u(t; \hat{q})$ and furthermore $u^{N_i}(t; q) \rightarrow u(t; q)$ for each $q \in Q$, with convergence in both cases uniform in t on $[0, T]$. This implies $J^{N_i}(\hat{q}^{N_i}) \rightarrow J(\hat{q}, u(\cdot; \hat{q}), \hat{y})$ and $J^{N_i}(q) \rightarrow J(q, u(\cdot; q), \hat{y})$ as $j \rightarrow \infty$ and hence from the above inequality we obtain, by passing to the limit, $J(\hat{q}, u(\cdot; \hat{q}), \hat{y}) \leq J(q, u(\cdot; q), \hat{y})$ for all $q \in Q$. Thus \hat{q} is a solution of (ID) and Theorem 2.1 is established.

Remark 2.2. If $v \rightarrow J(q, v, \hat{y})$ as a mapping on $C(0, T; X)$ actually depends only on a finite number of values $v(t_i)$, $t_i \in [0, T]$, as, for example, in (2.3) or (2.4), then the hypotheses of Theorem 2.1 involving uniformity in t can be relaxed to statements holding only for each fixed $t \in [0, T]$. The above arguments remain unchanged except that the uniform (in t) convergence remarks are replaced with pointwise convergence statements (see especially the term $\rho_1(j)$ in the proof).

We conclude this section with a brief explanation of how (2.5) (or (2.7)) is to be used in actual computations. We adopt notation very similar to that found in [6] in the development of spline methods for delay systems. We assume that X^N is finite-dimensional and choose a basis independent of q (recall that (H2) is a standing hypothesis) by

$$\hat{\beta}^N = (\beta_1, \dots, \beta_{d(N)})$$

where $d(N) = \dim X^N(q)$. From (2.7) under (H5) we see that the solution u^N of (2.5) satisfies $u^N(t; q) \in X^N$ for all t and hence there exists a representation $u^N(t; q) = \hat{\beta}^N w^N(t; q)$ with $w^N(t; q) = \text{col}(w_1^N(t; q), \dots, w_{d(N)}^N(t; q)) \in \mathbb{R}^{d(N)}$. We let $[A^N(q)]$ and $[F^N(q, t, w^N)]$ denote the matrix and coordinate representations relative to $\hat{\beta}^N$ of $A^N(q)$ and $P^N(q)F(q, t, u^N)$, respectively. The coordinate representation of (2.5) is then given by

$$(2.8) \quad \begin{aligned} \dot{w}^N(t; q) &= [A^N(q)]w^N(t; q) + [F^N(q, t, w^N(t; q))], \quad t > 0, \\ w^N(0; q) &= \gamma^N, \end{aligned}$$

where γ^N is defined through $P^N(q)u_0(q) = \hat{\beta}^N w^N(0; q)$. For any $z \in X^N$ the associated coordinate vector $\alpha^N \in \mathbb{R}^{d(N)}$ in $P^N(q)z = \hat{\beta}^N \alpha^N$ is determined uniquely by the condition $(P^N(q)z - z) \perp X^N$, or equivalently $\langle \hat{\beta}^N, \hat{\beta}^N \rangle_q \alpha^N = \langle \hat{\beta}^N, z \rangle_q$. Thus we have

$$(2.9) \quad \alpha^N = (Q^N)^{-1} R^N z$$

where Q^N is the $d(N) \times d(N)$ matrix with elements $\langle \beta_i^N, \beta_j^N \rangle_q$ and $(R^N z)_i = \langle \beta_i^N, z \rangle_q$ for $i = 1, 2, \dots, d(N)$. Arguing exactly as in [6, pp. 508–509], we therefore find

$$(2.10) \quad [A^N(q)] = (Q^N)^{-1} K^N,$$

where K^N is the $d(N) \times d(N)$ matrix with elements $K_{ij}^N = \langle \beta_i^N, A(q)\beta_j^N \rangle_q$, and

$$(2.11) \quad [F^N(q, t, w^N)] = (Q^N)^{-1} R^N F(q, t, u^N).$$

We thus arrive at the final form of the approximating system for (2.1) in $X^N(q)$ as

$$(2.12) \quad \begin{aligned} Q^N \dot{w}^N(t) &= K^N w^N(t) + R^N F(q, t, \hat{\beta}^N w^N(t)), \quad t > 0, \\ w^N(0) &= (Q^N)^{-1} R^N u_0(q). \end{aligned}$$

3. Approximation theorems for abstract systems. In this section we shall focus our attention on the condition “ $q^j \rightarrow \bar{q}$ implies $u^j(t; q^j) \rightarrow u(t; \bar{q})$ ” of Theorem 2.1(ii) and present results on the convergence of solutions of the approximating systems (2.7) to solutions of (2.1). We state and prove two theorems; the first is applicable to nonlinear parameter identification problems (§ 4) while the second will be used in connection with linear boundary control problems in § 5.

THEOREM 3.1. *Suppose hypotheses (H1)–(H3) and (H5)–(H9) hold and let q^N, \bar{q} be arbitrary in Q such that $q^N \rightarrow \bar{q}$. Further, suppose that $|u_0(q^N) - u_0(\bar{q})|_{q^N} \rightarrow 0$ as $N \rightarrow \infty$. Then the mild solutions $u^N(t; q^N)$ of*

$$(3.1) \quad \begin{aligned} \dot{u}^N(t) &= A^N(q^N)u^N(t) + F^N(q^N, t, u^N(t)), \\ u^N(0) &= P^N(q^N)u_0(q^N) \end{aligned}$$

converge to the mild solution $u(t; \bar{q})$ of (2.1) for each $t \in [0, T]$. If $(t, v) \rightarrow F(\bar{q}, t, v)$ is continuous on $[0, T] \times X$, then the convergence $|u^N(t; q^N) - u(t; \bar{q})|_{q^N} \rightarrow 0$ is uniform in t on $[0, T]$.

Proof. Let $q^N \rightarrow \bar{q}$ be arbitrary as hypothesized. Recalling the proof of Theorem 2.1, we observe that one easily argues existence of a constant K such that $|u^N(t; q^N)|_{q^N} \leq K, |u(t; \bar{q})|_{q^N} \leq K$ for all N and $t \in [0, T]$. Further, we see that for $t \in [0, T]$ we have (where all norms are $|\cdot|_{q^N}$)

$$\begin{aligned} & |u^N(t; q^N) - u(t; \bar{q})| \\ & \leq |T^N(t; q^N)P^N(q^N)u_0(q^N) - T^N(t; q^N)P^N(q^N)u_0(\bar{q})| \\ & \quad + |T^N(t; q^N)P^N(q^N)u_0(\bar{q}) - T^N(t; q^N)u_0(\bar{q})| \\ & \quad + |T^N(t; q^N)u_0(\bar{q}) - T(t; \bar{q})u_0(\bar{q})| \\ & \quad + \int_0^t |T^N(t-s; q^N)P^N(q^N)\{F(q^N, s, u^N(s; q^N)) - F(q^N, s, u(s; \bar{q}))\}| ds \\ & \quad + \int_0^t |T^N(t-s; q^N)P^N(q^N)\{F(q^N, s, u(s; \bar{q})) - F(\bar{q}, s, u(s; \bar{q}))\}| ds \\ & \quad + \int_0^t |T^N(t-s; q^N)\{P^N(q^N) - I\}F(\bar{q}, s, u(s; \bar{q}))| ds \\ & \quad + \int_0^t |\{T^N(t-s; q^N) - T(t-s; \bar{q})\}F(\bar{q}, s, u(s; \bar{q}))| ds \\ & = \varepsilon_1(N) + \varepsilon_2(N) + \varepsilon_3(N) \\ & \quad + \int_0^t |T^N(t-s; q^N)P^N(q^N)\{F(q^N, s, u^N(s; q^N)) - F(q^N, s, u(s; \bar{q}))\}| ds \\ & \quad + \varepsilon_4(N) + \varepsilon_5(N) + \varepsilon_6(N). \end{aligned}$$

By (H8), our hypotheses and the definition of $P^N(q^N)$ in (H5), we find $|\varepsilon_1(N)| \leq \mathcal{M} e^{\omega T} |u_0(q^N) - u_0(\bar{q})| \rightarrow 0$ as $N \rightarrow \infty$. Also, $|\varepsilon_2(N)| \leq \mathcal{M} e^{\omega T} |P^N(q^N) - I| u_0(\bar{q}) \rightarrow 0$ by

(H8) and (H7). That $|\varepsilon_3(N)| \rightarrow 0$ uniformly in t on $[0, T]$ follows directly from (H9). Moreover,

$$|\varepsilon_4(N)| \leq \mathcal{M} e^{\omega T} \int_0^T |F(q^N, s, u(s; \bar{q})) - F(\bar{q}, s, u(s; \bar{q}))| ds \rightarrow 0$$

by (H6)(iv), (H6)(iii) and dominated convergence, while

$$|\varepsilon_5(N)| \leq \mathcal{M} e^{\omega T} \int_0^T |\{P^N(q^N) - I\}F(\bar{q}, s, u(s; \bar{q}))| ds \rightarrow 0$$

by (H7), (H6)(iii) and dominated convergence. Finally,

$$|\varepsilon_6(N)| = \int_0^t |\{T^N(t-s; q^N) - T(t-s; \bar{q})\}F(\bar{q}, s, u(s; \bar{q}))| ds \rightarrow 0$$

by (H9) and dominated convergence ((H8) with (H6)(iii)) for each fixed $t \in [0, T]$.

We note that the convergence in all of the terms above, except ε_6 , is uniform in t on $[0, T]$. If, in addition, the continuity hypothesis of the theorem obtains, we find that $\{F(\bar{q}, s, u(s; \bar{q})) | s \in [0, T]\}$ is a compact subset of X and the convergence in the integrand of ε_6 is uniform in t and s ; hence in this case $\varepsilon_6 \rightarrow 0$ uniformly in t also.

We have thus established the following estimate:

$$\begin{aligned} |u^N(t; q^N) - u(t; \bar{q})| &\leq \sum_{i=1}^6 \varepsilon_i + \mathcal{M} e^{\omega t} \int_0^t |F(q^N, s, u^N(s; q^N)) - F(q^N, s, u(s; \bar{q}))| ds \\ &\leq \varepsilon^N(t) + \mathcal{M} e^{\omega T} \int_0^t k_1(s) |u^N(s; q^N) - u(s; \bar{q})| ds \end{aligned}$$

where $\varepsilon^N \rightarrow 0$ as $N \rightarrow \infty$. An application of Gronwall's inequality then yields that $|u^N(t; q^N) - u(t; \bar{q})| \rightarrow 0$ as $N \rightarrow \infty$, where the convergence is uniform in t under the added continuity hypothesis of the theorem.

COROLLARY 3.1. *Under the hypotheses of Theorem 3.1, $u^N(t; q) \rightarrow u(t; q)$ for each fixed $q \in Q$, uniformly in t on $[0, T]$ if, in addition, $(t, v) \rightarrow F(q, t, v)$ is continuous on $[0, T] \times X$.*

We turn next to convergence results needed for optimal control problems. Consider for fixed $q \in Q$ the system

$$\begin{aligned} \dot{u}(t) &= A(q)u(t) + G(q, t), \\ u(0) &= u_0, \end{aligned} \tag{3.2}$$

and the approximating system

$$\begin{aligned} \dot{u}^N(t) &= A^N(q)u^N(t) + P^N(q)G(q, t), \\ u^N(0) &= P^N(q)u_0, \end{aligned} \tag{3.3}$$

where G has the form $G(q, t) = \gamma(q)\sigma(t)$. We assume $\gamma(q) \in \Gamma$ where Γ is a subset of $\{\tilde{\gamma} = (\tilde{\gamma}_1, \dots, \tilde{\gamma}_\mu) | \tilde{\gamma}_i : Q \rightarrow \hat{X} \subset X\}$ with \hat{X} a given subset of X . We further assume $\sigma \in \Sigma$, Σ a given subset of $L_2(0, T; R^\mu)$.

THEOREM 3.2. *Assume (H1)–(H3), (H5), (H7)–(H9). Suppose moreover that \hat{X} is compact and Σ is bounded. Then for each fixed $q \in Q$, mild solutions u^N of (3.3) converge to the mild solutions of (3.2), uniformly in $\sigma \in \Sigma$, $\gamma \in \Gamma$ and $t \in [0, T]$.*

Proof. We consider the estimates in the proof of Theorem 3.1 with $F(q, s, v) = G(q, s)$ and $q^N = \bar{q} = q$ fixed. Then

$$|u^N(t; q) - u(t; q)| \leq \varepsilon_2(N) + \varepsilon_3(N) + \varepsilon_5(N) + \varepsilon_6(N)$$

($\varepsilon_1 = \varepsilon_4 = 0$), where ε_2 and ε_3 are as before and

$$\varepsilon_5 = \int_0^t |T^N(t-s; q)\{P^N(q) - I\}G(q, s)| ds,$$

$$\varepsilon_6 = \int_0^t |\{T^N(t-s; q) - T(t-s; q)\}G(q, s)| ds,$$

and, as usual, all X norms are $|\cdot|_q$. We have immediately (using (H7), (H8), (H9)) that ε_2 and $\varepsilon_3 \rightarrow 0$ as $N \rightarrow \infty$, uniformly in σ , γ and $t \in [0, T]$. Also,

$$\begin{aligned} |\varepsilon_5(N)| &\leq \mathcal{M} e^{\omega T} \int_0^T |(P^N(q) - I)\gamma(q)\sigma(s)| ds \\ &\leq \mathcal{M} e^{\omega T} \max_{1 \leq i \leq \mu} |(P^N(q) - I)\gamma_i(q)| \int_0^T |\sigma(s)| ds. \end{aligned}$$

But since \hat{X} is compact and Σ is bounded, this latter term $\rightarrow 0$ as $N \rightarrow \infty$ uniformly in $\gamma \in \Gamma$, $\sigma \in \Sigma$ and $t \in [0, T]$. Finally,

$$\begin{aligned} |\varepsilon_6(N)| &\leq \max_{1 \leq i \leq \mu} \int_0^t |\{T^N(t-s; q) - T(t-s; q)\}\gamma_i(q)| |\sigma(s)| ds \\ &\leq \max_{1 \leq i \leq \mu} \left\{ \int_0^t |\{T^N(t-s; q) - T(t-s; q)\}\gamma_i(q)|^2 ds \right\}^{1/2} |\sigma|_{L_2(0, T)}, \end{aligned}$$

and this last estimate yields $\varepsilon_6(N) \rightarrow 0$ uniformly in $\gamma \in \Gamma$, $\sigma \in \Sigma$ and $t \in [0, T]$, again from the compactness of \hat{X} , (H9), and the boundedness of Σ .

As we have previously noted, the main purpose of (H6) is to allow us to guarantee existence of solutions of (2.1) and (2.5) on fixed finite intervals $[0, T]$ (see Proposition 2.1). The condition (H6)(iii) is used in the proofs of Theorem 2.1 and 3.1 only to establish uniform bounds on the u^N . This permits us to employ the local Lipschitz condition (H6)(ii) and to appeal to the dominated convergence theorem in certain arguments. We have already noted that (H6)(iii) can be relaxed to “affine growth at ∞ ” (see Remark 2.1). With an alternative approach, one can relax this growth condition even further and still obtain the conclusions of Theorems 2.1 and 3.1 (with the other hypotheses remaining unchanged). Specifically, for N sufficiently large, the initial data and defining operators T^N, P^N and T for u^N and u , respectively, are close. Thus if one *assumes* (in place of (H6)(iii))

- (A6) (i) For each $q \in Q$, there exists a solution $u(t; q)$ of (2.1) on $[0, T]$, and
 (ii) There exists $k_3 \in L_2(0, T)$ such that $|F(q, t, 0)|_{\bar{q}} \leq k_3(t)$, for $q, \bar{q} \in Q$ and $t \in [0, T]$,

it is rather tedious but not difficult to show that for N sufficiently large, all u^N defined by (2.7) exist on $[0, T]$ and lie in some bounded neighborhood of u , the solution of (2.1). (The arguments involve use of classical fixed point ideas to obtain solutions on some interval $[0, \delta_1]$ and then continuation to $[\delta_1, 2\delta_1], \dots$, etc.) The

condition (A6)(ii) can then be combined with (H6)(ii) to obtain domination of terms such as $F(q, s, u^N(s; q))$. Thus all of the arguments behind Theorems 2.1 and 3.1 remain valid, the hypotheses being changed only in that (A6)(i), (A6)(ii) replace (H6)(iii), and the conclusions changed only in that they can be obtained only for N sufficiently large (which of course is not important *theoretically* in approximation results such as those discussed in this paper).

Regarding the assumption (A6)(i), we note that there are various conditions one might impose on F to insure existence. For example, monotonicity hypotheses might be assumed so that $-(A + F)$ is maximal monotone and one could then appeal to standard existence results [9], [19]. In § 6 we present numerical results for our approximation scheme for an example (Example 6.5) in which (H6)(iii) is not satisfied, yet (A6)(i) and (A6)(ii) do hold. However, we shall not pursue any of the theoretical ideas here, since this is really not the focus of our presentation.

4. Examples: Parameter identification in hyperbolic and parabolic equations.

We turn now to an application of the results developed in the preceding sections to identification problems for specific equations. A fundamental requirement in both Theorems 3.1 and 3.2 is that the conditions of (H7), (H8) and (H9) be verified. As we have indicated earlier, the convergence statement of (H9) can be obtained rather easily for our schemes from (H7), (H8) and some standard approximation results from linear semigroup theory. We state here, for our future reference, one version (due to Kurtz [24]) of these approximation theorems.

PROPOSITION 4.1. *Let $(\mathcal{B}, |\cdot|)$ and $(\mathcal{B}^N, |\cdot|_N)$, $N = 1, 2, \dots$, be Banach spaces and let $\pi^N : \mathcal{B} \rightarrow \mathcal{B}^N$ be bounded linear operators. Assume further that $\mathcal{T}(t)$ and $\mathcal{T}^N(t)$ are linear C_0 -semigroups on \mathcal{B} and \mathcal{B}^N with infinitesimal generators \mathcal{A} and \mathcal{A}^N , respectively. If*

(i) $\lim_{N \rightarrow \infty} |\pi^N z|_N = |z|$ for all $z \in \mathcal{B}$,

(ii) *there exist constants \tilde{M} , $\tilde{\omega}$ independent of N such that $|\mathcal{T}^N(t)| \leq \tilde{M} e^{\tilde{\omega} t}$, $t \geq 0$,*

(iii) *there exists a set $\mathcal{D} \subset \mathcal{B}$ such that $\mathcal{D} \subset \text{Dom}(\mathcal{A})$, $\tilde{\mathcal{D}} = \mathcal{B}$, and $(\lambda_0 - \mathcal{A})\tilde{\mathcal{D}} = \mathcal{B}$ for some $\lambda_0 > 0$,*

(iv) *for $z \in \mathcal{D}$ we have $\lim_{N \rightarrow \infty} |\mathcal{A}^N \pi^N z - \pi^N \mathcal{A} z|_N = 0$,*

then $\lim_{N \rightarrow \infty} |\mathcal{T}^N(t) \pi^N z - \pi^N \mathcal{T}(t) z|_N = 0$ for $z \in \mathcal{B}$, uniformly in t on compact subsets of $[0, \infty)$.

We note that the requirement in (iii) implies that \mathcal{D} is a core [21, p. 166] of \mathcal{A} ; this is easily seen using the fact that \mathcal{A} , being an infinitesimal generator, is closed and $(\lambda I - \mathcal{A})^{-1}$ is bounded for λ sufficiently large. The proposition then follows directly from Theorem 2.1 of [24] taken with subsequent remarks [24, p. 361] of that reference. Obviously, (iii) in our statement above could be replaced by the hypothesis that \mathcal{D} be a core of \mathcal{A} . Further, we remark that the requirement $\tilde{\mathcal{D}} = \mathcal{B}$ is superfluous in (iii) if one verifies that $(\lambda_0 - \mathcal{A})\mathcal{D} = \mathcal{B}$ for $\lambda_0 \in \rho(\mathcal{A})$. In this case one can easily demonstrate directly that \mathcal{D} is a core for the generator \mathcal{A} .

In the examples we discuss below, we shall use the notation A, F, A^N, P^N, F^N to denote the specific operators in each example, since this will facilitate reference back to the basic theorems of §§ 2 and 3 and should cause no confusion for readers. However, we shall adopt distinct notation for the various state spaces X, X^N within the context of each example.

Example 4.1. Hyperbolic equations. We consider the one-dimensional hyperbolic equation

$$(4.1) \quad v_{tt} = q_1 v_{xx} + q_2 v_t + q_3 v + f(q_6, t, x, v, v_t)$$

with initial and boundary conditions

$$(IC_1) \quad v(0, x) = \sum_{i=1}^m q_4^i \phi_i(x),$$

$$v_t(0, x) = \sum_{i=1}^m q_5^i \psi_i(x) \quad \text{for } 0 \leq x \leq 1,$$

$$(BC_1) \quad v(t, 0) = v(t, 1) = 0 \quad \text{for } t > 0,$$

where $v = v(t, x)$, $q_6 \in R^m$ and the remaining q_i, q_i^j are scalars. The vector parameter q of § 2 thus has dimension $k = 3m + 3$. (If the output maps such as Y or C in the fit-to-data functions (2.3) and (2.4) depend explicitly on some parameters q_7^1, \dots, q_7^m , we assume with no loss of generality that these have been embedded in the q_6 (or q_4 or q_5) vector.)

We remark that we do not formulate nontrivial boundary conditions, possibly depending on parameters, in (4.1)–(IC₁)–(BC₁); however, it is easily seen that by simple transformations such generalities actually can be included in our formulation above. Consider, for example,

$$(4.2) \quad v_{tt} = q_1 v_{xx}$$

with initial and boundary conditions

$$(IC_2) \quad v(0, x) = q_4 \phi(x),$$

$$v_t(0, x) = q_5 \psi(x),$$

$$(BC_2) \quad v(t, 0) = q_7 b_1(t), \quad v(t, 1) = q_8 b_2(t),$$

where b_1, b_2 are twice continuously differentiable functions. Employing the standard transformation $w(t, x) = v(t, x) - (1-x)q_7 b_1(t) - xq_8 b_2(t)$, we find that (4.2)–(IC₂)–(BC₂) can be reformulated as a special case of (4.1)–(IC₁)–(BC₁).

Returning to (4.1), we proceed in the usual manner to rewrite the equation with boundary and initial conditions as an abstract evolution equation. Let Δ denote the Laplacian operator $\partial^2/\partial x^2$ in $H^0 = L_2(0, 1; R)$; here and below the Sobolev spaces H^i consist of R^1 -valued functions on $[0, 1]$ taken with their usual inner products unless otherwise specified. It is well known that Δ with $\text{Dom}(\Delta) = H_0^1 \cap H^2$ is a self-adjoint operator in H^0 satisfying $\langle -\Delta z, z \rangle \geq |z|^2$ for all $z \in \text{Dom}(\Delta)$. We impose the following additional assumption on the coefficient q_1 in (4.1):

$$(HQ) \quad \text{There exist positive numbers } q_1^L \text{ and } q_1^U \text{ such that } q \in Q \subset R^k \text{ implies } q_1^L \leq q_1 \leq q_1^U.$$

For a given $q \in Q \subset R^{3m+3}$ we of course mean by q_1 the first coordinate of the vector $q = (q_1, \dots, q_6)$ where $q_j = (q_j^1, \dots, q_j^m)$, $j = 4, 5, 6$.

Having thus assumed hypothesis (HQ) for a given fixed parameter restraint set Q , we endow the set H_0^1 with a family of inner products

$$\langle w, z \rangle_{q_1} = \int_0^1 q_1 \frac{\partial w}{\partial x} \frac{\partial z}{\partial x} = \langle q_1 w_x, z_x \rangle$$

where q ranges over Q . In view of (HQ), the space $(H_0^1, \langle \cdot, \cdot \rangle_{q_1})$ is, for each $q \in Q$, a Hilbert space which we denote by $V(q_1)$. The space $X(q)$ of § 2 is chosen for this example to be $X(q) = \mathcal{H}(q) \equiv V(q_1) \times H^0$ with the usual product topology generated by $\langle (w_1, w_2), (z_1, z_2) \rangle_q = \langle w_1, z_1 \rangle_{q_1} + \langle w_2, z_2 \rangle$. Condition (H2) is obviously satisfied since

for any $\tilde{q}, q \in Q$ we find $|z|_{\tilde{q}} \leq \mathcal{K}|z|_q$ for all $z \in \mathcal{H}$ where $\mathcal{H} \equiv (q_1^U/q_1^L)^{1/2}$. Introducing the variable $w(t) \equiv v$, we may rewrite (4.1)–(IC₁)–(BC₁) in \mathcal{H} by

$$(4.3) \quad \begin{aligned} \frac{d}{dt} \begin{pmatrix} v(t) \\ w(t) \end{pmatrix} &= A(q) \begin{pmatrix} v(t) \\ w(t) \end{pmatrix} + F(q, t, v(t), w(t)), \quad t > 0, \\ \begin{pmatrix} v(0) \\ w(0) \end{pmatrix} &= \begin{pmatrix} \sum q_4^i \phi_i \\ \sum q_5^i \psi_i \end{pmatrix}, \end{aligned}$$

where $(\phi_i, \psi_i) \in \mathcal{H}$, $\text{Dom}(A(q)) = H_0^1 \cap H^2 \times H_0^1$,

$$A(q) = \begin{pmatrix} 0 & 1 \\ q_1 \Delta + q_3 & q_2 \end{pmatrix},$$

and

$$F(q, t, v(t), w(t)) = \begin{pmatrix} 0 \\ f(q_6, t, \cdot, v(t, \cdot), w(t, \cdot)) \end{pmatrix}.$$

Before turning to a careful discussion of (4.3), we define the operators, etc., needed in formulating the *modal approximation scheme* associated with (4.1).

Since the operator $-\Delta$ is self-adjoint with compact resolvent, standard results in spectral theory and the theory of bilinear forms (see [15, p. 1331], [39, p. 343], [33, pp. 247–254]) are applicable. Defining $\tilde{\Phi}_j(x) = (\sqrt{2}/j\pi) \sin j\pi x$ and $\Phi_j(x) = \sqrt{2} \sin j\pi x$, we find that $\{\tilde{\Phi}_j\}_{j=1}^\infty$ and $\{\Phi_j\}_{j=1}^\infty$ constitute complete orthonormal sets (CONS) for $V(1) (= H_0^1)$ and H^0 , respectively. The corresponding modal subspaces $X^N(q) = \mathcal{H}^N(q)$ of $\mathcal{H}(q)$ are then defined by $\mathcal{H}^N(q) = \text{span}\{\beta_1^N, \dots, \beta_{2N}^N\}$ where

$$\beta_j^N = \begin{pmatrix} \tilde{\Phi}_j \\ 0 \end{pmatrix}, \quad j = 1, \dots, N, \quad \beta_j^N = \begin{pmatrix} 0 \\ \Phi_{j-N} \end{pmatrix}, \quad j = N+1, \dots, 2N.$$

It is easily seen that $\bigcup_{N=1}^\infty \{\{\beta_j^N\}_{j=1}^{2N}\}$ forms a CONS for $\mathcal{H}(q^*)$ where $q^* = (1, 0, \dots, 0)$, and a complete orthogonal, but not normal, set for $\mathcal{H}(q)$, for $q = (q_1, \dots, q_6)$ with $q_1 > 0$, $q_1 \neq 1$. We note also that $\Phi_j, \tilde{\Phi}_j$ are eigenfunctions of Δ corresponding to the eigenvalues $\lambda = -j^2\pi^2$, $j = 1, 2, \dots$.

The subspaces $\mathcal{H}^N(q)$ and the corresponding orthogonal projections $P^N(q)$ (see (H5)) having been thus defined, the modal approximation operators $A^N(q)$ for $A(q)$ are determined as in (H5). The corresponding matrices Q^N and K^N of (2.10), which in this case are $2N \times 2N$ -matrices, are readily seen to be given by

$$(4.4) \quad Q^N = \text{diag}(q_1, \dots, q_1, 1, \dots, 1),$$

where the q_1 and 1 each appear N times, and

$$(4.5) \quad K^N = \begin{pmatrix} 0 & D_1^N \\ D_2^N & D_3^N \end{pmatrix},$$

where the D_j^N , $j = 1, 2, 3$, are $N \times N$ diagonal matrices defined by $D_1^N = \text{diag}(\pi q_1, 2\pi q_1, \dots, N\pi q_1)$, $D_2^N = \text{diag}(q_3/\pi, q_3/2\pi, \dots, q_3/N\pi) - D_1^N$, and $D_3^N = q_2 I$. Recalling (2.9), we observe that in this case the projection operators $P^N(q)$ are *actually independent of q* .

We are now in a position to verify that (H8) and (H9) obtain for the hyperbolic examples under consideration.

THEOREM 4.1. *Assume that (HQ) holds and let $\{q^N\}$ be an arbitrary sequence in $Q \subset \mathbb{R}^{3m+3}$ converging to $\tilde{q} \in Q$. Then the operators $A(\tilde{q})$ and $A^N(q^N)$ of (4.3) and*

the associated modal approximations described above generate C_0 -semigroups $T(t; \bar{q})$ and $T^N(t; q^N)$ on $\mathcal{H}(\bar{q})$ and $|T^N(t; q^N)z - T(t; \bar{q})z|_{q^N} \rightarrow 0$ for $z \in \mathcal{H}(\bar{q})$, uniformly in t on compact subsets of $[0, \infty)$. If we further assume (H4) (Q is compact), then there exists $\omega \in \mathbb{R}$, independent of $q \in Q$ and N , such that $|T(t; q)| \leq e^{\omega t}$ and $|T^N(t; q)| \leq e^{\omega t}$ for all $q \in Q$, $t \geq 0$, $N = 1, 2, \dots$.

Proof. For any $q \in Q$, a straightforward calculation shows that

$$iA_0(q) = \begin{pmatrix} 0 & i \\ iq_1\Delta & 0 \end{pmatrix}$$

with $\text{Dom}(A_0(q)) = H_0^1 \cap H^2 \times H_0^1$ is a symmetric operator in $\mathcal{H}(q)$. Furthermore, $\text{Dom}(A_0(q))$ is dense and A_0 is invertible. It follows [28, p. 97] that A_0 is skew adjoint and hence by Stone's theorem [26, p. 252], [41, p. 345] generates a unitary group on $\mathcal{H}(q)$. Defining the operator $\tilde{A}(q)$ on $\mathcal{H}(q)$ by $\tilde{A}(q)(z_1, z_2) = (0, q_3z_1 + q_2z_2)$, it is easily seen that $\tilde{A}(q)$ is bounded. Indeed, using the fact that the H_0^1 norm is stronger than the H^0 norm, one finds $|\tilde{A}(q)(z_1, z_2)|_q \leq c(q_1, q_2, q_3)|(z_1, z_2)|_q$ where the constant c is bounded above uniformly on Q if (HQ) and (H4) hold, say $c(q_1, q_2, q_3) \leq \omega$.

We thus find that $A(q) = A_0(q) + \tilde{A}(q)$ is the perturbation of A_0 by a bounded operator and hence [21, p. 495], [30, p. 80] generates a C_0 -semigroup $T(t; q)$ on $\mathcal{H}(q)$ satisfying

$$(4.6) \quad |T(t; q)| \leq \exp \{c(q_1, q_2, q_3)t\},$$

where there exist $\omega > 0$ independent of q such that $c(q_1, q_2, q_3) \leq \omega$ in case (HQ) and (H4) obtain (or in case (HQ) holds and q lies in a bounded subset of Q).

As we have pointed out earlier (see the remarks in § 2), $A^N(q)$ is a bounded linear operator for each N and hence generates a C_0 -semigroup on $\mathcal{H}(q)$. Assuming that (HQ) holds and q lies in a bounded subset of Q , we have that $A(q)$ is the infinitesimal generator of a C_0 -semigroup satisfying $|T(t; q)| \leq e^{\omega t}$ so that $A(q) - \omega I$ is the generator of a contraction semigroup and is hence dissipative [22, p. 90], [30, pp. 14–17]. That is,

$$\langle A(q)z, z \rangle \leq \omega \langle z, z \rangle$$

for all $z \in \text{Dom}(A(q))$. From the definition of A^N it follows that for $z \in \mathcal{H}(q)$

$$(4.7) \quad \langle A^N(q)z, z \rangle_q = \langle A(q)P^N z, P^N z \rangle_q \leq \omega |P^N z|_q^2 \leq \omega |z|_q^2,$$

since $P^N(q)$ is the orthogonal projection of $\mathcal{H}(q)$ onto $\mathcal{H}^N(q)$. Hence we find $|T^N(t; q)| \leq e^{\omega t}$, as desired in the second conclusion of the theorem.

We make use of Proposition 4.1 to establish the convergence results of the theorem. We take for our discussions $\mathcal{B} = \mathcal{H}(\bar{q})$ and $\mathcal{B}^N = \mathcal{H}(q^N)$, which of course satisfy the conditions of (H2). The above arguments yield immediately that (ii) of Proposition 4.1 holds for our family of semigroups $T^N(t; q^N)$. Letting $\pi^N = \mathcal{I}^N$ be the canonical isomorphism from $\mathcal{H}(\bar{q})$ to $\mathcal{H}(q^N)$, we see immediately from the hypothesis $q^N \rightarrow \bar{q}$ and the definition of the norms in $\mathcal{H}(q)$ that $|\pi^N z|_{q^N} \rightarrow |z|_{\bar{q}}$, so that (i) is satisfied.

We define $\mathcal{D} = \bigcup_{N=1}^{\infty} \mathcal{H}^N(\bar{q})$ and have at once that $\mathcal{D} \subset \text{Dom}(A(\bar{q}))$ and $\tilde{\mathcal{D}} = \mathcal{H}(\bar{q})$. From the definition of $A(\bar{q})$, the fact that $\lambda I - A(\bar{q})$ is invertible for λ sufficiently large and that the $\Phi_j, \tilde{\Phi}_j$ are eigenfunctions of Δ , it is easily argued that $(\lambda I - A(\bar{q}))\mathcal{D} = \mathcal{D}$ so that $(\lambda - A(\bar{q}))\tilde{\mathcal{D}} = \mathcal{H}(\bar{q})$; hence (iii) is satisfied.

Finally, suppressing the notation $\pi^N = \mathcal{J}^N$ for the canonical isomorphism (see our comments in § 2), we have for each $z \in \mathcal{D}$ and $N = N(z)$ sufficiently large (then $P^N z = z$)

$$\begin{aligned} |A^N(q^N)z - A(\bar{q})z| &= |P^N A(q^N)P^N z - A(\bar{q})z| \\ &\leq |P^N A(q^N)z - P^N A(\bar{q})z| + |P^N A(\bar{q})z - A(\bar{q})z| \\ &\leq |A(q^N)z - A(\bar{q})z| + |(P^N - I)A(\bar{q})z| \end{aligned}$$

where all norms are $|\cdot|_{q^N}$. Since $q^N \rightarrow \bar{q}$ by hypothesis, the form of $A(q)$ yields that the first term $\rightarrow 0$ as $N \rightarrow \infty$. From the completeness of the $\{\beta_j^N\}$ in $\mathcal{H}(q^*)$ (see our remarks above), the equivalence of norms and thus the strong convergence $P^N \rightarrow I$ in any of the norms, we obtain that the second term $\rightarrow 0$. This completes the proof of Theorem 4.1.

We return to the abstract nonlinear equation (4.3) to consider conditions on f of (4.1) so that F of (4.3) will satisfy (H6) and hence the results of §§ 2 and 3 will be applicable. Define $Q_6 = \{q_6 \in R^m | q \in Q\}$ where $Q \subset R^{3m+3}$ is given for (4.1). We impose the following hypotheses on f .

(H6*) The nonlinear function $f: Q_6 \times [0, T] \times [0, 1] \times R \times R \rightarrow R$ satisfies:

(i) For each $(q_6, v, w) \in Q_6 \times R^2$, the map $(t, x) \rightarrow f(q_6, t, x, v, w)$ is measurable.

(ii) For each constant $M > 0$, there exists a function $\tilde{k}_1 = \tilde{k}_1(M)$ in $L_2(0, T)$ such that for all $(q_6, t, x) \in Q_6 \times [0, T] \times [0, 1]$ we have

$$|f(q_6, t, x, v_1, w) - f(q_6, t, x, v_2, w)| \leq \tilde{k}_1(t)|v_1 - v_2|$$

for all $(v_i, w) \in R^2$ with $|v_i| \leq M$.

(iii) There exists a function \tilde{k}_2 in $L_2(0, T)$ such that

$$|f(q_6, t, x, v, w_1) - f(q_6, t, x, v, w_2)| \leq \tilde{k}_2(t)|w_1 - w_2|$$

for all $(q_6, t, x) \in Q_6 \times [0, T] \times [0, 1]$, and $v, w_1, w_2 \in R$.

(iv) There exists a function \tilde{k}_3 in $L_2(0, T)$ such that

$$|f(q_6, t, x, v, 0)| \leq \tilde{k}_3(t)\{|v| + 1\}$$

for all $(q_6, t, x, v) \in Q_6 \times [0, T] \times [0, 1] \times R$.

(v) For each $(t, x, v, w) \in [0, T] \times [0, 1] \times R^2$, the map $q_6 \rightarrow f(q_6, t, x, v, w)$ is continuous.

Employing rather standard arguments and results from analysis (e.g., see [15, Lem. 16(b), p. 196] in connection with (i)), it is quite straightforward to verify under (H2) that (H6*) for f implies (H6) for F in the example under consideration. We can therefore appeal to Proposition 2.1 to guarantee existence of a unique mild solution of (4.3) for any $q \in Q$.

Summarizing, we have shown that under (HQ), (H4) and (H6*) the conditions (H1)–(H9) hold for the abstract formulation (4.3) of (4.1)–(IC₁)–(BC₁) when considering the modal approximation scheme

$$\begin{aligned} \frac{d}{dt} \begin{pmatrix} v^N(t) \\ w^N(t) \end{pmatrix} &= A^N(q) \begin{pmatrix} v^N(t) \\ w^N(t) \end{pmatrix} + P^N F(q, t, v^N(t), w^N(t)), \\ \begin{pmatrix} v^N(0) \\ w^N(0) \end{pmatrix} &= P^N \begin{pmatrix} \sum q_4^i \phi_i \\ \sum q_5^i \psi_i \end{pmatrix} \end{aligned} \quad (4.8)$$

in $\mathcal{H}^N(q)$. The convergence of Theorems 3.1 and 3.2 is thus assured and we may,

when an appropriate fit-to-data function is defined, appeal to Theorem 2.1 to obtain a solution of the associated identification problem for (4.1)–(IC₁)–(BC₁). For example, suppose we are given observations $\hat{y}(t_i) \in \mathbb{R}^l$, $i = 1, \dots, r$, for $(v(t_i, x_1), \dots, v(t_i, x_l))$ in (4.1) where $0 \leq t_1 < \dots < t_r \leq T$. Let $u(t; q) = (u_1(t, \cdot; q), u_2(t, \cdot; q))$ denote the unique mild solution of (4.3) where we observe that $u_1(t; q) \in H_0^1$ for each t, q , so that pointwise evaluation in $[0, 1]$ is a meaningful operation. Let \tilde{J} have the form given in (2.4) where now $\xi(t_i, q) \equiv \text{col}(u_1(t_i, x_1; q), \dots, u_1(t_i, x_l; q))$ and $C(t_i, q)$ is an $l \times l$ -matrix depending continuously on q in Q for each i . Then clearly this \tilde{J} satisfies the continuity requirements of Theorem 2.1. Furthermore, it is easily seen that the initial data in (4.3) depend continuously on q . For the modal approximations, recall that $P^N(q)$ is independent of q and finally note that the continuity of $q \rightarrow T^N(t; q)z$ follows directly from the forms of $A^N(q)$, Q^N , and K^N given explicitly in (2.10), (4.4) and (4.5). The conclusions of our deliberations for Example 4.1 may thus be stated:

THEOREM 4.2. *If (HQ), (H4) and (H6*) obtain, then the problem (ID^N) for (4.8) with \tilde{J} as defined above has, for each N , a solution $\hat{q}^N \in Q \subset \mathbb{R}^{3m+3}$. Letting $\{\hat{q}^{N_i}\}$ be any subsequence of $\{\hat{q}^N\}$ converging to $\hat{q} \in Q$, then \hat{q} is a solution of the problem (ID) for (4.3) and moreover for each $t \in [0, T]$,*

$$\|(v^{N_i}(t; \hat{q}^{N_i}), w^{N_i}(t; \hat{q}^{N_i})) - (u_1(t; \hat{q}), u_2(t; \hat{q}))\| \rightarrow 0$$

as $N_i \rightarrow \infty$ where (u_1, u_2) is the solution of (4.3) and the norm is that of $H_0^1 \times H^0$.

Remark 4.1. We remark that the dependence of the norm on q in the above treatment of hyperbolic systems is somewhat artificial. While one cannot effectively rescale the time variable to remove the q_1 -dependence in problems where sampling times (observations) are important, one can rescale the state variables (use $w(t) = 1/\sqrt{q_1}v_t$ in place of the variable used in (4.3)) to avoid use of a weighted norm. We chose not to do that in our computations for Example 4.1. A preliminary consideration leads one to conjecture that such a rescaling does not result in simplification from a numerical viewpoint.

Example 4.2. Parabolic equations I. For our second class of examples we discuss scalar parabolic equations

$$(4.9) \quad v_t = \frac{q_1}{k} (pv_x)_x + q_2 v + f(q_4^1, \dots, q_4^m, t, x, v)$$

for $t > 0$, $x \in [0, 1]$ with initial and boundary conditions

$$(IC) \quad v(0, x) = \sum_{i=1}^m q_3^i \phi_i(x), \quad 0 \leq x \leq 1,$$

$$(BC) \quad R_j v(t, \cdot) = 0, \quad j = 1, 2.$$

Here we assume that $\phi_i \in H^0$, $v(t, x)$ (or $v(t, x; q)$) is in \mathbb{R} , and $q = (q_1, q_2, q_3, q_4) \in Q$ where $Q \subset \mathbb{R}^{2m+2}$ and $q_j = (q_j^1, \dots, q_j^m)$ for $j = 3, 4$. The operators R_1, R_2 defining the boundary conditions have domain H^2 and are given for $\psi \in H^2$ by

$$(4.10) \quad R_j \psi = \alpha_{j1} \psi(0) + \alpha_{j2} \psi'(0) + \alpha_{j3} \psi(1) + \alpha_{j4} \psi'(1),$$

where $\alpha_{ji} \in \mathbb{R}$. We impose the following conditions on k and p in (4.9) and the α_{ji} :

(HP1) The functions p, p_x and k are continuous with $k(x) > 0$, $p(x) > 0$ for $0 \leq x \leq 1$.

(HP2) The matrix

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \end{pmatrix}$$

has rank 2 and we have $p(0)\{\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21}\} = p(1)\{\alpha_{13}\alpha_{24} - \alpha_{14}\alpha_{23}\}$.

We shall also assume that our parameter set Q satisfies hypothesis (HQ) already given above.

To apply the results of §§ 2 and 3 we again rewrite our system as an abstract evolution equation. Although we consider here only a scalar equation, the results we present generalize immediately to coupled systems of parabolic equations. Indeed, we present numerical results in § 6 for examples of such systems which are of interest in biological applications (see [4]), among others.

We define the general Sturm–Liouville operator $A(q)$ (our operator here is the negative of the one found in the literature cited below) in H^0 by $\text{Dom}(A(q)) = \{\psi \in H^2 | R_j \psi = 0, j = 1, 2\}$ and $A(q)\psi = k^{-1}(q_1 p \psi_x)_x + q_2 \psi$. Then (4.9)–(IC)–(BC) can be rewritten as

$$(4.11) \quad \begin{aligned} \dot{v}(t) &= A(q)v(t) + F(q, t, v(t)), \quad t > 0, \\ v(0) &= \sum_{i=1}^m q_3^i \phi_i, \end{aligned}$$

where $F(q, t, v(t)) = f(q, t, \cdot, v(t, \cdot))$ and the equation is taken in the state space $X(q) = \mathcal{H} \equiv (H^0, \langle \cdot, \cdot \rangle)$ with $\langle \phi, \psi \rangle = \int_0^1 \phi(x)\psi(x)k(x)dx$. We note that, unlike in Example 4.1, here the state space actually doesn't depend on q .

Spectral results for the operator $A(q)$ are readily found in the literature—e.g., see [28, p. 182], [20, p. 126]. The hypotheses (HP1), (HP2) imply that $A(q)$ is self-adjoint and its spectrum consists of a countable number of real eigenvalues $\{\lambda_j(q)\}_{j=1}^\infty$, each of multiplicity less than or equal to 2, and, moreover, these eigenvalues can be ordered so that $-\infty < \dots \leq \lambda_j \leq \lambda_{j-1} \leq \dots \leq \lambda_1 < \infty$. The eigenfunctions $\{\Psi_j\}_{j=1}^\infty$ corresponding to $A(q^*)$ where $q^* = (1, 0, \dots, 0)$ form a CONS in \mathcal{H} .

We further observe that the eigenvalues $\{\lambda_j(q)\}$ of $A(q)$ are bounded above uniformly in q on bounded subsets of Q . This is easily seen as follows: Let $\tilde{\lambda}_j$ be the eigenvalues of $A_0 = A(q^*)$ (i.e., $A_0\psi = k^{-1}(p\psi_x)_x$) with corresponding eigenfunctions Ψ_j . From our remarks above we have $\tilde{\lambda}_j \leq \hat{\omega}, j = 1, 2, \dots$, for some positive finite constant $\hat{\omega}$. The eigenvalues for $A(q)$ are then found to be $\lambda_j(q) = q_1 \tilde{\lambda}_j + q_2$ (with eigenfunction Ψ_j) so that we find $\lambda_j(q) \leq \omega$ on bounded subsets of Q , where ω is independent of q (but depends, of course, on the particular bounded subset of Q involved).

We define the approximating modal subspaces of $X(q) = \mathcal{H}$ by $\mathcal{H}^N = \text{span}\{\Psi_1, \dots, \Psi_N\}$ and let $P^N: \mathcal{H} \rightarrow \mathcal{H}^N$ denote the associated canonical orthogonal projections. As before, we determine the operator $A^N(q)$ and F^N by $A^N(q) = P^N A(q) P^N$ and $F^N = P^N F$. We have the following convergence results.

THEOREM 4.3. *Suppose that (HQ), (HP1) and (HP2) hold and $q^N, \bar{q} \in Q \subset \mathbb{R}^{2m+2}$ are such that $q^N \rightarrow \bar{q}$ as $N \rightarrow \infty$. Then $A(\bar{q})$ and $A^N(q^N)$ generate C_0 -semigroups $T(t; \bar{q})$ and $T^N(t; q^N)$ on \mathcal{H} and $|T^N(t; q^N)z - T(t; \bar{q})z| \rightarrow 0$ for $z \in \mathcal{H}$ with the convergence uniform in t on compact subsets of $[0, \infty)$. Furthermore, if (H4) obtains, then there exists a constant ω independent of N and q such that $|T(t; q)| \leq e^{\omega t}$ and $|T^N(t; q)| \leq e^{\omega t}$ for $t > 0, q \in Q$, and $N = 1, 2, \dots$.*

Proof. Let \tilde{Q} be any bounded subset of Q . Then our remarks above imply existence of $\tilde{\omega} = \omega(\tilde{Q})$ such that the self-adjoint operator $A(q), q \in \tilde{Q}$ has its spectrum lying in $(-\infty, \tilde{\omega})$. Hence (see [36, p. 349]) $A(q) - \tilde{\omega}I$ is dissipative, i.e., $\langle (A(q) - \tilde{\omega}I)z, z \rangle \leq 0$ for all $z \in \text{Dom}(A(q))$ and $q \in \tilde{Q}$. For $\lambda \notin \sigma(A(q))$, we have [28, p. 180] that $A(q) - \lambda I$ has compact resolvent so that in particular we have $(A(q) - \lambda I) \text{Dom}(A(q)) = \mathcal{H}$ for $\lambda > 0$ properly chosen. It follows immediately [30, p. 17], [2, p. 175], [22, p. 87] that $A(q) - \tilde{\omega}I$ is maximal dissipative and generates a C_0 -semigroup

of contractions on $\tilde{\mathcal{H}}$. Arguing as in (4.7), we have that $A^N(q) - \tilde{\omega}I$ is dissipative, uniform in $q \in \tilde{Q}$ and $N = 1, 2, \dots$ (it is maximal, being defined on all of $\tilde{\mathcal{H}}$ —see [22, p. 86]), and hence is also the generator of a C_0 -semigroup of contractions.

The above remarks obviously apply if we choose $\tilde{Q} = \{q^N\} \cup \{\bar{q}\}$ where $q^N \rightarrow \bar{q}$ or if $\tilde{Q} = Q$ where Q itself is compact, i.e., (H4) holds. To obtain the convergence results of the theorem, take $\tilde{Q} = \{q^N\} \cup \{\bar{q}\}$ and use Proposition 4.1. Here $\mathcal{B} = \mathcal{B}^N = \tilde{\mathcal{H}}$ and conditions (i) and (ii) of the proposition clearly are satisfied. Let $\mathcal{D} = \bigcup_{N=1}^{\infty} \tilde{\mathcal{H}}^N$ so that $\mathcal{D} \subset \text{Dom}(A(\bar{q}))$ and \mathcal{D} is dense in $\tilde{\mathcal{H}}$. From our remarks above, we have $A(\bar{q})\Psi_j = \lambda_j(\bar{q})\Psi_j$. For $\lambda > \max_j \lambda_j(\bar{q})$ we have $\lambda I - A(\bar{q})$ invertible where $(\lambda I - A(\bar{q}))\Psi_j = (\lambda - \lambda_j(\bar{q}))\Psi_j$, and it follows that $(\lambda I - A(\bar{q}))\mathcal{D} = \mathcal{D}$. Hence (iii) of Proposition 4.1 is satisfied. The arguments that (iv) is satisfied are exactly analogous to those used to complete the proof of Theorem 4.1, here the completeness of the $\{\Psi_j\}$ yielding $P^N \rightarrow I$ strongly in $\tilde{\mathcal{H}}$. We thus have established the results of Theorem 4.3.

We turn finally to conditions on f in (4.9) that will insure that F of (4.11) satisfies (H6). Let $Q_4 \equiv \{q_4 \in R^m | q \in Q\}$ where Q is a given subset of R^{2m+2} .

(H6**) The nonlinear function $f: Q_4 \times [0, T] \times [0, 1] \times R \rightarrow R$ satisfies:

- (i) For each $(q_4, v) \in Q_4 \times R$, the map $(t, x) \rightarrow f(q_4, t, x, v)$ is measurable.
- (ii) There exists a function \tilde{k}_1 in $L_2(0, T)$ such that $|f(q_4, t, x, v_1) - f(q_4, t, x, v_2)| \leq \tilde{k}_1(t)|v_1 - v_2|$ for all $(q_4, t, x, v_i) \in Q_4 \times [0, T] \times [0, 1] \times R$.
- (iii) There exists a function \tilde{k}_2 in $L_2([0, T] \times [0, 1])$ such that $|f(q_4, t, x, 0)| \leq \tilde{k}_2(t, x)$ for all $q_4 \in Q_4$.
- (iv) For each (t, x, v) in $[0, T] \times [0, 1] \times R$, the map $q_4 \rightarrow f(q_4, t, x, v)$ is continuous on Q_4 .

It is an easy exercise to verify that (H6**) for f implies (H6) for F (note that in this example the condition (H2) is trivially satisfied). We thus may invoke Theorem 3.1 for convergence of our modal approximations defined in $\tilde{\mathcal{H}}^N$ by the equation

$$(4.12) \quad \begin{aligned} \dot{v}^N(t) &= A^N(q)v^N(t) + P^N F(q, t, v^N(t)), \\ v^N(0) &= P^N \sum_{i=1}^m q_3^i \phi_i. \end{aligned}$$

For these parabolic systems defined in H^0 , the question of an appropriate cost functional is somewhat more delicate, since in general, point evaluation may not be meaningful. One possibility (a different approach will be discussed below) is to choose a cost functional J as in (2.3) where now $u = v$ is the mild solution of (4.11) and we might assume, for example, that $(x, q) \rightarrow Y(t_i, x, q)$ is continuous for each t_i . We are again in a position to employ Theorem 2.1 (taking into account the comments in Remark 2.2) to establish the following results for this example.

THEOREM 4.4. Suppose (HQ), (H4), (H6**) and (HP1), (HP2) are satisfied. Then the problem (ID^N) for (4.12) with J as in (2.3) has a solution $\hat{q}^N \in Q \subset R^{2m+2}$ for each $N = 1, 2, \dots$. If $\{\hat{q}^{N_i}\}$ is any subsequence of $\{\hat{q}^N\}$ converging to $\hat{q} \in Q$, then \hat{q} is a solution of (ID) of (4.11) and for each $t \in [0, T]$ we have $|v^{N_i}(t; \hat{q}^{N_i}) - v(t; \hat{q})| \rightarrow 0$, as $N_i \rightarrow \infty$, where v, v^{N_i} are the mild solutions of (4.11), (4.12) respectively and the norm is that of H^0 .

Example 4.3. Parabolic equations II. We consider again the parabolic equation (4.9) with initial condition (IC) and boundary condition (BC) but with slightly more restrictive conditions on the boundary operators R_j than those given in (4.10)–(HP2). We treat problems with boundary operators chosen from the standard Sturm–

Liouville operators (see [14, p. 145], [13, p. 291])

$$\begin{aligned}
 (4.13) \quad & \text{(A)} \quad R_1\psi = \psi(0) \quad \text{or} \quad R_2\psi = \psi(1); \\
 & \text{(B)} \quad R_1\psi = \psi'(0) \quad \text{or} \quad R_2\psi = \psi'(1); \\
 & \text{(C)} \quad R_1\psi = \psi'(0) - \sigma_1\psi(0) \quad \text{or} \quad R_2\psi = \psi'(1) + \sigma_2\psi(1), \quad \sigma_1, \sigma_2 > 0; \\
 & \text{(D)} \quad R_1\psi = \psi(0) - \psi(1) \quad \text{and} \quad R_2\psi = p(0)\psi'(0) - p(1)\psi'(1).
 \end{aligned}$$

In the sequel, in referring to the standard Sturm–Liouville boundary conditions (SLBC), we shall mean any combination of choices for R_1 and R_2 from (4.13) (A)–(C) or the periodic boundary conditions arising from the choice of R_1, R_2 given in (4.13) (D).

Our main objective is to discuss the use of the simple pointwise fit-to-data criteria

$$(4.14) \quad J_1(q, v(\cdot; q), \hat{y}) \equiv \sum_{i=1}^r |\hat{y}(t_i) - C(t_i, q)\xi(t_i, q)|^2$$

as defined in § 2 (see (2.4) and the discussions thereof). When treating (4.9) in \mathcal{H} as we did in Example 4.2, it is by no means clear that the associated ID problem with (4.14) is well posed. Indeed, one must first justify the pointwise evaluation (in the spatial coordinate) of v involved in defining ξ ; assuming that this can be done, one must entertain a second difficulty in that the convergence (of Theorems 3.1 and 4.3) of $v^N(t; q^N)$ to $v(t; \bar{q})$ is in the \mathcal{H} (i.e. H^0) norm. Since $J_1(\cdot, \cdot, \hat{y})$ is *not* continuous on $Q \times C(0, T; \mathcal{H})$, the results of Theorem 2.1 are not directly applicable.

We turn first to the difficulty raised by point evaluations in (4.14). In this regard we note that the mapping $w \rightarrow J_1(q, w, \hat{y})$ from $C(0, T; \mathcal{H}^N)$ to R , where J_1 is given in (4.14) and \mathcal{H}^N are the modal subspaces defined in Example 4.2, is well defined for each $N = 1, 2, \dots$, and, in particular, is well defined on solutions of (4.12). Hence the approximating problems (ID) ^{N} associated with J_1 are well posed in any event. Justification of point evaluation for (ID) with J_1 depends heavily on the smoothing properties of (4.9) or, equivalently, (4.11). Roughly speaking, for $t > 0$ the solution values $v(t; q)$ will be contained in $\text{Dom}(A(q))$ if only $t \rightarrow f(t) = F(q, t, v(t))$ is smooth enough [8, p. 192]. However, since we wish to avoid additional smoothness hypotheses, we choose a slightly more technical route to the same end.

Recalling the arguments for Theorem 4.3, we have that $A(q) - \omega I$ is self-adjoint and maximal dissipative where ω can be chosen independent of $q \in Q$. It therefore follows [9, p. 47] that $-A(q) + \omega I = \partial\phi^*(q)$ where $\partial\phi^*(q)$ denotes the subdifferential of the functional $\phi^*(q)$ is given by

$$(4.15) \quad \phi^*(q)(u) = \begin{cases} \frac{1}{2} |(\omega I - A(q))^{1/2} u|^2 & \text{if } u \in \text{Dom}(\omega I - A(q))^{1/2}, \\ +\infty & \text{otherwise.} \end{cases}$$

Here $(\omega I - A(q))^{1/2}$ denotes the square root, which by standard results [21, p. 281] is known to exist. Under assumption (H6**) for f , we have that (H6) holds for $F(q, t, w) = f(q, t, \cdot, w)$ and in particular (see (H6)(iii)) the mapping $t \rightarrow F(q, t, v(t; q))$ is in $L_2(0, T; \mathcal{H})$ for $t \rightarrow v(t; q)$ the solution of (4.9). Thus (see [9, p. 72] and note that a weak solution in the sense of [9] is in fact the unique mild solution in our sense, which is, moreover, also a strong solution) it follows that the map $t \rightarrow \phi^*(q)(v(t; q))$ is in $L_1(0, T; \mathcal{H})$ and is absolutely continuous on all subintervals $[\delta, T]$, $\delta > 0$, of $(0, T]$. In light of (4.15) this implies $v(t; q) \in \text{Dom}(\omega I - A(q))^{1/2}$ for all $t > 0$ and we observe that this holds for arbitrary initial data in \mathcal{H} . It remains to note that $\text{Dom}(\omega I - A(q))^{1/2} \subset H^1$. To see this one only needs make standard arguments using elementary

results on interpolating spaces—see [27, p. 9]. Briefly, by defining X (in the notation of [27]) as either H^1 with appropriate boundary conditions (in the event of B.C.'s using combinations from (A) and (B) or B.C.(D) from (4.13)) or H^1 with an appropriate energy norm (in the event elastic boundary conditions from (C) of (4.13) are involved) and $Y = \mathcal{H}$, one can make the identification $\Lambda = (\omega I - A(q))^{1/2}$ with $\text{Dom}(\Lambda) = X$.

We can therefore summarize by stating that conditions (HQ), (H4), (H6**), (HP1) in the case of (SLBC) are sufficient to allow point evaluation in J_1 .

If we wish to relax the continuity hypothesis on J in Theorem 2.1, it is necessary to consider in more detail the fit-to-data criterion $J^N(q) \equiv J(q, u^N(\cdot; q), \hat{y})$ of (ID^N). The following proposition will be useful in our deliberations; we state and prove it using the notation of Theorem 2.1.

PROPOSITION 4.2. *We suppose there exist maps \mathcal{J} and \mathcal{J}^N , $N = 1, 2, \dots$, from the compact set Q to R satisfying:*

- (i) *for each $N = 1, 2, \dots$, the map $q \rightarrow \mathcal{J}^N(q)$ is continuous on Q ;*
- (ii) *for any $q \in Q$ and any sequence $\{N_k\}$ with $N_k \rightarrow \infty$, there exists a subsequence N_{k_i} such that $\mathcal{J}^{N_{k_i}}(q) \rightarrow \mathcal{J}(q)$;*
- (iii) *for any $q^N \rightarrow \bar{q}$, there exists a subsequence $\{q^{N_k}\}$ such that $\mathcal{J}^{N_k}(q^{N_k}) \rightarrow \mathcal{J}(\bar{q})$.*

Then for each N there exists $q^N \in Q$ that minimizes \mathcal{J}^N over Q and, moreover, for any convergent subsequence $\{q^{N_k}\}$ of $\{q^N\}$ with $q^{N_k} \rightarrow \bar{q}$, \mathcal{J} is a minimum over Q at \bar{q} .

Proof. Let q^N denote the minimizer (whose existence is guaranteed by (i) and the compactness of Q) of \mathcal{J}^N so that $\mathcal{J}^N(q^N) \leq \mathcal{J}^N(q)$ for all $q \in Q$. Suppose $q^{N_k} \rightarrow \bar{q}$; then by (iii) (reindexing for convenience in notation) we have $\mathcal{J}^{m_i}(q^{m_i}) \rightarrow \mathcal{J}(\bar{q})$ for some subsequence $\{q^{m_i}\}$ of $\{q^N\}$. We use this to argue that $\mathcal{J}(\bar{q}) \leq \mathcal{J}(q)$ for $q \in Q$. If we assume that there exist $\tilde{q} \in Q$ and \tilde{j} such that $\mathcal{J}^{m_i}(q^{m_i}) > \mathcal{J}(\tilde{q})$ for $j \geq \tilde{j}$, then by (ii) there is yet another subsequence of m_i , denoted by m_{i_i} such that $\mathcal{J}^{m_{i_i}}(\tilde{q}) \rightarrow \mathcal{J}(\tilde{q})$. Hence, for sufficiently large i we have $\mathcal{J}^{m_{i_i}}(q^{m_{i_i}}) > \mathcal{J}^{m_{i_i}}(\tilde{q})$, which contradicts the definition of $q^{m_{i_i}}$ as a minimizer for $\mathcal{J}^{m_{i_i}}$.

To use Proposition 4.2 with our particular $J^N = \mathcal{J}^N$ defined using J_1 , we must consider special cases for which hypotheses (i)–(iii) of that proposition are readily verified. We discuss the homogeneous ($f \equiv 0$) version of (4.9) in \mathcal{H} in this regard.

PROPOSITION 4.3. *Consider (4.9) with $f \equiv 0$, initial conditions (IC) and boundary conditions (SLBC) from (4.13). Assume that (HQ), (H4) and (HP1) obtain. Then for the solutions $v^N(\cdot; q)$ of the approximating equations ((4.12) with $F \equiv 0$) we have $\{v^N(q) | q \in Q\}$ is a relatively compact subset of $C(t^*, T; C(0, 1; R))$ for each $t^* \in (0, T]$.*

Since the proof of this result consists of checking compactness criteria for a specific subset of $C(t^*, T; Y)$, where Y is a Banach space, we shall only sketch the ideas involved, leaving the details to the interested and determined reader.

First recall that equation (4.12) is in $\mathcal{H}^N = \text{span}\{\Psi_1, \dots, \Psi_N\}$ where $\{\Psi_i\}_{i=1}^\infty$ is the CONS of eigenfunctions of $A(q^*)$ with corresponding eigenvalues $\tilde{\lambda}_j = \lambda_j(q^*)$ (see the discussion immediately preceding Theorem 4.3). Solutions v^N of (4.12) have the representation (see (2.8)–(2.12) and the associated discussions) $v^N(t; q) = \sum_{i=1}^N w_i^N(t) \Psi_i$ where $w_i^N(t) = w_i^N(0) \exp\{(\tilde{\lambda}_i q_1 + q_2)t\}$. To verify relative compactness of the desired set, one can use the Ascoli theorem [25, p. 211] which requires equicontinuity of the set along with relative compactness of $\{v^N(t; q) | q \in Q, N = 1, 2, \dots\}$ in $C(0, 1; R)$ for each t in $[t^*, T]$. Use of the representation results along with standard estimates in Fourier analysis (Parseval, Cauchy–Schwarz, etc.) reduces the compactness criteria to the task of verifying

$$(a) \quad \sum_{i=1}^{\infty} e^{\tilde{\lambda}_i q_1^L t^*} < \infty,$$

$$(b) \quad \sum_{i=1}^{\infty} e^{2\tilde{\lambda}_i q_1^L t^*} |\Psi_i'|_{H^0} < \infty,$$

$$(c) \quad \sup \{ \|\Psi_i(x)\| | 0 \leq x \leq 1, i = 1, 2, \dots \} \text{ is finite.}$$

Here q_1^L is the lower bound for q_1 (see (HQ)).

From standard Fourier results [15, p. 1332] we have $\langle Af, \Psi_i \rangle \rightarrow 0$ for $f \in \text{Dom}(A(q^*))$ or that $\{A(q^*)\Psi_i\}$ is bounded in \mathcal{H} . Hence $\{\Psi_i\}$ is bounded in the graph $(A(q^*))$ norm and, consequently, after some calculations taking into account the (SLBC), in H^1 . It follows that (c) obtains and that $\{\Psi_i'\}$ is H^0 bounded. Thus if (a) holds, we immediately obtain (b). We have, therefore, reduced the compactness criteria to a requirement on the rate at which $\tilde{\lambda}_i \rightarrow -\infty$. But asymptotic estimates for the eigenvalues of Sturm–Liouville problems are readily available. For example, for (SLBC) from combinations of (A), (B), (C) of (4.13) we have $\tilde{\lambda}_i \sim -i^2$ [14, p. 153], which yields (a). For the periodic boundary conditions (D) a slight modification of the arguments in [12, § 8.3] yield the same rate estimates. In some higher-dimensional problems, similar rates for the eigenvalues are also available [13, § VI.4.1].

With the compactness results of Proposition 4.3 we are now able to give well-posedness results for the special case under consideration.

PROPOSITION 4.4. *Suppose (HQ), (H4), (HP1) obtain and $f \equiv 0$ in (4.9)–(IC)–(SLBC). Then (ID)^N with J_1 of (4.14) has a solution $\hat{q}^N \in Q$ for each $N = 1, 2, \dots$. Moreover, there exists a subsequence $\{\hat{q}^{N_k}\}$ such that $\hat{q}^{N_k} \rightarrow \hat{q}$ with \hat{q} a solution of (ID) with J_1 and $|v^{N_k}(t; \hat{q}^{N_k}) - v(t; \hat{q})|_C \rightarrow 0$ as $k \rightarrow \infty$, uniformly on compact subsets of $(0, T]$.*

Proof. We first verify that (i)–(iii) of Proposition 4.2 hold with $\mathcal{J}^N(q) \equiv J_1(q, v^N(\cdot; q), \hat{y})$ and $\mathcal{F}(q) \equiv J_1(q, v(\cdot; q), \hat{y})$. Since the matrices K^N and Q^N of (2.10) are given by $Q^N = I$ and $K_{ij}^N = \langle \Psi_i, A(q)\Psi_j \rangle_{\mathcal{H}} = (q_1 \tilde{\lambda}_j + q_2) \delta_{ij}$, continuity of the map

$$q \rightarrow \sum_{i=1}^r |C(t_i, q) \text{col}(v^N(t_i, x_1; q), \dots, v^N(t_i, x_l; q)) - \hat{y}(t_i)|^2$$

can be demonstrated by elementary arguments; this implies (i) and hence the existence of a minimizer \hat{q}^N for \mathcal{J}^N .

By compactness of Q there exists a convergent subsequence \hat{q}^{N_k} with $\hat{q}^{N_k} \rightarrow \hat{q} \in Q$. Employing Proposition 4.3 with $t^* < t_1$, t^* otherwise arbitrarily chosen in $(0, T)$, we find that for a subsequence of $\{\hat{q}^{N_k}\}$, again denoted by $\{\hat{q}^{N_k}\}$, the sequence $\{v^{N_k}(\cdot; \hat{q}^{N_k})\}$ is Cauchy in $C(t^*, T; C(0, 1; R))$. Since we already have $|v^{N_k}(t; \hat{q}^{N_k}) - v(t; \hat{q})|_{\mathcal{H}} \rightarrow 0$, uniformly in t on $[0, T]$, we obtain $|v^{N_k}(t; \hat{q}^{N_k}) - v(t; \hat{q})|_C \rightarrow 0$, uniformly in t on $[t^*, T]$. It follows that (ii) and (iii) of Proposition 4.2 hold and thus the conclusions of the present proposition are established.

The special cases considered above involved homogeneous ($f \equiv 0$) equations. One can also obtain results in the case $f \neq 0$ by formulating the (ID) problem in a “smoother” space than H^0 . We illustrate these ideas with an example involving a special case of (4.9). Consider

$$(4.16) \quad v_t = q_1 v_{xx} + q_2 v + f(q_4^1, \dots, q_4^m, t, \cdot, v)$$

with boundary conditions

$$(DBC) \quad v(t, 0) = v(t, 1) = 0$$

and initial condition (IC) with $\phi_i \in H_0^1$. We again rewrite (4.16)–(DBC)–(IC) as an abstract equation, but this time in the Hilbert space $\mathcal{H} = (H_0^1, \langle \cdot, \cdot \rangle_1)$ with inner product $\langle \phi, \psi \rangle_1 \equiv \int_0^1 \phi' \psi' dx$ (e.g., see [40, pp. 94, 105]) and define the operator $A(q)$ by $\text{Dom}(A(q)) = \{\phi | \phi \in H_0^1, \phi'' \in H_0^1\}$ and $A(q)\phi = q_1 \phi'' + q_2 \phi$. Assuming (HQ) and (H4),

it is easy to verify existence of an ω independent of $q \in Q$ such that $A(q) - \omega I$ is dissipative and symmetric in \mathcal{H} , that $\text{Dom}(A(q))$ is dense and $R(A(q) - \lambda I) = \mathcal{H}$ for some λ , chosen sufficiently large and independent of $q \in Q$. In particular, $A(q)$ is therefore self-adjoint [28, p. 97]. The approximating subspaces \mathcal{H}^N are defined by $\mathcal{H}^N = \text{span} \{\tilde{\Phi}_1, \dots, \tilde{\Phi}_N\}$ where $\tilde{\Phi}_j(x) = (\sqrt{2}/j\pi) \sin j\pi x$ as in Example 4.1. We recall that $\{\tilde{\Phi}_j\}_1^\infty$ is a CONS in \mathcal{H} . As before we let $P^N: \mathcal{H} \rightarrow \mathcal{H}^N$ denote the canonical orthogonal projections and $A^N(q) = P^N A(q) P^N$, $F^N(q, t, v) = P^N F(q, t, v)$ where $F: Q \times [0, T] \times \mathcal{H} \rightarrow \mathcal{H}$ is given by $F(q, t, v) = f(q, t, \cdot, v)$. Formulation in \mathcal{H} will require additional assumptions on f , to be detailed in (H6^{***}) below. We then have:

THEOREM 4.5. *Suppose that (HQ) holds and let $q^N, \bar{q} \in Q$ be such that $q^N \rightarrow \bar{q}$ as $N \rightarrow \infty$. Then $A(\bar{q})$ and $A^N(q^N)$ generate C_0 -semigroups $T(t; \bar{q})$ and $T^N(t; q^N)$ on \mathcal{H} and $|T^N(t; q^N)z - T(t; \bar{q})z|_{\mathcal{H}} \rightarrow 0$ for each $z \in \mathcal{H}$, with the limit uniform in t on compact subsets of $[0, \infty)$. Furthermore, if Q is compact, then there exists a constant ω independent of N and q , such that $|T(t; q)| \leq e^{\omega t}$ and $|T^N(t; q)| \leq e^{\omega t}$ for $t > 0$, $q \in Q$, and $N = 1, 2, \dots$.*

The proof of this theorem is quite similar to that of Theorem 4.3 and will therefore not be given here.

We shall simply list conditions (compare with (H6^{**})) on f that will insure that F satisfies (H6).

(H6^{***}) The nonlinear function $f: Q_4 \times [0, T] \times [0, 1] \times R \rightarrow R$ satisfies:

- (0) For each $v \in \mathcal{H}$, the map $x \rightarrow f(q, t, x, v(x))$ is in \mathcal{H} .
- (i) For each $(q_4, v) \in Q_4 \times R$, the maps $(t, x) \rightarrow f(q_4, t, x, v)$, $(t, x) \rightarrow f_x(q_4, t, x, v)$, and $(t, x) \rightarrow f_v(q_4, t, x, v)$ are measurable.
- (ii) There exists a function \tilde{k}_1 in $L_2(0, T)$ such that $|f_v(q_4, t, x, v)| \leq \tilde{k}_1(t)$ for all $(q_4, x, v) \in Q_4 \times [0, 1] \times R$; for each $M > 0$ there exists \tilde{k}_1^M in $L_2(0, 1)$ such that $|f_x(q_4, t, x, v_1) - f_x(q_4, t, x, v_2)| \leq \tilde{k}_1^M(t)|v_1 - v_2|$ and $|f_v(q_4, t, x, v_1) - f_v(q_4, t, x, v_2)| \leq \tilde{k}_1^M(t)|v_1 - v_2|$ for all $(q_4, x) \in Q_4 \times [0, 1]$ and v_1, v_2 with $|v_1| \leq M, |v_2| \leq M$.
- (iii) There exists \tilde{k}_2 in $L_2(0, T)$ such that $|f(q_4, t, x, 0)| \leq \tilde{k}_2(t)$ and $|f_x(q_4, t, x, v)| \leq \tilde{k}_2(t)\{1 + |v|\}$ for all $(q_4, x, v) \in Q_4 \times [0, 1] \times R$.
- (iv) For each (t, x, v) in $[0, T] \times [0, 1] \times R$, the maps $q_4 \rightarrow f(q_4, t, x, v)$, $q_4 \rightarrow f_x(q_4, t, x, v)$ and $q_4 \rightarrow f_v(q_4, t, x, v)$ are continuous on Q_4 .

The fit-to-data criterion (4.14) together with the state equation in \mathcal{H} are such that the map $(q, v) \rightarrow J_1(q, v, \hat{y})$ from $Q \times C(0, T; \mathcal{H}) \rightarrow R$ is continuous and, therefore, Theorems 3.1 and 2.1 are readily applicable. We leave a precise statement of the theorem for (4.16)–(IC)–(DBC) with (H6^{***}), analogous to Theorem 4.4, to the reader.

In concluding this discussion, we remark that numerical implementation of a scheme formulated as above in \mathcal{H} (the projections in defining the approximations in (2.8)–(2.12) are now in the H_0^1 inner product) is, of course, somewhat more tedious from a technical viewpoint than that for schemes such as those in Example 4.3 where the state space is H^0 .

Example 4.4. Diffusion-convection equations. For the final example of this section, we return to the setting of Example 4.2 and indicate how, in (4.9), one might include convection (or advection) terms that are independent of the Sturm–Liouville operator $(pu_x)_x$. To illustrate the ideas, we, for ease in exposition only, take a simple linear example (nonlinearities of the type discussed previously present no essential difficulties) involving only diffusion and convection terms. Consider then

$$(4.17) \quad v_t = q_1 v_{xx} + q_2 v_x$$

with initial conditions (IC) and the Dirichlet boundary conditions

$$(DBC) \quad v(t, 0) = v(t, 1) = 0.$$

As state space we choose H^0 with its usual inner product. We define the operator $A(q)$ with $\text{Dom}(A(q)) = H^2 \cap H_0^1$ by $A(q)\psi = q_1\psi'' + q_2\psi'$. We then have $A(q)$ is dissipative since (again we assume (HQ) holds) for $z \in \text{Dom}(A(q))$,

$$\begin{aligned} \langle A(q)z, z \rangle &= q_1 \langle z_{xx}, z \rangle + q_2 \langle z_x, z \rangle \\ &= -q_1 |z_x|^2 + q_2 \int_0^1 (z^2/2)_x dx \\ &= -q_1 |z_x|^2 \leq 0. \end{aligned}$$

Thus if we assume (H4), we find that $A(q)$ is dissipative uniformly in $q \in Q$.

In considering modal approximations, the question of existence of a complete set of eigenfunctions for the operator $A(q)$ arises naturally. Standard spectral results for nonself-adjoint operators allow one to answer this question in the affirmative. First, $A(q)$ is a relatively bounded perturbation of a discrete spectral operator and is itself a discrete spectral operator (see [15, Thm. XIX.4.16, p. 2347]—in this case the boundary conditions (DBC) are easily seen to satisfy the necessary regularity hypotheses—see [15, p. 2341–2344]). It follows that $\sigma(A(q))$ consists of point spectrum and that the eigenprojections $\{E_{\lambda_j}\}$ (see [15, p. 2292]) of the resolution of identity for $A(q)$ satisfy $\sum_{j=1}^N E_{\lambda_j} z \rightarrow z$ for $z \in H^0$ (see [15, Cor. XVIII.2.33, p. 2257], along with the properties of the projection operators—e.g., [15, Lem. XVIII.2.31, p. 2255]). One can easily argue for our example that the generalized eigenmanifolds are one-dimensional so that the eigenfunctions $\tilde{\Psi}_j(q) = \exp(-q_2 x/2q_1) \sin j\pi x$ corresponding to the eigenvalues $\lambda_j(q) = -j^2 \pi^2 q_1 - q_2^2/2q_1$ form a complete (but not orthogonal) set in H^0 .

We thus also have that $\lambda \in \rho(A(q))$ if $\lambda > 0$ so that $(A(q) - I) \text{Dom}(A(q)) = H^0$ for $\lambda > 0$ and hence $A(q)$ is maximal dissipative [22, p. 87], [30, p. 17]. The operator $A(q)$ generates a C_0 -semigroup $T(t; q)$ satisfying $|T(t; q)| \leq e^{\omega t}$ for $q \in Q$.

For a modal approximation scheme, it might be tempting at first thought to use the finite-dimensional subspaces $\tilde{X}^N(q) = \text{span}\{\tilde{\Psi}_1(q), \dots, \tilde{\Psi}_N(q)\}$, but of course this would prove rather difficult computationally in identification problems. Here we choose to use the basis elements $\Phi_j(x) = \sqrt{2} \sin j\pi x$ since we know $\{\Phi_j\}_1^\infty$ forms a CONS in H^0 and $\Phi_j \in \text{Dom}(A(q))$. We thus define $H^N = \text{span}\{\Phi_1, \dots, \Phi_N\}$ and remind the reader that “modal” is something of a misnomer for this scheme (actually, we took a similar approach in Example 4.1 in choosing basis elements corresponding to $q^* = (1, 0, \dots, 0)$ fixed).

As usual, we define $A^N(q) = P^N A(q) P^N$ where P^N are the orthogonal projectors $P^N z = \sum_{j=1}^N \langle z, \Phi_j \rangle \Phi_j$ onto H^N which converge strongly to the identity on H^0 .

To develop approximation results similar to those given in Theorems 4.3 and 4.4, the essential effort remaining in our example is to verify the stability and consistency hypotheses (ii), (iii), (iv) of Proposition 4.1 with $\mathcal{A} = A(\bar{q})$ and $\mathcal{A}^N = A^N(q^N)$ where $q^N \rightarrow \bar{q}$ in Q . Stability (i.e., (ii)) is immediate while consistency is slightly more delicate. A natural choice (the one we have used in previous examples) for the set \mathcal{D} is $\bigcup_{N=1}^\infty H^N$ since then (iv) is trivial to verify. However, it is not apparent to us that this choice of \mathcal{D} satisfies (iii) of Proposition 4.1. We choose instead $\mathcal{D} = \bigcup_{N=1}^\infty \tilde{X}^N(\bar{q})$, where $\tilde{X}^N(\bar{q})$ is defined above in terms of the true modes $\tilde{\Psi}_j(\bar{q})$ for

$A(\bar{q})$. Since $(\lambda I - A(\bar{q}))\tilde{\Psi}_j(\bar{q}) = (\lambda - \lambda_j)\tilde{\Psi}_j(\bar{q})$ and the set $\{\tilde{\Psi}_j(\bar{q})\}$ is complete, (iii) is easily established and it only remains to show $A^N(q^N)z \rightarrow A(\bar{q})z$ for $z \in \mathcal{D}$.

We first note that from

$$|A^N(q^N)z - A(\bar{q})z| \leq |P^N(A(q^N)P^N z - A(\bar{q})z)| + |P^N A(\bar{q})z - A(\bar{q})z|$$

and the strong convergence of P^N to I , it is sufficient to argue $A(q^N)P^N z \rightarrow A(\bar{q})z$ for $z \in \mathcal{D}$. It suffices to argue this latter convergence for $z = \psi_k \equiv \tilde{\Psi}_k(\bar{q})$ fixed. For this choice of z we find

$$\begin{aligned} A^N(q^N)P^N \psi_k &= A(q^N) \sum_{j=1}^N \langle \psi_k, \Phi_j \rangle \Phi_j = \sum_{j=1}^N \langle \psi_k, \Phi_j \rangle A(q^N) \Phi_j \\ &= \sum_{j=1}^N \langle \psi_k, \Phi_j \rangle \{q_1^N \Phi_j'' + q_2^N \Phi_j'\} \\ &= \sum_{j=1}^N \langle \psi_k, \Phi_j \rangle \{q_1^N (-j^2 \pi^2) \Phi_j + j\pi q_2^N \chi_j\} \\ &= q_1^N \sum_{j=1}^N \langle \psi_k, -j^2 \pi^2 \Phi_j \rangle \Phi_j + q_2^N \sum_{j=1}^N \langle \psi_k, j\pi \Phi_j \rangle \chi_j \\ &= q_1^N \sum \langle \psi_k, \Phi_j'' \rangle \Phi_j + q_2^N \sum \langle \psi_k, -\chi_j' \rangle \chi_j, \end{aligned}$$

where $\chi_j(x) = \sqrt{2} \cos j\pi x$ and we have used the facts that $\chi_j' = -j\pi \Phi_j$ and $\Phi_j'' = -j^2 \pi^2 \Phi_j$.

Integration by parts twice (using the fact that ψ_k and Φ_j are in H_0^1) yields

$$\langle \psi_k, \Phi_j'' \rangle = \langle \psi_k'', \Phi_j \rangle$$

while a single integration by parts establishes (again use $\psi_k \in H_0^1$)

$$\langle \psi_k, -\chi_j' \rangle = \langle \psi_k', \chi_j \rangle.$$

We thus have

$$A^N(q^N)P^N \psi_k = q_1^N \sum_{j=1}^N \langle \psi_k'', \Phi_j \rangle \Phi_j + q_2^N \sum_{j=1}^N \langle \psi_k', \chi_j \rangle \chi_j,$$

where both $\{\Phi_j\}$ and $\{\chi_j\}$ constitute CONS in H^0 . Since $q_1^N \rightarrow \bar{q}_1$, $q_2^N \rightarrow \bar{q}_2$ we thus obtain $A^N(q^N)P^N \psi_k \rightarrow \bar{q}_1 \psi_k'' + \bar{q}_2 \psi_k' = A(\bar{q})\psi_k$, as was desired.

The theorems for these approximation ideas for the (ID) and (ID^N) problems with (4.17)–(IC)–(DBC) are so similar in statement to Theorems 4.3 and 4.4 that we shall not prolong our discussion by giving a precise statement here.

With regard to implementation of this scheme, we point out that $A(q)$ does not leave the subspaces H^N invariant and hence the matrix representation of $A^N = P^N A P^N$ (see (2.8)–(2.10)) is not a simple diagonal matrix. However, for equations such as (4.17), it is rather easily seen that (2.10) is given by

$$[A^N(q)]_{ij} = \begin{cases} -q_1 i^2 \pi^2 & \text{for } i = j, \\ 0 & \text{for } i \neq j \text{ and } i+j \text{ even,} \\ 2jq_2 \left[\frac{2i}{i^2 - j^2} \right] & \text{for } i \neq j \text{ and } i+j \text{ odd.} \end{cases}$$

While this is not a simple matrix, it does allow a rather straightforward implementation of the scheme in actual computations.

5. A boundary control problem. The theory developed in §§ 2 and 3 can also be applied to optimal control problems governed by partial differential equations. We shall demonstrate this by means of a specific example. Consider as a special case of (4.1)–(IC₁)–(BC₁) the problem

$$(5.1) \quad \tilde{v}_{tt} = \tilde{v}_{xx}$$

for $t > 0$, $x \in [0, 1]$ with initial and boundary conditions

$$(IC_3) \quad \tilde{v}(0, x) = \phi(x), \quad \tilde{v}_t(0, x) = \psi(x),$$

$$(BC_3) \quad \tilde{v}(t, 0) = s_1(t), \quad \tilde{v}(t, 1) = s_2(t),$$

where the boundary control functions s_i are chosen in $\mathcal{S} = \{s | s \in H^2(0, T; \mathbf{R}), s(0) = s'(0) = 0\}$ and $(\phi, \psi) \in H_0^1 \times H^0$. The transformation $v(t, x) = \tilde{v}(t, x) - (1-x)s_1(t) - xs_2(t)$ applied to (5.1)–(IC₃)–(BC₃) leads to

$$(5.2) \quad v_{tt} = v_{xx} - (1-x)(s_1)_{tt} - x(s_2)_{tt},$$

$$(IC_4) \quad v(0, x) = \phi(x), \quad v_t(0, x) = \psi(x),$$

$$(BC_4) \quad v(t, 0) = v(t, 1) = 0.$$

We let $w = v_t$ and reformulate (5.2)–(IC₄)–(BC₄) as in Example 4.1 in the Hilbert space $\mathcal{H} = H_0^1 \times H^0$ with the usual inner product. This leads to a special case of (4.3) given by

$$(5.3) \quad \frac{d}{dt} \begin{pmatrix} v(t) \\ w(t) \end{pmatrix} = A \begin{pmatrix} v(t) \\ w(t) \end{pmatrix} + \gamma \sigma(t), \quad \begin{pmatrix} v(0) \\ w(0) \end{pmatrix} = \begin{pmatrix} \phi \\ \psi \end{pmatrix},$$

where

$$A = \begin{pmatrix} 0 & I \\ \Delta & 0 \end{pmatrix}, \quad \gamma = \gamma(x) = \begin{pmatrix} 0 & 0 \\ -x & x-1 \end{pmatrix}, \quad \sigma(t) = \text{col}((s_2)_{tt}, (s_1)_{tt}), \quad (\phi, \psi) \in \mathcal{H}.$$

The finite-dimensional subspaces $\mathcal{H}^N = \mathcal{H}^N(q^*)$, $q^* = (1, 0, \dots, 0)$, are chosen as in Example 4.1 and again we take $A^N = P^N A P^N$, where $P^N: \mathcal{H} \rightarrow \mathcal{H}^N$ denote the canonical orthogonal projections. For the convenience of the reader we repeat the family of approximating equations given by

$$(5.4) \quad \frac{d}{dt} \begin{pmatrix} v^N(t) \\ w^N(t) \end{pmatrix} = A^N \begin{pmatrix} v^N(t) \\ w^N(t) \end{pmatrix} + P^N \gamma \sigma(t), \quad \begin{pmatrix} v^N(0) \\ w^N(0) \end{pmatrix} = P^N \begin{pmatrix} \phi \\ \psi \end{pmatrix}.$$

In the light of Theorems 3.2 and 4.1 (with $q^N = q^*$ for all N), the solutions $(v^N(t; \sigma), w^N(t; \sigma))$ and $(v(t; \sigma), w(t; \sigma))$ of (5.4) and (5.3), respectively, satisfy $\lim_N (v^N(t; \sigma), w^N(t; \sigma)) = (v(t; \sigma), w(t; \sigma))$ in \mathcal{H} uniformly in $t \in [0, T]$, for any $T > 0$ and uniformly in σ , as σ varies in bounded subsets Σ of $L_2(0, T; \mathbf{R}^2)$. We shall also need the following technical result.

LEMMA 5.1. *The operator $\mathcal{F}: L_2(0, T; \mathbf{R}^2) \rightarrow C(0, T; \mathcal{H})$ defined via $(\mathcal{F}\sigma)(t) = \int_0^t T(t-\tau)\gamma\sigma(\tau) d\tau$ is compact.*

Defining the maps $(\mathcal{F}^N\sigma)(t) = \int_0^t T^N(t-\tau)P^N\gamma\sigma(\tau) d\tau$ and using the convergence of the semigroups $T^N(t)$ to $T(t)$, generated by A^N and A , respectively, it is easily seen that $\mathcal{F}^N \rightarrow \mathcal{F}$ in the operator norm topology. The proof is completed once one argues that the maps \mathcal{F}^N themselves are compact.

The above remarks provide the technical tools that can be applied to a variety of optimal control problems, one of which will be outlined below. For a more complete

discussion concerning approximation of optimal control problems for infinite-dimensional systems by sequences of optimization problems for finite-dimensional systems, we refer to [7] and the references given there.

We let Σ_{ad} be any fixed closed convex subset of $L_2(0, T; R^2)$ (possibly L_2 itself) and choose nonnegative continuous functions $g_0: \mathcal{H} \rightarrow R$, $g_1: C(0, T; \mathcal{H}) \rightarrow R$ and $g_2: L_2(0, T; R^2) \rightarrow R$. The functions g_i define the cost functional $\hat{J}: \Sigma_{\text{ad}} \rightarrow R$ by

$$(5.5) \quad \hat{J}(\sigma) = g_0((v(T; \sigma), w(T; \sigma))) + g_1((v(\cdot; \sigma), w(\cdot; \sigma))) + g_2(\sigma).$$

The optimal boundary value control problem associated with (5.1)–(IC₃)–(BC₃), (5.5) is then taken to be:

$$(\mathcal{P}) \quad \text{minimize } \hat{J} \text{ over } \Sigma_{\text{ad}}.$$

Suppose that a solution $\hat{\sigma} = \text{col}(\hat{\sigma}_1, \hat{\sigma}_2) \in \Sigma_{\text{ad}}$ of (\mathcal{P}) is found; this will uniquely determine boundary controls \hat{s}_1 and \hat{s}_2 in \mathcal{S} . The approximate optimization problems are defined by

$$(\mathcal{P}^N) \quad \text{minimize } \hat{J}^N \text{ over } \Sigma_{\text{ad}},$$

where

$$(5.6) \quad \hat{J}^N(\sigma) = g_0((v^N(T; \sigma), w^N(T; \sigma))) + g_1((v^N(\cdot; \sigma), w^N(\cdot; \sigma))) + g_2(\sigma).$$

Notice that (5.6) is an optimization problem associated with an ordinary differential equation. We shall need two standard assumptions on the functions g_i :

(G1) The continuous functions g_i are convex,

(G2) $g_2(\sigma) \rightarrow \infty$ as $|\sigma| \rightarrow \infty$.

As a consequence of (G1), the maps $\sigma \rightarrow \hat{J}(\sigma)$ and $\sigma \rightarrow \hat{J}^N(\sigma)$ are convex, which together with (G2) implies the existence of solutions of (\mathcal{P}) and (\mathcal{P}^N) ; these solutions are in addition unique if one of the g_i is strictly convex. Let σ^N denote a solution of (\mathcal{P}^N) . Then by (G2) it follows that $\{\sigma^N\}$ must be a bounded subset of $L_2(0, T; R^2)$. Indeed, the assumption $|\sigma^{N_k}| \rightarrow \infty$ for some subsequence $\{N_k\}$ contradicts the inequalities $g_2(\sigma^{N_k}) \leq \hat{J}^{N_k}(\sigma^{N_k}) \leq \hat{J}^{N_k}(\sigma) \rightarrow \hat{J}(\sigma) < \infty$ for all $\sigma \in L_2(0, T; R^2)$. The convergence of $\hat{J}^{N_k}(\sigma) \rightarrow \hat{J}(\sigma)$ is a consequence of $(v^N(t; \sigma), w^N(t; \sigma)) \rightarrow (v(t; \sigma), w(t; \sigma))$ uniformly in $t \in [0, T]$ and (G1). Since Σ_{ad} is convex and closed it is weakly closed so that there exists a weakly convergent subsequence $\{\sigma^{N_k}\}$ of $\{\sigma^N\}$ with σ^{N_k} converging weakly to some $\hat{\sigma} \in \Sigma_{\text{ad}}$. By Theorem 3.2, Lemma 5.1 and the estimates

$$\begin{aligned} & |(v^{N_k}(t; \sigma^{N_k}), w^{N_k}(t; \sigma^{N_k})) - (v(t; \hat{\sigma}), w(t; \hat{\sigma}))| \\ & \leq |(v^{N_k}(t; \sigma^{N_k}), w^{N_k}(t; \sigma^{N_k})) - (v(t; \sigma^{N_k}), w(t; \sigma^{N_k}))| \\ & \quad + |(v(t; \sigma^{N_k}), w(t; \sigma^{N_k})) - (v(t; \hat{\sigma}), w(t; \hat{\sigma}))|, \end{aligned}$$

it follows that

$$(v^{N_k}(t; \sigma^{N_k}), w^{N_k}(t; \sigma^{N_k})) \rightarrow (v(t; \hat{\sigma}), w(t; \hat{\sigma}))$$

in \mathcal{H} uniformly in $t \in [0, T]$. Since convexity and continuity together imply weak lower semicontinuity, we obtain the following string of inequalities:

$$\begin{aligned} \hat{J}(\hat{\sigma}) & \leq \liminf \{g_0((v^{N_k}(T; \sigma^{N_k}), w^{N_k}(T; \sigma^{N_k}))) \\ & \quad + g_1((v^{N_k}(\cdot; \sigma^{N_k}), w^{N_k}(\cdot; \sigma^{N_k}))) + g_2(\sigma^{N_k})\} \\ & = \liminf \hat{J}^{N_k}(\sigma^{N_k}) \leq \limsup \hat{J}^{N_k}(\sigma) = \hat{J}(\sigma) \end{aligned}$$

for every $\sigma \in \Sigma_{\text{ad}}$. This implies that $\hat{\sigma}$ is a solution of (\mathcal{P}) . Further standard arguments can be used to show that strict convexity of \hat{J} implies that σ^N itself converges weakly to the unique solution $\hat{\sigma}$ of (\mathcal{P}) and that σ^N converges strongly in $L_2(0, T; R^2)$ to $\hat{\sigma}$ if \hat{J} is strongly convex (see [7]). We finally summarize some of the above discussion in

THEOREM 5.1. *Suppose that (G1) and (G2) hold. If $\{\sigma^N\}$ denotes a sequence of solutions of (\mathcal{P}^N) , then there exists a subsequence $\{\sigma^{N_k}\}$ converging weakly to a solution $\hat{\sigma}$ of (\mathcal{P}) . Furthermore, $\hat{J}^{N_k}(\sigma^{N_k}) \rightarrow \hat{J}(\hat{\sigma})$ and $(v^{N_k}(t; \sigma^{N_k}), w^{N_k}(t; \sigma^{N_k})) \rightarrow (v(t; \hat{\sigma}), w(t; \hat{\sigma}))$ uniformly in $t \in [0, T]$. Moreover, $\hat{\sigma}$ determines uniquely boundary controls \hat{s}_1, \hat{s}_2 in \mathcal{S} .*

6. Numerical examples. In this section we briefly summarize our numerical findings when applying the modal approximation algorithms to some of the identification problems that were outlined in § 4. The aim here is to demonstrate the feasibility of the method for both hyperbolic and parabolic systems. As it turns out, modal approximations appear to be very well suited for hyperbolic systems, while for certain identification problems for parabolic systems we encountered some essential difficulties which one should take into consideration before attempting any practical use of the method for this type of equation. This will be explained further below. In developing our software packages, no great attention was given to maximizing efficiency in implementing the algorithms, or to minimizing computer time. The ordinary differential equations (see (2.12)) that arise were integrated by a simple fourth-order Runge–Kutta method (with step size varying from one example to the next from .0125 to .05), and the coefficients of the nonlinearity and the initial data (see (2.9) and (2.11)) were computed by employing Simpson's rule. The minimization problem arising in the identification problem for the approximating ordinary differential equations was numerically solved by using an IMSL package (ZXSSQ) employing the Levenberg–Marquardt algorithm. The “exact” solutions, which were used for the “data” \hat{y} in the fit-to-data criterion J , were generated by a Crank–Nicolson algorithm whenever solutions in closed form were not available. These solutions were generated with fixed known values of the parameters in the equations; these values will be referred to in the sequel as the “true” parameter values.

In the examples below, a fit-to-data criterion of the type (2.4) with $C(t_i, q) = I$ was used throughout. Further, we usually (except in Example 6.5) let $T = 2$ and chose t_i and x_j equally spaced in $[0, 2]$ and $[0, 1]$, respectively, so that $|t_i - t_{i-1}| = 0.2$ and $|x_j - x_{j-1}| = 0.25$.

Example 6.1. Here we return to Example 4.1 and consider the linear one-dimensional hyperbolic equation, which we repeat for convenience:

$$\begin{aligned} v_{tt} &= q_1 v_{xx} + q_2 v_t + q_3 v & \text{for } t > 0, \\ v(0, x) &= q_4 x(1 - x) & \text{for } 0 \leq x \leq 1, \\ v_t(0, x) &= q_5 \hat{\psi}(x) & \text{for } 0 \leq x \leq 1, \\ v(t, 0) &= v(t, 1) = 0, & \text{for } t > 0, \end{aligned}$$

where $\hat{\psi}(x) = 2x$ for $x \in [0, .5]$ and $\hat{\psi}(x) = 2(1 - x)$ for $x \in [.5, 1]$. Below, we present numerical results which are typical of those obtained in making numerous runs with this example. The startup values $q^{N,0}$ for use in the Levenberg–Marquardt algorithm are recorded in the bottom row of the tables, whereas the next-to-last row contains the true parameter values. The tables contain only those parameters on which a search was performed, whereas the remaining parameters were assumed known and therefore were held fixed at the true values.

In a first run (Table 1) we assumed that $q_3 = 1$ and $q_5 = 0$ were known and we were searching for $q = (q_1, q_2, q_4)$ with the true parameter values chosen to be $\hat{q} = (1.414, -1, 2)$ and with startup values $q^{N,0} = (1, 0, 1)$.

TABLE 1

N	\hat{q}_1^N	\hat{q}_2^N	\hat{q}_4^N
4	1.4103	-0.9961	1.9978
8	1.4126	-1.0021	2.0031
16	1.4129	-0.9968	2.0005
32	1.4129	-0.9992	2.0000
true value	1.414	-1	2
$q^{N,0}$	1	0	1

A feature of interest for these models used with the Levenberg–Marquardt algorithm is the range of convergence for the parameter q . For this specific example, we carried out computations keeping two of the parameters q_1, q_2, q_4 in addition to q_3, q_5 fixed while identifying one of q_1, q_2, q_4 . It was observed that for $q_1^{N,0}, q_2^{N,0}, q_4^{N,0}$ taken in the ranges $1 \leq q_1^{N,0} \leq 5, -5 \leq q_2^{N,0} \leq 0, .5 \leq q_4^{N,0} \leq 5$, respectively, rapid convergence was still obtained. (The actual range of convergence may be much larger; these were merely the ranges of values we tested.)

In a second run (Table 2), $q_1 = 1.414$ and $q_5 = 1$ were assumed to be known and the search was performed on $q = (q_2, q_3, q_4)$ with true value $\hat{q} = (-5, 4, 2)$ and startup values $q^{N,0} = (0, 0, 1)$.

TABLE 2

N	\hat{q}_2^N	\hat{q}_3^N	\hat{q}_4^N
4	-4.9930	4.0242	1.9997
8	-4.9803	4.0372	2.0025
true value	-5	4	2
$q^{N,0}$	0	0	1

Example 6.2. This is the nonlinear example (again a special case of Example 4.1):

$$\begin{aligned}
 v_{tt} &= q_1 v_{xx} + v + q_6(1+v)^{-1} & \text{for } t > 0, \\
 v(0, x) &= q_4 x(1-x) & \text{for } 0 \leq x \leq 1, \\
 v_t(0, x) &= 0 & \text{for } 0 \leq x \leq 1, \\
 v(t, 0) &= v(t, 1) = 0 & \text{for } t > 0.
 \end{aligned}$$

We chose the true model parameters $\hat{q} = (1.414, 2, 1)$, whereas the startup values were taken to be $q^{N,0} = (1, 1, 0)$. For the numerical solutions we refer to Table 3.

Example 6.3. This is another nonlinear equation of the form

$$\begin{aligned}
 v_{tt} &= q_1 v_{xx} + q_3 v + q_6 v^3 & \text{for } t > 0, \\
 v(0, x) &= q_4 x(1-x) & \text{for } 0 \leq x \leq 1, \\
 v_t(0, x) &= 0 & \text{for } 0 \leq x \leq 1, \\
 v(t, 0) &= v(t, 1) = 0 & \text{for } t > 0.
 \end{aligned}$$

TABLE 3

N	\hat{q}_1^N	\hat{q}_4^N	\hat{q}_6^N
4	1.4141	1.9990	0.9735
8	1.4148	2.0013	0.9790
16	1.4152	2.0009	0.9788
true value	1.414	2	1
$q^{N,0}$	1	1	0

Although this nonlinearity does not satisfy (H6*) of Example 4.1, we report in Table 4 on calculations carried out in the subspaces \mathcal{H}^N of \mathcal{H} . It is clear that the algorithm is converging in this case; indeed, one can relax the assumptions (H6*) so as to prove convergence for such nonlinearities; see the discussion involving (A6)(i), (ii) in § 3.

TABLE 4

N	\hat{q}_1^N	\hat{q}_2^N	\hat{q}_4^N	\hat{q}_6^N
4	1.3835	0.6774	1.9999	1.2368
8	1.4107	0.9875	2.0001	0.8973
16	1.4138	0.9983	2.0001	1.0016
true value	1.414	1	2	1
$q^{N,0}$	1	0	1	0

We turn now to some special cases of the parabolic problem (4.9)–(IC)–(BC). As pointed out earlier, parabolic equations can be more formidable than hyperbolic ones to handle via modal approximations. The difficulties are more than just a simple lack of identifiability (however this concept is defined), which, of course, can lead to substantial numerical embarrassment. Indeed, parabolic equations can lead to stiff systems of approximating ordinary differential equations. The reader can quickly convince himself of this fact by taking Dirichlet boundary conditions and putting $p \equiv k \equiv 1$ and $f = q_2 = 0$. In our computational pursuits we did not make an effort to use specific numerical methods for the stiff systems that can arise, but we simply decreased the step size in the Runge–Kutta algorithm to effect numerical stability. A perhaps more reasonable approach to avoiding these difficulties due to modal approximations is to take a completely different approximation scheme, say for example spline-based methods. We have pursued this idea successfully for parabolic systems and the details of those investigations will be reported elsewhere.

The fit-to-data criterion is chosen to be (4.14) with $C(t, q) = I$ in all the scalar examples below. In the two-dimensional system of Example 6.7, we used the obvious analogue of (4.14) for a coupled system of equations.

Example 6.4. We consider the linear equation

$$\begin{aligned} v_t &= q_1 v_{xx} + q_2 v && \text{for } t > 0, \\ v(0, x) &= \hat{\psi}(x) && \text{for } 0 \leq x \leq 1, \\ v(t, 0) &= v(t, 1) = 0 && \text{for } t > 0, \end{aligned}$$

where $\hat{\psi}$ is the “hat”-function defined in Example 6.1. The modal approximation scheme failed to identify q_1 and q_2 simultaneously, although it did identify each of them individually so long as the other one was fixed. This is by no means surprising; the exact solution of the above problem has the explicit representation $v(t, x) = \sum_{j=1}^{\infty} v_j(t) \sin j\pi x$, where $v_j(t) = v_j(0) \exp((q_2 - q_1(j\pi)^2)t)$ and $v_j(0), j = 1, 2, \dots$, are the Fourier coefficients of the sine series for $\hat{\psi}$. At time $t = 0$, the values of q_1, q_2 have

no influence and at the following times $t = .2, .4, \dots$ the decay of the exponential term in addition to the decreasing magnitude of the coefficients $v_j(0)$ cause successive terms to contribute less to the criterion J . Moreover, in this example, $v_{2j}(0) = 0$ for $j = 1, 2, \dots$, so that the criterion uses essentially only one mode to fit the model to the data. The results from the search on both parameters simultaneously are presented in Table 5.

TABLE 5

N	\hat{q}_1^N	\hat{q}_2^N
4	0.0236	0.2313
8	0.0335	0.3289
16	0.0336	0.3296
true value	0.1	0.986
$q^{N,0}$	0.25	0.25

Keeping $q_2 = .2$ fixed and searching for q_1 , when the true parameter value is $\hat{q}_1 = .1$ and $q_1^{N,0} = 0.25$, we find $\hat{q}_1^4 = 0.09999$. Similarly, when $q_1 = .1$ is kept fixed and q_2 is to be identified, with $\hat{q}_2 = .986$ and $q_2^{N,0} = .25$, the algorithm yields $\hat{q}_2^4 = .986004$.

Example 6.5. We next consider the nonlinear parabolic equation

$$\begin{aligned} v_t &= q_1 v_{xx} - q_4 v^3 & \text{for } t > 0, \\ v(0, x) &= q_3 \hat{\psi} & \text{for } 0 \leq x \leq 1, \\ v(t, 0) &= v(t, 1) = 0 & \text{for } t > 0. \end{aligned}$$

It is well known that for $q_4 \geq 0$ the above system has a global solution and we are therefore again in a situation where hypotheses (A6)(i), (ii) of § 3 must be used in any theoretical considerations of convergence. Our findings for this example are given in Table 6. Here we choose $T = 1$, while keeping the increments between the “data” points the same as before ($\Delta t = .2$ and $\Delta x = .25$).

TABLE 6

N	\hat{q}_1^N	\hat{q}_3^N	\hat{q}_4^N
2	.5030	4.8271	.8539
4	.4976	5.3001	1.2370
8	.4985	5.1774	1.1482
16	.5021	5.0845	1.0443
true value	.5	5	1
$q^{N,0}$.25	1	0

Example 6.6. We consider

$$\begin{aligned} v_t &= q_1 v_{xx} + 2q_4(1+v)^{-1} & \text{for } t > 0, \\ v(0, x) &= q_3 \hat{\psi} & \text{for } 0 \leq x \leq 1, \\ v(t, 0) &= v(t, 1) = 0 & \text{for } t > 0. \end{aligned}$$

In this and the next example we solved the approximating identification problem both without and with noise. When noise was added, then the Crank–Nicolson data which were used in the fit-to-data criterion were perturbed by Gaussian noise with zero mean and variance $\sigma^2 = .01$. It is accurate to report that in these two examples the scheme behaves in a stable manner under the influence of noise. In Tables 7 and 8 below, a blank indicates that this parameter was kept fixed at the true parameter value. The estimates obtained for this example are recorded in Table 7.

TABLE 7

	N	\hat{q}_1^N	\hat{q}_3^N	\hat{q}_4^N
no noise; search on q_3, q_4	4		5.2275	1.9254
	8		5.1374	1.9741
	16		5.0668	1.9845
noise $\sigma^2 = .01$; search on q_3, q_4	4		5.2362	1.9335
	8		5.1459	1.9813
	16		5.0749	1.9917
no noise; search on q_1, q_3, q_4	4	0.2472	5.2846	2.5221
	8	0.2301	5.1706	2.3584
	16	0.2150	5.0823	2.1746
noise $\sigma^2 = .01$; search on q_1, q_3, q_4	4	0.2443	5.2903	2.4941
	8	0.2272	5.1760	2.3301
	16	0.2120	5.0873	2.1442
	true value	0.2	5	2
	$q^{N,0}$	0.1	1	0

Example 6.7. As a final example we consider the coupled parabolic system

$$v_t = q_1 v_{xx} + 2(1 + q_4 w + v)^{-1},$$
$$w_t = q_2 w_{xx} \qquad \qquad \qquad \text{for } t > 0,$$
$$v(0, x) = \hat{\psi}(x) \qquad \qquad \qquad \text{for } 0 \leq x \leq 1,$$
$$w(0, x) = \hat{\psi}(x) \qquad \qquad \qquad \text{for } 0 \leq x \leq 1,$$
$$v(t, 0) = v(t, 1) = w(t, 0) = w(t, 1) = 0 \qquad \text{for } t > 0,$$

for which the numerical results are given in Table 8.

TABLE 8

	N	\hat{q}_1^N	\hat{q}_2^N	\hat{q}_4^N
no noise; search on q_1, q_4	4	.2011		1.9933
	8	.1987		2.0226
noise $\sigma^2 = .01$; search on q_1, q_4	4	.1982		2.1105
	8	.1960		2.1246
no noise; search on q_2, q_4	4		.0500	2.0514
	8		.0498	1.9551
noise $\sigma^2 = .01$; search on q_2, q_4	4		.0522	2.0349
	8		.0520	1.9385
no noise; search on q_1, q_2, q_4	4	.2011	.0500	1.9931
	8	.1988	.0499	2.0187
noise $\sigma^2 = .01$; search on q_1, q_2, q_4	4	.1973	.0522	2.1776
	8	.1949	.0521	2.2066
	true value	.2	.05	2
	$q^{N,0}$.1	.1	0

7. Concluding remarks. The contributions of the discussions in this paper are twofold. First, we have developed a general approximation framework in the context of semigroups that allows treatment of identification and control problems for a wide class of distributed parameter systems. Our second contribution is a development, using this framework, of “modal” approximation schemes in the spirit of those often proposed in the engineering literature. In addition to providing a solid theoretical foundation for such schemes, we have systematically tested them numerically on a number of examples and reported some of our findings. One result of these investigations has been our efforts to develop alternate schemes. The approximation framework can be used efficiently to develop a class of schemes based on spline or “finite-element” ideas. A discussion of our findings in this regard will appear in a manuscript that is currently in preparation.

We close with several further remarks that we have added in the final version of this paper, partly in response to referees’ queries and partly as a result of our subsequent efforts and findings in related investigations. First, as we noted in Remark 4.1, the generality of our theoretical framework (q dependent spaces, norms, etc.) is not essential to treat Example 4.1 or, indeed, any of the specific examples discussed above. However, if one considers a parabolic system as in Example 4.2 for which the function k is parameter dependent, the q dependence of the appropriate inner product is essential. In fact, such problems arise naturally in estimation questions for porous media problems, where one of the parameters to be estimated is the function k (the field porosity) itself. A treatment using the theoretical framework developed above in connection with cubic spline approximations is outlined for such problems in [42].

With regard to general spline approximation schemes, we have, since this paper was first written, completed certain efforts on spline-based techniques (referred to several times above) in the context of the theoretical framework given above. Second-order parabolic and hyperbolic systems [43], as well as higher-order equations arising in elasticity [44], have been treated and our findings have been most positive from both computational and theoretical viewpoints.

Acknowledgment. We would like to express our sincere appreciation to James Crowley who, in addition to developing and managing the software packages extremely well, was also helpful in various discussions on the practical use of the algorithms employed in connection with the numerical results of § 6.

REFERENCES

- [1] K. J. ÅSTROM AND P. EYKHOFF, *System identification—A survey*, Automatica, 7 (1971), pp. 123–162.
- [2] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer, New York, 1976.
- [3] A. BAMBERGER, G. CHAVENT AND P. LAILLY, *About the stability of the inverse problem in 1-D wave equations—Application to the interpolation of seismic problems*, Appl. Math. Opt., 5 (1979), pp. 1–47.
- [4] H. T. BANKS, *Modeling and Control in the Biomedical Sciences*, Lecture Notes in Biomathematics 6, Springer, New York, 1975.
- [5] ———, *Parameter identification techniques for physiological control systems*, in Lectures in Applied Mathematics Vol. 19, American Mathematical Society, Providence, RI, 1981, pp. 361–383.
- [6] H. T. BANKS AND F. KAPPEL, *Spline approximations for functional differential equations*, J. Differential Equations, 34 (1979), pp. 496–522.

- [7] H. T. BANKS AND A. MANITIUS, *Projection series for functional differential equations with applications to optimal control problems*, J. Differential Equations, 18 (1975), pp. 296–332.
- [8] V. BARBU, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff, Bucuresti, 1976.
- [9] H. BREZIS, *Operateurs maximaux monotones*, North-Holland, Amsterdam, 1973.
- [10] R. BURRIDGE, *The Gelfand–Levitan, the Marchenko, and the Gopinath–Sondhi integral equations of inverse scattering theory, regarded in the context of inverse impulse-response problems*, Wave Motion, 2 (1980), pp. 305–323.
- [11] G. CHAVENT, M. DUPUY AND P. LEMONNIER, *History matching by use of optimal theory*, Soc. Petroleum Eng. J., 15 (1975), pp. 74–86.
- [12] A. E. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
- [13] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, Vol. I, Springer, New York, 1953.
- [14] J. W. DETTMAN, *Mathematical Methods in Physics and Engineering*, McGraw-Hill, New York, 1962.
- [15] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, Vols. I–III, John Wiley, New York, 1957, 1963, 1971.
- [16] R. E. EWING AND R. S. FALK, *On some ill-posed problems arising in glaciology*, preprint.
- [17] G. R. GAVALAS AND J. H. SEINFELD, *Reservoirs with spatially varying properties: Estimation of volume from late transient pressure data*, Soc. Petroleum Eng. J., 13 (1973), pp. 335–342.
- [18] F. GRANT AND G. WEST, *Interpretation Theory in Applied Geophysics*, McGraw-Hill, New York, 1965.
- [19] A. HARAUX, *Nonlinear Evolution Equations—Global Behavior of Solutions*, Lecture Notes in Mathematics 841, Springer-Verlag, Berlin, 1981.
- [20] G. HELLWIG, *Differential Operators of Mathematical Physics*, Addison-Wesley, Reading, MA, 1964.
- [21] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
- [22] S. G. KREIN, *Linear Differential Equations in Banach Space*, Transl. of Math. Monographs 29, American Mathematical Society, Providence, RI, 1971.
- [23] C. S. KUBRUSLY, *Distributed parameter system identification, a survey*, Int. J. Control, 26 (1977), pp. 509–535.
- [24] T. G. KURTZ, *Extensions of Trotter's operator semigroup approximation theorem*, J. Functional Anal., 3 (1969), pp. 354–375.
- [25] S. LANG, *Analysis II*, Addison-Wesley, Reading, MA, 1969.
- [26] P. D. LAX AND R. S. PHILLIPS, *Scattering Theory*, Academic Press, New York, 1967.
- [27] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. I, Springer, New York, 1972.
- [28] R. H. MARTIN, *Nonlinear Operators and Differential Equations in Banach Spaces*, John Wiley & Sons, New York, 1976.
- [29] A. OKUBO, *Diffusion and Ecological Problems: Mathematical Models*, Biomathematics, 10, Springer, New York, 1980.
- [30] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Lecture Notes 10, University of Maryland, College Park, 1974.
- [31] M. P. POLIS AND R. E. GOODSON, *Parameter identification in distributed systems: A synthesizing overview*, Proc. IEEE, 64 (1976), pp. 43–61.
- [32] W. H. RAY, *Some recent applications of distributed parameter systems theory—A survey*, Automatica, 14 (1978), pp. 281–287.
- [33] F. RIESZ AND B. SZ.-NAGY, *Functional Analysis*, F. Ungar, New York, 1955.
- [34] G. A. ROSENBERG, W. T. KYNER AND E. ESTRADA, *Bulk flow of brain interstitial fluid under normal and hyperosmolar conditions*, Amer. J. Physiol., 238 (1980), pp. F42–F49.
- [35] A. RUBERTI, ed., *Distributed Parameters Systems: Modelling and Identification*, Lecture Notes in Control and Information Sciences 1, Springer, Berlin, 1978.
- [36] W. RUDIN, *Functional Analysis*, McGraw-Hill, New York, 1973.
- [37] P. C. SABATIER, ed., *Applied Inverse Problems*, Lecture Notes in Physics 85, Springer, Berlin, 1978.
- [38] P. C. SHAH, G. R. GABALAS AND J. H. SEINFELD, *Error analysis in history matching: The optimum level of parametrization*, Soc. Petroleum Eng. J., 18 (1978), pp. 219–228.
- [39] A. E. TAYLOR, *Introduction to Functional Analysis*, John Wiley, New York, 1967.
- [40] J. A. WALKER, *Dynamical Systems and Evolution Equations, Theory and Applications*, Plenum Press, New York, 1980.
- [41] K. YOSIDA, *Functional Analysis*, Springer, Berlin, 1971.

- [42] H. T. BANKS, *A survey of some problems and recent results for parameter estimation and optimal control in delay and distributed parameter systems*, LCDS Rep. 81-19, Brown Univ., July, 1981; to appear in Proc. Conf. on Volterra and Functional Differential Equations, Blacksburg, VA, June 10-13, 1981.
- [43] H. T. BANKS, J. M. CROWLEY AND K. KUNISCH, *Cubic spline approximation techniques for parameter estimation in distributed systems*, LCDS Rep. 81-25, Brown Univ., Nov., 1981.
- [44] H. T. BANKS AND J. M. CROWLEY, *Parameter estimation for distributed systems arising in elasticity*, LCDS Rep. 81-24, Brown Univ., Nov. 1981; to appear in Proc. Symposium on Engineering Sciences and Mechanics, National Cheng Kung University, Tainan, Taiwan, Dec. 28-31, 1981.

PROPERTIES OF MIN-MAX CONTROLLERS IN UNCERTAIN DYNAMICAL SYSTEMS*

S. GUTMAN[†] AND Z. PALMOR[†]

Abstract. We discuss properties of min-max controllers (previously obtained) for uncertain dynamical systems. In particular, we present a new model for switching action and re-examine asymptotic stability. We demonstrate the attractiveness of switching surfaces and comment on the insensitivity of solutions in those surfaces to the uncertainty.

1. Introduction. Since the pioneering work of Lyapunov concerning the stability of differential equations, many papers have been written about the so-called “second method” of Lyapunov. Roughly speaking, this method consists of constructing a positive definite test function \mathcal{V} which decreases along a trajectory. The analysis of differential equations using the “second method” is well known and can be found in the literature [1]–[3]. The synthesis problem, on the other hand, is much younger. Here, one starts with a positive definite function \mathcal{V} and then chooses a control function such that \mathcal{V} decreases along a trajectory. One way for achieving this is by minimizing \mathcal{V} with respect to the control variable as was done in [3]–[5]. This process leads to a discontinuous control action. A review of the synthesis approach up to the mid-1960s can be found in [6]. During the same period of time, a different motivation [7]–[9], called “variable structure systems” (VSS), led to discontinuous control action, which is similar to that in [6]. Besides the effect of discontinuous control on stability it was demonstrated, using both the VSS [7]–[9] and the “second method” [5], that this type of control may overcome uncertainties and disturbances for a limited class of systems. As noted in [10]–[12], discontinuous control action imposes difficulties on the existence of solutions to differential equations. This difficulty leads one [3], [5] to replace a relay, for example, by a saturated linear control which is continuous. A different approach to relay action is presented in [13]. Here, the notion of a solution to differential equations is extended via “generalized dynamical systems” (GDS) in which the discontinuous control function is modeled as a set-valued function. This approach allows one to use the “second method” directly in spite of the discontinuity. In [14]–[16] an attempt has been made to use GDS [17]–[21] for uncertain systems. By observing [16], one concludes that both stability of uncertain systems and the VSS concept can be unified in a systematic approach using the “second method” of Lyapunov for GDS. Since GDS approach models relay action (on the switching surface) as a set-valued function, one may argue that although this is acceptable as a mathematical model, its validity in real systems is not clear. Thus, it was suggested [16] the control action on the switching surface be replaced by a short time open loop control. Another approach was introduced in [5], [22], where the discontinuous function is replaced by linear saturated action. However, one may argue that approximating a relay action by a linear saturated control is not allowed, since in the neighborhood of the switching surface, the action is completely unpredictable. Thus in the present paper, we relax the approximation and allow a relay (or any other switching) action to take on any single-valued continuous function in the neighborhood of the switching surface such that the control action is continuous everywhere. In this regard, see also [23]. In addition, we investigate the attractivity of this surface and the insensitivity of trajectories (in that surface) to a system’s uncertainty (invariance property). By so

* Received by the editors June 9, 1981, and in revised form February 1, 1982.

[†] Department of Mechanical Engineering, Technion—Israel Institute of Technology, Haifa, Israel.

doing, we show that the “second method” is a natural approach to dynamical systems with uncertainty in three major directions: stability, attractiveness and sensitivity. Moreover, it turns out that the GDS model [16] is the limit case of our approximation. This might settle the arguments of [12], [24] and provides a complete model for the switching action.

2. Background. In the Euclidean space \mathcal{R}^n , the *distance* between two points $x, y \in \mathcal{R}^n$ is $d(x, y) = \|x - y\|$. The distance between a point x and a set $A \subset \mathcal{R}^n$ is $d(x, A) = \inf \{d(x, y) : y \in A\}$. The *separation* of $A \subset \mathcal{R}^n$ from $B \subset \mathcal{R}^n$ is $d^*(A, B) = \sup \{d(x, B) : x \in A\}$. The ε -neighborhood of a set $A \subset \mathcal{R}^n$ is $N_\varepsilon(A) = \{x \in \mathcal{R}^n : d(x, A) < \varepsilon\}$. A *set-valued* (multivalued) function $F(x, t)$ is a mapping from \mathcal{R}^{n+1} to the set of all nonempty compact subsets of \mathcal{R}^n . $F(\cdot)$ is *upper semicontinuous* (u.s.c.) at (x_0, t_0) , if given any $\varepsilon > 0$, there exists a $\delta > 0$ such that $(x, t) \in N_\delta(x_0, t_0) \Rightarrow d^*(F(x, t), F(x_0, t_0)) < \varepsilon$. We say that $F_n \rightarrow F$ *uniformly on compacta* if given any $\varepsilon > 0$ and any compact subset K of \mathcal{R}^{n+1} , there exists an N such that for $n \geq N$, $d^*(G_n, G) < \varepsilon$, where G_n, G are the graphs of the restrictions of F_n, F to K .

A *multivalued differential equation* is the relation

$$(1) \quad \dot{x} \in F(x, t).$$

A function $x(\cdot)$ is a *solution* of (1) in GDS sense if it is absolutely continuous and if for almost all t

$$\frac{dx(t)}{dt} \in F(x(t), t).$$

The following is well known and is due to Marchaud and Zaremba.

THEOREM 1. [16], [21]. *Let $F(x, t) \subset \mathcal{R}^n$, defined on some closed neighborhood \bar{N} of (x_0, t_0) , be compact and convex, and let $F(\cdot)$ be u.s.c. Then there exists in \bar{N} at least one solution to (1) passing through (x_0, t_0) .*

The relation between multivalued and ordinary differential equations is established in the following:

THEOREM 2. [25]. *Let $F(x, t)$ defined on some open set D be convex, and let $F(\cdot)$ be u.s.c. in D . Let the continuous single-valued function $f(\cdot)$ approximate the set-valued function $F(\cdot)$ in the sense that for the sequence $\{f_i\}$ of continuous single-valued functions, $f_n \rightarrow F$ uniformly on compacta. Then the solutions $x_n(\cdot) \rightarrow x(\cdot)$ uniformly on any finite time interval.*

Remark 1. Theorem 2 is presented here informally. For more details, see [25].

Next, consider the following uncertain dynamical system [16]:

$$(2) \quad \begin{aligned} \dot{x} &= f(x, t) + \Delta f(x, t, v) + B(x, t)u + \Delta B(x, t, v)u + C(x, t)w, \\ x(t_0) &= x_0, \quad (v, \nu, w) \in \Omega, \end{aligned}$$

where $x \in \mathcal{R}^n$ is the state of the system, $B \in \mathcal{R}^{n \times m}$, $C \in \mathcal{R}^{n \times r}$, $u \in \mathcal{R}^m$ is the control vector, (v, ν, w) is the uncertainty triple (defined on $\mathcal{R}^n \times \mathcal{R}^1$), $f(\cdot)$, $\Delta f(\cdot)$, $B(\cdot)$, $\Delta B(\cdot)$ are continuous in all their arguments, and $\Omega \subset \mathcal{R}^n \times \mathcal{R}^1$ is the uncertainty set.

We suppose that no statistical information is known for the uncertainty behavior; we thus refer to the uncertainty as deterministic. The *objective* is to select u such that the uncertain system (2) has certain properties (such as asymptotic stability). In [16] we have required the following:

Assumption 1 (matching conditions). For all $(x, t) \in \mathcal{R}^n \times \mathcal{R}^1$ there exist a continuous vector function $h(x, t, v) \in \mathcal{R}^m$ and continuous matrix functions $D(x, t) \in$

$\mathcal{R}^{m \times r}$, $E(x, t, \nu) \in \mathcal{R}^{m \times m}$, such that: a) $\Delta f(x, t, \nu) = B(x, t)h(x, t, \nu)$, b) $\Delta B(x, t, \nu) = B(x, t)E(x, t, \nu)$, and c) $C(x, t) = B(x, t)D(x, t)$.

Remark 2. The matching conditions are properties of the system's structure only. One finds similar conditions in model-following control [26] and in adaptive control [27].

Following Assumption 1, system (2) takes the form

$$(3) \quad \dot{x} = f(x, t) + B(x, t)[(I + E(x, t, \nu))u + \eta],$$

where $\eta = h(x, t, \nu) + D(x, t)w$. Note that for $\Delta B \equiv 0$,

$$(3a) \quad \dot{x} = f(x, t) + B(x, t)(u + \eta).$$

The objective is to choose u depending on (x, t) such that for any η (in some class) the closed loop system has the following properties: (i) asymptotic stability, (ii) attractivity, and (iii) insensitivity. These will be discussed in the sequel.

3. Asymptotic stability [16]. We say that a \mathcal{C}^1 function $\mathcal{V}(\cdot): \mathcal{R}^n \times \mathcal{R}^1 \rightarrow \mathcal{R}^1$ is *positive definite* (p.d.), if a) $\mathcal{V}(0, t) = 0$ for all $t \geq 0$, and b) there exist continuous, increasing scalar functions $\gamma(\cdot), \beta(\cdot)$ with $\gamma(0) = \beta(0) = 0$, such that for all $(x, t) \in \mathcal{R}^n \times \mathcal{R}^1$, $\gamma(\|x\|) \leq \mathcal{V}(x, t) \leq \beta(\|x\|)$. We say that $\mathcal{W}(\cdot)$ is *negative definite* (n.d.) if $-\mathcal{W}(\cdot)$ is p.d. Following [16] we make

Assumption 2. The free system (without control and uncertainty) is asymptotically stable in the sense of Lyapunov. (If this is not the case we first stabilize it.) Thus, there exists a $\mathcal{V}(\cdot)$ such that

- (i) $\mathcal{V}(x, t)$ is p.d.,
- (ii) $\gamma(\|x\|) \rightarrow \infty$ as $\|x\| \rightarrow \infty$,
- (iii) $\mathcal{W}_0(x, t) = \frac{\partial \mathcal{V}}{\partial t} + \nabla_x \mathcal{V} \cdot f$ is n.d.

Next, we choose the nominal $\mathcal{V}(\cdot)$ for system (3a). Then,

$$(4) \quad \mathcal{W}(t) = \frac{\partial \mathcal{V}}{\partial t} + \nabla_x \mathcal{V} \cdot \dot{x} = \mathcal{W}_0(t) + \alpha'(x, t)(u + \eta),$$

where $\mathcal{W}_0(t) = \mathcal{W}_0(x(t), t)$ and

$$(5) \quad \alpha'(x, t) = \nabla_x \mathcal{V}(x, t)B(x, t).$$

If for all $(x, t) \in \mathcal{R}^n \times \mathcal{R}^1$,

$$(6) \quad \min_{u \in U} \max_{\eta \in V} \alpha'(u + \eta) \leq 0,$$

then, for a proper choice of u , $\mathcal{V}(\cdot)$ decreases along a solution $x(\cdot)$ of (3a) for every value of η . Now let

$$(7) \quad U = V = \{\eta \in \mathcal{R}^m: \|\eta\| \leq \rho(x, t)\}.$$

Then (6) is satisfied with $p(\cdot): \mathcal{R}^n \times \mathcal{R}^1 \rightarrow \mathcal{R}^m$ such that $u(t) = p(x(t), t)$, and

$$(8) \quad p(x, t) = -\rho(x, t) \frac{\alpha(x, t)}{\|\alpha(x, t)\|}.$$

To apply Theorem 1, we rewrite (8) as follows:

$$(9) \quad p^*(x, t) = \begin{cases} -\rho(x, t) \frac{\alpha(x, t)}{\|\alpha(x, t)\|} & \forall (x, t) \notin \mathcal{S}, \\ \{u \in \mathcal{R}^m: \|u\| \leq \rho(x, t)\} & \forall (x, t) \in \mathcal{S}, \end{cases}$$

where

$$(10) \quad \mathcal{S} = \{(x, t): \alpha(x, t) = 0\}.$$

Because of the fact that $p^*(\cdot)$ in (9) is u.s.c. and convex, Theorem 1 implies that (3a) with (9) has at least one solution in the neighborhood of (x_0, t_0) . This solution can be continued, since there is no finite escape time ($\mathcal{V}(\cdot)$ decreases and $\gamma(\|x\|) \rightarrow \infty$ as $\|x\| \rightarrow \infty$). As a consequence, we have

THEOREM 3. [16]. *Consider the uncertain system (3a), (7), and suppose that Assumption 2 holds. Then the feedback controller given in (9) assures uniform asymptotic stability in the large of $x = \{0\}$ for all admissible uncertainties.*

Remark 3. If $\Delta B \neq 0$, that is, we consider (3), the controller given in (9) remains the same with the exception that $\rho(x, t)$ is replaced by $\rho(x, t)/(1 + \underline{\lambda})$, where $\underline{\lambda} = \min_v \{\lambda_{\min}[\frac{1}{2}(E + E')]\}$. In this case, $1 + \underline{\lambda} > 0$ must hold. Note that $\lambda_{\min}[\frac{1}{2}(E + E')]$ is evaluated for each (x, t) . Here, $\lambda_{\min}[\frac{1}{2}(E + E')]$ is the smallest eigenvalue of the symmetric matrix $\frac{1}{2}(E + E')$.

4. Models for discontinuous control. In the previous section we have presented a *set-valued controller* (9) which is a mathematical model for the discontinuous function (7). This model allows one to use the “second method” of Lyapunov directly without tedious arguments for proving existence of solutions.

A different approach uses the *approximated controller*. The need of such a controller is justified on the ground that set-valued controllers do not exist in reality. Moreover, one cannot write an accurate model for a discontinuous controller in the neighborhood of the switching surface. We thus say that *in a sufficiently small neighborhood of the switching surface, the controller takes on any continuous function (from the admissible set) such that the overall control function is continuous.*

Fortunately, Theorem 2 states that if the controller $p^*(\cdot)$ is u.s.c. and convex, the desired approximation is allowed. In that case, the approximated solution lies in an arbitrarily small closed neighborhood of any set-valued one. In particular, consider the *approximated controller*

$$(11) \quad \hat{p}(x, t) = \begin{cases} -\rho(x, t) \frac{\alpha(x, t)}{\|\alpha(x, t)\|} & \forall (x, t) \notin \mathcal{N}_\varepsilon(\mathcal{S}), \\ \left(\begin{array}{l} \text{any function such that} \\ \hat{p}(x, t) \in U, \text{ and } \hat{p}(\cdot) \text{ is} \\ \text{continuous everywhere} \end{array} \right) & \forall (x, t) \in \mathcal{N}_\varepsilon(\mathcal{S}). \end{cases}$$

We now state:

THEOREM 4. *Consider the uncertain system (3a), (7) and suppose Assumption 2 holds. Then, given $x_0 \in \mathcal{R}^n$ and any compact time interval, there exists $\varepsilon > 0$ such that the solution $x_\varepsilon(\cdot)$ generated by $\hat{p}(\cdot)$ given by (11) is arbitrarily close to the ideal solution $x(\cdot)$ generated by $p^*(\cdot)$.*

Remark 4. Recall that $\hat{p}(\cdot)$ given in (11) is an approximation to $p^*(\cdot)$ given in (9). Since a *real* controller cannot execute (9) on \mathcal{S} , but agrees with *some* continuous function in the neighborhood $\mathcal{N}_\varepsilon(\mathcal{S})$, the controller $\hat{p}(\cdot)$ is a suitable model for $p^*(\cdot)$.

That is to say, one has to specify $p^*(\cdot)$ outside a small neighborhood of \mathcal{S} , and the physics takes care of the rest.

Proof. Consider a point (x_0, t_0, u_0) in $N_\varepsilon(\mathcal{S}) \times \mathcal{U}(x, t) \subset \mathbb{R}^{n+1} \times \mathbb{R}^m$. Let $d[(x_0, t_0, u_0), \Gamma] = \|(x_0, t_0, u_0) - s_0\|$, where $\Gamma = \{(x, t, u) : (x, t) \in \mathcal{S}\}$, and let $L = [(x_0, t_0, u_0), s_0]$ be the line from (x_0, t_0, u_0) to s_0 . By construction, its length is equal to or less than ε . See also Fig. 1. Thus, (i) $L \cap G^* = \emptyset \Rightarrow s_0 \in \mathcal{U}(\mathcal{S}) \Rightarrow$

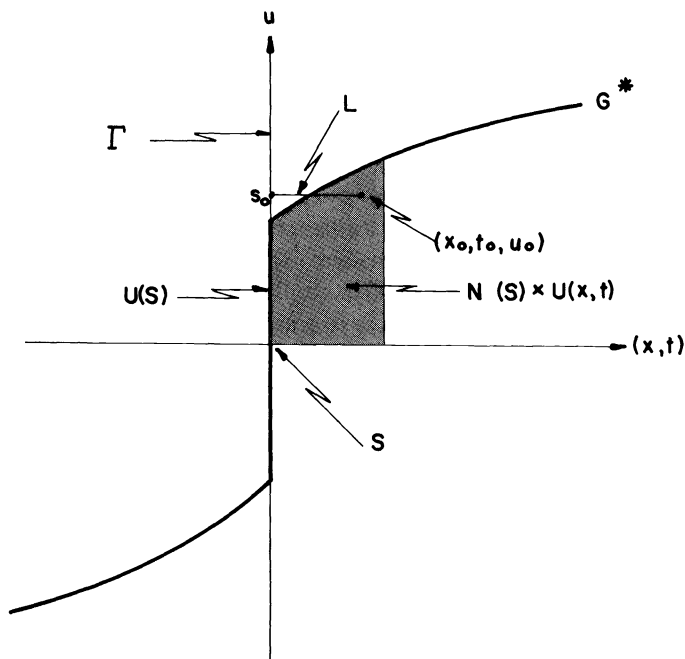


FIG. 1. Proof of Theorem 4.

$d[(x_0, t_0, u_0), \mathcal{U}(\mathcal{S})] < \varepsilon \Rightarrow d[(x_0, t_0, u_0), G^*] < \varepsilon$, where G^* is the graph of $p^*(\cdot)$, and (ii) $L \cap G^* \neq \emptyset \Rightarrow d[(x_0, t_0, u_0), G^*] < \varepsilon$, since $L \cap G^* \in \text{int}[L]$. Next note that in $N_\varepsilon(\mathcal{S})$, the graph \hat{G} of $\hat{p}(\cdot)$ lies in $N_\varepsilon(\mathcal{S}) \times \mathcal{U}(x, t)$. Thus, consider the sequence $\varepsilon_1 > \varepsilon_2 > \dots > \varepsilon_i \dots$, $\lim_{i \rightarrow \infty} \varepsilon_i = 0$. It is clear that the continuous single-valued function $\hat{p}(\cdot)$ approximates the set-valued function $p^*(\cdot)$ in the sense that $\hat{G}_n \rightarrow G^*$ uniformly on compacta. This last fact allows the direct application of Theorem 2, which concludes the proof.

5. The attractiveness of \mathcal{S} . In order to make results simple, we shall discuss uncertain dynamical systems which are nominally linear. That is, equation (2) takes the form

$$(12) \quad \dot{x} = (A + \Delta A(v))x + (B + \Delta B(v))u + Cw.$$

The matching conditions are

$$a) \Delta A(v) = BH(v), \quad b) \Delta B(v) = BE(v), \quad c) C = BD,$$

where H, E, D are matrices of proper dimensions, and B is of full rank.

Equation (3) takes the form

$$(13) \quad \dot{x} = Ax + B[(I + E(v))u + \eta],$$

where

$$(14) \quad \eta = H(v)x + Dw.$$

Upon choosing

$$(15) \quad \mathcal{V}(x, t) = x'Px,$$

where P satisfies Lyapunov's equation

$$(16) \quad PA + A'P = -Q, \quad Q \text{ is p.d.,}$$

the switching surface \mathcal{S} becomes

$$(17) \quad \mathcal{S} = \{(x, t): B'Px = 0\}$$

and the control function (8) becomes

$$(18) \quad p(x, t) = -\rho(x, t) \frac{B'Px}{\|B'Px\|},$$

where

$$(19) \quad \rho(x, t) = \max_v \|H(v)x\| + \max_w \|Dw\|.$$

We now investigate the attractiveness of the hyperplane \mathcal{S} defined in (17). We first need

DEFINITION 1. A surface $\mathcal{S} \subset \mathcal{R}^{n+1}$ is *asymptotically attractive* if a) any solution starting in it remains there, and b) any solution starting outside the surface tends to it at least asymptotically.

Let $\mathcal{V}(\cdot): \mathcal{R}^n \rightarrow \mathcal{R}^1$ be a distance of a point in the state space to \mathcal{S} such that

$$(20) \quad \mathcal{V}(x) = \|\alpha(x)\|^2$$

and consider the rate of change of $\mathcal{V}(\cdot)$ along a solution $x(\cdot)$ of (13). Then

$$(21) \quad \mathcal{W}(t) = \nabla_x \mathcal{V} \cdot \dot{x} \equiv 0 \quad \text{on } \mathcal{S},$$

and for all $x \notin \mathcal{S}$,

$$(22) \quad \begin{aligned} \mathcal{W}(t) &= 2\alpha' B' P \{Ax + B[(I + E(v))u + \eta]\} \\ &= 2\alpha' B' P Ax + 2\alpha' (B' PB)[I + E(v)]u + 2\alpha' (B' PB)\eta. \end{aligned}$$

Let u be given as in (8) or (18); that is,

$$(23) \quad u = -\bar{\rho}(x, t) \frac{\alpha(x, t)}{\|\alpha(x, t)\|},$$

where $\bar{\rho}(x, t)$ is to be determined. Then

$$(24) \quad \mathcal{W}(t) \leq -2[\bar{\Delta}\rho - \|B'PAx\|]\|\alpha\|$$

where

$$(25) \quad \begin{aligned} \text{(i)} \quad & \bar{\Delta}\rho = \bar{\rho}(\lambda + \lambda_{\min}(B'PB)) - \rho\lambda_{\max}(B'PB), \\ \text{(ii)} \quad & \lambda = \min_v \{\lambda_{\min}[\tfrac{1}{2}(B'PBE(v) + E'(v)B'PB)]\}, \\ \text{(iii)} \quad & \rho = \max_v \|F(v)x\| + \max_w \|Dw\|. \end{aligned}$$

Conditions necessary for the above to be meaningful are

$$(26) \quad \begin{aligned} \text{(iv)} \quad & \lambda + \lambda_{\min}(B'PB) > 0, \\ \text{(v)} \quad & \bar{\rho} \geq [\lambda_{\min}(B'PB) + \lambda]^{-1} \lambda_{\max}(B'PB)\rho. \end{aligned}$$

Define

$$(27) \quad \Gamma = \{x : \|B'PAx\| < \bar{\Delta}\rho\}$$

and

$$(28) \quad \tilde{P} = B'PB \text{ is p.d.,} \quad \hat{P} = PBB'P \text{ is p.s.d.,} \quad -\hat{Q} = \hat{P}A + A'\hat{P}.$$

Recalling that instead of (24) one can obtain

$$(29) \quad \mathcal{W}(t) \leq -x'\hat{Q}x - 2\bar{\Delta}\rho\|\alpha\|,$$

we conclude with

THEOREM 5. *Consider the uncertain dynamical system (13)–(14) driven by the feedback control*

$$\bar{p}(x, t) = \begin{cases} -\bar{\rho}(x, t) \frac{B'Px}{\|B'Px\|} & \forall (x, t) \notin \mathcal{S}, \\ \{u \in \mathbb{R}^m : \|u\| \leq \bar{\rho}\} & \forall (x, t) \in \mathcal{S}, \end{cases}$$

where \mathcal{S} is given by (17). Suppose the matrix A is Hurwitz and that conditions (26) hold. Then

- (i) *If $\bar{\rho} \geq \rho$ then the origin $\{x = 0\}$ is uniformly asymptotically stable in the large for any admissible uncertainty (Theorem 3).*
- (ii) *If \hat{Q} is p.s.d. and*

$$\bar{\rho} > \rho \frac{\lambda_{\max}(\tilde{P})}{\lambda + \lambda_{\min}(\tilde{P})},$$

then \mathcal{S} is asymptotically attractive everywhere.

- (iii) *If*

$$\bar{\rho} > \rho \frac{\lambda_{\max}(\tilde{P})}{\lambda + \lambda_{\min}(\tilde{P})},$$

then \mathcal{S} is asymptotically attractive in Γ .

- (iv) *If*

$$\bar{\rho} > \rho \frac{\lambda_{\max}(\tilde{P})}{\lambda + \lambda_{\min}(\tilde{P})} + \frac{\|B'PAx\|}{\lambda + \lambda_{\min}(\tilde{P})},$$

then \mathcal{S} is asymptotically attractive everywhere.

COROLLARY 1. *Theorem 5 remains unchanged if $\bar{p}(\cdot)$ is replaced by an approximation like (11). In that case we talk about the asymptotic attractiveness of $N_e(\mathcal{S})$.*

Example 1. Let

$$A = \begin{bmatrix} 0 & 1 \\ -0.5 & -1.5 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

and choose $Q = I$. Then

$$P = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}, \quad \hat{P} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \hat{Q} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix},$$

$\tilde{P} = 1$, and thus $\lambda_{\min}(\tilde{P}) = \lambda_{\max}(\tilde{P}) = 1$. For $\Delta B = \begin{bmatrix} 0 \\ q \end{bmatrix}$, $q \geq 0$, we find $\lambda = 0$.

We conclude

$$\mathcal{V}(x) = \|\alpha(x)\|^2 \text{ is positive outside } \mathcal{S} = \{x : x_1 + x_2 = 0\},$$

$$\mathcal{W}(t) \leq -\|\alpha(x(t))\|^2 - 2(\bar{\rho} - \rho)\|\alpha(x(t))\|.$$

Thus, provided $\bar{\rho} \geq \rho$, \mathcal{S} is asymptotically attractive everywhere.

6. Insensitivity of trajectories in \mathcal{S} . In this section we show that the concept of *average motion*, which is commonly used in regular relay control, can be extended to uncertain dynamical systems. We stress that in order to make this generalization, we need Theorems 1 and 2 for GDS. According to the previous section (in particular, Theorem 5), if the control magnitude $\bar{\rho}$ is properly chosen, the surface \mathcal{S} is asymptotically attractive. According to Theorem 1, a solution exists in \mathcal{S} ; that is, a sliding mode is possible. According to Theorem 2, the approximated solution lies in the neighborhood of the original one, and thus in the neighborhood of \mathcal{S} . In the following, therefore, we shall investigate the ideal motion in \mathcal{S} . For simplicity we discuss the linear version as in § 5. According to (17), motion in \mathcal{S} satisfies

$$(30) \quad B'Px(t) = 0.$$

Thus, by (3a),

$$B'P\dot{x} = B'P(Ax + B(u + \eta)) = B'PAx + B'PB(u + \eta) = 0.$$

Since $B'PB$ is nonsingular, u “chooses” its values in the GDS sense according to

$$u + \eta = -[(B'PB)^{-1}B'PA]x.$$

Substituting in (3a), one obtains

$$(31) \quad \dot{x} = [I - B(B'PB)^{-1}B'P]Ax$$

This result is clearly independent of the uncertainty vector η .

Remark 5. One should look upon the motion defined by (31) as a limit process of a sequence of approximated solutions as the approximation tends to zero. In this respect (31) defines the *average motion*.

7. Model-following control. The previous sections have demonstrated the systematic use of the min-max controller and its flexibility. In this section we apply the results to the adaptive model-following control (AMFC).

Let a system be given by

$$(32) \quad \dot{x}_p = A_p x_p + B_p(u + r)$$

and a model

$$(33) \quad \dot{x}_m = A_m x_m + B_m r,$$

where r is a reference signal.

It is assumed that:

- (i) the whole *state* vector is available for *measurement*,
- (ii) the *perfect model matching conditions* [26], [27] are satisfied.

In fact, conditions (ii) are equivalent to the *matching conditions* of § 5.

Let

$$(34) \quad e = x_p - x_m.$$

Then

$$(35) \quad \dot{e} = A_m e + B_p u + \Delta A x_p + \Delta B r,$$

where

$$(36) \quad \Delta A = A_p - A_m, \quad \Delta B = B_p - B_m.$$

If the matching conditions of § 5 are met, (35) becomes

$$(37) \quad \dot{e} = A_m e + B_p(u + \eta),$$

where

$$(38) \quad \eta = H(\nu)x_p + E(\nu)r.$$

In equation (38) we have taken into account the possibility of uncertainty in A_p . If we define

$$(39) \quad \rho = \max_n \|H(\nu)x_p\| + \max_\nu \|E(\nu)r\|,$$

our previous results, and in particular Theorem 5, are applicable. Note that outside $N_e(\mathcal{S})$ (see (11)), the control function has the form

$$(40) \quad \bar{p}(e, x_p) = -\bar{\rho}(r, x_p) \frac{B_p' P e}{\|B_p' P e\|},$$

with $\bar{\rho} \geq \rho$ and $PA_m + A_m'P = -Q$, where Q is p.d. The invariant surface is a function of the error only and is given by

$$(41) \quad \mathcal{S} = \{e: B_p' P e = 0\}.$$

8. An illustration. The aim of the present example is to illustrate the computational procedure of § 7. Toward this end we have chosen an example from the literature. This will help the reader to compare procedures previously examined [28]–[30] with the present min-max controller. The data of the problem are as follows:

$$A_m = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 5.318E-7 & -0.4179 & -0.1202 & 2.319E-3 \\ -4.619E-9 & 1.0 & -0.7523 & -2.387E-2 \\ -0.5614 & 0 & 0.3002 & -1.743E-2 \end{bmatrix},$$

$$B_m = \begin{bmatrix} 0 & 0 \\ -0.1717 & 7.451E-6 \\ -0.0238 & -7.783E-5 \\ 0 & 3.685E-3 \end{bmatrix},$$

$$A_p = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1.401E-4 & \sigma & -1.9513 & 0.0133 \\ -2.505E-4 & 1.0 & -1.3239 & -0.0238 \\ -0.561 & 0 & 0.358 & -0.0279 \end{bmatrix},$$

$$B_p = \begin{bmatrix} 0 & 0 & 0 \\ -5.3307 & 6.447E-3 & -0.2669 \\ -0.16 & -1.155E-2 & -0.2511 \\ 0 & 0.106 & 0.0862 \end{bmatrix},$$

$$r = [r_1 \ r_2 \ 0]'$$

We now calculate the control function (40):

$$\Delta A = A_p - A_m = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1.3957E-4 & 0.4179 + \sigma & -1.8311 & 1.0981E-2 \\ -2.505E-4 & 0 & -0.5716 & 7.0E-5 \\ 4.0E-4 & 0 & 0.0578 & -1.047E-2 \end{bmatrix},$$

$$\sigma \in [-0.558, -3.558].$$

To calculate ΔB , we augment B_m with an additional zero column:

$$\Delta B = B_p - [B_m \mid 0] = \begin{bmatrix} 0 & 0 & 0 \\ -5.159 & 6.4395E-3 & -0.2669 \\ -0.1362 & -1.1472E-2 & -0.2511 \\ 0 & 0.10232 & 0.0862 \end{bmatrix}.$$

Since

$$B_p = \begin{bmatrix} 0 \\ \hat{B}_p \end{bmatrix}$$

and \hat{B}_p is nonsingular, matching conditions are satisfied such that

$$H(v) = \begin{bmatrix} 0 \\ \hat{H}(v) \end{bmatrix}, \quad E = \begin{bmatrix} 0 \\ \hat{E} \end{bmatrix},$$

and with $\sigma = -2.058 + 1.5v$, $|v| \leq 1$, we find

$$\hat{H}(v) = \begin{bmatrix} -6.76E-5 & 0.319 - 0.291v & 0.232 & 2.49E-3 \\ 3.04E-3 & 0.171 - 0.157v & -1.231 & -0.104 \\ 1.071E-3 & -0.21 + 0.193v & 2.209 & 1.608E-3 \end{bmatrix},$$

$$\hat{E} = \begin{bmatrix} 0.971 & -1.145E-4 & -4.314E-5 \\ 6.49E-2 & 0.964 & -1.474E-5 \\ -7.978E-2 & 4.51E-2 & 1.037 \end{bmatrix},$$

$$\rho(x_p, r) = \max_v \left\| \begin{bmatrix} \hat{H}(v) \\ \hat{E} \end{bmatrix} \begin{bmatrix} x_2 \\ x_3 \\ x_4 \end{bmatrix}_p + \begin{bmatrix} r_1 \\ r_2 \\ 0 \end{bmatrix} \right\|.$$

From $PA_m + A'_m P = -Q$, $Q = I$, we obtain

$$P = \begin{bmatrix} 2718.8 & 4617.5 & -732.94 & 0.89 \\ 4617.5 & 8043.3 & -1256.7 & 100.03 \\ -732.94 & -1256.7 & 199.07 & -5.99 \\ 0.89 & 100.03 & -5.99 & 50.21 \end{bmatrix},$$

$$B'_p P = \begin{bmatrix} -2333 & -4068 & 635.1 & -52.13 \\ 38.33 & 76.97 & -11.04 & 6.036 \\ -1048 & -1823 & 284.9 & -2086 \end{bmatrix}.$$

Finally, recall the control function (40).

Acknowledgment. The authors wish to thank Professor G. Leitmann and M. Corless for their valuable comments.

REFERENCES

- [1] J. LASALLE AND S. LEFSCHETZ, *Stability by Liapunov's Direct Method with Applications*, Academic Press, New York, 1961.
- [2] W. HAHN, *Theory and Application of Liapunov's Direct Method*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1963.
- [3] R. E. KALMAN AND J. E. BERTRAM, *Control system analysis and design via the "second method" of Lyapunov*, ASME Trans. Basic Eng., 82 (1960), pp. 371-393.
- [4] R. W. BASS, *Discussion of a paper by A. M. Letov*, in Proc. Heidelberg Conference of Automatic Control, Instruments Pub. Co., Pittsburgh, PA, 1957, pp. 209-210.
- [5] R. V. MONOPOLI, *Engineering aspects of control systems design via the direct method of Lyapunov*, NASA Rep. CR-654, 1966.
- [6] L. P. GRAYSON, *The status of synthesis using Lyapunov's method*, Automatica, 3 (1965), pp. 91-121.
- [7] V. I. UTKIN, *Sliding Regimes and their Applications in Variable Structure Systems*, Moscow, Nauka, 1974. (In Russian.)
- [8] ———, *Variable structure systems with sliding modes*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 212-222.
- [9] U. ITKIS, *Control Systems of Variable Structure*, Keter Pub. House, Jerusalem, 1976.
- [10] J. ANDRÉ AND P. SEIBERT, *Über stückweise lineare Differentialgleichungen die bei Regelungsproblemen auftreten*, I, II, Arch. Math., 7 (1956), pp. 148-156 and 157-165.
- [11] ———, *After end-point motions of general discontinuous control systems and their stability properties*, 1st Congress of IFAC, June 1960, pp. 919-922.
- [12] R. V. MONOPOLI, *Discussion on two theorems on the second method*, IEEE Trans. Automat. Control, AC-11 (1966), pp. 140-141.
- [13] Y. I. ALIMOV, *On the application of Lyapunov's direct method to differential equations with ambiguous right sides*, Automat. Remote Control, 22 (1961), pp. 817-830.
- [14] S. GUTMAN, *Uncertain dynamical systems—A differential game approach*, NASA Tech. Memo. TMX-73,135, Apr. 1976.
- [15] S. GUTMAN AND G. LEITMANN, *Stabilizing feedback control for dynamical systems with bounded uncertainty*, presented at the IEEE Conference on Decision and Control, 1976.
- [16] S. GUTMAN, *Uncertain dynamical systems—a Lyapunov min-max approach*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 437-443.
- [17] E. ROXIN, *Axiomatic foundation of the theory of control systems*, presented at the 2nd Internat. Conference of IFAC, Basel, Switzerland, 1963.
- [18] ———, *Stability in general control systems*, J. Differential Equations, 1 (1969), pp. 115-150.
- [19] ———, *On stability in control systems*, this Journal, 3 (1966), pp. 357-372.
- [20] ———, *On asymptotic stability in control systems*, Rend. Circ. Mat. Palermo ser. II, Vol. XV (1966), pp. 193-208.
- [21] ———, *On generalized dynamical systems defined by contingent equations*, J. Differential Equations, 1 (1965), pp. 188-205.

- [22] G. LEITMANN, *Guaranteed asymptotic stability for some linear systems with bounded uncertainties*, J. Dynamic Systems Measurement Control, 101 (1979), pp. 212–216.
- [23] M. CORLESS AND G. LEITMANN, *Continuous state feedback guaranteeing uniform ultimate boundedness for uncertain dynamic systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 1139–1144.
- [24] L. P. GRAYSON, *Two theorems on the second method*, IEEE Trans. Automat. Control, AC-9 (1964), p. 587.
- [25] A. CELLINA, *Multivalued differential equations and ordinary differential equations*, SIAM J. Appl. Math., 18 (1970), pp. 533–538.
- [26] H. ERZBERGER, *Analysis and design of model following systems by state space techniques*, in Proc. Joint Automat. Control Conference, 1968, pp. 572–581.
- [27] K. S. NARENDRA AND P. KUDVA, *Stable adaptive schemes for system identification and control, I*, IEEE Trans. Systems Man Cybernet., SMC-4 (1974), pp. 542–551.
- [28] K. K. D. YOUNG, *Design of variable structure model-following control systems*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 1079–1085.
- [29] C. A. WINSOR AND R. J. ROY, *The application of specific optimal control to design of desensitized model following control systems*, IEEE Trans. Automat. Control, AC-15 (1970), pp. 326–333.
- [30] I. D. LANDAU AND B. COURTIOL, *Design of multivariate adaptive model following control systems*, Automatica, 10 (1974), pp. 483–494.

BOUNDARY STABILIZATION OF HYPERBOLIC SYSTEMS WITH NO DISSIPATIVE CONDITIONS*

SUN SHUNHUA†

Abstract. In this paper a theory of boundary stabilization of hyperbolic systems is given in such a way that the resulting system is not necessarily dissipative. Further, we obtain the uniform boundedness of a class of semigroups whose infinitesimal generator is not necessarily dissipative. The key point of this paper is that in contrast to most other works on stabilization of hyperbolic systems we do not require our generator to be dissipative.

1. Introduction. There has been much work (cf. [1], [4], [7]–[12] and references in [8]) on the stabilization of hyperbolic systems. Most of the papers, with the exception of [4] and [12], are concerned only with feedback laws such that the resulting systems are dissipative. As is pointed out at the beginning of [7], the dissipative property demands that the control forces be applied at those points of the system where the sensors are located and that they be proportional to the measured data. This leads to many more restrictions on the control devices. Moreover, for some practical control systems, the locations of the sensors and of the points where the control forces are applied are restricted to a preassigned finite number of places in the system. All this leads us to consider feedback laws for the stabilization problem such that the resulting system is not necessarily dissipative. We shall see later that the stabilization problem with no dissipative condition is also interesting even from the purely mathematical point of view.

Let us begin with some examples.

Example 1. The boundary stabilization of the wave equation. Let $\Omega \subset \mathbb{R}^n$ be a bounded and connected domain with smooth boundary Γ . The boundary control (forces) can be applied only on a finite number of subsets $\Gamma_i (1 \leq i \leq N)$ of Γ . The system that we are concerned with in this paper is the following:

$$\begin{aligned} (1) \quad & \frac{\partial^2 W}{\partial t^2} = \Delta W \quad \text{in } \Omega, \quad t > 0, \\ & W|_{\Gamma_0} = 0, \\ & \frac{\partial W}{\partial \eta} \Big|_{\Gamma/\Gamma_0} = \begin{cases} -g_i U_i(t) & \text{on } \Gamma_i, \quad 1 \leq i \leq N, \\ 0 & \text{on } \Gamma \setminus \left(\Gamma_0 \bigcup_{i=1}^N \Gamma_i \right), \end{cases} \\ & W|_{t=0} = W_0, \quad \frac{\partial W}{\partial t} \Big|_{t=0} = W_1. \end{aligned}$$

Here $\Gamma_i \cap \Gamma_j = \emptyset$ (empty set) $1 \leq i, j \leq N$, $g_i \in H^{1/2}(\Gamma_i)$ and $U_i(\cdot) \in L^2(0, T)$ ($\forall T > 0$) is a scalar control function, and $H^m(\cdot)$ is Sobolev space of order m . For simplicity, we assume that the $(n-1)$ dimensional Lebesgue measure of Γ_i is not zero, $0 \leq i \leq N$, and that the boundary $\partial\Gamma_i$ of Γ_i is also smooth, $0 \leq i \leq N$.

* Received by the editors November 11, 1980, and in final revised form, December 1, 1981. This research was performed while the author was a Visiting Professor at Purdue University, W. Lafayette, Indiana.

† Sichuan University, Chengdu, Sichuan, People's Republic of China.

Let $V = H_{\Gamma_0}^1 = \{\phi \in H^1(\Omega) | \phi|_{\Gamma_0} = 0 \text{ in } H^{1/2}(\Gamma)\}$. The sesquilinear functional on V ,

$$\Pi(\phi, \psi) = \int_{\Omega} \sum_{i=1}^n \phi_{x_i} \bar{\psi}_{x_i} d\Omega,$$

defines a bounded linear operator $A \in \mathcal{L}(V, V')$ such that $\Pi(\phi, \psi) = (A\phi, \psi)_{(V, V')}$, where V' is the dual space of V and $(\cdot, \cdot)_{(V, V')}$ is the dual product of V and V' . Let $\hat{\mathcal{D}} = \{\phi \in V | A\phi \in L^2(\Omega)\}$. It is not difficult to verify that the restriction of A on $\hat{\mathcal{D}}$ (still denoted by A) is an unbounded positive operator with its domain $\mathcal{D}(A) = \hat{\mathcal{D}}$ and $A = -\Delta$ on $\hat{\mathcal{D}}$. Indeed, $\hat{\mathcal{D}} = \{\phi \in H_{\Gamma_0}^1 | \partial\phi/\partial\eta = 0 \text{ on } \Gamma/\Gamma_0 \text{ (in } H^{-1/2}(\Gamma/\Gamma_0) \text{), cf. [6, p. 24]}\}$ with $\Delta\phi \in L^2(\Omega)$ (in the distribution sense). In the remainder of this paper, the operator A is always understood in this sense. We choose the feedback as follows:

$$(2) \quad U_i(t) = \langle \dot{W}(\cdot, t), b_i(\cdot) \rangle = \int_{\Omega} \dot{W}(x, t) \overline{b_i(x)} d\Omega, \quad 1 \leq i \leq N,$$

where $b_i \in L^2(\Omega)$, $(1 \leq i \leq N)$.

Our problem is whether we can choose b_i ($1 \leq i \leq N$) to stabilize the system (1) under the feedback (2).

Let \hat{g}_i ($1 \leq i \leq N$) be the unique solution of the following problem:

$$(3) \quad \begin{aligned} \Delta \hat{g}_i &= 0 \quad \text{in } \Omega, \\ \hat{g}_i|_{\Gamma_0} &= \frac{\partial \hat{g}_i}{\partial \eta} \Big|_{\Gamma/(\Gamma_i \cup \Gamma_0)} = 0, \\ \frac{\partial \hat{g}_i}{\partial \eta} \Big|_{\Gamma_i} &= g_i. \end{aligned}$$

It is known ([5, p. 18]) that $g_i \in \mathcal{D}(A^{1/2}) = V$. From (1)–(3) it follows that $(W + \sum_{i=1}^N \hat{g}_i \langle \dot{W}, b_i \rangle)|_{\Gamma_0} = 0$, $(\partial/\partial\eta)(W + \sum_{i=1}^N \hat{g}_i \langle \dot{W}, b_i \rangle)|_{\Gamma/\Gamma_0} = 0$ (in $H^{-1/2}(\Gamma/\Gamma_0)$) and that $\Delta(W + \sum_{i=1}^N \hat{g}_i \langle \dot{W}, b_i \rangle) = \Delta W \in L^2(\Omega)$ (in the distribution sense), so we have

$$\left(W + \sum_{i=1}^N \hat{g}_i \langle \dot{W}, b_i \rangle \right) \in \mathcal{D}(A)$$

and

$$(1') \quad \begin{aligned} \frac{d^2 W}{dt^2} &= -A \left(W + \sum_{i=1}^N \hat{g}_i \langle \dot{W}, b_i \rangle \right), \\ W|_{t=0} &= W_0, \quad \dot{W}|_{t=0} = W_1. \end{aligned}$$

Making the substitution

$$Y = \begin{pmatrix} A^{1/2} W \\ \dot{W} \end{pmatrix},$$

we obtain

$$(4) \quad \frac{dY(t)}{dt} = \mathcal{A}_0(I + S)Y(t), \quad Y(0) = Y_0 \triangleq \begin{pmatrix} A^{1/2} W_0 \\ W_1 \end{pmatrix},$$

where

$$\mathcal{A}_0 \triangleq \begin{pmatrix} 0 & A^{1/2} \\ -A^{1/2} & 0 \end{pmatrix}, \quad SX = \sum_{i=1}^N \begin{pmatrix} A^{1/2} \hat{g}_i \\ 0 \end{pmatrix} \left\langle X, \begin{pmatrix} 0 \\ b_i \end{pmatrix} \right\rangle,$$

for all $X \in L^2(\Omega) \times L^2(\Omega)$. Let $\mathcal{A} = \mathcal{A}_0(I + S)$. Then

$$\mathcal{D}(\mathcal{A}) = \{X \in L^2(\Omega) \times L^2(\Omega) | (1 + S)X \in \mathcal{D}(\mathcal{A}_0)\}.$$

Example 2. The boundary stabilization of a vibrating string. For a string with one end fixed and other end controllable, it is known [2] that the vibration is described as follows:

$$(5) \quad \begin{aligned} \frac{\partial^2 y(x, t)}{\partial t^2} - \frac{\partial^2 y(x, t)}{\partial x^2} &= 0, & 0 < x < l, \\ y(0, t) &= 0, & \frac{\partial y}{\partial x} \Big|_{x=l} &= -u(t), \\ y(0) &= y_0(x), & \frac{\partial y}{\partial t} \Big|_{t=0} &= y_1(x), \quad p < x < l, \end{aligned}$$

where $u(t)$ is scalar control function, y_0, y_1 and $y(\cdot, t)$ are all in $L^2(0, l)$. We choose the feedback as follows

$$(6) \quad u(t) = \langle \dot{y}, b \rangle = \int_0^l \frac{\partial y(x, t)}{\partial t} \overline{b(x)} dx,$$

where $b \in L^2(0, l)$.

Let $V = H_0^1(0, l) \triangleq \{\phi \in H^1(0, l) | \phi(0) = 0\}$. Similarly, the sesquilinear functional on V

$$\Pi(\phi, \Phi) = \int_0^l \phi_x \bar{\psi}_x dx$$

defines a bounded linear operator $A \in \mathcal{L}(V, V')$ such that $\Pi(\phi, \psi) = (A\phi, \psi)_{(V, V')}$. Let $\hat{\mathcal{D}} = \{\phi \in V | A\phi \in L^2(0, l)\}$. It is also not difficult to verify that the restriction of A on $\hat{\mathcal{D}}$ (still denoted by A) is an unbounded positive operator on $L^2(0, l)$ with domain $\mathcal{D}(A) = \hat{\mathcal{D}}$. Indeed, $\hat{\mathcal{D}} = \{\phi \in H^2(0, l) | \phi(0) = 0, \phi'(l) = 0\}$ and the operation $A\phi = \phi''$ for all $\phi \in \hat{\mathcal{D}}$. Let g be the unique solution of the following problem in

$$\frac{d^2 g}{dx^2} = 0, \quad 0 < x < l, \quad g(0) = 0, \quad g'(l) = 1.$$

It is easy to see that $g = x$, hence, $g \in V = \mathcal{D}(A^{1/2})$. As in Example 1, we have that $(y(\cdot, t) + \langle \dot{y}, b \rangle g) \in \mathcal{D}(A)$. Then the system (5) with feedback (6) is reduced to the following

$$(5') \quad \frac{d^2 y(t)}{dt^2} = -A(y + \langle \dot{y}, b \rangle g), \quad y(0) = y_0, \quad \frac{dy}{dt} \Big|_{t=0} = y_1.$$

Making the substitution

$$y = \begin{pmatrix} A^{1/2} y \\ \dot{y} \end{pmatrix},$$

we also obtain

$$(4') \quad \frac{dY(t)}{dt} = \mathcal{A}_0(I + S)Y(t), \quad Y(0) = Y_0 \triangleq \begin{pmatrix} A^{1/2} y_0 \\ y_1 \end{pmatrix},$$

where \mathcal{A}_0 has the same form as in Example 1, and in present case, $SX = \langle X, \begin{pmatrix} 0 \\ b \end{pmatrix} \rangle \cdot \begin{pmatrix} A^{1/2} \\ 0 \end{pmatrix} g$ for all $X \in L^2(0, l) \times L^2(0, l)$. For $\mathcal{A} = \mathcal{A}_0(I + S)$, we have a form of $\mathcal{D}(\mathcal{A})$ similar to that in Example 1. Now our problem is whether we can choose $b \in L^2(0, l)$ to stabilize the system (4).

All this leads us to consider the abstract Cauchy problem (4) in an appropriate separate Hilbert space $H \times H$. In particular, in Example 1 we take $H = L^2(\Omega)$. In Example 2 we take $H = (\text{span}\{1, x\})^\perp$ (in $L^2(0, l)$).

In the second part of this paper, we shall show that under the assumptions (H_1) , (H_2) and (SH_2) the semigroup $e^{\mathcal{A}t}$ is uniformly bounded. Then the standard argument using a Lyapunov function yields the asymptotic stability of the solution of the abstract system (4). In the last part of this paper, we shall go back to the examples to exhibit the stabilizability condition into more concrete form.

Evidently, the systems (1) and (5) together with the feedbacks (2) and (6), respectively, are not required to be dissipative.

2. Uniform boundedness of semigroups and stabilization. We now give some general results for the operator $\mathcal{A} = \mathcal{A}_0(I + S)$ and the associated semigroup $e^{\mathcal{A}t}$ with the operator S , an operator of rank N , given as in Example 1. It is worthwhile to note that the operator stated above is general enough to cover some other examples, for instance, the vibrating beam cantilevered on one end (as in Example 2) and the vibrating plate with some simple boundary conditions (as in Example 1).

We assume that the operator A (in (4)) has the spectral resolution $A = \sum_{j=1}^{\infty} \nu_j F_j$, $\nu_j > 0$ ($\forall j \geq 1$) and $\{F_j\}_1^{\infty}$ is the resolution of the identity associated with A in H . It is easy to see that [4] $\mathcal{A}_0 = \sum_{j=-\infty}^{\infty} \lambda_j E_j$, where $\lambda_j = i \cdot (\text{sgn } j) \cdot \sqrt{\nu_{|j|}}$ ($\forall j \neq 0$), Σ' means that the term with index zero is excluded in the summation and $\{E_j\}_{-\infty}^{\infty}$ is the resolution of the identity (in $H \times H$) associated with \mathcal{A}_0 .

We now assume:

(H_1) . $\sup_{k \neq 0} \dim E_k = N < \infty$ (or equivalently $\sup_{k \geq 1} \dim F_k = N$) and there exist constants α and γ such that

$$\sup_{l \neq k} \left| \frac{\lambda_k}{\lambda_l^{2\alpha} (\lambda_l - \lambda_k)} \right| < \infty, \quad \inf_{l \neq k} |\lambda_l^{2\gamma} (\lambda_l - \lambda_k)| > 0;$$

(H_2) . With the constants α and γ given in (H_1) , $A^\alpha b_k$, $A^{1/2+2\gamma} b_k \in H$ ($1 \leq k \leq N$), and the $(N \times N)$ matrices $\langle \langle F_l \hat{g}_i, F_l b_j \rangle \rangle_{N \times N}$ ($\forall l \geq 1$) are all nonnegative with their ranks equal to rank F_l ($\equiv \dim F_l$), respectively.

(SH_2) . It will be shown in the remark in § 2 that the operators $\Lambda_k: F_k \hat{g}_i \rightarrow F_k b_i$, $1 \leq i \leq N$, are well defined and positive on $F_k H$, $k \geq 1$, respectively. We now assume

$$\sup_{k \geq 1} \frac{\max \{|\lambda| : \lambda \in \sigma(\Lambda_k)\}}{\min \{|\lambda| : \lambda \in \sigma(\Lambda_k)\}} < \infty.$$

LEMMA 1. $\mathcal{D}(\mathcal{A})$ is dense in $H \times H$.

Proof. By definition $\mathcal{D}(\mathcal{A}) = \{x \in H \times H | (I + S)x \in \mathcal{D}(\mathcal{A}_0)\} = (1 + S)^{-1} \mathcal{D}(\mathcal{A}_0)$ in the sense of the inverse image of the set $\mathcal{D}(\mathcal{A}_0)$ under the operator $(I + S)$. Since S is compact, $(1 + S)$ is one-one from $H \times H \ominus \ker(I + S)$ onto $H \times H \ominus \ker(I + S^*)$. So $(1 + S)^{-1}$ is bounded from $H \times H \ominus \ker(I + S^*)$ onto $H \times H \ominus \ker(I + S)$. Evidently, $(I + S)^{-1} \mathcal{D}(\mathcal{A}_0) \supset \ker(I + S)$ and $\mathcal{D}(\mathcal{A}_0) \ominus \ker(I + S^*)$ is dense in $H \times H \ominus \ker(I + S^*)$; these lead to $(I + S)^{-1} \mathcal{D}(\mathcal{A}_0) \supset (\text{a dense subset of } H \times H \ominus \ker(I + S)) \oplus \ker(I + S)$ which is dense in $H \times H$.

LEMMA 2 [12]. For any $x, y \in H \times H$, we have

$$(7) \quad \langle R(\lambda; \mathcal{A})x, y \rangle = \langle R(\lambda; \mathcal{A}_0)x, y \rangle - \left\langle (I + W(\lambda))^{-1} \left\langle R(\lambda; \mathcal{A}_0)x, \begin{pmatrix} 0 \\ b_1 \\ \vdots \\ 0 \\ b_N \end{pmatrix} \right\rangle, \right. \\ \left. \left\langle y, A^{1/2} R(\lambda; \mathcal{A}_0) \begin{pmatrix} 0 \\ A^{1/2} \hat{g}_1 \\ \vdots \\ 0 \\ A^{1/2} \hat{g}_N \end{pmatrix} \right\rangle \right\rangle_{\mathbb{C}^N} \quad \forall \operatorname{Re} \lambda > 0,$$

where $R(\lambda; \cdot) \triangleq (\lambda - \cdot)^{-1}$ and

$$(8) \quad \begin{aligned} W(\lambda) &= \left(\left\langle A^{1/2} R(\lambda; \mathcal{A}_0) \begin{pmatrix} 0 \\ A^{1/2} \hat{g}_i \end{pmatrix}, \begin{pmatrix} 0 \\ b_j \end{pmatrix} \right\rangle \right)_{N \times N}^T \\ &= \sum_{l=-\infty}^{\infty} \frac{|\lambda_l|}{\lambda - \lambda_l} \left(\left\langle E_l \begin{pmatrix} 0 \\ A^{1/2} \hat{g}_i \end{pmatrix}, \begin{pmatrix} 0 \\ b_j \end{pmatrix} \right\rangle \right)_{N \times N}^T \\ &= \frac{1}{2} \sum_{l=-\infty}^{\infty} \frac{|\lambda_l|^2}{\lambda - \lambda_l} \langle F_{||} \hat{g}_i, b_j \rangle_{N \times N}^T, \end{aligned}$$

where T means the transpose of the matrix (of order $N \times N$). Also, we use the same symbol I to denote the identity matrix (of order $N \times N$).

Proof. Verify directly. In fact,

(the right-hand side of (7) with $(\lambda - \mathcal{A})x$ instead of x)

$$= \langle R(\lambda; \mathcal{A}_0)(\lambda - \mathcal{A}_0(I + S))x, y \rangle$$

$$- \left\langle (I + W(\lambda))^{-1} \left\langle R(\lambda; \mathcal{A}_0)(\lambda - \mathcal{A}_0(I + S))x, \begin{pmatrix} 0 \\ b_1 \\ \vdots \\ 0 \\ b_N \end{pmatrix} \right\rangle, \right. \\ \left. \left\langle y, A^{1/2} R(\lambda; \mathcal{A}_0) \begin{pmatrix} 0 \\ A^{1/2} \hat{g}_1 \\ \vdots \\ 0 \\ A^{1/2} \hat{g}_N \end{pmatrix} \right\rangle \right\rangle_{\mathbb{C}^N},$$

$$\begin{aligned}
&= \langle (I - R(\lambda; \mathcal{A}_0) \mathcal{A}_0 \mathcal{S})x, y \rangle \\
&= \left\langle (I + W(\lambda))^{-1} \left\langle (I - R(\lambda; \mathcal{A}_0) \mathcal{A}_0 \mathcal{S})x, \begin{pmatrix} 0 \\ b_1 \\ \vdots \\ 0 \\ b_N \end{pmatrix} \right\rangle, \right. \\
&\quad \left. \left\langle y, A^{1/2} R(\lambda; \mathcal{A}_0) \begin{pmatrix} 0 \\ A^{1/2} \hat{g}_1 \\ \vdots \\ 0 \\ A^{1/2} \hat{g}_N \end{pmatrix} \right\rangle \right\rangle_{\mathbb{C}^N} \\
&= \langle (I - R(\lambda; \mathcal{A}_0) \mathcal{A}_0 \mathcal{S})x, y \rangle \\
&= \left\langle (I + W(\lambda))^{-1} (I + W(\lambda)) \left\langle x, \begin{pmatrix} 0 \\ b_1 \\ \vdots \\ 0 \\ b_N \end{pmatrix} \right\rangle, \left\langle y, A^{1/2} R(\lambda; \mathcal{A}_0) \begin{pmatrix} 0 \\ A^{1/2} \hat{g}_1 \\ \vdots \\ 0 \\ A^{1/2} \hat{g}_N \end{pmatrix} \right\rangle \right\rangle_{\mathbb{C}^N} \\
&= \langle x, y \rangle \quad \forall \operatorname{Re} \lambda > 0, \quad x, y \in H \times H.
\end{aligned}$$

Similarly, we have $(\lambda - \mathcal{A}) \cdot$ (the corresponding terms of the right-hand side of (7)) = I .

We can verify (8) easily by noting the following. Every normal eigenvector ϕ of A associated with eigenvalue ν gives rise to two orthonormal eigenvectors $\psi_+ = 1/\sqrt{2}(\phi)$ and $\psi_- = 1/\sqrt{2}(\bar{\phi})$ of \mathcal{A}_0 associated with the eigenvalues of $i\sqrt{\nu}$ and $-i\sqrt{\nu}$, respectively.

COROLLARY. *Under the assumptions (H₁) and (H₂), \mathcal{A}^{-1} is compact and*

$$(9) \quad \sigma(\mathcal{A}) = \{\lambda \mid \det(I + W(\lambda)) = 0\} \subset \{\lambda \mid \operatorname{Re} \lambda < 0\}.$$

Proof. The proof is as given in [4] and [12]. Indeed, (7) holds for all $\lambda \notin \sigma(\mathcal{A}_0) \cup \sigma(\mathcal{A})$, because $R(\lambda; \mathcal{A})$ is analytic in $\rho(\mathcal{A})$. It follows from the proof of Lemma 2 that $\lambda \in \rho(\mathcal{A})$ whenever $\lambda \notin \sigma(\mathcal{A}_0)$ and $\det(I + W(\lambda)) \neq 0$. In particular, \mathcal{A}^{-1} exists and is compact. As in [12] it is not difficult to verify that $\sigma(\mathcal{A}) \cap \sigma(\mathcal{A}_0) = \emptyset$ (empty set). Thus, only the second inclusion relation remains to be proved. Assume now that there exists λ_0 such that $\det(I + W(\lambda_0)) = 0$. Hence, there would be a nonzero vector $\xi \in \mathbb{C}^N$ such that

$$0 = \langle (1 + W(\lambda_0))\xi, \xi \rangle_{\mathbb{C}^N} = \|\xi\|^2 + \frac{1}{2} \sum_{-\infty}^{\infty} \frac{|\lambda_l|^2}{\lambda_0 - \lambda_l} \langle (F_l \hat{g}_i, b_j) \rangle_{N \times N}^T \xi, \xi \rangle_{\mathbb{C}^N}.$$

If we now let $r = \operatorname{Re} \lambda_0$ and notice that the λ_l 's are imaginary, we get

$$0 = \|\xi\|^2 + \frac{r}{2} \sum_{-\infty}^{\infty} \frac{|\lambda_l|^2}{|\lambda_0 - \lambda_l|^2} \langle (F_l \hat{g}_i, b_j) \rangle_{N \times N}^T \xi, \xi \rangle_{\mathbb{C}^N} + \text{imaginary part}.$$

By assumption (H₂), the matrices $\langle (F_l \hat{g}_i, b_j) \rangle_{N \times N}^T$ are all nonnegative. This together with the above equality leads to $r < 0$, i.e., $\operatorname{Re} \lambda_0 < 0$.

The proof of the corollary is thus complete.

The following remark is needed for the remainder of this paper.

Remark. Because of assumption (H₂), we can assume, without loss of generality, that for any l , $\text{span}\{F_l\hat{g}_1, \dots, F_l\hat{g}_{\rho_l}\} = F_lH$, where $\rho_l = \dim F_l$. Define an operator Λ_l in F_lH as follows:

$$(10) \quad \Lambda_l(F_l\hat{g}_k) = F_lb_k \quad 1 \leq k \leq \rho_l.$$

The operator Λ_l is positive because for any $x = \sum_{i,j=1}^{\rho_l} \alpha_i F_l\hat{g}_i \in F_lH$ we get from (H₂) that

$$(11) \quad \langle \Lambda_l x, x \rangle = \sum_{i,j=1}^{\rho_l} \alpha_i \bar{\alpha}_j \langle \Lambda_l F_l\hat{g}_i, F_l\hat{g}_j \rangle = \sum_{i,j=1}^{\rho_l} \alpha_i \bar{\alpha}_j \langle F_lb_i, F_l\hat{g}_j \rangle > 0,$$

whenever $\|x\| \neq 0$. If now $N > \rho_l$, we claim there must be

$$(12) \quad \Lambda_l F_l\hat{g}_k = F_lb_k, \quad \rho_l + 1 \leq k \leq N.$$

We prove (12) only for $k = \rho_l + 1$. Since $\text{span}\{F_l\hat{g}_1, \dots, F_l\hat{g}_{\rho_l}\} = F_lH$. We have $F_l\hat{g}_{\rho_l+1} = \sum_{i=1}^{\rho_l} r_i F_l\hat{g}_i$ for some $\{r_i\}_1^{\rho_l}$. The nonnegativity of the matrix $(\langle F_l\hat{g}_i, F_lb_j \rangle)_{N \times N}$ implies that $\langle F_l\hat{g}_{\rho_l+1}, F_lb_k \rangle = \langle F_lb_{\rho_l+1}, F_l\hat{g}_k \rangle$, $1 \leq k \leq \rho_l$. For the same reason, $F_lb_{\rho_l+1} = \sum_{k=1}^{\rho_l} r'_k F_lb_k$. Thus we obtain

$$(13) \quad (\langle F_lb_i, F_l\hat{g}_j \rangle)_{\rho_l \times \rho_l} \begin{pmatrix} \bar{r}_1 \\ \vdots \\ \bar{r}_{\rho_l} \end{pmatrix} = (\langle F_l\hat{g}_i, F_lb_j \rangle) \begin{pmatrix} \bar{r}'_1 \\ \vdots \\ \bar{r}'_{\rho_l} \end{pmatrix}.$$

By the positivity (and also the selfadjointness) of the matrix $(\langle F_lb_i, F_l\hat{g}_j \rangle)_{N \times N}$ we have $r_i = r'_i$, $1 \leq i \leq \rho_l$. Hence (12) is well defined; i.e., (12) is compatible with (10). We now set $\Lambda = \sum_{i=1}^{\infty} \Lambda_i^{1/2} F_i$. The operator Λ is positive in H with $\mathcal{D}(\Lambda) = \{x \in H \mid \sum_{i=1}^{\infty} \|\Lambda_i^{1/2} F_i x\|^2 < \infty\}$ and commutes with A (at least on a dense set in $L^2(\Omega)$) because A is a constant multiple ν_k of the identity on every eigenspace F_kH . Moreover, from (12) we get

$$(14) \quad \Lambda \hat{g}_i = \Lambda^{-1} b_i, \quad 1 \leq i \leq N.$$

Let

$$(15) \quad C_i = \Lambda(A^{1/2} \hat{g}_i) = \Lambda^{-1}(A^{1/2} b_i), \quad 1 \leq i \leq N.$$

This is also well defined because $\hat{g}_i \in \mathcal{D}(A^{1/2})$ and (H₂) holds. Thus we have

$$(16) \quad \begin{aligned} \|C_i\| &= (\langle \Lambda(A^{1/2} \hat{g}_i), \Lambda^{-1}(A^{1/2} b_i) \rangle)^{1/2} \\ &= (\langle A^{1/2} \hat{g}_i, A^{1/2} b_i \rangle)^{1/2} < (\|A^{1/2} \hat{g}_i\| \cdot \|A^{1/2} b_i\|)^{1/2} < \infty. \end{aligned}$$

Let $\Lambda_0 = \Lambda A^{-1/2}$ with domain $\mathcal{D}(\Lambda_0) = \{x \in H \mid \sum_{i=1}^{\infty} (1/\nu_i) \|\Lambda_i^{1/2} F_i x\|^2 < \infty\}$. Then Λ_0 is positive, but may be unbounded in H . This is because $\Lambda_0 x = \sum_{i=1}^{\infty} (1/\sqrt{\nu_i}) \Lambda_i^{1/2} F_i x$, and on F_iH the operator Λ_i is unitarily equivalent to a diagonal positive matrix. Then (15) becomes formally

$$(17) \quad C_i = \Lambda_0 A \hat{g}_i = \Lambda_0^{-1} b_i, \quad 1 \leq i \leq N.$$

We now can restate (SH₂) in the following way:

(SH₂). There exists a sequence of positive numbers $\{\varepsilon_i\}_1^{\infty}$ such that $\sup_{i \in \mathbb{N}} \{\|\varepsilon_i \Lambda_i^{1/2}\| + \|\varepsilon_i^{-1} \Lambda_i^{-1/2}\|\} < \infty$, where the Λ_i 's are as defined in (10).

LEMMA 3. Under the assumptions (H₁) and (H₂),

$$(18) \quad \sum_{n=0}^{\infty} \left\| \frac{1}{n!} \frac{d^n}{d\lambda^n} \left\{ (1+W(\lambda))^{-1} \frac{\hat{C}_j}{\lambda - \lambda_j} \right\} \right\|^2 \leq \frac{1}{\lambda^{2(n+1)}} \sum_{l=1}^N \|C_l\|^2 \quad \forall \lambda > 0, \quad n \geq 0,$$

where $\|(a_{ij})_{m \times n}\|^2 \triangleq \sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2$ and

$$(19) \quad \hat{C}_j = \begin{pmatrix} E_j \begin{pmatrix} 0 \\ C_1 \end{pmatrix} \\ \vdots \\ E_j \begin{pmatrix} 0 \\ C_N \end{pmatrix} \end{pmatrix}_{N \times \dim E_j}$$

The same symbol $E_j \begin{pmatrix} 0 \\ C_k \end{pmatrix}$ is used to denote the vector $E_j \begin{pmatrix} 0 \\ C_k \end{pmatrix}$ and the coordinate vector of the vector $E_j \begin{pmatrix} 0 \\ C_k \end{pmatrix}$ with respect to a chosen orthonormal basis of the space $E_j(H \times H)$; i.e., $E_j \begin{pmatrix} 0 \\ C_k \end{pmatrix}$ also denotes a vector in the complex space of dimension E_j and also the set of $\dim E_j$ complex numbers which are just the components of the vector $E_j \begin{pmatrix} 0 \\ C_k \end{pmatrix}$ with respect to some chosen orthonormal basis in $E_j(H \times H)$.

Proof. Let $\mathcal{A}_C = \mathcal{A}_0 - \sum_{i=1}^N \begin{pmatrix} 0 \\ C_i \end{pmatrix} \langle \cdot, \begin{pmatrix} 0 \\ C_i \end{pmatrix} \rangle$ with $\mathcal{D}(\mathcal{A}_C) = \mathcal{D}(\mathcal{A}_0)$. We have $\mathcal{A}_C = \Lambda_0 \mathcal{A} \Lambda_0^{-1}$ (at least on a dense subset), where $\begin{pmatrix} \Lambda_0 & 0 \\ 0 & \Lambda_0 \end{pmatrix}$ is denoted by the same symbol Λ_0 . Since $\operatorname{Re} \mathcal{A}_C = -\sum_{i=1}^N \begin{pmatrix} 0 \\ C_i \end{pmatrix} \langle \cdot, \begin{pmatrix} 0 \\ C_i \end{pmatrix} \rangle \leq 0$, $e^{\mathcal{A}_C t}$ is a contraction semigroup (cf. [14]). If in (7) we replace b_i and $A\hat{g}_i$ by C_i , we get

$$\begin{aligned} & \left\langle R(\lambda; \mathcal{A}_C) \begin{pmatrix} \begin{pmatrix} 0 \\ C_1 \end{pmatrix} \\ \vdots \\ \begin{pmatrix} 0 \\ C_N \end{pmatrix} \end{pmatrix}, E_j x \right\rangle \quad (\text{matrix of order } N \times 1) \\ &= \frac{1}{\lambda - \lambda_j} \left\langle \begin{pmatrix} E_j \begin{pmatrix} 0 \\ C_1 \end{pmatrix} \\ \vdots \\ E_j \begin{pmatrix} 0 \\ C_N \end{pmatrix} \end{pmatrix}, E_j x \right\rangle \\ &= \frac{1}{\lambda - \lambda_j} \left\langle \left((1+W(\lambda))^{-1} \left(\left\langle \begin{pmatrix} 0 \\ C_i \end{pmatrix}, R(\lambda; \mathcal{A}_0) \begin{pmatrix} 0 \\ C_j \end{pmatrix} \right\rangle \right)^*_{(\text{matrix of order } N \times N)} \right)^T, \right. \\ & \quad \left. \begin{pmatrix} \langle E_j x, E_j \begin{pmatrix} 0 \\ C_1 \end{pmatrix} \rangle \\ \vdots \\ \langle E_j x, E_j \begin{pmatrix} 0 \\ C_N \end{pmatrix} \rangle \end{pmatrix} \right\rangle \quad \mathbb{C}^N (\text{matrix of order } N \times 1) \end{aligned} \quad (20)$$

$$\begin{aligned}
&= \frac{1}{\lambda - \lambda_j} \begin{pmatrix} \left\langle E_j \begin{pmatrix} 0 \\ C_1 \end{pmatrix}, x \right\rangle \\ \vdots \\ \left\langle E_j \begin{pmatrix} 0 \\ C_N \end{pmatrix}, x \right\rangle \end{pmatrix} \\
&\quad - \frac{1}{\lambda - \lambda_j} \left(\left\langle R(\lambda; \mathcal{A}_0) \begin{pmatrix} 0 \\ C_i \end{pmatrix}, \begin{pmatrix} 0 \\ C_i \end{pmatrix} \right\rangle \right)_{N \times N} (I + W(\lambda))^{-1} \begin{pmatrix} \left\langle E_j \begin{pmatrix} 0 \\ C_1 \end{pmatrix}, E_j x \right\rangle \\ \vdots \\ \left\langle E_j \begin{pmatrix} 0 \\ C_N \end{pmatrix}, E_j x \right\rangle \end{pmatrix}
\end{aligned}$$

If we now use (8), (17) and (H₂) we can write the right-most side of (20) as follows:

$$\begin{aligned}
&\frac{1}{\lambda - \lambda_j} (I - W(\lambda)(I + W(\lambda))^{-1}) \begin{pmatrix} \left\langle E_j \begin{pmatrix} 0 \\ C_1 \end{pmatrix}, E_j x \right\rangle \\ \vdots \\ \left\langle E_j \begin{pmatrix} 0 \\ C_N \end{pmatrix}, E_j x \right\rangle \end{pmatrix} \\
&= \frac{1}{\lambda - \lambda_j} (I + W(\lambda))^{-1} \begin{pmatrix} \left\langle E_j \begin{pmatrix} 0 \\ C_1 \end{pmatrix}, E_j x \right\rangle \\ \vdots \\ \left\langle E_j \begin{pmatrix} 0 \\ C_N \end{pmatrix}, E_j x \right\rangle \end{pmatrix}.
\end{aligned}$$

Let $\{e_{jk}\}$ ($1 \leq k \leq \dim E_j$) be an orthonormal basis of $E_j(H \times H)$. Replacing x by e_{jk} in (20) and letting the index k run from 1 to $\dim E_j$, we obtain

$$\begin{aligned}
&\left\langle R(\lambda, \mathcal{A}_C) \begin{pmatrix} \begin{pmatrix} 0 \\ C_1 \end{pmatrix} \\ \vdots \\ \begin{pmatrix} 0 \\ C_N \end{pmatrix} \end{pmatrix} (e_{j1}, \dots, e_{j \dim E_j}) \right\rangle_{N \times \dim E_j} \\
&= \frac{1}{\lambda - \lambda_j} (I + W(\lambda))^{-1} \begin{pmatrix} \left\langle E_j \begin{pmatrix} 0 \\ C_1 \end{pmatrix}, e_{j1} \right\rangle & \cdots & \left\langle E_j \begin{pmatrix} 0 \\ C_1 \end{pmatrix}, e_{j \dim E_j} \right\rangle \\ \vdots & & \vdots \\ \left\langle E_j \begin{pmatrix} 0 \\ C_N \end{pmatrix}, e_{j1} \right\rangle & \cdots & \left\langle E_j \begin{pmatrix} 0 \\ C_N \end{pmatrix}, e_{j \dim E_j} \right\rangle \end{pmatrix} \\
&\triangleq \frac{1}{\lambda - \lambda_j} (I + W(\lambda))^{-1} \hat{C}_j.
\end{aligned}$$

Summing both sides of (21) with respect to the index j , and using known results on $R(\lambda, \mathcal{A}_C)$ (cf. [10, p. 250]), we get

$$\begin{aligned}
 & \sum'_{-\infty}^{\infty} \left\| \frac{1}{n!} \frac{d^n}{d\lambda^n} \left\{ (1 + W(\lambda))^{-1} \frac{\hat{C}_j}{\lambda - \lambda_j} \right\} \right\|^2 \\
 &= \sum'_{-\infty}^{\infty} \left\| \left\langle R^{n+1}(\lambda, \mathcal{A}_C) \begin{pmatrix} 0 \\ C_1 \\ \vdots \\ 0 \\ C_N \end{pmatrix}, (e_{j1}, \dots, e_{j \dim E_j}) \right\rangle \right\|^2 \\
 (22) \quad &= \sum'_{-\infty}^{\infty} \sum_{l=1}^N \left\| E_j R^{n+1}(\lambda; \mathcal{A}_C) \begin{pmatrix} 0 \\ C_l \end{pmatrix} \right\|^2 = \sum_{l=1}^N \left\| R^{n+1}(\lambda; \mathcal{A}_C) \begin{pmatrix} 0 \\ C_l \end{pmatrix} \right\|^2 \\
 &\leq \frac{1}{\lambda^{2(n+1)}} \sum_{l=1}^N \|C_l\|^2 \quad \forall \lambda > 0, \quad n \geq 0 \text{ (} n \text{ integer)}.
 \end{aligned}$$

This is just (18).

We are now at a position to prove our main lemma.

LEMMA 4. *Under the assumptions (H₁), (H₂) and (SH₂), the semigroup $e^{\mathcal{A}t}$ is uniformly bounded.*

Proof. By well-known results (cf. [14, Chap. IX]), we only need to prove that there exists a constant M such that $\|(1/n!)(d^n/d\lambda^n)R(\lambda; \mathcal{A})\| \leq M/\lambda^{n+1}$, $\lambda > 0$, $n \geq 0$. We shall first express $\langle R(\lambda; \mathcal{A})x, y \rangle$ as a sum of certain expressions and then differentiate both sides n times with respect to λ . On the right we shall be able to estimate the derivatives of each term.

In the following we write

$$R(\lambda; \mathcal{A}) \begin{pmatrix} 0 \\ A\hat{g}_l \end{pmatrix} \quad \text{instead of} \quad A^{1/2}R(\lambda; \mathcal{A}_0) \begin{pmatrix} 0 \\ A^{1/2}\hat{g}_l \end{pmatrix}$$

for convenience. By (7) we have

$$\langle R(\lambda; \mathcal{A})x, y \rangle = \langle R(\lambda; \mathcal{A}_0)x, y \rangle$$

$$- \sum'_{l,k=-\infty}^{\infty} \frac{1}{(\lambda - \lambda_l)(\lambda - \lambda_k)} \left\langle (I + W(\lambda))^{-1} \left\langle E_l x, E_l \begin{pmatrix} 0 \\ b_1 \\ \vdots \\ 0 \\ b_N \end{pmatrix} \right\rangle, \left\langle E_k y, E_k \begin{pmatrix} 0 \\ A\hat{g}_1 \\ \vdots \\ 0 \\ A\hat{g}_N \end{pmatrix} \right\rangle \right\rangle_{\mathbb{C}^N}.$$

Now let

$$\Pi_{k,l} = \left\langle (1 + W(\lambda))^{-1} \left\langle E_l x, E_l \begin{pmatrix} 0 \\ b_1 \\ \vdots \\ 0 \\ b_N \end{pmatrix} \right\rangle, \left\langle E_k y, E_k \begin{pmatrix} 0 \\ A\hat{g}_1 \\ \vdots \\ 0 \\ A\hat{g}_N \end{pmatrix} \right\rangle \right\rangle_{\mathbb{C}^N}.$$

Then from the preceding equality, we get

$$\begin{aligned}
 \langle R(\lambda; \mathcal{A})x, y \rangle &= \langle R(\lambda; \mathcal{A}_0)x, y \rangle \\
 &\quad - \sum_{l=-\infty}^{\infty} \frac{1}{(\lambda - \lambda_l)^2} (1 + W(\lambda))^{-1} \Pi_{l,l} - \sum'_{l \neq k} \frac{1}{\lambda_l - \lambda_k} \left(\frac{1}{\lambda - \lambda_l} - \frac{1}{\lambda - \lambda_k} \right) \Pi_{k,l} \\
 &= \langle R(\lambda; \mathcal{A}_0)x, y \rangle \\
 &\quad - \sum'_{l=-\infty}^{\infty} \frac{1}{(\lambda - \lambda_l)^2} \left\langle (1 + W(\lambda))^{-1} \left\langle \varepsilon_l \Lambda_l^{1/2} E_l x, E_l \begin{pmatrix} 0 \\ C_1 \\ \vdots \\ 0 \\ C_N \end{pmatrix} \right\rangle \right\rangle, \\
 &\quad \left\langle \varepsilon_l^{-1} \Lambda_l^{-1/2} E_l y, E_l \begin{pmatrix} 0 \\ C_1 \\ \vdots \\ 0 \\ C_N \end{pmatrix} \right\rangle \Bigg\rangle_{\mathbb{C}^N} \\
 &\quad - \sum'_{l \neq k} \frac{1}{\lambda_l - \lambda_k} \left(\frac{1}{\lambda - \lambda_l} - \frac{1}{\lambda - \lambda_k} \right) \Pi_{k,l}.
 \end{aligned}$$

Here we have used (14), (15) and $AF_l = |\lambda_l| A^{1/2} F_l$ to get the last equality in the above chain of equalities. What we want to do in what follows is to estimate $\|(1/n!)(d^n/d\lambda^n)R(\lambda; \mathcal{A})\|$ term by term by using (23). Before doing this we need first the following equalities.

$$\begin{aligned}
 &\frac{1}{\lambda - \lambda_n} (I + W(\lambda))^{-1} \\
 &= \frac{1}{\lambda - \lambda_n} - \frac{W(\lambda)(1 + W(\lambda))^{-1}}{\lambda - \lambda_n} \\
 (24) \quad &= \frac{I}{\lambda - \lambda_n} - \sum'_{\nu=-\infty}^{\infty} \frac{\bar{C}_\nu \hat{C}_\nu^T}{(\lambda - \lambda_n)(\lambda - \lambda_\nu)} (1 + W(\lambda))^{-1} \\
 &= \frac{I}{\lambda - \lambda_n} - \frac{\bar{C}_n \hat{C}_n^T}{(\lambda - \lambda_n)^2} (1 + W(\lambda))^{-1} \\
 &\quad - \sum'_{\substack{\nu=-\infty \\ \nu \neq n}}^{\infty} \frac{1}{\lambda_n - \lambda_\nu} \left(\frac{1}{\lambda - \lambda_n} - \frac{1}{\lambda - \lambda_\nu} \right) \bar{C}_\nu \hat{C}_\nu^T (1 + W(\lambda))^{-1}.
 \end{aligned}$$

Hence,

$$\begin{aligned}
 \frac{(I + W(\lambda))^{-1}}{\lambda - \lambda_n} &= \left(+ \sum'_{\substack{\nu=-\infty \\ \nu \neq n}}^{\infty} \frac{\bar{C}_\nu \cdot \hat{C}_\nu^T}{\lambda_n - \lambda_\nu} \right)^{-1} \left\{ \frac{1}{\lambda - \lambda_n} + \frac{\bar{C}_n \cdot \hat{C}_n^T}{(\lambda - \lambda_n)^2} (1 + W(\lambda))^{-1} \right. \\
 (25) \quad &\quad \left. + \sum'_{\substack{\nu=-\infty \\ \nu \neq n}}^{\infty} \frac{\bar{C}_\nu \cdot \hat{C}_\nu^T}{(\lambda_n - \lambda_\nu)(\lambda - \lambda_\nu)} (I + W(\lambda))^{-1} \right\}.
 \end{aligned}$$

Now we turn to the term by term estimate of $\|(1/n!)(d^n/d\lambda^n)R(\lambda; \mathcal{A})\|$ by using (23) in the following steps.

Step 1. Estimate of any term in the last summation of the rightmost side of (23). For any term in the last summation, by using (25) we have (l and k being symmetric)

$$\begin{aligned}
 & \left| \frac{1}{n!} \frac{d^n}{d\lambda^n} \left\{ \frac{\Pi_{k,l}}{(\lambda_l - \lambda_k)(\lambda - \lambda_l)} \right\} \right| \\
 &= \left| \frac{1}{n!} \frac{d^n}{d\lambda^n} \frac{1}{\lambda_l - \lambda_k} \left\langle \left(1 + \sum_{\substack{\nu=-\infty \\ \nu \neq l}}^{\infty} \frac{\bar{\tilde{C}}_\nu \hat{C}_\nu^T}{\lambda_l - \lambda_\nu} \right)^{-1} \right. \right. \\
 & \quad \cdot \left[\frac{1}{\lambda - \lambda_l} + \frac{\bar{\tilde{C}}_l \hat{C}_l^T}{(\lambda - \lambda_l)^2} (1 + W(\lambda))^{-1} + \sum_{\substack{\nu=-\infty \\ \nu \neq l}}^{\infty} \frac{\bar{\tilde{C}}_\nu \hat{C}_\nu^T}{(\lambda_l - \lambda_\nu)(\lambda - \lambda_\nu)} (1 + W(\lambda))^{-1} \right] \\
 & \quad \left. \left. \left\langle E_l x, E_l \begin{pmatrix} 0 \\ b_1 \\ \vdots \\ 0 \\ b_N \end{pmatrix} \right\rangle, \left\langle E_k y, E_k \begin{pmatrix} 0 \\ A \hat{g}_1 \\ \vdots \\ 0 \\ A \hat{g}_N \end{pmatrix} \right\rangle \right\rangle_{\mathbb{C}^N} \right|.
 \end{aligned}
 \tag{26}$$

To estimate (26) we need to estimate each term on the right-hand side separately. Before doing this we note the following.

Let

$$\left(1 + \sum_{\substack{\nu=-\infty \\ \nu \neq n}}^{\infty} \frac{\bar{\tilde{C}}_\nu \hat{C}_\nu^T}{\lambda_n - \lambda_\nu} \right)^{-1} \triangleq (I + T_n)^{-1}.$$

Since $\bar{\tilde{C}}_\nu \hat{C}_\nu^T$ are all nonnegative and the λ_k 's are all imaginary, it must be that $T_n^* = -T_n$. Therefore, for any $x \in \mathbb{C}^N$ we have

$$(27) \quad \|(1 + T_n) \times\|_{\mathbb{C}^N}^2 = \|x\|^2 + \|T_n x\|^2 \geq \|x\|_{\mathbb{C}^N}^2,$$

so

$$(27') \quad \|(1 + T_n)^{-1}\| \leq 1 \quad \forall n \neq 0.$$

We now return to the estimate of the terms on the right in (26):

(i)

$$\begin{aligned} & \left| \frac{1}{n!} \frac{d^n}{d\lambda^n} \frac{1}{\lambda_l - \lambda_k} \left\langle \frac{1}{\lambda - \lambda_l} (1 + T_l)^{-1} \left\langle E_l x, E_l \begin{pmatrix} 0 \\ b_1 \\ \vdots \\ 0 \\ b_N \end{pmatrix} \right\rangle, \left\langle E_k y, E_k \begin{pmatrix} 0 \\ A \hat{g}_1 \\ \vdots \\ 0 \\ A \hat{g}_N \end{pmatrix} \right\rangle \right\rangle_{\mathbb{C}^N} \right| \\ & \leq \frac{1}{\lambda^{n+1}} \left\| (1 + T_l)^{-1} \left\langle E_l x, E_l \begin{pmatrix} 0 \\ A^\alpha b_1 \\ \vdots \\ 0 \\ A^\alpha b_N \end{pmatrix} \right\rangle \right\|_{\mathbb{C}^N} \\ & \quad \cdot \left\| \left\langle E_k y, \frac{\lambda_k}{\lambda_l^{2\alpha} (\lambda_l - \lambda_k)} E_k \begin{pmatrix} 0 \\ A^{1/2} \hat{g}_N \\ \vdots \\ A^1 \\ {}^2 \hat{g}_N \end{pmatrix} \right\rangle \right\|_{\mathbb{C}^N} \\ & \leq \frac{1}{\lambda^{n+1}} \sup_{\forall l \neq k} \left| \frac{\lambda_k}{\lambda_l^{2\alpha} (\lambda_l - \lambda_k)} \right| \cdot \left(\sum_{i=1}^N \|E_l (A^\alpha b_i)\|^2 \right)^{1/2} \left(\sum_{j=1}^N \|E_k (A^{1/2} \hat{g}_j)\|^2 \right)^{1/2} \\ & \quad \cdot \|E_l x\| \cdot \|E_k y\|, \end{aligned}$$

(ii)

$$\begin{aligned} & \left| \frac{1}{n!} \frac{d^n}{d\lambda^n} \frac{1}{\lambda_l - \lambda_k} \left\langle (1 + T_l)^{-1} \frac{\bar{C}_l \hat{C}_l^T}{(\lambda - \lambda_l)^2} (1 + W(\lambda))^{-1} \left\langle E_l x, E_l x \begin{pmatrix} 0 \\ b_1 \\ \vdots \\ 0 \\ b_N \end{pmatrix} \right\rangle, \right. \\ & \quad \left. \left\langle E_k y, E_k \begin{pmatrix} 0 \\ A \hat{g}_1 \\ \vdots \\ 0 \\ A \hat{g}_N \end{pmatrix} \right\rangle \right\rangle_{\mathbb{C}^N} \right| \end{aligned}$$

$$= \left| \frac{1}{n!} \frac{d^n}{d\lambda^n} \left\langle \frac{(1+W(\lambda))^{-1}}{(\lambda-\lambda_l)^2} \left\langle \varepsilon_l \Lambda_l^{1/2} E_l x, E_l \begin{pmatrix} 0 \\ C_1 \\ \vdots \\ 0 \\ C_N \end{pmatrix} \right\rangle \right\rangle, \right.$$

$$\left. \bar{C}_l \frac{\hat{C}_l^T}{|\lambda_l| \varepsilon_l (\bar{\lambda}_l - \bar{\lambda}_k)} (1+T_l^*)^{-1} \left\langle E_k y, E_k \begin{pmatrix} 0 \\ A \hat{g}_1 \\ \vdots \\ 0 \\ A \hat{g}_N \end{pmatrix} \right\rangle \right|_{\mathbb{C}^N}.$$

Letting

$$\xi = \frac{\bar{C}_l^T}{|\lambda_l| \varepsilon_l (\bar{\lambda}_l - \bar{\lambda}_k)} (1+T_l^*)^{-1} \left\langle E_k y, E_k \begin{pmatrix} 0 \\ A \hat{g}_1 \\ \vdots \\ 0 \\ A \hat{g}_N \end{pmatrix} \right\rangle,$$

we get that the right-hand side of (ii) is equal to the following:

$$(ii') \quad \left| \frac{1}{n!} \frac{d^n}{d\lambda^n} \left\langle \frac{(1+W(\lambda))^{-1}}{(\lambda-\lambda_l)^2} \bar{C}_l \varepsilon_l \Lambda_l^{1/2} E_l x, \bar{C}_l \xi \right\rangle_{\mathbb{C}^N} \right|.$$

From

$$\begin{aligned} & \langle R(\lambda; \mathcal{A}_C) E_l x, E_l y \rangle \\ &= \frac{1}{\lambda - \lambda_l} \langle E_l x, E_l y \rangle \\ (28) \quad & - \frac{1}{(\lambda - \lambda_l)^2} \left\langle (1+W(\lambda))^{-1} \left\langle E_l x, E_l \begin{pmatrix} 0 \\ C_1 \\ \vdots \\ 0 \\ C_N \end{pmatrix} \right\rangle, \left\langle E_l y, E_l \begin{pmatrix} 0 \\ C_1 \\ \vdots \\ 0 \\ C_N \end{pmatrix} \right\rangle \right\rangle_{\mathbb{C}^N} \\ &= \frac{1}{\lambda - \lambda_l} \langle E_l x, E_l y \rangle - \frac{1}{(\lambda - \lambda_l)^2} \langle (1+W(\lambda))^{-1} \bar{C}_l E_l x, \bar{C}_l E_l y \rangle_{\mathbb{C}^N}, \end{aligned}$$

with $\xi = E_l y$ and the well-known results on the dissipative operator \mathcal{A}_C (cf. [10, p.250]) and by using the following in the expression of ξ :

$$\begin{aligned} \frac{\hat{C}_l^T}{|\lambda| \varepsilon_l} &= \frac{1}{|\lambda_l| \varepsilon_l} \begin{pmatrix} E_l \begin{pmatrix} 0 \\ C_1 \end{pmatrix} \\ \vdots \\ \begin{pmatrix} 0 \\ C_N \end{pmatrix} \end{pmatrix}^T = \frac{1}{|\lambda_l| \varepsilon_l} \begin{pmatrix} E_l \begin{pmatrix} 0 \\ \Lambda_l^{-1/2} A^{1/2} b_1 \end{pmatrix} \\ \vdots \\ E_l \begin{pmatrix} 0 \\ \Lambda_l^{-1/2} A^{1/2} b_N \end{pmatrix} \end{pmatrix}^T \\ &= \begin{pmatrix} \varepsilon_l^{-1} \Lambda_l^{-1/2} E_l \begin{pmatrix} 0 \\ b_1 \end{pmatrix} \\ \vdots \\ \varepsilon_l^{-1} \Lambda_l^{-1/2} E_l \begin{pmatrix} 0 \\ b_N \end{pmatrix} \end{pmatrix}^T, \end{aligned}$$

we get that (ii') is less than or equal to

$$\begin{aligned} &\frac{2}{\lambda^{n+1}} \|\varepsilon_l \Lambda_l^{1/2} E_l x\| \cdot \|\xi\|_{\mathbb{C}^N} \\ &= \frac{2}{\lambda^{n+1}} \|\varepsilon_l \Lambda_l^{1/2} E_l x\| \\ &\quad \cdot \left\| \frac{\lambda_k}{\lambda_l^{2\alpha} (\lambda_l - \lambda_k)} \left(\varepsilon_l^{-1} \Lambda_l^{-1/2} E_l \begin{pmatrix} 0 \\ A^\alpha b_1 \end{pmatrix}, \dots, \right. \right. \\ &\quad \left. \left. \varepsilon_l^{-1} \Lambda_l^{-1/2} E_l \begin{pmatrix} 0 \\ A^\alpha b_N \end{pmatrix} \right) \right\|_{\dim E_l \times N} (1 + T_l^*)^- \\ &\quad \cdot \left\langle E_k y, E_k \begin{pmatrix} 0 \\ A^{1/2} \hat{g}_1 \\ \vdots \\ 0 \\ A^{1/2} \hat{g}_N \end{pmatrix} \right\rangle \Bigg\|_{\mathbb{C}^N} \\ &\leq \frac{2}{\lambda^{n+1}} \|\varepsilon_l \Lambda_l^{1/2}\| \cdot \|\varepsilon_l^{-1} \Lambda_l^{-1/2}\| \sup_{\forall l \neq k} \left| \frac{\lambda_k}{\lambda_l^{2\alpha} (\lambda_l - \lambda_k)} \right| \cdot \left(\sum_{i=1}^N \left\| E_l \begin{pmatrix} 0 \\ A^\alpha b_i \end{pmatrix} \right\|^2 \right)^{1/2} \\ &\quad \cdot \left(\sum_{j=1}^N \left\| E_k \begin{pmatrix} 0 \\ A^{1/2} \hat{g}_j \end{pmatrix} \right\|^2 \right)^{1/2} \|E_l x\| \cdot \|E_k y\|. \end{aligned}$$

(iii) Using the transpose of (18) at the third step below, using (15) at the next to the last below and using (H_1) , (H_2) at the last step below, we get

$$\begin{aligned}
 & \left| \frac{1}{n!} \frac{d^n}{d\lambda^n} \left\langle \frac{(1+T_l)^{-1}}{\lambda_l - \lambda_k} \cdot \sum_{\substack{\nu=-\infty \\ \nu \neq l}}^{\infty} \frac{\tilde{\mathcal{C}}_\nu \hat{\mathcal{C}}_\nu^T}{(\lambda_l - \lambda_\nu)(\lambda_l - \lambda_\nu)} \right. \right. \\
 & \qquad \qquad \qquad (1+W(\lambda))^{-1} \left\langle E_l x, E_l \begin{pmatrix} 0 \\ b_1 \\ \vdots \\ 0 \\ b_N \end{pmatrix} \right\rangle, \left. \left\langle E_k y, E_k \begin{pmatrix} 0 \\ A \hat{g}_1 \\ \vdots \\ 0 \\ A \hat{g}_N \end{pmatrix} \right\rangle \right\rangle_{\mathbb{C}^N} \\
 & = \left| \frac{1}{n!} \frac{d^n}{d\lambda^n} \sum_{\substack{\nu=-\infty \\ \nu \neq l}}^{\infty} \left\langle \frac{\hat{\mathcal{C}}_\nu^T}{\lambda - \lambda_\nu} (1+W(\lambda))^{-1} \left\langle E_l x, E_l \begin{pmatrix} 0 \\ A^\alpha b_1 \\ \vdots \\ 0 \\ A^\alpha b_N \end{pmatrix} \right\rangle \right. \right. \\
 & \qquad \qquad \qquad \frac{|\lambda_k| \hat{\mathcal{C}}_\nu^T}{|\lambda_l|^{2\alpha} (\bar{\lambda}_l - \bar{\lambda}_k)(\bar{\lambda}_l - \bar{\lambda}_\nu)} (1+T_l^*)^{-1} \left. \left\langle E_k y, E_k \begin{pmatrix} 0 \\ A^{1/2} \hat{g}_1 \\ \vdots \\ 0 \\ A^{1/2} \hat{g}_N \end{pmatrix} \right\rangle \right\rangle_{\mathbb{C}^N} \\
 & \leq \left(\sum_{\substack{\nu=-\infty \\ \nu \neq l}}^{\infty} \left\| \frac{1}{n!} \frac{d^n}{d\lambda^n} \frac{\hat{\mathcal{C}}_\nu^T}{\lambda - \lambda_\nu} (1+W(\lambda))^{-1} \left\langle E_l x, E_l \begin{pmatrix} 0 \\ A^\alpha b_1 \\ \vdots \\ 0 \\ A^\alpha b_N \end{pmatrix} \right\rangle \right\|_{\mathbb{C}^N}^2 \right)^{1/2} \\
 & \quad \cdot \left(\sum_{\substack{\nu=-\infty \\ \nu \neq l}}^{\infty} \left\| \frac{|\lambda_k| \hat{\mathcal{C}}_\nu \cdot (I+T_l^*)^{-1}}{|\lambda_l|^{2\alpha} (\lambda_l - \lambda_k)(\lambda_l - \lambda_\nu)} \left\langle E_k y, E_k \begin{pmatrix} 0 \\ A^{1/2} \hat{g}_1 \\ \vdots \\ 0 \\ A^{1/2} \hat{g}_N \end{pmatrix} \right\rangle \right\|_{\mathbb{C}^N}^2 \right)^{1/2}
 \end{aligned}$$

$$\begin{aligned}
& \leq \frac{1}{\lambda^{n+1}} \sum_{l=1}^N (\|C_l\|^2)^{1/2} \left\| \left\langle E_l x, E_l \begin{pmatrix} 0 \\ A^\alpha b_1 \\ \vdots \\ 0 \\ A^\alpha b_N \end{pmatrix} \right\rangle \right\|_{\mathbb{C}^N} \left(\sum_{\substack{\nu=-\infty \\ \nu \neq l}}^{\infty} \left\| \frac{\hat{C}_\nu^T}{\lambda_l - \lambda_\nu} \right\|^2 \right)^{1/2} \\
& \quad \cdot \frac{\lambda_k}{|\lambda_l|^{2\alpha} (\lambda_l - \lambda_k)} \left\| \left\langle E_k y, E_k \begin{pmatrix} 0 \\ A^{1/2} \hat{g}_1 \\ \vdots \\ 0 \\ A^{1/2} \hat{g}_N \end{pmatrix} \right\rangle \right\|_{\mathbb{C}^N} \\
& \leq \frac{1}{\lambda^{n+1}} \left(\sum_{l=1}^N \|C_l\|^2 \right)^{1/2} \cdot \left(\sum_{i=1}^N \left\| E_i \begin{pmatrix} 0 \\ A^\alpha b_i \end{pmatrix} \right\|^2 \right)^{1/2} \|E_l x\| \\
& \quad \sum_{\substack{\nu=-\infty \\ \nu \neq l}}^{\infty} \frac{\left(E_\nu \begin{pmatrix} 0 \\ C_1 \end{pmatrix}, \dots, E_\nu \begin{pmatrix} 0 \\ C_N \end{pmatrix} \right)^2}{(\lambda_l - \lambda_\nu)} \cdot \left(\sum_{i=1}^N \left\| E_k \begin{pmatrix} 0 \\ A^{1/2} \hat{g}_i \end{pmatrix} \right\|^2 \right)^{1/2} \\
& \quad \cdot \|E_k y\| \sup_{\forall l \neq k} \left| \frac{\lambda_k}{\lambda^{2\alpha} (\lambda_l - \lambda_k)} \right| \\
& = \frac{1}{\lambda^{n+1}} [\text{the other factors}] \\
& \quad \cdot \left(\sum_{\substack{\nu=-\infty \\ \nu \neq l}}^{\infty} \left\| \frac{\left(E_\nu \begin{pmatrix} 0 \\ A^r C_1 \end{pmatrix}, \dots, E_\nu \begin{pmatrix} 0 \\ A^r C_N \end{pmatrix} \right)}{|\lambda_\nu|^{2r} (\lambda_l - \lambda_\nu)} \right\|^2 \right)^{1/2} \\
& \leq \frac{1}{\lambda^{n+1}} [\text{as above}] \left(\inf_{\forall l \neq \nu} |\lambda_\nu|^{2r} (\lambda_l - \lambda_\nu) \right)^{-1} \\
& \quad \cdot \left(\sum_{\substack{\nu=-\infty \\ \nu \neq l}}^{\infty} \left\| \left(E_\nu \begin{pmatrix} 0 \\ A^r C_1 \end{pmatrix}, \dots, E_\nu \begin{pmatrix} 0 \\ A^r C_N \end{pmatrix} \right) \right\|^2 \right)^{1/2} \\
& \leq \frac{1}{\lambda^{n+1}} [\text{as above}] \left(\inf_{\forall l \neq \nu} |\lambda_\nu|^{2r} (\lambda_l - \lambda_\nu) \right)^{-1} \left(\sum_{j=1}^N \left\| \begin{pmatrix} 0 \\ A^r C_j \end{pmatrix} \right\|^2 \right)^{1/2} \\
& = \frac{1}{\lambda^{n+1}} [\text{as above}] \left(\inf_{\forall l \neq \nu} |\lambda_\nu|^{2r} (\lambda_l - \lambda_\nu) \right)^{-1} \left(\sum_{j=1}^N \|A^r C_j\|^2 \right)^{1/2} \\
& = \frac{1}{\lambda^{n+1}} [\text{as above}] \left(\inf_{\forall l \neq \nu} |\lambda_\nu|^{2r} (\lambda_l - \lambda_\nu) \right)^{-1} \left(\sum_{j=1}^N \langle A^{1/2} \hat{g}_j, A^{(1/2)+r} b_j \rangle \right)^{1/2} \\
& \leq \frac{M_1}{\lambda^{n+1}} \left(\sum_{i=1}^N \left\| E_i \begin{pmatrix} 0 \\ A^\alpha b_i \end{pmatrix} \right\|^2 \right)^{1/2} \cdot \left(\sum_{i=1}^N \left\| E_k \begin{pmatrix} 0 \\ A^{1/2} \hat{g}_i \end{pmatrix} \right\|^2 \right)^{1/2} \|E_l x\| \cdot \|E_k y\|,
\end{aligned}$$

where M_1 is a constant independent of λ (>0) and n .

Putting (i)–(iii) into (26), we obtain

$$\begin{aligned}
 & \left| \frac{1}{n!} \frac{d^n}{d\lambda^n} \left\{ \frac{\Pi_{k,l}}{(\lambda_l - \lambda_k)(\lambda - \lambda_l)} \right\} \right| \\
 (29) \quad & \leq \frac{M_2}{\lambda^{n+1}} \left(\sum_{i=1}^N \left\| E_l \begin{pmatrix} 0 \\ A^{\alpha} b_i \end{pmatrix} \right\|^2 \right)^{1/2} \cdot \left(\sum_{i=1}^N \left\| E_k \begin{pmatrix} 0 \\ A^{1/2} \hat{g}_i \end{pmatrix} \right\|^2 \right)^{1/2} \cdot \|E_l x\| \cdot \|E_k y\|, \\
 & \lambda > 0, \quad n \geq 0,
 \end{aligned}$$

where M_2 is a constant independent of λ (>0) and n (≥ 0 , integer).

Step 2. The estimate of any term in the second summation of the rightmost side of (23).

It follows immediately from (28) and (SH₂) that

$$\begin{aligned}
 (30) \quad & \left| \frac{1}{n!} \frac{d^n}{d\lambda^n} \left\{ \frac{1}{(\lambda - \lambda_l)^2} \left\langle (1 + W(\lambda))^{-1} \left\langle \varepsilon_l \Lambda_l^{1/2} E_l x, E_l \begin{pmatrix} 0 \\ C_1 \\ \vdots \\ 0 \\ C_N \end{pmatrix} \right\rangle \right\rangle \right. \right. \\
 & \left. \left. \left\langle \varepsilon_l^{-1} \Lambda_l^{-1/2} E_l y, E_l \begin{pmatrix} 0 \\ C_1 \\ \vdots \\ 0 \\ C_N \end{pmatrix} \right\rangle \right\rangle \right| \\
 & \leq \frac{2}{\lambda^{n+1}} \|\varepsilon_l \Lambda_l^{1/2} E_l x\| \cdot \|\varepsilon_l^{-1} \Lambda_l^{-1/2} E_l y\| \leq \frac{M_3}{\lambda^{n+1}} \|E_l x\| \cdot \|E_l y\|, \quad \lambda > 0, \quad n \geq 0,
 \end{aligned}$$

where M_3 is a constant independent of λ (>0) and n (≥ 0).

The estimate of the first term of the rightmost side of (23) is trivial.

Putting (29), (30) into (23) and using (H₂) we can easily verify that

$$\begin{aligned}
 (31) \quad & \left| \frac{1}{n!} \frac{d^n}{d\lambda^n} \langle R(\lambda; \mathcal{A})x, y \rangle \right| \\
 & \leq \frac{\|x\| \cdot \|y\|}{\lambda^{n+1}} + \frac{M_3}{\lambda^{n+1}} \sum_{l=1}^{\infty} \|E_l x\| \cdot \|E_l y\| \\
 & \quad + \frac{M_2}{\lambda^{n+1}} \sum_{l \neq k}^{\infty} \left(\sum_{i=1}^N \left\| E_l \begin{pmatrix} 0 \\ A^{\alpha} b_i \end{pmatrix} \right\|^2 \right)^{1/2} \cdot \left(\sum_{i=1}^N \left\| E_k \begin{pmatrix} 0 \\ A^{1/2} \hat{g}_i \end{pmatrix} \right\|^2 \right)^{1/2} \\
 & \quad \cdot \|E_l x\| \cdot \|E_k y\| \leq \frac{M}{\lambda^{n+1}} \|x\| \cdot \|y\|, \quad \lambda > 0, \quad n \geq 0, \quad x, y \in H \times H,
 \end{aligned}$$

where M is a constant independent of λ (>0) and n (≥ 0 , integer).

By well-known results (cf. [10]), (31) implies

$$(32) \quad \|e^{\mathcal{A}t}\| \leq M \quad \forall t \geq 0.$$

This completes the proof of the lemma.

Our main theorem is the following.

THEOREM 1. *Under the assumptions (H_1) , (H_2) and (SH_2) ,*

$$\|e^{\mathcal{A}t}x\| \rightarrow 0 \quad (\text{as } t \rightarrow \infty) \quad \forall x \in H \times H.$$

Proof. Let [12]

$$H_\Lambda = \{x \in H \times H \mid \|x\|_{H_\Lambda}^2 = \langle x, x \rangle + \langle \Lambda_0 x, \Lambda_0 x \rangle \text{ is finite.}\}$$

By the definition of Λ_0 , it is easy to see that $\mathcal{D}(\Lambda_0)$ is dense in $H \times H$. Hence H_Λ is dense. Also, H_Λ is a Hilbert space because Λ_0 is positive, hence, closed (cf. [10, pp. 196–197]). On H_Λ we have

$$\begin{aligned} \|e^{\mathcal{A}t}x\|_{H_\Lambda}^2 &= \|e^{\mathcal{A}t}x\|^2 + \langle \Lambda_0 e^{\mathcal{A}t}x, \Lambda_0 e^{\mathcal{A}t}x \rangle \\ &= \|e^{\mathcal{A}t}x\|^2 + \langle \Lambda_0 e^{\mathcal{A}t} \Lambda_0^{-1} \Lambda_0 x, \Lambda_0 e^{\mathcal{A}t} \Lambda_0^{-1} \Lambda_0 x \rangle \\ &= \|e^{\mathcal{A}t}x\|^2 + \|e^{\mathcal{A}C^t} \Lambda_0 x\|^2. \end{aligned}$$

Hence, by Lemma 4 and noticing $e^{\mathcal{A}C^t}$ being a contraction semigroup, we obtain

$$(33) \quad \|e^{\mathcal{A}t}x\|_{H_\Lambda}^2 \leq M^2 \|x\|^2 + \|\Lambda_0 x\|^2 \leq (1 + M^2) \|x\|_{H_\Lambda}^2, \quad x \in H_\Lambda.$$

Considering $T(t) = e^{\mathcal{A}t}$ as a semigroup on H_Λ , from (33) we have (cf. [14, Chap. IX, § 3, Thm. 1]) that $H_\Lambda \cap \mathcal{D}(\mathcal{A}) = \{x \in H_\Lambda \mid \lim_{h \downarrow 0} (1/h)(T(h) - I)x \text{ exists in } H_\Lambda\}$ is dense in H_Λ . Hence, $H_\Lambda \cap \mathcal{D}(\mathcal{A})$ is also dense in $H \times H$.

Let $V(x) = \|\Lambda_0 x\|^2$ on H_Λ . It is a Lyapunov function [3] because

$$\begin{aligned} \dot{V}(x) &= \lim_{t \rightarrow 0^+} \frac{1}{t} (V(e^{\mathcal{A}t}x) - V(x)) \\ &= \langle \Lambda_0 \mathcal{A}x, \Lambda_0 x \rangle - \langle \Lambda_0 x, \Lambda_0 \mathcal{A}x \rangle \\ (34) \quad &= \langle \mathcal{A}_C \Lambda_0 x, \Lambda_0 x \rangle - \langle \Lambda_0 x, \mathcal{A}_C \Lambda_0 x \rangle \\ &= 2 \operatorname{Re} \langle \mathcal{A}_C \Lambda_0 x, \Lambda_0 x \rangle \leq 0 \quad \forall x \in H_\Lambda \cap \mathcal{D}(\mathcal{A}). \end{aligned}$$

Lemma 4 and (34) guarantee the existence of an invariant set of the dynamical system $T(t)x$ ($x \in H_\Lambda \cap \mathcal{D}(\mathcal{A})$). Assumption (H_2) then implies that the singleton $\{0\}$ is the unique invariant set of $T(t)x$, $x \in H_\Lambda \cap \mathcal{D}(\mathcal{A})$. Therefore, by the standard technique (cf. [9]–[12]), we get

$$\|e^{\mathcal{A}t}x\|_{H_\Lambda} \rightarrow 0 \quad (\text{as } t \rightarrow \infty) \quad \forall x \in H_\Lambda \cap \mathcal{D}(\mathcal{A}).$$

In particular $\|e^{\mathcal{A}t}x\| \rightarrow 0$ (as $t \rightarrow \infty$) $\forall x \in H_\Lambda \cap \mathcal{D}(\mathcal{A})$. Lemma 4 then implies $\|e^{\mathcal{A}t}x\| \rightarrow 0$ (as $t \rightarrow \infty$) $\forall x \in H \times H$.

Theorem 1 is thus proved.

3. Applications. We now apply the theory given in § 2 to the concrete boundary stabilization problems given in § 1.

First we consider the system (5) with the feedback (6). It is known from [2] that the corresponding operator A (see (5')) has simple point spectrum $\{\mu_n\}_1^\infty$ with $\mu_n = O(n^2)$ (n large). Let $\{\phi_n\}_1^\infty$ be the corresponding system of orthonormal eigenvectors

of A . It can be easily verified that

$$\langle g, \phi_n \rangle = -\frac{1}{\mu_n} \phi_n(l) \quad \forall n \geq 1,$$

also $\phi_n(l) \neq 0 \quad \forall n \geq 1$.

The assumption (H_1) is valid if we choose $\alpha = \frac{1}{2}$ and $r = 0$. The assumption (H_2) is reduced to

$$(35) \quad \langle b, \phi_n \rangle \phi_n(l) < 0.$$

The assumption (SH_2) is satisfied automatically in the present case.

CONCLUSION. Under the conditions, (35) and $A^{1/2}b \in L^2(0, l)$ holds. Then for any $y_1 \in L^2(0, l)$ and $y_0 \in \mathcal{D}(A^{1/2}) = H_0^1(0, l)$ the solution of (5) together with (6) is strongly stable.

This is just a consequence of Theorem 1.

Now we go back to the boundary stabilization of the hyperbolic system (1) in $\Omega \subset \mathbb{R}^2$ with the feedback (2).

We assume that the corresponding elliptic operator A (see (1')) has the property

$$(36) \quad \dim F_k \leq N \quad \forall k \leq 1.$$

By well-known results, we have (cf. [2])

$$\nu_n = O(n), \quad n \text{ large.}$$

Hence assumption (H_1) is satisfied if we choose $\alpha = 1$ and $r = \frac{1}{2}$.

The assumption (H_2) is then in the following form:

$$(37_1) \quad \text{rank}(\langle F_l \hat{g}_i, F_l b_i \rangle)_{N \times N} = \text{rank } F_l, \quad l \geq 1,$$

$$\begin{aligned} \langle F_l \hat{g}_i, F_l b_j \rangle)_{N \times N} &= \frac{1}{\nu_l} \langle A \hat{g}_i, F_l b_j \rangle \\ &= \frac{1}{\nu_l} \left(\left\langle A \hat{g}_i, \sum_{k=1}^{\dim F_l} \langle b_j, \phi_{lk} \rangle \phi_{lk} \right\rangle \right) \\ (37_2) \quad &= \left(\sum_{k=1}^{\dim F_l} \langle \phi_b, b_j \rangle \int_{\Gamma_i} g_i \bar{\phi}_{lk} d\Gamma \right) \\ &= \left(\int_{\Gamma_i} g_i \phi_{lk} d\Gamma \right)_{N \times \dim F_l} \cdot \left(\int_{\Omega} \phi_{lk} \bar{b}_j d\Omega \right)_{\dim F_l \times N} \\ &\geq 0 \quad \forall l \geq 1, \end{aligned}$$

where $\{\phi_{lk}\} \quad 1 \leq k \leq \dim F_l$ is the orthonormal basis of $F_l\{L^2(\Omega)\}$. In the present case, the assumption (SH_2) reduces to the assumption

$$(38) \quad \sup_{k \geq 1} \frac{\max \{\lambda : \lambda \in \sigma(\Lambda_k)\}}{\min \{\lambda : \lambda \in \sigma(\Lambda_k)\}} < \infty,$$

where

$$\Lambda_k : \frac{1}{\nu_k} \begin{pmatrix} \int_{\Gamma_i} g_i \bar{\phi}_{k,1} d\Gamma \\ \vdots \\ \int_{\Gamma_i} g_i \bar{\phi}_{k, \dim F_k} d\Gamma \end{pmatrix} \rightarrow \begin{pmatrix} \int_{\Omega} b_i \bar{\phi}_{k,1} d\Omega \\ \vdots \\ \int_{\Omega} b_i \bar{\phi}_{k, \dim F_k} d\Omega \end{pmatrix} \quad 1 \leq i \leq \dim F_k, \quad k \geq 1.$$

CONCLUSION. Under the condition (36)–(38) and $b_i \in \mathcal{D}(A^{3/2})$, $1 \leq i \leq N$, the solution of the system (1) together with (2) is strongly stable for any $W_1 \in L^2(\Omega)$ and $W_0 \in H_{\Gamma_0}^1$.

Remark 1. A natural question is whether there exist b_i 's and g_i 's such that the assumptions (H₁), (H₂) and (SH₂) hold. For the system (5) with the feedback (6), the fact that $\phi_n(l) \neq 0$, $n \geq 1$, implies the existence of b satisfying all the assumptions we required. For the system (1) with the feedback (2), it follows immediately from assumption (H₂) that

$$(39) \quad \text{rank} \left(\int_{\Gamma_i} g_i \bar{\phi}_{l,k} d\Gamma \right)_{N \times \dim F_l} = \dim F_l, \quad l \geq 1.$$

The condition (39) is indeed, in a sense, equivalent to the controllability of the system (1) or the corresponding parabolic system. We are not going into the details; however, some information can be obtained in [13] or [15]. Hence, it is natural to assume that (39) holds.

CONCLUSION. If (39) holds with $g_i \in H^{1/2}(\Gamma_i)$, $1 \leq i \leq N$, then there always exist $\{b_i\}_1^N$ such that assumptions (H₂) and (SH₂) (or equivalently, the assumptions (37) and (38)) hold.

Proof. Let

$$(40) \quad B_l = \frac{\varepsilon_l}{\nu_l} \left(\int_{\Gamma_i} g_i \bar{\phi}_{l,k} d\Gamma \right)_{N \times \dim F_l}, \quad l \geq 1,$$

where $\varepsilon_l > 0$ ($l \geq 1$), are sufficiently small. Also, let

$$(41) \quad b_k = \sum_{\substack{1 \leq j \leq \dim F_l \\ l \geq 1}} \left(\frac{\varepsilon_l}{\nu_l} \int_{\Gamma_k} g_k \bar{\phi}_{l,j} d\Gamma \right) \phi_{l,j}, \quad 1 \leq k \leq N.$$

Then it is easy to verify that the corresponding operator Λ_k (see (38)) is equal to $\varepsilon_k^{-1} I_k$, where I_k is the identity operator in $F_k L^2(\Omega)$. Hence, whatever $\varepsilon_l > 0$ we choose, we always have

$$(42) \quad \sup_{k \geq 1} \frac{\max \{ \lambda : \lambda \in \sigma(\Lambda_k) \}}{\min \{ \lambda : \lambda \in \sigma(\Lambda_k) \}} = 1.$$

This implies (38). Moreover, for any $\alpha > 0$, using the trace theorem (cf. [5], p. 39) we have

$$(43) \quad \begin{aligned} \|A^\alpha b_k\|^2 &= \sum_{\substack{1 \leq j \leq \dim F_l \\ l \geq 1}} \left| \varepsilon_l \nu_l^{\alpha-1} \int_{\Gamma_k} g_k \cdot \bar{\phi}_{l,j} d\Gamma \right|^2 \\ &\leq \sum_{\substack{1 \leq j \leq \dim F_l \\ l \geq 1}} |\varepsilon_l \nu_l^{\alpha-1}|^2 \|g_k\|_{L^2(\Gamma_k)}^2 \cdot \|\phi_{l,j}\|_{L^2(\Gamma_k)}^2 \\ &\leq C_1 \sum_{\substack{1 \leq j \leq \dim F_l \\ l \geq 1}} |\varepsilon_l \nu_l^{\alpha-1}|^2 \|g_k\|_{L^2(\Gamma_k)}^2 \cdot \|\phi_{l,j}\|_{H^{1/2}(\Gamma_k)}^2 \\ &\leq C_2 \cdot \|g_k\|_{L^2(\Gamma_k)}^2 \cdot \sum_{\substack{1 \leq j \leq \dim F_l \\ l \geq 1}} |\varepsilon_l \nu_l^{\alpha-1}|^2 \cdot \|\phi_{l,j}\|_{H_{\Gamma_0}^1(\Omega)}^2 \\ &\leq C_3 \|g_k\|_{L^2(\Gamma_k)}^2 \cdot \sum_{\substack{1 \leq j \leq \dim F_l \\ l \geq 1}} |\varepsilon_l \nu_l^{\alpha-1}|^2 \|A^{1/2} \phi_{l,j}\|_{L^2(\Omega)}^2 \\ &= C_3 \|g_k\|_{L^2(\Gamma_k)}^2 \cdot \sum_{\substack{1 \leq j \leq \dim F_l \\ l \geq 1}} \varepsilon_l^2 \nu_l^{2\alpha-1} < \infty, \quad 1 \leq k \leq N \end{aligned}$$

whenever we choose $\varepsilon_l > 0$ such that

$$(44) \quad \sum_{l \geq 1} \varepsilon_l^2 \nu_l^{2\alpha-1} < \infty,$$

where C_1 , C_2 and C_3 are constant. For any given $\alpha > 0$, the existence of $\{\varepsilon_l\}_1^\infty$ satisfying (44) is trivial. Furthermore, (37) holds automatically by virtue of (39)–(40) (also (42)).

Our conclusion is thus proved.

Remark 2. Another natural question is whether the semigroup given in Lemma 4 is still a contraction semigroup. In general, the answer is NO. Indeed, for the system (5) with feedback (6), the following formal calculation can be fully justified.

$$(45) \quad \operatorname{Re} \langle \mathcal{A}X, X \rangle = \operatorname{Re} \left\{ \left\langle \begin{pmatrix} 0 \\ Ag \end{pmatrix}, X \right\rangle \cdot \left\langle X, \begin{pmatrix} 0 \\ b \end{pmatrix} \right\rangle \right\} = \operatorname{Re} \{ \langle x_2, b \rangle \cdot \overline{\langle x_2, l \rangle} \}$$

$$\forall x_2 \in H^1(0, l), \quad X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathcal{D}(\mathcal{A}).$$

Hence, $\operatorname{Re} \langle \mathcal{A}X, X \rangle \leq 0$ would imply $b = \varepsilon Ag$, where $\varepsilon > 0$ is a constant. But Ag is only in $(H_0^1(0, l))'$ and not in $L^2(0, l)$. This contradicts $b \in \mathcal{D}(A^{1/2})$.

Acknowledgments. The author wishes to thank his colleague Dr. Wang Falun for the proofs of Lemma 3 and Lemma 4, which are modified versions of those given by Wang in an unpublished paper. The author is especially indebted very much to Professor L. Berkovitz for his kind suggestions and help on writing this version of this paper. The author would also like to thank the referees for many important suggestions and help on this paper.

REFERENCES

- [1] G. CHEN, *Energy decay estimates and exact boundary value controllability for the wave equation in a bounded domain*, J. Math. Pure et Appl., 58 (1979), pp. 249–273.
- [2] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, Vol. I, John Wiley, New York, 1953.
- [3] J. K. HALE, *Dynamical systems and stability*, J. Math. Anal. Appl., 26 (1969), pp. 39–59.
- [4] CHAO-CHIH KWAN AND KANG-NING WANG, *Sur la stabilisation de la vibration élastique*, Scientia Sinica, 17 (1974), pp. 446–467.
- [5] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, I, Springer-Verlag, New York, 1972.
- [6] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.
- [7] D. L. RUSSELL, *Linear stabilization of the linear oscillator in Hilbert space*, J. Math. Anal. Appl., 25 (1969), pp. 663–675.
- [8] ———, *Controllability and stability theory for linear partial differential equations: Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 630–739.
- [9] M. SLEMROD, *Stabilization of boundary control systems*, J. Differential Equations, 22 (1976), pp. 402–415.
- [10] ———, *A note on complete controllability and stability for linear control system in Hilbert space*, SIAM J. Control, 12 (1974), pp. 500–508.
- [11] SHUNHUA SUN, *A stabilization problem of linear control system in Hilbert space*, J. Sichuan University, 1 (1975), pp. 51–56.
- [12] ———, *A note on stabilization of linear control systems in Hilbert space*, Acta Mathematica Sinica, 18 (1975), pp. 297–299.
- [13] ———, *Boundary observability and boundary controllability of parabolic equations*, J. Sichuan University, 3 (1979), pp. 31–60.
- [14] K. YOSIDA, *Functional Analysis*, Springer-Verlag, New York, 1965.
- [15] SAKAWA YOSIYUKI, *Observability and related problems for partial differential equations of parabolic type*, SIAM J. Control, 12 (1974), pp. 389–400.

ASYMPTOTIC PROPERTIES OF A FINITE STATE CONTINUOUS TIME MARKOV DECISION PROCESS*

GERD RODÉ†

Abstract. We consider a continuous time Markov decision process with a finite state space. There is a specified terminal reward, but the reward or cost rate is always zero. The maximum expected final gain can then be obtained by means of the exponential of a certain sublinear operator on R^n . This representation allows us to describe the asymptotic properties of the reward vector. We prove that the expected reward always tends to a limit as the time parameter $t \rightarrow \infty$. If we assume that it is allowed to stop the process in any state, then we can construct an almost optimal stationary control. Finally, we characterize the case where the asymptotic gain is independent of the initial state.

Key words. Markov decision processes, nonlinear semigroups

1. Introduction. The theory of Markov decision processes with continuous time and finite state space (see, e.g., [2], [3], [4]) is formally identical with the discussion of the Bellman differential equation [1]

$$\frac{d}{dt}x(t) = \sup_{u \in U} (A^u x(t) + a^u) \quad \forall t \geq 0,$$

where, for $u \in U$, $A^u: R^n \rightarrow R^n$ is a linear operator generating a semigroup $\{\exp(tA^u)\}_{t \geq 0}$ of Markov operators on R^n and where $a^u \in R^n$ describes the reward rate of the process specific for the strategy $u \in U$.

In this note, we discuss the special case $a^u = 0$ for all $u \in U$. This means that our only goal is to reach a good state at the end of the control time; any costs or rewards during the control are out of consideration.

Our main results (Theorems 1, 2, and 3 in § 4.) describe the asymptotic properties of the final gain $x(t)$ as $t \rightarrow \infty$.

2. Notation and definitions. A mapping $F: R^n \rightarrow R^n$ with

$$F(x + y) \leq F(x) + F(y) \quad \forall x, y \in R^n$$

and

$$F(cx) = cF(x) \quad \forall x \in R^n, \quad c \geq 0,$$

is called a sublinear operator on R^n . We shall need three special sublinear operators:

$$\text{Id: Id}(x) = x \quad \forall x \in R^n,$$

$$\text{Max: Max}(x) = \max_{1 \leq k \leq n} x_k \quad \forall x \in R^n,$$

$$\text{Gen: Gen}(x) = \text{Max}(x) - x \quad \forall x \in R^n.$$

A sublinear operator F on R^n is called a Markov operator, if it fulfills

$$x \leq y \Rightarrow F(x) \leq F(y) \quad \forall x, y \in R^n$$

and

$$F(c) = c \quad \forall \text{ constant } c \in R^n.$$

* Received by the editors March 3, 1981, and in revised form January 12, 1982.

† Kraichgastrasse 4, 6908 Wiesloch, West Germany.

F is called a Markov generator if it satisfies the condition

$$x \in R^n, 1 \leq k \leq n, x_k = \max_{1 \leq i \leq n} x_i \Rightarrow (F(x))_k \leq 0.$$

It is easy to see that a sublinear operator F on R^n is a Markov operator if and only if

$$F(x) \leq \text{Max}(x) \quad \forall x \in R^n$$

and that F is a Markov generator if and only if there exists a $c \geq 0$ such that

$$F(x) \leq c \text{ Gen}(x) \quad \forall x \in R^n.$$

Finally, a continuous semigroup on R^n is a family $\{P(t)\}_{t \geq 0}$ of mappings $P(t): R^n \rightarrow R^n$ such that

- 1) $P(s+t) = P(s)P(t) \quad \forall s, t > 0$ and $P(0) = \text{Id}$,
- 2) $P(\cdot)x: R^+ \rightarrow R^n$ is continuous $\forall x \in R^n$.

3. Semigroups of sublinear Markov operators. It is a standard fact that each Markov process with continuous time and finite state space $\{1, \dots, n\}$ can be described by a semigroup of linear Markov operators on R^n . Replacing the linearity by sublinearity, we obtain the analogous description of Markov decision processes.

The basic analytic properties of these sublinear semigroups are contained in the following proposition:

PROPOSITION 1. (a) Let θ be a sublinear Markov generator on R^n . The following definitions of the exponential of θ , $\exp(\theta): R^n \rightarrow R^n$, are equivalent:

$$(D1) \quad \exp(\theta)x = \sup \{ \exp(p_1 A^1) \cdots \exp(p_m A^m)x \mid m \in N, p_1, \dots, p_m \geq 0, \\ \sum p_k = 1, A^k: R^n \rightarrow R^n \text{ linear with } A^k \leq \theta \quad \forall 1 \leq k \leq m \}.$$

$$(D2) \quad \exp(\theta)x = \lim_{m \rightarrow \infty} (\text{Id} + \theta/m)^m x.$$

$$(D3) \quad \exp(\theta)x = \lim_{m \rightarrow \infty} (\text{Id} - \theta/m)^{-m} x.$$

$$(D4) \quad \text{Let } x(t) \in R^n, t \geq 0, \text{ be the solution of the differential equation}$$

$$\frac{d}{dt} x(t) = \theta x(t) \quad \forall t \geq 0, \quad x(0) = x,$$

$$\text{and define } \exp(\theta)x = x(1).$$

The mappings $\{\exp(t\theta)\}_{t \geq 0}$ form a continuous semigroup of sublinear Markov operators on R^n .

(b) Let $\{P(t)\}_{t \geq 0}$ be a continuous semigroup of sublinear Markov operators on R^n . Then the infinitesimal generator θ ,

$$\theta x = \lim_{t \rightarrow 0+} \frac{1}{t} (P(t)x - x) \quad \forall x \in R^n,$$

is well defined on R^n . θ is a sublinear Markov generator. The semigroup fulfills the differential equation

$$\frac{d}{dt} P(t)x = \theta P(t)x \quad \forall x \in R^n, \quad t \geq 0.$$

(c) Each continuous semigroup of sublinear Markov operators on R^n is the exponential (as defined in (a)) of its infinitesimal generator (as defined in (b)).

Proof. (a) The existence of the limit in (D2) and the uniqueness of the solution of the differential equation in (D4) follow from the fact that θ is Lipschitz on R^n since it is sublinear. The existence of $(\text{Id} - (\theta/m))^{-1}$ and of the limit in (D3) can be deduced from the general theory of dissipative operators in Banach spaces (see, e.g., [8]), since each sublinear Markov generator is dissipative with respect to the norm

$$\|\cdot\|_\infty: R^n \rightarrow R, \quad \|x\|_\infty = \max_{1 \leq k \leq n} |x_k|.$$

It is a standard result in the theory of ordinary differential equations that (D2) and (D3) define the value $x(1)$ of the solution $x(t)$, $t \geq 0$, of the differential equation in (D4). Hence (D2), (D3) and (D4) are equivalent. The equivalence of (D1) and (D4) can be shown directly and is essentially contained in [5], [6] or [7].

(b) This is a special case of the following result (see [9]): Each continuous semigroup of convex operators on R^n is continuously differentiable with respect to the parameter t .

(c) From (D4) and the uniqueness theorem for differential equations.

Assume now that a Markov decision process is given by a collection of linear transition semigroups on R^n . Let A^U be the corresponding set of infinitesimal generators. We assume

(A1) A^U is a compact set of linear operators on R^n .

The operator

$$\theta: R^n \rightarrow R^n, \quad \theta x = \sup_{A \in A^U} Ax$$

is the supremum of linear Markov generators and hence a sublinear Markov generator. From Proposition 1 (a), we obtain the semigroup $P(t) = \exp(t\theta)$, $t \geq 0$. This semigroup describes the maximum expected gain of the Markov decision process (see [5], [6] and [7] for more details).

It is in general not possible to rediscover the set A^U from θ . In the case of decision processes, however, it is reasonable to assume

(A2) $A, B \in A^U, p \in R^n, 0 \leq p \leq 1 \Rightarrow (1-p)A + pB \in A^U$.

We then have, from the bipolar theorem,

$$A^U = \{A: R^n \rightarrow R^n \text{ linear} \mid A \leq \theta \text{ on } R^n\}.$$

4. Asymptotic properties of sublinear Markov semigroups. Let $\{P(t)\}_{t \geq 0}$ be a continuous semigroup of sublinear Markov operators on R^n . From Proposition 1 we obtain the representation $P(t) = \exp(t\theta)$ for all $t \geq 0$. We have the following results concerning the behavior of $P(t)$ as $t \rightarrow \infty$.

THEOREM 1. *The limit*

$$\lim_{t \rightarrow \infty} P(t) =: P(\infty)$$

exists on R^n , uniformly on bounded sets. $P(\infty)$ is a sublinear Markov operator with

$$P(\infty) = P(\infty)P(\infty) = P(t)P(\infty) = P(\infty)P(t) \quad \forall t \geq 0$$

and

$$\theta P(\infty) = 0.$$

THEOREM 2. If $\theta \geq 0$, then for each $a \in R^n$ and each $\varepsilon > 0$ there exists a linear operator $A: R^n \rightarrow R^n$ with $A \leq \theta$ on R^n , such that

$$\lim_{t \rightarrow \infty} \exp(tA)a \geq \lim_{t \rightarrow \infty} \exp(t\theta)a - \varepsilon.$$

Remark. (a) The condition $\theta \geq 0$ is equivalent to $P(t)x \geq P(s)x$ for all $x \in R^n$, $s \leq t$. The interpretation in the decision process context is just the possibility of stopping in any state.

(b) We shall give an example showing that the condition $\theta \geq 0$ cannot be dropped.

THEOREM 3. $P(\infty)a$ is a constant for each $a \in R^n$ if and only if there do not exist two sets $S, T \subset \{1, \dots, n\}$ such that

$$S, T \neq \emptyset, \quad S \cap T = \emptyset, \quad \theta(\chi_S) \geq 0 \quad \text{and} \quad \theta(-\chi_T) \leq 0,$$

where χ_S and χ_T are the characteristic functions of S and T .

The proof of Theorem 1 is based on the following proposition.

PROPOSITION 2. Let $X = \{x \in R^n \mid \|x\|_\infty \leq 1\}$ and assume that $F: X \rightarrow X$ is non-expansive:

$$\|Fx - Fy\|_\infty \leq \|x - y\|_\infty \quad \forall x, y \in X.$$

Let $0 < c < 1$. Then for each starting-value $x^0 \in X$, the inductively defined sequence (x^0, x^1, \dots) ,

$$x^m = (1 - c)x^{m-1} + cF(x^{m-1}) \quad \forall m \in N$$

converges to a fixed-point of F .

Proof of the proposition. We proceed by induction over n . Assume that the proposition holds in R^k for $k \in N$, $k < n$.

Let y^0 be a cluster point of (x^0, x^1, \dots) . Define (y^0, y^1, \dots) by

$$y^m = (1 - c)y^{m-1} + cF(y^{m-1}) \quad \forall m \in N.$$

We first show the convergence of (y^0, y^1, \dots) . Let $z \in X$ be any fixed point of F . (The existence of a fixed point follows from Banach's fixed-point theorem; choose for $0 < t < 1$ a fixed point z_t of the strict contraction tF and let z be a cluster point of z_t as $t \rightarrow 1$.) Since $\|x^m - z\|_\infty$ is monotone decreasing in m and since each y^k is a cluster point of the sequence (x^0, x^1, \dots) , we have

$$\|y^k - z\|_\infty = \lim_{m \rightarrow \infty} \|x^m - z\|_\infty =: r \geq 0 \quad \forall k \in N_0.$$

If $1 \leq i \leq n$ with

$$|y_i^m - z_i| = r$$

for some $m \geq 1$, then

$$|y_i^{m-1} - z_i| = r \quad \text{and} \quad y_i^m = y_i^{m-1},$$

since

$$r = |(1 - c)(y_i^{m-1} - z_i) + c(F(y^{m-1})_{i-z_i})|,$$

with

$$|y_i^{m-1} - z_i| \leq r \quad \text{and} \quad |F(y^{m-1})_{i-z_i}| \leq r.$$

Hence there exists an i , $1 \leq i \leq n$, with $y_i^m = y_i^0$ for all $m \in N$. We may assume $i = n$.

If $n = 1$, then (y^0, y^1, \dots) is constant and hence convergent. If $n > 1$, then we apply the induction hypothesis:

Define

$$W = \{x \in R^{n-1} \mid \|x\|_\infty \leq 1\}, \quad P: W \rightarrow X, \quad (Pw)_i = \begin{cases} w_i & \text{if } 1 \leq i \leq n-1, \\ y_n^0 & \text{if } i = n \end{cases}$$

$$G: W \rightarrow W, \quad (Gw)_i = (F(Pw))_i \quad \forall 1 \leq i \leq n-1, w \in W.$$

Then G is nonexpansive. Choose $w^0 \in W$, with $Pw^0 = y^0$, and define

$$w^m = (1-c)w^{m-1} + cGw^{m-1} \quad \forall m \in N.$$

Then the sequence (w^0, w^1, \dots) converges by the induction hypothesis. This shows that (y^0, y^1, \dots) converges in the first $(n-1)$ coordinates as well, since $Pw^m = y^m$ for all $m \in N$.

Let y^∞ be the limit point of (y^0, y^1, \dots) . From

$$(1-c)y^\infty + cF(y^\infty) = \lim_{m \rightarrow \infty} ((1-c)y^{m-1} + cF(y^{m-1})) = \lim_{m \rightarrow \infty} y^m = y^\infty,$$

we see that y^∞ is a fixed-point of F . Hence, $\|x^m - y^\infty\|_\infty$ is decreasing in m , and

$$\lim_{m \rightarrow \infty} \|x^m - y^\infty\|_\infty = \|y^k - y^\infty\|_\infty \quad \forall k \in N_0.$$

As $k \rightarrow \infty$, we obtain $\lim_{m \rightarrow \infty} \|x^m - y^\infty\|_\infty = 0$.

Proof of Theorem 1. We first show the convergence of $P(mt)a$, $m \in N$, where $t > 0$ is fixed. We may assume $\|a\|_\infty \leq 1$.

Choose $r > 0$ with $\theta \leq r \cdot \text{Gen}$ on R^n and define

$$X = \{x \in R^n \mid \|x\|_\infty \leq 1\},$$

$$c = 1 - \exp(-tr),$$

$$F: X \rightarrow X, \quad F(x) = c^{-1}(P(t)x - (1-c)x).$$

From

$$P(t) = \exp(t\theta) \leq \exp(rt \text{Gen}) = (1 - \exp(-rt)) \text{Max} + \exp(-rt) \text{Id},$$

it is evident that F is the restriction of a sublinear Markov operator on X , especially F is nonexpansive and $F(X) \subset X$. By Proposition 2,

$$\lim_{m \rightarrow \infty} ((1-c) \text{Id} + cF)^m a$$

exists. Since

$$(1-c) \text{Id} + cF = P(t) \quad \text{on } X,$$

this proves that $\lim_{m \rightarrow \infty} P(mt)a$ exists.

Define $a^\infty = \lim_{m \rightarrow \infty} P(m)a$. Then $a^\infty = \lim_{m \rightarrow \infty} P(m/k)a$ for all $k \in N$, since a^∞ is a cluster point of the converging sequence $P(m/k)a$, $m \in N$. This implies

$$P\left(\frac{m}{k}\right)a^\infty = a^\infty \quad \forall m, k \in N.$$

By continuity we then obtain $P(t)a^\infty = a^\infty$ for all $t \geq 0$. Hence $\|P(t)a - a^\infty\|_\infty$ is monotone decreasing in t . Since $\|P(m)a - a^\infty\|_\infty \rightarrow 0$ as $m \rightarrow \infty$, we have

$$P(t)a \rightarrow a^\infty \quad \text{as } t \rightarrow \infty.$$

The pointwise convergence of the operators $P(t)$ as $t \rightarrow \infty$ is clearly equivalent with the uniform convergence on bounded subsets of R^n , since the operators $P(t)$ are nonexpansive.

The stated properties of $P(\infty)$ follow by standard arguments.

In the following example we consider a simple Markov decision process with state space $\{1, 2, 3\}$. This process has the property that no stationary optimal control exists for certain final rewards $x \in R^n$, one has to change the strategy exactly $\log \frac{3}{2}$ time units before the end of the control time. This is a consequence of the fact that there exists a state where one cannot stop the process.

Example. Define a sublinear Markov generator θ on R^3 by

$$\theta(x_1, x_2, x_3) = (2(x_2 - x_1)^+, x_3 - x_2, 0).$$

Then one verifies by differentiation that

$$\begin{aligned} & \exp(t\theta)(0, 1, -1) \\ &= \begin{cases} (4 \exp(-t) - 3 \exp(-2t) - 1, 2 \exp(-t) - 1, 1) & 0 \leq t \leq \log \frac{3}{2} \\ (\frac{1}{3}, 2 \exp(-t) - 1, -1) & \log \frac{3}{2} \leq t. \end{cases} \end{aligned}$$

Hence

$$P(\infty)(0, 1, -1) = (\frac{1}{3}, -1, -1).$$

Each linear $A \leq \theta$ is of the form

$$A^c = (1 - c)A^0 + cA^1, \quad 0 \leq c \leq 1,$$

where

$$A^0(x_1, x_2, x_3) = (0, x_3 - x_2, 0), \quad A^1(x_1, x_2, x_3) = (2(x_2 - x_1), x_3 - x_2, 0).$$

We have

$$\lim_{t \rightarrow \infty} \exp(tA^c)(0, 1, -1) = \begin{cases} (0, -1, -1) & \text{if } c = 0, \\ (-1, -1, -1) & \text{if } 0 < c \leq 1. \end{cases}$$

We prepare the proof of Theorem 2 with the following two lemmas.

LEMMA 1. Let $\theta \geq 0$ be a sublinear Markov generator. Then

$$N(\theta) = \{x \in R^n \mid \theta x = 0\}$$

is a convex cone and

$$\lim_{t \rightarrow \infty} \exp(t\theta)x = \min \{a \in N(\theta) \mid x \leq a\} \quad \forall x \in R^n.$$

Proof. Since θ is sublinear and since $N(\theta) = \{x \mid \theta x \leq 0\}$, $N(\theta)$ is a convex cone in R^n . We have $P(\infty)x \in N(\theta)$ by Theorem 1. If $a \in N(\theta)$ and $x \leq a$, then $P(t)x \leq P(t)a = a$ for all $t \geq 0$, hence $P(\infty)x \leq a$.

LEMMA 2. Let $\theta \geq 0$ be a sublinear Markov generator, $x \in R^n$ and $\varepsilon > 0$. Then there exist finitely many linear operators $A^0, \dots, A^m \leq \theta$ with $A^0 = 0$ such that

$$\lim_{t \rightarrow \infty} \exp(t \max(A^0, \dots, A^m))x \geq \lim_{t \rightarrow \infty} \exp(t\theta)x - \varepsilon.$$

Proof. This can easily be deduced from definition (D1) in Proposition 1.

Proof of Theorem 2. We may assume, using Lemma 2, that θ is the maximum of a finite number of linear operators,

$$\theta = \max(A^0 = 0, A^1, \dots, A^m).$$

We fix i , $1 \leq i \leq n$, and define for $0 \leq k \leq m$

$$\theta^k: R^n \rightarrow R^n, \quad (\theta^k x)_j = \begin{cases} (\theta x)_j & \text{if } j \neq i, \\ (A^k x)_j^+ & \text{if } j = i, \end{cases}$$

$$P^k(t) = \exp(t\theta^k),$$

$$v^k = \min \{v \geq 0 \mid (A^k P^0(\infty)(a + v\chi_i))_i \leq 0\}.$$

Choose K , $1 \leq K \leq m$, such that

$$v^K = \max_{1 \leq k \leq m} v^k.$$

We now prove

$$P^K(\infty)a = P(\infty)a.$$

This is done in two steps:

$$(a) \quad P(\infty)a \leq P^0(\infty)(a + v^K\chi_i),$$

$$(b) \quad P^0(\infty)(a + v^K\chi_i) \leq P^K(\infty)(a).$$

Proof of (a): Since $a \leq P^0(\infty)(a + v^K\chi_i)$, it suffices to show

$$P^0(\infty)(a + v^K\chi_i) \in N(\theta).$$

If $1 \leq j \leq n$ and $j \neq i$, then

$$(\theta P^0(\infty)(a + v^K\chi_i))_j = (\theta^0 P^0(\infty)(a + v^K\chi_i))_j = 0.$$

If $j = i$, then

$$(\theta P^0(\infty)(a + v^K\chi_i))_i = (A^k P^0(\infty)(a + v^K\chi_i))_i$$

for some k , $1 \leq k \leq m$. We have

$$a + v^K\chi_i \leq a + v^k\chi_i + v^K - v^k,$$

hence

$$P^0(\infty)(a + v^K\chi_i) \leq P^0(\infty)(a + v^k\chi_i) + v^K - v^k.$$

The i -components of these two vectors are identical $= a_i + v^K$. From the fact that A^k is a Markov generator, it then follows that

$$(A^k P^0(\infty)(a + v^K\chi_i))_i \leq (A^k P^0(\infty)(a + v^k\chi_i))_i = 0$$

and hence (a).

Proof of (b): Define $v = (P^K(\infty)(a) - a)_i \geq 0$. Then

$$P^0(\infty)(a + v\chi_i) \leq P^K(\infty)(a + v\chi_i) = P^K(\infty)a.$$

The i -components of these vectors are identical. Hence

$$(A^K P^0(\infty)(a + v\chi_i))_i \leq (A^K P^K(\infty)(a))_i \leq 0.$$

This implies $v^K \leq v$ and hence

$$P^0(\infty)(a + v^K\chi_i) \leq P^0(\infty)(a + v\chi_i) \leq P^K(\infty)a.$$

We now have constructed a new sublinear Markov generator $\theta^K \leq \theta$ that is almost linear in the component i , such that

$$\lim_{t \rightarrow \infty} \exp(t\theta^K)a = \lim_{t \rightarrow \infty} \exp(t\theta)a.$$

Applying the same method for all components i , $1 \leq i \leq n$, we obtain a linear Markov generator $B \leq \theta$ such that

$$\lim_{t \rightarrow \infty} \exp(tB^+)a = \lim_{t \rightarrow \infty} \exp(t\theta)a,$$

with $B^+ = \max(B, 0)$. Now we have only to define the optimal stopping strategy for the Markov process described by B . Put $a^\infty = P(\infty)a$ and define

$$\Omega = \{i | 1 \leq i \leq n, a_i < a_i^\infty\},$$

$$A: R^n \rightarrow R^n, \quad (Ax)_i = \begin{cases} 0 & \text{if } i \notin \Omega, \\ (Bx)_i & \text{if } i \in \Omega. \end{cases}$$

Then A is linear and $\leq \theta$. Put $b^\infty = \lim_{t \rightarrow \infty} \exp(tA)a$. Since $b_i^\infty = a_i^\infty$ for $i \notin \Omega$, there exists a c , $0 < c < 1$, such that

$$a \leq (1-c)b^\infty + ca^\infty.$$

We have $b^\infty \in N(B^+)$: If $i \in \Omega$, then $B^+(b^\infty)_i = A^+(b^\infty)_i = 0$, and if $i \notin \Omega$, then it follows from $b^\infty \leq a^\infty$ and $a_i^\infty = b_i^\infty$ that $B^+(b^\infty)_i \leq B^+(a^\infty)_i \leq 0$.

Since $a \in N(B^+)$, and since $N(B^+)$ is a convex cone, we have $(1-c)b^\infty + ca^\infty \in N(B^+)$. This implies $a^\infty \leq (1-c)b^\infty + ca^\infty$ or $a^\infty \leq b^\infty$. Hence, we have proved that

$$\lim_{t \rightarrow \infty} \exp(tA)a = \lim_{t \rightarrow \infty} \exp(t\theta)a.$$

The proof of Theorem 3 is essentially contained in the next two lemmas.

LEMMA 3. Let $S, T \subset \{1, \dots, n\}$ be nonvoid with $\theta(\chi_S) \geq 0$ and $\theta(-\chi_T) \leq 0$. Then

$$\min_S P(t)x \geq \min_S x \quad \forall x \in R^n$$

and

$$\max_T P(t)x \leq \max_T x \quad \forall x \in R^n.$$

Proof. We may assume $x \geq 0$ and $\min_S x = 1$. Since $(\text{Id} + t\theta/m)$ is monotone for large $m \in N$, we have

$$a \in R^n, a \geq \chi_S \Rightarrow \left(\text{Id} + \frac{t\theta}{m}\right)a \geq \left(\text{Id} + \frac{t\theta}{m}\right)\chi_S \geq \chi_S.$$

From $x \geq \chi_S$ it hence follows that $(\text{Id} + t\theta/m)^m x \geq \chi_S$ for large $m \in N$. As $m \rightarrow \infty$ we obtain $\exp(t\theta)x \geq \chi_S$ or $\min_S P(t)x \geq 1$.

The proof of $\max_T P(t)x \leq \max_T x$ is similar.

LEMMA 4. Let $x \in R^n$ with $\theta(x) = 0$ and define

$$S = \left\{i | 1 \leq i \leq n, x_i = \max_{1 \leq k \leq n} x_k\right\},$$

$$T = \left\{i | 1 \leq i \leq n, x_i = \min_{1 \leq k \leq n} x_k\right\}.$$

Then $\theta(\chi_S) \geq 0$ and $\theta(-\chi_T) \leq 0$.

Proof. If $i \notin S$, then $(-\chi_S)_i = \max(-\chi_S)$, hence $\theta(-\chi_S)_i \leq 0$ and $\theta(\chi_S)_i \geq -\theta(-\chi_S)_i \geq 0$. If $i \in S$, then $(x - c\chi_S)_i = \max(x - c\chi_S)$ for small $c > 0$, hence $\theta(x - c\chi_S)_i \leq 0$ and $c\theta(\chi_S)_i \geq \theta(x)_i - \theta(x - c\chi_S)_i \geq 0$. The proof of $\theta(-\chi_T) \leq 0$ is similar.

Proof of Theorem 3. Let $S, T \subset \{1, \dots, n\}$ be two nonvoid disjoint subsets with $\theta(\chi_S) \geq 0$ and $\theta(-\chi_T) \leq 0$. Let $a \in R^n$ be $= 1$ on S and $= 0$ on T . From Lemma 3 we obtain $P(t)a \geq 1$ on S and $P(t)a \leq 0$ on T for all $t \geq 0$, hence $P(\infty)a$ is not constant.

Conversely, if $P(\infty)a$ is not constant for some $a \in R^n$, we can define

$$S = \{i | 1 \leq i \leq n, P(\infty)(a)_i = \max P(\infty)(a)\},$$

$$T = \{i | 1 \leq i \leq n, P(\infty)(a)_i = \min P(\infty)(a)\}.$$

Then S and T are nonvoid disjoint sets, and from Lemma 4, we obtain

$$\theta(\chi_S) \geq 0 \quad \text{and} \quad \theta(-\chi_T) \leq 0,$$

since $P(\infty)a \in N(\theta)$ according to Theorem 1.

REFERENCES

- [1] R. BELLMAN, *Dynamic Programming*, Princeton Univ. Press, Princeton, NJ, 1957.
- [2] R. A. HOWARD, *Dynamic Programming and Markov Processes*, John Wiley, New York, 1960.
- [3] B. L. MILLER, *Finite state continuous time Markov decision processes with a finite planning horizon*, SIAM J. Control, 6 (1968), pp. 266–280.
- [4] ———, *Finite state continuous time Markov decision processes with an infinite planning horizon*, J. Math. Anal. Appl., 22 (1968), pp. 552–569.
- [5] M. NISIO, *On a non-linear semi-group attached to stochastic optimal control*, Publ. RIMS, Kyoto Univ., 13 (1976), pp. 513–537.
- [6] S. R. PLISKA, *A semigroup representation of the maximum expected reward vector in continuous parameter Markov decision theory*, SIAM J. Control, 13 (1975), pp. 1115–1129.
- [7] ———, *Accretive operators and Markov decision processes*, Math. Oper. Res., 5 (1980), pp. 444–459.
- [8] G. DA PRATO, *Applications croissantes et équations d'évolution dans les espaces de Banach*, Academic Press, London and New York, 1976.
- [9] G. RODÉ, *Differentiability of a convex semigroup on R^n* , submitted.

CORRIGENDUM: DISCRETE TIME STOCHASTIC ADAPTIVE CONTROL*

GRAHAM C. GOODWIN[†], PETER J. RAMADGE[‡] AND PETER E. CAINES[¶]

Line 12, p. 837 should read

$$Z(t) \rightarrow Z < \infty \quad \text{a.s.}$$

Lines 15–18, p. 837. The second part of the step below was missing; however, Theorem 5.1 is true as stated on p. 835. The same argument should be inserted at the corresponding places in the proofs of Theorems 5.2, 6.1, 7.1, which are also true as stated.

$$\sum_{t=1}^{\infty} \frac{z^2(t)}{r(t)} < \infty \quad \text{a.s.} \quad (\text{line 16})$$

implies

$$\lim_{N \rightarrow \infty} \frac{N}{r(N)} \frac{1}{N} \sum_{t=1}^N z^2(t) = 0 \quad \text{a.s.} \quad ((5.14), \text{line 18})$$

in case $r(N) \uparrow \infty$. In case $r(N) \uparrow r_{\infty} < \infty$ on $\Omega_1 \subset \Omega$ with $P(\Omega_1) < 1$, then $\|\phi(N)\| \rightarrow 0$ a.s. on Ω_1 . But $Z(t) \rightarrow Z < \infty$ a.s. implies $\|\hat{\theta}(N)\| < 2Z$ a.s. for all $N > N_1(\omega)$. Since $C(z)$ has all zeros outside the unit circle, and since $C(q^{-1})z(t) = b(t) = -\phi(t-1)^T \tilde{\theta}(t-1)$, it follows that $z(t) \rightarrow 0$ a.s. on Ω_1 and so $(1/N) \sum_{t=1}^N z^2(t) \rightarrow 0$ a.s. on Ω_1 , which is statement (5.22). The argument given in the paper, showing that (5.14) implies (5.22), is valid on Ω_1^C . We conclude that (5.22) holds on Ω as required.

* This Journal, 19 (1981), pp. 829–853.

[†] Department of Electrical Engineering, University of Newcastle, N.S.W. 2308, Australia.

[‡] Department of Electrical Engineering, University of Toronto, Toronto, Ontario, Canada.

[¶] Division of Applied Sciences, Harvard University, Pierce Hall, Cambridge, Massachusetts 02138; now with the Department of Electrical Engineering, McGill University, Montreal, P.Q., Canada, H3A 2A7.