

A NEW STOCHASTIC TIME OPTIMAL CONTROL PROBLEM*

U. G. HAUSSMANN†, W. J. ANDERSON‡ AND A. BOYARSKY§

Abstract. Consider a system described by the stochastic differential equation $dx = f(t, x, u) dt + \sigma(t, x) dw$ where w is a Wiener process and u lies in a compact set. If solutions are defined in the sense of Girsanov rather than of Itô, then it is natural to work not with the state at time t but rather with the distribution of $x(t)$. It is shown by an extension of the standard argument that if the state reaches a target set, i.e. a specified set of probability distributions, in finite time then there is a first such time. Next it is shown that if the system is linear in u and if $-1 \leq u \leq 1$, then the attainable distributions arising from bang-bang controls are weakly dense in the set of attainable distributions. Finally some geometric properties of the bang-bang attainable densities are discussed; for example, the exposed points of the attainable densities are the bang-bang attainable densities.

1. Introduction. In this article we shall analyze some stochastic control problems by manipulating not the actual physical state of the system, but rather the probability distribution of the trajectories. These latter are defined by the method of Girsanov. As the "state" is now infinite-dimensional the general nature of the problem becomes more function theoretic than probabilistic; for example in section three we consider the problem of driving these densities to a target set (a set of probability distributions) in minimum time. A method of proof analogous to the one for the deterministic finite-dimensional time optimal control problem yields the existence of an optimal control. It must be emphasized that we are not minimizing the expectation of a random time, but rather a deterministic time and our result does not overlap with any of the usual stochastic existence theorems [1], [2], [3], [4]. Because the Girsanov solutions are used, we demand that the controls be in feedback form.

The example (taken up in § 3 again) of driving the state $x(t)$ to lie in a closed set B motivates us. If the system is deterministic this is the standard optimal time problem, which has usually been stochastized by minimizing $E\tau^u$, the expectation of the first entrance time into B under the control u . However, the engineer working with the system still has no idea when the system will hit the target; in fact he may want to a confidence level, say a time t such that $P(\tau^u > t) < .05$. Rather than merely accept an upper bound on t we shall look for the first such t . The problem can be posed directly as $\min \{t: P(\tau^u > t) < .05, u \text{ admissible}\}$. It is quite possible that for some u $P(\tau^u < \infty) > .95$ and hence the optimal time is finite (even if $E\tau^u = \infty$; cf. [5]).

If the system is linear in u , one would like to have the "bang-bang" principle that the bang-bang controls generate all the attainable densities; however this is

* Received by the editors May 31, 1974, and in final revised form February 9, 1977.

† Department of Mathematics, University of British Columbia, Vancouver, British Columbia, Canada V6T 1W5. The work of this author was supported by the National Research Council of Canada under Grant A 8051.

‡ Department of Mathematics, McGill University, Montreal, Quebec, Canada H3C 3G1. The work of this author was supported in part by the National Research Council of Canada Grant A 8466.

§ Department of Mathematics, Concordia University, Sir George Williams Campus, Montreal, Quebec, Canada H3G 1M8. The work of this author was supported by the National Research Council of Canada under Grant A 9072.

false. But in § 4 we do show that for a Markovian system the densities generated by the bang-bang controls, which we call the bang-bang densities, are weakly dense in the set of attainable densities. In the last section we characterize the bang-bang densities as the exposed points of the attainable densities. We also conclude that if there is a time optimal control which drives the system to a closed half-space (in L_1), then there is a bang-bang control which is also optimal. Optimal controls which are bang-bang also arise from the stochastic maximum principle [6], (but not [7] because there the optimal control must be differentiable in the state) if one is minimizing $EL[x(\cdot)]$ where L is a functional of the trajectories.

2. The problem. b is an n -dimensional separable Brownian motion on $(\Omega, \mathcal{F}, P_0)$. The system is described by

$$(2.1) \quad \begin{aligned} dx(t) &= f(t, x, u(t, x)) dt + \sigma(t, x) db(t), \\ 0 \leq t \leq T < \infty, \quad x(0) &= x_0. \end{aligned}$$

Let \mathcal{C} be the space of R^n valued continuous functions defined on $[0, T]$, under the sup norm. Let \mathcal{H}_t be the σ -algebra generated by $\{y \in \mathcal{C}: y(s) \in B\}$, $0 \leq s \leq t$, B a Borel set in R^n . Let Γ be a compact metric space of control points, with Borel sets \mathcal{B}_Γ . \mathcal{U} , the set of *admissible controls* is the set of all measurable functions $u: ([0, T] \times \mathcal{C}, \mathcal{B}_T \otimes \mathcal{H}_T) \rightarrow (\Gamma, \mathcal{B}_\Gamma)$, adapted to $\{\mathcal{H}_t\}$. \mathcal{B}_T is the Borel algebra of $[0, T]$. We stress that we are using feedback controls.

Assume

- (i) $f(\cdot, \cdot, u): [0, T] \times \mathcal{C} \rightarrow R^n$ is measurable and $\{\mathcal{H}_t\}$ adapted, and $f(t, y, \cdot)$ is continuous for each $(t, y) \in [0, T] \times \mathcal{C}$,
- (ii) $f(t, y, \Gamma)$ is convex for each (t, y) in $[0, T] \times \mathcal{C}$,
- (iii) $\sigma(\cdot, \cdot): [0, T] \times \mathcal{C} \rightarrow R^{n \times n}$ is measurable and $\{\mathcal{H}_t\}$ adapted,
- (iv) $\int_0^T |\sigma(t, y)|^2 dt \leq k < \infty$ for all y in \mathcal{C} ,
- (v) $\sigma(t, y)^{-1}$ exists for all (t, y) in $[0, T] \times \mathcal{C}$, and $|\sigma(t, y)^{-1} f(t, y, u)|^2 \leq K(1 + |y|_t^2)$ where K is a finite constant independent of t, y, u , and where $|y|_t = \sup_{0 \leq s \leq t} |y(s)|$, with $|\cdot|$ denoting the usual Euclidean vector norm or compatible matrix norm, as the case may be,
- (vi) the equation

$$(2.2) \quad x(t) = x_0 + \int_0^t \sigma(s, x) db(s), \quad 0 \leq t \leq T,$$

has a solution adapted to $\mathcal{F}_t = b^{-1}(\mathcal{H}_t)$.

Sufficient conditions for (vi) to hold are discussed in [8, Thm. 5.6 and Cor. 3.2], and in [9, Thm. 1].

We define

$$\alpha^u(t, \omega) = \exp \left\{ \int_0^t [\sigma(s, \omega)^{-1} f^u(s, \omega)]^* db(s, \omega) - \frac{1}{2} \int_0^t |\sigma(s, \omega)^{-1} f^u(s, \omega)|^2 ds \right\},$$

where we have set

$$\sigma(s, \omega) = \sigma(s, x(\cdot, \omega)), \quad f^u(s, \omega) = f(s, x(\cdot, \omega), u(s, x(\cdot, \omega)))$$

with x and b as given by (2.2) and with $u \in \mathcal{U}$. a^* denotes the transpose of a . If E

denotes expectation with respect to P_0 , then $E\alpha^u(T, \cdot) = 1$ according to [6, Lemma 2.1]; hence $(\Omega, \mathcal{F}_t, P_0^u, \{x(t)\})$, with $dP_0^u(\omega) = \alpha^u(T, \omega) dP_0(\omega)$, is a weak solution of (2.1) according to [10, Thm. 1]. In [11, Thm. 6.5], conditions are given to ensure that this solution is unique in law, a property which we do not require. The controls in \mathcal{U} are of course feedback controls.

It is well known that α^u satisfies

$$(2.3) \quad \alpha^u(t, \omega) = 1 + \int_0^t \alpha^u(s, \omega) [\sigma(s, \omega)^{-1} f^u(s, \omega)]^* db(s, \omega),$$

and that $(\Omega, \mathcal{F}_t, P_0, \{\alpha^u(t)\})$ is a martingale since $E\alpha^u(T) = 1$. Moreover by [6, Cor. 4.2], there exist $p > 1$, $K_0 < \infty$ such that for all $u \in \mathcal{U}$, $t \leq T$,

$$(2.4) \quad E\{\alpha^u(t)^p\} \leq K_0.$$

All of the above requires only the assumptions (i), (iii)–(vi).

As it will be convenient to work in the space $(\mathcal{C}, \mathcal{H}_T, P_0 \circ x^{-1})$ we shall do so from now on writing P for $P_0 \circ x^{-1}$, and

$$(2.5) \quad \alpha^u(t, y) = \exp \left\{ \int_0^t [\sigma(s, y)^{-1} f^u(s, y)]^* dw(s, y) - \frac{1}{2} \int_0^t |\sigma(s, y)^{-1} f^u(s, y)|^2 ds \right\}$$

with $f^u(s, y) = f(s, y, u(s, y))$ and with a new Brownian motion $w(t, y) = \int_0^t \sigma^{-1}(s, y) dy(s) = \int_0^t \sigma^{-1}(s, y) dx(s)$. Observe that $x(t, y) = y(t)$, and set $dP^u = \alpha^u(T) dP$, and $w^u(t) = w(t) - \int_0^t \sigma^{-1}(s, y) f^u(s, y) ds$.

Define $\pi_s: \mathcal{C} \rightarrow R^n$ by $\pi_s(y) = y(s)$, and $S_t: L_1(\mathcal{C}, P) \rightarrow \mathcal{M}$ by

$$(S_t \alpha)(B) = \int_{\pi_t^{-1} B} \alpha dP$$

where \mathcal{M} is the Banach space of all regular signed measures on R^n with variation norm. Define the set of *attainable densities* at time t by

$$\mathcal{D}(t) = \{\alpha^u(t): u \in \mathcal{U}\} \subset L_1(\mathcal{C}, P)$$

and the *attainable set* at time t by

$$\mathcal{A}(t) = S_t \mathcal{D}(t).$$

We remark that $S_t \mathcal{D}(t) = S_t \mathcal{D}(v)$ for all $v \in [t, T]$, and that $\mu(\cdot)$ is in $\mathcal{A}(t)$ if and only if it is the distribution of $x(t)$ under P^u for some u in \mathcal{U} . Henceforth we shall write $\mu(B; t, u)$ for $\mu(B) \in \mathcal{A}(t)$. Under (ii) $\mathcal{D}(t)$ is weakly closed in $L_1(\mathcal{C}, P)$. This last result is contained in [12, Thm. 2], for the case $\sigma = I$, but the proof can be generalized to the case $\sigma \neq I$ fairly easily. Theorem 6 of [1] also gives the result when $\sigma = I$.

We remark that one can also treat systems of the form

$$dx(t) = \begin{pmatrix} f_1(t, x) \\ f_2(t, x, u(t, x)) \end{pmatrix} dt + \begin{pmatrix} 0 & 0 \\ 0 & \sigma(t, x) \end{pmatrix} d \begin{pmatrix} 0 \\ b \end{pmatrix}$$

where b is still n -dimensional, but $x(t) \in R^{m+n}$; cf. [6]. This allows for (systems of) higher order equations with noise in the forcing term.

3. Existence of time optimal controls. Our aim in this section is to give conditions for the existence of a time \hat{t} such that $K(t) \cap \mathcal{D}(t)$ (respectively $K(t) \cap \mathcal{A}(t)$) is empty for $t < \hat{t}$, but nonempty for $t = \hat{t}$. Here $K(t)$ is the target set at time t , lying in $L_1(\mathcal{C}, P)$ (respectively \mathcal{M}). We use the standard method of first proving that $\mathcal{D}(t)$ is compact and $t \mapsto \mathcal{D}(t)$ is continuous. The existence follows then as in the deterministic case. In this section we assume all the hypotheses (i)–(vi).

Let Z be the space of probability measures on R^n , and let \mathcal{W} be the topology of weak convergence; cf. [13, p. 236]. Hence $\mu_n \rightarrow \mu$ in (Z, \mathcal{W}) if and only if $\int_{R^n} \phi d\mu_n \rightarrow \int_{R^n} \phi d\mu$ for all bounded continuous real functions ϕ on R^n . Moreover [13, p. 239], \mathcal{W} is a metric topology with the Prohorov metric p where

$$p(\mu, \nu) = \inf \{ \varepsilon > 0 : \mu(B^\varepsilon) + \varepsilon \geq \nu(B), \nu(B^\varepsilon) + \varepsilon \geq \mu(B), \text{ all Borel sets } B \text{ in } R^n \}$$

with B^ε being the open ε neighborhood of B .

Observe that $Z \subset \mathcal{M} = \mathcal{C}_0^*$, where \mathcal{C}_0^* is the dual of \mathcal{C}_0 , the space of continuous functions on R^n which are zero at ∞ . Hence on \mathcal{M} there are the weak topology w and the weak $*$ topology w^* . Considering the induced topologies on Z , we have, [14, V.3.9],

$$w \supset \mathcal{W} \supset w^*.$$

LEMMA 3.1. *For each fixed $t \leq T$, $\mathcal{D}(t)$ is weakly compact in $L_1(\mathcal{C}, P)$, and $\mathcal{A}(t)$ is compact in (Z, \mathcal{W}) .*

Proof. $\mathcal{D}(t)$ is bounded and uniformly integrable [(2.4)], and hence weakly sequentially compact in $L_1(\mathcal{C}, P)$ [14, IV.8.11]. As remarked above $\mathcal{D}(t)$ is weakly closed and so by the Eberlein–Šmulian theorem it is weakly compact. Since S_t is a continuous linear map of $L_1(\mathcal{C}, P)$ into \mathcal{M} , it is continuous $(L_1, w) \rightarrow (\mathcal{M}, w)$ [14, p. 422], and so $S_t \mathcal{D}(t) = \mathcal{A}(t)$ is weakly compact. As $w \supset \mathcal{W}$, $\mathcal{A}(t)$ is \mathcal{W} compact and the result is established.

Z is a metric space so we can consider a Hausdorff metric d on the closed subsets:

$$d(A, B) = \max \{ d_0(A, B), d_0(B, A) \}$$

where

$$d_0(A, B) = \sup \{ d_0(x, B) : x \in A \} \quad \text{and} \quad d_0(x, B) = \inf \{ p(x, y) : y \in B \}.$$

Similarly we can define the Hausdorff metric on the closed sets of $L_1(\mathcal{C}, P)$ using the norm of L_1 in place of p . This Hausdorff metric is used in the next result.

LEMMA 3.2. *$t \mapsto \mathcal{D}(t)$, $t \mapsto \mathcal{A}(t)$ are continuous mappings on $[0, T]$.*

Proof. $\mathcal{D}(t)$ is weakly closed, hence strongly closed. We shall show that $t \rightarrow \alpha^u(t)$ is continuous in t , uniformly in u . This will establish the result for \mathcal{D} .

Consider

$$\alpha^u(t, y) - \alpha^u(s, y) = \alpha^u(t, y) \left[1 - \exp \left\{ \int_t^s (\sigma^{-1} f^u)^* dw - \frac{1}{2} \int_t^s |\sigma^{-1} f^u|^2 d\tau \right\} \right].$$

From (iv) and (v) it follows that the term inside the braces converges to zero in

$L_2(\mathcal{C}, P)$ uniformly in $u \in \mathcal{U}$ as $s \rightarrow t$. Hence $[1 - \alpha^u(s)/\alpha^u(t)] \rightarrow 0$ in probability uniformly in \mathcal{U} . This together with (2.4), the uniform integrability of α^u , gives the result.

Now we show the same result for $t \mapsto \mu(\cdot; t, u)$. For any bounded continuous function ϕ ,

$$\begin{aligned} & \left| \int_{R^n} \phi(x) \mu(dx; t, u) - \int_{R^n} \phi(x) \mu(dx; s, u) \right| \\ & \leq \int |\phi(\pi_t y) - \phi(\pi_s y)| \alpha(T, y; u) P(dy) \\ & \leq \left\{ \int |\phi(\pi_t y) - \phi(\pi_s y)|^q P(dy) \right\}^{1/q} \left\{ \int |\alpha|^p P(dy) \right\}^{1/p} \\ & \leq \|\phi(\pi_t \cdot) - \phi(\pi_s \cdot)\|_q K_0 \end{aligned}$$

where p is given by (2.4). But $|\phi(\pi_t y)| \leq \sup_x |\phi(x)|$, and so by the dominated convergence theorem and the continuity of ϕ the result follows.

For the next theorem we need a new condition which we state first for the case $K(t) \subset L_1(\mathcal{C}, P)$ and then $K(t) \subset Z$.

(H₁) $K(t)$ is weakly closed in $L_1(\mathcal{C}, P)$ and $\lim_{s \downarrow t} d_0(K(s) \cap \mathcal{D}(s), K(t)) = 0$.

(H₂) $K(t)$ is a closed set in (Z, \mathcal{W}) and $\lim_{s \downarrow t} d_0(K(s) \cap \mathcal{A}(s), K(t)) = 0$.

We observe that if $K(t)$ is closed and if $t \rightarrow K(t)$ is right continuous then (H₁) or (H₂) hold.

THEOREM 3.1. *Assume (H₁) (respectively (H₂)) holds. If $K(T) \cap \mathcal{D}(T)$ (respectively $K(T) \cap \mathcal{A}(T)$) $\neq \emptyset$, then there is a smallest $t \geq 0$ such that $K(t) \cap \mathcal{D}(t)$ (respectively $K(t) \cap \mathcal{A}(t)$) $\neq \emptyset$.*

The proof is standard, along the lines of the result on p. 127 of [15]. We do not attempt to answer the question of when $K(T) \cap \mathcal{D}(T) \neq \emptyset$; the notion of controllability for our systems should arise here; however we will give two examples.

Example 1. A closed set B in R^n is given, as well as a right continuous decreasing function $c(t)$, $1 \geq c(t) \geq 0$. The problem is to find the first time t such that $P^u(x(t) \in B) \geq c(t)$. We set

$$K(t) = \{\mu \in Z : \mu(B) \geq c(t)\}, \quad 0 \leq t \leq 1.$$

[For example if $c(1) = 0$, $c(t) = .9$ for $t < 1$, then we are trying to drive the system to B with probability .9 as fast as possible, but at time $t = 1$ the system is shut off. The condition $c(1) = 0$ implies $K(1) = Z$, so that $K(1) \cap \mathcal{A}(1) \neq \emptyset$ hence controllability is not an issue here.]

$K(t)$ is closed [13, Thm. 2.1]. We shall now show that $t \rightarrow K(t)$ is right continuous. Since $K(t) \subset K(s)$ for $t < s$, then $d_0(K_t, K_s) = 0$. Now for $\mu \in K_s$, if $\mu(B) \geq c(t)$, set $\bar{\mu} = \mu$, but if $c(s) \leq \mu(B) < c(t) < 1$, take $s > t$ so close to t that

$c(t) \geq c(s) > c(t)^2$ and set

$$\bar{\mu} = \frac{c(t)}{c(s)} \mu \Big|_B + \frac{1 - [c(t)/c(s)] \mu(B)}{\mu(B^c)} \mu \Big|_{B^c}$$

where $B^c = R^n - B$ and $\mu|_B$ is μ restricted to B . Finally if $c(s) \leq \mu(B) < c(t) = 1$, set

$$\bar{\mu} = [\mu(B)]^{-1} \mu|_B.$$

Then $\bar{\mu} \in K(t)$ and if $c(t) \geq \varepsilon/(c(t) - \varepsilon)$, then

$$p(\mu, \bar{\mu}) \leq \varepsilon/(c(t) - \varepsilon)$$

and so $d_0(K_s, K_t) \leq \varepsilon/(c(t) - \varepsilon)$. This establishes the required continuity. The next example is less artificial.

Example 2. Again B is a closed set in R^n . One wishes to drive the state $x(t)$ to the set B . Let $\tau(y)$ be the first entrance time of $x(t, y)$ into B , i.e.

$$\tau(y) = \inf \{t: y(t) \in B\}.$$

A standard problem is to minimize

$$E^u(\tau) = \int_C \tau(y) \alpha^u(T; y) dP(y).$$

In general however this expression is not finite (see [3] for an example where it is because τ is the first exit time from a bounded set) so one replaces τ by $\tau \wedge T$ rather arbitrarily. Let us consider the problem slightly differently: suppose we are satisfied if the state lies in B with a probability at least $c < 1$, but we wish to achieve this as fast as possible. Hence we minimize over the set

$$A = \{t \leq T: P^u(\tau \leq t) \geq c, u \in \mathcal{U}\}.$$

If for some u , $P^u(\tau < \infty) > c$, then the set A is nonempty for T sufficiently large. Lyapunov conditions for a process to satisfy $P^u(\tau < \infty) = 1$ have been given in [5]

We define the target set $K(t)$ by

$$K(t) = \left\{ \beta \in L_1(\mathcal{C}, P): \int_{\mathcal{C}} 1_{\tau \leq t} \beta dP \geq c \right\}.$$

From the above discussion it follows (under conditions as in [3] or [5] assuming unique weak solutions) that $K(T) \cap \mathcal{D}(T) \neq \emptyset$ for T sufficiently large.

Clearly $K(t)$ is weakly closed. It remains to show that for any $\varepsilon > 0$, any $\beta \in K(s) \cap \mathcal{D}(s)$, there is a $\bar{\beta} \in K(t)$ such that if s is sufficiently close to t , $s > t$, then $|\beta - \bar{\beta}| < \varepsilon$. We set $B_1 = \{y: \tau(y) \leq t\}$, $B_2 = \{y: t < \tau(y) \leq s\}$, $B_3 = \mathcal{C} - B_1 - B_2$. If $\int_{B_1} \beta dP \geq c$ set $\bar{\beta} = \beta$. Otherwise given $\varepsilon > 0$, choose s so small that $P(B_2) = P\{\omega: t < \tau(y) \leq s\} < (\varepsilon K_0^{-1})^q$. Hence $\int_{B_2} \beta dP \leq \varepsilon$ and $c > \int_{B_1} \beta dP \geq c - \varepsilon$. Now set

$$\bar{\beta}(y) = \begin{cases} \beta(y) c \left(\int_{B_1} \beta dP \right)^{-1} & \text{if } y \in B_1, \\ \beta(y) & \text{otherwise.} \end{cases}$$

Then $\bar{\beta} \in K(t)$. Moreover if $\beta \neq \bar{\beta}$ then

$$E|\beta - \bar{\beta}| = \left| 1 - c \left(\int_{B_1} \beta dP \right)^{-1} \int_{B_1} \beta dP - c - \int_{B_1} \beta dP \right| \leq \varepsilon.$$

Hence (H₁) is satisfied and this time optimal problem has a solution.

Note that q is the index conjugate to p in (2.4) and K_0 is as in (2.4).

4. A weak bang-bang principle. In the deterministic linear theory the bang-bang principle states that if a point π can be reached by the system using an admissible control then it can be reached using a bang-bang control. It is well known that this result fails in infinite dimensions (the case with any stochastic problem), but we shall find a weak version of this principle. We define the bang-bang controls \mathcal{U}_b for the case where

$$\Gamma = \{u \in R^r : |u_i| \leq 1, i = 1, 2, \dots, r\}$$

by

$$\mathcal{U}_b = \{u \in \mathcal{U} : |u_i(t, y)| = 1 \text{ a.e. } dt \times dP, i = 1, 2, \dots, r\}.$$

However we must restrict ourself to the Markov case, so we replace \mathcal{U} by $\bar{\mathcal{U}}$ and \mathcal{U}_b by $\bar{\mathcal{U}}_b$ where

$$\bar{\mathcal{U}} = \{u : [0, T] \times R^n \rightarrow \Gamma : u \text{ Borel measurable}\},$$

$$\bar{\mathcal{U}}_b = \{u \in \bar{\mathcal{U}} : |u_i(t, x)| = 1 \text{ a.e. } dt \times dx, i = 1, 2, \dots, r\}.$$

It is clear how to imbed $\bar{\mathcal{U}}$ in \mathcal{U} , $\bar{\mathcal{U}}_b$ in \mathcal{U}_b since $P \circ \pi_t^{-1}$ is absolutely continuous with respect to Lebesgue measure for almost all t [2, Thm. III4]. Note that Thm. III4 can be extended to unbounded drifts satisfying (v). We write $\bar{\mathcal{D}}(t) = \{\alpha^u(t) : u \in \bar{\mathcal{U}}\}$, $\bar{\mathcal{D}}_b(t) = \{\alpha^u(t) : u \in \bar{\mathcal{U}}_b\}$, $\bar{\mathcal{A}}(t) = S_t \bar{\mathcal{D}}(t)$, $\bar{\mathcal{A}}_b(t) = S_t \bar{\mathcal{D}}_b(t)$.

We shall show that for the system given by

$$(4.1) \quad dx(t) = [g_0(t, x(t)) + g(t, x(t))u(t, x(t))] dt + \sigma(t, x(t)) dw(t)$$

$\bar{\mathcal{D}}_b(t)$ is weakly dense in $\bar{\mathcal{D}}(t)$, and $\bar{\mathcal{A}}_b(t)$ is dense in $\bar{\mathcal{A}}(t)$.

Proofs are only given for the case $t = 1$ since they are identical for all t . To apply the results of [2], we assume (v), (vi), and we strengthen (i), (iii) and (iv) by demanding

- (i)' $f(\cdot, \cdot, u) : [0, T] \times R^n \rightarrow R^n$ is Borel measurable and $f(t, x, \cdot)$ is continuous,
- (iii)' $\sigma : [0, T] \times R^n \rightarrow R^{n \times n}$ is continuous,
- (iv)' $|\sigma(t, x)| \leq K_1, \forall(t, x) \in [0, T] \times R^n$.

The next lemma, due to R. S. Phelps, gives the essential denseness. (X, \mathcal{B}, m) is a measure space. $A \in \mathcal{B}$ is an *atom* if $m(A) > 0$ and if for any $\bar{A} \in \mathcal{B}$, $\bar{A} \subset A \Rightarrow \{m(\bar{A}) = 0, \text{ or } m(\bar{A}) = m(A)\}$.

Let

$$\|\phi\|_\infty = \text{ess sup}_x \max_{1 \leq i \leq r} |\phi_i(x)|$$

and put

$$\|\phi\|_1 = \max_{1 \leq i \leq r} \int_X |\phi_i(x)| m(dx).$$

Assuming equivalent functions are identified, set

$$\mathcal{L}'_\infty = \{\phi: X \rightarrow R', \phi \text{ is } \mathcal{B} \text{ measurable}, \|\phi\|_\infty < \infty\},$$

and

$$\mathcal{L}'_1 = \{\phi: X \rightarrow R', \phi \text{ is } \mathcal{B} \text{ measurable}, \|\phi\|_1 < \infty\}.$$

LEMMA 4.1 (R. S. Phelps). *Let (X, \mathcal{B}, m) be a finite measure space. Let*

$$S = \{\phi \in \mathcal{L}'_\infty: \|\phi\|_\infty \leq 1\}$$

and let

$$S^\wedge = \{\phi \in \mathcal{L}'_\infty: |\phi_i(x)| = 1, \text{ a.e. } i = 1, 2, \dots, r\},$$

*i.e., S^\wedge is the set of extreme points of S . Then S^\wedge is weak * dense in S if and only if (X, \mathcal{B}, m) has no atoms.*

Proof. Assume that (X, \mathcal{B}, m) has an atom $A \in \mathcal{B}$. Then

$$N = \left\{ \phi \in S: \left| \int_X \phi_i 1_A dm \right| < \frac{1}{2} m(A), i = 1, \dots, r \right\}$$

is a weak * neighborhood of 0. If $\phi \in N \cap S^\wedge$ then $\phi_i(x) = +1$ on $\tilde{A}_i \subset A$ and $\phi_i(x) = -1$ on $A - \tilde{A}_i$, $\tilde{A}_i \in \mathcal{B}$. Since A is an atom, $m(\tilde{A}_i) = m(A)$ or $m(A - \tilde{A}_i) = m(A)$. In either case $|\int_X \phi_i 1_A dm| = m(A)$ hence $\phi \notin N$. Thus $N \cap S^\wedge$ is empty and S^\wedge is not weak * dense in S .

The proof in the other direction makes use of the bang-bang principle. Given $f^1, f^2, \dots, f^n \in \mathcal{L}'_1$, define $T: S \rightarrow R^{n \times r}$ by

$$T(\phi)_{ij} = \int_X f_j^i \phi_i dm.$$

Since m is nonatomic, it follows from the bang-bang principle of LaSalle [16, Thm. 8.2] with the use of the general proof of [17], that

$$T(S^\wedge) = T(S)$$

for any f^1, f^2, \dots, f^n in \mathcal{L}'_1 . Consider any ϕ_0 in S and a weak * neighborhood N given by

$$N = \left\{ \phi \in S: \left| \int_X f^i \cdot (\phi - \phi_0) dm \right| < \varepsilon_i, f^i \in \mathcal{L}'_1, i = 1, \dots, n \right\}.$$

Let $\phi \neq \phi_0$ be in N . Since $T(S^\wedge) = T(S)$ there exists $\tilde{\phi}$ in S^\wedge such that

$$\int_X f^i \cdot \phi dm = \int_X f^i \cdot \tilde{\phi} dm, \quad i = 1, 2, \dots, n.$$

Hence $\phi \in N$ and so $N \cap S^\wedge \neq \emptyset$, or S^\wedge is weak * dense in S . This completes the proof.

Let $X = [0, 1] \times R^n$, $X_M = [0, 1] \times \{x \in R^n: |x| \leq M\}$, let \mathcal{B} be the Borel sets in X_M , let $\mathcal{U}^M = \{u \cdot 1_{|x| \leq M}: u \in \tilde{\mathcal{U}}\}$, $\mathcal{U}_b^M = \mathcal{U}^M \cap \tilde{\mathcal{U}}_b$. Lemma 4.1 now states that \mathcal{U}_b^M is dense in $\mathcal{U}^M \in (L_\infty(X_M, dt \times dx), w^*)$.

We give now an important continuity result which follows essentially from the footnote on p. 500 of [8], a proof of which is supplied in [2, Thm. IV-3].

THEOREM 4.1. *The mapping $u \rightarrow \alpha^u$ of $(L_\infty(X, dt \times dx), w^*)$ into $(L_1(\mathcal{C}, P), w)$ is continuous.*

Proof. If $u_k \rightarrow u_0$ weak $*$ then the result follows from [8], cf. [2], provided

$$f(t, x, u_k) = g_0(t, x) + g(t, x)u_k(t, x)$$

is bounded for each k , and converges to $f(t, x, u_0)$ in $(L_\infty(X, dt \times dx), w^*)$. The next lemma now completes the proof.

For $N < \infty$, set $f_N(t, x, u_k)1_{x \leq N}$ and write α_k for α^{u_k} , α_k^N for the density generated by $f_N(t, x, u_k)$.

LEMMA 4.2. (i) $|f_N(t, x, u_k)|^2 \leq K_1 K_0(1 + N^2)$,

(ii) $f_N(\cdot, \cdot, u_k)$ converges to $f_N(\cdot, \cdot, u_0)$ in (L_∞, w^*) ,

(iii) $\|\alpha_k^N - \alpha_k\|_1 \rightarrow 0$ uniformly in k as $N \rightarrow \infty$.

Proof. Part (i) follows from assumptions (iv)' and (v). Part (ii) is obvious.

Now consider (iii).

Let

$$E_N = \{y \in \mathcal{C} : |y|_1 > N\}.$$

Then $\alpha_k^N = \alpha_k$ on the complement of E_N , and

$$\begin{aligned} \|\alpha_k^N - \alpha_k\| &= \int_{E_N} |\alpha_k^N - \alpha| dP \\ &\leq |\alpha_k^N - \alpha|_p [P(E_N)]^{(p-1)/p} \leq 2K_0^{1/p} [P(E_N)]^{(p-1)/p} \end{aligned}$$

by (2.4). Note that although $\alpha_k^N \notin \mathcal{D}(1)$ in general, nevertheless (v) is still satisfied so that α_k^N also satisfies (2.4). But now $P(E_N) \rightarrow 0$ as $N \rightarrow \infty$ and the result is established.

THEOREM 4.2. $\bar{\mathcal{D}}_b(t)$ is dense in $\overline{\mathcal{D}(t)} \in (L_1(\mathcal{C}, P), w)$.

Proof. For u_0 in $\bar{\mathcal{U}}$, set $u^N = u_0 1_{x \leq N}$. Let $\{v_k^N\} \subset \mathcal{U}_b^N$ be a sequence converging to u^N in the weak $*$ topology. Such sequences exist according to Lemma 4.1. Define

$$u_k = \sum_{N=1}^{\infty} v_k^N \cdot (1_{|x| \leq N} - 1_{|x| < N-1}).$$

Then $u_k \in \bar{\mathcal{U}}_b$ and in the proof of Theorem 4.1, $f_N(\cdot, \cdot, u_k) \rightarrow f_N(\cdot, \cdot, u_0)$ in $(L_\infty(X), w^*)$ because

$$f_N(t, x, u_k) = \sum_{M=1}^N f_n(t, x, v_k^M(t, x))(1_{|x| \leq M}(x) - 1_{|x| < M-1}(x)).$$

But now $\alpha_k^N \rightarrow \alpha_0^N$ weakly according to [2, Thm. IV-3]. An application of Lemma 4.2 completes the proof.

COROLLARY. $\bar{\mathcal{A}}_b(t)$ is dense in $\bar{\mathcal{A}}(t)$ in (Z, \mathcal{W}) .

Proof. This follows because the map S_t is continuous.

In [2] and [3] Theorem 4.1 is used to obtain existence of a solution to the standard stochastic optimal control problem with cost $E^u l(y)$ where l is in $L_\infty(\mathcal{C}, P)$. Theorem 4.2 tells us (if f is linear in u) that there is an ε -optimal control which is bang-bang. If the coefficients of the problem are sufficiently smooth, then dynamic programming yields the existence of an optimal control which is in fact

bang-bang, cf. [9]. The maximum principle [7] also gives the result *formally*, but unfortunately the optimal feedback control must be differentiable to apply the principle. The version in [6] can be applied rigorously to yield a bang-bang control, which however is $\{\mathcal{H}_t\}$ adapted, i.e. not Markovian a priori.

5. Bang-bang controls. We characterize here the bang-bang densities $\mathcal{D}_b(t)$ in terms of the geometry of $\mathcal{D}(t)$ when the drift f is linear in u . We show that if α^u is a support point of $\mathcal{D}(t)$, then u is bang-bang except possibly on a (t, x) set where the support functional is “degenerate”. This leads to three conclusions: if α^u is an exposed point of $\mathcal{D}(t)$ then u is bang-bang (and conversely under added assumptions), if $\hat{\alpha}$ minimizes $El(y)\alpha^u(y)$ then there is a bang-bang control which also gives the minimum, if we are considering the time optimal problem of § 3 with $K(t)$ a closed half-space then there is a bang-bang control which is time optimal. Again $t = 1$ (the results hold for all finite t by the same proofs). The equation is

$$(5.1) \quad dx = [g_0(t, x) + g(t, x)u(t, x)] dt + \sigma(t, x) dw,$$

under assumptions (i), (iii)–(vi), $\Gamma = \{u \in R^r : |u_i| \leq 1, i = 1, \dots, r\}$. We emphasize that the first of the three conclusions states that the bang-bang densities, far from being all the densities, are only the exposed points of the attainable densities in the space $L_{p'}(\mathcal{C}, P)$ with $p' > 1$ sufficiently small.

Let us now look at support points of $\mathcal{D}(1)$, i.e. points α_0 such that for some continuous linear functional l , $l\alpha_0 \geq l\alpha$, $\alpha \in \mathcal{D}(1)$. For $l \in L_\infty(\mathcal{C}, P)$ and $u_0 \in \mathcal{U}$ if $\alpha_0(t) = \alpha(t; u_0)$ then $l_0 \equiv l\alpha_0(1) \in L_1(\mathcal{C}, P)$. Define $l_0(t) = E\{l\alpha_0(1) | \mathcal{H}_t\}$. By Theorem 3 of [18] there exists a unique measurable function $h(t, y)$ such that

$$P\left\{\int_0^1 h^2 dt < \infty\right\} = 1$$

and

$$l_0(t) = E\{l\alpha_0(1)\} + \int_0^t h(s)^* dw(s).$$

Since $l_0(t) = \alpha(t, u_0)E^{u_0}(l | \mathcal{H}_t)$ and since

$$\begin{aligned} d\alpha(t; u_0)^{-1} &= -\alpha(t; u_0)^{-1}\sigma(t, y)^{-1}f^{u_0}(t, y)dw + \alpha(t; u_0)^{-1}|\sigma(t, y)^{-1}f^{u_0}(t, y)|^2 dt \\ &= -\alpha(t; u_0)^{-1}\sigma(t, y)^{-1}f^{u_0}(t, y)dw^{u_0} \end{aligned}$$

then from Itô's lemma it follows that

$$E^{u_0}(l | \mathcal{H}_t) = E^{u_0}(l) + \int_0^t \mathcal{X}(s, y) dw^{u_0}(s)$$

with $\mathcal{X}(s, y) = \alpha(s, y; u_0)^{-1}[h(s, y) - l_0(s, y)\sigma(s, y)^{-1}f^{u_0}(s, y)]^*$.

LEMMA 5.1. *If u_0 and $u_0 + v$ are in \mathcal{U} , then*

$$\begin{aligned} (5.2) \quad & E\{l[\alpha(1; u_0 + v) - \alpha(1; u_0)]\} \\ &= E\left\{\alpha(1; u_0 + v) \int_0^1 \mathcal{X}(t, y)\sigma(t, y)^{-1}g(t, y)v(t, y) dt\right\}. \end{aligned}$$

Proof. This follows from the results of [6, § 3] if we observe that $\alpha(1; u_0)^{-1}\alpha(1; u)$, the density of P^u relative to P^{u_0} , also satisfies a bound as in (2.4).

THEOREM 5.1. *If l is a support functional of $\mathcal{D}(1)$ at $\alpha_0 = \alpha(u_0)$, then for all $u \in \mathcal{U}$*

$$(5.3) \quad [h(t, y) - l_0(t, y)\sigma(t, y)^{-1}f^{u_0}(t, y)]^*\sigma(t, y)^{-1}g(t, y)[u(t, y) - u_0(t, y)] \leq 0, \quad \text{a.e. } dt \times dP.$$

Proof. Set $u - u_0 = v$ and consider the control $u_0 + \lambda 1_N v = u_\lambda$ where $0 < \lambda \leq 1$,

$$1_N(t, y) = \begin{cases} 1 & \text{if } \max \left\{ \int_0^t |\mathcal{X}(s)|^2 ds, |y|^2 \right\} \leq N^2, \\ 0 & \text{otherwise.} \end{cases}$$

Since \mathcal{U} is convex, this control is admissible. From Lemma 5.1 we obtain

$$0 \geq E\alpha(1; u_\lambda)\lambda \int_0^1 \mathcal{H}\sigma^{-1}gv 1_N dt.$$

We divide by λ and let $\lambda \rightarrow 0$. Since $\alpha(1, u_\lambda) \rightarrow \alpha_0$ with probability 1, since $\alpha(1, u_\lambda) - \alpha_0$ is uniformly integrable, and since

$$\left| \int_0^1 \mathcal{H}\sigma^{-1}gv 1_N dt \right| \leq N[K(1 + N^2)]^{1/2},$$

then

$$0 \geq E\alpha_0 \int_0^1 \mathcal{H}\sigma^{-1}gv 1_N dt = E \int_0^1 \alpha_0(t)\mathcal{H}_t\sigma_t^{-1}g_tv_t 1_N(t) dt.$$

Since the integrand is measurable and $\{\mathcal{H}_t\}$ adapted, and the inequality holds for all v such that $u_0 + v \in \mathcal{U}$, then

$$\alpha_0\mathcal{H}\sigma^{-1}gv 1_N \leq 0 \quad \text{a.e. } dt \times dP.$$

Since $E \int_0^1 |\mathcal{X}|^2 dt < \infty$ and $E|y|_1^2 < \infty$ it follows that $1_N \nearrow 1$ a.e. $dt \times dP$. This establishes the theorem.

It follows from (5.3) that each component of u_0 is bang-bang on the set where the corresponding component of $(\sigma^{-1}g)^*[h - l_0\sigma^{-1}f^{u_0}]$ is not zero. If l is a support functional at α_0 , set

$$S_i^l = \{(t, y) \in [0, 1] \times \mathcal{C} : [(\sigma^{-1}g)^*(h - l_0\sigma^{-1}f^{u_0})]_i = 0\},$$

$$S_i(u_0) = \bigcap_l S_i^l, \quad S(u_0) = \bigcup_i S_i(u_0),$$

where the intersection is over all l which support $\mathcal{D}(1)$ at $\alpha(u_0)$. We call $S(u_0)$ the singularity set of u_0 , and we say that u_0 is *singular* if $S(u_0)$ has positive measure.

COROLLARY 5.1. *If $\alpha(1; u_0)$ is a support point of $\mathcal{D}(1)$ and if u_0 is not singular, then u_0 is bang-bang.*

We remark that all points of $\mathcal{D}(1)$ are support points with the support functional $l = 1$. However $S_i^1 = [0, 1] \times \mathcal{C}$ for all i , so that if the point has no

support functionals other than the constants, we can say nothing about whether its control u is ever bang-bang. Corollary 5.1 is not very illuminating as we cannot readily identify the nonsingular controls.

To characterize \mathcal{D}_b more explicitly we proceed as follows. An exposed point α of a set A is a point in A such that for some continuous linear functional l , $l(\alpha) > l(\alpha')$ for all α' in A with $\alpha' \neq \alpha$. We write $\exp [A]_q$ for the set of exposed points of A , $A \subset L_q(\mathcal{C}, P)$.

COROLLARY 5.2. $\mathcal{D}_b(1) \supset \exp [\mathcal{D}(1)]_{p'}$ for $1 \leq p' \leq 2p/(p+1)$, p as in (2.4).

Proof. We need only show that if α_0 is exposed, then u_0 is not singular. If α_0 is exposed, but $[(\sigma^{-1}g)^*(h - l_0\sigma^{-1}f^{u_0})]_i = 0$ on B , a set of positive measure, choose v to have support on B with only $v_i \neq 0$. Note that now $l_0 = l \cdot \alpha_0 \in L_1$ with $l \in L_{q'}$, $q' \geq 2q$ where q is conjugate to p . The factor 2 comes from [6] as used in Lemma 5.1. By (5.2) with $\alpha(1) = \alpha(1; u_0 + v)$,

$$E\{l[\alpha(1) - \alpha_0(1)]\} = 0$$

and thus $\alpha(1) = \alpha_0(1)$ and $gv = 0$ a.e. We can now change u_0 on B to u_1 , which is bang-bang, and $\alpha(1; u_0) = \alpha(1; u_1)$, so that $\alpha_0 \in \mathcal{D}_b(1)$.

Now assume

- (vii) for all extreme points u of Γ , $u + \ker g(t, y)$ is an extremal set of Γ a.e. $dt \times dP$ ($\ker g$ is the null space of g) (cf. [14] for extremal sets);
- (viii) if $\overline{g(t, y)}$ is the pseudo inverse of $g(t, y)^*$ then for some $m < \infty$

$$|\overline{g(t, y)}| \equiv \sup_{|b|=1} \min \{|a| : g(t, y)^*a = b\} \leq K(1 + |y|^m).$$

THEOREM 5.2. Under hypotheses (vii), (viii), $\mathcal{D}_b(1) = \exp [\mathcal{D}(1)]_{p'}$ for all $1 < p' \leq 2p/(p+1)^{-1}$.

Proof. We already know that $\mathcal{D}_b(1) \supset \exp [\mathcal{D}(1)]_{p'}$. Suppose now that $u_0 \in \mathcal{U}_b$. According to (viii) and [6, Lemma 3.1], $E^{u_0}(\int_0^1 |\mathcal{X}|^2 dt)^{q'} < \infty$ if we set $\mathcal{X} = (u_0 - v)^* \bar{g}^* \sigma$ where v is the orthogonal projection of u_0 on $\ker g$ and where q' is conjugate to p' . We set $l = \int_0^1 \mathcal{X} dw^{u_0}$ to obtain

$$\begin{aligned} [E|l|^{q'}]^{1/q'} &\leq [E^{u_0}(\alpha(u_0)^{-1})^p]^{1/pq'} [E^{u_0}|l|^{qq'}]^{1/qq'} \\ &\leq K(q)E^{u_0} \left[\left(\int_0^1 |\mathcal{X}|^2 dt \right)^{qq'} \right]^{1/qq'} < \infty \end{aligned}$$

by (2.4) and the Burkholder inequality. Note that

$$\alpha^{-1} = \exp \left\{ \int -\sigma^{-1}f \cdot dw^{u_0} - \frac{1}{2} \int |\sigma^{-1}f|^2 dt \right\}, \quad dw^{u_0} = dw - \sigma^{-1}f dt,$$

so that [6] implies that α^{-1} satisfies (2.4).

Hence for $u \in \mathcal{U}$ with the use of Girsanov's transformation

$$\begin{aligned}
 El(\alpha^u - \alpha^{u_0}) &= El\alpha^u \\
 &= E\alpha^u \int_0^1 \mathcal{X} dw^{u_0} \\
 &= E\alpha^u \left\{ \int_0^1 \mathcal{X}\sigma^{-1}g(u - u_0) dt + \int_0^1 \mathcal{X} dw^u \right\} \\
 &= E\alpha^u(1) \int_0^1 (u_0 - v)^*(u - u_0) dt \leq 0
 \end{aligned}$$

where the inequality follows from (vii). Moreover we have equality if and only if $u - u_0 \in \ker g$ a.e. $dt \times dP$, hence if $\alpha(u) = \alpha(u_0)$. This shows that $\alpha(u_0)$ is an exposed point.

We observe that (vii) means that if $\ker g$ is a subspace of dimension $d < r$, then $\ker g$ is parallel to a d -dimensional hyperplane which forms part of the boundary of Γ .

Although this description of the bang-bang densities cannot be extended to $\mathcal{A}(1)$, Corollary 5.2 can.

We say that μ_0 in $\mathcal{A}(1)$ is a \mathcal{W} exposed point if there exists a bounded measurable function ϕ such that for all $\mu \in \mathcal{A}(1)$, $\mu \neq \mu_0$,

$$\int_{R^n} \phi d\mu < \int_{R^n} \phi d\mu_0.$$

COROLLARY 5.3. $\mathcal{W} \exp [\mathcal{A}(1)] \subset \mathcal{A}_b(1)$.

Proof. Let $u_0 \in \mathcal{U}$ be such that $\mu(\alpha(1; u_0)) = \mu_0$ and let $[u_0] = \{u \in \mathcal{U} : \mu(\alpha(1; u)) = \mu_0\}$. Assume μ_0 is a \mathcal{W} exposed point. Then

$$\int_{\mathcal{C}} \phi[\pi_1(y)]\alpha(1; u) dP \leq \int_{\mathcal{C}} \phi[\pi_1(y)]\alpha(1; u_0) dP$$

with equality if and only if $u \in [u_0]$. Now $(u_0)_i$ is bang-bang except on S_i^l where $l = \phi \circ \pi_1$. Choose v such that v_i has support on S_i^l and $u^0 + v$ is bang-bang. By (5.2), $u_0 + v \in \mathcal{A}_b(1)$.

We have now characterized \mathcal{D}_b as the exposed points of \mathcal{D} in L_p . Let us next use Lemma 5.1 to show that bang-bang controls are optimal for many problems. First, if u_0 solves the problem of minimizing $E^u l(x) = El(x^u)$ over u in \mathcal{U} , where $l \in L_{2q}(\mathcal{C}, P)$, and where the time interval is fixed, say $0 \leq t \leq 1$, then l supports $\mathcal{D}(1)$ at α_0 . If u_0 is singular then we can change it to \hat{u} where $u_0 = \hat{u}$ if $(t, y) \notin S(u_0)$ such that $\hat{u} \in \mathcal{U}_b$, and \hat{u} is also optimal, i.e. $E^{u_0}l = E^{\hat{u}}l$; cf. (5.2). Hence the second of the three conclusions also holds.

Finally we apply Theorem 5.1 to the problem of § 3 with the equation (5.1). We assume

(C₁) for each t in $[0, T]$, $K(t)$ is an at most countable union of closed half-spaces, i.e.

$$K(t) = \bigcup_{i=1}^{\infty} \{\beta \in L_1(\mathcal{C}, P) : l_i^t \beta \geq c^i\}$$

such that for each i

$$\sup_t |l_t^i|_\infty < \infty, \lim_{s \nearrow t} \sup_{\alpha \in \mathcal{D}(s)} |l_t^i \alpha - l_s^i \alpha| = 0.$$

LEMMA 5.2. *Under hypothesis (C₁) if there exists a time optimal control \hat{u} , then $\alpha(\hat{t}; \hat{u})$ is a support point of $\mathcal{D}(\hat{t})$.*

Proof. By the weak compactness of $\mathcal{D}(\hat{t})$, $\sup \{l_t^i \alpha : \alpha \in \mathcal{D}(\hat{t})\}$ is attained, at say $\bar{\alpha}$.

Given $\varepsilon > 0$ and i we can choose $\delta^i > 0$ so that for each s , $\hat{t} - \delta^i < s \leq \hat{t}$, there is an $\alpha_s \in \mathcal{D}(s)$ with $\|\alpha_s - \bar{\alpha}\|_1 < \varepsilon / \sup_t |l_t^i|_\infty$, and with $|l_t^i \alpha_s - l_s^i \alpha_s| < \varepsilon$. Then

$$(5.4) \quad l_t^i \bar{\alpha} = l_t^i (\bar{\alpha} - \alpha_s) + (l_t^i - l_s^i) \alpha_s + l_s^i \alpha_s \leq 2\varepsilon + \sup_{\alpha \in \mathcal{D}(s)} l_s^i \alpha.$$

Since \hat{t} is the optimal time, then there exists at least one i such that $l_t^i \hat{\alpha} \geq c^i$, and $\sup_{\alpha \in \mathcal{D}(s)} l_s^i \alpha < c^i$ for $s < t$. This fact together with (5.4) implies that

$$(5.5) \quad c^i \leq l_t^i \hat{\alpha} \leq l_t^i \bar{\alpha} \leq c^i$$

and so l_t^i supports $\mathcal{D}(\hat{t})$ at $\hat{\alpha}$.

COROLLARY 5.4. *Under condition C₁, if there exists a time optimal control then it can be chosen to be bang-bang.*

Proof. The optimal density $\hat{\alpha}$ lies in the hyperplanes $\{\beta \in L_1 : l_t^i \beta = c^i\}$, $i \in J$. Hence \hat{u} is bang-bang except on a (t, y) set where each of these l_t^i are “degenerate”, i.e. $[(\sigma^{-1}g)^*(h^i - l_0^i \sigma^{-1}f^a)]_j = 0$, the same component j for all $i \in J$. We can change the values of \hat{u} on this set to ± 1 without changing the values of $l_t^i \alpha$, $i \in J$.

The analogous result holds in (Z, \mathcal{W}) using the half-spaces $\{\mu \in \mathcal{M} : \int_{R^n} l_t^i(x) d\mu(x) \geq c^i\}$ with l_t^i bounded, Borel measurable, because $t \mapsto \mathcal{A}(t)$ is also continuous when $\mathcal{A}(t)$ is considered in \mathcal{M} , i.e. we take the Hausdorff metric in \mathcal{M} .

We can now consider the examples from § 3 again. By taking $K(t) = \{\mu \in \mathcal{M} : \int_{R^n} 1_B c(t)^{-1} d\mu \geq 1\}$ in Example 1 we have not changed $K(t) \cap \mathcal{A}(t)$, the major point. (H₂) still holds. (C₁) holds if c is left continuous, $\inf_t c(t) > 0$.

As for Example 2: $l_t = 1_{\tau < t}$ and

$$\left| \int_{\mathcal{G}} (1_{\tau \leq t} - 1_{\tau \leq s}) \alpha dP \right| \leq K_0^{1/p} P\{s < \tau \leq t\}^{1/q} \rightarrow 0 \text{ as } s \nearrow t.$$

Hence (C₁) again holds and in either example an optimal control can be taken bang-bang. In fact from Lemma 5.2 it follows that if \hat{u} is optimal then a.e. $dt \times dP$

$$\max_{u \in \Gamma} H(\mathcal{X}, t, y, u) = H(\mathcal{X}, t, y, \hat{u}(t, y))$$

with $H(\mathcal{X}, t, y, u) = \mathcal{X}(t, y) \sigma(t, y)^{-1} g(t, y) u$ where for some i

$$l_t^i = E(l_t^i) + \int_0^t \mathcal{X}(t, y) dw^a(t, y),$$

i.e. we have a maximum principle.

REFERENCES

- [1] V. E. BENES, *Existence of optimal stochastic control laws*, this Journal, 9 (1971), pp. 446–472.
- [2] J. M. BISMUT, *Théorie probabiliste du contrôle des diffusions*, Mem. Amer. Math. Soc., 4 (1976), no. 167, pp. 1–130.
- [3] K. YAMADA, *Continuity of cost functionals in diffusion processes*, IEEE Conference on Decision and Control (Adaptive Processes), San Diego, 1973.
- [4] H. J. KUSHNER, *Existence results for optimal stochastic controls*, J. Optimization Theory Appl., 15 (1975), pp. 347–359.
- [5] W. M. WONHAM, *Liapunov criteria for weak stochastic stability*, J. Differential Equations, 2 (1966), pp. 195–207.
- [6] U. G. HAUSSMANN, *General necessary conditions for optimal control of stochastic systems*, Proceedings of the 1975 International Symposium on Stochastic Systems, Lexington, Kentucky; Mathematical Programming Studies, 5 (1976).
- [7] H. J. KUSHNER, *Necessary conditions for continuous parameter stochastic optimization problems*, this Journal, 10 (1972), pp. 550–565.
- [8] D. STROOCK AND S. R. S. VARADHAN, *Diffusion processes with continuous coefficients*, Comm. Pure Appl. Math., 22 (1969), pp. 345–400, 479–530.
- [9] M. NISIO, *On stochastic optimal control laws*, Nagoya Math. J., 52 (1973), pp. 1–30.
- [10] I. V. GIRSANOV, *On transforming a certain class of stochastic processes by absolutely continuous substitution of measures*, Theor. Probability Appl., 5 (1960), pp. 285–301.
- [11] J. MÉMIN, *Sur quelques problèmes fondamentaux de la théorie du filtrage*, thèse du troisième cycle, U.E.R. Mathématiques et Informatique, l'Université de Rennes, Rennes, France, 1974.
- [12] T. DUNCAN AND P. VARAIYA, *On the solution of a stochastic control system*, this Journal, 9 (1971), pp. 354–371.
- [13] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.
- [14] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, Part I, Interscience, New York, 1966.
- [15] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1968.
- [16] H. HERMES AND J. P. LASALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.
- [17] J. A. LINDENSTRAUSS, *A short proof of Lyapunov's convexity theorem*, J. Math. Mech., 15 (1966), pp. 971–972.
- [18] J. M. C. CLARK, *The representation of functionals of Brownian motion by stochastic integrals*, Ann. Math. Statist., 41 (1970), pp. 1282–1295.

THE OPTIMAL RECOURSE PROBLEM IN DISCRETE TIME: L^1 -MULTIPLIERS FOR INEQUALITY CONSTRAINTS*

R. T. ROCKAFELLAR† AND R. J-B. WETS‡

Abstract. An optimal recourse problem is an optimization problem with both stochastic and dynamic aspects, involving the interplay of observations and responses. In discrete time (with a finite horizon), there are finitely many stages, at each of which a decision is selected on the basis of prior observations of random events and subject to costs and constraints affected by these observations as well as past decisions. The goal is to minimize expected cost, taking into account the known distribution of future random events. This paper is concerned with the derivation of necessary and sufficient conditions for optimality in the case of convex costs and constraints.

It is shown that if the recourse problem is strictly feasible and satisfies a new condition called *essentially complete recourse*, optimal solutions can be characterized by a “pointwise” Kuhn–Tucker property involving L^1 -multipliers. Applications to multistage stochastic programs with special structures are developed in the last two sections of the paper. In particular, the relation between the general model and discrete-time stochastic control models is brought out by applying the basic results to a linear stochastic problem with state constraints.

1. Introduction. For $k = 1, \dots, N$, let $\xi_k \in R^{\nu_k}$ and $u_k \in R^{n_k}$ represent the observation and decision (control) associated with stage k of a sequential decision process. The sequence of observations

$$\xi = (\xi_1, \xi_2, \dots, \xi_N) \in R^{\nu_1} \times R^{\nu_2} \times \dots \times R^{\nu_N} = R^{\nu}$$

and the sequence of decisions

$$u = (u_1, u_2, \dots, u_N) \in R^{n_1} \times R^{n_2} \times \dots \times R^{n_N} = R^n$$

determine a “cost” denoted $f_0(\xi, u)$. The objective is to find a *recourse function* (or *policy*, or *decision rule*, or *control law*) $\xi \mapsto u(\xi)$ which minimizes the expected value of this cost subject to certain constraints, including a kind of nonanticipativity, i.e. the property that $u_k(\xi)$ essentially depends only on ξ_1, \dots, ξ_k . This is an *optimal recourse problem in discrete time*. Our aim here is to derive necessary and sufficient conditions for the optimality of a recourse function in the case of a problem satisfying convexity assumptions with respect to the decision variables.

To give a precise formulation, let $(\Xi, \mathcal{F}, \sigma)$ denote the sample space associated with the random elements of the problem; Ξ is a Borel subset of R^{ν} , \mathcal{F} is the Borel field on Ξ , and σ is a Borel probability measure on (Ξ, \mathcal{F}) . The corresponding expectation operator is denoted simply by E .

A function $u: \Xi \rightarrow R^n$ is said to be *nonanticipative* in the sequential framework described above if it is of the form

$$u(\xi) = (u_1(\xi_1), u_2(\xi_1, \xi_2), \dots, u_N(\xi_1, \dots, \xi_N));$$

it is *essentially nonanticipative* if it is measurable (with respect to \mathcal{F}) and differs

* Received by the editors September 9, 1976.

† Department of Mathematics, University of Washington, Seattle, Washington 98105. This research was sponsored by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under AFOSR Grant 72-2269.

‡ Department of Mathematics, University of Kentucky, Lexington, Kentucky 40506, and Université Scientifique et Médicale de Grenoble, Grenoble, France. This research was sponsored by the National Science Foundation under Grant MPS 75-07028.

only on a set of measure zero (with respect to σ) from some measurable nonanticipative function.

It is useful, for purposes of comparison with other work in stochastic optimization, to recognize that this concept of essential nonanticipativity can also be formulated in terms of a nest of sigma-fields. Let \mathcal{F}' denote the class of all sets in \mathcal{F} of measure zero with respect to σ , and for $k = 1, \dots, N$ let \mathcal{F}_k be the sigma-field generated by ξ_1, \dots, ξ_k completed with respect to σ , i.e. the class of all sets of \mathcal{F} of the form

$$((A \times [R^{v_{k+1}} \times \dots \times R^n]) \cap \Xi) \Delta B,$$

where A is a Borel set in $R^{v_1} \times \dots \times R^{v_k}$, B is a set in \mathcal{F}' , and Δ denotes symmetric difference. Then each \mathcal{F}_k is a sigma-field.

$$\mathcal{F}' \subset \mathcal{F}_1 \subset \dots \subset \mathcal{F}_N = \mathcal{F},$$

and a function $u: \Xi \rightarrow R^n$ is essentially nonanticipative if and only if for $k = 1, \dots, N$ the function $u_k: \Xi \rightarrow R^{n_k}$ is \mathcal{F}_k -measurable.

In fact, everything that follows remains valid for an arbitrary choice of sigma-fields $\mathcal{F}_1, \dots, \mathcal{F}_N$ nesting as indicated, if the latter property is adopted as the generalized definition of essential nonanticipativity. We therefore work mainly in this notational framework.

For the conditional expectation given \mathcal{F}_k , we write E^k . This is taken to be a *regular* conditional expectation, i.e. representable as an indefinite integral with respect to a regular conditional probability. (Such regular conditional probabilities exist, even for a general choice of \mathcal{F}_k , because \mathcal{F} is the Borel field on Ξ and σ is a regular Borel probability measure.)

The optimal recourse problem considered here consists of minimizing the expected cost

$$(1.1) \quad I_{f_0}(u) = E\{f_0(\xi, u(\xi))\}$$

over all essentially nonanticipative functions $u: \Xi \rightarrow R^n$ satisfying *almost surely* (a.s.)

$$(1.2) \quad f_i(\xi, u(\xi)) \leq 0, \quad i = 1, \dots, m,$$

and the abstract constraint $u(\xi) \in U(\xi)$. It is assumed that for every $\xi \in \Xi$ the set $U(\xi)$ is closed and convex with nonempty interior, and the functions $u \mapsto f_i(\xi, u)$, $i = 0, 1, \dots, m$, are defined for all $u \in U(\xi)$ (finite, i.e. real-valued), convex and lower semicontinuous. It is assumed further that for each $u \in R^n$ the set

$$U^{-1}(u) = \{\xi \in \Xi | u \in U(\xi)\}$$

is Borel measurable (i.e. belongs to \mathcal{F}) and the functions $\xi \mapsto f_i(\xi, u)$ are all Borel measurable relative to $U^{-1}(u)$. Setting

$$f_i(\xi, u) = +\infty \quad \text{if } u \notin U(\xi),$$

we obtain from these assumptions that each f_i is a normal convex integrand on $\Xi \times R^n$ [1, Lemma 2] and the multifunction $U: \Xi \rightarrow R^n$ is measurable [2, Cor. 3.1].

It follows that $f_i(\xi, u(\xi))$ is Borel measurable in $\xi \in \Xi$ when $u(\xi)$ is measurable

[1, Cor. to lemma 5]. Moreover, the multifunction

$$(1.3) \quad D: \xi \rightarrow D(\xi) = \{u \in U(\xi) \mid f_i(\xi, u) \leq 0, i = 1, \dots, m\}$$

is measurable [2, Cors. 4.1 and 4.3]. This multifunction with closed, convex values provides an abstract description of the constraint structure, and it is crucial in what follows.

We assume that the sets $D(\xi)$ are uniformly bounded (i.e. their union for all $\xi \in \Xi$ is a bounded subset of R^n). This enables us to restrict our attention in the recourse problem to functions u belonging to the space $L_n^\infty = L^\infty(\Xi, \mathcal{F}, \sigma; R^n)$. We suppose in addition that to each bounded set $K \subset R^n$ there corresponds a summable function $\alpha: \Xi \rightarrow R$ and a constant $\beta \in R$ such that

$$(1.4) \quad |f_0(\xi, u)| \leq \alpha(\xi) \quad \text{for all } u \in U(\xi) \cap K,$$

$$(1.5) \quad |f_i(\xi, u)| \leq \beta \quad \text{for all } u \in U(\xi) \cap K, \quad i = 1, \dots, m.$$

These “growth” conditions imply that for every function u in the class

$$\mathcal{U} = \{u \in L_n^\infty \mid u(\xi) \in U(\xi) \text{ a.s.}\}$$

the functions $f_i(\cdot, u(\cdot))$, $i = 1, \dots, m$, are essentially bounded, while $f_0(\cdot, u(\cdot))$ is summable.

With these assumptions the optimal recourse problem introduced above is well-defined and can be stated as:

P Minimize the functional (1.1) over all $u \in \mathcal{U} \cap \mathcal{N}_\infty$ satisfying (1.2) a.s., where \mathcal{N}_∞ represents the constraint of nonanticipativity:

$$\begin{aligned} \mathcal{N}_\infty &= \{u = (u_1, \dots, u_N) \in L_n^\infty \mid u_k \text{ is } \mathcal{F}_k\text{-measurable}, k = 1, \dots, N\} \\ &= L_{n_1}^\infty(\Xi, \mathcal{F}_1, \sigma) \times L_{n_2}^\infty(\Xi, \mathcal{F}_2, \sigma) \times \dots \times L_{n_N}^\infty(\Xi, \mathcal{F}_N, \sigma). \end{aligned}$$

Clearly \mathcal{N}_∞ is a linear subspace of \mathcal{L}_n^∞ , while \mathcal{U} is a convex set, as is the class of all $u \in \mathcal{U}$ satisfying (1.2) a.s. The functional (1.1) is convex and finite on \mathcal{U} . Thus we are dealing with a convex optimization problem. In such a setting, it is typical to find multiplier characterizations of optimality which are always sufficient but not necessary without some “constraint qualification.”

A natural constraint qualification to consider is that P be *strictly feasible*. This is taken to mean that there exist $\tilde{u} \in \mathcal{N}_\infty$ and $\varepsilon > 0$ such that

$$(1.6) \quad f_i(\xi, \tilde{u}(\xi)) \leq -\varepsilon \quad \text{a.s. for } i = 1, \dots, m,$$

and

$$(1.7) \quad \tilde{u}(\xi) + \varepsilon B \subset D(\xi) \quad \text{a.s.,}$$

where B is the closed unit ball in R^n . However, strict feasibility is not enough in itself. What we need for our characterization of optimality, as it turns out, is for P also to have the property of *essentially complete recourse*, in the sense that for $k = 1, \dots, N$ the multifunction

$$\begin{aligned} (1.8) \quad D^k: \xi \rightarrow D^k(\xi) &= \{(u_1, \dots, u_k) \mid u \in D(\xi)\} \\ &= \text{projection of } D(\xi) \text{ on } R^{n_1} \times \dots \times R^{n_k} \end{aligned}$$

is \mathcal{F}_k -measurable. (In this case, the constraint multifunction D is said to be *essentially nonanticipative*.) Henceforth, we assume the problem P to be endowed with both strict feasibility and essentially complete recourse, as well as all other properties of U, f_i and D already mentioned.

The optimality condition to be studied below involves the function

$$h: \Xi \times R^n \times R_+^m \times R^n \rightarrow R$$

defined by

$$(1.9) \quad h(\xi, u, y, p) = f_0(\xi, u) + \sum_{i=1}^m y_i f_i(\xi, u) - u \cdot p.$$

This acts much like the Hamiltonian in control theory.

The *Lagrangian* associated with the problem P is defined to be the function

$$(1.10) \quad I_h(u, y, p) = E\{h(\xi, u(\xi), y(\xi), p(\xi))\} \quad \text{for } (u, y, p) \in \mathcal{U} \times \mathcal{Y} \times \mathcal{M}_1,$$

where

$$\mathcal{Y} = \{y = (y_1, \dots, y_m) \in L_m^1 | y_i(\xi) \geq 0 \text{ a.s. for } i = 1, \dots, m\},$$

$$\mathcal{M}_1 = \{p = (p_1, \dots, p_N) \in L_N^1 | E^k\{p_k(\xi)\} = 0 \text{ a.s. for } k = 1, \dots, N\}.$$

(Here $p_k(\xi) \in R^{n_k}$.) The set \mathcal{Y} is convex, while \mathcal{M}_1 is a linear subspace. In fact, as is easy to verify from the definitions, \mathcal{M}_1 and \mathcal{N}_∞ are complementary to each other with respect to the natural pairing between L_n^1 and L_n^∞ .

$$\mathcal{M}_1 = \mathcal{N}_\infty^\perp \quad \text{and} \quad \mathcal{N}_\infty = \mathcal{M}_1^\perp.$$

Our growth conditions on the functions f_i imply that $I_h(u, y, p)$ is finite throughout $\mathcal{U} \times \mathcal{Y} \times \mathcal{M}_1$, and, of course, convex in u and affine in (y, p) .

A *saddle point* of I_h with respect to minimization in u and maximization in (y, p) is an element $(\bar{u}, \bar{y}, \bar{p})$ of $\mathcal{U} \times \mathcal{Y} \times \mathcal{M}_1$ satisfying

$$(1.11) \quad I_h(\bar{u}, y, p) \leq I_h(\bar{u}, \bar{y}, \bar{p}) \leq I_h(u, \bar{y}, \bar{p}) \quad \text{for all } (u, y, p) \in \mathcal{U} \times \mathcal{Y} \times \mathcal{M}_1.$$

We shall prove in § 2 that the regularity conditions imposed on P ensure the existence of such a saddle point $(\bar{u}, \bar{y}, \bar{p})$, with \bar{u} an optimal solution to P and (\bar{y}, \bar{p}) an optimal solution to an associated dual problem. (See [3] for a general exposition of the relation between the saddle points of a Lagrangian and the optimal solutions of the corresponding convex program and its dual.)

As is also shown in § 2, the saddle points $(\bar{u}, \bar{y}, \bar{p})$ of I_h are characterized by the following Kuhn–Tucker conditions, whose satisfaction for some (\bar{y}, \bar{p}) is therefore necessary and sufficient for the optimality of \bar{u} in P:

(a) $\bar{u} \in \mathcal{N}_\infty$ and

$$(1.12) \quad \bar{u}(\xi) \in U(\xi) \text{ almost surely}$$

$$(1.13) \quad f_i(\xi, \bar{u}(\xi)) \leq 0 \quad \text{for } i = 1, \dots, m \quad \text{almost surely};$$

(b) $\bar{y} = (\bar{y}_1, \dots, \bar{y}_m) \in L_m^1$ and

$$(1.14) \quad \bar{y}_i(\xi) \geq 0 \quad \text{for } i = 1, \dots, m \quad \text{almost surely},$$

$$(1.15) \quad \bar{y}_i(\xi) f_i(\xi, \bar{u}(\xi)) = 0 \quad \text{for } i = 1, \dots, m \quad \text{almost surely};$$

(c) $\bar{p} \in \mathcal{M}_1$ and

$$(1.16) \quad h(\xi, \bar{u}(\xi), \bar{y}(\xi), \bar{p}(\xi)) = \lim_{u \in U(\xi)} h(\xi, u, \bar{y}(\xi), \bar{p}(\xi)) \quad \text{almost surely.}$$

The Kuhn–Tucker conditions show that if \bar{y} and \bar{p} , the multipliers associated with P, are known or can be generated by an algorithmic procedure, a function $\bar{u} \in L_n^\infty$ is optimal for P if and only if it is nonanticipative and $\bar{u}(\xi)$ satisfies certain constraints “pointwise” for each $\xi \in \Xi$, namely (1.12), (1.16), and (1.13) with equality holding when $\bar{y}_i(\xi) > 0$. *Moreover, if P is such that the pointwise minimum in (1.16) is almost surely unique, as is true for example if $f_0(\xi, \cdot)$ is almost surely convex on $U(\xi)$, then the function $\bar{u} \in L_n^\infty$ is optimal if it merely satisfies (1.12) and (1.16), without regard to nonanticipativity and the other constraints.* Indeed, these other properties must then hold automatically for \bar{u} , since according to the above there does exist at least one optimal recourse function characterized by the Kuhn–Tucker conditions. This is discussed further in a more specialized context in § 3.

Essentially complete recourse plays a vital role in the derivation of these results. The importance of this kind of property was first brought out in [4] in connection with our work on a special case of P. It was shown in [4] that if a stochastic program with a two-stage constraint structure has *relatively complete recourse*, the multipliers appearing in the Kuhn–Tucker conditions may be chosen to be L^1 -functions; one has to rely on esoteric elements of $(L^\infty)^*$ when this condition is not satisfied. It can be shown that essentially complete recourse is implied by relatively complete recourse in that setting (see the remarks in § 3 following Theorem 6). Essentially complete recourse is a more general and abstract condition demanding that at each stage k the set from which the decision u_k must be chosen, namely

$$D_k(\xi, u_1, \dots, u_{k-1}) = \{u_k \in R^{n_k} | (u_1, \dots, u_{k-1}, u_k) \in D^k(\xi)\},$$

really depends only on past decisions and observations, and one therefore does not have to restrict further to an intersection relative to all possible future observations (an implicit constraint induced by the need to maintain availability of recourse under all circumstances).

In a companion paper [5], essentially complete recourse was used extensively, first in the justification of the dynamic programming technique for optimal recourse problems, but then also to obtain a system of L^1 -multipliers, in fact a summable martingale, that can be associated with the nonanticipativity restriction on the recourse functions. However, our concern in [5] was only with such multipliers. The model was formulated directly in terms of the nonanticipative constraint multifunction D ; no structure of D in terms of inequality constraints as in (1.3) was explicitly introduced, and hence there was no multiplier vector $y(\xi)$. The existence of multipliers associated with the nonanticipativity restriction was first pointed out in [6].

2. Basic results. Our first theorem shows that the regularity conditions imposed on the recourse problem P guarantee the existence of an optimal solution \bar{u} , and that such functions \bar{u} correspond to saddle-points $(\bar{u}, \bar{y}, \bar{p})$ of I_h . We proceed by observing that the question can be settled through reducing P to an

equivalent problem without explicit inequality constraints. We then utilize the key result of [5] to complete the proof. The second theorem demonstrates that the saddle points of I_h can be characterized by the Kuhn–Tucker conditions, and these therefore furnish necessary and sufficient conditions for optimality. The third theorem brings in the corresponding dual problem D.

THEOREM 1. *The Lagrangian I_h has at least one saddle point $(\bar{u}, \bar{y}, \bar{p})$ relative to $\mathcal{U} \times \mathcal{Y} \times \mathcal{M}_1$. Moreover, the components \bar{u} of such saddle points are precisely the optimal recourse functions in P.*

Proof. First observe that P consists of minimizing over \mathcal{N}_∞ the functional

$$I_f(u) = E\{f(\xi, u(\xi))\} = E\{f_0(\xi, u(\xi)) + \psi_{D(\xi)}(u(\xi))\},$$

where $\psi_{D(\xi)}$ is the indicator of $D(\xi)$. Since D is a measurable multifunction and f_0 is a normal convex integrand, we know f is a normal convex integrand [2, Thm. 2 and Cor. 4.2].

According to our assumptions, $D(\xi)$ is uniformly bounded and there is a summable function $\alpha: \Xi \rightarrow R$ such that

$$u \in D(\xi) \Rightarrow |f(\xi, u)| \leq \alpha(\xi).$$

Furthermore, by strict feasibility there exist $\tilde{u} \in \mathcal{N}_\infty$ and $\varepsilon > 0$ such that (1.7) holds.

These facts put us in the framework of [5, Thm. 2] and furnish not only the existence of an optimal solution \bar{u} to P but also the characterization of such a function \bar{u} as the first component of a saddle point (\bar{u}, \bar{p}) of the *reduced Lagrangian*

$$(2.1) \quad L(u, p) = E\{f(\xi, u(\xi)) - u(\xi) \cdot p(\xi)\} = I_f(u) - \langle u, p \rangle \quad \text{for } (u, p) \in L_n^1 \times \mathcal{M}_1.$$

The existence of an optimal solution is seen as follows. The subspace \mathcal{N}_∞ , being representable as

$$\mathcal{M}_1^\perp = \{u \in L_n^1 \mid \langle u, p \rangle = 0 \text{ for all } p \in \mathcal{M}_1\},$$

is closed in the weak topology $w(L_n^\infty, L_n^1)$. The functional I_f on L_n^∞ is lower semicontinuous in this topology, because it is representable as the conjugate of the functional I_{f^*} on L_n^1 , where f^* is the conjugate integrand ([1, Thm. 2] and [7, Thm. 2]). The sets

$$\{u \in \mathcal{N}_\infty \mid I_f(u) \leq \mu\}, \quad \mu \in R,$$

are therefore closed in this topology, in fact compact by the uniform boundedness of $D(\xi)$, since

$$(2.2) \quad I_f(u) < +\infty \Rightarrow u(\xi) \in D(\xi) \quad \text{a.s.}$$

The nonempty sets in this nest of compact sets therefore have a nonempty intersection, and this consists obviously of optimal solutions to P.

The existence of the multiplier \bar{p} in [5, Thm. 2] is obtained by a more subtle argument, the details of which will not be repeated here. By our hypothesis, the convex functional I_f is finite and norm-continuous at a certain point \tilde{u} of \mathcal{N}_∞ , and this furnishes by Fenchel's duality theorem a norm-continuous linear functional φ

on L_n^∞ such that φ vanishes on \mathcal{N}_∞ and

$$\inf_{u \in \mathcal{N}_\infty} I_f(u) = \inf_{u \in L_n^\infty} \{I_f(u) - \varphi(u)\}.$$

The property of essentially complete recourse enters in showing that φ can actually be taken to be of the form $\varphi(u) = \langle u, \bar{p} \rangle$ for some \bar{p} in L_n^1 (and hence in $\mathcal{M}_1 = \mathcal{N}_\infty^\perp$). This yields the existence of at least one saddle point (\bar{u}, \bar{p}) of L in (2.1), and it follows then by the usual reasoning in minimax theory that such saddle points characterize the optimal solutions \bar{u} to P.

To complete the proof of Theorem 1, we must show that a pair (\bar{u}, \bar{p}) is a saddle point of the reduced Lagrangian L if and only if there exists $\bar{y} \in \mathcal{Y}$ such that $(\bar{u}, \bar{y}, \bar{p})$ is a saddle point of the Lagrangian I_h . The sufficiency of this condition is obvious from the fact that

$$(2.3) \quad L(u, p) = \sup_{y \in \mathcal{Y}} I_h(u, y, p).$$

(In view of (2.2), there is no loss of generality in replacing L_n^∞ by \mathcal{U} in discussing saddle points of I_h).

Now consider any saddle point (\bar{u}, \bar{p}) of L . We have $\bar{u} \in \mathcal{U}$ and

$$(2.4) \quad L(\bar{u}, \bar{p}) = \sup_{p \in \mathcal{M}_1} L(\bar{u}, p) = \sup_{p \in \mathcal{M}_1} \sup_{y \in \mathcal{Y}} I_h(\bar{u}, y, p),$$

while on the other hand, using the fact already noted that the conjugate of I_f is I_{f^*} on L_n^1 , we have

$$(2.5) \quad I_f(\bar{u}) - \langle \bar{u}, \bar{p} \rangle = L(u, p) = \inf_{u \in L_n^\infty} L(u, \bar{p}) = \inf_{u \in L_n^\infty} \{I_f(u) - \langle u, \bar{p} \rangle\} = -I_{f^*}(\bar{p}),$$

where by definition

$$(2.6) \quad -f^*(\xi, \bar{p}(\xi)) = \inf_{u \in \mathbb{R}^n} \{f(\xi, u) - u \cdot \bar{p}(\xi)\}.$$

In order to verify for some $\bar{y} \in \mathcal{Y}$ that $(\bar{u}, \bar{y}, \bar{p})$ is a saddle point of I_h , it suffices in view of (2.4) to establish that

$$I_h(u, \bar{y}, \bar{p}) \geq L(\bar{u}, \bar{p}) \quad \text{for all } u \in \mathcal{U},$$

or in other words that

$$(2.7) \quad E\{h(\xi, u(\xi), \bar{y}(\xi), \bar{p}(\xi))\} \geq E\{f(\xi, \bar{u}(\xi)) - \bar{u}(\xi) \cdot \bar{p}(\xi)\} \quad \text{for all } u \in \mathcal{U}.$$

We know from (2.5) and (2.6) that

$$f(\xi, \bar{u}(\xi)) - \bar{u}(\xi) \cdot \bar{p}(\xi) = \inf_{u \in \mathbb{R}^n} \{f(\xi, u) - u \cdot \bar{p}(\xi)\} \quad \text{almost surely.}$$

Thus $\bar{u}(\xi)$ is almost surely an optimal solution to the convex programming problem

$$(2.8) \quad \begin{aligned} &\text{minimize } f_0(\xi, u) - u \cdot p(\xi) \quad \text{over all } u \in U(\xi) \\ &\text{satisfying } f_i(\xi, u) \leq 0 \quad \text{for } i = 1, \dots, m. \end{aligned}$$

However, this problem is strictly feasible almost surely, due to the assumed existence of $\tilde{u} \in \mathcal{U}$ and $\varepsilon > 0$ satisfying (1.6), and it therefore has almost surely a Kuhn–Tucker vector, i.e. a vector $y \in R_+^m$ such that (cf. (1.9)):

$$\inf_{u \in U(\xi)} h(\xi, u, y, \bar{p}(\xi)) = \inf \text{ in (2.8)} = f(\xi, \bar{u}(\xi)) - \bar{u}(\xi) \cdot \bar{p}(\xi).$$

Let $Y(\xi)$ denote the set of all vectors $y \in R_+^m$ such that

$$(2.9) \quad h(\xi, u, y, \bar{p}(\xi)) \geq f(\xi, \bar{u}(\xi)) - \bar{u}(\xi) \cdot \bar{p}(\xi) \quad \text{for all } u \in U(\xi).$$

As we have just seen, $Y(\xi) \neq \emptyset$ almost surely. Let A denote a countable dense subset of R^n . Since for each $y \in R_+^m$ the function $h(\xi, \cdot, y, \bar{p}(\xi))$ is finite, lower semicontinuous (l.s.c.), and convex on $U(\xi)$ (a convex set with nonempty interior), it is continuous on the interior of $U(\xi)$ and relative to all line segments in $U(\xi)$, and hence

$$\inf_{u \in U(\xi)} h(\xi, u, y, \bar{p}(\xi)) = \inf_{u \in U(\xi) \cap A} h(\xi, u, y, \bar{p}(\xi)).$$

Thus $U(\xi)$ can be replaced by $U(\xi) \cap A$ in (2.9) without affecting the nature of the condition on y . This yields the representation

$$(2.10) \quad Y(\xi) = \bigcap_{a \in A} Y_a(\xi),$$

where $Y_a(\xi)$ denotes the set of all $y \in R_+^m$ satisfying

$$h(\xi, a, y, \bar{p}(\xi)) \geq f(\xi, \bar{u}(\xi)) - \bar{u}(\xi) \cdot \bar{p}(\xi),$$

or more specifically, is given for each ξ in the (Borel measurable) set $U^{-1}(a)$ by

$$Y_a(\xi) = \left\{ y \in R_+^m \mid \sum_{i=1}^m y_i f_i(\xi, a) \geq f_0(\xi, \bar{u}(\xi)) - f_0(\xi, a) \right\},$$

while for other $\xi \in \Xi$ simply $Y_a(\xi) = R_+^m$. Each of the multifunctions $Y_a: \xi \rightarrow Y_a(\xi)$ is close-valued and Borel measurable [2, Cor. 4.3], and hence so is Y as the intersection of a countable collection in (2.10) [2, Cor. 1.3]. It follows that Y has a Borel measurable selection where it is nonempty-valued [2, Cor. 1.1]. Since $Y(\xi) \neq \emptyset$ almost surely, we therefore have the existence of a Borel measurable function $\bar{y}: \Xi \rightarrow R_+^m$ such that almost surely $\bar{y}(\xi) \in Y(\xi)$, i.e.

$$(2.11) \quad h(\xi, u, \bar{y}(\xi), \bar{p}(\xi)) \geq f(\xi, \bar{u}(\xi)) - \bar{u}(\xi) \cdot \bar{p}(\xi) \quad \text{for all } u \in U(\xi).$$

We claim (2.11) implies $\bar{y}(\xi)$ is summable in ξ , so that actually $\bar{y} \in \mathcal{Y}$. Indeed, for the function \tilde{u} in our strict feasibility assumption we can set $u = \tilde{u}(\xi)$ in (2.11) to obtain (almost surely)

$$\begin{aligned} f_0(\xi, \tilde{u}(\xi)) - \varepsilon \sum_{i=1}^m \bar{y}_i(\xi) - \tilde{u}(\xi) \cdot \bar{p}(\xi) &\geq f_0(\xi, \tilde{u}(\xi)) - \tilde{u}(\xi) \cdot \bar{p}(\xi) + \sum_{i=1}^m \bar{y}_i(\xi) f(\xi, \tilde{u}(\xi)) \\ &\geq f(\xi, \bar{u}(\xi)) - \bar{u}(\xi) \cdot \bar{p}(\xi) \\ &= f_0(\xi, \bar{u}(\xi)) - \bar{u}(\xi) \cdot \bar{p}(\xi) \end{aligned}$$

and thus for $i = 1, \dots, m$ (almost surely)

$$(2.12) \quad 0 \leq \varepsilon \bar{y}_i(\xi) \leq f_0(\xi, \bar{u}(\xi)) - f_0(\xi, \bar{u}(\xi)) - (\bar{u}(\xi) - \bar{u}(\xi)) \cdot \bar{p}(\xi).$$

The right side of (2.12) is, of course, summable in ξ , and hence so is $\bar{y}_i(\xi)$.

We have thus established the existence of $\bar{y} \in \mathcal{Y}$ satisfying (2.11). But (2.11) implies (2.7) and therefore, as already argued, that $(\bar{u}, \bar{y}, \bar{p})$ is a saddle point of I_h . This ends the proof of Theorem 1.

COROLLARY. *The restricted Lagrangian*

$$(2.13) \quad I_\ell(u, y) = E\{\ell(\xi, u(\xi), y(\xi))\} \quad \text{for } (u, y) \in (\mathcal{U} \cap \mathcal{N}_\infty) \times \mathcal{Y},$$

where

$$(2.14) \quad \ell(\xi, u, y) = f_0(\xi, u) + \sum_{i=1}^m y_i f_i(\xi, u),$$

has at least one saddle point (\bar{u}, \bar{y}) relative to $(\mathcal{U} \cap \mathcal{N}_\infty) \times \mathcal{Y}$. Moreover, the components \bar{u} of such saddle points are precisely the optimal recourse functions in P.

Proof. Let $(\bar{u}, \bar{y}, \bar{p})$ be a saddle point of I_h relative to $\mathcal{U} \times \mathcal{Y} \times \mathcal{M}_1$, as exists by Theorem 1. Since $\bar{p} \in \mathcal{M}_1 = \mathcal{N}_\infty^\perp$, we have

$$I_h(u, y, \bar{p}) = I_\ell(u, y) \quad \text{for } (u, y) \in (\mathcal{U} \cap \mathcal{N}_\infty) \times \mathcal{Y},$$

and hence (\bar{u}, \bar{y}) is a saddle point of I_ℓ relative to $(\mathcal{U} \cap \mathcal{N}_\infty) \times \mathcal{Y}$. The existence of at least one such saddle point, together with the fact that P is equivalent to minimizing the functional

$$I_f(u) = \sup_{y \in \mathcal{Y}} I_\ell(u, y) \quad \text{for } u \in \mathcal{U} \cap \mathcal{N}_\infty,$$

yields the characterization of solutions \bar{u} by the usual minimax considerations.

THEOREM 2. *An element $(\bar{u}, \bar{y}, \bar{p})$ is a saddle point of the Lagrangian I_h relative to $\mathcal{U} \times \mathcal{Y} \times \mathcal{M}_1$ if and only if the Kuhn–Tucker conditions (a), (b), (c) are satisfied.*

Proof. In either case we have $\bar{u} \in \mathcal{U}$, $\bar{y} \in \mathcal{Y}$ and $\bar{p} \in \mathcal{M}_1$. If $(\bar{u}, \bar{y}, \bar{p})$ is a saddle point, then \bar{u} is optimal for P by Theorem 1, and in particular $\bar{u} \in \mathcal{N}_\infty$. Thus in showing the equivalence we can limit attention to the case where also $\bar{u} \in \mathcal{N}_\infty$. Then $\langle \bar{u}, p \rangle = 0$ for all $p \in \mathcal{M}_1$, so that $I_h(\bar{u}, y, p) = I_h(\bar{u}, y, \bar{p})$, and the saddle point condition can just as well be expressed as

$$(2.15) \quad \sup_{y \in \mathcal{Y}} I_h(\bar{u}, y, \bar{p}) = I_h(\bar{u}, \bar{y}, \bar{p}) = \inf_{u \in \mathcal{U}} I_h(u, \bar{y}, \bar{p}).$$

The left half of (2.13) is trivially equivalent to

$$\sup_{y \in R^n} h(\xi, \bar{u}(\xi), y, \bar{p}(\xi)) = h(\xi, \bar{u}(\xi), \bar{y}(\xi), \bar{p}(\xi)) \quad \text{a.s.,}$$

and this is identical to (1.13) plus (1.15).

It remains only to show that the second equality in (2.15) implies (1.16), the opposite implication being immediate. Define the integrand j on $\Xi \times R^n$ by

$$j(\xi, u) = h(\xi, u, \bar{y}(\xi), \bar{p}(\xi)),$$

this value being interpreted as $+\infty$ for $u \notin U(\xi)$, so that

$$U(\xi) = \{u \in R^n \mid j(\xi, u) < +\infty\}.$$

Our hypotheses say that $j(\xi, u)$ is l.s.c. convex in u and Borel measurable in ξ , hence (since $\text{int } U(\xi) \neq \emptyset$) j is a (Borel) normal convex integrand [1, Lemma 2]. Furthermore, the "growth" conditions on the functions f_i imply for each bounded set $K \subset R^n$ the existence of a summable function $\gamma: \Xi \rightarrow R$ such that

$$|j(\xi, u)| \leq \gamma(\xi) \quad \text{for all } u \in U(\xi) \cap K.$$

The right half of (2.15) thus can be regarded as the assertion that

$$(2.16) \quad I_j(\bar{u}) = \inf_{u \in L_n^\infty} I_j(u),$$

where

$$I_j(u) = E\{j(\xi, u(\xi))\}.$$

On the other hand, (1.16) can be restated as

$$(2.17) \quad j(\xi, \bar{u}(\xi)) = \inf_{u \in R^n} j(\xi, u) \quad \text{a.s.}$$

The question is thus reduced to that of the equivalence of (2.16) and (2.17), which is answered affirmatively by the theory of normal integrands and integral functionals. (In particular, the two properties can be expressed in terms of $0 \in \partial I_j(\bar{u})$ and $0 \in \partial j(\xi, \bar{u}(\xi))$, and then [7, Cor. 1B and Thm. 2] can be invoked.) Theorem 2 is thereby established.

We have mentioned in § 1 that the multipliers \bar{y} and \bar{p} for P solve a certain dual problem. This will now be described. Define the function g on $\Xi \times R^m \times R^n$ by

$$(2.18) \quad g(\xi, y, p) = \begin{cases} \inf_{u \in U(\xi)} h(\xi, u, y, p) & \text{if } y \in R_+^m, \\ -\infty & \text{if } y \notin R_+^m. \end{cases}$$

It will be shown below that $-g$ is a normal convex integrand. Let

$$(2.19) \quad I_g(y, p) = E\{g(\xi, y(\xi), p(\xi))\} \quad \text{for } (y, p) \in L_m^1 \times L_n^1.$$

The *dual problem* associated with P is taken to be:

$$D \quad \text{Maximize } I_g(y, p) \text{ over all } (y, p) \in \mathcal{Y} \times \mathcal{M}_1.$$

THEOREM 3. *The functional I_g in D is well-defined and concave, with*

$$(2.20) \quad I_g(y, p) = \inf_{u \in \mathcal{U}} I_h(u, y, p) \quad \text{for all } (y, p) \in \mathcal{Y} \times \mathcal{M}_1.$$

Thus optimal solutions to D exist, and they are precisely the components (\bar{y}, \bar{p}) of the saddle points $(\bar{u}, \bar{y}, \bar{p})$ of the Lagrangian I_h . In particular,

$$\min P = \max D.$$

Proof. We begin by proving that $-g$ is a (Borel-)normal convex integrand. There exists in \mathcal{U} a countable subcollection \mathcal{U}' such that $U(\xi)$ is almost surely the closure of the set $\{u(\xi) | u \in \mathcal{U}'\}$. (This follows from the measurability of the multifunction U via Castaing's theorem; cf. [2, Thm. 1].) Then by convexity

$$(2.21) \quad g(\xi, y, p) = \inf_{u \in \mathcal{U}'} h(\xi, u(\xi), y, p) \quad \text{a.s.} \quad \text{for } y \in R_+^m.$$

For each $u \in \mathcal{U}'$, define

$$g_u(\xi, y, p) = \begin{cases} h(\xi, u(\xi), y, p) & \text{if } y \in R_+^m, \\ -\infty & \text{if } y \notin R_+^m. \end{cases}$$

Then $-g$ is a normal convex integrand by virtue of our regularity assumptions, and we have from (2.21) the representation

$$g(\xi, y, p) = \inf_{u \in \mathcal{U}'} g_u(\xi, y, p) \quad \text{a.s.}$$

Since the collection is countable, this implies $-g$ is a normal convex integrand [2, Cor. 4.1].

Normality ensures that $g(\xi, y(\xi), p(\xi))$ is measurable in ξ whenever $y(\cdot)$ and $p(\cdot)$ are. On the other hand, fixing any $u \in \mathcal{U}$ we have for all $y \in L_m^1$ and $p \in L_n^1$ the bound

$$g(\xi, y(\xi), p(\xi)) \leq f_0(\xi, u(\xi)) + \sum_{i=1}^m y_i(\xi) f_i(\xi, u(\xi)) - u(\xi) \cdot p(\xi),$$

where the right side is summable. Thus $I_g(y, p)$ is always unambiguously a real number or $-\infty$. The concavity of I_g is obvious.

We establish (2.20) by fixing any (y, p) in $\mathcal{Y} \times \mathcal{M}_1$ and considering the integrand

$$(2.22) \quad j(\xi, u) = \begin{cases} h(\xi, u, y(\xi), p(\xi)) & \text{if } u \in U(\xi), \\ +\infty & \text{if } u \notin U(\xi). \end{cases}$$

The situation is extremely close to the one at the end of the proof of Theorem 2; j is a normal convex integrand, and we get from the theory of integral functionals that

$$(2.23) \quad \sup_{u \in L_n^\infty} \{ \langle q, u \rangle - I_j(u) \} = I_j^*(q) \quad \text{for all } q \in L_n^1,$$

where

$$j^*(\xi, q(\xi)) = \sup_{u \in R^n} \{ q(\xi) \cdot u - j(\xi, u) \}.$$

Taking $q = 0$, we turn the latter into

$$-j^*(\xi, 0) = g(\xi, p(\xi), y(\xi))$$

by (2.22) and (2.18), and then (2.23) becomes the equation in (2.20).

The rest of Theorem 3 is evident from (2.20) and the existence of a saddle point of I_h in Theorem 1.

3. Special structures. So far, it has been convenient and useful to endow P with as little structure as possible. This level of generality is rarely, if ever, needed in practice. The main purpose of this section, and the next one, is to consider recourse problems that possess some of the structural characteristics most commonly encountered in applications.

An initial observation may be made about the differentiable case, i.e. where $U(\xi) = R^n$ and the functions $u \mapsto f_i(\xi, u)$ are all differentiable with gradients denoted by $\nabla f_i(\xi, u)$. Then (1.16) of the Kuhn–Tucker conditions becomes

$$(3.1) \quad \nabla f_0(\xi, \bar{u}(\xi)) + \sum_{i=1}^m \bar{y}_i(\xi) \nabla f_i(\xi, \bar{u}(\xi)) = \bar{p}(\xi) \quad \text{a.s.},$$

and hence part (c) of the conditions asserts simply that

$$(3.2) \quad E^k \left\{ \nabla f_0(\xi, \bar{u}(\xi)) + \sum_{i=1}^m \bar{y}_i(\xi) \nabla f_i(\xi, \bar{u}(\xi)) \right\} = 0 \quad \text{a.s.} \quad \text{for } k = 1, \dots, N.$$

A. The separable case. By SP we denote a version of P that satisfies all the regularity conditions laid out in § 1 and is also *separable*, by which we mean that

$$(i) \quad U(\xi) = \bigtimes_{k=1}^N U_k(\xi),$$

$$(ii) \quad f_i(\xi, u) = \sum_{k=1}^N f_{ik}(\xi, u_k) \quad \text{for } i = 0, 1, \dots, m,$$

where the multifunctions $U_k: \xi \mapsto U_k(\xi) \subset R^{n_k}$ are \mathcal{F}_k -measurable, and the functions $\xi \mapsto f_{ik}(\xi, u_k)$ are \mathcal{F}_k -measurable relative to the set

$$U_k^{-1}(u_k) = \{\xi \in \Xi \mid u_k \in U_k(\xi)\} \in \mathcal{F}_k.$$

The function h (as defined by (1.9)) is also separable, in the sense that

$$(3.3) \quad h(\xi, u, y, p) = \sum_{k=1}^N [\ell_k(\xi, u_k, y) - u_k \cdot p_k],$$

where

$$(3.4) \quad \ell_k(\xi, u_k, y) = f_{0k}(\xi, u_k) + \sum_{i=1}^m y_i f_{ik}(\xi, u_k)$$

and the functions $\xi \mapsto \ell_k(\xi, u_k, y)$ are \mathcal{F}_k -measurable relative to $U_k^{-1}(u_k)$.

Since SP possesses all the properties of P, the problem is solvable and the Kuhn–Tucker conditions (a), (b), (c) are necessary and sufficient for optimality. We shall show that (c) can be replaced by:

(sc) for $k = 1, \dots, N$ one has

$$(3.5) \quad \ell_k(\xi, \bar{u}_k(\xi), E^k \bar{y}(\xi)) = \min_{u_k \in U_k(\xi)} \ell_k(\xi, u_k, E^k \bar{y}(\xi)) \quad \text{a.s.}$$

where

$$(3.6) \quad (E^k \bar{y})(\xi) = E^k \{\bar{y}(\xi)\} \quad (\text{conditional expectation given } \mathcal{F}_k).$$

Of course $E^k \bar{y}$ is \mathcal{F}_k -measurable by definition, so *the process* $\{E^k \bar{y}, k = 1, \dots, N\}$ is *nonanticipative*. Note that everything in the expression (3.5) is \mathcal{F}_k -measurable, and therefore the “almost surely” can be interpreted with respect to the restriction of the probability σ to \mathcal{F}_k . Thus the minimization is entirely in terms of information pertinent to stage k and independent of the future. In particular, for the nest of sigma-fields \mathcal{F}_k corresponding to the sequential notation $\xi = (\xi_1, \dots, \xi_N)$ at the beginning of § 1, ξ can be replaced essentially by $\xi^k = (\xi_1, \dots, \xi_k)$ throughout (3.5). The decision taken at stage k is then represented as a solution $\bar{u}_k(\xi^k)$ to an optimization problem depending only on the past information ξ^k and a vector $E^k \bar{y}(\xi^k)$ of expected “prices.”

THEOREM 4. *A function \bar{u} solves the separable optimal recourse problem SP if and only if there is a multiplier function \bar{y} such that (\bar{u}, \bar{y}) satisfies (a) and (b) of the general Kuhn–Tucker conditions and (sc) above.*

Proof. From the Corollary to Theorem 1, we know that \bar{u} is optimal if and only if $\bar{u} \in \mathcal{U} \cap \mathcal{N}_\infty$ and there exists $\bar{y} \in \mathcal{Y}$ such that

$$(3.7) \quad \sup_{y \in \mathcal{Y}} I_\ell(\bar{u}, y) = I_\ell(\bar{u}, \bar{y}) = \inf_{u \in \mathcal{U} \cap \mathcal{N}_\infty} I_\ell(u, \bar{y}).$$

The left half of (3.7) is equivalent to

$$\sup_{y \in R^n} \ell(\xi, \bar{u}(\xi), y) = \ell(\xi, \bar{u}(\xi), \bar{y}(\xi)),$$

which means that (1.13) and (1.15) hold (and hence all of (a) and (b)). It remains only to show that the right half of (3.7) is equivalent to (sc). But separability implies

$$(3.8) \quad I_\ell(u, \bar{y}) = \sum_{k=1}^N I_{\ell_k}(u_k, \bar{y}) \quad \text{for all } u \in \mathcal{U} \cap \mathcal{N}_\infty,$$

where

$$(3.9) \quad I_{\ell_k}(u_k, \bar{y}) = E\{\ell_k(\xi, u_k(\xi), \bar{y}(\xi))\} = E\{\ell_k(\xi, u_k(\xi), E^k \bar{y}(\xi))\},$$

the last equality being true because the function u_k is \mathcal{F}_k -measurable and ℓ_k is affine in the multiplier y . For $k = 1, \dots, N$, define the integrand r_k on $\Xi \times R^{n_k}$ by

$$(3.10) \quad r_k(\xi, u_k) = \begin{cases} \ell_k(\xi, u_k, E^k \bar{y}(\xi)) & \text{if } u_k \in U_k(\xi), \\ +\infty & \text{if } u_k \notin U_k(\xi). \end{cases}$$

Then for functions $u_k \in L_{n_k}^\infty(\Xi, \mathcal{F}_k, \sigma)$ we have from (3.9)

$$E\{r_k(\xi, u_k(\xi))\} = \begin{cases} I_{\ell_k}(u_k, \bar{y}) & \text{if } u_k(\xi) \in U_k(\xi) \quad \text{a.s.}, \\ +\infty & \text{otherwise.} \end{cases}$$

The right half of (3.7) is therefore identical to the assertion that for $k = 1, \dots, N$:

$$(3.11) \quad \begin{aligned} &\text{the minimum of } I_{r_k}(u_k) = E\{r_k(\xi, u_k(\xi))\} \text{ over all} \\ &u_k \in L_{n_k}^\infty(\Xi, \mathcal{F}_k, \sigma) \text{ is attained at } \bar{u}_k, \end{aligned}$$

while condition (3.5) is the same as

$$(3.12) \quad \begin{aligned} & \text{the minimum of } r_k(\xi, u_k) \text{ over all} \\ & u_k \in R^{n_k} \text{ is attained at } \bar{u}_k(\xi) \text{ almost surely.} \end{aligned}$$

The equivalence of (3.11) and (3.12) follows from our regularity assumptions exactly as did the equivalence of (2.16) and (2.17) in the proof of Theorem 2: each r_k is an \mathcal{F}_k -normal convex integrand. This completes the proof of Theorem 4.

The Kuhn–Tucker conditions in this “decomposed” form have a number of significant features that render them attractive from a computational viewpoint. Notably, if at stage k the multiplier function \bar{y}^k is known and the minimum in (3.5) is uniquely attained almost surely, then the minimizing points must be the values $\bar{u}_k(\xi)$ of the unique optimal decision function \bar{u}_k associated with this stage. In other words, the requirement of \mathcal{F}_k -measurability is automatically taken care of, and there is no need to worry about the ultimate satisfaction of the constraints $f_i(\xi, \bar{u}(\xi)) \leq 0$.

We remark also that in the differentiable case, with $U_k(\xi) = R^{n_k}$ for all k , condition (3.5) takes on the form

$$(3.13) \quad \nabla f_{0k}(\xi, \bar{u}_k(\xi)) + \sum_{i=1}^m E^k \bar{y}_i(\xi) \nabla f_{ik}(\xi, \bar{u}_k(\xi)) = 0 \quad \text{a.s. } (\mathcal{F}_k).$$

The structure of separability also leads to a special dual problem associated with SP. For $k = 1, \dots, N$, define the function g_k on $\Xi \times R^m$ by

$$(3.14) \quad g_k(\xi, y) = \inf_{u_k \in U_k(\xi)} \ell_k(\xi, u_k, y) \quad \text{if } y \in R^m.$$

Then $-g_k$ is an \mathcal{F}_k -normal convex integrand, and the functional

$$(3.15) \quad I_{g_k}(y) = E\{g_k(\xi, y(\xi))\} \quad \text{for } y \in L_m^1$$

is well-defined, concave (with $-\infty$ as a possible value) and satisfies

$$(3.16) \quad I_{g_k}(y) = \inf_{u_k \in \mathcal{U}_k} I_{\ell_k}(u_k, y) \quad \text{for all } \mathcal{F}_k\text{-measurable } y \in \mathcal{Y},$$

where

$$(3.17) \quad \mathcal{U}_k = \{u_k \in L_{n_k}^\infty(\Xi, \mathcal{F}_k, \sigma) \mid u_k(\xi) \in U_k(\xi) \text{ a.s. } (\mathcal{F}_k)\}.$$

These facts are established almost exactly as they were for g and I_g in the proof of Theorem 3.

As the special dual problem for SP, we introduce:

$$\text{SD} \quad \text{Maximize } \sum_{k=1}^N I_{g_k}(E^k y) \text{ over all } y \in \mathcal{Y}.$$

The following result is then immediate from the decomposition

$$(3.18) \quad I_\ell(u, y) = \sum_{k=1}^N I_{\ell_k}(u_k, E^k y) \quad \text{for } (u, y) \in \mathcal{U} \times \mathcal{Y}$$

and the fact that the Kuhn–Tucker conditions (a), (b), (sc) in Theorem 4 characterize the saddle points of this expression.

THEOREM 5. *The dual problem SD has optimal solutions, and they are precisely the components \bar{y} of the pairs (\bar{u}, \bar{y}) satisfying the Kuhn–Tucker conditions (a), (b), (sc). In particular,*

$$\min \text{SP} = \max \text{SD}.$$

B. Linear recourse models. By LP we denote a version of SP that can be formulated as follows:

$$\begin{aligned} \text{LP} \quad & \text{Minimize } E \left\{ \sum_{k=1}^N c_k \cdot u_k(\xi^k) \right\} \\ & \text{subject to } \sum_{k=1}^j A_{jk} u_k(\xi^k) \geq b_j \quad \text{a.s. for } j = 1, \dots, N, \end{aligned}$$

where $c_k \in R^{n_k}$, $b_j \in R^{m_j}$, $A_{jk} \in R^{m_j \times n_k}$ and $\xi^k = (\xi_1, \dots, \xi_k)$ with $\xi_k = (c_k, A_{k1}, \dots, A_{kN}, b_k)$. Thus the vectors c_k and b_k and matrices A_{jk} are random variables whose values become known in stage k , and we are in the sequential notational setting at the beginning of § 1 with $\xi = (\xi_1, \dots, \xi_N)$. It is required that

$$(3.19) \quad u_k \in L_{n_k}^\infty(\Xi^k, \mathcal{F}^k, \sigma^k),$$

where $(\Xi^k, \mathcal{F}^k, \sigma^k)$ is the marginal probability space of the random variable ξ^k , i.e. of the random elements observed in the first k stages.

This formulation differs slightly from the previous pattern in having (3.19) in place of the \mathcal{F}_k -measurability of u_k as a function of Ξ (with \mathcal{F}_k the “cylindrical extension” of \mathcal{F}^k relative to Ξ , as introduced in § 1 for the setting where $\xi = (\xi_1, \dots, \xi_N)$). In simpler terms, the recourse function is taken to be nonanticipative, rather than just essentially nonanticipative. However, the two formulations are equivalent as long as we are not concerned with the multipliers $\bar{p}(\xi)$, and this is justified in the present context by Theorem 3. (In introducing $\bar{p} \in L_n^1$, we need to regard the recourse function u as an element of L_n^∞ and therefore must admit, as negligible, alterations of $u_k(\xi^k)$ on a set of ξ -values of probability zero, even if these involve ξ_{n+1}, \dots, ξ_N .) Incidentally, in contrast to this equivalence, one cannot change the “almost surely” in the constraints of LP without risking a disastrous effect on the problem. This is shown by counterexamples in [8], where a condition on the probability measure σ is also developed which ensures against the discrepancy.

As with SP, we assume that LP satisfies all the regularity conditions we have imposed on P. Actually, the convexity, lower semicontinuity and measurability conditions are trivially satisfied; note that $U_k(\xi) = R^{n_k}$, while each f_i is an affine function of u_k with random variables as coefficients. The uniform boundedness assumption requires that for all realizations of ξ the polyhedron generated by the constraints of LP lies within a fixed ball. For the case where the matrices A_{jk} are nonrandom—or equivalently, have a degenerate distribution—a sufficient condition for uniform boundedness is given by Olsen [9, Lemma 2.4]; cf. also [10]. Various sufficient conditions for strict feasibility can easily be found. For example, one such criterion can be derived from the results of Isofescu and Theodorescu [11] for systems of stochastic linear inequalities.

Problem LP has a block-triangular structure which makes it easy to see more specifically when the property of essentially complete recourse is present. Consider the following decision procedure. In the first stage (having observed $\xi_1 = (c_1, A_{11}, b_1)$) we choose u_1 satisfying $A_{11}u_1 \geq b_1$. In the second stage (having observed ξ_2) we choose u_2 satisfying $A_{22}u_2 \geq b_2^\dagger$, where $b_2^\dagger = b_2 - A_{21}u_1$. And so forth: in the k th stage (having observed ξ_k) we choose u_k satisfying

$$(3.20) \quad A_{kk}u_k \geq b_k^\dagger, \quad \text{where } b_k^\dagger = b_k - \sum_{j=1}^{k-1} A_{kj}u_j.$$

One says that *relatively complete recourse* is present if this procedure can almost surely be continued to the end (i.e. to the choice of u_N), or in other words, if with probability one we will not encounter a stage where we are stymied by the emptiness of the u_k -polyhedron defined by the constraint system (3.20).

THEOREM 6. *Relatively complete recourse implies essentially complete recourse.*

Proof. Let us denote by $\Delta^k(\xi^k)$ the set of all (u_1, \dots, u_k) which can be generated by the first k stages of this procedure. Relatively complete recourse means that each element of $\Delta^k(\xi^k)$ is contained almost surely (with respect to the conditional distribution of $(\xi_{k+1}, \dots, \xi_N)$ given (ξ_1, \dots, ξ_k)) in the set $D^k(\xi)$ in (1.8), which consists of all (u_1, \dots, u_k) such that the procedure can be continued to the end when the total outcome of the random variable is $\xi = (\xi_1, \dots, \xi_k, \xi_{k+1}, \dots, \xi_N)$. Representing $\Delta^k(\xi^k)$ as the closure of a countable set, to each element of which this fact can be applied, we see from the closedness of $D^k(\xi)$ that

$$\Delta^k(\xi^k) \subset D^k(\xi)$$

almost surely (conditionally, given ξ^k). But trivially, the opposite inclusion is universally valid by the definition of $D^k(\xi)$. Therefore, relatively complete recourse is equivalent to the property that

$$(3.21) \quad D^k(\xi) = \Delta^k(\xi^k) \quad \text{a.s.}$$

(in the sense of the overall distribution of ξ). Of course, (3.21) implies that $D^k(\xi)$ essentially depends only on ξ^k , which is the property of essentially complete recourse.

Remark. The concept of relatively complete recourse, and with it Theorem 6, can easily be extended to SP and even to the general context of P, thereby also covering our use of the term in [4]. The multifunction U is itself assumed nonanticipative (as is true for instance in SP): the projection $U^k(\xi)$ consisting of all components (u_1, \dots, u_k) of elements u of $U(\xi)$ is thus assumed \mathcal{F}_k -measurable. The index set $\{1, \dots, m\}$ is partitioned into subsets J_k such that, for $i \in J_k$, $f_i(\xi, u)$ is \mathcal{F}_k -measurable in ξ and depends only on the components (u_1, \dots, u_k) of u . Let $\Delta^k(\xi)$ consist of all elements (u_1, \dots, u_k) of $U^k(\xi)$ satisfying the constraints $f_i(\xi, u) \leq 0$ for all indices $i \in J_1 \cup \dots \cup J_k$. Then $\Delta^k(\xi)$ is \mathcal{F}_k -measurable in ξ . Relatively complete recourse is the property that each element of $\Delta^k(\xi)$ belongs to $D^k(\xi)$ almost surely (conditional probability given \mathcal{F}_k). This can

also be expressed as above in terms of the almost sure feasibility of a “block-triangular” procedure for generating u_1, \dots, u_N sequentially. The proof of Theorem 6 remains valid in this case.

Our assumption of strict feasibility appears needed for the validity of the Kuhn–Tucker conditions (a), (b), (sc) of Theorem 4 in the case of LP, despite the linearities. This may be attributed to the (infinite-dimensional) constraint of nonanticipativity, even though the corresponding multipliers are suppressed in (sc).

The optimal recourse functions for LP, which exist according to Theorem 1 under the regularity conditions which have been imposed, are characterized as follows.

THEOREM 7. *In order that the function $\bar{u} = (\bar{u}_1, \dots, \bar{u}_N)$ with $\bar{u}_k \in L_{n_k}^\infty(\Xi^k, \mathcal{F}^k, \sigma^k)$ be an optimal solution to LP, it is necessary and sufficient that the following conditions be satisfied for some function $\bar{y} = (\bar{y}_1, \dots, \bar{y}_N)$ with $\bar{y}_k \in L_{m_k}^1(\Xi^k, \mathcal{F}^k, \sigma^k)$, $k = 1, \dots, N$:*

$$(3.22) \quad A_{kk}\bar{u}_k(\xi^k) \geq b_k^+(\xi^k) \quad \text{a.s.},$$

$$(3.23) \quad \bar{y}_k(\xi^k) \geq 0 \quad \text{a.s.},$$

$$(3.24) \quad \bar{y}_k(\xi^k) \cdot [b_k^+(\xi^k) - A_{kk}\bar{u}_k(\xi^k)] = 0 \quad \text{a.s.},$$

$$(3.25) \quad \bar{y}_k(\xi^k)A_{kk} = c_k^+(\xi^k) \quad \text{a.s.},$$

where

$$(3.26) \quad b_k^+(\xi^k) = b_k - \sum_{j=1}^{k-1} A_{kj}\bar{u}_j(\xi^j),$$

$$(3.27) \quad c_k^+(\xi^k) = c_k - \sum_{j=k+1}^N (E^k \bar{y}_j)(\xi^k)A_{jk}.$$

Proof. When conditions (a), (b) and (sc) of Theorem 4 are specialized to the present context, we get something slightly different. Namely, each \bar{y}_k would appear in (3.23) and (3.24) as a function of all of ξ , while the expression in (3.25) would instead be $(E^k \bar{y}_k)(\xi^k)$. However, these conditions on \bar{u} really involve only the latter expressions (and their expectations in earlier stages). Therefore, we can just as well apply E^k to (3.23) and (3.24), so that only $E^k \bar{y}_k$ is relevant throughout; it is a mere change of notation to then call this function \bar{y}_k , instead of the original function.

The dual problem in this context may be stated as:

$$\begin{aligned} & \text{Maximize } E \left\{ \sum_{k=1}^N b_k \cdot y_k(\xi^k) \right\} \text{ over all summable} \\ \text{LD} \quad & y_k(\xi^k) \geq 0, \quad k = 1, \dots, N, \text{ satisfying} \\ (3.28) \quad & \sum_{k=j}^N (E^j y_k)(\xi^k)A_{kj} = c_j \quad \text{a.s. for } j = 1, \dots, N. \end{aligned}$$

Note that the function $y = (y_1, \dots, y_N)$ may be called a *nonanticipative* element of L_m^1 . However, LD does not fit the same mold as LP, since in determining the

component y_k for stage k we need consider the conditional expectations of the future components y_j , $k < j \leq N$. Looked at another way, LD involves certain special *chance constraints*, in contrast to LP, because if the expected values of the multipliers y_j associated with future stages are treated as variables to be determined at stage k , then the decision which is taken poses a subsequent constraint on expectations that y_j must live up to.

THEOREM 8. *The dual problem LD has optimal solutions, and they are precisely the components $\bar{y} = (\bar{y}_1, \dots, \bar{y}_N)$ of the pairs (\bar{u}, \bar{y}) satisfying the Kuhn–Tucker conditions in Theorem 7. In particular,*

$$\min \text{LP} = \max \text{LD}.$$

Proof. This follows as a specialization of Theorem 5 via a slight change in notation as in the proof of Theorem 7.

Problem LD resembles the dual obtained by Eisner and Olsen [12] for linear recourse models formulated in L^p -spaces, $1 < p < \infty$. The approach developed here, however, yields a min = max duality theorem with corresponding Kuhn–Tucker conditions, whereas [12] only allows for min = sup duality results.

4. A discrete time stochastic control problem. The purpose of this section is to illustrate, by an example, the relations between the recourse model and certain types of stochastic control problems in discrete time. The optimality conditions developed here can then be used to characterize optimal solutions to these stochastic control problems. The goal is not to give a description of the most general stochastic control problem that can be handled in the framework of the recourse model; it is easy to see how the problem described below can be generalized in many directions and still fit our pattern.

While there are a number of substantial contributions to the theory of necessary and sufficient conditions for stochastic control problems in discrete time, e.g. [13] and [14], there does not seem to be a treatment that allows for the inclusion of state-space constraints when seeking *pointwise optimality conditions*. Several papers do deal with state-space constraints in the continuous case; see [15], [16], [17] and [18]. The difference between the present approach and the one taken by Kushner [15], Haussmann [16] and Ichikawa [17] is that they seek an “expected maximum principle,” in which case the multipliers associated with the state-space constraints (at a finite number of time periods) turn out to be elements of R . It is when seeking pointwise optimality conditions that the difficulties do arise, as illustrated in [18] where Bismut must rely on an $(L^\infty)^*$ -multiplier rather than L^1 -multiplier. Even for continuous-time deterministic problems with state-space constraints these exotic multipliers cannot always be avoided [19].

Let $(\xi_k, k = 1, \dots, N)$ denote a vector-valued (discrete time) stochastic process; for $k = 1, \dots, N$, the realizations of ξ_k are elements of $R^{v'}$ denoted by ξ_k . The state of the system at time k is denoted by x_k , also an element of $R^{v'}$. The dynamics are given by the relations

$$(4.1) \quad x_1 = \xi_1$$

and for $k = 1, \dots, N-1$,

$$(4.2) \quad x_{k+1} = Ax_k + Bu_k + \xi_{k+1},$$

where A is a $(\nu' \times \nu')$ -matrix, B is a $(\nu' \times n')$ -matrix, and $u_k \in R^{n'}$ is the recourse (or control) selected at time k . To be consistent with our earlier notation, we set $\nu = N\nu'$ and $n = Nn'$. The recourse is selected on the basis of *complete information* and *total recall*, by which we mean that the recourse decision u_k is selected in complete knowledge of the past history of the system, i.e. up to and including x_k , the state of the system at time k . (Note that a number of problems with incomplete observation and partial recall can actually be cast as problems with complete information and total recall, see for example [20], [21].) In this set-up, it is equivalent to assert that the decision maker observes ξ_k and recalls past observations and past decisions, since clearly from (4.2) it follows that knowledge of earlier states and decisions, and observation of x_{k+1} uniquely determines ξ_{k+1} , the "noise" of the system.

Moreover, the particular form of the dynamics of this system (4.2) allows us to short-circuit the state component in the description of the model. Indeed, combining (4.1) and (4.2) we obtain

$$(4.3) \quad x_{k+1} = \sum_{q=0}^k A^q \xi_{k+1-q} + \sum_{q=0}^{k-1} B u_{k-q},$$

i.e. the state of the system is a linear function of the past recourse decisions and realizations.

For performance criterion we take a real-valued functional φ_0 defined on $R^\nu \times R^n \times R^\nu$ such that for all $\xi \in \Xi$ the function $(u, x) \mapsto \varphi_0(\xi, u, x)$ is convex, and for all $(u, x) \in R^\nu \times R^\nu$ the function $\xi \mapsto \varphi_0(\xi, u, x)$ is (Borel) measurable. The problem in rough form is to minimize $E\{\varphi_0(\xi, u(\xi), x(\xi))\}$ subject to (4.3) and the further constraints that

$$(4.4) \quad u(\xi) \in U(\xi) \quad \text{a.s.}$$

and for $i = 1, \dots, m$

$$(4.5) \quad \varphi_i(\xi, u(\xi), x(\xi)) \leq 0 \quad \text{a.s.}$$

The multifunction U is assumed to have closed, convex values with nonempty interior. For $i = 1, \dots, m$, the functions φ_i on $R^\nu \times R^n \times R^\nu$, are required to satisfy the same assumptions as φ_0 .

Relation (4.3) allows us to formulate this stochastic control problem as a problem of the type P. Let us write (4.3) (regarded as including (4.1)) in the form

$$x = S\xi + Tu$$

and define

$$f_i(\xi, u) = \varphi_i(\xi, u, S\xi + Tu) \quad \text{for all } u \in U(\xi) \quad i = 0, 1, \dots, m.$$

Suppose in fact that $\varphi_i(\xi, u, S\xi + Tu)$ is (for each fixed u) summable in ξ when $i = 0$ and essentially bounded in ξ when $i = 1, \dots, m$. It can be verified that f_0, \dots, f_m and U satisfy all the conditions imposed on them in § 1, and the corresponding optimal recourse problem P represents the present situation. If in addition P is strictly feasible and the abstract constraint multifunction D corresponding to f_1, \dots, f_m and U is uniformly bounded and nonanticipative, all our general assumptions are satisfied and the above results can be applied. In this way we

obtain necessary and sufficient conditions for optimality from the basic Kuhn–Tucker conditions (a), (b), (c).

Many of the regularity conditions in question are “standard” for control problems. For example, it is common to assume uniform boundedness of the set $\{U(\xi), \xi \in \Xi\}$, and this ensures the uniform boundedness of the multifunction D . As far as nonanticipative feasibility is concerned, we have already explained its relation to the notion of relatively complete recourse that has played an important role in the literature devoted to stochastic programming [4]. This concept has also recently surfaced in stochastic control theory [22], [23]. For a system without state constraints, Striebel [22] introduced the concept of *optimality from time t onward*, requiring essentially that for each control satisfying (4.4) and each time t —whatever be the resulting state—there is a control which is optimal from time t onward. Striebel and Rishel [23] use this condition in their study of optimality criteria for continuous time stochastic control problems. Their motivation for introducing “optimality from time t onward” is quite different from ours but seems to be required by technical considerations that are akin to those that lead us to essentially complete recourse. In particular, Rishel shows that this condition allows him to obtain an explicit form for the generator applied to the value function.

Finally, we note that certain classes of stochastic problems yield separable recourse problems. This is certainly the case if

- (i) $U(\xi) = \bigtimes_{k=1}^N U_k(\xi)$ where U_k is \mathcal{F}_k -measurable,
- (ii) for $i = 0, 1, \dots, m$, $\varphi_i(\xi, u, x) = \sum_{k=1}^N \varphi_{ik}(\xi, u_k) + x \cdot r_i$,

where the functions $\xi \mapsto \varphi_{ik}(\xi, u_k)$ are \mathcal{F}_k -measurable and $r_i \in \mathbb{R}^v$. In this case we can rely on the sharper results of § 3A, if not in fact 3B, in deriving optimality conditions.

REFERENCES

- [1] R. T. ROCKAFELLAR, *Integrals which are convex functionals*, Pacific J. Math., 24 (1968), pp. 525–539.
- [2] ———, *Measurable dependence of convex sets and functions on parameters*, J. Math. Anal. Appl., 28 (1969), pp. 4–25.
- [3] ———, *Conjugate Duality and Optimization*, Monograph Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1974.
- [4] R. T. ROCKAFELLAR AND R. WETS, *Stochastic convex programming: relatively complete recourse and induced feasibility*, this Journal, 14 (1976), pp. 574–589.
- [5] ———, *Nonanticipativity and L^1 -martingales in stochastic optimization problems*, Mathematical programming Studies, 6 (1976); also in Stochastic Systems: Modeling, Identification and Optimization, R. Wets, ed., North-Holland, 1976, pp. 170–187.
- [6] ———, *Stochastic convex programming: Kuhn–Tucker conditions*, J. Math. Economics, 2 (1975), pp. 349–370.
- [7] R. T. ROCKAFELLAR, *Integrals which are convex functionals, II*, Pacific J. Math., 39 (1971), pp. 439–469.
- [8] R. T. ROCKAFELLAR AND R. WETS, *Continuous versus measurable recourse in N -stage stochastic programming*, J. Math. Anal. Appl., 48 (1974), pp. 836–859.

- [9] P. OLSEN, *Discretization of multistage stochastic programming problems*, Mathematical Programming Studies, 6 (1976); also in Stochastic Systems: Modeling, Identification and Optimization, R. Wets, ed., North-Holland, Amsterdam, 1976, pp. 111–124.
- [10] R. GRINOLD, *Symmetric duality for continuous linear programs*, SIAM J. Appl. Math., 18 (1970), pp. 84–96.
- [11] M. ISOFESCU AND R. THEODORESCU, *Linear programming under uncertainty*, Colloquium on Applications of Mathematics to Economics, A. Prekopa, ed., Publishing House of the Hungarian Academy of Sciences, Budapest, 1965, pp. 133–140.
- [12] M. EISNER AND P. OLSEN, *Duality for stochastic programming interpreted as L.P. in L_p -space*, SIAM J. Appl. Math., 28 (1974), pp. 779–792.
- [13] C. STRIEBEL, *Optimal control of discrete time stochastic systems*, Tech. Report, Univ. Minnesota, Minneapolis, 1974.
- [14] K. HINDERER, *Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter*, Lecture Notes in Operations Research and Mathematical Systems, 33, Springer-Verlag, Berlin, 1970.
- [15] H. KUSHNER, *Necessary conditions for continuous parameter stochastic optimization problems*, this Journal, 10 (1972), pp. 550–565.
- [16] U. HAUSSMANN, *General necessary conditions for optimal control of stochastic systems*, Mathematical Programming Studies, 6, 1976; also in Stochastic Systems: Modeling, Identification and Optimization, R. Wets, ed., North-Holland, Amsterdam, 1976, pp. 30–48.
- [17] A. ICHIKAWA, *Notes on a maximum principle of Hausmann*, Tech. Report, Univ. British Columbia, Vancouver, 1975.
- [18] J.-M. BISMUT, *An example of optimal stochastic control with constraints*, this Journal, 12 (1974), pp. 401–418.
- [19] R. T. ROCKAFELLAR, *State constraints in convex problems of Bolza*, this Journal, 10 (1972), pp. 691–715.
- [20] Y. C. HO AND K. C. CHU, *Information structure in dynamic multi-person control problems*, Automatica, 10 (1974), pp. 341–351.
- [21] J.-M. BISMUT, *An introductory approach to duality in optimal stochastic control*, SIAM Rev., to appear.
- [22] C. STRIEBEL, *Martingale conditions for optimal stochastic control of continuous time stochastic systems*, Tech. Report, Univ. of Minnesota, Minneapolis, 1974.
- [23] R. RISHEL, *Controls optimal from time t onward and dynamic programming for systems of controlled jump processes*, Mathematical Programming Studies, 6, 1976; also in Stochastic Systems: Modeling, Identification and Optimization, R. Wets, ed., North-Holland, Amsterdam, 1976, pp. 125–144.

PERIODIC SYSTEMS: CONTROLLABILITY AND THE MATRIX RICCATI EQUATION*

S. BITTANTI,[†] G. GUARDABASSI,[†] C. MAFFEZZONI[‡] AND L. SILVERMAN[¶]

Abstract. Referring to recently published results, a few problems apparently playing a basic role in periodic control theory are discussed in this paper. Specifically, the problems dealt with are the controllability of linear periodic systems and the existence of periodic solutions for periodic matrix Riccati equations.

Introduction. The existence of periodic solutions for a periodic matrix Riccati equation has been recently considered in [1] by G. A. Hewer whose main purpose was to prove the following statement.

For any $\omega > 0$, let $A(\cdot)$, $B(\cdot)$ and $H(\cdot)$ be given ω -periodic real matrices; then, the matrix Riccati equation (MRE)

$$\dot{R}(t) + A'(t)R(t) + R(t)A(t) + H'(t)H(t) - R(t)B(t)B'(t)R(t) = 0, \\ -\infty < t < +\infty,$$

has one and only one positive semidefinite ω -periodic solution such that

$$\dot{x}(t) = (A(t) - B(t)B'(t)R(t))x(t)$$

is asymptotically stable if and only if $(A(\cdot), B(\cdot))$ is stabilizable and $(H(\cdot), A(\cdot))$ is detectable.

Unfortunately, the proof given in [1] is based crucially on a number of preliminary results, some of which turn out to be wrong. Two of them play in fact a fundamental role and are discussed in the following.

1. Controllability. Consider the linear ω -periodic system

$$(1a) \quad \dot{x}(t) = A(t)x(t) + B(t)u(t),$$

$$(1b) \quad y(t) = H(t)x(t)$$

where $x(t) \in R^n$, $u(t) \in R^m$, $y(t) \in R^p$ and $A(\cdot)$, $B(\cdot)$ and $H(\cdot)$ are ω -periodic matrices integrable over $[0, \omega]$.

Brunovsky [2, Proposition 3] proved that (1a) is controllable if and only if it is controllable on $[0, n\omega]$. Hewer claims [1, Thm. 2.15] that the following stronger result is true: system (1a) is controllable if and only if it is controllable on $[0, \omega]$ (this proposition was also stated without proof by Kalman [3, Proposition 2.26]).

The proof given by Hewer is erroneous because the functional linear dependence of the rows of any matrix $L(\cdot)$ on some interval $[0, T]$ is not equivalent to

* Received by the editors July 21, 1976, and in revised form March 7, 1977. This work was supported by Centro di Teoria dei Sistemi, Consiglio Nazionale delle Ricerche (CNR), and by the National Science Foundation under Grant Eng. 76-14379.

[†] Istituto di Elettrotecnica ed Elettronica, Politecnico di Milano, Milano, Italy.

[‡] ENEL-Centro Ricerca di Automatica (CRA), Milano, Italy.

[¶] Department of Electrical Engineering-Systems, University of Southern California, Los Angeles, California 90007.

the existence of a (nonzero) vector z such that

$$L(t)z = 0 \quad \text{for a.e. } t \in [0, T],$$

as assumed in his proof of the preliminary Theorem 2.13. Moreover, a simple counterexample shows that the statement itself is not correct. Assume that, in system (1a), $n \geq 2$ while the matrices $A(\cdot)$ and $B(\cdot)$ are given by

$$(2a) \quad A(t) \triangleq A = \text{diag}(\lambda), \quad \lambda \in \mathbb{R}^n, \quad \lambda_i \neq \lambda_j, \quad \forall i \neq j, \quad \forall t \in \mathbb{R},$$

$$(2b) \quad B(t) = B(t+k), \quad \forall(t, k) \in \mathbb{R} \times \mathbb{Z},$$

$$(2c) \quad B'(\sigma) \triangleq |e^{-\lambda_1(1-\sigma)} e^{-\lambda_2(1-\sigma)} \dots e^{-\lambda_n(1-\sigma)}| \sin \pi \sigma, \quad \forall \sigma \in [0, 1].$$

The 1-periodic system (1), (2) is controllable. In fact, its controllability matrix on $[0, n]$ is given by

$$C(0, n) = \frac{1}{2} V_n V_n'$$

where

$$V_n \triangleq |v^1 v^2 \dots v^n|,$$

$$v^k \triangleq |e^{-k\lambda_1} e^{-k\lambda_2} \dots e^{-k\lambda_n}|'.$$

Since V_n is a Vandermonde matrix, the conclusion that $C(0, n)$ is nonsingular directly follows.

Nevertheless, system (1), (2) is generally not controllable on one period. In fact, for any $\tau \in [0, 1)$,

$$C(\tau, \tau+1) = V_2 M V_2'$$

where

$$M \triangleq \begin{vmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{vmatrix},$$

$$\mu_1 \triangleq \int_{\tau}^1 \sin^2 \pi t \, dt, \quad \mu_2 \triangleq \int_1^{\tau+1} \sin^2 \pi t \, dt.$$

Hence

$$\text{rank}(C(\tau, \tau+1)) = \begin{cases} 1, & \tau = 0, \\ 2, & \tau \in (0, 1), \end{cases}$$

so that system (1), (2) is controllable on one period if and only if $n = 2$ and $\tau \in (0, 1)$.

Finally, the above example enables one to point out that, in general, Brunovsky's result [2, Proposition 3] cannot be significantly strengthened.

As a matter of fact, it can be proved that system (1), (2) is controllable on (τ, t) if and only if $j - i = n$, where i and j are respectively the largest and the smallest integers such that $i \leq \tau, j \geq t$. Of course, when $A(t), B(t)$ are also analytic it follows from Silverman and Meadows [4] that the proposition is true since, in this special case, controllability is interval independent.

2. The Lyapunov differential equation. By the Lyapunov differential equation (LDE) reference is made to the equation

$$\dot{K}(t) + A'(t)K(t) + K(t)A(t) + H'(t)H(t) = 0$$

which coincides, up to the linear terms, with the matrix Riccati equation dealt with in the introduction. Apparently, the ω -periodic solutions of the LDE are of primary interest in the analysis of the ω -periodic solutions of the MRE.

Let $\Phi(t)$ denote the $(n \times n)$ matrix solution of

$$\dot{\Phi}(t) = A(t)\Phi(t), \quad \Phi(0) = I.$$

As is well known, $\Phi(t)$ can be expressed in the form

$$\Phi(t) = F(t) e^{Gt}, \quad \forall t,$$

where $F(\cdot)$ is a nonsingular matrix of ω -periodic functions while G is a constant matrix.

Another result claimed in [1, Thm. 3.7] is the following: If, for each eigenvalue γ of G with $\operatorname{Re}[\gamma] \geq 0$ and each associated nonzero eigenvector η ,

$$(3) \quad (H(t)F(t)\eta = 0 \text{ for a.e. } t \in [0, \omega]) \Rightarrow \eta = 0,$$

then the existence of a ω -periodic solution of the LDE is equivalent to G being Hurwitz.

However, the above proposition is not correct as the following simple example shows. Precisely, the existence of a ω -periodic solution of the LDE does not imply that G is Hurwitz. With reference to system (1), let $n \triangleq 1$, $A(t) = a \neq 0$ and $H(t) \triangleq \sin t$. Then, $\omega = 2\pi$, $F(t) = 1$ and $G = a$ so that (3) is trivially verified. For any $a \neq 0$, the ω -periodic solution of

$$\dot{K}(t) + 2aK(t) + \sin^2(t) = 0$$

exists and is unique. However, if $a > 0$, G is not Hurwitz.

To the best of the present authors' knowledge, the most powerful condition for the LDE to admit a unique ω -periodic solution is given in [5, Remark 5.1].

3. Concluding remarks. The periodic solutions of the MRE play an important role in periodic optimization theory [6]–[8] as specifically pointed out in [5] and [9]. However, in view of the discussion carried on in § 1, it can be concluded that in [1] detectability and stabilizability have been improperly defined (Definition 3.1 and Definition 4.1) so that the meaning of Hewer's main result (Theorem 1.3) turns out to be ambiguous and, as such, cannot even be taken as an appealing conjecture. So far, the only substantial results on the existence of periodic solutions of the MRE are those given in [10], for the scalar case, and in [5], for the general case.

REFERENCES

- [1] G. A. HEWER, *Periodicity, detectability and the matrix Riccati equation*, this Journal, 13 (1975), pp. 1235–1251.
- [2] P. BRUNOVSKY, *Controllability and linear closed-loop controls in linear periodic systems*, J. Differential Equations, 6 (1969), pp. 296–313.

- [3] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.
- [4] L. SILVERMAN AND H. E. MEADOWS, *Controllability and observability in time-variable linear systems*, this Journal, 5 (1967), pp. 64–72.
- [5] S. BITTANTI, A. LOCATELLI AND C. MAFFEZZONI, *Second-variation methods in periodic optimization*, J. Optimization Theory Appl., 14 (1974), no. 1, pp. 31–49.
- [6] G. GUARDABASSI, A. LOCATELLI AND S. RINALDI, *The status of periodic optimization of dynamical systems*, Ibid., 14 (1974), no. 1, pp. 1–20.
- [7] E. J. NOLDUS, *A survey of optimal periodic control of continuous systems*, Journal A, 16 (1975), pp. 11–16.
- [8] G. GUARDABASSI, *The optimal periodic control problem*, Journal A, 17 (1976), no. 2, pp. 75–83.
- [9] C. MAFFEZZONI, *Hamilton–Jacobi theory for periodic control problems*, J. Optimization Theory Appl., 14 (1974), no. 1, pp. 21–30.
- [10] D. A. SANCHEZ, *A note on periodic solutions of Riccati-type equations*, SIAM J. Appl. Math., 17 (1969), pp. 957–959.

A MAXIMIZATION PROBLEM RELATED TO PARAMETRIC LINEAR COMPLEMENTARITY*

IKUYO KANEKO†

Abstract. The programming problem considered here is that of finding the maximal value of α such that the solution z of $q + \alpha p + Mz \geq 0$, $z \geq 0$ and $z^T(q + \alpha p + Mz) = 0$ satisfies $z \leq a$. In this problem, q , p , a are n -vectors such that $q \geq 0$, $a > 0$, M is an $n \times n$ P -matrix and α is a scalar. This problem has an important application in structural mechanics. In this paper it is first explained that a certain local optimum of the above problem can be obtained more easily than the global optimum, and then necessary and sufficient conditions are determined under which the local and global optima coincide. Relationships are examined between these conditions and those on the isotonicity of the solutions of the parametric linear complementarity problem.

1. Introduction. For an n -vector r and $n \times n$ matrix M , the *linear complementarity problem* is that of finding an n -vector z such that

$$r + Mz \geq 0, \quad z \geq 0, \quad z^T(r + Mz) = 0.$$

We shall denote this problem by symbol (r/M) . A fundamental theorem in linear complementarity due to Samelson, Thrall and Wesler [19] states that the problem (r/M) has a unique solution for every r if and only if M is a P -matrix, i.e., it has positive principal minors. In what follows, we shall denote by $P(n)$ the set of $n \times n$ P -matrices. When the order of a matrix is clear from context, we simply write P . For n -vectors q , p and an $n \times n$ matrix M , the *parametric linear complementarity problem*, $(q/p/M)$, is the family of linear complementarity problems of the form

$$(1) \quad \{(q + \alpha p/M): \alpha \geq 0\},$$

where α is a scalar parameter. If $M \in P$, then (1) has a unique solution for every α , which we denote by $z(\alpha)$. We note that if, in addition, $q \geq 0$, then $z(0) = 0$ and the *parametric principal pivoting algorithm* (PPPA) which is a parametric version of Graves' algorithm [6] (see Cottle [1] for a description) generates $z(\alpha)$ for increasing values of $\alpha \geq 0$. Under the assumptions, the PPPA proves, constructively, that $z(\alpha)$ is continuous and piecewise linear in α .

In this paper, we are concerned with the following programming problem:

Find

$$\alpha^* = \max \alpha$$

such that there exists z satisfying

$$(2.1) \quad q + \alpha p + Mz \geq 0,$$

$$(2.2) \quad z \geq 0,$$

$$(2.3) \quad a \geq z,$$

$$(2.4) \quad z^T(q + \alpha p + Mz) = 0.$$

* Received by the editors September 24, 1975, and in final revised form March 4, 1977.

† Department of Industrial Engineering, University of Wisconsin—Madison, Madison, Wisconsin 53706.

Here, q, p, a are n -vectors with $q \geq 0, a > 0, M \in P(n)$, and α is a scalar. In what follows, this problem will be referred to as the *max problem* ($q/p/a/M$). The conditions (2.1), (2.2), and (2.4) give the parametric linear complementarity problem which is said to be *associated* with the max problem. Under the assumptions it is not difficult to show (see Kaneko [11]) that α^* is nonnegative and that α^* is finite if and only if $p \not\geq 0$.

The max problem is a special case of a class of problems considered by Kirchgässner [14] and Ibaraki [7], [8] who described iteration procedures for solving them. A more natural approach is by using the PPPA. Solution methods for the max problem will be discussed in § 3.

Using the notation defined above, the max problem can be rewritten as:

$$\text{Find} \quad \alpha^* = \max \{ \alpha : z(\alpha) \leq a \}.$$

We define $\hat{\alpha}$ by

$$(3) \quad \hat{\alpha} = \max \{ \alpha : z(\xi) \leq a, \xi \in [0, \alpha] \}.$$

The reader can verify that $\hat{\alpha}$ is a local optimum for the max problem. As will be explained later (§ 3), a local minimum $\hat{\alpha}$ can be obtained more quickly than the global optimum α^* . The purpose of the present paper is to determine necessary and sufficient conditions on data q, p, a and M such that $\hat{\alpha} = \alpha^*$ holds.

The max problem arises in a certain problem in structural mechanics addressed by Maier [15], where the abovementioned assumptions on the data are automatically satisfied. The problem may be explained very briefly as follows. Every structure is designed so that it carries some reserve strength to allow for an unexpected overload and errors. The *safety factor* is a measure for such a reserve strength. Maier demonstrated that the problem of determining the behavior of a certain fairly broad class of structures can be formulated as a linear complementarity problem with a P -matrix, and based on the formulation, showed that the safety factor of such structures can be identified as the optimal solution of the max problem. For the details, we refer to Maier [15], De Donato and Maier [5] and the author's report [13].

Let $z(\alpha)$ be the solution map of the problem ($q/p/M$) where $M \in P$. We say that the parametric linear complementarity problem has *isotone solutions* if $z(\alpha)$ is isotone, i.e., monotone nondecreasing in $\alpha \geq 0$, componentwise. It is easy to see that the isotonicity of the solutions of the associated parametric linear complementarity problem provides a sufficient condition for $\hat{\alpha} = \alpha^*$ to hold. In fact, the isotonicity is necessary for the local optimum to be globally optimal in a certain sense (see Property (22) in § 5). In [1] Cottle considered *strong* and *uniform monotonicity* conditions on q and/or M characterizing the isotonicity of the solutions of the problem ($q/p/M$). Analogously (but noting that the max problem has an extra parameter a not present in ($q/p/M$)) we shall pose the following questions:

(A-1) What are necessary and sufficient conditions on $q \geq 0, a > 0$ and $M \in P$ such that $\hat{\alpha} = \alpha^*$ holds in the max problem ($q/p/a/M$) for every p ?

(A-2) What are necessary and sufficient conditions on $q \geq 0$ and $M \in P$ such that $\hat{\alpha} = \alpha^*$ holds in the max problem ($q/p/a/M$) for every p and every $a > 0$?

(B-1) What are necessary and sufficient conditions on $a > 0$ and $M \in P$ such that $\hat{\alpha} = \alpha^*$ holds in the max problem $(q/p/a/M)$ for every $q \geq 0$ and every p ?

(B-2) What are necessary and sufficient conditions on $M \in P$ such that $\hat{\alpha} = \alpha^*$ holds in the max problem $(q/p/a/M)$ for every $q \geq 0$, every p and every $a > 0$?

Obviously, the first two correspond to the strong monotonicity and the last two to the uniform monotonicity. A result in Cottle's paper [2, p. 7] "almost" answers (B-2). We shall explain it in § 4.

The organization of the rest of the paper is as follows. The remainder of the present section specifies some notation and terminology to be used in later sections. In the next section we summarize existing solution procedures to compute $\hat{\alpha}$ and/or α^* and indicate that $\hat{\alpha}$ may be obtained much more quickly than α^* . In § 3 we introduce the concept of convex directions by which we characterize the situation where $\hat{\alpha} = \alpha^*$ for a given set of data, and using the concept give an answer to (A-1) and (A-2). Questions (B-1) and (B-2) will be answered in § 4. The final section relates the results obtained in the previous sections to the isotonicity of the solutions of the parametric linear complementarity problem. There, we also pose and give an answer to questions similar to (A-1) and (A-2) but with the role of q and p reversed.

Euclidean n space is denoted by R^n , its nonnegative orthant by R_+^n . By $R^{n \times m}$, we denote the space of real $n \times m$ matrices. We use I to denote the identity matrix of appropriate order. A vector is regarded as a column and superscript T is used to denote transposition. The symbol e denotes the summation vector $(1, \dots, 1)^T$ of appropriate size. For n -vectors x and y , $x^T y$ is the usual inner product.

For positive integer n we define $((n))$ to be the set of all (ordered) sequences of integers $\gamma = (\gamma_1, \dots, \gamma_t)$ such that $1 \leq \gamma_1 < \dots < \gamma_t \leq n$ and $1 \leq t \leq n$. For $\gamma = (\gamma_1, \dots, \gamma_t) \in ((n))$, we write $\bar{\gamma} = (\delta_1, \dots, \delta_s)$ if and only if $\{\gamma_1, \dots, \gamma_t, \delta_1, \dots, \delta_s\} = \{1, 2, \dots, n\}$, $n = s + t$ and $\bar{\gamma} \in ((n))$. Note that given $\gamma \in ((n))$, $\bar{\gamma}$ is uniquely determined.

Let $M \in R^{n \times n}$, $i, j \in \{1, \dots, n\}$, $\delta = (\delta_1, \dots, \delta_s) \in ((n))$ and $\gamma = (\gamma_1, \dots, \gamma_t) \in ((n))$. We denote the (i, j) th element of M by M_{ij} , the i th row and j th column of M , respectively, by M_i and M_j . Also we define

$$M_{\delta} = \begin{bmatrix} M_{\delta_1} \\ M_{\delta_2} \\ \dots \\ M_{\delta_s} \end{bmatrix}, \quad M_{\gamma} = [M_{\gamma_1}, M_{\gamma_2}, \dots, M_{\gamma_t}]$$

and

$$M_{\delta\gamma} = (M_{\gamma})_{\delta..}$$

For matrices $A \in R^{m \times s}$ and $B \in R^{m \times t}$ we denote by $\text{pos}\{A, B\}$ the polyhedral cone spanned by columns of A and B . Let $M \in R^{n \times n}$. For $\gamma \in ((n))$,

$$C(\gamma; M) = \text{pos}\{I_{\delta}, -M_{\gamma}\},$$

where $\delta = \bar{\gamma}$ is called a *complementary cone* (Murty [17]). It is easy to show that if $M \in P(n)$, then $C(\gamma; M)$ has a nonempty interior (relative to R^n) for each $\gamma \in ((n))$. Finally, $S \subset R^n$ is said to be *star-shaped* on $T \subset S$ if for every $r \in T$ and every $r' \in S$ we have

$$(1 - \lambda)r + \lambda r' \in S$$

for each $\lambda \in [0, 1]$.

2. Solution procedures. As Maier [15] pointed out, the max problem is a special case of the following programming problem

$$\begin{aligned} & \text{maximize } c_1^T x + c_2^T y + c_3^T u \\ & \text{subject to} \\ & A_1 x + A_2 y + A_3 u = b, \\ & x \geq 0, \quad y \geq 0, \quad u \geq 0, \\ & x^T y = 0, \end{aligned}$$

considered by Kirchgässner [14] who described a cutting plane algorithm for solving the problem. Here, c_1, c_2, c_3 are n_1, n_1, n_2 -vectors, respectively, A_1, A_2, A_3 are $m \times n_1, m \times n_1, m \times n_2$ matrices, respectively, and b is an m -vector. Independently, Ibaraki [7], [8] proposed branch-and-bound and hybrid algorithms using cuts, for solving the same problem.

A more natural approach taking advantage of the special structure of the max problem is by using the PPPA. As Cottle [2] suggested, we could use the PPPA to generate $z(\alpha)$ for all $\alpha \geq 0$ and determine α^* by examining $\{z(\alpha): \alpha \geq 0\}$. The computational effort required by this approach is expected to be considerably less than that by Kirchgässner's Ibaraki's method, which may take solving a number of linear programs of size at least n .

A motivation to examine conditions for $\hat{\alpha} = \alpha^*$ is that in the PPPA approach, the local minimum $\hat{\alpha}$ can be computed more quickly than α^* by furnishing the algorithm with a certain "optimality criterion". Namely, one generates $z(\alpha), \alpha \geq 0$ until

$$(4) \quad z(\alpha) \leq a$$

is violated *for the first time*, when the algorithm is terminated. For the purpose of reference, we shall call this the *modified* PPPA. Obviously, the original PPPA, which requires $z(\alpha)$ for all $\alpha \geq 0$, takes more (no fewer) pivots than the modified version. In particular, if the upper bounds $z \leq a$ are tight, then the first violation of the bounds is expected to occur for a small value of α , which implies that the modified PPPA takes considerably fewer pivots.

In the structural engineering application mentioned in the preceding section, the upper bounds do tend to be tight. The following table (Table 1) demonstrates the number of pivots needed to compute $\hat{\alpha}$ and α^* in the PPPA for two examples of the max problem arising from the structural engineering situation (see Kaneko [10], [13] for the details).

TABLE 1

	The order of M	To compute $\hat{\alpha}$	To compute α^*
Problem 1	10	4	10
Problem 2	34	17	34

Under the assumption that $\hat{\alpha} = \alpha^*$, there is another efficient method to solve the max problem. Maier [15] mentioned that a subroutine used in Kirchgässner's iterative procedure suffices to solve the max problem *if* the associated parametric linear complementarity problem has isotone solutions. The statement is true under a weaker assumption $\hat{\alpha} = \alpha^*$. The subroutine is essentially the simplex method applied to a linear program

$$\begin{aligned}
 &\text{maximize } \alpha \\
 &\text{subject to} \\
 &Iw - Mz - \alpha p = q, \\
 &w \geq 0, \quad z \geq 0, \quad a \geq z,
 \end{aligned}$$

with a restricted pivot choice to keep the complementarity condition $w^T z = 0$. the subroutine is terminated when the objective value can no longer be increased without violating the constraints.

Cottle [2] called this subroutine the *restricted basis simplex method* (RBSM) and gave an example showing that the RBSM need not compute α^* if $\hat{\alpha} \neq \alpha^*$. If $\hat{\alpha} = \alpha^*$, both the modified PPPA and RBSM solve the max problem and their computational efficiencies are about the same. In fact it can be shown that it takes the both algorithms exactly the same number of pivots to solve the problem provided the upper bounding technique (Dantzig [4]) is used in the RBSM, and all bases encountered are nondegenerate. This can be reasoned as follows. In both the PPPA and RBSM, all constraints in the problem are maintained including the complementarity. Since M is a P -matrix, there exists a unique solution to (2.1)–(2.4) for each $\alpha \in [0, \hat{\alpha}]$. In both algorithms, the value of α is monotonically increased from 0 to $\hat{\alpha}$. A pivot occurs (in both algorithms) when the half line $\{q + \alpha p : \alpha \geq 0\}$ passes from one complementarity cone to an adjacent one.

3. Convex directions and questions (A-1) and (A-2). For $a \in R^n$, $a > 0$ and $M \in P(n)$, we define

$$\mathcal{R}(a, M) = \{r \in R^n : \text{the solution } z \text{ of } (r/M) \text{ satisfies } z \leq a\}.$$

By using this symbol, the max problem $(q/p/a/M)$ can be restated as:

$$\text{Find } \alpha^* = \max \{\alpha : q + \alpha p \in \mathcal{R}(a, M)\}$$

and $\hat{\alpha}$ defined by (3) is given by

$$\hat{\alpha} = \max \{\alpha : q + \xi p \in \mathcal{R}(a, M), \xi \in [0, \alpha]\}.$$

We say that $p \in R^n$ is a *convex direction* of $\mathcal{R}(a, M)$ at $q \in R_+^n$ if the following holds:

$$\left[\begin{array}{c} q + \alpha' p \in \mathcal{R}(a, M) \\ \alpha' > 0 \end{array} \right] \text{ implies } \left[\begin{array}{c} q + \alpha p \in \mathcal{R}(a, M) \\ \text{for all } \alpha \in [0, \alpha'] \end{array} \right].$$

The concept of convex directions characterized the situation where $\hat{\alpha} = \alpha^*$ holds. The following result is direct from the definition.

(5) **PROPERTY.** *Let $q \in R_+^n$, $p \in R^n$, $a \in R^n$, $a > 0$ and $M \in P(n)$. Then $\hat{\alpha} = \alpha^*$ if and only if p is a convex direction of $\mathcal{R}(a, M)$ at q .*

Using the terminology and the above property, we can rephrase questions (A-1) and (A-2) as follows.

(A-1) What are necessary and sufficient conditions on $q \geq 0$, $a > 0$ and $M \in P$ such that every p is a convex direction of $\mathcal{R}(a, M)$ at q ?

(A-2) What are necessary and sufficient conditions on $q \geq 0$ and $M \in P$ such that every p is a convex direction of $\mathcal{R}(a, M)$ at q for every $a > 0$?

We give an answer to (A-1) and (A-2) stated this way after we prove the following lemma.

(6) **LEMMA.** *Let $q \in R_+^n$, $a \in R^n$, $a > 0$ and $M \in P(n)$ be given. Then every $p \in R^n$ is a convex direction of $\mathcal{R}(a, M)$ at q if and only if for each $\gamma \in ((n))$, every $b \in R^{|\gamma|}$ is a convex direction of $\mathcal{R}(a_\gamma, M_{\gamma\gamma})$ at q_γ .*

Proof. The if part is trivial. To show the only if part, suppose there exist $\gamma \in ((n))$ and $b \in R^{|\gamma|}$ such that b is not a convex direction of $\mathcal{R}(a_\gamma, M_{\gamma\gamma})$ at q_γ . Then there exist $0 < \alpha_1 < \alpha_2$ such that

$$(7) \quad q_\gamma + \alpha_1 b \notin \mathcal{R}(a_\gamma, M_{\gamma\gamma})$$

and

$$(8) \quad q_\gamma + \alpha_2 b \in \mathcal{R}(a_\gamma, M_{\gamma\gamma}).$$

Let $x(\alpha) \in R^{|\gamma|}$ be the solution of $(q_\gamma + \alpha b / M_{\gamma\gamma})$ for each $\alpha \in [0, \alpha_2]$. Let $\delta = \bar{\gamma}$ and let c be a $|\delta|$ -vector of the form

$$c = \theta e, \quad \theta > 0,$$

where e is the $|\delta|$ -dimensional summation vector. We can choose θ sufficiently large so that for every α in $[0, \alpha_2]$,

$$q_\delta + \alpha c + M_{\delta\gamma} x(\alpha) \geq 0.$$

This is possible since $x(0) = 0$ and $x(\alpha)$ is continuous and piecewise linear. Let $p \in R^n$ be defined by

$$p_\gamma = b \quad \text{and} \quad p_\delta = c.$$

Then we have that $z(\alpha) \in R^n$ given by

$$(9) \quad z_\gamma(\alpha) = x(\alpha)$$

and

$$(10) \quad z_\delta(\alpha) = 0$$

is the solution of $(q/p/M)$ for $\alpha \in [0, \alpha_2]$. From (7)–(10) we see that

$$q + \alpha_1 p \notin \mathcal{R}(a, M) \quad \text{and} \quad q + \alpha_2 p \in \mathcal{R}(a, M),$$

which implies that the p defined above is not a convex direction of $\mathcal{R}(a, M)$ at q . \square

(11) THEOREM. *Let $q \in R^n$, $a \in R^n$, $a > 0$ and $M \in P(n)$. Then every $p \in R^n$ is a convex direction of $\mathcal{R}(a, M)$ at q if and only if for every $\gamma \in ((n))$,*

$$(M_{\gamma\gamma})^{-1}q_\gamma \geq -a_\gamma.$$

Proof. Assume that every $p \in R^n$ is a convex direction of $\mathcal{R}(a, M)$ at q . Let $\gamma \in ((n))$. By the preceding lemma we have that every $b \in R^{|\gamma|}$ is a convex direction of $\mathcal{R}(a_\gamma, M_{\gamma\gamma})$ at q_γ . From this we show that

$$(M_{\gamma\gamma})^{-1}q_\gamma \geq -a_\gamma.$$

Define b by

$$b = -q_\gamma - M_{\gamma\gamma}a_\gamma$$

and let

$$r(\alpha) = q_\gamma + \alpha b, \quad \alpha \geq 0.$$

Since $a_\gamma > 0$, we have that

$$r(1) = -M_{\gamma\gamma}a_\gamma \in \text{int pos } \{-M_{\gamma\gamma}\}$$

and hence for a sufficiently small $\theta > 0$,

$$r(1 - \theta) \in \text{pos } \{-M_{\gamma\gamma}\}.$$

This implies that

$$-(M_{\gamma\gamma})^{-1}r(1 - \theta)$$

is the solution of the problem $(r(1 - \theta)/M_{\gamma\gamma})$. Since b is a convex direction of $\mathcal{R}(a_\gamma, M_{\gamma\gamma})$ at q_γ and since $r(1)$ belongs to $\mathcal{R}(a_\gamma, M_{\gamma\gamma})$, $r(1 - \theta)$ also belongs to $\mathcal{R}(a_\gamma, M_{\gamma\gamma})$. This implies that

$$-(M_{\gamma\gamma})^{-1}r(1 - \theta) \leq a_\gamma.$$

By using the definition of $r(\alpha)$, we then have

$$-(M_{\gamma\gamma})^{-1}(-\theta b - M_{\gamma\gamma}a_\gamma) \leq a_\gamma \quad \text{or} \quad \theta(M_{\gamma\gamma})^{-1}(-q_\gamma - M_{\gamma\gamma}a_\gamma) \leq 0.$$

Since $\theta > 0$, we have

$$(M_{\gamma\gamma})^{-1}q_\gamma \geq -a_\gamma.$$

To show the converse, suppose there exists $p \in R^n$ such that p is not a convex direction of $\mathcal{R}(a, M)$ at q . Let $r(\alpha) = q + \alpha p$, $\alpha \geq 0$. It follows from the supposition that there exist $0 < \alpha_1 < \alpha_2$ for which $r(\alpha_2)$ belongs to $\mathcal{R}(a, M)$ but $r(\alpha_1)$ does not. We denote by $z(\alpha)$ the solution of $(r(\alpha)/M)$ for each $\alpha \geq 0$. Then we have that for $\alpha = \alpha_1$, there exists $j \in \{1, \dots, n\}$ such that

$$z_j(\alpha_1) > a_j.$$

Choose $\alpha_3 \in (\alpha_1, \alpha_2)$ so that

$$(12) \quad z_j(\alpha_3) > a_j$$

and

$$(13) \quad z_j(\alpha) \leq z_j(\alpha_3), \quad \alpha \in [\alpha_3, \alpha_3 + \varepsilon_1]$$

for some positive number ε_1 . Such α_3 and ε_1 exist since $\alpha_1 < \alpha_2$, $z_j(\alpha_1) > a_j$, $z_j(\alpha_2) \leq a_j$ and $z_j(\alpha)$ is continuous. Since $M \in P$, there exists $\gamma \in ((n))$ such that $r(\alpha_3)$ belongs to $C(\gamma; M)$. We shall note that γ necessarily contains j since $z_j(\alpha_3)$ is positive. Since $M \in P$, every complementary cone has a nonempty interior. Thus, we may assume without loss of generality that $r(\alpha)$ is in $C(\gamma; M)$ for $\alpha \in [\alpha_3, \alpha_3 + \varepsilon_2]$ for some positive ε_2 . Let ε^* be the minimum of ε_1 and ε_2 . From pivot theory (see e.g., Parsons [18]) we have for $\alpha \in [\alpha_3, \alpha_3 + \varepsilon^*]$,

$$z_\gamma(\alpha) = \hat{q}_\gamma + \alpha \hat{p}_\gamma,$$

where

$$\hat{q}_\gamma = -(M_{\gamma\gamma})^{-1}q_\gamma \quad \text{and} \quad \hat{p}_\gamma = -(M_{\gamma\gamma})^{-1}p_\gamma.$$

By (13), \hat{p}_γ must be nonpositive. From this, $\alpha_3 > 0$ and (12), we have

$$a_j < z_j(\alpha_3) = \hat{q}_j + \alpha_3 \hat{p}_j \leq \hat{q}_j,$$

or

$$((M_{\gamma\gamma})^{-1}q_\gamma)_j < -a_j. \quad \square$$

(14) COROLLARY. Let $q \in R_+^n$ and $M \in P(n)$. Then every $p \in R^n$ is a convex direction of $\mathcal{R}(a, M)$ at q for every $a > 0$ if and only if for each $\gamma \in ((n))$

$$(M_{\gamma\gamma})^{-1}q_\gamma \geq 0.$$

4. Questions (B-1) and (B-2). Let $a \in R^n$, $a > 0$ and $M \in P(n)$. By Property (5) and by noting the condition that every $p \in R^n$ is a convex direction of $\mathcal{R}(a, M)$ at all $q \in R_+^n$ is equivalent to the condition that $\mathcal{R}(a, M)$ is star-shaped on R_+^n , we can rephrase questions (B-1) and (B-2) as follows:

(B-1) What are necessary and sufficient conditions on $a > 0$ and $M \in P$ such that $\mathcal{R}(a, M)$ is star-shaped on R_+^n ?

(B-2) What are necessary and sufficient conditions on $M \in P$ such that $\mathcal{R}(a, M)$ is star-shaped on R_+^n for every $a > 0$?

In [2], Cottle proved the following result which “almost” answers (B-2).

(15) THEOREM (Cottle). Let $M \in P$. Then $\mathcal{R}(a, M)$ is convex for every $a > 0$ if and only if M has nonpositive off-diagonal elements.

Obviously, a convex set S is star-shaped on every $T \subset S$, but a set S which is star-shaped on one particular $T \subset S$ need not be convex. For the set $\mathcal{R}(a, M)$, however, its convexity and star-shapedness on R_+^n are in fact equivalent. Thus the above theorem due to Cottle essentially gives an answer to (B-2). Moreover, the convexity (or equivalently star-shapedness on R_+^n) of $\mathcal{R}(a, M)$ is independent of $a > 0$. Hence M having nonpositive off-diagonal elements is a necessary and sufficient condition answering (B-1) as well as (B-2). Formally we prove:

- (16) **THEOREM.** *Let $M \in P$. The following five statements are equivalent.*
 (17) *M has nonpositive off-diagonal elements.*
 (18) *For every $a > 0$, $\mathcal{R}(a, M)$ is convex.*
 (19) *There exists $a > 0$ such that $\mathcal{R}(a, M)$ is convex.*
 (20) *For every $a > 0$, $\mathcal{R}(a, M)$ is star-shaped on R_+^n .*
 (21) *There exists $a > 0$ such that $\mathcal{R}(a, M)$ is star-shaped on R_+^n .*

Proof. By definitions and Theorem (15) we have the following implications:

$$\begin{array}{ccc}
 & (19) & \\
 & \nearrow & \searrow \\
 (17) \Leftrightarrow (18) & & (21). \\
 & \searrow & \nearrow \\
 & (20) &
 \end{array}$$

It remains to show that (21) implies (17). Assume M has a positive off-diagonal element. Since we are assuming $M \in P$, every principal submatrix has the inverse. Thus by Proposition 1 in Cottle and Veinott [3, p. 247] there exists $\gamma \in ((n))$ such that

$$(M_{\gamma\gamma})^{-1} \not\leq 0,$$

or for some i and j

$$((M_{\gamma\gamma})^{-1})_{ij} < 0.$$

We consider $|\gamma|$ -vector b of the following form

$$b = \theta e_j, \quad \theta > 0,$$

where e_j is the j th unit vector of dimension $|\gamma|$. Given $a > 0$, we can choose θ sufficiently large so that

$$(M_{\gamma\gamma})^{-1}b \not\geq -a_\gamma.$$

If we let $q_\gamma = b$ and $q_\delta = 0$, then by Theorem (11) there exists p such that p is not a convex direction of $\mathcal{R}(a, M)$ at this $q \in R_+^n$. Thus $\mathcal{R}(a, M)$ is not star-shaped on R_+^n . \square

We refer to Kaneko [9], [12] for amplifications of the above result.

5. Relationships to isotone solutions. In this final section, we relate the results obtained in the previous two sections to isotonicity properties of the solutions of the parametric linear complementarity problem. We also consider a strong monotonicity property of the solutions of the parametric linear complementarity problem with respect to $p < 0$ and M , and the corresponding optimality conditions for the max problem.

We first prove the following.

(22) **PROPERTY.** *Let $q \in R_+^n$, $p \in R^n$ and $M \in P(n)$. Then $(q/p/M)$ has isotone solutions if and only if $\hat{\alpha} = \alpha^*$ in the max problem $(q/p/a/M)$ for every $a > 0$.*

Proof. The only if part is straightforward. To show the if part, suppose $(q/p/M)$ does not have isotone solutions. Then there exist $0 \leq \alpha_1 < \alpha_2$ and j for which

$$z_j(\alpha_1) > z_j(\alpha_2).$$

where $z(\alpha)$ is the solution of $(q + \alpha p/M)$ for each $\alpha \geq 0$. Define a_j by

$$a_j = \frac{1}{2}(z_j(\alpha_1) + z_j(\alpha_2)) > 0.$$

For $k \neq j$ choose a_k so that

$$a_k > z_k(\alpha_2) \geq 0.$$

Then we have that $a > 0$,

$$z(\alpha_1) \not\leq a \quad \text{and} \quad z(\alpha_2) \leq a.$$

This implies that for this a

$$\hat{\alpha} < \alpha_1 < \alpha_2 \leq \alpha^*$$

in the max problem $(q/p/a/M)$. \square

Comparing Corollary (14), Theorem (16) in this paper and Theorems 1, 2 in Cottle [1], we have the following relationships.

(23) THEOREM. Let $M \in P(n)$.

(i) Every $p \in R^n$ is a convex direction of $\mathcal{R}(a, M)$ at $q \in R_+^n$ for every $a > 0$ if and only if $(q/p/M)$ has isotone solutions for every $p \in R^n$.

(ii) The set $\mathcal{R}(a, M)$ is convex (star-shaped on R_+^n) for every (some) $a > 0$ if and only if $(q/p/M)$ has isotone solutions for every $q \in R_+^n$ and every $p \in R^n$.

Clearly, (i) and parts of (ii) could be obtained by Theorems 1, 2 in [1] and using Property (22).

The strong monotonicity property of the parametric linear complementarity as stated in Cottle [1] is with respect to $q \in R_+^n$ and $M \in P(n)$. It may be of interest to consider a similar property with respect to $p \in R^n$ and $M \in P(n)$, i.e., conditions on $p \in R^n$ and $M \in P(n)$ under which $(q/p/M)$ has isotone solutions for every $q \in R_+^n$. This may be done by using a result appearing in Megiddo [16]. For $q \in R_+^n$, $p \in R^n$ and $M \in P(n)$, define

$$\Gamma(q, p, M) = \{\gamma \in ((n)): q + \alpha p \in C(\gamma; M) \text{ for more than one } \alpha \geq 0\},$$

and

$$\Gamma(p, M) = \bigcup_{q \geq 0} \Gamma(q, p, M).$$

Lemma 3.4 in [16] implies¹:

(24) THEOREM. Let $q \in R_+^n$, $p \in R^n$ and $M \in P(n)$. Then $(q/p/M)$ has isotone solutions if and only if

$$(M_{\gamma\gamma})^{-1}p_\gamma \leq 0$$

¹ The proof given in [16] contains an error. Specifically, in the part (b) of the proof of Lemma 3.4 (p. 9), it is stated that (for a given set of data $\{q, p, M\}$) if p_γ belongs to $\text{pos}\{-M_{\gamma\gamma}\}$ and if γ is in $\Gamma(q, p, M)$, then the problem (p/M) has a solution. This is certainly untrue for $q = (0, 1)^T$, $p = (-1, -1)^T$, $M_{.1} = (1, 0)^T$, $M_{.2} = (1, 1)^T$ and $\gamma = (1)$. Still, the assertion is true, at least under the present assumptions, and can be proved easily. In fact, if $q + \alpha p$ belongs to $C(\gamma; M)$, then the corresponding values of the basic z -variables are given by $z_\gamma(\alpha) = -(M_{\gamma\gamma})^{-1}q_\gamma - \alpha(M_{\gamma\gamma})^{-1}p_\gamma$. The solutions of $(q/p/M)$ are isotone if and only if $z_\gamma(\alpha)$ is isotone for all α such that $q + \alpha p \in C(\gamma; M)$, i.e., $-(M_{\gamma\gamma})^{-1}p_\gamma \geq 0$ for all $\gamma \in \Gamma(q, p, M)$.

for every γ in $\Gamma(q, p, M)$.

An immediate consequence is:

(25) COROLLARY. Let $p \in R^n$ and $M \in P(n)$. Then $(q/p/M)$ has isotone solutions for every $q \in R_+^n$ if and only if

$$(M_{\gamma\gamma})^{-1}p_\gamma \leq 0$$

for every γ in $\Gamma(p, M)$.

The above characterization of the strong monotonicity with respect to p and M is not satisfactory because $\Gamma(p, M)$ (or $\Gamma(q, p, M)$) is not well represented. By restricting the choice of p to negative vectors, we can have a sharper characterization which is "symmetric" to that of the strong monotonicity with respect to q and M (Theorem 1 in [1]).

(26) THEOREM. Let $p \in R^n$, $p < 0$ and $M \in P(n)$. Then $(q/p/M)$ has isotone solutions for every $q \in R_+^n$ if and only if for each $\gamma \in ((n))$

$$(M_{\gamma\gamma})^{-1}p_\gamma \leq 0.$$

Proof. Assume $p < 0$. It suffices to show that for every $\gamma \in ((n))$, $\gamma \in \Gamma(p, M)$. Let $\gamma \in ((n))$. Since $M \in P$, $C(\gamma; M)$ has a nonempty interior. Let $x \in R^n$ be an interior point of $C(\gamma; M)$. Then for $\theta > 0$ sufficiently small, we have that

$$x + \theta p \in C(\gamma; M).$$

Since $p < 0$, we can choose $\lambda > 0$ sufficiently large so that $x - \lambda p$ is nonnegative. Letting

$$q = x - \lambda p \geq 0$$

we have that

$$q + \lambda p = x \in C(\gamma; M) \quad \text{and} \quad q + (\lambda + \theta)p = x + \theta p \in C(\gamma; M).$$

Thus γ belongs to $\Gamma(p, M)$. \square

Restricting p to negative vectors corresponds to considering a special case in the structural engineering problem. In fact, an important special case of the engineering problem including reinforced concrete frame problems has $p < 0$ by its physical nature in addition to $q \geq 0$ and $M \in P$. (See Kaneko [10], [13].) Just as questions (A-1) and (A-2) correspond to the strong monotonicity with respect to $q \geq 0$ and $M \in P$, we could pose the following questions corresponding to the strong monotonicity with respect to $p > 0$ and $M \in P$.

(C-1) What are necessary and sufficient conditions on $p < 0$, $a > 0$ and $M \in P$ such that $\hat{\alpha} = a^*$ in the max problem $(q/p/a/M)$ for every $q \geq 0$?

(C-2) What are necessary and sufficient conditions on $p < 0$ and $M \in P$, such that $\hat{\alpha} = \alpha^*$ in the max problem $(q/p/a/M)$ for every $q \geq 0$ and every $a > 0$?

In the light of Property (22), it is clear that an answer to (C-2) is given by the same conditions for the strong monotonicity with respect to $p < 0$ and $M \in P$ proved in Theorem (26), i.e.,

$$(M_{\gamma\gamma})^{-1}p_\gamma \leq 0$$

for every $\gamma \in ((n))$. Interestingly, this same set of conditions gives an answer to (C-1) also. In closing this paper we prove the following:

(27) THEOREM. Let $p \in R^n$, $p < 0$ and $M \in P(n)$. Then the following are equivalent.

(i) There exists $a > 0$ such that p is a convex direction of $\mathcal{R}(a, M)$ at every $q \in R_+^n$.

(ii) For every $a > 0$, p is a convex direction of $\mathcal{R}(a, M)$ at every $q \in R_+^n$.

(iii) For every $\gamma \in ((n))$,

$$(M_{\gamma\gamma})^{-1}p_\gamma \leq 0.$$

Proof. It suffices to show the equivalence between (i) and (ii). For $a > 0$, we define

$$\mathcal{P}(a) = \{p < 0: p \text{ is a convex direction of } \mathcal{R}(a, M) \text{ for every } q \geq 0\},$$

and prove that $\mathcal{P}(a)$ is the same for every $a > 0$. To do so, let $a > 0$ and assume that $p \notin \mathcal{P}(a)$. Then there exists $q \geq 0$ such that p is not a convex direction of $\mathcal{R}(a, M)$ at q . Letting $z(\alpha)$ be the solution of $(q + \alpha p/M)$ for each $\alpha \geq 0$, it follows that there exist $0 < \alpha_1 < \alpha_2$ for which

$$(28) \quad z(\alpha_1) \not\leq a$$

and

$$(29) \quad z(\alpha_2) \leq a.$$

Now let $a' > 0$ be given. By considering cases, we show that there exists $q' \geq 0$ such that (the same) p is not a convex direction of $\mathcal{R}(a', M)$ at q' . i.e., $p \notin \mathcal{P}(a')$.

Case 1. $a' = \theta a$, $\theta > 0$. For $\alpha \geq 0$ let $z'(\alpha)$ solve $(\theta q + \alpha p/M)$. Clearly, $z'(\theta\alpha) = \theta z(\alpha)$ for each $\alpha \geq 0$. Then by (28) and (29) we have

$$z'(\theta\alpha_1) \not\leq a' \quad \text{and} \quad z'(\theta\alpha_2) \leq a'$$

which implies that p is not a convex direction of $\mathcal{R}(a', M)$ at θq . Thus $p \notin \mathcal{P}(a')$.

Case 2. $a' \geq a$, $a' \neq a$. It is clear that we need consider only the case where $a'_1 > a_1$ and $a'_k = a_k$, $k \neq 1$. By (28), (29) and the continuity of $z(\alpha)$, there exists $\alpha^* \in (\alpha_1, \alpha_2]$ such that

$$(30) \quad z(\alpha^*) \leq a$$

and

$$(31) \quad z(\alpha) \not\leq a, \quad \alpha \in [\alpha_1, \alpha^*).$$

Since $a' \geq a$, obviously

$$z(\alpha^*) \leq a'.$$

Thus, if

$$z(\alpha) \not\leq a'$$

holds for some α in $[\alpha_1, \alpha^*)$, then p is not a convex direction of $\mathcal{R}(a', M)$ at (the same) $q \geq 0$. So assume otherwise; i.e., assume

$$(32) \quad z(\alpha) \leq a', \quad \alpha \in [\alpha_1, \alpha^*].$$

It follows from (31), (32) and the fact that $a'_k = a_k$ for $k \neq 1$ that

$$(33) \quad a_1 < z_1(\alpha) (< a'_1), \quad \alpha \in [\alpha_1, \alpha^*].$$

This, together with (30) and the continuity of $z(\alpha)$ implies that

$$(34) \quad z_1(\alpha^*) = a_1.$$

Since $p < 0$, we can choose $\theta > 0$ sufficiently large so that

$$q' = q + (a'_1 - a_1)(-M_{\cdot 1}) - \theta p \geq 0.$$

Our goal is to show that p is not a convex direction of $\mathcal{R}(a', M)$ at this q' .

To this end we define $z'(\alpha)$, $\alpha \in [\alpha_1 + \theta, \alpha^* + \theta]$, by

$$(35) \quad z'_1(\alpha) = z_1(\alpha - \theta) + (a'_1 - a_1)$$

and

$$(36) \quad z'_k(\alpha) = z_k(\alpha - \theta),$$

for $k \neq 1$. We show that $z'(\alpha + \theta)$ is a solution of $(q' + (\alpha + \theta)p/M)$ for $\alpha \in [\alpha_1, \alpha^*]$. Let $\alpha \in [\alpha_1, \alpha^*]$. By (35), (36) and the fact that $z(\alpha)$ solves $(q + \alpha p/M)$, we have that

$$(37) \quad z'(\alpha + \theta) \geq 0$$

and

$$(38) \quad q' + (\alpha + \theta)p + Mz'(\alpha + \theta) = q + \alpha p + Mz(\alpha) \geq 0.$$

From (33) and (34) we see that $z_1(\alpha)$ is positive, thus by the complementarity between $z_1(\alpha)$ and $(q + \alpha p + Mz(\alpha))_1$,

$$0 = (q + \alpha p + Mz(\alpha))_1 = (q' + (\alpha + \theta)p + Mz'(\alpha + \theta))_1.$$

From (36) and (38), the complementarity with respect to $k \neq 1$ is preserved. Thus $z'(\alpha + \theta)$ solves $(q' + (\alpha + \theta)p/M)$. Now by (33) with $\alpha = \alpha_1$, we have that

$$a_1 < z_1(\alpha_1)$$

and so

$$a'_1 < z'_1(\alpha_1 + \theta);$$

thus

$$z'(\alpha_1 + \theta) \not\leq a'.$$

On the other hand, by (34) and (35)

$$z'_1(\alpha^* + \theta) = a'_1,$$

and for $k \neq 1$

$$z'_k(\alpha^* + \theta) \leq a'_k$$

by (32), (36) and the fact that $a'_k = a_k$; hence

$$z'(\alpha^* + \theta) \leq a'.$$

Case 3. $a' \leq a$, $a' \neq a$. We only consider the case where $a'_1 < a_1$ and $a'_k = a_k$, $k \neq 1$. Let $\theta = a'_1/a_1$ and $a'' = \theta a$. Then by case 1, $p \notin \mathcal{P}(a'')$. Since $0 > \theta > 1$, we have $a'' \leq a'$. Then by Case 2 we have that $p \notin \mathcal{P}(a')$.

Now for an arbitrary $a' > 0$, let $a_\eta \leq a'_\eta$ and $a_\xi > a'_\xi$, where $\xi = \bar{\eta}$. Define a^* by $a_\eta^* = a'_\eta$ and $a_\xi^* = a_\xi$. Then we have

$$a \leq a^* \leq a'.$$

Thus by Cases 2 and 3 we conclude that $p \notin \mathcal{P}(a')$. \square

Acknowledgments. A portion of this paper is a part of the author's doctoral dissertation at the Operations Research Department, Stanford University. The author expresses his gratitude to his principal advisor Professor Richard W. Cottle for the guidance and discussions. Also, he is thankful to referees for a number of valuable comments on the early version of this paper. In particular, he is indebted for the simplified definition of convex directions (§ 3) and a rearrangement of the proof of Theorem (11) (§ 3) which makes the proof much clearer.

REFERENCES

- [1] R. W. COTTLE, *Monotone solutions of the parametric linear complementarity problem*, Math. Programming, 3 (1972), pp. 210–224.
- [2] ———, *On Minkowski matrices and the linear complementarity problem*, Tech. Rep. SOL 75–2, Dept. of Operations Research, Stanford Univ., Jan. 1975; Optimization and Optimal Control, Lecture Notes in Mathematics 477, R. Bulirsch, H. Oettli, and J. Stoer (eds.), Springer-Verlag, Berlin, 1975.
- [3] R. W. COTTLE AND A. F. VEINOTT, JR., *Polyhedral sets having a least element*, Math. Programming, 3 (1972), pp. 238–249.
- [4] G. B. DANTZIG, *Linear Programming and Extensions*, Princeton University Press, Princeton, NJ, 1963.
- [5] O. DE DONATO AND G. MAIER, *Mathematical programming methods for the inelastic analysis of concrete frames allowing for limited rotation capacity*, Internat. J. Numer. Methods Engrg., 4 (1972), pp. 306–329.
- [6] R. L. GRAVES, *A principal pivoting simplex algorithm for linear and quadratic programming*, Operations Res., 15 (1967), pp. 482–494.
- [7] T. IBARAKI, *Complementary programming*, Ibid., 19 (1971), pp. 1523–1529.
- [8] ———, *The use of cuts in complementary programming*, Ibid., 21 (1973), pp. 353–359.
- [9] I. KANEKO, *Isotone solutions of parametric linear complementarity problems*, Math. Programming, to appear.
- [10] ———, *A mathematical programming method for the inelastic analysis of reinforced concrete frames*, Internat. J. Numer. Methods Engrg., to appear.
- [11] ———, *The parametric linear complementarity problem in the De Donato–Maier analysis of reinforced concrete beams*, Tech. Rep. Sol 75–13, Dept. of Operations Research, Stanford Univ., Stanford, CA, May, 1975.
- [12] ———, *Linear complementarity problems and characterizations of Minkowski matrices*, Linear Algebra and Appl., to appear.
- [13] ———, *Linear complementarity problems in plastic structural analysis*, to appear.
- [14] K. KIRCHGÄSSNER, *Ein Verfahren zur Maximierung Linearer Funktionen in Nichtkonvexen Bereichen*, Z. Angew. Math., (1962), pp. T22–24.
- [15] G. MAIER, *A matrix structural theory of piecewise linear elastoplasticity with interacting yield planes*, Meccanica–J. Italian Assoc. Theoret. Appl. Mech., 5 (1970), pp. 54–66.
- [16] N. MEGIDDO, *On Monotonicity in Parametric Linear Complementarity Problems*, Dept. of Statistics, Tel Aviv Univ., Israel, July 1975.

- [17] K. G. MURTY, *On the number of solutions of the complementarity problems and spanning properties of complementary cones*, Linear Algebra and Appl., 5 (1972), pp. 65–108.
- [18] T. D. PARSONS, *A combinatorial approach to convex quadratic programming*, Doctoral dissertation, Dept. of Math., Princeton Univ., Princeton, NJ, May 1966.
- [19] H. SAMELSON, R. M. THRALL AND O. WESLER, *A partition theorem for Euclidean N -space*, Proc. Amer. Math. Soc., 9 (1958), pp. 805–807.

CUTTING-PLANES FOR COMPLEMENTARITY CONSTRAINTS*

R. G. JEROSLOW†

Abstract. A characterization is given of all the cutting-planes for a generalized linear complementarity problem, in terms of rules whose repeated application yields exactly these valid implied inequalities.

This report is a revision of our paper (1976), and our earlier proofs have been substantially simplified.

Introduction. We provide a characterization of the set of all valid inequalities for a constraint system (see (GLC) below) that simultaneously generalizes the linear complementarity problem and bivalent integer programming. This characterization is in terms of rules for generating cutting-planes, which are easily proven to be valid rules. We also show that the repeated application of these rules can obtain any valid cutting-plane.

Our characterization does yield a finite procedure for generating a set of linear inequalities that define the convex hull of feasible solutions to the constraint system, as does Balas' result (1974) for facial constraints (see § 1), on which it is based. When instead one feasible solution is desired, which maximizes a linear form, there is the further issue of how to "activate" the cutting-planes "as needed". We will discuss this issue in a later paper.

Our result is comparable to Chvátal's characterization of all valid cutting-planes for a bounded integer program (1973), and it is essentially an extension of Blair's characterization of the cutting-planes for a bivalent integer program (1976). All three characterizations are in terms of rules for generating cutting-planes, and include the rule of taking nonnegative combinations of given inequalities. They differ in the additional rule or rules.

Chvátal's one additional rule involves translating a hyperplane; the additional rules of Blair, and those given here, involve certain kinds of "simultaneous rotations" of pairs of hyperplanes which are rotated until they coincide. Blair's additional rules, and ours also, can be sequenced, in that each rule is applied once to the set of previously derived cuts, in any fixed order of application of the rules.

Related work, in which cutting-planes are developed for a nonconvex objective function, is in Konno (1976); here our emphasis is on certain nonconvex constraints sets.

This paper is a revision of Jeroslow (1976). The formulations here, in terms of sets of inequalities, as opposed to the proof trees used in Jeroslow (1976), were prompted by the referee.

A point of terminology deserves mention: we use the term "cutting-plane" synonymously with "valid implied linear inequality".

* Received by the editors June 29, 1976, and in revised form April 4, 1977.

† Department of Mathematics and Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213. Revised at the Center for Operations Research and Econometrics, Université Catholique de Louvain, Louvain, Belgium.

1. The main result. We first consider constraint sets over the reals of the following nature, which involve logical constraints on linear inequalities:

$$Dx \geq d$$

and for each $h = 1, \dots, t$, at least one of the constraints

(1)

$$d^i x \geq d_{i0}$$

holds for some $i \in J_h$.

In (1), each set J_h is a nonempty, finite set, and each $d^i x \geq d_{i0}$ is a linear inequality; $x = (x_1, \dots, x_r)$, D is a $p \times r$ matrix, and d is a $p \times 1$ vector.

The constraint set (1) is called *facial* (Balas (1974)) if, for each $h = 1, \dots, t$ and $i \in J_h$, $\{x | d^i x \geq d_{i0}, Dx \geq d\}$ is a face (possibly empty) of $\{x | Dx \geq d\}$ (see Rockafellar (1970) for terminology).

For example, the following constraint set is facial:

$$(GLC) \quad Ay + Bz \geq d, \quad y \geq 0, \quad z \geq 0, \quad y \cdot z = 0,$$

in which $y = (y_1, \dots, y_s)$ and $z = (z_1, \dots, z_s)$ have s variables. One has $D = [A : B]$, $x = (y, z)$, $r = 2s$, $t = s$, $J_h = \{h, h + s\}$, and $d^h x \geq d_{h0}$ is $-x_h \geq 0$. Thus, the requirement that at least one constraint $d^i x \geq d_{i0}$ holds for $i \in J_h$ is equivalent to the requirement that either $y_h \leq 0$ or $z_h \leq 0$. Both inequalities yield faces of $\{(y, z) | Ay + Bz \geq d, y \geq 0, z \geq 0\}$ because of the nonnegatives $y, z \geq 0$; and for the same reason this requirement is equivalent to $y_h z_h = 0$. Since such a requirement is imposed for $h = 1, \dots, t$, we have $y \cdot z = 0$.

Note that (GLC) includes the linear complementary problem Cottle and Dantzig (1968), Eaves (1971), Lemke (1965) and Lemke and Howson (1964) as a special case. More general problems fit in the framework (GLC), for indeed the special case of (GLC) given by

$$(2) \quad \begin{aligned} Ay &\geq d, \quad y \geq 0, \quad z \geq 0, \quad y \cdot z = 0, \\ y_j + z_j &= 1 \quad \text{for } j = 1, \dots, s, \end{aligned}$$

is clearly equivalent to the constraints of the bivalent integer program, specifically

$$(BIP) \quad Ay \geq d, \quad y_j = 0 \text{ or } 1 \quad \text{for } j = 1, \dots, s.$$

In what follows, $\text{clconv}(S)$ denotes the closure of the convex hull of S (see Rockafellar (1970)) and the convex hull of the empty set is empty.

Define inductively the convex polyhedra:

$$(3) \quad K_0 = \{x | Dx \geq d\},$$

$$(4) \quad K_{h+1} = \text{clconv} \bigcup_{i \in J_h} (K_h \cap \{x | d^i x \geq d_{i0}\}) \quad \text{for } 0 \leq h \leq t-1.$$

We shall need the following result of Balas.

THEOREM 1 (Balas (1974)). *If K_0 is bounded and (1) is facial, then*

$$(5) \quad K_t = \text{clconv} \{x | (1) \text{ holds}\}.$$

We shall first state our main result under a boundedness assumption, which we will later remove in § 2. It may be worth noting that the hypothesis, that $\{x | Dx \geq d'\}$ is nonempty and bounded for some r.h.s. d' , implies $\{x | Dx \geq 0\} = \{0\}$. Consequently, this hypothesis implies, that, for any d , $\{x | Dx \geq d\}$ is bounded.

Our main result announced in Jeroslow (1976) is the following.

THEOREM 2. *If*

$$(6) \quad \{(y, z) | Ay + Bz \geq d', y \geq 0, z \geq 0\}$$

is bounded and nonempty for some d' , then any valid implied inequality for (GLC) is obtained by starting from the linear defining inequalities

$$(7) \quad Ay + Bz \geq d, \quad y \geq 0, \quad z \geq 0,$$

and applying, finitely often, the following two rules (the second for $h = 1, \dots, s$):

(i) *Take nonnegative combinations of given inequalities, and possibly reduce the right-hand side.*

(ii)_h *Having already obtained the two inequalities*

$$(8a) \quad \alpha_1 y_1 + \dots + u y_h + \dots + \alpha_s y_s + \beta_1 z_1 + \dots + t z_h + \dots + \beta_s z_s \geq \alpha_0,$$

$$(8b) \quad \alpha_1 y_1 + \dots + u' y_h + \dots + \alpha_s y_s + \beta_1 z_1 + \dots + t' z_h + \dots + \beta_s z_s \geq \alpha_0,$$

one may deduce

$$(8c) \quad \alpha_1 y_1 + \dots + u y_h + \dots + \alpha_s y_s + \beta_1 z_1 + \dots + t' z_h + \dots + \beta_s z_s \geq \alpha_0.$$

Conversely, any inequality thus obtained is valid for the complementarity constraints (GLC).

Before proving the theorem, we give an example to fix the idea of the rule (ii)_h, with $s = 3$, $h = 2$. From the two inequalities

$$y_1 + 2y_2 + 3y_3 + 4z_1 + 5z_2 + 6z_3 \geq 7$$

and

$$y_1 + 8y_2 + 3y_3 + 4z_1 + 3z_2 + 6z_3 \geq 7$$

one may deduce

$$y_1 + 2y_2 + 3y_3 + 4z_1 + 3z_2 + 6z_3 \geq 7.$$

In this example, we have made coefficients distinct whenever possible, e.g. the coefficients of y_1 , y_3 , z_1 , and z_3 in both inequalities, as well as the constant terms, must agree.

Proof. First, we show that the rules (i), (ii)_h yield valid implied inequalities. The validity of (i) holds in any ordered field; it suffices to show that (ii)_h is valid.

Suppose that (y, z) satisfies (GLC), and that (8a), (8b) are valid for all solutions to (GLC). Since $y \cdot z = 0$, we have $y_h z_h = 0$, hence either $y_h = 0$ or $z_h = 0$. If $y_h = 0$, then since (8b) is true, so is (8c). If $z_h = 0$, then since (8a) is true, so is (8c). In either case, (8c) is true. Since (y, z) was an arbitrary solution to (GLC), we see that (8c) is also valid for all solutions to (GLC). Thus the application of rule (ii)_h preserves validity, and hence any finite number of applications does also.

Next, we must show that any valid cutting-plane can be obtained in finitely many applications of (i), (ii)_h starting from (7). This is done by induction. First, we show the “ground step” that any valid inequality for K_0 is obtained via rule (i) above. Second, we show the “inductive step”, that any valid inequality for K_h , $h \geq 1$, is obtained via one application of rule (ii)_h, with two applications of rule (i), to a suitable set of defining inequalities for K_{h-1} . It then will follow that the repeated application of rules (i) and (ii)_h, $h = 1, \dots, s$, with applications of (ii)_{h-1} sequenced to precede those of (ii)_h ($h \geq 1$), yields any valid inequality of K_r . By Theorem 1, our proof will then be complete.

To establish the “ground step”, we differentiate two cases, accordingly as $K_0 \neq \emptyset$ or $K_0 = \emptyset$.

If $K_0 \neq \emptyset$, and

$$(9) \quad \pi y + \sigma z \geq \pi_0$$

is a valid inequality for K_0 , then the optimal value of the linear program

$$(10) \quad \begin{aligned} & \text{minimize } \pi y + \sigma z \\ & \text{subject to } Ay + Bz \geq d, \quad y, z \geq 0 \end{aligned}$$

is $\pi'_0 \geq \pi_0$, and (10) is consistent. Hence the dual linear program to (10) has value π'_0 , i.e., there are $\lambda, \theta, \gamma \geq 0$ with

$$(11) \quad \lambda A + \theta = \pi, \quad \lambda B + \gamma = \sigma, \quad \lambda d = \pi'_0 \geq \pi_0.$$

Therefore, (9) arises by rule (i).

If $K_0 = \emptyset$, and (9) is any inequality, by the fact that (6) is nonempty and bounded, reasoning similar to the case for $K_0 \neq \emptyset$ obtains $\lambda, \theta, \gamma \geq 0$ with

$$(12) \quad \lambda A + \theta = \pi, \quad \lambda B + \gamma = \sigma.$$

Since $K_0 = \emptyset$, there are also $\lambda^0, \theta^0, \gamma^0 \geq 0$ with

$$(13) \quad \lambda^0 A + \theta^0 = 0, \quad \lambda^0 B + \gamma^0 = 0, \quad \lambda^0 d > 0.$$

Thus for all real scalars $\rho \geq 0$,

$$(14) \quad (\lambda + \rho \lambda^0)A + (\theta + \rho \theta^0) = \pi, \quad (\lambda + \rho \lambda^0)B + (\gamma + \rho \gamma^0) = \sigma$$

and for $\rho = \rho_0$ large enough, $(\lambda + \rho_0 \lambda^0)d \geq \pi_0$. Again, (9) arises by rule (i).

To establish the “inductive step”, we proceed as follows. We suppose that some defining set of inequalities for K_{h-1} are given:

$$(15) \quad K_{h-1} = \{(y, z) | A^0 y + B^0 z \geq d^0, y, z \geq 0\}.$$

Regarding this defining set, we shall require that

$$(16) \quad \{(y, z) | A^0 y + B^0 z \geq d^*, y, z \geq 0\}$$

is bounded and nonempty for some d^* . This is no restriction. If $K_{h-1} \neq \emptyset$ then the fact that $K_{h-1} \subseteq K_0$ and K_0 is bounded (as (6) is bounded and nonempty by assumption) shows that we can take $d^* = d^0$. On the other hand, if $K_{h-1} = \emptyset$, we may assume that $y \leq 0$ and $z \leq 0$ are included in the defining inequalities for K_{h-1} in (15), and we may take $d^* = 0$.

Suppose that (8c) is valid for K_h . It suffices to show that, for suitably large t , (8a) is valid for K_{h-1} , and that, for suitably large u' , (8b) is valid for K_{h-1} . For then, by the kind of reasoning used for K_0 , (8a) and (8b) are obtained from the defining inequalities of K_{h-1} in (15) by rule (i), and (8c) is then obtained by one application of (ii)_h. Without loss of generality, we show (8a) is valid for K_{h-1} for certain t ; if $K_{h-1} = \emptyset$ we may take $t = t'$, so we also may assume $K_{h-1} \neq \emptyset$.

We distinguish two cases, accordingly as the minimum M of z_h for $(y, z) \in K_{h-1}$ is zero or positive.

If $M > 0$, let θ be the minimum of the linear form $l(y, z) = \alpha_1 y_1 + \dots + u y_h + \dots + \alpha_s y_s + \beta_1 z_1 + \dots + 0 \cdot z_h + \dots + \beta_s z_s$ subject to $(y, z) \in K_{h-1}$. Then for $(y, z) \in K_{h-1}$ we have for $\alpha'_0 \geq \theta$,

$$(17) \quad \begin{aligned} \alpha_1 y_1 + \dots + u y_h + \dots + \alpha_s y_s + \beta_1 z_1 + \dots + \frac{\alpha'_0 - \theta}{M} z_h + \dots + \beta_s z_s \\ = l(y, z) + \left(\frac{\alpha'_0 - \theta}{M} \right) z_h \geq \theta + \left(\frac{\alpha'_0 - \theta}{M} \right) M \geq \alpha'_0. \end{aligned}$$

Hence with $t = (\alpha'_0 - \theta)/M$, (8a) is valid for K_{h-1} , by taking $\alpha'_0 \geq \alpha_0$.

If $M = 0$, let $\theta(v)$ denote the minimum of $l(y, z)$ subject to $(y, z) \in K_{h-1}$ and $z_h = v$. Recall that the perturbation function of a linear program is a convex function with subgradients wherever it is finite (Rockafellar (1970)). Also note that the validity of (8c) for K_h implies that $\theta(0) \geq \alpha_0$, due to the fact that $K_{h-1} \cap \{(y, z) | z_h = 0\} \subseteq K_h$.

Therefore there is a scalar t , representing the negative of the v -component of a subgradient to the perturbation function at $(d^0, 0, 0, 0)$ (the last zero for $z_h = v$ at $v = 0$), such that

$$(18) \quad \theta(v) \geq \theta(0) - t(v - 0) \geq \alpha_0 - tv$$

for all v . From (18), if $(y, z) \in K_{h-1}$ we have

$$(19) \quad l(y, z) \geq \theta(z_h) \geq \alpha_0 - tz_h,$$

and (19) shows that (8a) is valid for K_{h-1} .

This completes the "inductive step" of our proof. Q.E.D.

The construction in the proof of Theorem 2 shows somewhat more than is claimed in the theorem, and we next state this additional information. To verify Corollary 3, let S_{h-1} for $h = 1, \dots, t+1$ denote the set of defining inequalities in (15) for K_{h-1} , with the requirement given at that point of the proof, i.e., that (16) is bounded and nonempty for some d^* .

COROLLARY 3. *Suppose that (6) is bounded and nonempty for some d' .*

There are finite sets of linear inequalities S_0, \dots, S_t with the following properties:

1. S_0 consists of the inequalities $Ay + Bz \geq d$, $y \geq 0$, $z \geq 0$.
2. Any inequality in S_h , $h \geq 1$, arises by two applications of rule (i) to the inequalities of S_{h-1} , followed by one application of rule (ii)_h.
3. Any valid inequality for (GLC) arises by an application of rule (i) to the inequalities in S_t .

We remark here that Blair's characterization (Blair (1976)) of the valid inequalities for (BIP) can be obtained by an argument virtually identical to the proof of Theorem 2. This comment serves to unify the two characterizations conceptually, though a proof of Blair's result by our methods would be longer than the elegant proof in Blair (1976).

2. Generalizations. Somewhat more general results than Theorem 2 and Corollary 3 can be obtained by the same methods.

Consider the constraint system

$$(CMP) \quad \begin{aligned} Ax &\geq b, & x &\geq 0, \\ \sum_{h=1}^t \prod_{i \in J_h} \sum_{k \in K(i)} x_k &= 0. \end{aligned}$$

This is of the form (1) with

$$(20) \quad d^i x = - \sum_{k \in K(i)} x_k, \quad d_{i0} = 0,$$

as $x \geq 0$ is included in the constraints of (CMP), and for similar reasons (CMP) is facial.

For each $h = 1, \dots, t$ define the function $h(j)$, where $h(j) = i$ if i is the j th element in some linear order in J_h . The domain of $h(j)$ is all integers $j = 1, \dots, |J_h|$.

The rule (CMC) $_h$ of Fig. 1 is clearly valid for (CMP), where the line means that the inequality below it can be obtained from those above it. Indeed, if x

$$\begin{array}{c} a_{11}x_1 + \dots + a_{1r}x_r \geq a_0, \dots, a_{j1}x_1 + \dots + a_{jr}x_r \geq a_0, \dots, a_{h'1}x_1 + \dots + a_{h'r}x_r \geq a_0 \\ \hline a_1x_1 + \dots + a_rx_r \geq a_0 \\ \text{where } a_{jk} = a_k \text{ if } k \notin K(h(j)) \end{array} \quad (h' = |J_h|)$$

FIG. 1. The rule (CMC) $_h$ for J_h

satisfies the constraints of (CMP), then for some $i \in J_h$ we have $x_k = 0$ for all $k \in K(i)$, and for some j we have $h(j) = i$. Supposing that all the inequalities above the line in Fig. 1 are valid for (CMP), then $a_{j1}x_1 + \dots + a_{jr}x_r \geq a_0$ holds. Therefore so does $a_1x_1 + \dots + a_rx_r \geq a_0$, since $a_{jk} = a_k$ if $x_k \neq 0$.

By an analysis similar to that in our proof of Theorem 2, the next result can be obtained. Basically, one modifies primarily the inductive step of that proof, using the linear form $-d^i x$ of (20) in place of z_h in the argument concerning subgradients.

THEOREM 4. Suppose that

$$(21) \quad \{x \geq 0 \mid Ax \geq b'\}$$

is bounded and nonempty for some b' .

There are finite sets of linear inequalities S_0, \dots, S_t with the following properties:

1. S_0 consists of the inequalities $Ax \geq b, x \geq 0$.
2. Any inequality in $S_h, h \geq 1$, arises by $|J_h|$ applications of rule (i) to the inequalities of S_{h-1} , followed by one application of the rule $(CMC)_h$.
3. Any valid inequality for (CMP) arises by an application of rule (i) to the inequalities in S_i .

In Jeroslow (1976) we discuss how the hypotheses, that (6) or (21) is bounded and nonempty, can be removed. To summarize that discussion in the more general setting of (CMP), one adds, to the inequalities $Ax \geq b, x \geq 0$ in S_0 , the new inequality

$$(22) \quad x_1 + \cdots + x_r \leq M$$

where M is literally an infinite quantity of the ordered field extension $R(M)$ of M .

The boundedness assumption is then satisfied in $R(M)$, and a close examination of suitable proofs of Theorem 4 will reveal that the theorem is true in any ordered field (the existence of subgradients for linear programs is valid in any ordered field). Then one applies Theorem 4 in $R(M)$. See Jeroslow (1976) for details.

Thus, while the "phantom" quantity M will appear among many inequalities in the sets S_h , M disappears, when a valid inequality for (CMP) involving only real numbers is obtained from S_i by rule (i).

REFERENCES

- E. BALAS (1974), *Disjunctive programming: Properties of the convex hull of feasible points*, MSRR 348, GSIA, Carnegie-Mellon Univ., Pittsburgh, PA.
- C. E. BLAIR (1976), *Two rules for deducing valid inequalities for 0-1 problems*, SIAM J. Appl. Math., 31, pp. 614-617.
- V. CHVÁTAL (1973), *Edmonds polytopes and a heirarchy of combinatorial problems*, Discrete Math., 4, pp. 305-337.
- R. W. COTTLE (1974), *Complementarity and variational problems*, Tech. Rep. SOL 74-76, Systems Optimization Laboratory, Stanford Univ., Stanford, CA.
- R. W. COTTLE AND G. B. DANTZIG (1968), *Complementary pivot theory of mathematical programming*, Linear Algebra and Appl., 1, pp. 103-125.
- B. C. EAVES (1971), *The linear complementarity problem*, Management Sci., 17, pp. 612-634.
- F. GLOVER (1975), *Polyhedral annexation in mixed integer and combinatorial programming*, Math. Programming, 9, pp. 161-188.
- R. G. JEROSLOW (1976), *Cutting-planes for complementarity constraints*, MSRR 394, GSIA, Carnegie-Mellon Univ., Pittsburgh, PA.
- (1976), *Cutting-planes for complementarity constraints*, Notices AMS, 23, p. A-364.
- H. KONNO (1976), *A cutting-plane algorithm for solving bilinear programs*, Math. Programming, 11, pp. 14-27.
- C. E. LEMKE (1965), *Bimatrix equilibrium points and mathematical programming*, Management Sci., 11, pp. 681-689.
- C. E. LEMKE AND J. T. HOWSON (1964), *Equilibrium points of bimatrix games*, J. Soc. Indust. Appl. Math., 12, pp. 413-423.
- R. T. ROCKAFELLAR (1970), *Convex Analysis*, Princeton University Press, Princeton, NJ.

SUFFICIENCY THEOREM FOR DISCONTINUOUS OPTIMAL COST SURFACES*

HAROLD STALFORD†

Abstract. A V type function is introduced for the purpose of modeling the optimal cost surfaces (OCS) of a very general class of optimal control processes, especially state-constrained processes. This classification of the OCS builds from the assumption that the OCS is the countable union of disjoint submanifolds. The V -type function is particularly suitable for handling OCSs that are discontinuous at submanifolds of codimension 1 of the state space. A discontinuous nontransversality condition is presented which keeps trajectories from entering any submanifold of the OCS from a discontinuous edge. A sufficiency theorem utilizing the V -type function is given. This theorem states that if certain conditions are met by a control policy, then it is optimal. The defining of these conditions, together with the sufficiency theorem, provides a classification of the OCS for general optimal control processes with state constraints. An example is provided to demonstrate the implementation of the results.

1. Introduction. The optimal cost surface (OCS) is the plot of the values of the optimal value function above the state space. This surface is obtained as part of the solution of an optimal control problem. Even though the properties of this surface are dependent entirely on the given elements of the problem, researchers have found it profitable to hypothesize about its properties in deriving results in optimal control theory. For instance, current field-type sufficiency theorems require that the OCS be continuous [1]–[8], [16]–[18]. Additional discussion on works [16]–[18] using Young's approach to develop field-type sufficiency theorems is given at the end of the paper. The derivation of the fundamental partial differential equation of dynamic programming in [9] of Dreyfus are based on the OCS having bounded second partial derivatives. These properties, however, are not always met, even in rather simple control problems, i.e., Pontryagin [10]. Recent Lotka–Volterra models of prey predator control systems provide simple examples where the OCS is discontinuous and has unbounded first partial derivatives, e.g., Vincent [11], [12]. A discontinuous OCS arises particularly in state-constrained optimal control processes, i.e., Litt [13].

An assumption which has survived all known examples is that the OCS is the countable union of disjoint submanifolds. In this paper we examine the problem of arranging these submanifolds in such a way that the properties of the resulting OCS can be utilized in establishing optimality over a very general class of optimal control processes, especially state-constrained processes. In particular, we model the optimal value function with a V -type function. This function may be locally discontinuous on submanifolds of codimension 1. A nontransversality condition and a discontinuous nontransversality condition are presented. The latter condition keeps a trajectory of the control process from entering any submanifold of the OCS from a discontinuous edge. These properties and conditions are formulated in a sufficiency theorem which states that, if certain conditions are met by a control policy, then it is optimal. The defining of these properties and conditions, together with the sufficiency theorem, provides a classification of the OCS for general

* Received by the editors October 21, 1975, and in revised form March 28, 1977.

† Systems Analysis Department, Dynamics Research Corporation, Wilmington, Massachusetts 01887. The work of this author was supported by the Office of Naval Research.

optimal control processes with state constraints. An example is given to demonstrate the implementation of the results.

2. Definition of optimal control processes. The optimal control process under investigation has its dynamical behavior governed by a system of ordinary differential equations and has its evolution of state described by the motion of a point in n dimensional Euclidean space E^n . The seven basic elements needed in defining the optimal control process are four functions (f , U , f_0 , and g_0), two sets (X and Θ), and a function space Ω . These elements are described subsequently.

The dynamic behavior of the optimal control process is modeled by the state velocity function f in the state equation

$$(1) \quad \dot{x} = f(x, \nu), \quad x \in E^n, \quad \nu \in E^m,$$

where f is a function with domain $E^n \times E^m$. We let $\varphi(\cdot)$, an absolutely continuous function, represent a solution of (1) when controlled by a control policy $u(\cdot)$, a Lebesgue measurable function of time. With initial time t_0 the initial state satisfies

$$(2) \quad \varphi(t_0) = x_0.$$

The state space X is a subset of E^n . It is considered to be a topological space possessing the induced topology from E^n . The terminal set Θ is a closed subset of X .

The controller of the process is equipped with the elements Ω and U . The control function space Ω is the space of all Lebesgue measurable functions of time defined on bounded intervals with range in E^m . Constraints on the control functions in Ω are given implicitly by the set-valued function

$$(3) \quad U: X \rightarrow \text{power set of } E^m.$$

Given $x \in X$, the set $U(x)$ is a set of control values available to the controller at the state x . $U(x)$ is nonempty for all $x \in X$.

A solution of (1) for some control $u \in \Omega$, $u: [t_0, t_f] \rightarrow E^m$, and given initial conditions is called a *trajectory*. A trajectory $\varphi: [t_0, t_f] \rightarrow E^n$ is said to be *admissible* if it lies entirely in the state space X for all times t contained in $[t_0, t_f]$. An admissible trajectory is said to be *terminating* if $\varphi(t_f)$ is contained in Θ . The time t_f is called the *terminating* or *final time for a terminating admissible trajectory*. The time t_f belongs to the interval $[t_0, \infty]$. t_f does not have to be the same terminating time for distinct trajectories unless it is constrained to be fixed by the terminal set Θ . For nonautonomous systems (that is, f an explicit function of t), one component of φ is the time t itself.

A control $u \in \Omega$, $u: [t_0, t_f] \rightarrow E^m$ is said to be admissible if it has at least one corresponding admissible trajectory $\varphi: [t_0, t_f] \rightarrow X$ such that

$$(4) \quad u(t) \in U[\varphi(t)] \quad \forall t \in [t_0, t_f].$$

Here, the trajectory φ corresponds to the control u if

$$(5) \quad \varphi(t) - \varphi(t_0) = \int_{t_0}^t f[\varphi(\tau), u(\tau)] d\tau$$

for all $t \in [t_0, t_f]$.

Let x_0 be contained in X . Let $C(x_0)$ denote the set of all admissible controls having at least one terminating admissible trajectory emanating from x_0 . For $u \in C(x_0)$, let $T(x_0; u)$ denote the set of all terminating admissible trajectories φ emanating from x_0 , corresponding to the control u , and satisfying (4). The domains of u and φ coincide for all φ contained in $T(x_0; u)$.

DEFINITION 1. A *denumerable decomposition* D of a set $X \subset E^n$ is defined to be a denumerable collection of pairwise disjoint subsets whose union is X . This is written as $D = \{X_j; j \in J\}$, where J is a denumerable index set of the disjoint subsets.

Let B be a subset of E^n . A mapping $F: B \rightarrow R$ is said to be *differentiable locally Lipschitzian* if and only if there is an open set W containing B such that F may be extended to a function which is differentiable and locally Lipschitzian on W .

DEFINITION 2. Let X be a subset of E^n and D a denumerable decomposition of X . A real-valued function $V: X \rightarrow E^1$ is said to be *piecewise differentiable locally Lipschitzian with respect to D* if, for $j \in J$, the restriction $V|X_j: X_j \rightarrow E^1$ is differentiable and locally Lipschitzian; that is, there exists a collection $\{(W_j, V_j): j \in J\}$ such that W_j is an open set containing X_j , $V_j: W_j \rightarrow E^1$ is differentiable locally Lipschitzian, and $V_j(x) = V(x)$ for $x \in X_j$. We say the collection $\{(W_j, V_j): j \in J\}$ is associated with V and D .

If the process is in the state $x_0 \in X$, then it is to be controlled during a transfer of the process to the terminal set Θ so as to render the criterion

$$(6) \quad J(x_0, \varphi, u) \triangleq g_0[\varphi(t_f)] + \int_{t_0}^{t_f} f_0[\varphi(\tau), u(\tau)] d\tau$$

a minimum value, where g_0 , a real-valued function, is continuous and piecewise differentiable locally Lipschitzian with respect to a decomposition over a neighborhood of the terminal set Θ ; f_0 is a real-valued bounded Borel-measurable function with domain $E^n \times E^m$; $u \in C(x_0)$; and $\varphi \in T(x_0; u)$.

In summary, the control process is represented by the septuple $(f, U, f_0, g_0, X, \Theta, \Omega)$ where f is a function, U is a set-valued map, f_0 is bounded Borel-measurable, g_0 is continuous and piecewise differentiable locally Lipschitzian, X is a subset of E^n , Θ is closed in X , and Ω is the space of Lebesgue measurable controls. In particular, $C(x_0)$ represents the set of admissible controls at $x_0 \in X$, and $T(x_0; u)$ denotes the set of all terminating admissible trajectories emanating from x_0 due to the control $u \in C(x_0)$. We will let Γ denote this control process.

DEFINITION 3 (Optimality). Let $x_0 \in X$, $u^* \in C(x_0)$, and $\varphi^* \in T(x_0; u^*)$. The pair (u^*, φ^*) is said to be *optimal* at x_0 if and only if, for all $u \in C(x_0)$ and for all $\varphi \in T(x_0; u)$, the following inequality is satisfied:

$$(7) \quad J(x_0, \varphi^*, u^*) \leq J(x_0, \varphi, u).$$

If the pair (u^*, φ^*) is optimal at x_0 , then the value $J(x_0, \varphi^*, u^*)$ is arbitrarily defined to be $V(x_0)$. If an optimal pair (u^*, φ^*) exists for every $x_0 \in X$, then we have a real-valued function defined on X :

$$(8) \quad V: X \rightarrow E^1.$$

It is for this function—the optimal value function (OVF)—that we want to prescribe very general properties, those that will hold over a large class of optimal control processes. We do this in the next section.

3. Formulation of discontinuous optimal value function. This section is devoted to setting up some very general properties of the optimal value function that are compatible with the dynamics of optimal control processes which admit continuous or discontinuous optimal value functions.

LEMMA 1 (A monotonicity lemma for a discontinuous optimal value function). *Let $D = \{X_j: j \in J\}$ be a denumerable decomposition of X . Let $\varphi: [t_0, t_f] \rightarrow X$ be absolutely continuous and $h_0: [t_0, t_f] \rightarrow E^1$ be integrable. Let $V: X \rightarrow E^1$ be piecewise differentiable locally Lipschitzian with respect to D . Let $\{(W_j, V_j): j \in J\}$ be a collection associated with V and D . Let $T_j = \{t \in [t_0, t_f]: \varphi(t) \in X_j\}$ for $j \in J$. Suppose that*

(i) *for each $j \in J$*

$$h_0(t) + \frac{d}{dt}(V_j \circ \varphi)(t) \geq 0 \quad \text{almost everywhere in } T_j;$$

(ii) *there exists a countable compact subset T of $[t_0, t_f]$ such that $V \circ \varphi$ is continuous on the open set $\theta = (t_0, t_f) \sim T$;*

(iii) *if t is a point of discontinuity of $V \circ \varphi$, then $\inf \{\sup \{(V \circ \varphi)(\tau): 0 < t - \tau < r, \tau \in [t_0, t_f]\}: r > 0\} \leq (V \circ \varphi)(t) \leq \sup \{\inf \{(V \circ \varphi)(\tau): 0 < \tau - t < r, \tau \in [t_0, t_f]\}: r > 0\}$. For the case $t = t_0$ or $t = t_f$ only one inequality is used.*

Then the function

$$g(t) = \int_{t_0}^t h_0(\tau) d\tau + (V \circ \varphi)(t)$$

defined for $t \in [t_0, t_f]$ is monotone nondecreasing.

Proof. The set θ , being open, is the denumerable union of disjoint open intervals

$$\theta = \bigcup \{\theta_i: i \in I\}$$

where I is the index set of these disjoint open intervals. For each $i \in I$ let the interval (a_i, b_i) designate θ_i . Let $[d_i, e_i] \subset (a_i, b_i)$. Since the function $V \circ \varphi$ is continuous on $[d_i, e_i]$, it follows from Theorem 3.1 of Stalford [6] that g is monotone nondecreasing on $[d_i, e_i]$; an hypothesis of Theorem 3.1 requires V to be continuous over X , but its proof only uses the continuity of $V \circ \varphi$.

Consequently, it follows that g is monotone nondecreasing over (a_i, b_i) . From condition (iii) it follows that g is monotone nondecreasing over $[a_i, b_i]$.

Suppose that the intervals (a_i, b_i) and (a_{i+1}, b_{i+1}) are such that $b_i = a_{i+1}$. Thus, b_i is an isolated point in T . From the above analysis we know that g is monotone nondecreasing over $[a_i, b_i]$ and $[b_i, b_{i+1}]$. Thus g is monotone nondecreasing over $[a_i, b_{i+1}]$. In this way, we can remove all isolated points of T . Every countable compact subset of the real line contains an isolated point; i.e., see Stalford and Leitmann [14]. Let I_0 be the set of all isolated points of T . Let $T_1 = T \sim I_0$. Let $\theta_1 = (t_0, t_f) \sim T_1$. The set T_1 is countable compact and θ_1 is open. Let I_1 be the isolated points of T_1 .

Let

$$\theta_1 = \bigcup_{k \in K} (a_{1k}, b_{1k}).$$

From the above analysis we have established that g is monotone nondecreasing over (a_{1k}, b_{1k}) , $k \in K$. If $b_{1k} \in I_1$, then by evoking condition (iii), we have g monotone nondecreasing over (a_{1k}, b_{1k+1}) where $b_{1k} = a_{1k+1}$. By repeated use of condition (iii), we establish that the monotonicity of g is extendable through all the isolated points of T_1 . Let $T_2 = T_1 \sim I_1$ and $\theta_2 = (t_0, t_f) \sim T_2$. Let I_2 be the isolated points of T_2 .

Using the process of transfinite induction as developed by Stalford and Leitmann [14], we extend the monotonicity of g to the entire interval $[t_0, t_f]$ by removing, in the above fashion, the isolated points I_k from T_k . If T_{k+1} is nonempty then it contains isolated points. Since points of T are being removed at each stage and T is countable, the entire set T is evacuated by transfinite induction. This completes the proof of the lemma.

In this section we speak of a manifold M to mean an embedded continuously differential $(n-1)$ dimensional submanifold with boundary (the boundary may be empty) of E^n such that there exists a manifold \tilde{M} containing M where \tilde{M} is an embedded, continuously differentiable, $(n-1)$ dimensional submanifold without boundary of E^n . That is, $M \subset \tilde{M}$ and the boundary of \tilde{M} , $\partial\tilde{M}$, is empty. For the case that the boundary of M , ∂M , is empty we take $\tilde{M} = M$. The formal definitions of manifolds can be found in Munkres [15].

All neighborhoods are considered open.

DEFINITION 4 (θ^+ and θ^-). Let $x \in M$ and let θ be a sufficiently small neighborhood of x in E^n such that \tilde{M} slices θ into parts θ^+ , θ^- , and $\tilde{M} \cap \theta$. The open set θ^+ is that part of θ belonging to one side of \tilde{M} , and the open set θ^- is that part belonging to the other side of \tilde{M} . In particular, the intersections $\theta^- \cap \theta^+$, $\theta^+ \cap (\tilde{M} \cap \theta)$, and $\theta^- \cap (\tilde{M} \cap \theta)$ are empty and $\theta = \theta^+ \cup (\tilde{M} \cap \theta) \cup \theta^-$.

We say that two sets are *Hausdorff separated* if and only if they are contained in disjoint open sets.

In the next two definitions, let $X \subset E^n$ and let V^* denote a function $V^*: X \rightarrow E^1$.

DEFINITION 5 (Locally discontinuous side). A manifold $M \subset X$ has at $x \in M$ a *locally discontinuous side with respect to V^** if and only if there exists a neighborhood θ of x in E^n such that the set $\{(y, V^*(y)): y \in \tilde{M} \cap \theta \cap X\}$ is Hausdorff separated from either the set $\{(y, V^*(y)): y \in \theta^+ \cap X\}$ or the set $\{(y, V^*(y)): y \in \theta^- \cap X\}$. Moreover, we say that M has at x two locally discontinuous sides with respect to V^* if and only if the set is Hausdorff separated from both the latter two sets.¹ We call θ^+ the *positive side* and θ^- the *negative side* of M at x .

DEFINITION 6 (Locally continuous side). A manifold M has at $x \in M$ a *locally continuous side with respect to V^** if and only if there exists a neighborhood θ of x in E^n such that V^* is continuous on either $(\theta^+ \cap X) \cup (\tilde{M} \cap \theta \cap X)$ or $(\theta^- \cap X) \cup (\tilde{M} \cap \theta \cap X)$. That is, V^* is continuous on either the plus side or the minus side of M at x .

¹ V^* is continuous over $\tilde{M} \cap \theta \cap X$, restrictively.

Let $V^*: X \rightarrow E^1$ be a function and let $B \subset X$. The function V^* when restricted to the subset B is denoted as $V^*|_B$; that is,

$$V^*|_B: B \rightarrow E^1$$

and

$$V^*|_B(x) = V^*(x) \quad \forall x \in B.$$

DEFINITION 7 (*V-type function*). A function $V^*: X \rightarrow E^1$ is called a *V-type* function if and only if there exist a denumerable decomposition $D = \{X_j: j \in J\}$ of X and a denumerable collection of disjoint manifolds $\{M_i: i \in I\} \subset D$ such that

(i) V^* is a piecewise differentiable locally Lipschitzian with respect to D (let $\{(W_j, V_j^*): j \in J\}$ be the associated collection);

(ii) $V^*|_{X \sim \Theta \cup \bigcup_{i \in I} M_i}$ is continuous over $X \sim \Theta \cup \bigcup_{i \in I} M_i$;

(iii) $V^*|_{\bigcup_{i \in I} M_i}$ is continuous over $\bigcup_{i \in I} M_i$;

(iv) $\Theta \cup \bigcup_{i \in I} M_i$ is closed in X ;

(v) $M_i, i \in I$, has at each $x \in M_i$ either a locally continuous side and a locally discontinuous side or two locally discontinuous sides, all with respect to V^* .

The collections $\{X_j: j \in J\}$, $\{M_i: i \in I\}$, and $\{(W_j, V_j^*): j \in J\}$ are said to be *associated with the V-type function V^** .

Note that each manifold $M_i, i \in I$, is by definition a member of the decomposition D . Here, I denotes the index set of the manifolds.

Remark 1. Property (iv) is needed in some of the proofs of the theorems and lemmas contained in this paper. It is included so that the set $X \sim \Theta \cup \bigcup_{i \in I} M_i$ is open in the topology of X .

The next definition gives a condition for the impossibility of a trajectory entering a point on a manifold. This is established in the lemma following it.

DEFINITION 8 (*Nontransversality condition*). A manifold M satisfies at a point $x_0 \in M$ the *nontransversality condition on the positive side (on the negative side)* with respect to the control process Γ if and only if the following condition holds:

NT condition. There exist a manifold \tilde{M} and a neighborhood θ of x_0 in E^n such that, for every $y \in \theta \cap \tilde{M} \cap X$, the inequality

$$(9) \quad N^+(y) \cdot f(x, \nu) \geq 0 \quad \forall x \in \theta^+ \cap X, \quad \forall \nu \in U(x)$$

holds, where $N^+(y)$ is the unit normal to M at y that points into θ^+ . The “negative side” version of this definition replaces θ^+ and $N^+(y)$ with θ^- and $N^-(y)$, the unit normal pointing into θ^- . If $x_0 \in M \sim \partial M$ then \tilde{M} can be replaced with M .

LEMMA 2 (*Nontransversality of trajectory*). *Let x_0 be contained in a manifold M without boundary where $M \subset X$. Suppose the NT condition is satisfied, say, on the positive side of M at x_0 with respect to the control process Γ . Then there exists no control policy $u: [t_1, t_2] \rightarrow E^m$ with corresponding solution $\varphi: [t_1, t_2] \rightarrow \theta^+ \cup (M \cap \theta)$ of (1) satisfying (4) such that*

$$(10) \quad \varphi(t_1) \in \theta^+,$$

$$(11) \quad \varphi(t_2) = x_0.$$

Proof. Suppose, to the contrary, there exists such a control u with corresponding trajectory φ . Let $t_3 \in (t_1, t_2)$ be the first time that $\varphi(t_3) \in M \cap \theta$. Denote $y_0 = \varphi(t_3)$. Thus, $\varphi(t) \in \theta^+$ for all $t \in (t_1, t_3)$. Note that the NT condition is satisfied on the positive side of M at y_0 with respect to the control process Γ . It then suffices to show the impossibility of the control policy $u: [t_1, t_3] \rightarrow E^m$ with corresponding trajectory $\varphi: [t_1, t_3] \rightarrow \theta^+ \cup (M \cap \theta)$ of (1) satisfying (4) and

$$(12) \quad \varphi(t) \in \theta^+ \quad \forall t \in (t_1, t_3),$$

$$(13) \quad \varphi(t_3) = y_0.$$

From (9) it follows that, for all $y \in \theta \cap M \cap X$,

$$(14) \quad N^+(y) \cdot \dot{\varphi}(t) \geq 0 \quad \text{almost everywhere } t \in (t_1, t_3).$$

Integrating (14) over the interval (t_1, t_3) , $t \in (t_1, t_3)$, we have, for all $y \in \theta \cap M \cap X$,

$$(15) \quad N^+(y) \cdot [y_0 - \varphi(t)] \geq 0 \quad \forall t \in (t_1, t_3).$$

Note that, in particular, (15) implies that

$$(16) \quad N^+(y_0) \cdot [y_0 - \varphi(t)] \geq 0 \quad \forall t \in (t_1, t_3).$$

This is a statement that the trajectory φ belongs to the closed (flat) half space on the negative side of the tangent plane to M at y_0 . Denote this half space by $H(y_0, y_0)$.

For $y \in \theta \cap M \cap X$, let $T_y(M)$ represent the tangent plane to M at y . Translate this tangent plane by the vector $y_0 - y$ to the point y_0 and let $H(y, y_0)$ denote the closed (flat) half space on the negative side of the translated tangent plane. Equation (15) is a statement that the trajectory φ belongs to $H(y, y_0)$ for all $y \in \theta \cap M \cap X$.

Define the cone H as

$$H = \bigcap_{y \in \theta \cap M \cap X} H(y, y_0).$$

Thus,

$$(17) \quad \varphi(t) \in H \quad \forall t \in (t_1, t_3).$$

This is a contradiction, since the tip of the cone H (a sufficiently small neighborhood of y_0 intersected with H) is contained in $\theta^- \cup (\theta \cap M)$; see Lemma 3. That is, it lies on the negative side of M at y_0 rather than on the positive side. This completes the proof of the lemma.

LEMMA 3. *The tip of the cone H is contained in $\theta^- \cup (\theta \cap M)$.*

Proof. Let $B(y_0, \delta)$ denote an open ball in E^n having center y_0 and radius $\delta > 0$. For what follows it is convenient to let z represent the vector $(x_1, x_2, \dots, x_{n-1})$ of E^{n-1} . For a sufficiently small ball $B(y_0, \delta)$ contained in $\theta \cap X$ and for a specially selected local coordinate system (x_1, x_2, \dots, x_n) the manifold $M \cap B(y_0, \delta)$ can be described by a C^1 mapping $F: E^{n-1} \rightarrow E^n$ where $F(z) = (z, f(z))$ and where $f: E^{n-1} \rightarrow E^1$ is a C^1 mapping. That is, given (z, x_n) contained in $M \cap B(y_0, \delta)$ we have $x_n = f(z)$.

Define the cone H_1 as

$$H_1 = \bigcap_{y \in B(y_0, \delta) \cap M} H(y, y_0).$$

The cone H_1 contains H since $\theta \cap X$ contains $B(y_0, \delta)$. Consequently it suffices to show that the tip of the cone H_1 belongs to $\theta^- \cup (\theta \cap M)$.

For each $y \in M \cap B(y_0, \delta)$ and for each $z \in E^{n-1}$ let $g(y, z)$ denote the x_n coordinate value at z of the translated tangent plane of $T_y(M)$. This translated tangent plane forms the boundary of $H(y, y_0)$. That is,

$$H(y, y_0) = \{(z, x_n) : z \in E^{n-1}, x_n \leq g(y, z)\}.$$

For each $y \in M \cap B(y_0, \delta)$ and for each $z \in E^{n-1}$ define

$$G(y, z) = \{(z, x_n) : x_n \leq g(y, z)\}.$$

Then

$$H(y, y_0) = \bigcup_{z \in E^{n-1}} G(y, z).$$

Since the sets $G(y, z_1)$ and $G(y, z_2)$ are disjoint for $z_1 \neq z_2$ we can rewrite H_1 as

$$H_1 = \bigcup_{z \in E^{n-1}} \bigcap_{y \in B(y_0, \delta) \cap M} G(y, z).$$

Suppose to the contrary of our assertion that there is some $y^* \triangleq (z^*, x_n^*)$ contained in $\theta^+ \cap H_1 \cap B(y_0, \delta)$. Note that $x_n^* > f(z^*)$.

Let the $(x_1, x_2, \dots, x_{n-1})$ coordinates of y_0 be denoted by z_0 and the x_n coordinate of y_0 by x_n^* . Define the function $h: [0, 1] \rightarrow E^{n-1}$ as

$$h(\alpha) = (1 - \alpha)z_0 + \alpha z^*.$$

Define

$$\beta = \min \left\{ \frac{d(f \circ h)}{d\alpha}(\alpha) : \alpha \in [0, 1] \right\}$$

where the derivative is defined from the right for $\alpha = 0$ and from the left for $\alpha = 1$. Since

$$f(z^*) = f(z_0) + \int_0^1 \frac{d(f \circ h)}{d\alpha}(\alpha) d\alpha$$

we have

$$f(z^*) \geq f(z_0) + \beta.$$

Consider the point y_α of $B(y_0, \delta) \cap M$ where y_α is defined as $(h(\alpha), (f \circ h)(\alpha))$. Since y^* belongs to H_1 we have

$$(z^*, x_n^*) \in \bigcap_{\alpha \in [0, 1]} G(y_\alpha, z^*).$$

By the definition of the function g we have

$$g(y_\alpha, z^*) = f(z_0) + df(h(\alpha)) \cdot (z^* - z_0)$$

and this together with the definition of β gives

$$g(y_\alpha, z^*) \geq f(z_0) + \beta.$$

Therefore,

$$\bigcap_{\alpha \in [0,1]} G(y_\alpha, z^*) = \{(z^*, x_n): x_n \leq f(z_0) + \beta\},$$

which implies that $x_n^* \leq f(z_0) + \beta$. This is a contradiction since $x_n^* > f(z^*)$ and $f(z^*) \geq f(z_0) + \beta$. This completes the proof that there is no y^* belonging to the intersection of $\theta^+ \cap B(y_0, \delta)$ and the tip of the cone H .

DEFINITION 9 (Discontinuous nontransversality condition). A V -type function $V^*: X \rightarrow E^1$ is said to satisfy the *discontinuous nontransversality (DNT) condition* with respect to the control process Γ if and only if, for each $x_0 \in M_i$, $i \in I$, where M_i has at x_0 a locally discontinuous side in the positive direction (in the negative direction) with respect to V^* , the manifold M_i satisfies at x_0 the nontransversality condition on the positive side (on the negative side) with respect to the control process Γ .

LEMMA 4 (Countable compactness of discontinuities). *Let $x_0 \in X$. Let $u \in C(x_0)$, $\varphi \in T(x_0; u)$, and t_f be the terminating time for the trajectory φ . Let $V^*: X \rightarrow E^1$ be a V -type function satisfying the discontinuous nontransversality condition with respect to the control process Γ . Let S denote the set of all discontinuities of the function $V^* \circ \varphi: [t_0, t_f] \rightarrow E^1$. Let $T = \{t_0\} \cup \{t_f\} \cup S$. Then T is a countable compact subset of $[t_0, t_f]$. Moreover, $V^* \circ \varphi$ is continuous from the left except possibly at $t = t_f$.*

Proof. We assert if t is a point of continuity of $V^* \circ \varphi$ then there is a $\delta > 0$ such that $V^* \circ \varphi$ is continuous on $(t - \delta, t + \delta)$. This assertion is obviously true if $\varphi(t)$ does not belong to $\bigcup_{i \in I} M_i$. Suppose $\varphi(t)$ belongs to M_i for some $i \in I$. Assume for all $\delta > 0$ there is some $\tau \in (t - \delta, t + \delta)$ such that $\varphi(\tau)$ belongs to a discontinuous side of M_i at $\varphi(t)$. Consequently, as such τ 's converge to t the $(V^* \circ \varphi)(\tau)$'s converge to $(V^* \circ \varphi)(t)$ from a discontinuous side; recall t is a point of continuity of $V^* \circ \varphi$. The "Hausdorff separated" part of Definition 5 implies that the $(V^* \circ \varphi)(\tau)$'s cannot converge to $(V^* \circ \varphi)(\tau)$ from a discontinuous side. Since the above assumption results in a contradiction, we conclude there is some $\delta > 0$ for which $\varphi(\tau)$, $\tau \in (t - \delta, t + \delta)$, belongs to either M_i or to a continuous side of M_i at $\varphi(t)$, proving our assertion.

It now suffices to show that for each $t \in S \cap (t_0, t_f)$ there exists $\delta > 0$ such that $V^* \circ \varphi$ is continuous on $(t - \delta, t)$. For, in this case we can write

$$[t_0, t_f] = \{t_0\} \cup \{t_f\} \cup \{(t_{\alpha(k)}, t_k): k \in K\}$$

where K indexes the points of $S \cap (t_0, t_f)$ and where $t_{\alpha(k)}$ represents the last discontinuity before t_k ; for $k = 0$ we take $t_{\alpha(0)} = t_0$ if $t_0 \notin S$ and for $k = 1$ we take $t_1 = t_f$ if $t_f \notin S$.

Since

$$\bigcup_{k \in K} (t_{\alpha(k)}, t_k)$$

is an open subset of the real line, it is the countable union of disjoint open intervals. Since the intervals $(t_{\alpha(k)}, t_k)$ are already disjoint it follows that the index K is countable. Thus S and, therefore, T are countable.

Let $t \in (t_0, t_f) \cap S$. That $\varphi(t)$ belongs to $\bigcup_{i \in I} M_i$ follows from the properties (ii) and (iv) of Definition 7. Suppose $\varphi(t) \in M_i$. It follows from Lemma 2 that the trajectory φ cannot enter M_i at $\varphi(t)$ from a discontinuous side since V^* satisfies the DNT condition. If M_i has at $\varphi(t)$ two locally discontinuous sides with respect to V^* , then there is $\delta > 0$ such that $\varphi(t_1) \in M_i$ for all $t_1 \in (t - \delta, t)$. Property (iii) of Definition 7 states that V^* is continuous over M_i . Thus $V^* \circ \varphi$ is continuous over $(t - \delta, t)$.

If M_i has at $\varphi(t)$ only one locally discontinuous side with respect to V^* , then, according to property (v) of Definition 7, M_i has at $\varphi(t)$ a locally continuous side. From Definition 6 it follows that there exists a neighborhood of θ of $\varphi(t)$ in E^n such that V^* is continuous on, say, $(\theta^+ \cap X) \cup (\tilde{M}_i \cap \theta \cap X)$. There exists a $\delta > 0$ such that $\theta(t_1)$ belongs to the latter set for all $t_1 \in (t - \delta, t)$. Consequently, $V^* \circ \varphi$ is continuous over $(t - \delta, t)$. This shows that $V^* \circ \varphi$ is continuous from the left except possibly at $t = t_f$.

4. Sufficiency theorem for discontinuous optimal value function. The sufficiency theorem given below is designed to establish the optimality of a pair (u^*, φ^*) in a control process that may not have a continuous optimal value function. A discontinuous optimal value function arises particularly in state-constrained optimal control processes, e.g., Litt [13].

A function $V^*: X \rightarrow E^1$ is *lower semicontinuous* if and only if, for all $x_0 \in X$,

$$V^*(x_0) \leq \sup \{ \inf \{ V^*(x) : 0 < \|x - x_0\| < r, x \in X \} : r > 0 \}.$$

For a V -type function, the V^* lower semicontinuity, for each $x \in M_i$, $i \in I$, is equivalent to

$$(18) \quad V^*(x) \leq \lim_{x_m \rightarrow x} \{ \inf [V^*(x_k) : k \geq m] \}$$

for all sequences $(x_m) \subset X$ converging to x .

THEOREM 1 (Sufficiency for discontinuous optimal value function). *Let $x_0 \in X$. Let $u^* \in C(x_0)$ and $\varphi^* \in T(x_0; u^*)$. For the optimality in X of the pair (u^*, φ^*) , it is sufficient that there exists a V -type function $V^*: X \rightarrow E^1$ with associated collections $\{X_j : j \in J\}$, $\{M_i : i \in I\}$, and $\{(W_j, V_j^*) : j \in J\}$ such that the following conditions are satisfied:*

(i) V^* satisfies the discontinuous nontransversality condition with respect to the control process Γ ;

(ii) V^* is lower semicontinuous;

(iii) for each $u \in C(x_0)$ and each $\varphi \in T(x_0; u)$ there exists a convergent sequence (t_k) contained in $[t_0, t_f]$ of φ such that

$$(19) \quad g_0[\varphi(t_f)] \geq \lim_{t_k \rightarrow t_f} V^*[\varphi(t_k)].$$

$$(iv) \quad V^*(x_0) = g_0[\varphi^*(t_f^*)] + \int_{t_0}^{t_f^*} f_0[\varphi^*(\tau), u^*(\tau)] d\tau$$

where t_f^* is the terminating time for φ^* ;

(v) for all $x \in X_j$, $\nu \in U(x)$, $j \in J$,

$$(20) \quad f_0(x, \nu) + \text{grad } V_j^*(x) \cdot f(x, \nu) \geq 0.$$

Proof. Let $u \in C(x_0)$, $\varphi \in T(x_0; u)$, and t_f be the terminating time for the trajectory φ . In view of condition (v) we want to show that

$$(21) \quad V^*(x_0) \leq g_0[\varphi(t_f)] + \int_{t_0}^{t_f} f_0[\varphi(\tau), u(\tau)] d\tau.$$

Let

$$g(t) = \int_{t_0}^t f_0[\varphi(\tau), u(\tau)] d\tau + (V^* \circ \varphi)(t)$$

for all $t \in [t_0, t_f - \delta]$, where $\delta > 0$. Condition (v) implies that

$$f_0[\varphi(t), u(t)] + \frac{d}{dt}(V_j^* \circ \varphi)(t) \geq 0 \quad \text{almost everywhere in } T_j, \quad j \in J,$$

where $T_j = \{t \in [t_0, t_f]: \varphi(t) \in X_j\}$. Thus condition (i) of Lemma 1 is satisfied.

As a result of Lemma 4, condition (i) implies that $T = \{t_0\} \cup \{t_f\} \cup S$ is countable compact where S are the discontinuities of $V^* \circ \varphi$. Condition (ii) of Lemma 1 is, therefore, satisfied. Invoking Lemma 4 we have that $V^* \circ \varphi$ is continuous from the left except possibly at $t = t_f$. This satisfies the left inequality of condition (iii) of Lemma 1 for $t \in [t_0, t_f]$. Condition (ii) of Theorem 1 together with condition (ii) of Definition 7, implies that the right inequality of condition (iii) of Lemma 1 is satisfied for $t \in [t_0, t_f]$. With all conditions of Lemma 1 met, it follows that $g: [t_0, t_f - \delta] \rightarrow E^1$ is monotone nondecreasing for all $\delta > 0$.

Consequently, $V^*(x_0) \leq g(t)$ for all $t \in [t_0, t_f]$ and, therefore,

$$(22) \quad V^*(x_0) \leq \int_{t_0}^{t_f} f_0[\varphi(\tau), u(\tau)] d\tau + V^*[\varphi(t_f)].$$

From condition (iii) of Theorem 1, we can write

$$(23) \quad \lim_{t_m \rightarrow t_f} V^*[\varphi(t_m)] \leq g_0[\varphi(t_f)]$$

for some convergent sequence (t_m) .

Inequality (21) follows from (22) and (23). This completes the proof.

5. Control process with discontinuous optimal value function. Consider the bilinear control process with state equations

$$(24a) \quad \dot{x}_1 = -\nu + x_2(1 - \nu), \quad \nu \in [0, 1],$$

$$(24b) \quad \dot{x}_2 = -x_1(1 - \nu)$$

with state space $X = E^2$ and terminal set $\Theta = \{(0, 0)\}$. For a given initial state $x_0 = (x_1^0, x_2^0)$, we desire to minimize the transfer time to the origin:

$$(25) \quad J(x_0, \varphi, u) = \int_{t_0}^{t_f} d\tau.$$

Thus, $f_0 = 1$ and $g_0 = 0$. The set-valued function is given by

$$(26) \quad U(x) = [0, 1] \quad \forall x \in E^2.$$

The control function space Ω is the space of all Lebesgue measurable functions of time defined on bounded intervals with range in E^1 . An admissible control policy $u \in C(x_0)$, $x_0 \in E^2$, with $u: [t_0, t_f] \rightarrow E^1$, satisfies the relation

$$(27) \quad u(t) \in [0, 1] \quad \forall t \in [t_0, t_f]$$

and has a terminating trajectory of $\varphi \in T(x_0; u)$.

Let Γ_0 represent the above control process.

We seek to find a pair (u^*, φ^*) , $u^* \in C(x_0)$, $\varphi^* \in T(x_0; u)$ for each $x_0 \in E^2$ and to discover a V -type function $V^*: E^2 \rightarrow E^1$ with associated collections

$$\{X_j: j \in J\}, \quad \{M_i: i \in I\} \quad \text{and} \quad \{(W_j, V_j^*): j \in J\}$$

such that conditions (i)–(v) of Theorem 1 are satisfied with respect to the control process Γ_0 .

For this purpose we make the following definitions. Let $J = \{1, 2, \dots, 5\}$ and $I = \{1\}$. Define $h: E^2 \rightarrow E^1$ with

$$h(x_1, x_2) = 1 + \frac{x_2}{x_1^2 + x_2^2} - \frac{x_1}{\sqrt{x_1^2 + x_2^2}}$$

for all $(x_1, x_2) \in E^2 \sim \{(0, 0)\}$ and with $h(0, 0) = 0$. Write $p: \{x_2: x_2 < 0 \text{ and } x_2 > -2\} \rightarrow E^1$ such that

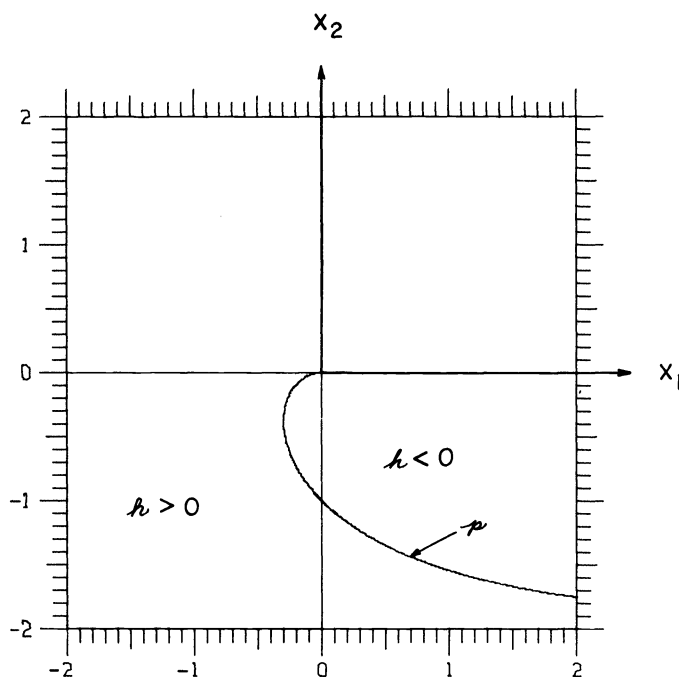
$$(28) \quad h[p(x_2), x_2] = 0 \quad \forall x_2 < 0.$$

The solutions $p(x_2)$ of (28) for $x_2 < 0$ are unique, and p is a smooth function. The function p is plotted in Fig. 1 for $p(x_2) \leq 2$. Define

$$\begin{aligned} X_1 &= \{(x_1, x_2): x_1 > 0, x_2 = 0\}, & X_2 &= \{(x_1, x_2): h(x_1, x_2) > 0\}, \\ X_3 &= \{(x_1, x_2): h(x_1, x_2) = 0, x_2 < 0\}, & X_4 &= \{(x_1, x_2): h(x_1, x_2) < 0\}, \\ X_5 &= \{(0, 0)\}, & M_1 &= X_1. \end{aligned}$$

Let $W_1 = W_5 = E^2$, $W_2 = X_2$, $W_4 = X_4$, and $W_3 = \{(x_1, x_2): x_2 < 0\}$. Define $A: E^2 \sim \{(0, 0)\} \rightarrow E^1$ as the arctangent function such that

$$A(x_1, x_2) = \begin{cases} \tan^{-1}(x_2/x_1), & x_1 > 0, \quad x_2 \geq 0, \\ (\pi/2) + \tan^{-1}(-x_1/x_2), & x_1 \leq 0, \quad x_2 > 0, \\ \pi + \tan^{-1}(-x_2/-x_1), & x_1 < 0, \quad x_2 \leq 0, \\ (3\pi/2) + \tan^{-1}(x_1/-x_2), & x_1 \geq 0, \quad x_2 < 0. \end{cases}$$

FIG. 1. The curve of p

Define V_j^* , $j \in J$, as follows:

$$V_1^*(x_1, x_2) = x_1 \quad \forall (x_1, x_2) \in E^2,$$

$$V_2^*(x_1, x_2) = A(x_1, x_2) + \sqrt{x_1^2 + x_2^2} \quad \forall (x_1, x_2) \in X_2,$$

$$V_3^*(x_1, x_2) = A(x_1, x_2) + \sqrt{x_1^2 + x_2^2} \quad \forall (x_1, x_2) \in W_3,$$

$$V_4^*(x_1, x_2) = x_1 - p(x_2) + V_2^*[p(x_2), x_2] \quad \forall (x_1, x_2) \in X_4,$$

$$V_5^*(x_1, x_2) = 0 \quad \forall (x_1, x_2) \in E^2.$$

We define a V -type function candidate $V^*: E^2 \rightarrow E^1$ as

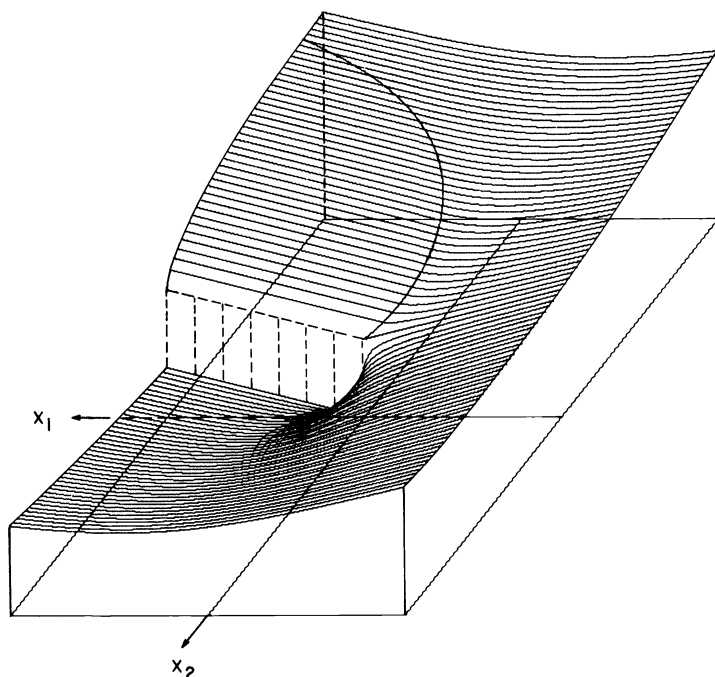
$$V^*(x) = V_j^*(x) \quad \forall x \in X_j, \quad j \in J.$$

This function is plotted in Fig. 2. Note the orientation of the axes. The curve defined by $[p(x_2), x_2, V^*(p(x_2), x_2)]$ is plotted in the surface; this is the intersection of p with the optimal cost surface.

It is easily checked that V^* is piecewise differentiable, locally Lipschitzian with respect to D since the collection $\{(W_j, V_j^*): j \in J\}$ has the necessary properties satisfied; the W_j 's are open and the V_j 's are smooth. Note that the restrictions

$$V^*|_{E^2 \sim (M_1 \cup \Theta)} \quad \text{and} \quad V^*|_{M_1}$$

are continuous. The set $\Theta \cup M_1$ is closed in E^2 . Let $N^+(x) = (0, 1)$ and

FIG. 2. Optimal value function V^*

$N^-(x) = (0, -1)$ for all $x \in M_1$. It follows that M_1 has at each $x \in M_1$ a locally continuous side in the positive direction and a locally discontinuous side in the negative direction. Actually, these two directions have the values of $V^*(x)$ separated by the amount π . This completes the proof that V^* is a V-type function.

V^* satisfies the DNT condition with respect to Γ_0 , provided M_1 satisfies at each $x_0 = (x_1^0, x_2^0) \in M_1$ the NT condition on the negative side with respect to Γ_0 . Since ∂M is empty, we take $\tilde{M}_1 = M_1$. Let θ denote the open cube

$$\left\{ (x_1, x_2): |x_1^0 - x_1| < \frac{x_1^0}{2} \text{ and } |x_2^0 - x_2| < \frac{x_1^0}{2} \right\}.$$

Let $(y_1, y_2) \in \theta \cap M$. Let $(x_1, x_2) \in \theta^- \cap E^2$ and $\nu \in [0, 1]$. Since $N^-(y_1, y_2) = (0, -1)$, the inequality (9) becomes

$$(29) \quad x_1(1 - \nu) \geq 0.$$

This inequality holds since $x_1 > 0$ and $\nu \leq 1$. Consequently, condition (i) of Theorem 1 is satisfied.

Since M_1 has at each $x \in M_1$ a locally continuous side in the positive direction, the inequality (18) is satisfied for all sequences converging to x from the positive side. Let (x_m) be a sequence converging to $x = (x_1, 0)$ from the negative side. In this case we have $V^*(x_m) = V_4^*(x_{1m}, x_{2m})$ and, therefore,

$$V^*(x_m) = x_{1m} - p(x_{2m}) + V_2^*[p(x_{2m}), x_{2m}],$$

with

$$V_2^*[p(x_{2m}), x_{2m}] = \pi + \tan^{-1} \frac{-x_{2m}}{p(x_{2m})} + \sqrt{p(x_{2m})^2 + x_{2m}^2}.$$

Verify that

$$\begin{aligned} p(x_{2m}) &\rightarrow 0 \quad \text{as } x_{2m} \rightarrow 0, \\ \tan^{-1} \frac{-x_{2m}}{p(x_{2m})} &\rightarrow 0 \quad \text{as } x_{2m} \rightarrow 0. \end{aligned}$$

Consequently,

$$(30) \quad \pi + x_1 = \lim_{x_m \rightarrow x} V^*(x_m).$$

Since $V^*(x) = x_1$, we see that condition (ii) is satisfied.

For condition (iii) it suffices to show, for $u \in C(x_0)$, $\varphi \in T(x_0; u)$, that

$$(31) \quad \lim_{t_k \rightarrow t_f} V^*[\varphi(t_k)] = 0$$

for some sequence (t_k) converging to t_f where $\varphi(t_f) = (0, 0)$. It follows from the dynamics (24) that the origin cannot be reached from the second, third, or fourth quadrants. This implies that $\varphi_1(t_k) > 0$ and $\varphi_2(t_k) > 0$ for all such sequences (t_k) . There is a $\delta > 0$ such that $\varphi(t)$ belongs to the first quadrant for all times $t \in [t_f - \delta, t_f]$. From (24) we have $\varphi_1(t)\dot{\varphi}_1(t) + \varphi_2(t)\dot{\varphi}_2(t) = -\varphi_1(t) - \dot{\varphi}_2(t)$ almost everywhere $t \in [t_0, t_f]$. Integrating this equation over $[t_k, t_f]$ gives

$$\varphi_1^2(t_k) + \varphi_2^2(t_k) + 2\varphi_2(t_k) = 2 \int_{t_k}^{t_f} \varphi_1(\tau) d\tau.$$

There exists a sequence (t_k) converging to t_f such that

$$\int_{t_k}^{t_f} \varphi_1(\tau) d\tau \leq \varphi_1(t_k) \int_{t_k}^{t_f} d\tau;$$

consequently, these two expressions imply that

$$(32) \quad \varphi_1(t_k) + \frac{\varphi_2(t_k)}{\varphi_1(t_k)} [\varphi_2(t_k) + 2] \leq 2(t_f - t_k)$$

for all t_k belonging to the sequence (t_k) . This implies that

$$(33) \quad \lim_{t_k \rightarrow t_f} \frac{\varphi_2(t_k)}{\varphi_1(t_k)} = 0,$$

since the left-hand side of (32) is bounded above zero. In the first quadrant, we have

$$V^*[\varphi_1(t_k), \varphi_2(t_k)] \leq \tan^{-1} \left(\frac{\varphi_2(t_k)}{\varphi_1(t_k)} \right) + \sqrt{[\varphi_1(t_k)]^2 + [\varphi_2(t_k)]^2}.$$

This inequality, together with $V^* \geq 0$ and (33), implies (31).

We show that condition (v) is met and then define (u^*, φ^*) and show that (iv) is satisfied.

Let $j = 1$, $x \in X_1$, and $\nu \in [0, 1]$. In this case $x_2 = 0$ and $\text{grad } V_1^*(x) = (1, 0)$. The left-hand side of (20) reduces to $1 - \nu$. Thus (20) is met for $j = 1$.

Let $j = 2$, $x \in X_2$, and $\nu \in [0, 1]$. In this case we have

$$(34a) \quad \frac{\partial V_2^*(x)}{\partial x_1} = \frac{-x_2}{x_1^2 + x_2^2} + \frac{x_1}{\sqrt{x_1^2 + x_2^2}}$$

and

$$(34b) \quad \frac{\partial V_2^*(x)}{\partial x_2} = \frac{x_1}{x_1^2 + x_2^2} + \frac{x_2}{\sqrt{x_1^2 + x_2^2}}.$$

The inequality (20) reduces to

$$(35) \quad \nu h(x_1, x_2) \geq 0,$$

which is satisfied for all $(x_1, x_2) \in X_2$.

Let $j = 3$, $x \in X_3$, and $\nu \in [0, 1]$. The partial derivatives of V_3^* are given by (34) with $x_1 = p(x_2)$. The left-hand side of (20) reduces to the left-hand side of (35), with $h[p(x_2), x_2] = 0$.

Let $j = 4$, $x \in X_4$, and $\nu \in [0, 1]$. Check that

$$\frac{\partial V_4^*}{\partial x_1} = 1$$

and

$$\frac{\partial V_4^*}{\partial x_2} = \frac{p(x_2)}{[p(x_2)]^2 + x_2^2} + \frac{x_2}{\sqrt{[p(x_2)]^2 + x_2^2}}.$$

The left-hand side of (20) reduces to

$$(1 - \nu) \left[1 + x_2 - \frac{x_1 p(x_2)}{[p(x_2)]^2 + x_2^2} - \frac{x_1 x_2}{\sqrt{[p(x_2)]^2 + x_2^2}} \right].$$

After adding and subtracting the term

$$\frac{p(x_2)x_2}{\sqrt{[p(x_2)]^2 + x_2^2}}$$

inside the above brackets and making use of $h[p(x_2), x_2] = 0$, we obtain

$$(36) \quad (1 - \nu)[p(x_2) - x_1] \left[\frac{p(x_2)}{[p(x_2)]^2 + x_2^2} + \frac{x_2}{\sqrt{[p(x_2)]^2 + x_2^2}} \right].$$

For $(x_1, x_2) \in X_4$, we have

$$(37) \quad p(x_2) < x_1.$$

In order for (20) to hold, it suffices then to have

$$(38) \quad \frac{p(x_2)}{\sqrt{[p(x_2)]^2 + x_2^2}} + x_2 < 0.$$

For $x_2 \geq -1$, $p(x_2) \leq 0$, so that (38) holds. For $x_2 < -1$, (38) holds since it is always true that

$$\frac{p(x_2)}{\sqrt{[p(x_2)]^2 + x_2^2}} < 1.$$

For $j = 5$, $x = (0, 0)$, $\nu \in [0, 1]$ the inequality (20) holds since $\text{grad } V_5^*(0, 0) = (0, 0)$. This completes the proof that condition (v) is satisfied.

Define the closed-loop control policy $\sigma: E^2 \rightarrow [0, 1]$ as

$$\sigma(x) = \begin{cases} 1 & \text{if } x \in X_1 \cup X_4, \\ 0 & \text{if } x \in X_2 \cup X_3 \cup X_5. \end{cases}$$

The trajectories resulting from this control are plotted in Fig. 3.

Let $x_0 \in X_1$. Define the pair (u^*, φ^*) as

$$\begin{aligned} u^*(t) &= \begin{cases} 1 & \forall t \in [t_0, t_f], \\ 0, & t = t_f, \end{cases} \\ \varphi_1^*(t) &= x_1^0 - (t - t_0) \quad \forall t \in [t_0, t_f], \\ \varphi_2^*(t) &= 0 \quad \forall t \in [t_0, t_f], \end{aligned}$$

where

$$(39) \quad t_f - t_0 = x_1^0.$$

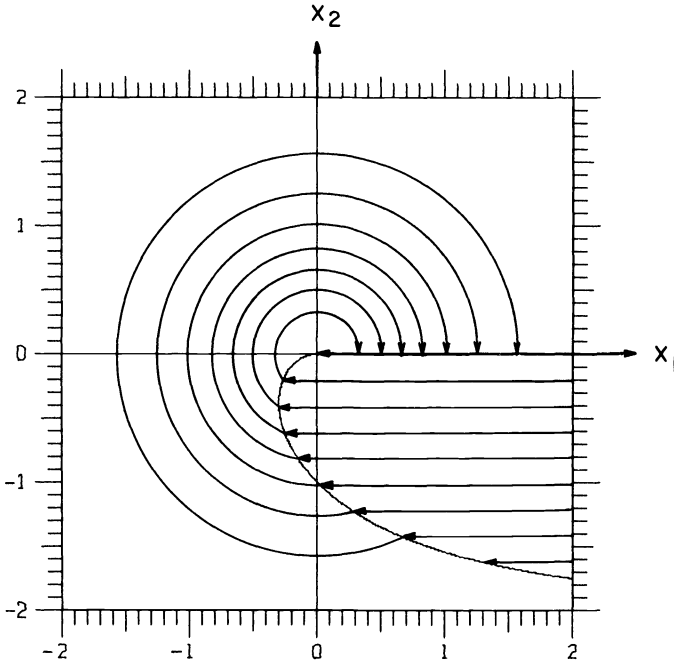


FIG. 3. Optimal trajectories

Let $x_0 \in X_2 \cup X_3$. Define the pair (u^*, φ^*) as

$$\begin{aligned} u^*(t) &= \begin{cases} 0 & \text{if } t \in [t_0, t_1) \cup \{t_f\}, \\ 1 & \text{if } t \in [t_1, t_f], \end{cases} \\ \varphi_1^*(t) &= \begin{cases} \sqrt{(x_1^0)^2 + (x_2^0)^2} \cos(t_1 - t) & \forall t \in [t_0, t_1), \\ \sqrt{(x_1^0)^2 + (x_2^0)^2} - (t - t_1) & \forall t \in [t_1, t_f]; \end{cases} \\ \varphi_2^*(t) &= \begin{cases} \sqrt{(x_1^0)^2 + (x_2^0)^2} \sin(t_1 - t) & \forall t \in [t_0, t_1), \\ 0 & t \in [t_1, t_f], \end{cases} \end{aligned}$$

where $t_1 - t_0 \in [0, 2\pi)$ satisfies

$$\cos(t_1 - t_0) = \frac{x_1^0}{\sqrt{(x_1^0)^2 + (x_2^0)^2}}, \quad \sin(t_1 - t_0) = \frac{x_2^0}{\sqrt{(x_1^0)^2 + (x_2^0)^2}}$$

and where

$$(40) \quad t_f - t_0 = A(x_1^0, x_2^0) + \sqrt{(x_1^0)^2 + (x_2^0)^2}.$$

Let $x_0 \in X_4$. Define (u^*, φ^*) as

$$\begin{aligned} u^*(t) &= \begin{cases} 1 & \text{if } t \in [t_0, t_1) \cup [t_2, t_f], \\ 0 & \text{if } t \in [t_1, t_2) \cup \{t_f\}, \end{cases} \\ \varphi_1^*(t) &= \begin{cases} x_1^0 - (t - t_0) & \forall t \in [t_0, t_1), \\ \sqrt{[p(x_2^0)]^2 + (x_2^0)^2} \cos(t_2 - t) & \forall t \in [t_1, t_2), \\ \sqrt{[p(x_2^0)]^2 + (x_2^0)^2} - (t_1 - t_2) & \forall t \in [t_2, t_f], \end{cases} \\ \varphi_2^*(t) &= \begin{cases} x_2^0 & \forall t \in [t_0, t_1), \\ \sqrt{[p(x_2^0)]^2 + (x_2^0)^2} \sin(t_2 - t) & \forall t \in [t_1, t_2), \\ 0 & \forall t \in [t_2, t_f], \end{cases} \end{aligned}$$

where t_1 satisfies

$$t_1 = t_0 + x_1^0 - p(x_2^0)$$

and $t_2 - t_1 \in [0, 2\pi)$ satisfies

$$\cos(t_2 - t_1) = \frac{p(x_2^0)}{\sqrt{[p(x_2^0)]^2 + (x_2^0)^2}}, \quad \sin(t_2 - t_1) = \frac{x_2^0}{\sqrt{[p(x_2^0)]^2 + (x_2^0)^2}},$$

with

$$(41) \quad t_f - t_0 = x_1^0 - p(x_2^0) + A[p(x_2^0), x_2^0] + \sqrt{[p(x_2^0)]^2 + (x_2^0)^2}.$$

Since X_5 is the terminal set, nothing needs to be undertaken there.

From (39)–(41), the definitions of V^* , u^* , φ^* , it follows that condition (iv) is met. Since all conditions of Theorem 1 are met, the pairs (u^*, φ^*) defined above are optimal. Equation (30) shows that the optimal value function V^* is discontinuous along the positive x_1 axis.

6. Discussion. The purpose of this discussion is to show that the works of Young [16], Armstrong [17] and Hack [18] are only applicable to optimal control problems having continuous optimal value functions. First, the precise work of Hack takes the continuity of the optimal value function as a postulate in his sufficiency theorems. Second, Young [16, p. 281] has established a sufficiency theorem under suitable hypotheses for a concourse of flights that satisfies a strengthened form of Pontryagin maximum principle. It is interesting to note that the example of the previous section provides trajectories, controls and conjugate vector functions (i.e. these are obtained by taking the partial derivatives of V_1^*, \dots, V_4^*) that form a concourse of flights satisfying Young's strengthened form of Pontryagin maximum principle. To be sure, Young's sufficiency theorem [16, p. 281] is based on this fundamental theorem which, in turn, is based on the validity of equation (29.2) of [16]. Hack has shown in Lemma 2.14 of [18, p. 94] that equation (29.2) of [16] holds if the optimal value function is continuous. That equation does not hold for a discontinuous optimal value function; verify this by taking a sufficiently short rectifiable curve having one endpoint at a discontinuity of the optimal value function such that the composition of the rectifiable curve with the optimal value function is discontinuous at that endpoint. The integral on the left hand side of equation (29.2) of [16] goes to zero with the length of the rectifiable curve; the right hand side will not converge to zero due to the discontinuity. Consequently, Young's fundamental theorem is premised on the optimal value function being continuous.

Third, Armstrong [17, p. 652] has shown in Theorem 1 that the continuity of the optimal value function follows from Young's fundamental theorem. In addition, Armstrong has shown under suitable hypotheses in Theorem 3 of [17] that the optimal value function is continuous for problems having weak lines of flight. His proof is based on the assumption that $x_i(t^1)$ converges to $x_0(t^1)$ as i goes to infinity; $x_0(\cdot)$ is a weak line of flight and $x_i(\cdot)$ is a line of flight satisfying Young's strengthened form of Pontryagin maximum principle. The time $-t^1$ is the transfer time from $x_0(t^1)$ to the target, taken along the $x_0(\cdot)$ line of flight. The point $x_i(t^1)$ is that point on the $x_i(\cdot)$ line of flight requiring the same transfer time $-t^1$. For the case that the optimal value function is discontinuous at $x_0(t^1)$ a sequence of flight lines ($x_i(\cdot)$) can be selected so that the sequence ($x_i(t^1)$) converges to some point other than $x_0(t^1)$; that is, the points along $x_i(\cdot)$ that lie in a small neighborhood of $x_0(t^1)$ have associated transfer times that do not belong to a sufficiently small neighborhood of $-t^1$. Consequently the work of Armstrong [17] is not applicable to optimal control problems having discontinuous optimal value functions.

REFERENCES

- [1] R. E. KALMAN, *The theory of optimal control and the calculus of variations*, Mathematical Optimization Techniques, R. Bellmann, ed., RAND Rep. R-396-PR, 1963, pp. 309–331.

- [2] V. G. BOLTYANSKII, *Sufficient conditions for optimality and the justification of the dynamic programming method*, this Journal, 4 (1966) pp. 326–361.
- [3] G. LEITMANN, *Sufficiency theorems for optimal control*, J. Optimization Theory Appl., 2 (1968), p. 285.
- [4] ———, *A Note on a sufficiency theorem for optimal control*, Ibid., 3 (1969), p. 76.
- [5] ———, *Sufficiency for optimal control with state and control constraints*, Internat. J. Non-Linear Mech. 5 (1970), p. 577.
- [6] H. L. STALFORD, *Sufficient conditions for optimal control with state and control constraints*, J. Optimization Theory Appl., 7 (1971), p. 118.
- [7] G. LEITMANN AND H. L. STALFORD, *A sufficiency theorem for optimal control*, J. Optimization Theory Appl., 8 (1971), p. 169.
- [8] V. G. BOLTYANSKII, *Mathematical Methods of Optimal Control*, Holt, Rinehart and Winston, New York, 1971.
- [9] S. E. DREYFUS, *Dynamic Programming and the Calculus of Variations*, Academic Press, New York, 1965.
- [10] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.
- [11] T. L. VINCENT, *Pest Management Programs via Optimal Control Theory*, 13th Joint Automatic Control Conference of the American Automatic Control Council, Stanford University, Stanford, Cal., Aug. 16–18, 1972.
- [12] T. L. VINCENT, E. M. CLIFF AND BEAN-SAN GOH, *Optimal direct control programs for a prey-predator system*, Trans. ASME, 71 (1974), p. 966.
- [13] F. X. LITT, *Some aspects of state constrained optimal control problems*, Ph.D. dissertation, University of California, Berkeley, 1971.
- [14] H. L. STALFORD AND G. LEITMANN, *On integrals of a class of measurable functions*, J. Franklin Inst., 290 (1970), p. 155.
- [15] J. R. MUNKRES, *Elementary Differential Topology*, Annals of Mathematics Studies, no. 54, Princeton University Press, Princeton, NJ, 1963.
- [16] L. C. YOUNG, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.
- [17] G. M. ARMSTRONG, *An extension of an optimal control sufficiency theory*, this Journal, 12 (1974), pp. 650–654.
- [18] T. G. HACK, *Sufficient conditions in the theory of optimal control and differential games*, Ph. D. dissertation, Mathematics Dept., Purdue University, W. Lafayette, IN, 1970.

LINEAR FEEDBACK—AN ALGEBRAIC APPROACH*

M. L. J. HAUTUS† AND MICHAEL HEYMANN‡

Abstract. The algebraic theory of linear input-output maps is reexamined with the objective of accommodating the concept of (state) feedback in this theory. The concepts of extended and restricted linear i/o maps (and linear i/s maps) are introduced and investigated. It is shown how “fraction representations” of transfer matrices arise naturally in this new theoretical framework.

Conditions are given for when the change caused to a linear input-output map by an (open loop) “cascade compensator” can also be accomplished by utilization of (closed loop) state feedback. In particular, it is shown that the change caused to a linear input-output map by cascading (composing) it with an input space isomorphism, can also be effected by feedback, provided the input space isomorphism is “bicausal”, i.e. it does not change the causal structure of the input-output map. Further detailed characterizations of feedback are also given especially in connection with the newly introduced concepts of degree chain and degree list.

1. Introduction. Probably one of the most important contributions to linear systems theory since the introduction of the concepts of controllability and observability, has been the discovery by R. E. Kalman (1965) (see also Kalman (1968) and Kalman et al. (1969, Chap. 10)) that the theory of linear systems can be naturally accommodated in classical module theory. This observation led to a completely satisfactory theory of realization, i.e. the theory that links (external) input-output descriptions with (internal) state space descriptions of systems, the most recent complete discussion of which can be found in Eilenberg (1974, Chap. 16).

Yet, despite the power of the module theoretic approach in attacking the realization problem, there seemed to be no apparent contact between this theory and even some of the most elementary control theoretic questions of linear systems especially insofar as the concept of feedback is concerned.

Two, completely unrelated, approaches were used to study feedback problems: One is the so called “geometric” approach, forwarded and promoted by Wonham and Morse (see e.g. Wonham (1974)), which has been successfully applied to solve such problems as decoupling, regulator design, design of model following systems, and investigating feedback invariant structures, (see. e.g. Wonham and Morse (1970), (1972), Morse and Wonham (1970), (1971), Morse (1973), (1975), Wonham and Pearson (1974) and Wonham (1973)). The second approach, which was widely used and was developed mainly by Rosenbrock (1970) and by Wolovich (see, e.g., Wolovich (1974)), used polynomial matrix techniques for the study of a variety of control theoretic questions. This latter approach, whose primary power derives from the surprising usefulness of fraction

* Received by the editors May 21, 1976, and in revised form January 25, 1977. This work was supported in part by the U.S. Army under Research Grant DAHCO4-76-G-0011 through the Center for Mathematical System Theory, University of Florida, Gainesville, Florida.

† Department of Mathematics, Technological University, Eindhoven, The Netherlands. This work was completed while the author was on leave at the Center for Mathematical System Theory, University of Florida, Gainesville, Florida.

‡ Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa, Israel. This work was completed while the author was on leave at the Center for Mathematical System Theory, University of Florida, Gainesville, Florida.

representations of transfer function matrices, seemed to be especially successful in providing convenient and quite powerful computational algorithms with a capability of yielding various abstract results. Also, fraction representations of transfer matrices provide a convenient vehicle for studying such problems as minimal realization (see e.g. Heymann (1972), Forney (1975) and Fuhrmann (1976)) and feedback invariants as in Heymann (1972).

Probably the most striking paradox in this state of affairs is the fact that, historically, the module theoretic approach to linear systems seemed to support the prevailing viewpoint that transfer function matrices in the form $H(z)/\psi(z)$ ($\psi(z)$ a polynomial) are *the* natural (and theoretically sound) concrete representations of linear input-output maps (see Kalman et al. (1969, Chap. 10)). The representation of transfer function matrices as matrix fractions seemed therefore to be nothing more than a useful technical trick. This discrepancy has been recognized notably by Eckberg (1974) and by Fuhrmann (1976) who attempted to reconcile the two representations within the module theoretic framework. Yet, both of these attempts provide a rather ad-hoc accommodation. Specifically, while both Eckberg and Fuhrmann make a very sound case for viewing transfer matrices as matrix fractions, no successful contact is made with the theory of input-output maps. More importantly, there is no satisfactory contact with feedback theory. While in Fuhrmann (1976) there is no attempt in this direction, in Eckberg (1974) the treatment of feedback is not very successful in that it fails in exhibiting module theory as a powerful or even a convenient framework for dealing with the feedback concept altogether.

The main purpose of the present paper is to reexamine the module theoretic setting of linear input-output maps with the explicit objective of accommodating the concept of "state feedback" within this framework.

In the theory of realization, the concept "canonical" (equivalently reachable and observable) plays a very central and fundamental role in that it defines what is essentially a unique state space. Yet, the property of being a *canonical* realization is *not* invariant under feedback (i.e. a canonical state space can be modified by state feedback to become noncanonical and vice versa). Since the input-output map defines uniquely (or essentially uniquely) only a canonical state space, it is clear why the concept of state feedback somehow seems incompatible with the "classical" module theoretic setting of linear input-output maps. It is easily seen however that reachability *is* invariant under feedback. The importance of the matrix fraction representation for the input-output maps derives from the fact that the representation essentially fixes a reachable state space. Specifically, the representation determines uniquely a reachable realization. As a consequence, the concept of feedback (and especially its effects) can be studied at the level of input-output maps without going through the process of constructing state space descriptions first. Hence, in the study of feedback from an input-output point of view, the process of (concrete) *realization* can essentially be bypassed.

It has been technically well known for quite some time that the modification caused to the input-output map of a linear system by state feedback can be accomplished by cascading the system with a linear device (sometimes called a compensator). Conversely, the effect on a linear input-output map by cascading it with certain types of linear "devices" can sometimes also be accomplished by a

feedback implementation. Yet this converse problem is much less well understood even on a purely “technical” (rather than theoretical) level.

In this paper we give necessary and sufficient conditions (in a module theoretic framework) for a cascade “compensator” to be realizable by a feedback implementation. To formulate these conditions certain revisions in the way an input-output map is viewed are necessary in order to accommodate feedback. We shall adopt here a point of view that was already taken previously in the unpublished notes of Wyman (1972) (see also Sontag (1976)).

The paper is organized as follows: In § 2 the concepts of extended and restricted linear i/o maps are introduced and their relation is investigated (the latter concept coinciding with the standard linear input-output map defined in Kalman et al. (Chap. 10)). In § 3 the concepts of abstract realization and semirealization are discussed and the concept of linear i/s map associated with a (reachable) semirealization is introduced and investigated. Theorems 3.5 and 3.9 give characterizations of linear i/s maps. In § 4 the results of § 3 are specialized to the case where the input and output spaces are finite dimensional and in particular for the case in which the realization or semirealization (i.e. the associated state space) is finite dimensional. Specifically, it is shown that with every reachable realization there is associated a concrete representation of the extended linear i/o map in terms of a (matrix) fraction. This establishes the naturalness of fraction representations and explains their usefulness in the study of feedback. In a similar manner it is shown how linear i/s maps are concretely represented as matrix fractions in association with reachable semirealizations of linear i/o maps. In § 5 the concept of feedback is abstractly introduced and its relation with linear i/s maps is investigated. The notion of a *bicausal* isomorphism (in the extended input space) is defined and investigated. It is seen that every feedback transformation can be implemented in “open loop” through a bicausal isomorphism of the input space. Conversely, conditions are found for the implementability of a bicausal isomorphism of the input space as a feedback transformation (Theorems 5.7 and 5.10). The section is concluded with Theorem 5.13, which states essentially that if the linear i/o map is rational (“rational” being appropriately defined) then every bicausal isomorphism of the input space can be implemented as feedback in some finite dimensional reachable (although not necessarily observable) state space. This result has the intuitive interpretation that the change caused to a linear i/o map by (externally) modifying its input structure can also be accomplished by (internally) implementing feedback provided the external input change does not alter the causal structure of the linear i/o map and is reversible. The paper is concluded in § 6, where the study of feedback is further expanded. In particular, it is noted that feedback can be investigated by studying the structure of the kernel of a (restricted) linear i/s map. Since this kernel is a submodule of the input module, the study of feedback is generalized by studying the structure of an essentially general submodule. In this connection, the concept of the *degree chain* of a submodule is introduced and investigated. The main result of § 6 is Theorem 6.10, which gives a complete characterization of “feedback equivalent” submodules. This reestablishes from a new viewpoint the central role of certain feedback invariants which have been introduced previously. In addition, it is shown how certain previously known (but not well understood) facts find a natural

accommodation within this new theoretical framework (in particular Forney's concept of minimal basis and Wolovich's concept of column properness).

Throughout this paper it will be assumed that the reader is familiar with the now classical module theory of linear systems as can be found for example in Kalman (1968), and Kalman et al. (1969, Chap. 10).

2. Extended and restricted linear i/o maps. We shall begin by introducing some notation. Throughout the paper K will denote a field and U and Y will denote K -linear spaces. The space U will be referred to as the *input value space* and Y as the *output value space* (of an underlying dynamical system Σ). We shall make finite dimensionality assumptions on U and Y only when explicitly stated.

We let \mathbb{Z} denote the set of integers. If S is a K -linear space (in particular U and Y), we consider the set of all sequences $s = (s_t)_{t \in \mathbb{Z}} = (\cdots, s_{-1}, s_0, s_1, \cdots)$ possessing the following properties: (i) $s_t \in S$ for all $t \in \mathbb{Z}$, and (ii) there exists $t_0 \in \mathbb{Z}$ such that $s_t = 0$ for all $t < t_0$. These sequences will be identified with (formal) *S-Laurent series* in z^{-1} , i.e. series of the form

$$(2.1) \quad s = \sum_{t=t_0}^{\infty} s_t z^{-t}.$$

We shall denote the set of S -Laurent series by $S((z^{-1}))$ or alternatively by ΛS . It is then well known that the set $\Lambda K = K((z^{-1}))$ of (scalar) K -Laurent series is a field with convolution as scalar multiplication and the obvious (coefficientwise) addition. Also, with convolution as scalar multiplication and the usual addition, ΛS is a ΛK -linear space. This, in particular, implies that ΛS is also K -linear and also a $K[z]$ -module (where $K[z]$ is the ring of polynomials in z). For an element $s \in \Lambda S$ given by (2.1), the *order* of s is defined by

$$(2.2) \quad \text{ord } s := \begin{cases} \min \{t \in \mathbb{Z} | s_t \neq 0\} & \text{if } s \neq 0, \\ \infty & \text{if } s = 0. \end{cases}$$

Furthermore, for an element $s \in \Lambda S$, multiplication by z results in a shift of the sequence (s_t) to the left, that is

$$(s_t)_{t \in \mathbb{Z}} \mapsto z \cdot (s_t)_{t \in \mathbb{Z}} = (s_{t+1})_{t \in \mathbb{Z}}.$$

We now introduce the *extended input space* ΛU , and the *extended output space* ΛY . If a map $\bar{f}: \Lambda U \rightarrow \Lambda Y$ is K -linear, we say that \bar{f} is *time invariant* provided it satisfies

$$\bar{f}(z \cdot w) = z \bar{f}(w) \quad \text{for all } w \in \Lambda U.$$

The following elementary but important result then follows (see also Wyman (1972)):

THEOREM 2.3. *A K -linear map $\bar{f}: \Lambda U \rightarrow \Lambda Y$ is time invariant if and only if it is ΛK -linear.*

A K -linear map $\bar{f}: \Lambda U \rightarrow \Lambda Y$ is called *causal* if $\text{ord } \bar{f}(w) \geq \text{ord } w$ and *strictly causal* if $\text{ord } \bar{f}(w) > \text{ord } w$ for all $w \in \Lambda U$. We now introduce the following

DEFINITION 2.4. A map $\bar{f}: \Lambda U \rightarrow \Lambda Y$ is called an *extended linear input-output map* (or *extended linear i/o map*) if it is K -linear, strictly causal, and time invariant.

Theorem 2.3 provides an algebraic characterization of time invariance. In order to complete the characterization of extended linear i/o maps we now turn to the question of causality. Let us denote by L the K -linear space of all K -linear maps $U \rightarrow Y$ and consider the space ΛL of all L -Laurent series. This space can be identified with the space of ΛK -linear maps $\Lambda U \rightarrow \Lambda Y$ as follows: We define the K -linear maps

$$\bar{i}_u: U \rightarrow \Lambda U: u \mapsto u \quad (\text{canonical injection}),$$

$$\bar{p}_k: \Lambda Y \rightarrow Y: \sum y_t z^{-t} \mapsto y_k,$$

and for every ΛK -linear map $\bar{f}: \Lambda U \rightarrow \Lambda Y$ and every $k \in \mathbb{Z}$ we define the K -linear map $A_k: U \rightarrow Y$ by

$$(2.5) \quad A_k := A_k(\bar{f}) := \bar{p}_k \circ \bar{f} \circ \bar{i}_u.$$

The L -Laurent series associated with the map \bar{f} is then given by

$$(2.6) \quad Z_{\bar{f}}(z^{-1}) := \sum A_t(\bar{f}) z^{-t}$$

and is called the *impulse response* or the *transfer function* of \bar{f} . Now, if $Z = \sum A_t z^{-t}$ is any element of ΛL we define the action of Z on $w = \sum u_t z^{-t} \in \Lambda U$ by

$$(2.7) \quad Zw := \sum_t \sum_k (A_k u_{t-k}) z^{-t}.$$

If $\bar{f}: \Lambda U \rightarrow \Lambda Y$ is defined as the map whose action is given by (2.7), it is then easily verified that

$$Z_{\bar{f}}(z^{-1}) = Z.$$

We now have the following immediate characterization of causality: *The map $\bar{f}: \Lambda U \rightarrow \Lambda Y$ is causal if and only if $A_k(\bar{f}) = 0$ for $k < 0$ and is strictly causal if and only if $A_k(\bar{f}) = 0$ for $k \leq 0$.* The following is then proved:

THEOREM 2.8. *A map $\bar{f}: \Lambda U \rightarrow \Lambda Y$ is an extended linear i/o map if and only if it is ΛK -linear and $A_k(\bar{f}) = 0$ for all $k \leq 0$.*

For a linear space S we denote by $S[z]$ or alternatively by ΩS the set of all polynomials $\sum_{k=0}^N s_k z^k$ with coefficients s_k in S . Obviously, ΩS is a subset of ΛS and it is easily seen that it is also a K -linear subspace and even a $K[z]$ -submodule of ΛS . It is not, however, a ΛK -linear subspace.

In the development of the module theoretic treatment of linear systems, one of Kalman's primary objectives was to provide a rigorous framework for relating linear input-output maps to state space descriptions, (i.e. the problem of realization). In his treatment Kalman considered the following "experimental" setup (see e.g. Kalman (1968) and Kalman et al. (1969, Chap. 10)): All inputs terminate (i.e. become identically zero) at $t = 0$ and the outputs are observed only for $t \geq 1$. This special set of inputs ΩU will henceforth be called the *restricted input space*, and elements $w \in \Omega U$ will be called *restricted inputs*. Since in this setting outputs are observed only for $t \geq 1$, it follows that two outputs are indistinguishable if their difference is in ΩY . Thus, if we introduce the notation

$$(2.9) \quad \Gamma S := \Lambda S / \Omega S$$

for any K -linear space S , then according to the foregoing we define ΓY to be the *restricted output space*. Let $\Lambda_+ Y$ denote the subset of ΛY consisting of all elements of the form $\sum_{t=1}^{\infty} y_t z^{-t}$. It is readily verified that there is a bijective correspondence between ΓY and $\Lambda_+ Y$ with the property that with every $\gamma \in \Gamma Y$ is associated a unique $\tilde{\gamma} \in \Lambda_+ Y$ satisfying $\pi \tilde{\gamma} = \gamma$ where $\pi: \Lambda Y \rightarrow \Gamma Y$ is the canonical projection. Furthermore, ΓY induces a unique module structure on $\Lambda_+ Y$ by the requirement that the above bijection become a $K[z]$ -module homomorphism. We shall henceforth identify elements of ΓY with those of $\Lambda_+ Y$ and refer to the latter as *restricted outputs*. (It is precisely the module $\Lambda_+ Y$ which was defined in Kalman et al. (1969, Chap. 10) as the output module of a linear system.)

DEFINITION 2.10. A map $f: \Omega U \rightarrow \Gamma Y$ is called a *restricted linear input-output map* (or *restricted linear i/o map*) if it is a $K[z]$ -module homomorphism.

Remark 2.11. In the above definition of restricted linear i/o maps the properties of causality and time invariance are automatically built in. Wyman (1972) called the same a *Kalman linear i/o map* and it is, in fact, precisely the input-output map derived in Kalman's work (see e.g. Kalman et al. (1969, Chap. 10)). We shall later make extensive use both of extended and of restricted linear i/o maps.

In addition to the extended and restricted i/o maps, we introduce the concept of *linear i/o value map*

$$(2.12) \quad f: \Omega U \rightarrow Y$$

as follows: If \tilde{f} is an extended linear i/o map, we define f (associated with it) by

$$(2.13) \quad f(w) := \bar{p}_1 \circ \tilde{f}(w); \quad w \in \Omega U.$$

Alternatively, if \tilde{f} is a restricted linear i/o map, we construct the i/o value map f by

$$(2.14) \quad f := p_1 \circ \tilde{f}$$

where (with ΓY identified with $\Lambda_+ Y$)

$$(2.15) \quad p_1: \Gamma Y \rightarrow Y: \sum_{t=1}^{\infty} y_t z^{-t} \mapsto y_1.$$

Conversely, if $f: \Omega U \rightarrow Y$ is any K -linear map, we can regard it as a linear i/o map by associating with it the maps \tilde{f} and \bar{f} which are constructed from f using the conditions of time invariance and causality. In particular, we have

$$(2.16) \quad \tilde{f}(w) = \sum_{t \geq 0} f(z^t w) z^{-t-1}$$

and

$$(2.17) \quad \bar{f}(w) = \sum_{t \in \mathbb{Z}} f(\mathcal{S}(z^t w)) z^{-t-1}$$

where the *truncation operator* $\mathcal{S}: \Lambda U \rightarrow \Omega U$ is defined by

$$(2.18) \quad \mathcal{S}(\sum u_t z^{-t}) := \sum_{t \leq 0} u_t z^{-t}.$$

(Compare also Kalman and Hautus (1972, formulas (2) and (4)).) It is easily verified that the maps \tilde{f} and \bar{f} as defined in (2.16) and (2.17) are, respectively, a restricted and an extended linear i/o map and that the formulas (2.13) and (2.14) hold. The relation between \bar{f} , \tilde{f} and f is indicated in the following commutative diagram:

$$\begin{array}{ccc}
 \Lambda U & \xrightarrow{\bar{f}} & \Lambda Y \\
 \uparrow i & & \downarrow \pi \\
 \Omega U & \xrightarrow{\tilde{f}} & \Gamma Y \\
 \uparrow i & & \downarrow p_1 \\
 \Omega U & \xrightarrow{f} & Y
 \end{array}
 \begin{array}{c}
 \\
 \\
 \bar{p}_1
 \end{array}$$

where i is the identity map, j is the canonical injection, and π is the canonical projection. The maps f , p_1 and \bar{p}_1 are K -linear, the maps j , π and \tilde{f} are $K[z]$ -homomorphisms and \bar{f} is a ΛK -linear map. The maps \tilde{f} and \bar{f} are called the *restricted and extended i/o maps associated with f* .

3. Linear i/s maps. Let $\tilde{f}: \Omega U \rightarrow \Gamma Y$ be a restricted linear i/o map. By an *abstract realization* of \tilde{f} we refer to a triple (X, g, h) where X is a $K[z]$ -module and $g: \Omega U \rightarrow X$ and $h: X \rightarrow \Gamma Y$ are $K[z]$ -module homomorphisms such that the diagram

$$\begin{array}{ccc}
 \Omega U & \xrightarrow{\tilde{f}} & \Gamma Y \\
 g \searrow & & \nearrow h \\
 & X &
 \end{array}$$

commutes. For the system theoretic interpretation of an abstract realization, the reader is referred to Kalman (1968) and Kalman et al. (1969, Chap. 10). The module X is called the *state space* (and is sometimes regarded only as a K -linear space).

If (X, g, h) is a given abstract realization of a restricted linear i/o map \tilde{f} , one can construct from it a concrete realization of \tilde{f} (see Kalman et al. (1969, Chap. 10)). Such a concrete realization is uniquely determined by the abstract realization (X, g, h) and we shall henceforth call (X, g, h) simply a *realization* of \tilde{f} . In keeping with standard systems terminology we then call a realization *reachable* if g is surjective and *observable* if h is injective. A realization is called *canonical* if it is both reachable and observable.

Consider now a restricted linear i/o map $\tilde{f}: \Omega U \rightarrow \Gamma Y$. Let X be a $K[z]$ -module and let $g: \Omega U \rightarrow X$ be a $K[z]$ -module homomorphism. The pair (X, g) will be called a *semirealization* of \tilde{f} if it can be extended to a realization (X, g, h) with some $K[z]$ -homomorphism $h: X \rightarrow \Gamma Y$. (X, g) will be called a *reachable semirealization* if g is surjective, and *canonical* if every extension realization (X, g, h) of (X, g) is canonical. The following characterization of reachable semirealizations is easily verified:

THEOREM 3.1. *Let $\tilde{f}: \Omega U \rightarrow \Gamma Y$ be a restricted linear i/o map, let X be a $K[z]$ -module, and let $g: \Omega U \rightarrow X$ be a surjective $K[z]$ -homomorphism. Then*

- (i) *(X, g) is a (reachable) semirealization of \tilde{f} if and only if $\ker g \subset \ker \tilde{f}$.*
- (ii) *If (X, g) is a reachable semirealization of \tilde{f} there is a unique h such that (X, g, h) is a realization of \tilde{f} .*
- (iii) *(X, g) is a canonical semirealization of \tilde{f} if and only if $\ker g = \ker \tilde{f}$.*

Consider now a $K[z]$ -homomorphism $g: \Omega U \rightarrow X$ where X is a $K[z]$ -module. If we refer only to the K -linear structure of X and regard g as a K -linear map then, as in (2.16) and (2.17), we can construct the restricted and extended linear i/o maps

$$(3.2) \quad \tilde{g}: \Omega U \rightarrow \Gamma X,$$

$$(3.3) \quad \bar{g}: \Lambda U \rightarrow \Lambda X$$

associated with g . In view of the fact that X is a $K[z]$ -module and not just a K -linear space and g is a $K[z]$ -homomorphism and not just a K -linear map, the maps \tilde{g} and \bar{g} have properties that distinguish them from ordinary i/o maps. For this reason we will call \tilde{g} and \bar{g} , respectively, *a restricted* and *an extended linear i/s map* (where i/s stands for input-state). More generally, a restricted linear i/o map $\tilde{f}: \Omega U \rightarrow \Gamma Y$ and the corresponding extended i/o map $\bar{f}: \Lambda U \rightarrow \Lambda Y$ are called *i/s maps* if Y can be endowed with a $K[z]$ -module structure, compatible with its K -vector space structure, such that the associated i/o value map

$$f := \bar{p}_1 \circ \bar{f} \circ j = p_1 \circ \tilde{f}$$

is a $K[z]$ -module homomorphism. It is easily verified that if \tilde{f} is a restricted linear i/s map and f the associated value map then

$$(3.4) \quad \ker \tilde{f} = \ker f.$$

Indeed, for $w \in \Omega U$, $\tilde{f}(w) = 0$ if and only if $f(z^t w) = 0$ for all $t \geq 0$ (see (2.16)). However, in view of the fact that f is a $K[z]$ -homomorphism, the latter is equivalent to $f(w) = 0$ since $f(z^t w) = z^t f(w)$.

Assume now that $\tilde{f}: \Omega U \rightarrow \Gamma Y$ is a (restricted) linear i/s map and that its associated value map f is surjective. Then (3.4) and Theorem 3.1 imply that (Y, f) is a reachable (in fact even a canonical) semirealization of \tilde{f} . We shall henceforth call a restricted linear i/s map \tilde{f} *reachable* whenever the pair (Y, f) is a reachable semirealization of \tilde{f} or, equivalently, whenever the associated i/o value map f is surjective. (The name i/s was adopted precisely for the reason that in these special i/o maps the output value space qualifies as state space.)

Assume next that $\tilde{f}: \Omega U \rightarrow \Gamma Y$ is a linear i/o map such that its associated i/o value map f is surjective and that (3.4) holds. Then Y is isomorphic as a K -linear space to $\Omega U / \ker \tilde{f}$ which is a $K[z]$ -module. This isomorphism induces a compatible $K[z]$ -module structure on Y and it is easily seen that f is then a $K[z]$ -homomorphism. We summarize the situation in the following

THEOREM 3.5. *Let $\tilde{f}: \Omega U \rightarrow \Gamma Y$ be a restricted linear i/o map such that the associated i/o value map $f = p_1 \circ \tilde{f}: \Omega U \rightarrow Y$ is surjective. Then \tilde{f} is an i/s map if and only if (3.4) holds.*

If $H: X \rightarrow Y$ is a K -linear map then it induces in a natural way a ΛK -linear map

$$(3.6) \quad H: \Lambda X \rightarrow \Lambda Y: \sum x_i z^{-i} \mapsto \sum (Hx_i) z^{-i}$$

and also a $K[z]$ -homomorphism

$$(3.7) \quad H: \Gamma X \rightarrow \Gamma Y: \sum_{i \geq 1} x_i z^{-i} \mapsto \sum_{i \geq 1} (Hx_i) z^{-i}$$

where we have identified $\Lambda_+ X$ with ΓX and $\Lambda_+ Y$ with ΓY . The maps defined by (3.6) and (3.7) are called *static* maps and it will be convenient to denote them by the same symbol H . We now have the following

THEOREM 3.8. *Let $\tilde{f}: \Omega U \rightarrow \Gamma Y$ be a restricted linear i/o map, let $g: \Omega U \rightarrow X$ be a $K[z]$ -homomorphism and let \tilde{g} be the restricted linear i/s map associated with g .*

(i) *If (X, g) is a semirealization of \tilde{f} , then there exists a K -linear map $H: X \rightarrow Y$ such that $\tilde{f} = H \cdot \tilde{g}$ (correspondingly $f = H \cdot g$).*

(ii) *Conversely, if g is surjective and there exists $H: X \rightarrow Y$ such that $\tilde{f} = H \cdot \tilde{g}$, then (X, g) is a semirealization of \tilde{f} .*

Proof. (i) Let $h: X \rightarrow \Gamma Y$ be a $K[z]$ -homomorphism such that (X, g, h) is a realization of \tilde{f} , and define

$$H: X \rightarrow Y: x \mapsto p_1 \circ h(x),$$

where p_1 is defined in (2.15). Then, if $f := p_1 \circ \tilde{f}$, we have

$$f = p_1 \circ \tilde{f} = p_1 \circ h \circ g = H \circ g$$

and consequently for $w \in \Omega U$ we have by (2.16) and (3.7)

$$\begin{aligned} \tilde{f}(w) &= \sum_{i \geq 0} f(z^i w) z^{-i-1} = \sum_{i \geq 0} H \circ g(z^i w) z^{-i-1} \\ &= H \circ \sum_{i \geq 0} g(z^i w) z^{-i-1} = H \cdot \tilde{g}, \end{aligned}$$

and the proof of (i) is complete.

(ii) The existence of H such that $\tilde{f} = H \cdot \tilde{g}$ implies that $\ker g = \ker \tilde{g} \subset \ker \tilde{f}$. Hence, since g is surjective the result follows from Theorem 3.1. \square

The following result gives another characterization of the linear i/s maps amongst all linear i/o maps.

THEOREM 3.9. *Let $\tilde{g}: \Omega U \rightarrow \Gamma Y$ be a restricted linear i/o map. Then \tilde{g} is a reachable restricted linear i/s map if and only if the following condition holds: for every restricted linear i/o map $\tilde{f}: \Omega U \rightarrow \Gamma S$ satisfying $\ker \tilde{g} \subset \ker \tilde{f}$ (where S is a K -linear space), there exists a unique K -linear map $H: Y \rightarrow S$ such that $\tilde{f} = H \cdot \tilde{g}$.*

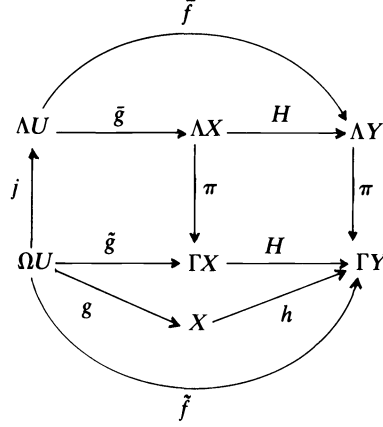
Proof. Assume that \tilde{g} is a reachable restricted linear i/s map and that \tilde{f} satisfies $\ker \tilde{g} \subset \ker \tilde{f}$. Then the existence of H such that $\tilde{f} = H \cdot \tilde{g}$ follows from Theorems 3.1 and 3.8. Also, $f = H \cdot g$ where $f = p_1 \circ \tilde{f}$ and $g = p_1 \circ \tilde{g}$, and from the surjectivity of g (the reachability of \tilde{g}) it follows that H is unique.

Conversely, assume the condition holds. Define $S := \Omega U / \ker \tilde{g}$ and let $f: \Omega U \rightarrow S$ be the canonical projection. Let $\tilde{f}: \Omega U \rightarrow \Gamma S$ be the restricted linear i/o map associated with f . Then \tilde{f} is an i/s map and $\ker \tilde{f} = \ker f = \ker \tilde{g}$. It follows that

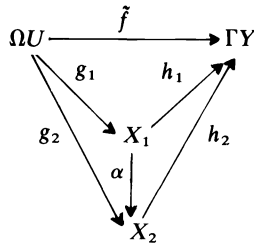
there exists a unique K -linear map $H: Y \rightarrow S$ such that $\tilde{f} = H \cdot \tilde{g}$ or, equivalently, $f = H \cdot g$. Hence $\ker g \subset \ker f = \ker \tilde{g}$ and, since $\ker \tilde{g} \subset \ker g$ is obvious, it follows that $\ker \tilde{g} = \ker g$. From the uniqueness of H we have that g is surjective and the result follows from Theorem 3.5. \square

We conclude this section with the following result (due to Wyman (1972)) which summarizes in a transparent way our current point of view.

THEOREM 3.10. *Let $\tilde{f}: \Omega U \rightarrow \Gamma Y$ be a restricted linear i/o map. Then every reachable realization (X, g, h) of \tilde{f} induces a unique commutative diagram*



4. Finite dimensionality and matrix representations. Let $\tilde{f}: \Omega U \rightarrow \Gamma Y$ be a restricted linear i/o map. Two reachable realizations (X_1, g_1, h_1) and (X_2, g_2, h_2) of \tilde{f} are called *isomorphic* if there exists a $K[z]$ -isomorphism $\alpha: X_1 \rightarrow X_2$ such that the diagram



commutes. Let $\Delta \subset \ker \tilde{f}$ be a submodule. There is associated with Δ (uniquely) a reachable realization $(X_\Delta, g_\Delta, h_\Delta)$ of \tilde{f} where $X_\Delta := \Omega U / \Delta$, $g_\Delta: \Omega U \rightarrow X_\Delta$ is the canonical projection and $h_\Delta: X_\Delta \rightarrow \Gamma Y$ is defined via Theorem 3.1(ii). Suppose now that (X, g, h) is any reachable realization of \tilde{f} and let $\Delta = \ker g$. It is easily verified that (X, g, h) is then isomorphic to $(X_\Delta, g_\Delta, h_\Delta)$. It follows that the class of reachable realizations of \tilde{f} is in one-one correspondence (modulo isomorphisms) with the class of submodules of $\ker \tilde{f}$. Specifically, with every submodule $\Delta \subset \ker \tilde{f}$ is associated a state space $\Omega U / \Delta$, and the structural properties of all reachable

realizations (X, g, h) for which $\Delta = \ker g$ are essentially completely determined by Δ . This fact will be particularly useful in our study of feedback where it will turn out to be easier to deal with (the submodule) Δ than with (the quotient module) $\Omega U/\Delta$.

So far we have imposed no finite dimensionality conditions on either U , X or Y . In this section we shall specialize our results to the finite dimensional case and shall henceforth assume that U and Y are finite dimensional linear spaces. More specifically, we let

$$U = K^m, \quad Y = K^p$$

so that $\Omega U = K^m[z]$ and $\Gamma Y = K^p((z^{-1}))/K^p[z]$.

Thus, in this case ΩU is a free $K[z]$ -module and has a basis of m elements. Since $K[z]$ is a principal ideal domain, every submodule Δ of ΩU has at most m free generators and their number d , which we call rank Δ , is independent of their choice (see e.g. Hartley and Hawkes (1970, Thm. 7.8)).

Our main interest is in the representation of finite dimensional linear systems, and we are concerned with the case in which for a submodule $\Delta \subset \Omega U$, $\Omega U/\Delta$ is a finite dimensional K -linear space, or equivalently, a torsion module (see e.g. Lang (1965, p. 388)). Thus we shall make use of the following standard but important result (the proof of which can be found in Fuhrmann (1976)).

THEOREM 4.1. *Let $\Delta \subset \Omega U$ be a submodule. Then $\Omega U/\Delta$ is a torsion module (or equivalently a finite dimensional K -linear space) if and only if rank $\Delta = m$.*

Let $\Delta \subset \Omega U$ be a submodule of rank m and let d_1, \dots, d_m be a basis for Δ . Define the $K[z]$ -homomorphism $D: \Omega U \rightarrow \Delta$ by $De_i = d_i$, $i = 1, \dots, m$, where e_1, \dots, e_m denotes the natural basis in K^m , (as well as in ΩU). We can view D also as an $m \times m$ polynomial matrix (i.e. a matrix with entries in $K[z]$) by regarding $d_i \in K^m[z]$ as the i th column of D . Conversely, if $D = D(z)$ is an $l \times m$ polynomial matrix, we can regard D as a $K[z]$ -homomorphism $K^m[z] \rightarrow K^l[z]: e_i \mapsto d_i$, $i = 1, \dots, m$, where $d_i = d_i(z) \in K^l[z]$ is the i th column of D . If in particular $l = m$, then d_1, \dots, d_m are elements of $\Omega U = K^m[z]$ and are thus generators of a submodule $\Delta \subset \Omega U$ defined by

$$\Delta = D\Omega U := \{Dw \mid w \in \Omega U\}.$$

Clearly rank $\Delta = \text{rank } D$ (rank D being the matrix rank of D), and rank $\Delta = m$ if and only if D is nonsingular (i.e. $0 \neq \det D \in K[z]$).

Consider now the $K[z]$ -homomorphism $\Omega U \rightarrow \Omega U$ defined by an $m \times m$ nonsingular polynomial matrix D . It is easily verified that there exists a unique ΛK -linear map $\bar{D}: \Lambda U \rightarrow \Lambda U$ such that the diagram

$$\begin{array}{ccc} \Lambda U & \xrightarrow{\bar{D}} & \Lambda U \\ \uparrow j & & \uparrow j \\ \Omega U & \xrightarrow{D} & \Omega U \end{array}$$

commutes, where j , as usual, denotes the canonical injection. It is readily noted that the transfer function of \bar{D} is given by the (polynomial) matrix D . [We shall

refer to a ΛK -linear map whose transfer function is a polynomial matrix as a *polynomial map*.] Since D is nonsingular, it is obviously invertible over ΛK , and thus \bar{D} is an invertible map. We shall henceforth denote both the maps D and \bar{D} and their associated polynomial matrix by the single symbol D . The meaning will always be clear from the context.

An $m \times m$ polynomial matrix R is called *unimodular* if its determinant is a nonzero constant, i.e. if its inverse is also a polynomial matrix. The following theorem whose elementary proof is omitted and the corresponding immediate corollary will be useful:

THEOREM 4.2. *Let $\Delta, \Delta^* \subset \Omega U$ be submodules given by $\Delta = D\Omega U$ and $\Delta^* = D^*\Omega U$ where D and D^* are nonsingular. Then $\Delta^* \subset \Delta$ if and only if there exists a polynomial matrix R such that $D^* = DR$.*

COROLLARY 4.3. *Under the conditions of Theorem 4.2, $\Delta^* = \Delta$ if and only if there exists a unimodular matrix R such that $D^* = DR$.*

We now turn to the study of representations of linear i/o maps (and their realizations). Let $\bar{f}: \Lambda U \rightarrow \Lambda Y$ be an extended linear i/o map, let \tilde{f} be the associated restricted linear i/o map and let (X, g, h) be a reachable realization of \tilde{f} with X a finite dimensional K -vector space (i.e. $\text{rank ker } g = m$). Let $\ker g = \Delta = D\Omega U$ and define the map

$$(4.4) \quad N: \Lambda U \rightarrow \Lambda Y: w \mapsto \bar{f}(Dw).$$

Since N is a composite of two ΛK -linear maps, it is clearly also ΛK -linear. Let

$$Z_N(z) = \sum N_t z^{-t}$$

denote the transfer function of N . We observe that Z_N is a polynomial matrix (i.e. $N_t = 0$ for all $t > 0$). Indeed, for $w \in \Omega U$

$$\pi(Nw) = \pi \circ \bar{f}(Dw) = \tilde{f}(Dw) = 0,$$

with the last equality holding since $Dw \in \ker g \subset \ker \tilde{f}$. It follows that $Nw \in K^p[z]$. Since $w \in \Omega U$ was chosen arbitrarily, we see (compare (2.7)) that

$$Z_N(z) = \sum_{t \leq 0} N_t z^{-t}$$

as claimed. We shall now (just as we have done for D) identify (notationally) Z_N with N so that N will denote both the map and its associated polynomial matrix. In view of the relation $Nw = \bar{f}(Dw)$ for every $w \in \Lambda U$, and the fact that D is invertible we obtain the following expression:

$$\bar{f}(w) = ND^{-1}w \quad (w \in \Lambda U).$$

Thus, the transfer function of \bar{f} has the representation

$$(4.5) \quad Z_{\bar{f}}(z) = \sum_{t=1}^{\infty} A_t z^{-t} = N(z)D^{-1}(z).$$

A ΛK -linear map $\bar{f}: \Lambda U \rightarrow \Lambda Y$ and its L -Laurent series $Z_{\bar{f}}(z)$ are called *rational* if and only if there exists a (nonsingular) polynomial matrix $D(z)$ such

that $Z_{\tilde{f}}D$ is a polynomial matrix. Thus we have proved the necessity part of the following

THEOREM 4.6. *A restricted linear i/o map $\tilde{f}: \Omega U \rightarrow \Gamma Y$ has a reachable realization with finite dimensional state space if and only if \tilde{f} is rational.*

Proof (sufficiency). If \tilde{f} is rational, there exist polynomial matrices N and D such that $\tilde{f} = ND^{-1}$. Then $(X_{\Delta}, g_{\Delta}, h_{\Delta})$, where $\Delta = D\Omega U$, is a reachable realization of \tilde{f} with finite dimensional state space. \square

Two matrices N and D are called *right coprime* if every $m \times m$ polynomial matrix R satisfying $N = N_1R$ and $D = D_1R$ for some polynomial matrices N_1 and D_1 , is necessarily unimodular.

THEOREM 4.7. *Let (X, g, h) be a reachable realization of a restricted linear i/o map \tilde{f} . Let $\ker g = \Delta = D\Omega U$ and assume D is nonsingular. Let $N = Z_{\tilde{f}}D$. Then (X, g, h) is observable (i.e. $\ker \tilde{f} = \ker g$) if and only if N and D are right coprime.*

Proof. Let D_1 be chosen such that $\ker \tilde{f} = D_1\Omega U$. Then $D\Omega U = \ker g \subset \ker \tilde{f} = D_1\Omega U$, and by Theorem 4.2 there exists a polynomial matrix R such that $D = D_1R$. D_1 is then nonsingular and if N_1 is defined such that $N_1D_1^{-1} = Z_{\tilde{f}}$, it is readily seen that $N = N_1R$. If we assume that N and D are right coprime, then the matrix R must be unimodular and by Corollary 4.3 $\ker g = \ker \tilde{f}$ and observability holds. Conversely, let $N = N_1R$ and $D = D_1R$. Then R is nonsingular and $Z_{\tilde{f}} = ND^{-1} = N_1D_1^{-1}$. It follows that $D_1\Omega U \subset \ker \tilde{f}$ and we have

$$\ker g = D\Omega U \subset D_1\Omega U \subset \ker \tilde{f}.$$

If (X, g, h) is observable, then $\ker g = \ker \tilde{f}$ so that $D\Omega U = D_1\Omega U$ and by Corollary 4.3 R is unimodular, whence N and D are right coprime. \square

Remark 4.8. The representation $Z_{\tilde{f}} = ND^{-1}$ of a transfer function is widely known as a *fraction representation*. We have just seen how fraction representations arise naturally from reachable (abstract) realizations (X, g, h) of an input-output map. In fact, there is essentially a one-one correspondence between the class of all finite dimensional reachable realizations of \tilde{f} and the class of fraction representations. It is important to observe that by writing $Z_{\tilde{f}} = ND^{-1}$ we essentially have (modulo isomorphism) a specific reachable realization in mind. The observability along with the associated right coprimeness is essentially of no consequence in our study of feedback as we shall see later.

Let $X = K^n$. Assume X is also endowed with a module structure and let $g: \Omega U \rightarrow X$ be a surjective $K[z]$ -homomorphism. Let $\tilde{g}: \Omega U \rightarrow \Gamma X$ be the restricted (reachable) linear i/s map associated with g . Then (X, g) is a canonical semirealization of \tilde{g} and there exists $h: X \rightarrow \Gamma X$ such that (X, g, h) is a (canonical) realization of \tilde{g} . According to the foregoing we can write

$$(4.9) \quad Z_{\tilde{g}} = SD^{-1},$$

where $Z_{\tilde{g}}$ is the transfer function of \tilde{g} , with D being a (nonsingular) polynomial matrix such that $\ker \tilde{g} = D\Omega U$, and S being an $n \times m$ polynomial map. In view of the observability of (X, g, h) it follows from Theorem 4.7 that D and S are right coprime, and moreover, $\deg \det D = n$.

If $Z_{\tilde{f}}$ is the transfer function of an extended linear i/o map \tilde{f} , then the strict causality of \tilde{f} is equivalent to $Z_{\tilde{f}}$ being a *strictly proper* rational matrix; that is, each entry of $Z_{\tilde{f}}$ is a fraction of polynomials with the degree of the denominator higher

than the degree of the numerator. We now specialize Theorem 3.9 to the finite dimensional case.

THEOREM 4.10. *Let S and D be right coprime polynomial matrices with D nonsingular, and assume that $Z := SD^{-1}$ is strictly proper (or equivalently strictly causal). Then Z is the transfer function of a reachable extended linear i/s map \tilde{g} if and only if for every polynomial matrix N satisfying the condition ND^{-1} is strictly proper, there exists a unique constant matrix H (with entries in K) such that $N = HS$.*

Proof. Assume that $Z = Z_{\tilde{g}}$ where \tilde{g} is a reachable extended linear i/s map. By the coprimeness of S and D we have $\ker \tilde{g} = D\Omega U$ and if \tilde{f}_N is the restricted linear i/o map associated with $Z_N = ND^{-1}$ (ND^{-1} strictly proper), then $\ker \tilde{g} = D\Omega U \subset \ker \tilde{f}_N$. By Theorem 3.9 there exists a unique $H: X \rightarrow Y$ such that $\tilde{f}_N = H \cdot \tilde{g}$ so that if H also denotes the corresponding matrix we have $N = (ND^{-1})D = (HSD^{-1})D = HS$. Conversely, let \tilde{f} be a restricted linear i/o map with $\ker \tilde{g} \subset \ker \tilde{f}$. Then $D\Omega U \subset \ker \tilde{f}$ and there exists a polynomial matrix N such that $\tilde{f} = ND^{-1}$ (\tilde{f} denoting the extended linear i/o map associated with \tilde{f}). By hypothesis there exists a unique constant matrix H such that $N = HS$, and consequently $\tilde{f} = H \cdot \tilde{g}$ and by Theorem 3.9 the proof is complete. \square

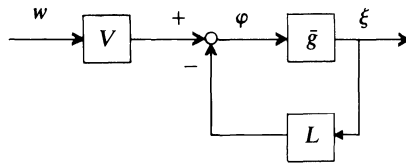
COROLLARY 4.11. *Let D be a nonsingular $m \times m$ polynomial matrix and let $\det D$ have degree n . The set \hat{X} of m -row polynomial vectors $v(z)$ such that $v(z)D^{-1}(z)$ is strictly proper, is an n dimensional vector space over K . If S is a polynomial matrix such that SD^{-1} is strictly proper, then SD^{-1} is the transfer function of a reachable linear i/s map if and only if the rows of S form a basis of \hat{X} .*

Proof. That \hat{X} is a K -vector space is obvious. Suppose SD^{-1} is the transfer function of a linear i/s map. If $v(z) \in \hat{X}$ is any vector, then by Theorem 4.10 there exists a (row) matrix H such that $v = HS$ and thus, the rows of S generate \hat{X} . The uniqueness condition of H implies that the rows of S are linearly independent (over K) and thus form a basis for \hat{X} . Conversely, if the rows of S form a basis for \hat{X} then SD^{-1} is a strictly proper matrix and for each matrix N whose rows are elements of \hat{X} there exists a unique K -matrix H such that $N = HS$ whence, by Theorem 4.10, SD^{-1} is the transfer function of a linear i/s map. Since the number of rows in S is equal to n , it follows that $\dim \hat{X} = n$. \square

5. Feedback. Consider a restricted linear i/o map $\tilde{f}: \Omega U \rightarrow \Gamma Y$, let (X, g) be a reachable semirealization of \tilde{f} and let $\tilde{g}: \Lambda U \rightarrow \Lambda X$ be the extended linear i/s map associated with g . Suppose we modify \tilde{g} in the following way. For each $w \in \Lambda U$ instead of letting \tilde{g} act on w we let \tilde{g} act on $\varphi := V \cdot w - L \cdot \xi$, where $V: U \rightarrow U$ and $L: X \rightarrow U$ are K -linear (static) maps with V invertible, and where ξ is such that $\xi = \tilde{g}(\varphi)$. Then w , φ , and ξ are related by the equations

$$(5.1) \quad \varphi = V \cdot w - L \cdot \xi; \quad \xi = \tilde{g}(\varphi),$$

which schematically can be described by the following *feedback* block diagram:



In line with the above block diagram we call the pair (L, V) a *feedback pair*. The map L will be called a *feedback map* and V will be called a *(static) input transformation*.

From (5.1) we can eliminate ξ , to obtain

$$(5.2) \quad (I + L \cdot \bar{g}) \cdot \varphi = V \cdot w.$$

Since $L \cdot \bar{g}$ is strictly causal, it is easily verified that $\text{ord } (I + L\bar{g})\varphi = \text{ord } \varphi$ for each $\varphi \in \Lambda U$. This implies that $\ker(I + L\bar{g}) = 0$, from which it follows that $I + L\bar{g}$ is invertible. Thus, (5.2) has a unique solution for φ given by

$$\varphi = \bar{l} \cdot w,$$

where

$$(5.3) \quad \bar{l} := \bar{l}_{L,V} := (I + L\bar{g})^{-1}V.$$

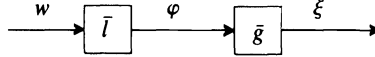
Upon substituting for φ in the second formula of (5.1), we obtain the expression

$$(5.4) \quad \xi = \bar{f}_{L,V}(w),$$

where

$$(5.5) \quad \bar{f}_{L,V} := \bar{g} \circ \bar{l}.$$

The schematic interpretation of (5.5) is given by the block diagram



where $\bar{l}: \Lambda U \rightarrow \Lambda U$ is regarded as a *dynamic input transformation*. It is readily noted (see also Lemma 5.6 below) that the map \bar{l} has the following properties:

- (i) \bar{l} is an invertible ΛK -linear map.
- (ii) Both \bar{l} and \bar{l}^{-1} are causal maps.

We shall henceforth call a map $\bar{l}: \Lambda U \rightarrow \Lambda U$ which satisfies both (i) and (ii) a *bicausal isomorphism* on ΛU .

It is obvious that $\bar{f}_{L,V}$ is an extended linear i/o map. This follows immediately from the fact that the composite of ΛK -linear maps is ΛK -linear, and that the composite of a causal map with a strictly causal one is strictly causal. In fact, it will be seen later that $\bar{f}_{L,V}$ is even a reachable linear i/s map.

We have seen that if we can construct $\bar{f}_{L,V}$ from \bar{g} by feedback (as in our first interpretation) then we can also construct it by cascading \bar{g} with a bicausal isomorphism $\bar{l} = \bar{l}_{L,V}$ (which is an “open loop” construction). We shall now turn to the more difficult question: when can a ΛK -linear map $\bar{l}: \Lambda U \rightarrow \Lambda U$ be expressed as in (5.3) for some L and V . If this is the case we call \bar{l} a *feedback transformation* (corresponding to (L, V)).

LEMMA 5.6. *Let $\bar{l}: \Lambda U \rightarrow \Lambda U$ be a causal ΛK -linear map and let $Z_{\bar{l}}(z^{-1}) = \sum_{k=0}^{\infty} A_k(\bar{l})z^{-k}$ be its transfer function. Then \bar{l} has a causal inverse if and only if $A_0(\bar{l}): U \rightarrow U$ is invertible, in which case $A_0(\bar{l}^{-1}) = (A_0(\bar{l}))^{-1}$.*

The easy proof of Lemma 5.6 is by direct calculation and is omitted.

THEOREM 5.7. *Let $\tilde{g}: \Lambda U \rightarrow \Lambda X$ be a reachable extended linear i/s map. If $\bar{l}: \Lambda U \rightarrow \Lambda U$ is a ΛK -linear map, then there exists a feedback pair (L, V) such that $\bar{l} = \bar{l}_{L,V}$ (as in (5.3)) if and only if*

- (i) \bar{l} is a bicausal isomorphism,
- (ii) for $w \in \Omega U$, $\tilde{g}(w) \in \Omega X$ implies $\bar{l}^{-1}(w) \in \Omega U$ (equivalently, $\bar{l}^{-1}(\ker \tilde{g}) \subset \Omega U$).

Proof. Assume first that $\bar{l} = \bar{l}_{L,V}$. The bicausality of \bar{l} has been noted before so that (i) holds. Also, since $\bar{l}^{-1} = V^{-1} + V^{-1}L\tilde{g}$ it is readily seen that if $w \in \Omega U$, then $\tilde{g}(w) \in \Omega X$ implies that $\bar{l}^{-1}(w) \in \Omega U$ so that (ii) also holds.

Conversely, assume \bar{l} satisfies (i) and (ii). Let $V: U \rightarrow U$ be defined by

$$V := A_0(\bar{l}).$$

By (i) and Lemma 5.6, V is invertible and $V^{-1} = A_0(\bar{l}^{-1})$. Hence, the map $\bar{l}^{-1} - V^{-1}$ (where V^{-1} is regarded as a static map) is strictly causal and

$$(5.8) \quad \tilde{f} := \bar{l}^{-1} - V^{-1}$$

is an extended linear i/o map $\Lambda U \rightarrow \Lambda U$. We claim that $\ker \tilde{g} \subset \ker \tilde{f}$ (where \tilde{f} is the restricted linear i/o map associated with \tilde{f}). Indeed $\tilde{g}(w) = 0$ implies $\tilde{g}(w) \in \Omega X$ and by (ii) $\bar{l}^{-1}(w) \in \Omega U$. By (5.8) it then follows that $\tilde{f}(w) \in \Omega U$ so that $\tilde{f}(w) = 0$. Upon employing Theorem 3.9, we conclude the existence of a K -linear map $H: X \rightarrow U$ such that $\tilde{f} = H \cdot \tilde{g}$ (or equivalently $\bar{F} = H \cdot \bar{g}$). Letting $L: X \rightarrow U$ be defined by $L = VH$ we obtain upon substituting into (5.8)

$$\bar{l}^{-1} = V^{-1} + V^{-1}L\tilde{g}$$

and the proof is complete. \square

COROLLARY 5.9. *Let $\tilde{g}: \Lambda U \rightarrow \Lambda X$ be a reachable extended linear i/s map. Then for every feedback pair (L, V) the map $\tilde{f}_{L,V}$ defined by (5.5) is a reachable extended linear i/s map.*

Proof. We shall prove the corollary by showing that $\tilde{f}_{L,V}$ satisfies Theorem 3.9. Let $\tilde{f}: \Omega U \rightarrow \Gamma S$ be any restricted linear i/o map such that $\ker \tilde{f}_{L,V} \subset \ker \tilde{f}$. Let \bar{f} be the extended linear i/o map associated with \tilde{f} and define $\bar{f}_1: \Lambda U \rightarrow \Lambda S$ by

$$\bar{f}_1: \Lambda U \rightarrow \Lambda S: w \mapsto \bar{f}\bar{l}^{-1}(w).$$

If we can show that $\ker \tilde{g} \subset \ker \bar{f}_1$, then by Theorem 3.9 there exists a unique $H: X \rightarrow S$ such that $\bar{f}_1 = H \cdot \tilde{g}$ (or equivalently $\bar{f}_1 = H \cdot \bar{g}$), whence $\bar{f} = \bar{f}_1 \circ \bar{l} = H \cdot \bar{g} \circ \bar{l} = H \cdot \tilde{f}_{L,V}$ (or equivalently $\tilde{f} = H \cdot \tilde{f}_{L,V}$) showing that $\tilde{f}_{L,V}$ satisfies Theorem 3.9.

Hence, the proof will be complete upon showing that $\ker \tilde{g} \subset \ker \bar{f}_1$. If $w \in \Omega U$ and $\tilde{g}(w) = 0$ then $\tilde{g}(w) \in \Omega X$ and by Theorem 5.7(ii) $w_1 := \bar{l}^{-1}(w) \in \Omega U$. It follows that

$$\tilde{f}_{L,V}(w_1) = \tilde{g} \circ \bar{l}(w_1) = \tilde{g} \circ \bar{l} \circ \bar{l}^{-1}(w) = \tilde{g}(w) \in \Omega X$$

and consequently $\tilde{f}_{L,V}(w_1) = 0$ from which $\tilde{f}(w_1) = 0$ by assumption. Thus, $\bar{f}(w_1) \in \Omega S$ and

$$\bar{f}_1(w) = \bar{f} \circ \bar{l}^{-1}(w) = \bar{f}(w_1) \in \Omega S,$$

so that $\bar{f}_1(w) = 0$ and $\ker \tilde{g} \subset \ker \bar{f}_1$. \square

We shall now specialize our results to the finite dimensional case and assume, just as in § 4, that $U = K^m$ and $Y = K^p$. Moreover, if $\tilde{f}: \Omega U \rightarrow \Gamma Y$ is a restricted linear i/o map, we shall assume that \tilde{f} is rational, i.e. that it has a realization with finite dimensional state space.

First we have the following specialization of Theorem 5.7.

THEOREM 5.10. *Let $\tilde{g}: \Lambda U \rightarrow \Lambda X$ be a reachable extended finite dimensional (i.e. $X = K^n$) linear i/s map, and let $D\Omega U = \ker \tilde{g}$. If $\bar{l}: \Lambda U \rightarrow \Lambda U$ is a ΛK -linear map, there exists a pair (L, V) such that $\bar{l} = \bar{l}_{L,V}$ (see (5.3)) if and only if \bar{l} is a bicausal isomorphism and $Z_{\bar{l}}^{-1}D$ is a polynomial matrix (where $Z_{\bar{l}}$ denotes the transfer function of \bar{l}).*

Remark 5.11. The condition that $Z_{\bar{l}}^{-1}D$ is a polynomial matrix is obvious upon examination of (5.3). In the finite dimensional case we have $Z_{\tilde{g}} = SD^{-1}$ (compare Theorem 4.10), and consequently (5.3) implies that

$$Z_{\bar{l}}^{-1}D = V^{-1}(I + LSD^{-1})D = V^{-1}(D + LS).$$

It is interesting to observe that if \bar{l} is given by $\tilde{g} \circ \bar{l}$ (where $\bar{l} = \bar{l}_{L,V}$), then $Z_{\bar{l}} = S(D + LS)^{-1}V$ and consequently

$$(5.12) \quad \ker \tilde{f}_{\bar{l}} = V^{-1}(D + LS)\Omega U.$$

Alternatively, (5.12) can also be written in the form

$$\ker \tilde{f}_{\bar{l}} = V^{-1}(D + Q)\Omega U,$$

where Q is an arbitrary polynomial matrix such that QD^{-1} is strictly proper (the matrix L being of course uniquely specified through Theorem 4.10). The interesting point in this alternate formulation is the fact that we have eliminated L from explicit consideration.

It is also interesting to make explicit the possible i/o maps (or transfer functions) obtainable by feedback. If $\tilde{f} = ND^{-1}$ where it is understood that the factorization defines a realization of \tilde{f} , then the transfer functions obtainable using feedback are given by

$$\tilde{f}_{L,V} = N(D + Q)^{-1}V,$$

where Q is any polynomial matrix such that QD^{-1} is strictly proper (L being, as before, specified through Theorem 4.10).

The reader should notice the similarity of the situation with the single input single output case.

Let $\tilde{f}: \Omega U \rightarrow \Gamma Y$ be a rational restricted linear i/o map, and let $\bar{l}: \Lambda U \rightarrow \Lambda U$ be a rational causal ΛK -linear map. Let (X, g, h) be a reachable finite dimensional ($X = K^n$) realization of \tilde{f} and let \tilde{g} be the extended linear i/s map associated with g . Theorem 5.10 then states that there exists a feedback pair (L, V) such that $\bar{l} = (I + L\tilde{g})^{-1}V$ provided \bar{l} is a bicausal isomorphism and $Z_{\bar{l}}^{-1}D$ is a polynomial matrix, where $Z_{\bar{l}}$ is the transfer function of \bar{l} , and D is a polynomial matrix satisfying $D\Omega U = \ker \tilde{g}$. While the bicausality of \bar{l} is clearly a necessary condition for its feedback implementation in a given realization, it is not sufficient. We shall conclude this section by showing that there always exists a reachable realization in which \bar{l} can be implemented by feedback if and only if \bar{l} is a rational bicausal isomorphism.

THEOREM 5.13. *Let $\tilde{f}: \Omega U \rightarrow \Gamma Y$ be a rational restricted linear i/o map and let $\bar{l}: \Lambda U \rightarrow \Lambda U$ be a rational causal ΛK -linear map. A necessary and sufficient condition for the existence of a finite dimensional reachable realization (X, g, h) of \tilde{f} such that*

$$Z_{\bar{l}} = (I + LSD^{-1})^{-1}V$$

for some feedback pair (L, V) (where SD^{-1} is the transfer function of the linear i/s map associated with g) is that \bar{l} is a bicausal isomorphism.

Proof. The necessity of the condition is obvious from Theorem 5.10. We will prove sufficiency. If \tilde{f} is rational, then $Z_{\tilde{f}} = ND^{-1}$ for polynomial matrices N and D and $D\Omega U \subset \ker \tilde{f}$. Also if \bar{l} is a bicausal isomorphism, then it can be expressed as $\bar{l} = (I + PQ^{-1})^{-1}V$ (with PQ^{-1} strictly proper). Let $R := \det Q \cdot I$. Then $Z_{\tilde{f}} = NR(DR)^{-1}$ and consider the realization $(X_{\Delta}, g_{\Delta}, h_{\Delta})$ where $\Delta = DR\Omega U$. Clearly $Z_{\bar{l}}^{-1}DR = V^{-1}(I + PQ^{-1})D \cdot \det Q \cdot I$ is a polynomial matrix satisfying the condition of Theorem 5.10. \square

Remark 5.14. Theorem 5.13 gives a complete characterization of those dynamical compensators which can be implemented by “state feedback” in a (possibly unobservable) reachable state-space. This characterization problem received some attention in the literature and the reader is referred to Wolovich (1974).

6. Chains and invariants. Throughout this section we shall assume that $U = K^m$ and $Y = K^p$. Let $\Delta \subset \Omega U$ be a submodule of rank m and let (X_{Δ}, g_{Δ}) be the (finite dimensional) reachable semirealization given by $X_{\Delta} := \Omega U / \Delta$ and $g_{\Delta}: \Omega U \rightarrow X_{\Delta}$ the canonical projection. Let $\tilde{g} = \tilde{g}_{\Delta}$ be the (restricted) linear i/s map associated with g_{Δ} . Then $\ker \tilde{g} = \Delta$ and, as we have seen at the beginning of § 4, all reachable realizations and hence also all reachable linear i/s maps whose kernel is Δ are isomorphic. Thus, modulo state space isomorphism, \tilde{g} is uniquely determined by Δ . If feedback \bar{l} is applied to \tilde{g} , then \tilde{g} is transformed into a new reachable linear i/s map $\tilde{f}_{\bar{l}}$ (see (5.5)) and correspondingly the submodule Δ is transformed into a submodule $\Delta(\bar{l}) := \ker \tilde{f}_{\bar{l}}$ (see (5.12)). Two submodules Δ and Δ' of ΩU of rank m will be called *feedback equivalent* if there exists a feedback \bar{l} such that if $\Delta = \ker \tilde{g}$, then $\Delta' = \ker \tilde{f}_{\bar{l}}$. Clearly feedback equivalence is an equivalence relation and as we have just seen, each feedback equivalence class of submodules characterizes a class of feedback equivalent linear i/s maps. In the present section we study this equivalence relation of submodules.

For $w \in \Omega U$ define the (polynomial) degree of w by $\deg w := -\text{ord } w$. Let Δ and Δ' be submodules of ΩU . A $K[z]$ -homomorphism $q: \Delta \rightarrow \Delta'$ is called *degree preserving* if for each $w \in \Delta$, $\deg q(w) = \deg w$.

LEMMA 6.1. *Two submodules Δ and Δ' of ΩU of rank m are feedback equivalent if and only if there exists a degree preserving $K[z]$ -isomorphism $q: \Delta \rightarrow \Delta'$.*

Proof. If Δ and Δ' are feedback equivalent then $\Delta = \ker \tilde{g}$ and $\Delta' = \ker \tilde{f}_{\bar{l}}$ where $\tilde{f}_{\bar{l}} = \tilde{g} \circ \bar{l}$ for a feedback transformation \bar{l} . Let q denote the restriction of \bar{l}^{-1} to Δ . We see that q maps Δ into Δ' . Indeed, if $w \in \Delta$, then $\tilde{g}(w) \in \Omega X$, and by Theorem 5.7(ii) we also have $v := q(w) = \bar{l}^{-1}(w) \in \Omega U$. Thus $\tilde{f}_{\bar{l}}(v) = \tilde{g} \circ \bar{l} \circ \bar{l}^{-1}(w) = \tilde{g}(w) \in \Omega X$, whence $\tilde{f}_{\bar{l}}(v) = 0$ and $q(w) = v \in \Delta'$. From the bicausality of \bar{l} , it is clear that $\text{ord } \bar{l}^{-1}(w) = \text{ord } w$ for all $w \in \Lambda U$ and consequently the homomorphism $q: \Delta \rightarrow \Delta': w \mapsto \bar{l}^{-1}(w)$ is a degree preserving isomorphism.

Conversely, assume that Δ and Δ' are submodules of ΩU of rank m and that a degree preserving $q: \Delta \rightarrow \Delta'$ exists. Let d_1, \dots, d_m be a basis for Δ and denote $d'_i := q(d_i)$, $i = 1, \dots, m$. Let D and D' be the polynomial matrices $D := [d_1, \dots, d_m]$ and $D' := [d'_1, \dots, d'_m]$. Then surely D is nonsingular and for each $w \in \Delta$ we have $q(w) = D'D^{-1}w$. By formally regarding the matrix $D'D^{-1}$ as a transfer function, we see that q extends uniquely to a ΛK -linear map

$$\bar{q}: \Lambda U \rightarrow \Lambda U; \quad Z_{\bar{q}} := D'D^{-1}.$$

We will complete the proof by showing that \bar{q} is a bicausal isomorphism and that the map $\bar{l} := \bar{q}^{-1}$ is the desired feedback transformation such that $\Delta' = \Delta(\bar{l})$. First note that \bar{q} (and hence also \bar{l}) is a bicausal isomorphism if and only if

$$(6.2) \quad \text{ord } \bar{q}(w) = \text{ord } w$$

for all $w \in \Lambda U$. Recall that (6.2) holds for $w \in \Delta$ by assumption. Since Δ is of rank m (and hence $\Omega U/\Delta$ is a torsion module) there exists a polynomial ψ such that $\psi w \in \Delta$ for all $w \in \Omega U$. Thus $\text{ord } \bar{q}(\psi w) = \text{ord } \psi w$ for all $w \in \Omega U$ and it follows that (6.2) also holds for all $w \in \Omega U$. Next, for $w \in \Lambda U$ let k be an integer such that $w_1 := \mathcal{S}(z^k w) \neq 0$. Then $w_1 \in \Omega U$ and $\text{ord } \bar{q}(w_1) = \text{ord } w_1 = \text{ord } z^k w$. Since for k sufficiently large $\text{ord } \bar{q}(w_1) = \text{ord } \bar{q}(z^k w)$, it follows that (6.2) holds for all $w \in \Lambda U$ and \bar{q} is a bicausal isomorphism. Finally, to see that $\bar{l} = \bar{q}^{-1}$ is the desired feedback transformation we need to show (see Theorem 5.10) that $Z_{\bar{l}}^{-1}D$ is a polynomial matrix. Indeed this is true since $Z_{\bar{l}}^{-1}D = Z_{\bar{q}}^{-1}D = D'D^{-1}D = D'$ and the proof is complete. \square

In order to apply Lemma 6.1 in checking as to whether two submodules Δ and Δ' are feedback equivalent one has to verify the existence (or nonexistence) of a degree preserving homomorphism $q: \Delta \rightarrow \Delta'$. To this end we introduce the concept of *degree chain*.

DEFINITION 6.3. Let Δ be a submodule of ΩU and for each $i = 0, 1, 2, \dots$, let $\Delta_i \subset \Delta$ be the submodule generated by the elements of Δ whose polynomial degree is less than or equal to i . Let $\nu_i := \nu_i(\Delta) := \text{rank } \Delta_i$. The sequence $(\Delta_i)_{i=0}^{\infty}$ of submodules is called the *degree chain* of Δ and the sequent $(\nu_i)_{i=0}^{\infty}$ is called the *degree list*.

It follows immediately that $\Delta_0 \subset \Delta_1 \subset \Delta_2 \subset \dots \subset \Delta$, and $\nu_0 \leq \nu_1 \leq \nu_2 \leq \dots \leq \nu := \text{rank } \Delta$. Since ΩU is Noetherian, it satisfies the ascending chain condition and we have $\Delta_r = \Delta$ for some finite r . Hence

$$\Delta_0 \subset \Delta_1 \subset \dots \subset \Delta_r = \Delta, \quad 0 \leq \nu_0 \leq \nu_1 \leq \dots \leq \nu_r = \nu.$$

We also introduce the following notation. Let $a \neq 0$ be a polynomial or a polynomial vector. We denote by \hat{a} the leading coefficient (vector) of a . That is, if $a = \sum_{i=0}^k a_i z^i$ and $a_k \neq 0$ then $\hat{a} = a_k$. Also, if $a = 0$ we define $\hat{a} := 0$. Recall that if d_1, \dots, d_k are elements of a $K[z]$ -module which generate a submodule M , they are called *free generators of M* , or simply *free* if they are $K[z]$ -linearly independent. We now introduce the following

DEFINITION 6.4. Let d_1, \dots, d_k be elements of a finitely generated $K[z]$ -module (specifically of ΩU). Then d_1, \dots, d_k are called *properly free* if $\hat{d}_1, \dots, \hat{d}_k$ are K -linearly independent.

LEMMA 6.5. *If d_1, \dots, d_k are properly free then they are free.*

Proof. We prove the lemma by negation. If d_1, \dots, d_k are not free, there exist polynomials $\alpha_1, \dots, \alpha_k$, not all zero, such that $\sum_{i=1}^k \alpha_i d_i = 0$. Let $r := \max \deg(\alpha_i d_i)$ and define

$$\varepsilon_i := \begin{cases} 0 & \text{if } \deg(\alpha_i d_i) < r, \\ 1 & \text{if } \deg(\alpha_i d_i) = r. \end{cases}$$

Adding the terms of degree r in $\sum \alpha_i d_i$ yields

$$\sum_{i=1}^k \varepsilon_i \hat{\alpha}_i \hat{d}_i = 0,$$

implying that $\hat{d}_1, \dots, \hat{d}_k$ are K -linearly dependent since not all $\varepsilon_i \hat{\alpha}_i$ are zero. \square

DEFINITION 6.6. Let Δ be a free $K[z]$ -module. A basis d_1, \dots, d_k of Δ is called *proper* if d_1, \dots, d_k are properly free.

The following is an instrumental result.

LEMMA 6.7. *Let Δ be a submodule of ΩU of rank $\nu(>0)$. Then there exists a proper basis d_1, \dots, d_ν of Δ such that*

$$(6.8) \quad \deg d_j = i \quad \text{for } \nu_{i-1} < j \leq \nu_i \text{ and } i = 0, 1, 2, \dots$$

where $(\nu_i)_{i=0}^\infty$ is the degree list of Δ and $\nu_{-1} := 0$.

Proof. First observe that if $k \geq 0$ is the first integer such that $\nu_k \neq 0$, we can choose $0 \neq d_1 \in \Delta_k$ such that $\deg d_1 = k$. Then d_1 is clearly properly free and satisfies (6.8). We proceed stepwise and assume that for $l > 0$, d_1, \dots, d_l are properly free elements of Δ satisfying (6.8) and let S denote the submodule generated by d_1, \dots, d_l . If $l \neq \nu$ (and hence $S \neq \Delta$), then there exists an integer t such that

$$\Delta_{t-1} \subset S \subsetneq \Delta_t.$$

Thus, there exists an element $d_{l+1} \in \Delta_t$ such that $d_{l+1} \notin S$ and $\deg d_{l+1} = t$. Obviously d_1, \dots, d_{l+1} satisfies (6.8) and we complete the proof by showing that this set is also properly free. Assume to the contrary that \hat{d}_{l+1} is a K -linear combination of $\hat{d}_1, \dots, \hat{d}_l$:

$$\hat{d}_{l+1} = \sum_{i=1}^l \lambda_i \hat{d}_i.$$

Consider the element

$$p := d_{l+1} - \sum_{i=1}^l \lambda_i z^{\delta_i} d_i$$

where $\delta_i := t - \deg d_i$, $i = 1, \dots, l$. Obviously $p \in \Delta$ and it is easily verified that $\deg p \leq t - 1$. Thus $p \in \Delta_{t-1} \subset S$ and we conclude that $d_{l+1} \in S$, a contradiction to our assumption. \square

COROLLARY 6.9. *Let $\Delta \subset \Omega U$ be a submodule of rank ν and let $(\Delta_i)_{i=0}^\infty$ and $(\nu_i)_{i=0}^\infty$ be the degree chain and degree list of Δ respectively. There exists a proper basis d_1, \dots, d_ν of Δ such that for each i satisfying $\nu_i \neq 0$ the set d_1, \dots, d_{ν_i} is a basis for Δ_i .*

Proof. The basis constructed in Lemma 6.7 satisfies the desired property. \square

DEFINITION 6.10. Two submodules Δ and Δ' of ΩU are called *chain isomorphic* if there exists a $K[z]$ -isomorphism $q: \Delta \rightarrow \Delta'$ such that $q(\Delta_i) = \Delta'_i$ for $i = 0, 1, 2, \dots$.

The following is the main result of the present section.

THEOREM 6.11. *Let Δ and Δ' be two submodules of ΩU of rank m . Then the following statements are equivalent:*

- (i) Δ and Δ' are feedback equivalent.
- (ii) There exists a degree preserving $K[z]$ -homomorphism $q: \Delta \rightarrow \Delta'$.
- (iii) Δ and Δ' are chain isomorphic.
- (iv) $\nu_i(\Delta) = \nu_i(\Delta')$ for $i = 0, 1, 2, \dots$.

Proof. The equivalence of (i) and (ii) follows from Lemma 6.1. We prove the remaining implications.

(ii) \Rightarrow (iii). If $q: \Delta \rightarrow \Delta'$ is a degree preserving $K[z]$ -homomorphism then $q(\Delta_i) \subset \Delta'_i$. Indeed, if $w \in \Delta_i$, then $w = \sum \alpha_j w_j$ with $\alpha_j \in K[z]$, and $w_j \in \Omega U$ satisfying $\deg w_j \leq i$. Hence $q(w) = \sum \alpha_j q(w_j) \in \Delta'_i$ since $\deg q(w_j) = \deg w_j \leq i$. Similarly we have that $q^{-1}(\Delta'_i) \subset \Delta_i$ and hence $q(\Delta_i) = \Delta'_i$.

(iii) \Rightarrow (iv) is obvious.

(iv) \Rightarrow (ii). Suppose $\nu_i(\Delta) = \nu_i(\Delta') = \nu_i$ for all $i = 0, 1, 2, \dots$. Let d_1, \dots, d_m and d'_1, \dots, d'_m be bases of Δ and Δ' respectively as in Lemma 6.7 (see also Corollary 6.9). Define the $K[z]$ -homomorphism $q: \Delta \rightarrow \Delta'$ by $d_i \mapsto d'_i$ for $i = 1, \dots, m$. It is readily noted that this q is degree preserving. \square

Remark 6.12. For a module $\Delta \subset \Omega U$ of rank m , a sequence of submodules which is essentially equivalent to our degree chain has previously been introduced by Eckberg (1974) in his study of so called canonical matrices. Eckberg's motivation for the introduction of this chain was the construction of a certain unique "canonical" basis for Δ (see also Remark 6.13 below). It is interesting to observe that this chain was completely ignored in Eckberg's study of feedback and he apparently never recognized its feedback invariance properties.

Remark 6.13. As we have already mentioned earlier, fraction representations of the form ND^{-1} were used successfully over the past several years for the study of a variety of technical problems associated with feedback. While it is usually required (unnecessarily) that N and D be right coprime it was also recognized correctly that it is useful to select D in a special form. Specifically, it was first recognized by Wolovich that the matrix D should be taken to be "column proper", i.e. that the sum of the degrees of the columns of D equal the degree of the determinant of D (see Wolovich (1974)). Similar requirements were subsequently made by others (see e.g. Heymann (1972, Chap. 6) and Popov (1970), (1972)). The requirement of column properness has been introduced upon the technical observation that certain "canonical" forms arise naturally provided the matrix D is chosen to be column proper. However, there was no deep understanding of the reasons for this fact since everything was "technique" oriented and motivated. Eckberg (1974) tried to formulate the column properness in his module theoretic study and had moderate success. Forney (1975) developed a theory of *minimal bases* for rational vector spaces for the purpose of giving respectability to what is essentially also nothing else but column properness. The drawback in these approaches is that they depend on defining at the outset (and without any prior motivation) certain "canonical" structures. Yet, it is easily

verified that the column properness property is essentially the same as the property that the columns of D are proper basis for the submodule $\Delta = D\Omega U$. In view of Theorem 6.11 it thus becomes clear that if D is chosen to be column proper, then input isomorphisms are easily tested for the degree preserving property. This explains (in a natural way) the usefulness of this construction for the study of state feedback and it is precisely this important fact, that remained obscure in previous investigations.

REFERENCES

- A. E. ECKBERG, JR. (1974), *A characterization of linear systems via polynomial matrices and module theory*, M.I.T. Electronic Systems Laboratory Rep. ESL-R-528, Mass. Inst. of Tech., Cambridge, MA.
- S. EILENBERG (1974), *Automata, Languages and Machines*, vol. A, Academic Press, New York.
- G. D. FORNEY, JR. (1975), *Minimal bases of rational vector spaces, with applications to multivariable linear systems*, this Journal, 13, pp. 493–520.
- P. FUHRMANN (1976), *Algebraic system theory, an analyst's point of view*, J. Franklin Inst., 301, pp. 521–540.
- R. HARTLEY AND T. O. HAWKES (1970), *Rings, Modules and Linear Algebra*, Chapman and Hall Ltd., London.
- M. HEYMANN (1972), *Structure and realization problems in the theory of dynamical systems*, Lecture notes, International Center for Mechanical Sciences, Udine, Italy; also Springer-Verlag, New York, 1975.
- R. E. KALMAN (1965), *Algebraic structure of linear dynamical systems. I. The module of Σ* , Proc. Nat. Acad. Sci. USA, 54, pp. 1503–1508.
- (1968), *Lectures on controllability and observability*, Lecture notes, Centro Internazionale Matematico Estivo (CIME).
- R. E. KALMAN, P. L. FALB AND M. A. ARBIB (1969), *Topics in Mathematical System Theory*, McGraw-Hill, New York.
- R. E. KALMAN AND M. L. J. HAUTUS (1972), *Realization of continuous-time linear dynamical systems: rigorous theory in the style of Schwartz*, Ordinary differential equations, L. Weiss, ed., Academic Press, New York, pp. 151–164.
- S. LANG (1965), *Algebra*, Addison-Wesley, Reading MA.
- A. S. MORSE (1973), *Structural invariants of linear multivariable systems*, this Journal, 11, pp. 446–465.
- (1975), *System invariance under feedback and cascade control*, Proc. of the International Symposium on Mathematical System Theory, Udine, Italy; Lecture notes in Economics and Mathematical Systems No. 131, Springer-Verlag, New York, pp. 61–74.
- A. S. MORSE AND W. M. WONHAM (1970), *Decoupling and pole assignment by dynamic compensation*, this Journal, 8, pp. 317–337.
- (1971), *Status of noninteracting control*, IEEE Trans. Automatic Control, AC-16, pp. 568–580.
- V. M. POPOV (1970), *Some properties of control systems with irreducible matrix transfer functions*, Seminar of Differential Equations and Dynamical Systems, Lecture Notes in Mathematics no. 144, Springer-Verlag, New York, pp. 250–261.
- (1972), *Invariant description of linear, time-invariant, controllable systems*, this Journal, 10, pp. 252–264.
- H. H. ROSENBROCK (1970), *State Space and Multivariable Theory*, Nelson, London.
- E. D. SONTAG (1976), *On linear systems and noncommutative rings*, Math. Systems Theory, 9, pp. 327–344.
- W. A. WOLOVICH (1974), *Linear Multivariable Systems*, Applied Mathematical Sciences Series no. 11, Springer-Verlag, New York.
- W. M. WONHAM (1973), *Tracking and regulation in linear multivariable systems*, this Journal, 11, pp. 424–437.

- (1974), *Linear Multivariable Control: A Geometric Approach*, Lecture Notes in Economics and Mathematical Systems no. 101, Springer-Verlag, New York.
- W. M. WONHAM AND A. S. MORSE (1970), *Decoupling and pole assignment in linear multivariable systems: A geometric approach*, this Journal, 8, pp. 1–18.
- (1972), *Feedback invariants of linear multivariable systems*, Automatica, 8, pp. 93–100.
- W. M. WONHAM AND J. B. PEARSON (1974), *Regulation and internal stabilization in linear multivariable systems*, this Journal, 12, pp. 5–18.
- B. F. WYMAN (1972), *Linear systems over commutative rings*, Lecture notes, Stanford Univ., Stanford, CA.

CHARACTERIZATIONS OF SOME VARIATIONAL PERTURBATIONS OF THE ABSTRACT LINEAR-QUADRATIC PROBLEM*

T. ZOLEZZI†

Abstract. The coefficients of the plant are perturbed in an abstract linear quadratic problem on a Hilbert space. Many characterizations are obtained about the perturbations that give strongly convergent sequences of optimal controls. Some connections are shown to exist between this continuous dependence problem and some modes of variational convergence for sequences of convex functions, in particular Mosco's convergence. Two applications are given to the classical linear-quadratic problem, for ordinary and elliptic partial differential equations.

Introduction. The continuous dependence of the solutions of an optimization problem from the data is one of the most important features of the given problem. In this paper such a dependence is characterized for the abstract linear-quadratic problem in a Hilbert space setting. We take as fixed the quadratic cost functional, and we consider perturbations acting on the coefficients of the linear plant. In this way a simple theory is obtained with many necessary and sufficient conditions for strong convergence of optimal controls.

In the opinion of this author, the results presented here should be considered as preliminary to a deeper study of variational stability of classes of optimization problems, in that they answer the question for a relatively simple, yet very important, example.

It turns out that such smooth optimization problems inherit many quite strong variational stability properties. Moreover sequences of perturbations of the considered minimum problem generate sequences of strongly convergent optimal solutions if and only if Mosco's convergence obtains, thus showing interesting connections between this theory and the more abstract modes of variational convergence of convex functions.

Two applications to the classical linear quadratic problem for ordinary and elliptic partial differential equations are presented.

The results of this paper can be suitably extended to different or more general and more abstract optimization problems; moreover they are related to well-posedness properties. This will be shown in future works.

Proofs of some results are given here somewhat at length and in a detailed form to show the structure of the problems and to present clearly the possible generalizations.

Some classes of variational perturbations of optimal control problems have been studied in previous work, see [8a, b, c] and [5]. This last paper contains a study of the perturbations of the classical linear quadratic problem for ordinary differential equations, but in a different (non-Hilbert) setting than here.

Notations. The norm of any Banach or Hilbert space in the sequel is denoted by $\|\cdot\|$. In a Hilbert space X the scalar product between x and y is denoted by

* Received by the editors July 21, 1976, and in revised form February 7, 1977.

† Istituto di Matematica, Genova, Italy. This work was supported in part by "Laboratorio per la Matematica Applicata del C.N.R. Genova".

$(x, y)_X$, and simply by (x, y) if no confusion arises. The pairing between any Hilbert space X and its dual X^* is denoted by $\langle \cdot, \cdot \rangle$. Strong convergence is denoted by \rightarrow , weak convergence by \rightharpoonup . If X and Y are Hilbert space, then $X \oplus Y$ denotes their direct sum equipped with the scalar product

$$((x_1, y_1), (x_2, y_2))_{X \oplus Y} = (x_1, x_2)_X + (y_1, y_2)_Y$$

of the pairs $(x_1, y_1), (x_2, y_2)$.

Given a bounded linear operator $L: X \rightarrow Y$, its graph is denoted by

$$\text{gph } L = \{(z, L(z)): z \in X\} \subset X \oplus Y.$$

L^* denotes the Hilbert spaces adjoint of L , so that

$$(Lx, y) = (x, L^*y), \quad x \in X, \quad y \in Y,$$

and L^* acts between Y and X . I denotes the identity operator (on a given Hilbert space). Given a real-valued functional J on the Hilbert space X ,

$$DJ: X \rightarrow X^*$$

denotes its Fréchet differential. A prime denotes the transpose of a matrix. Subsequences are denoted as the original sequence.

Statement of the problem: Assumptions and preliminaries. We are given two real Hilbert spaces X and Y , a positive number c . We consider a given linear bounded operator

$$L_0: X \rightarrow Y,$$

and sequences of linear bounded operators

$$L_n: X \rightarrow Y, \quad n = 1, 2, \dots,$$

such that for any $n = 0, 1, 2, \dots$,

$$(1) \quad \|L_n\| \leq c.$$

Given $u^0 \in X, y^0 \in Y$ we set

$$I_n(u) = \|u - u^0\|^2 + \|L_n(u) - y^0\|^2, \quad u \in X, \quad n = 0, 1, 2, \dots$$

For easy notation no mention will be made of the dependence of I_n on u^0, y^0 .

Let us remark that for any n, u^0, y^0 , there exists exactly one absolute minimum point

$$\bar{u}_n = (I + L_n^* L_n)^{-1}(u^0 + L_n^* y^0), \quad I \text{ the identity operator,}$$

of I_n on the whole space X (depending on u^0, y^0). We write

$$\min I_n = I_n(\bar{u}_n), \quad n = 0, 1, 2, \dots$$

We shall consider the following problem; to characterize all sequences L_n , subject to (1), such that for all u^0, y^0 we have

$$\bar{u}_n \rightarrow \bar{u}_0.$$

We need the following known definitions (in the Hilbert space setting). Let Z be a real Hilbert space, $K_n, n = 0, 1, 2, \dots$, a sequence of closed convex subsets of Z . We say that $K_n \rightarrow K_0$ in the sense of Mosco, and write

$$K_n \xrightarrow{M} K_0$$

if and only if

$$(2) \quad n_i \text{ a subsequence of the positive integers, } x_i \in K_{n_i} \text{ for all } i, \\ x_i \rightharpoonup x_0 \text{ implies } x_0 \in K_0;$$

$$(3) \quad \text{for every } y_0 \in K_0 \text{ there exists a sequence } y_n \in K_n \text{ such that} \\ y_n \rightarrow y_0.$$

Let $J_n: Z \rightarrow (-\infty, +\infty)$ a sequence of functions. We say that $J_n \rightarrow J_0$ in the sense of Mosco and write

$$J_n \xrightarrow{M} J_0$$

if and only if

$$(4) \quad x_n \rightharpoonup x_0 \text{ implies } \liminf J_n(x_n) \geq J_0(x_0);$$

$$(5) \quad \text{for every } y_0 \in Z \text{ there exists a sequence } y_n \rightarrow y_0 \text{ such that} \\ \limsup J_n(y_n) \leq J_0(y_0).$$

This definition is equivalent to that given in [4] (as remarked in [3]), where instead of (4) the following was required:

$$(4') \quad n_i \text{ a subsequence of the positive integers, } z_i \rightharpoonup z_0 \text{ implies} \\ \liminf J_{n_i}(z_i) \geq J(z_0)$$

To show it, let as in (4'), $z_i \rightharpoonup z_0$ and n_i be a subsequence of the positive integers. Set $x_n = z_i$ whenever $n_i \leq n < n_{i+1}$. Then $x_n \rightharpoonup z_0$ and by (4), $\liminf J_{n_i}(z_i) \geq \liminf J_n(x_n) \geq J_0(x_0)$.

Given L_n as above, we consider the closed convex subsets $\text{gph } L_n$ of $X \oplus Y$. Let

$$p_n: X \oplus Y \rightarrow X \oplus Y$$

for any n , be the orthogonal projection over $\text{gph } L_n$. We remark that

$$(\bar{u}_n, L_n \bar{u}_n) = p_n(u^0, y^0) \quad \text{for all } n, u^0, y^0,$$

by the definition of I_n and p_n .

We shall reach the desired results concerning a variational convergence of I_n (in fact Mosco's convergence) by deriving a number of lemmas in the next section.

Results.

LEMMA 1. *The convergence $p_n(z) \rightarrow p_0(z)$ for every $z \in X \oplus Y$ implies*

$$\text{gph } L_n \xrightarrow{M} \text{gph } L_0.$$

Proof. (See [5, prop. 3.2].) Let $y_0 \in \text{gph } L_0$. Then $p_n(y_0) \in \text{gph } L_n$ and $p_n(y_0) \rightarrow p_0(y_0)$ so that (3) is true. Let $x_i \rightarrow x_0$ with $x_i \in \text{gph } L_{n_i}$. The projection $p_{n_i}(x_0)$ is characterized by [7, p. 104]

$$(6) \quad (x_0 - p_{n_i}(x_0), x - p_{n_i}(x_0)) \leq 0 \quad \text{for all } x \in \text{gph } L_{n_i}, \quad \text{any } i.$$

Putting $x = x_i$ in (6) we get

$$(7) \quad (x_0 - p_{n_i}(x_0), x_i - p_{n_i}(x_0)) \leq 0 \quad \text{for all } i.$$

Letting $i \rightarrow \infty$ in (7) we find

$$(x_0 - p_0(x_0), x_0 - p_0(x_0)) \leq 0,$$

that is, $x_0 \in \text{gph } L_0$ and (2) is valid. Q.E.D.

LEMMA 2. $\text{gph } L_n \xrightarrow{M} \text{gph } L_0$ implies $L_n(u) \rightarrow L_0(u)$ for all $u \in X$.

Proof. Mosco's convergence of the graphs of L_n is equivalent to the following (see (2), (3)):

$$(8) \quad \text{for any subsequence } n_i \text{ of the positive integers, } u_i \rightharpoonup u_0, \\ L_{n_i}u_i \rightarrow y_0 \text{ implies } y_0 = L_0u_0.$$

$$(9) \quad \text{for every } u_0 \in X, \text{ there exist } u_n \rightarrow u_0 \text{ such that} \\ L_nu_n \rightarrow L_0u_0.$$

Pick any $u_0 \in X$. Let u_n be as in (9). Then

$$L_n(u_0) = L_n(u_0 - u_n) + L_n(u_n), \quad \text{and} \quad L_n(u_n - u_0) \rightarrow 0$$

by (1), Q.E.D.

THEOREM 1. The following conditions are equivalent:

$$(10) \quad \text{gph } L_n \xrightarrow{M} \text{gph } L_0:$$

$$(11) \quad \text{for all } u \in X, L_nu \rightarrow L_0u; u_n \rightharpoonup u_0 \text{ implies } L_nu_n \rightharpoonup L_0u_0.$$

Proof. Condition (10) implies pointwise convergence of L_n to L_0 as we saw in Lemma 2. If $u_n \rightharpoonup u_0$ in X , then L_nu_n is bounded by (1), and for some subsequence $L_nu_n \rightharpoonup y_0$; therefore by (8), $y_0 = L_0u_0$. Then $L_nu_n \rightharpoonup L_0u_0$ since the same conclusion holds for any weakly convergent subsequence of L_nu_n . Conversely, for any u_0 we can take $u_n = u_0$ to get (9). If $u_i \rightharpoonup u_0$ and $L_nu_i \rightarrow y_0$ then $y_0 = L_0u_0$ by (11) and this gives (8). Q.E.D.

LEMMA 3. For any n, u^0, y^0 we have

$$(12) \quad \bar{u}_n = (I + L_n^* L_n)^{-1} (u^0 + L_n^* y^0).$$

Proof. Fix n, u^0, y^0 . By uniqueness of \bar{u}_n and the easily verified existence of DI_n we see that

$$(13) \quad DI_n(\bar{u}_n) = 0$$

characterizes the minimum point of I_n . For every $u, h \in X$

$$I_n(u + h) - I_n(u) = \|h\|^2 + 2(u - u^0, h) + \|L_n h\|^2 + 2(L_n u - y^0, L_n h).$$

Since

$$\langle DI_n(u), h \rangle = 2(u - u^0, h) + 2(L_n u - y^0, L_n h),$$

we get by (13)

$$(\bar{u}_n - u^0, h) = (y^0 - L_n \bar{u}_n, L_n h);$$

therefore

$$\bar{u}_n + L_n^* L_n \bar{u}_n = u^0 + L_n^* y^0.$$

The conclusion follows from the well-known isomorphic character of $I + L_n^* L_n$. Q.E.D.

LEMMA 4. *The convergences*

$$(14) \quad L_n u \rightarrow L_0 u \quad \text{for all } u, \quad L_n^* x \rightarrow L_0^* x \quad \text{for all } x$$

imply

$$(15) \quad \bar{u}_n \rightarrow \bar{u}_0 \quad \text{for all } u^0, y^0.$$

Proof. We remark that for any n the operator $I + L_n^* L_n$ has a bounded inverse on X . Moreover by (1), for any n and x

$$(16) \quad \|x\|^2 \leq ((I + L_n^* L_n)x, x) \leq \|x\|^2(1 + c^2),$$

so that the norms of both $I + L_n^* L_n$ and $(I + L_n^* L_n)^{-1}$ are uniformly bounded. Since by (1), $\|L_n^*\| \leq c$, we get by (14)

$$L_n^*(L_n u - L_0 u) \rightarrow 0 \quad \text{and} \quad L_n^* L_0 u \rightarrow L_0^* L_0 u \quad \text{for all } u.$$

This implies pointwise convergence of $I + L_n^* L_n$ to $I + L_0^* L_0$, and the same holds for $(I + L_n^* L_n)^{-1}$ because of the uniform bound on their norms. Since

$$L_n^*(y^0) \rightarrow L_0^*(y^0) \quad \text{for all } y^0$$

by (14), the conclusion follows from Lemma 3. Q.E.D.

LEMMA 5. *If*

$$(17) \quad \bar{u}_n \rightarrow \bar{u}_0 \quad \text{for all } u^0, y^0$$

then

$$L_n^* L_n x \rightarrow L_0^* L_0 x \quad \text{for all } x, \quad L_n^* y \rightarrow L_0^* y \quad \text{for all } y.$$

Proof. Since for any u^0, y^0 , by (12)

$$(18) \quad (I + L_n^* L_n)^{-1}(u^0 + L_n^* y^0) \rightarrow (I + L_0^* L_0)^{-1}(u^0 + L_0^* y^0),$$

taking $y^0 = 0$ we get

$$(I + L_n^* L_n)^{-1}(u^0) \rightarrow (I + L_0^* L_0)^{-1}(u^0), \quad u^0 \in X,$$

and by the uniform boundedness of the norms,

$$L_n^* L_n x \rightarrow L_0^* L_0 x, \quad x \in X.$$

Setting $u^0 = 0$ in (18), we find that

$$(I + L_n^* L_n)^{-1} L_n^* y^0 \rightarrow (I + L_0^* L_0)^{-1} L_0^* y^0, \quad y^0 \in Y.$$

By the uniform boundedness of $I + L_n^* L_n$ and their pointwise convergence to $I + L_0^* L_0$ we get

$$L_n^* y^0 \rightarrow L_0^* y^0, \quad y^0 \in Y, \quad \text{Q.E.D.}$$

THEOREM 2. *The following facts are equivalent:*

$$(19) \quad \bar{u}_n \rightarrow \bar{u}_0 \quad \text{for all } u^0, y^0;$$

$$(20) \quad L_n x \rightarrow L_0 x \quad \text{for all } x, \quad L_n^* y \rightarrow L_0^* y \quad \text{for all } y.$$

Proof. From (19) we get the pointwise convergence of the adjoints L_n^* to L_0^* and of $L_n^* L_n$ to $L_0^* L_0$ (Lemma 5). Given $x \in X$, for a subsequence $L_n x \rightarrow \bar{x}$, so that

$$(L_n x, y) = (x, L_n^* y) \rightarrow (\bar{x}, y) = (L_0 x, y), \quad y \in Y,$$

and $L_n x \rightarrow L_0 x$ for the original sequence. Moreover

$$\|L_n x\|^2 = (x, L_n^* L_n x) \rightarrow (L_0 x, L_0 x) = \|L_0 x\|^2.$$

This shows that (19) implies (20). The conclusion follows by Lemma 4. Q.E.D.

THEOREM 3. *Any of the conditions (10), (11), (19), (20) are equivalent to each other, and moreover to the following:*

$$(21) \quad \bar{u}_n \rightarrow \bar{u}_0 \quad \text{and} \quad L_n \bar{u}_n \rightarrow L_0 \bar{u}_0 \quad \text{for all } u^0, y^0.$$

Proof. It suffices to show the equivalence between (20) and (11). Given (20), if $u_n \rightarrow u_0$ then $(L_n u_n, y) = (u_n, L_n^* y) \rightarrow (L_0 u_0, y)$ for any y , so that $L_n u_n \rightarrow L_0 u_0$. The remaining half of (11) has been shown to hold in Theorem 2. Assume conversely that (11) holds. Then for any y, z ,

$$(L_n^* y, z) = (y, L_n z) \rightarrow (L_0^* y, z)$$

so that $L_n^* y \rightarrow L_0^* y$ for all y . Moreover for all x

$$\|L_n^* x\|^2 = (L_n L_n^* x, x) \rightarrow (L_0 L_0^* x, x) = \|L_0 x\|^2$$

and the equivalence is proved. Q.E.D.

Up to this point the strong convergence of the optimal states and controls for all desired pairs u^0, y^0 has been characterized in terms of pointwise convergence of both L_n, L_n^* to L_0, L_0^* respectively. Stated in different words, a condition equivalent to strong convergence of the minimum points turns out to be the pointwise convergence, and collectively weak convergence, of L_n to L_0 ; see (11). This last characterization, more topological in character, would be useful in that it avoids any use of the adjoints L_n^* . We remark that (21) implies the following: for all u^0, y^0 , $\min I_n \rightarrow \min I_0$.

We go further in the analysis of the variational convergence of I_n . In particular we wish to relate the above conditions (11), (19), (20), to the convergence of the values $\min I_n$ for all u^0, y^0 , and to Mosco's convergence of I_n to I_0 .

THEOREM 4. *The following conditions are equivalent: for all u^0, y^0 ,*

$$(22) \quad I_n \xrightarrow{M} I_0;$$

$$(23) \quad \bar{u}_n \rightarrow \bar{u}_0.$$

Proof. Assuming (23), consider any sequence $u_n \rightarrow u_0$ in X . Then $L_n u_n \rightarrow L_0 u_0$ by Theorem 3; therefore by weak sequential lower semi-continuity of the norm

$$\liminf I_n(u_n) \geq I_0(u_0).$$

This meets condition (4), while (5) is verified since given u_0 we take $v_n = u_0$ constantly so that, by pointwise convergence of L_n (Theorem 2) we get $I_n(v_n) \rightarrow I_0(u_0)$. Assume conversely (22). Having fixed u^0, y^0 , we remark, by (12) and (1), that

$$\sup \|\bar{u}_n\| < +\infty.$$

Taking some subsequence we have $\bar{u}_n \rightharpoonup u_0$. Any subsequence of I_n converges to I_0 (in the sense of Mosco); see [4, p. 521]. Therefore along the above subsequence

$$I_0(u_0) \leq \liminf I_n(\bar{u}_n).$$

Pick any $u \in X$. Consider $v_n \rightarrow u$ such that

$$\limsup I_n(v_n) \leq I_0(u).$$

Then

$$I_0(u_0) \leq \liminf I_n(\bar{u}_n) \leq \limsup I_n(v_n) \leq I_0(u)$$

and by uniqueness of \bar{u}_0 it follows $\bar{u}_0 = u_0$. Since any weakly convergent subsequence of \bar{u}_n behaves as above, we conclude that for the original sequence

$$(24) \quad \bar{u}_n \rightarrow \bar{u}_0.$$

Moreover given $w_n \rightarrow \bar{u}_0$ such that $\limsup I_n(w_n) \leq I_0(\bar{u}_0)$, we get $I_0(\bar{u}_0) \leq \liminf I_n(\bar{u}_n) \leq \limsup I_n(\bar{u}_n) \leq \limsup I_n(w_n) \leq I_0(\bar{u}_0)$, that is,

$$\min I_n \rightarrow \min I_0.$$

Now consider

$$(25) \quad \begin{aligned} \langle DI_n(u), u - \bar{u}_n \rangle &= 2(u - u^0, u - \bar{u}_n) + 2(L_n^*(L_n u - y^0), u - \bar{u}_n) \\ &= 2((I + L_n L_n^*)(u - \bar{u}_n), u - \bar{u}_n) \geq 2\|u - \bar{u}_n\|^2, \end{aligned}$$

for any n .

We use, in (25), Lemma 3 and a formula in its proof. (In passing, we remark that (25) establishes the following noteworthy property of our variational problem. Such problems are equi-well set in the sense of [8e], where a more general class of functionals is considered. In this case (25) relies on the uniform strong convexity of the sequence I_n .)

For every n , by (25) and convexity,

$$I_n(\bar{u}_0) - I_n\left(\frac{\bar{u}_n + \bar{u}_0}{2}\right) \geq \frac{1}{2} \left\langle DI_n\left(\frac{\bar{u}_n + \bar{u}_0}{2}\right), \bar{u}_0 - \bar{u}_n \right\rangle \geq \frac{1}{2} \|\bar{u}_0 - \bar{u}_n\|^2.$$

It follows that

$$(26) \quad \limsup \|\bar{u}_n - \bar{u}_0\|^2 \leq 2 \limsup \left(I_n(\bar{u}_0) - I_n\left(\frac{\bar{u}_n + \bar{u}_0}{2}\right) \right).$$

Let us show that

$$(27) \quad I_n(\bar{u}_0) \rightarrow I_0(\bar{u}_0).$$

Consider $v_n \rightarrow \bar{u}_0$ such that $\limsup I_n(v_n) \leq I_0(\bar{u}_0)$. By writing

$$I_0(\bar{u}_0) \leq \liminf I_n(v_n) \leq \limsup I_n(v_n) \leq I_0(\bar{u}_0)$$

we see that

$$(28) \quad I_n(v_n) \rightarrow I_0(\bar{u}_0).$$

Given x, z , an easy computation gives

$$\begin{aligned} I_n(x) - I_n(z) &= \|x\|^2 - \|z\|^2 - 2(x - z, u^0) + \|L_n x\|^2 - \|L_n z\|^2 - 2(L_n(x - z), y^0) \\ &\leq c_1 \|x - z\|^2 + c_2 \|x - z\|(\|z\| + \|u^0\| + \|y^0\|) \end{aligned}$$

for some constants c_1, c_2 (depending on c only). Therefore I_n is a sequence of equilocally Lipschitzian functions. In other words, given u^0, y^0 and a number $r > 0$ there exists a number $a > 0$ such that

$$|I_n(x) - I_n(z)| \leq a \|x - z\| \quad \text{if } \|x\| \leq r, \quad \|z\| \leq r,$$

and $\|x - z\| \leq 1$. This implies $I_n(v_n) - I_n(\bar{u}_0) \rightarrow 0$. By (28), condition (27) is proved. Thus by (26)

$$\limsup \|\bar{u}_n - \bar{u}_0\|^2 \leq I_0(\bar{u}_0) - \liminf I_n\left(\frac{\bar{u}_n + \bar{u}_0}{2}\right) \leq 0$$

because of (4). This shows $\bar{u}_n \rightarrow \bar{u}_0$. Q.E.D.

Very often (especially in the applications) no exact minimization is performed on I_n , only an approximate one. The following definition will be useful in this connection. Fix u^0, y^0 . A sequence $u_n \in X$ is called *asymptotically minimizing* for the sequence I_n if and only if

$$I_n(u_n) - \min I_n \rightarrow 0.$$

THEOREM 5. *The following are equivalent facts:*

(29) *for all u^0, y^0 , any asymptotically minimizing sequence converges strongly to \bar{u}_0 ;*

(30) *any of the equivalent conditions in Theorems 1, 2, 3, 4 hold.*

Proof. Assume (30). Having fixed u^0, y^0 , let u_n be an asymptotically minimizing sequence. Then for all large n

$$\|u_n - u^0\|^2 \leq I_n(u_n) \leq 1 + \min I_n \leq 1 + \|u^0\|^2 + \|y^0\|^2,$$

so that there exists $x_0 \in X$ such that for some subsequence $u_n \rightharpoonup x_0$. From Theorem 4 and (4) we get (along the subsequence)

$$\liminf I_n(u_n) \geq I_0(x_0).$$

From (21) we know that $\min I_n \rightarrow \min I_0$. This implies $x_0 = \bar{u}_0$ and $u_n \rightharpoonup \bar{u}_0$ for the

original sequence. In the Hilbert direct sum $X \oplus Y$ the pairs

$$(u_n, L_n u_n) \rightarrow (\bar{u}_0, L_0 \bar{u}_0)$$

by (11), and their squared norms

$$\|u_n\|^2 + \|L_n u_n\|^2 \rightarrow \|\bar{u}_0\|^2 + \|L_0 \bar{u}_0\|^2$$

as easily verified by writing down $\min I_n \rightarrow \min I_0$. Therefore (29) is proved. The converse implication is trivial. Q.E.D.

An extension of the approximate minimization of I_n is considered now. This includes a version of the epsilon technique for our problem, as will be shown in a moment. Given u^0, y^0 we set

$$J(u, y) = \|u - u^0\|^2 + \|y - y^0\|^2, \quad u \in X, \quad y \in Y.$$

THEOREM 6. *Any of the equivalent conditions in Theorems 1, 2, 3, 4, 5 is equivalent to the following: for any u^0, y^0 , any sequence u_n^*, y_n^* such that*

$$(31) \quad J(u_n^*, y_n^*) - \min I_n \rightarrow 0, \quad y_n^* - L_n u_n^* \rightarrow 0,$$

then $u_n^ \rightarrow \bar{u}_0, y_n^* \rightarrow L_0 \bar{u}_0$.*

The proof is similar to that of Theorem 5. As an application, consider the following situation. Given u^0, y^0 , and a sequence $c_n \rightarrow +\infty$, we take an approximate minimization of

$$J_n(u, y) = \|u - u^0\|^2 + \|y - y^0\|^2 + c_n \|y - L_n u\|^2$$

on the whole space $X \oplus Y$. This represents an application of the epsilon technique. The order of the penalization terms increases together with that of the approximation of L_n to L_0 .

Suppose that u_n^*, y_n^* are found such that

$$J_n(u_n^*, y_n^*) - \inf J_n \rightarrow 0.$$

This means that (u_n^*, y_n^*) is an asymptotically minimizing sequence for J_n . Then there exists a constant d such that

$$c_n \|y_n^* - L_n u_n^*\|^2 \leq J_n(u_n^*, y_n^*) \leq d + \|u^0\|^2 + \|y^0\|^2.$$

Therefore

$$\|y_n^* - L_n u_n^*\|^2 \rightarrow 0.$$

On the other hand suppose that any of the equivalent conditions in Theorems 1-5 hold. Then for some sequence $\bar{c}_n \rightarrow 0$,

$$J(u_n^*, y_n^*) \leq J_n(u_n^*, y_n^*) \leq \bar{c}_n + J_n(\bar{u}_n, L_n \bar{u}_n) = \bar{c}_n + \min I_n,$$

implying that

$$\limsup J(u_n^*, y_n^*) \leq \min I_0.$$

As easily seen, u_n^* and y_n^* are bounded, and a u_0^* exists such that for a subsequence $u_n^* \rightharpoonup u_0^*$. Therefore

$$L_n u_n^* \rightharpoonup L_0 u_0^* \quad \text{and} \quad y_n^* \rightharpoonup L_0 u_0^*;$$

moreover

$$\min \lim J(u_n^*, y_n^*) \geq J(u_0^*, L_0 u_0^*) \geq \min I_0.$$

These facts show that (31) is satisfied in this case.

In the next steps we wish to characterize the convergence $\bar{u}_n \rightarrow \bar{u}_0$ for all u^0, y^0 by using information about $\min I_n$ and the values of I_n in suitably selected points. This will lead to a use of the following definition. Given u^0, y^0 , we say that I_n is *G-convergent* to I_0 , written

$$I_n \xrightarrow{G} I_0$$

if and only if for every $v \in X$

$$\min (I_n + (v, \cdot)) \rightarrow \min (I_0 + (v, \cdot));$$

the above minima are taken on the whole space X . For the notion of *G*-convergence see [3], [6] and the references thereof. In [3] it is shown that the definition above is in fact equivalent to *G*-convergence of I_n since for all n, u^0, y^0, u .

$$\|u - u^0\|^2 \leq I_n(u) \leq \|u - u^0\|^2 + 2c^2\|u\|^2 + 2\|y^0\|^2$$

(see [3, p. 143, (6)]).

THEOREM 7. *The following are equivalent facts: for all u^0, y^0 ,*

$$(32) \quad \bar{u}_n \rightarrow \bar{u}_0;$$

$$(33) \quad I_n \xrightarrow{G} I_0 \quad \text{and} \quad I_n(x) \rightarrow I_0(x) \quad \text{for all } x;$$

$$(34) \quad I_n \xrightarrow{G} I_0 \quad \text{and} \quad I_n(\bar{u}_0) \rightarrow I_0(\bar{u}_0);$$

$$(35) \quad \liminf \min I_n \geq \min I_0 \quad \text{and} \quad \limsup I_n(\bar{u}_0) \leq I_0(\bar{u}_0).$$

Proof. By Theorem 4, prop. 8, p. 149 and prop. 7, p. 143 of [3], (32) is equivalent to (33). Moreover by remembering (26) we see that (32) is equivalent to (35). Now $I_n(\bar{u}_0) \rightarrow I_0(\bar{u}_0)$ holds if and only if

$$\|L_n \bar{u}_0 - y^0\|^2 \rightarrow \|L_0 \bar{u}_0 - y^0\|^2 \quad \text{for all } u^0, y^0.$$

This in turn is equivalent to

$$\begin{aligned} & \|L_n(I + L_0^* L_0)^{-1}(u^0 + L_0^* y^0)\|^2 - 2(L_n(I + L_0^* L_0)^{-1}(u^0 + L_0 y^0), y^0) \\ & \rightarrow \|L_0(I + L_0^* L_0)^{-1}(u^0 + L_0^* y^0)\|^2 - 2(L_0(I + L_0^* L_0)^{-1}(u^0 + L_0 y^0), y^0), \end{aligned}$$

that is,

$$(36) \quad \|L_n z\|^2 - 2(L_n z, y^0) \rightarrow \|L_0 z\|^2 - 2(L_0 z, y^0)$$

for all $z \in X, y^0 \in Y$, since $(I + L_0^* L_0)^{-1}$ is a surjective map. By putting $y^0 = 0$ in (36), we get

$$\|L_n z\| \rightarrow \|L_0 z\|;$$

therefore by (36), $L_n z \rightarrow L_0 z$ so that

$$L_n z \rightarrow L_0 z, \quad \text{any } z.$$

This shows that (33) follows from (34). Q.E.D.

Theorem 7 shows, among other things, the following fact. The strong variational convergence (32), is equivalent to the seemingly less restrictive variational condition (35). This condition in turn is an essentially weaker requirement than assumptions (1), (2) of Theorem 1, p. 248 in [8d], guaranteeing *variational convergence* (in a general setting).

The following theorem relates the strong convergence to the weak one for optimal states and controls.

THEOREM 8. *Any of the equivalent conditions in theorems 1–7 is equivalent to the following:*

$$\bar{u}_n \rightharpoonup \bar{u}_0, \quad L_n \bar{u}_n \rightarrow L_0 \bar{u}_0, \quad \limsup \min I_n \leq \min I_0 \quad \text{for all } u^0, y^0.$$

Proof. By weak convergence of \bar{u}_n and $L_n \bar{u}_n$ we get

$$\liminf \min I_n \geq \min I_0$$

and therefore

$$\|\bar{u}_n\|^2 + \|L_n \bar{u}_n\|^2 = \min I_n - 2(u_n^0, \bar{u}_n) - 2(y^0, L_n \bar{u}_n) - \|u^0\|^2 - \|y^0\|^2 \rightarrow \|\bar{u}_0\|^2 + \|L_0 \bar{u}_0\|^2.$$

This shows $\bar{u}_n \rightarrow \bar{u}_0$. Q.E.D.

We end this section with an explicit estimate of the error $\|\bar{u}_n - \bar{u}_0\|$. Let us set

$$d_n = \|(L_n^* L_n - L_0^* L_0) L_0^* L_0\|.$$

THEOREM 9. *Assume that $L_0^* L_0$ is a compact operator, and that any of the equivalent conditions in Theorems 1–8 hold. Then for any u^0, y^0 , and for sufficiently large n the following estimate holds:*

$$\begin{aligned} \|\bar{u}_n - \bar{u}_0\| \leq & \|L_n^* y^0 - L_0^* y^0\| + (1 - d_n)^{-1} (\|(L_n^* L_n - L_0^* L_0)(u^0 + L_0^* y^0)\| \\ & + d_n \|u^0 + L_0^* y^0\|). \end{aligned}$$

Proof. Since pointwise convergence of an equicontinuous sequence is in fact uniform on relatively compact sets, by the compactness of $L_0^* L_0$ and the pointwise convergence of $L_n^* L_n$ to $L_0^* L_0$ we see that $d_n \rightarrow 0$ (see [2, prop. 1.7, p. 7]). By (12), for every u^* , y^* and n we obtain

$$\begin{aligned} \|\bar{u}_n - \bar{u}_0\| \leq & \|(I + L_n^* L_n)^{-1} (L_n^* y^0 - L_0^* y^0)\| + \|[(I + L_n^* L_n)^{-1} \\ & - (I + L_0^* L_0)^{-1}] (u^0 + L_0^* y^0)\|. \end{aligned}$$

From the last formula in the statement of Theorem 1.10, p. 9, of [2], and remembering that for all n $\|(I + L_n^* L_n)^{-1}\| \leq 1$, we get the conclusion. Q.E.D.

Applications. We consider perturbations of the standard linear-quadratic problem described by

$$(37) \quad \dot{x} = A(t)x + B_n(t)u \quad \text{a.e. in } (0, T), \quad x(0) = 0,$$

assuming a complete knowledge of the uncontrolled plant. For easy notation we

denote simply by $L^p(a, b)$ the corresponding Lebesgue space over the compact interval $[a, b]$, consisting of vector-valued functions of the appropriate dimension.

We denote by $x_n(u)$ the unique solution of (37) for a given $u \in L^2(0, T)$ (see (38) below). Here the control u and the state x are of dimension k, m , respectively. Given a fixed $c^* > 0$, we assume.

$$(38) \quad A \in L^1(0, T), \quad \int_0^T |B_n|^2 dt \leq c^*.$$

The cost is given by

$$(39) \quad I_n(u) = \int_0^T \{(u - u^0)' Q(u - u^0) + [(x_n(u) - y^0)' P(x_n(u) - y^0)]\} dt,$$

where $Q(t), P(t)$, are square symmetric matrices of respective dimension k, m . We assume

$$(40) \quad Q, P \in L^\infty(0, T); \quad \lambda' Q(t) \lambda \geq \alpha |\lambda|^2, \quad \mu' P(t) \mu \geq \alpha |\mu|^2$$

for all $\lambda \in R^k, \mu \in R^m$ and some $\alpha > 0$.

THEOREM 10. *Under the hypotheses (38), (39), (40) the following are equivalent facts:*

$$(41) \quad \bar{u}_n \rightarrow \bar{u}_0 \quad \text{for all } u^0, y^0 \in L^2(0, T);$$

$$(42) \quad B_n \rightarrow B_0 \quad \text{in } L^2(0, T) \quad \text{and} \quad B_n \rightarrow B_0 \quad \text{in every } L^2(0, T - \varepsilon), \quad 0 < \varepsilon < T.$$

Proof. Let F a fundamental matrix for the homogeneous system

$$\dot{x} = A(t)x.$$

Then obviously

$$(L_n u)(t) = F(t) \int_0^t F^{-1} B_n u ds, \quad u \in L^2(0, T). \quad 0 \leq t \leq T.$$

Given $u, v \in L^2(0, T)$ of the appropriate dimension we compute

$$\begin{aligned} (L_n u, v) &= \int_0^T (L_n u)' v dt = \int_0^T \int_0^t [F(t) F^{-1}(s) B_n(s) u(s)]' ds v(t) dt \\ &= \int_0^T u'(s) B_n'(s) F'^{-1}(s) \int_s^T F'(t) v(t) dt ds = (u, L_n^* v) \\ &= \int_0^T u' L_n^* v ds. \end{aligned}$$

Therefore

$$(L_n^* v)(t) = B_n'(t) F'^{-1}(t) \int_t^T F' v ds.$$

By Theorem 2, the conclusion will follow if it can be shown that equivalence

between (42) and the following two conditions holds:

$$(43) \quad L_n u \rightarrow L_0 u \quad \text{in } L^2(0, T) \text{ for all } u \in L^2(0, T);$$

$$(44) \quad L_n^* v \rightarrow L_0^* v \quad \text{in } L^2(0, T) \text{ for all } v \in L^2(0, T).$$

As a matter of fact, we show that (43) is equivalent to

$$(45) \quad B_n \rightharpoonup B_0 \quad \text{in } L^2(0, T),$$

and that (44) amounts to

$$(46) \quad B_n \rightarrow B_0 \quad \text{in every } L^2(0, T - \varepsilon), \quad 0 < \varepsilon < T.$$

Clearly (45) implies (43). Denote by

$$\int_0^t w \, ds$$

the map $t \rightarrow \int_0^t w \, ds$. Let us denote briefly $L^2(0, T)$ by L^2 . Now let us assume (41). Let $u \in L^2$ be fixed. We see that

$$(47) \quad \int_0^t F^{-1} B_n u \, ds \rightarrow \int_0^t F^{-1} B_0 u \, ds \quad \text{in } L^2,$$

by remarking that $v \rightarrow Fv$ is a linear isomorphism of L^2 onto L^2 . Given t' , $t'' \in [0, T]$ we see by (38) that

$$\left| \int_{t'}^{t''} F^{-1} B_n u \, ds \right| \leq c_1 \left| \int_{t'}^{t''} |u|^2 \, ds \right|^{1/2}$$

for some constant c_1 . From (47) we find

$$\int_0^t F^{-1} B_n u \, ds \rightarrow \int_0^t F^{-1} B_0 u \, ds \quad \text{a.e. in } (0, T)$$

for some subsequence (depending on u). By a.e. convergence and equicontinuity we get

$$\int_0^t F^{-1} B_n u \, ds \rightarrow \int_0^t F^{-1} B_0 u \, ds \quad \text{uniformly on } [0, T]$$

for all $u \in L^2$ and the original sequence. In particular

$$\int_0^t F^{-1} B_n u \, ds \rightarrow \int_0^t F^{-1} B_0 u \, ds, \quad \text{every } t \text{ and } u,$$

so that $F^{-1} B_n \rightharpoonup F^{-1} B_0$ in L^2 , hence (45) follows.

Let us denote by

$$\int^T L^2$$

the subspace of L^2 consisting of all functions $\int^T w \, ds$, $w \in L^2$ (with a notation similar to that above), and by

$$F^{-1} \int^T L^2$$

its image by the linear operator $v \rightarrow F'^{-1}v$ on L^2 . By considering the isomorphism $v \rightarrow Fv$ on L^2 we see that (44) is equivalent to

$$(48) \quad B'_n z \rightarrow B'_0 z \quad \text{in } L^2 \text{ for all } z \in F'^{-1} \int_0^T L^2.$$

By integrating $(0, \dots, 0, 1, 0, \dots, 0)$ (only one 1) between t and T and using the result in (48) we see that (48) implies

$$(49) \quad (T-t)B'_n F'^{-1} \rightarrow (T-t)B'_0 F'^{-1} \quad \text{in } L^2;$$

hence given ε such that $0 < \varepsilon < T$ we get

$$B_n \rightarrow B_0 \quad \text{in } L^2(0, T-\varepsilon).$$

Conversely assume (46). Set $b_n = B'_n - B'_0$. Given $z \in F'^{-1} \int_0^T L^2$, ε with $0 < \varepsilon < T$, we have

$$(50) \quad \int_0^T |b_n z|^2 dt = \int_0^{T-\varepsilon} |b_n z|^2 dt + \int_{T-\varepsilon}^T |b_n z|^2 dt.$$

Clearly

$$\int_0^{T-\varepsilon} |b_n z|^2 dt \leq (\max |z|^2) \int_0^{T-\varepsilon} |b_n|^2 dt \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Moreover there exists a point t_ε , $T-\varepsilon \leq t_\varepsilon \leq T$, such that

$$\sup_n \int_{T-\varepsilon}^T |b_n z|^2 dt \leq |z(t_\varepsilon)|^2 \sup_n \int_0^T |b_n|^2 dt \rightarrow 0 \quad \text{as } \varepsilon \rightarrow 0.$$

The previous estimates combined with (50) show that (48) holds. Q.E.D.

In passing we remark that a necessary condition to (41) is a.e. convergence of B_n to B_0 in $(0, T)$, as easily seen by (49).

This example of variational perturbations, acting on A too, with uniformly convergent optimal states, has been studied in [5] in a different setting. In such a paper a sequence of perturbations is given as in (37) with $A(t)$ replaced by $A_n(t)$, and $x(0)$ not necessarily 0. The cost is given by (39), and the constraint $\|u\| \leq 1$ is added. Assuming boundedness of A_n in L^1 and of B_n in L^2 , then $\bar{u}_n \rightarrow \bar{u}_0$ together with uniform convergence of the optimal states, for all $u^0, y^0 \in L^2$ and all $x(0)$, if and only if $A_n \rightarrow A_0$ and $B_n \rightarrow B_0$ both in L^2 . The relationship with Theorem 10 is then clear from (42). Here the perturbations we consider belong to a more restricted class than in [5] but a less stringent convergence is required on the optimal states.

As a final application we consider a linear quadratic problem for a distributed control system governed by an elliptic Dirichlet problem. Here the perturbations act on the uncontrolled plant only.

Given $\alpha > 0$, $\omega > 0$, and an open bounded set $\Omega \subset R^p$ with boundary $\partial\Omega$ we consider sequences of functions

$$a_{ij}^n = a_{ji}^n \in L^\infty(\Omega), \quad n = 0, 1, 2, \dots, \quad 1 \leq i, j \leq p,$$

a_{ij}^0 given, such that the following uniform ellipticity assumption holds:

$$(51) \quad \alpha|t|^2 \leq \sum_{i,j=1}^p a_{ij}^n(x) t_i t_j \leq \omega|t|^2, \quad \text{all } t \in \mathbb{R}^p, \quad \text{a.e. } x \in \Omega.$$

We denote by $H^1(\Omega)$ the Hilbert space of real functions $z \in L^2(\Omega)$ whose distributional derivatives belong to $L^2(\Omega)$. Moreover $H_0^1(\Omega)$ denotes the closure in $H^1(\Omega)$ of the space of smooth functions with compact support contained in Ω . Both $H^1(\Omega)$ and $H_0^1(\Omega)$ are equipped with the usual norm. We shall denote by $H^{-1}(\Omega)$ the dual space of $H_0^1(\Omega)$. As well known, see [1], any element of $H^{-1}(\Omega)$ can be represented as a sum of derivatives of functions in $L^2(\Omega)$. Given a control $u \in L^2(\Omega)$ we shall denote by $z_n(u) \in H_0^1(\Omega)$ the weak solution of the following Dirichlet problem:

$$(52) \quad \begin{aligned} - \sum_{i,j=1}^p \frac{\partial}{\partial x_i} a_{ij}^n \frac{\partial z}{\partial x_j} &= u \quad \text{in } \Omega, \\ z &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

The cost is given by

$$I_n(u) = \int_{\Omega} (u - u^0)^2 dx + \int_{\Omega} (z_n(u) - y^0)^2 dx + \int_{\Omega} (\dot{z}_n(u) - \dot{y}^0)^2 dx,$$

$u^0 \in L^2(\Omega)$, $y^0 \in H_0^1(\Omega)$. Here a dot denotes the gradient.

In the statement of the next theorem we denote by $H_{\text{loc}}^{-1}(\Omega)$ the dual space of $H_{\text{loc}}^1(\Omega)$. This last space is defined as follows: $z \in H_{\text{loc}}^{-1}(\Omega)$ if and only if z and its distributional derivatives belong to $L^2(K)$ for every compact $K \subset \Omega$.

THEOREM 11. *Assume that (51) holds. A necessary and sufficient condition such that $\bar{u}_n \rightarrow \bar{u}_0$ in $L^2(\Omega)$ for all $u^0 \in L^2(\Omega)$, $y^0 \in H_0^1(\Omega)$ is that*

$$(53) \quad \text{and} \quad \begin{aligned} a_{ij}^n &\rightarrow a_{ij}^0 \quad \text{in } L^2(\Omega) \text{ for all } i, j, \\ \sum_{i=1}^p \frac{\partial}{\partial x_i} a_{ij}^n &\rightarrow \sum_{i=1}^p \frac{\partial}{\partial x_i} a_{ij}^0 \quad \text{in } H_{\text{loc}}^{-1}(\Omega) \text{ for all } j. \end{aligned}$$

Proof. For a given sequence a_{ij}^n as in (51) let us denote by

$$A_n = - \sum_{i,j=1}^p \frac{\partial}{\partial x_i} a_{ij}^n \frac{\partial}{\partial x_j}: H_0^1(\Omega) \rightarrow H^{-1}(\Omega)$$

the corresponding sequence of elliptic operators. Then, for any n , A_n is an isomorphism between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$, with an uniformly bounded inverse A_n^{-1} . Moreover by the symmetric character of a_{ij}^n , any A_n is a selfadjoint operator. Therefore L_n is the restriction of A_n^{-1} to $L^2(\Omega)$. The space $L^2(\Omega)$ is continuously imbedded in $H^{-1}(\Omega)$, as well known (see [1]); hence by (51) the norms of

$$L_n: L^2(\Omega) \rightarrow H_0^1(\Omega)$$

are uniformly bounded. Moreover, $L_n = L_n^*$ for all n . By Theorem 2, we see that $\bar{u}_n \rightarrow \bar{u}_0$ in $L^2(\Omega)$ for all u^0 , y^0 if and only if $L_n u \rightarrow L_0 u$ in $H_0^1(\Omega)$ for all $u \in L^2(\Omega)$. This in turn is equivalent to $L_n v \rightarrow L_0 v$ in $H_0^1(\Omega)$ for all $v \in H^{-1}(\Omega)$ since $L^2(\Omega)$ is densely imbedded in $H^{-1}(\Omega)$ and the norms of A_n^{-1} are uniformly bounded. But

this is equivalent to strong operator convergence of A_n to A_0 ([6, remark 10, p. 10]). By [6, remark 8, p. 10] we get the required equivalence between (53) and the convergence of the optimal controls. Q.E.D.

REFERENCES

- [1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
- [2] P. M. ANSELONE, *Collectively Compact Operator Approximation Theory and Applications to Integral Equations*, Prentice-Hall, Engelwood Cliffs, NJ, 1971.
- [3] P. MARCELLINI, *Su una convergenza di funzioni convesse*, Boll. Un. Mat. Ital., 8 (1973), pp. 137–158.
- [4] U. MOSCO, *Convergence of convex sets and of solutions of variational inequalities*, Advances in Math., 3 (1969), pp. 510–585.
- [5] G. PIERI, *Variational perturbations of the linear-quadratic problem*, J. Optimization Theory Appl., to appear.
- [6] S. SPAGNOLO, *Convergence in energy for elliptic operators*, Proc. Symp. Numerical Solutions of Partial Differentials Equations, Univ. of Maryland, College Park, 1975, to appear.
- [7] G. STAMPACCHIA, *Variational inequalities*, Theory and Applications of Monotone Operators, Proceedings of a N.A.T.O. Advanced study Institute, Venice, 1968, A. Ghizzetti, ed. (Oderisi, Gubbio (Italy), 1969), pp. 101–192.
- [8a] T. ZOLEZZI, *Su alcuni problemi debolmente ben posti di controllo ottimo*, Ricerche Mat., 21 (1972), pp. 184–203.
- [8b] ———, *Su alcuni problemi fortemente ben posti di controllo ottimo*, Ann. Mat. Pura. Appl., 95 (1972), pp. 148–160.
- [8c] ———, *Condizioni necessarie di stabilità variazionale per il problema lineare del minimo scarto finale*, Boll. Un. Mat. Ital., 7 (1973), pp. 142–150.
- [8d] ———, *On convergence of minima*, Ibid., 8 (1973), pp. 246–257.
- [8e] ———, *On equiwellposed minimum problems*, in preparation.

AN APPROXIMATION METHOD IN OPTIMAL STOCHASTIC CONTROL*

JEAN-MICHEL BISMUT†

Abstract. The purpose of this paper is to prove that an approximation scheme can be defined for the general problems of optimal stochastic control which we have solved in *Théorie probabiliste du contrôle des diffusions*, Mem. Amer. Math. Soc., 1976 [2].

1. Introduction. Let us consider the stochastic differential equation

$$(1.1) \quad \begin{aligned} dx &= b(t, x, u(t, x)) dt + \sigma(t, x) \cdot d\beta^b, \\ x_s &= x \end{aligned}$$

where β^b is a Brownian motion and where u is a control variable with values in U . We then consider a criterion

$$(1.2) \quad V_u(s, x) = E_{(s,x)}^{b_u} \int_s^T L(t, x_t, u(t, x_t)) dt$$

which we want to minimize for all (s, x) . We have used potential theory methods in [2] to prove existence results for this problem under very general conditions. In particular, it was proved in [2, (4.49)] that for a given control u , there is a Borel function H_u on $R^+ \times R^d$ such that:

$$(1.3) \quad V_u(t, x_t) - V_u(s, x_s) = - \int_s^t L(\sigma, x_\sigma, u(\sigma, x_\sigma)) d\sigma + \int_s^t H_u(\sigma, x_\sigma) \cdot d\beta^b.$$

Moreover, we found in [2] that there is a probability measure ν on $R^+ \times R^d$ such that for μ_0 to be an optimum, it is necessary and sufficient that, ν a.e.

$$(1.4) \quad \begin{aligned} &L(t, x, u_0(t, x)) + \langle H_{u_0}(t, x), \sigma^{-1}(t, x)b(t, x, u_0(t, x)) \rangle \\ &= \inf_{u \in U} L(t, x, u) + \langle H_{u_0}(t, x), \sigma^{-1}(t, x)b(t, x, u) \rangle. \end{aligned}$$

Moreover, in [4] we applied the theory of duality to stochastic control problems. In particular it appeared that it is possible to associate to each control problem a dual variable which is itself an adapted right continuous process, and a solution of a dual stochastic control problem.

Finally, we applied, in [6], the theory of duality to problems defined by (1.1)–(1.2). Although duality did not give us existence results for this problem—which had to be obtained by stronger methods in [1]—it appeared then in [5] that, in a loose sense, the “real” controlled state variable is the Cameron–Martin–Girsanov density of process (1.1) relative to the same process (1.1) with $b = 0$. This problem is then handled by convex analysis methods. Equation (1.4) appeared then to be a special case of the general extremality relations for two convex dual problems, and process $V_{u_0}(t, x_t)$ is then the optimal dual process when u_0 is

* Received by the editors June 6, 1975, and in final revised form February 22, 1977.

† 191 rue d'Alésia, 75014/Paris, France.

optimum for (1.2) This allowed us to reinterpret the results of Benes [1], Duncan–Varaiya [8], Davis–Varaiya [7] in terms of duality.

It is then natural to ask if, from relation (1.4), it is possible to derive an approximation scheme for the optimal control problem which would use the duality features of the problem. More precisely, we start from a given control u_1 . We find a control u_2 minimizing ν a.e.:

$$(1.5) \quad L(t, x, u) + \langle H_{u_1}(t, x), \sigma^{-1}(t, x)b(t, x, u) \rangle.$$

This defines V_{u_2} and consequently, H_{u_2} . u_3 is then defined in the same way from H_{u_2}, \dots, u_n from u_{n-1} .

This is precisely the iteration procedure studied by Fleming in [9] with partial differential equations techniques. When $a = \sigma\sigma^*$ is Lipschitz continuous and uniformly elliptic, Fleming considers the partial differential equation:

$$(1.6) \quad \frac{\partial V_u}{\partial s} + \frac{1}{2} \sum_{i,j} a_{ij} \frac{\partial^2 V_u}{\partial x_i \partial x_j} + \left\langle b_u, \frac{\partial V_u}{\partial x} \right\rangle = -L_u, \quad V_u(x, T) = 0$$

to define the cost function V_u . When (b_u, L_u) stays uniformly bounded in $L_\infty(R^+ \times R^d)$, Fleming uses results by Friedman and Gagliardo to prove that V_u and $\partial V_u / \partial x$ form a family of equi-continuous functions. Fleming can then prove the convergence of $(V_{u_n}, \partial V_{u_n} / \partial x)$ to the optimal $(V_{u_0}, \partial V_{u_0} / \partial x)$ by using a compactness argument in the space of continuous functions.

Our approach will be different:

1. We use the results of Stroock and Varadhan [12]–[13] on diffusions, which require only continuity and ellipticity on a .
2. As in a previous work [2], we do not use any continuity argument on V_u and $\partial V_u / \partial x$ ($\partial V_u / \partial x$ is not always well defined). Moreover the techniques used here apply to more general criteria of type:

$$e^{ps} E_{(s,x)}^{b_u} \int_s^{T_A} e^{-pt} L(t, x_t, u(t, x_t)) dt$$

where A is a Borel set, and where T_A is the hitting time on A . In this case, the cost function is generally not continuous, and compactness arguments do not work.

3. We use constantly a common feature of the Markov processes studied in stochastic control [2], [5]: they have a common reference measure μ , i.e. if L_u is 0 μ a.e. on $R^+ \times R^d$ then V_u is 0 everywhere on $R^+ \times R^d$.

4. The compactness argument is applied in the space of control functions \mathcal{L} where (b_u, L_u) varies, and not in the space of cost functions.

The main result of the paper is that the sequence of functions V_{u_n} decreases to the optimal V_{u_0} and that H_{u_n} converges in a L_1 -space to H_{u_1} . This result has important consequences because by (1.4), H_{u_0} characterizes entirely the optimal solutions of the problem. Moreover, it underlines the deep convex nature of the problem of optimal stochastic control, to which we apply a “classical” approximation scheme. Finally, it can be applied as in [5, Part III] for more general processes and problems, especially to the optimal stopping problem. This scheme may also apply to all the criteria considered in [2], which are more general than (1.2). It can

also be coupled with partial differential equation methods or with discretization of the state space for computational purposes.

It should be noted that, under the assumptions of Fleming [9], H_{u_n} is necessarily equal to $\partial V_{u_n} \sigma / \partial x$. The result which we obtain on the convergence of H_{u_n} to H is then weaker than the result of Fleming, which guarantees uniform convergence of $\partial V_{u_n} / \partial x$ to $\partial V_{u_0} / \partial x$. This is not unnatural, since the potentials for the diffusions of Stroock and Varadhan are solutions in a weak sense of parabolic partial differential equations, and do not always admit regular derivatives. However, if we apply our result under Fleming's assumptions, we reobtain the result of Fleming, since equicontinuity and convergence in an L_1 space of H_{u_n} to H_{u_0} guarantees uniform convergence of H_{u_n} to H_{u_0} .

All the main notation and techniques are taken from [2], to which we will refer constantly. In § 2 we state the main results of [2] which we use in this paper. In § 3 we prove a simple functional result. Finally, in § 4 the main result is proved.

2. Statement of the problem.

a is a function defined on $R^+ \times R^d$ with values in $R^d \otimes R^d$ which is

- (i) continuous and bounded,
- (ii) with positive definite values,
- (iii) uniformly elliptic.

σ is the positive square root of a .

b is a Borel bounded function defined on $R^+ \times R^d$ with values in R^d .

$Q_{(s,x)}^b$ is the measure on the space of continuous function solutions of the martingale problem of Stroock and Varadhan [12] associated to (a, b, s, x) .

$E_{(s,x)}^b$ is the expectation operator relative to $Q_{(s,x)}^b$.

K is a set-valued function defined on $R^+ \times R^d$ with values in $R^d \times R$ which is

- (i) Borel measurable and uniformly bounded,
- (ii) with nonempty compact values.

\mathcal{L} is the set of $dt \otimes dx$ classes of Borel selections of K , with the $\sigma(L_\infty(R^+ \times R^d), L_1(R^+ \times R^d))$ topology.

T is a positive constant.

For $0 \leq s \leq T$ and $c = (b, L) \in \mathcal{L}$ we consider the function V_c :

$$(2.1) \quad V_c(s, x) = E_{(s,x)}^b \int_s^T L(u, x_u) du.$$

$V_{(s,x)}^0$ is the measure

$$L \rightarrow E_{(s,x)}^0 \int_s^T L(u, x_u) du.$$

μ is a probability measure on $[0, T] \times R^d$. We consider the functional:

$$(2.2) \quad I_\mu(c) = \int V_c(s, x) d\mu(s, x).$$

By (4.49) and (3.17) in [2], we know that for any bounded b' , $V_c(t, x_t)$ can be written as:

$$(2.3) \quad V_c(t, x_t) = V_c(s, x_s) - \int_s^t L(u, x_u) du + \int_s^t \langle H_c(u, x_u), d\beta_u^{b'} \rangle - \int_s^t \langle H_c, \sigma^{-1}(b - b') \rangle(u, x_u) du$$

$\beta^{b'}$ being a Brownian motion for $Q^{b'}$, and H_c being a Borel function such that for $(s, x) \in [0, T] \times R^d$

$$(2.4) \quad E_{(s,x)}^b \int_s^T |H_c(u, x_u)|^2 du < +\infty.$$

By Theorem IV-5 of [2] we know that if $c \in \mathcal{L}$ and if $c' \in \mathcal{L}$ is such that, μV^0 a.e.,

$$(2.5) \quad L'(t, x) + \langle H_c(t, x), \sigma^{-1}(t, x)b'(t, x) \rangle \leq L(t, x) + \langle H_c(t, x), \sigma^{-1}(t, x)b(t, x) \rangle$$

then

$$(2.6) \quad I_\mu(c') \leq I_\mu(c).$$

If the inequality is strict on a non μV^0 negligible set,

$$(2.7) \quad I_\mu(c') < I_\mu(c).$$

If such a choice of c' is impossible, c is a minimum for I_μ .

We then define an iteration procedure: c_n is taken in \mathcal{L} . H_{c_n} is the associated element. c_{n+1} is one of the elements of \mathcal{L} such that

$$(2.8) \quad L_{n+1}(t, x) + \langle H_{c_n}(t, x), \sigma^{-1}(t, x)b_{n+1}(t, x) \rangle = \inf_{(b,L) \in K(t,x)} L + \langle H_{c_n}(t, x), \sigma^{-1}(t, x)b \rangle, \quad \mu V^0 \text{ a.e.}$$

If c_{n+1} can be taken to be equal to c_n , the iteration is stopped.

Then, we want to know if (c_n, H_{c_n}) converge. The answer will be positive for H_{c_n} , which will be proved to converge in a strong sense to the optimal H defined in [2, Corollary of Thm. IV-7] in the following way: if $\tilde{\mu}$ is a probability measure mutually absolutely continuous with the Lebesgue measure on $[0, T] \times R^d$, H is a Borel function defined $\tilde{\mu} V^0$ a.e., which is such that if q is the function

$$(2.9) \quad q = \min_{c' \in \mathcal{L}} V_{c'}.$$

then for any $c = (b, L)$ in \mathcal{L} such that $q = V_c$,

$$(2.10) \quad q(t, x_t) = q(s, x_s) - \int_s^t L(u, x_u) du + \int_s^t H(u, x_u) \cdot d\beta_u^b, \quad Q^b \text{ a.s.}$$

3. A simple functional result. We give first a simple functional result which is essential in the sequel.

THEOREM 3.1. *For any $b' \in L_\infty(\mathbf{R}^+ \times \mathbf{R}^d)$,*

$$E_{(s,x)}^{b'} \int_s^T |H_c(u, x_u)|^2 du$$

is finite and uniformly bounded when c and b' vary in a bounded set, and when (s, x) varies in $\mathbf{R}^+ \times \mathbf{R}^d$.

Proof. By (2.3) we know that

$$(3.1) \quad \varphi(s, x) = E_{(s,x)}^b \int_s^T |H_c(u, x_u)|^2 du$$

is uniformly bounded.

By the Markov property of Q^b , we have:

$$(3.2) \quad \begin{aligned} E_{(s,x)}^b \left(\int_s^T |H_c(u, x_u)|^2 du \right)^2 &= 2E_{(s,x)}^b \int_s^T |H_c(u, x_u)|^2 du \int_u^T |H_c(t, x_t)|^2 dt \\ &= 2E_{(s,x)}^b \int_s^T |H_c(u, x_u)|^2 \varphi(u, x_u) du. \end{aligned}$$

$E_{(s,x)}^b \left(\int_s^T |H_c(u, x_u)|^2 du \right)^2$ is then uniformly bounded. The density of $Q_{(s,x)}^{b'}$ relative to $Q_{(s,x)}^b$ on M_T^s can be written as

$$(3.3) \quad \begin{aligned} Z_T &= \exp \left\{ \int_s^t \langle (b' - b)(u, x_u), \sigma^{-1}(u, x_u) d\beta_u^b \rangle \right. \\ &\quad \left. - \frac{1}{2} \int_s^t \langle (b' - b), a^{-1}(b' - b) \rangle (u, x_u) du \right\}. \end{aligned}$$

Necessarily, if in (3.3) we replace $b' - b$ by $2(b' - b)$, (3.3) remains integrable with mean 1. Then

$$(3.4) \quad E_{(s,x)}^{Q^b} |Z_T|^2 \leq (T-s) \sup \|a^{-1}(t, x)\| \sup \text{ess} \|b' - b\|^2.$$

It follows immediately that:

$$(3.5) \quad E_{(s,x)}^{b'} \int_s^T |H_c(u, x_u)|^2 du$$

stays uniformly bounded. \square

Remark 3.1. The majoration of (3.2) can also be proved by the inequalities of [11].

4. Convergence. We now prove the main result on convergence.

THEOREM 4.1. $I_\mu(c_n)$ is a decreasing sequence converging to $\int q d\mu$. Any weak limit of a subsequence of c_n is such that $I_\mu(c) = \int q d\mu$. Moreover H_{c_n} converges to H in $L_1(\mu V^0)$.

If μ is mutually absolutely continuous with the Lebesgue measure on $\mathbf{R}^+ \times \mathbf{R}^d$, the sequence $\{V_{c_n}\}$ decreases to V_c , and H is entirely defined by the limit of H_{c_n} in $L_1(\mu V^0)$.

Proof. We consider the problem associated to \hat{K} , where $\hat{K}(t, x)$ is the closed convex hull of $K(t, x)$. By the results of [2, Chap. IV, Part 5], the optimization

problem is not changed, and in particular H stays the same for the two problems, while the two problems have the same minimal values.

Then \mathcal{L} associated to \hat{K} is compact. We take a subsequence c_{m_n} converging to $c \in \mathcal{L}$, such that c_{m_n+1} converges to $c' \in \mathcal{L}$. We rename these sequences $\{c_n\}$ and $\{c'_n\}$. Then we have

$$(4.1) \quad I_\mu(c_{n+1}) \leq I_\mu(c'_n) \leq I_\mu(c_n).$$

The sequences $I_\mu(c_n)$ and $I_\mu(c'_n)$ are both decreasing, have a lower bound, and then, by (4.1), have a common limit.

By Theorem IV-4 of [2], I_μ being continuous on \mathcal{L} , we find then

$$(4.2) \quad I_\mu(c) = I_\mu(c').$$

Moreover, by the same referenced result, V_{c_n} converges to V_c . We then have

$$(4.3) \quad \begin{aligned} V_{c_n}(s, x) &= \int_s^T L_n(u, x_u) du - \int_s^T H_{c_n} \cdot d\beta^{b_n}, \\ V_c(s, x) &= \int_s^T L(u, x_u) du - \int_s^T H_c \cdot d\beta^b. \end{aligned}$$

By (2.3), we have:

$$(4.4) \quad V_c(s, x) = \int_s^T (L + \langle H_c, \sigma^{-1}(b - b_n) \rangle)(t, x_t) dt - \int_s^T H_c \cdot d\beta^{b_n}.$$

We now prove that for $(s, x) \in [0, T] \times R^d$, when $n \rightarrow +\infty$,

$$(4.5) \quad K_n(s, x) = E_{(s,x)}^{b_n} \int_s^T \langle H_c, \sigma^{-1}(b_n - b) \rangle(u, x_u) du \rightarrow 0.$$

Let $p_n(s, x)$ (resp. $p(s, x)$) be the transition density for $Q_{(s,x)}^{b_n}$ (resp. $Q_{(s,x)}^b$) (see [12, § 8]).

By Theorem IV-4 in [2], we know that if R_n converges in the $\sigma(L_\infty, L_1)^1$ topology to R , then $\langle p_n(s, x), R_n \rangle$ converges to $\langle p(s, x), R \rangle$. Let $R_n(t, y) = \text{sgn}(p_n - p)(s, x, t, y)$ and let R_{n_k} be a subsequence of R_n converging weakly to R . We have then

$$(4.6) \quad |\langle p_{n_k}(s, x) - p(s, x), R_{n_k} \rangle| \leq |\langle p_{n_k}(s, x), R_{n_k} \rangle - \langle p, R \rangle| + |\langle p, R - R_{n_k} \rangle|.$$

Relation (4.6) implies that $\|p_{n_k}(s, x) - p(s, x)\|_{L_1}$ converges to zero, or equivalently that $p_{n_k}(s, x)$ converges to $p(s, x)$ in L_1 . By reasoning on all the subsequences of \mathbb{N} , we find that $p_n(s, x)$ converges to $p(s, x)$ in L_1 .

Moreover, by Theorem 3.1, we know that:

$$(4.7) \quad \int_s^T dt \int_{R^d} p_n(s, x, t, y) |H_c(t, y)|^2 dy$$

stays uniformly bounded by a constant M .

We have the following situation for any (s, x) :

$$(4.8) \quad \begin{aligned} p_n(s, x) &\rightarrow p(s, x) \quad \text{in } L_1, \\ p_n(s, x) |H_c|^2 &\quad \text{and } p(s, x) |H_c|^2 \text{ are uniformly bounded in } L_1. \end{aligned}$$

¹ $L_1 = L_1([0, T] \times R^d)$.

Then

$$(4.9) \quad \begin{aligned} \|(p_n - p)(s, x)H_c\|_{L_1} &\leq \|(p_n - p)(s, x)1_{|H_c| \leq a}H_c\|_{L_1} \\ &\quad + \|(p_n - p)(s, x)1_{|H_c| > a}H_c\|_{L_1}. \end{aligned}$$

But we have by Schwarz's inequality

$$(4.10) \quad \|p_n(s, x)1_{|H_c| > a}H_c\|_{L_1} \leq M^{1/2} \|p_n(s, x)1_{|H_c| > a}\|_{L_1}^{1/2}$$

and by Chebyshev's inequality, it follows that

$$(4.11) \quad \|p_n(s, x)1_{|H_c| > a}H_c\|_{L_1} \leq \frac{M}{a}.$$

By (4.9) and (4.11), it follows that $p_n(s, x)H_c$ converges to $p(s, x)H_c$ in L_1 , and necessarily,

$$(4.12) \quad K_n(s, x) \rightarrow 0.$$

We have as in (3.2)

$$(4.13) \quad \begin{aligned} E_{(s,x)}^{b_n} \left| \int_s^T \langle H_c, \sigma^{-1}(b_n - b) \rangle(u, x_u) du \right|^2 \\ = 2E_{(s,x)}^{b_n} \int_s^T \langle H_c, \sigma^{-1}(b_n - b) \rangle(u, x_u) K_n(u, x_u) du. \end{aligned}$$

Knowing that $p_n(s, x)H_c \rightarrow p(s, x)H_c$ in L_1 , that $b_n \rightarrow b$ weakly, and that $K_n \rightarrow K$ simply, while staying uniformly bounded, we see that (4.13) converges to 0 and stays uniformly bounded in (s, x) . Similarly

$$(4.14) \quad E_{(s,x)}^{b_n} \left| \int_s^T (L_n - L)(u, x_u) du \right|^2 \rightarrow 0$$

while staying uniformly bounded.

Knowing that $V_{c_n} \rightarrow V_c$ simply, we find from (4.3) and (4.4) that

$$(4.15) \quad E_{(s,x)}^{b_n} \left| \int_s^T (H_{c_n} - H_c)(u, x_u) d\beta_u^{b_n} \right|^2 \rightarrow 0.$$

This implies

$$(4.16) \quad E_{(s,x)}^{b_n} \int_s^T |H_{c_n} - H_c|^2(u, x_u) du \rightarrow 0.$$

But $p_n(s, x) \rightarrow p(s, x)$ in L_1 , and the measures $p_n(s, x) dy$ stay equivalent to the measure $V^0(s, x)$. This implies that $H_{c_n} \rightarrow H_c$ in "probability" for the measure μV^0 (i.e., from any subsequence of $\{H_{c_n}\}$, one can extract a subsequence converging μV^0 a.e. to H_c).

Moreover, by Theorem 3.1, $\{H_{c_n}\}$ stays bounded in $L_2(\mu V^0)$ and is then uniformly integrable in $L_1(\mu V^0)$. By Theorem II-21 of [10] $H_{c_n} \rightarrow H_c$ in $L_1(\mu V^0)$.

But, we know that c'_n is such that for any $\tilde{c} = (\tilde{b}, \tilde{L}) \in \mathcal{L}$

$$(4.17) \quad \int (L'_n + \langle H_{c_n}, \sigma^{-1} b'_n \rangle)(t, x) d(\mu V^0) \leq \int (\tilde{L} + \langle H_{c_n}, \sigma^{-1} \tilde{b} \rangle)(t, x) d(\mu V^0).$$

Then, with H_{c_n} converging to H_c in $L_1(\mu V^0)$, we have

$$(4.18) \quad \int (L' + \langle H_c, \sigma^{-1} b' \rangle)(t, x) d(\mu V^0) \leq \int (\tilde{L} + \langle H_c, \sigma^{-1} \tilde{b} \rangle)(t, x) d(\mu V^0).$$

This will imply immediately

$$(4.19) \quad L'(t, x) + \langle H_c, \sigma^{-1} b' \rangle(t, x) = \min_{(\tilde{b}, \tilde{L}) \in \mathcal{K}(t, x)} \tilde{L} + \langle H_c(t, x), \sigma^{-1}(t, x) \tilde{b} \rangle \mu V^0 \quad \text{a.e.}$$

If on a μV^0 nonnegligible set

$$(4.20) \quad L'(t, x) + \langle H_c, \sigma^{-1} b' \rangle(t, x) < L(t, x) + \langle H_c, \sigma^{-1} b \rangle(t, x)$$

then

$$(4.21) \quad I_\mu(c') < I_\mu(c).$$

But, by (4.2), this is impossible. Then, necessarily, μV^0 a.e.,

$$(4.22) \quad L(t, x) + \langle H_c, \sigma^{-1} b \rangle(t, x) = \min_{(\tilde{b}, \tilde{L}) \in \mathcal{K}(t, x)} \tilde{L} + \langle H_c(t, x), \sigma^{-1}(t, x) \tilde{b} \rangle.$$

By Theorem IV-5 of [2], c is an optimum for I_μ on \mathcal{L} . By Remark IV-2 of [2], if q is defined by

$$(4.23) \quad q = \inf_{\tilde{c} \in \mathcal{L}} V_{\tilde{c}}$$

if $c \in \mathcal{L}$, then $V_c - q$ is Q^b p -excessive. But, c being an optimum for I_μ ,

$$(4.24) \quad V_c = q, \quad \mu \text{ a.e.}$$

The processes $V_c(t, x_t)$ and $q(t, x_t)$ are then a.s. equal for the measure Q_p^0 . By (2.3) and (2.10)

$$(4.25) \quad H = H_c, \quad \mu V^0 \text{ a.e.}$$

By reasoning on all the possible subsequences of the initial $\{c_n\}$ the result follows.

If μ is mutually absolutely continuous with the Lebesgue measure on $R^+ \times R^d$, necessarily, by Theorem IV-5 of [2]

$$V_{c_n} \leq V_{c_{n-1}}.$$

The sequence V_{c_n} is then decreasing. For any weak limit c of a subsequence $\{c_{n_k}\}$, (4.23) holds. V_c and q being finely continuous, we have

$$(4.26) \quad V_c = q. \quad \square$$

Remark 4.1. The convergence of H_{c_n} to H is essential, even if c_n does not necessarily converge. Since the optimum solutions are obtained from H , the determination of H is then fundamental.

If a is only elliptic, then it is possible to prove the convergence of H_{c_n} to H in μV^0 probability.

Remark 4.2. When $a = I$, it is possible to choose $\mu = \delta(0, 0)$ to obtain the whole optimal H .

Remark 4.3. The results are extendable to all the optimization problems studied in [2].

REFERENCES

- [1] V. E. BENES, *Existence of optimal stochastic control laws*, this Journal, 9 (1971), pp. 446–472.
- [2] J. M. BISMUT, *Théorie probabiliste du contrôle des diffusions*, Mem. Amer. Math. Soc., 4 (Jan. 1976), no. 167, pp. 1–130.
- [3] ———, *Linear quadratic optimal stochastic control with random coefficients*, this Journal, 14 (1976), pp. 419–444.
- [4] ———, *Conjugate convex functions in optimal stochastic control*, J. Math. Anal. and Appl., 44 (1973), no. 2, pp. 384–404.
- [5] ———, *Control of jump processes and applications*, Bul. Soc. Math. France, to appear.
- [6] ———, *Duality methods in the control of densities*, to appear.
- [7] M. H. A. DAVIS AND P. VARAIYA, *Dynamic programming conditions for partially observable stochastic systems*, this Journal, 11 (1973), pp. 226–261.
- [8] T. DUNCAN AND P. VARAIYA, *On the existence of a stochastic control system*, this Journal, 9 (1971), pp. 354–371.
- [9] W. H. FLEMING, *Some Markovian optimization problems*, J. Math. Mech., 12 (1963), no. 1, pp. 131–140.
- [10] P. A. MEYER, *Probabilités et Potentiels*, Hermann, Paris; English translation, Blaisdell, Boston, 1966.
- [11] ———, *Une Majoration du Processus Croissant Naturel Associé à une Martingale*, Séminaire de Probabilités, No. 2, Lecture Notes in Mathematics, No. 51, Springer-Verlag, Berlin, 1968.
- [12] D. W. STROOCK AND S. R. S. VARADHAN, *Diffusion processes with continuous coefficients*, Comm. Pure Appl. Math., 22 (1969), pp. 345–400, pp. 479–530.
- [13] ———, *Diffusion processes with boundary conditions*, Comm. Pure Appl. Math., 24 (1971), pp. 147–225.

STABILIZATION OF BILINEAR CONTROL SYSTEMS WITH APPLICATIONS TO NONCONSERVATIVE PROBLEMS IN ELASTICITY*

MARSHALL SLEMROD†

Abstract. A criterion is presented which assures that a bilinear system may be stabilized with a feedback control which yields the feedback system globally asymptotically stable. Applications to stabilization of columns with follower load are presented.

1. Introduction. This paper deals with stabilization of bilinear control systems of the form

$$(\mathcal{B}) \quad \dot{x}(t) = Ax(t) + u(t)Bx(t)$$

where A and B are constant $m \times m$ matrices, u a scalar control. Feedback stabilization of bilinear systems of this type has recently been considered in the papers of Jacobson [1] and Jurdjevic and Quinn [2]. In [2] Jurdjevic and Quinn have proven the following theorem.¹

THEOREM. *If A has a purely imaginary distinct spectrum and*

$$\text{span}\{Ax, \text{ad}^0(A, B)x, \text{ad}^1(A, B)x, \dots\} = R^m$$

for all $x \in R^m - \{0\}$ then (\mathcal{B}) possesses a stabilizing feedback control $u(x)$.

The proof of the theorem is based on an invariance argument using the Lyapunov function $V(x) = x^T Q x$; Q is the positive definite solution to matrix equation $A^T Q + Q A = 0$. This paper extends Jurdjevic and Quinn's proof and indeed the similar argument given in [1] to reach a further and more easily applicable result (Theorem 2). The main idea is to partition the phase space R^m into a set Ω and its complement Ω' such that $\{0\}$ is the only subset of Ω' which is invariant under the uncontrolled system ($u = 0$). It then suffices to check that

$$\text{span}\{Ax, \text{ad}^0(A, B)x, \text{ad}^1(A, B)x, \dots\} = R^m$$

on Ω , rather than on $R^m - \{0\}$ as in the above theorem.

Stabilization problems for bilinear control systems arise naturally in the study of nonconservative elastic systems. In fact the control and stability of such systems is an important problem in structural engineering (see [3] and [7]). It is for this reason that a thorough discussion of two simple model problems is given in § 2. The reader uninterested in applications may proceed directly to § 3 where the stabilization problem is considered and the main results are given. Section 4 presents applications of the main stabilization result to the model problems. Also the easy applicability of the main result (Theorem 2) is stressed. Section 5 restates

* Received by the editors September 2, 1976, and in revised form March 3, 1977.

† Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, New York 12181. This research was supported in part by the National Science Foundation under Grant MCS76-07012.

¹ See *Notation* below.

the argument originally given by Jacobsen [1] on the optimality of the feedback control. Finally an Appendix provides a simple proof of a result used in the stability analysis.

Notation. For square matrices X, Y we use the notation

$$[X, Y] = XY - YX$$

to denote the standard matrix commutator. Also define $\text{ad}(X, Y) = [X, Y]$, $\text{ad}^k(X, Y) = \text{ad}(X, \text{ad}^{k-1}(X, Y))$, where $\text{ad}^0(X, Y) = Y$.

2. Model problems. In this section, we formulate two simple model problems. Consider the system illustrated in Fig. 1. This system, consisting of two rigid bars of length $l/2$ and two elastic hinges, was introduced by H. Ziegler [3]. It has been used by many investigators for various purposes; it is to be taken as an approximation of a real deformable column. Assume the load P is to follow the deflections with tangency coefficient η (P and η may be time dependent); ϕ_i are the deflection angles of the bars (ϕ_i are assumed to be small); c_i are the elastic constants of the hinges; m_i are the masses of the bars fixed at the distances αl and γl .

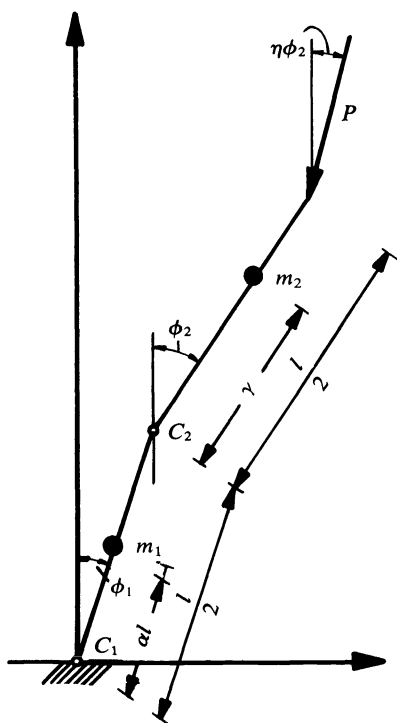


FIG. 1

The linearized equation of motion is

$$(1) \quad M\ddot{y} + Cy + D(t)y = 0$$

where

$$y = \begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix}, \quad M = \begin{bmatrix} (m_1 \alpha^2 + \frac{1}{4} m_2) l_2 & \frac{1}{2} m_2 \gamma l^2 \\ \frac{1}{2} m_2 l^2 \gamma & m_2 \gamma^2 l^2 \end{bmatrix},$$

$$C = \begin{bmatrix} c_1 + c_2 & -c_2 \\ -c_2 & c_2 \end{bmatrix}, \quad D(t) = \begin{bmatrix} -\frac{1}{2} P(t) l & \frac{1}{2} P(t) l \eta(t) \\ 0 & -\frac{1}{2} P(t) l (1 - \eta(t)) \end{bmatrix}.$$

M and C are positive definite matrices.

The case $\eta \equiv 0$ is a double pendulum model of the Euler column; the case $\eta \equiv 1$ is a double pendulum model of a column with follower load. Also, it is easily seen that when $\eta \equiv 0$ and P is constant, all solutions λ of the characteristic equation

$$(2) \quad \det(\lambda^2 M + C + D) = 0$$

are distinct, purely imaginary and nonzero if and only if

$$P < P_0 = \frac{(2c_2 + c_1) - \sqrt{4c_2^2 + c_1^2}}{l}.$$

P_0 is the critical Euler buckling load; $P \geq P_0$ yields the column Lyapunov unstable. Similarly, when $\eta \equiv 1$ and P is a constant, all solutions λ of the characteristic equation (2) are distinct, purely imaginary and nonzero if and only if

$$P < P_{cr} = \frac{c_2}{l} \left\{ \frac{(1 + 4\mu\alpha^2) + 4\gamma^2(1 + \psi) + 4\gamma - \alpha\gamma\sqrt{\mu\psi}}{2\gamma^2 + \gamma} \right\}$$

where $\mu = m_1/m_2$, $\psi = c_1/c_2$. P_{cr} is the critical load; $P \geq P_{cr}$ yields the column Lyapunov unstable. (See the paper of Zyczkowski and Gajewski [4] for a derivation of P_{cr} .)

We now formulate two stabilization problems:

- (PI) If a constant load P^* , $0 < P^* < P_{cr}$ is applied for all $t \geq 0$, find a feedback control law $\eta(t) = \eta(y(t), \dot{y}(t))$ for the loading angle which yields the zero equilibrium of (1) globally asymptotically stable. In order to restrict the angle of loading to a symmetric sector, we require $0 \leq \eta(t) \leq 2$. Also on physical grounds, we desire $\eta(t)$ to be a continuous function of t .
- (PII) If $\eta = 1$ for all $t \geq 0$, i.e., the column with follower load, find a feedback control law $P(t) = P(y(t), \dot{y}(t))$ for a nonnegative loading which yields the zero equilibrium of (1) globally asymptotically stable. We require the loading to be constrained to some region centered at some $P^* < P_{cr}$, i.e., $|P(t) - P^*| \leq \varepsilon$, $P^* > \varepsilon > 0$.

At first glance one might be tempted to try to resolve the above problems by linear time invariant analysis. Such an approach is doomed to failure. The reason is simple. To keep (1) linear and time invariant, one must have η and P constant for all time. In this case, analogous to the special cases with $\eta \equiv 0$ and $\eta \equiv 1$, there is a critical loading $P(\eta)$ such that the system can have two types of behavior:

- (i) $P < P(\eta)$ (the critical loading) and (1) exhibits undamped oscillations,
- (ii) $P \geq P(\eta)$ and (1) is unstable.

(A formula for $P(\eta)$ may be found in [4].)

In either case (i) or (ii), the zero equilibrium is not asymptotically stable. The choice of feedback mechanisms in the above problems will of necessity be nonconstant and (1) will become nonlinear.

We now rewrite the problems in a more convenient form. Let

$$v_1(t) = \eta(t) - 1, \quad v_2(t) = P(t) - P^*$$

$$G_1 = \begin{bmatrix} -1 & 1 \\ 0 & 0 \end{bmatrix}, \quad G_2 = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

From (1) we see that the state equations for (PI) and (PII) are of the form

$$(3) \quad \ddot{y}(t) + Jy(t) + u(t)Hy(t) = 0,$$

where

$$J = M^{-1}C + \frac{1}{2}P^*lM^{-1}G_1,$$

$u(t)$ satisfies constraints $u_1 \leq u(t) \leq u_2$, ($u_1 < 0$, $u_2 > 0$). In (PI) we have

$$H = \frac{1}{2}P^*lM^{-1}G_2, \quad u(t) = v_1(t),$$

$$u_1 = -1, \quad u_2 = +1.$$

In (PII) we have

$$H = \frac{1}{2}lM^{-1}G_1, \quad u(t) = v_2(t),$$

$$u_1 = -\varepsilon, \quad u_2 = +\varepsilon.$$

Also we note that since $P^* < P_{cr}$, J has distinct positive eigenvalues.

Remark 1. Problems (PI) and (PII) are presented as model problems. This does not mean, however, the study of these problems is of purely academic interest. Such models have been proposed by McDonough [7] in his study of stability problems arising in the control of launch vehicles.

3. Stabilization of bilinear systems. In the previous section, we derived a state equation (3), with constrained scalar control, that described both (PI) and (PII). In this section, we explore the stabilization problem for (3) in a more general form. Assuming no confusion to the reader, we use the same notation for the general problem as used in the previous section.

Let

- (i) J be a diagonalizable $n \times n$ matrix with positive eigenvalues;
- (ii) H be an $n \times n$ matrix;
- (iii) $y(t)$ be an n vector;
- (iv) $u(t)$ is a scalar control, $u_1 \leq u(t) \leq u_2$, $u_1 < 0$, $u_2 > 0$.

Remark 2. The obvious modifications of the subsequent theory will allow consideration of unconstrained problems, i.e., $u_1 = -\infty$, and/or $u_2 = +\infty$.

Remark 3. Note that in this general form, distinctness of the eigenvalues of J is *not* required, though in the model problems the eigenvalues are indeed distinct.

Our stabilization problem is as follows:

- (P) Find a continuous feedback control law $u(t) = u(y(t), \dot{y}(t))$, $u_1 \leq u(t) \leq u_2$, so that the zero equilibrium of the bilinear control system

$$(4) \quad \ddot{y}(t) + Jy(t) + u(t)Hy(t) = 0$$

is globally asymptotically stable.

Problems (PI) and (PII) are special cases of (P) with $n = 2$.
We now write (4) as the first order bilinear system on R^{2n}

$$(5) \quad \dot{x}(t) = Ax(t) + u(t)Bx(t)$$

where

$$x = \begin{bmatrix} y \\ \dot{y} \end{bmatrix}, \quad A = \begin{bmatrix} 0 & I \\ -J & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ -H & 0 \end{bmatrix}.$$

Remark 4. The results of [2] are given in terms of (5) where it is assumed that A has purely imaginary distinct spectrum. In this presentation assumption (i) allows consideration of cases where the spectrum of A is *not* distinct.

The construction of the feedback control is the same one as given by Jacobson [1] and Jurdjevic and Quinn [2], namely, we choose $u(t)$ so as to make the time derivative of a Lyapunov function negative-semidefinite. This is done as follows. For J satisfying (i), we know there exists a positive definite matrix Q , so that $A^T Q + QA = 0$. (A simple proof is provided in an Appendix.) Choose as a Lyapunov function $V(x(t)) = \frac{1}{2} x^T(t) Q x(t)$. Along solutions of (5), we have $\dot{V}(x(t)) = u(t) x^T(t) Q B x(t)$. This suggests that we choose

$$(6) \quad u(x) = \begin{cases} u_1 & \text{if } -x^T Q B x < u_1, \\ -x^T Q B x & \text{if } u_1 \leq -x^T Q B x \leq u_2, \\ u_2 & \text{if } -x^T Q B x > u_2. \end{cases}$$

With u given by (6), system (5) possesses unique solutions for all $t \geq 0$. In fact, since $\dot{V}(x(t)) \leq -u^2(x(t)) \leq 0$, we know the zero equilibrium is stable.

To study asymptotic stability, we employ the well known invariance principle of LaSalle [6]. In this case, LaSalle's theorem asserts that all solutions of (5), with u given by (6), will approach the set S . S is defined as the largest invariant set in the set $W = \{x \in R^{2n}; x^T Q B x = 0\}$. Invariance of S , here, is understood to mean that any solution of (5), with u given by (6), starting in S remains in S , for all time t , $-\infty < t < \infty$.

From the definition of S we immediately obtain the result

THEOREM 1. *If $S = \{0\}$, then the feedback control (6) yields the zero equilibrium of (5) globally asymptotically stable.*

While Theorem 1 is quite simple, it has a limitation in its present form. The determination of the set S is dependent on the choice of Q . Since Q is not unique, it is conceivable that a fortuitous choice of Q will yield $S = \{0\}$ or a bad choice of Q might result in a set bigger than $\{0\}$. We will try to develop a sufficient condition that is (i) computationally practical in applications, (ii) independent of the choice of Q .

From the definition of u in (6), we see that for a trajectory $x(t; x_0)$, $x_0 \in S$, we must have

$$(7) \quad x^T(t) Q B x(t) = 0, \quad \dot{x}(t) = Ax(t),$$

for all time t , $-\infty < t < \infty$.

It follows from (7) that for $x_0 \in S$

$$x_0^T e^{A^T t} Q B e^{A t} x_0 = 0$$

or, equivalently,

$$(8) \quad x_0^T Q e^{-A^T t} B e^{A t} x_0 = 0.$$

We now expand (8) in powers of t and obtain the elementary, well known identity

$$(9) \quad x_0^T Q \left[\sum_{k=0}^{\infty} \frac{(-1)^k \text{ad}^k(A, B) t^k}{k!} \right] x_0 = 0.$$

Since (9) must hold for all time t , we see that for $x_0 \in S$, the system

$$(10) \quad x_0^T Q \text{ad}^k(A, B) x_0 = 0, \quad k = 0, 1, \dots,$$

must be satisfied. In addition since $QA = -A^T Q$, we must also have

$$x_0^T Q A x_0 = 0.$$

In the bracket notation, S can now be written as

$$(11) \quad S = \{x_0 \in R^{2n}; x_0^T Q A x_0 = 0, x_0^T Q \text{ad}^k(A, B) x_0 = 0, k = 0, 1, \dots\}.$$

From this form of S , we easily obtain the result of Jurdjevic and Quinn [2], which was stated in the Introduction.

While Jurdjevic and Quinn's result is very neat, it is easy to see that its hypothesis is stronger than that of Theorem 1. This is important. In examples we shall see that it may be a difficult task to apply Jurdjevic and Quinn's result.

Let us return to the definition of S as given in (11). It is often an easy matter to compute a few brackets and see that

$$\{Ax, \text{ad}^0(A, B)x, \text{ad}^1(A, B)x, \dots, \text{ad}^K(A, B)x\}$$

spans R^{2n} for all x in a set $\Omega \subset R^{2n} - \{0\}$. So if S has any nonzero elements, they must be contained in the set Ω' (the complement of Ω). Hence if $\{0\}$ is the only subset of Ω' which is invariant under $e^{A^T t}$, we have $S = \{0\}$. We summarize this argument in the main theorem:

THEOREM 2 (Sufficiency condition for stabilization). *Assume there exists a nonempty set $\Omega \subset R^{2n} - \{0\}$, with the following two properties:*

(i) *for each $x \in \Omega$ there exists an integer K such that*

$$\text{span}\{Ax, \text{ad}^0(A, B)x, \text{ad}^1(A, B)x, \dots, \text{ad}^K(A, B)x\} = R^{2n},$$

(ii) *$\{0\}$ is the only subset of Ω' which is invariant under $e^{A^T t}$, $-\infty < t < \infty$.*

Under these assumptions, the feedback control (6) yields the zero equilibrium of (5) globally asymptotically stable.

We note three important features of Theorem 2. First, it requires K brackets to be computed where K is chosen in examples as small as possible. Second, any positive definite Q satisfying $QA + A^T Q = 0$ will work in the feedback law (6), since Theorem 2 is *not* dependent on the choice of Q . Third, since the only essential property of A is that there exist a positive definite matrix Q so that $QA + A^T Q = 0$, we see that the following corollary holds.

COROLLARY: *If A is such that there exists a positive definite matrix Q so that $QA + A^T Q = 0$ and if the hypotheses of Theorem 2 hold (where R^{2n} is replaced by R^m), then u given by (6) yields the zero equilibrium of (\mathcal{B}) globally asymptotically stable.*

In order to do some examples we need to compute a few brackets. Matrix multiplication yields

$$\begin{aligned} \text{ad}(A, B) &= \begin{bmatrix} -H & 0 \\ 0 & H \end{bmatrix}, & \text{ad}^2(A, B) &= \begin{bmatrix} 0 & 2H \\ JH + HJ & 0 \end{bmatrix}, \\ \text{ad}^3(A, B) &= \begin{bmatrix} JH + 3HJ & 0 \\ 0 & -3JH - HJ \end{bmatrix}. \end{aligned}$$

Computation of higher order brackets can, in general, become a very messy affair. Fortunately to apply Theorem 2 to some examples the above brackets suffice.

4. Examples. We will give two examples. In both examples we fix $m_1 = 2$, $m_2 = 1$, $l = 2$, $\alpha = \frac{1}{2}$, $\gamma = \frac{1}{2}$, $c_1 = 1$, $c_2 = 1$; since $P_{\text{cr}} = \frac{1}{2}(7 - 2\sqrt{2})$, $P^* = 1$ is an allowable simple choice. We easily see

$$\begin{aligned} M &= \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}, & M^{-1} &= \frac{1}{2} \begin{bmatrix} 1 & -1 \\ -1 & 3 \end{bmatrix}, \\ C &= \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}, & J &= \frac{1}{2} \begin{bmatrix} 2 & -1 \\ -4 & 3 \end{bmatrix}. \end{aligned}$$

(PI): For this problem, $H = \frac{1}{2}P^*IM^{-1}G_2$, so we have

$$H = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

If we use our previous bracket computations, we see that for $x = \text{col}(a, b, c, d) \in R^4$,

$$\det(Bx, \text{ad}(A, B)x, \text{ad}^2(A, B)x, \text{ad}^3(A, B)x) = \det \begin{bmatrix} 0 & 0 & 0 & -\frac{1}{2}b \\ 0 & -b & 2d & -6a + 6b \\ 0 & 0 & -b/2 & \frac{3}{2}d \\ -b & d & -2a + 3b & 2c - 2d \end{bmatrix} = \frac{b^4}{4}.$$

Thus $\{Bx, \text{ad}(A, B)x, \text{ad}^2(A, B)x, \text{ad}^3(A, B)x\}$ spans R^4 at every point $x \in R^4$ except at those points whose second component is zero. So $\Omega' = \{x \in R^4; b \neq 0\}$. Now we look for the largest set in Ω' invariant under e^{At} . If $(a, b, c, d) \in \Omega'$ and $e^{At}(a, b, c, d)$ remains in Ω' for all t , then the system

$$\begin{aligned} \dot{x}_1(t) &= x_3, & \dot{x}_2(t) &= x_4, \\ \dot{x}_3(t) &= -x_1 + \frac{1}{2}x_2, & \dot{x}_4(t) &= 2x_1 - \frac{3}{2}x_2 \end{aligned}$$

must be satisfied for all t , $-\infty < t < \infty$, with $x_1(0) = a$, $x_3(0) = c$, $x_4(0) = d$ and $x_2(t) = 0$. But by inspection of the above system, this means $x_1(t) = x_2(t) = x_3(t) = x_4(t) = 0$ for all t , $-\infty < t < \infty$, and $(a, b, c, d) = \{0\}$. Hence the hypotheses of Theorem 2 are the zero equilibrium is globally asymptotically stable with feedback law (6).

Notice that it is *not* true that

$$\{Ax, Bx, \text{ad}(A, B)x, \dots, \text{ad}^3(A, B)x\}$$

spans R^4 for every $x \in R^4 - \{0\}$. As a matter of practical usefulness, then, it seems that Theorem 2 is more advantageous than Jurdjevic and Quinn's theorem since fewer computations are needed.

(PII): In this case

$$H = \frac{1}{2} \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Anyone actually performing the matrix multiplications and determinant evaluation in the previous example will readily see the advantage of having H sparse and diagonal. It thus is convenient to make a change of variables in the control system (4) of the form

$$y = Tw \quad \text{where} \quad T = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

is a similarity transformation which diagonalizes H . System (4) becomes, in the new dependent variable w ,

$$\ddot{w} + J_1 w + u(t)H_1 w = 0$$

where

$$J_1 = T^{-1}JT = \frac{1}{2} \begin{bmatrix} 0 & 2 \\ -1 & 5 \end{bmatrix}, \quad H_1 = T^{-1}HT = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}.$$

Now we can proceed as before to see if

$$A_1 = \begin{bmatrix} 0 & I \\ -J_1 & 0 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 0 & 0 \\ -H_1 & 0 \end{bmatrix}$$

satisfy the hypotheses of Theorem 2.

We again use our bracket calculations to see that for $x = \text{col}(a, b, c, d) \in R^4$,

$$\det(B_1 x, \text{ad}(A_1, B_1)x, \text{ad}^2(A_1, B_1)x, \text{ad}^3(A_1, B_1)x) =$$

$$\det \begin{bmatrix} 0 & 0 & 0 & -b \\ 0 & -b & -2d & \frac{3}{2}a - 10b \\ 0 & 0 & -b & 3d \\ b & -d & \frac{1}{2}a - 5b & -\frac{1}{2}c + 10d \end{bmatrix} = -b^4.$$

Thus again $\{B_1 x, \text{ad}(A_1, B_1)x, \text{ad}^2(A_1, B_1)x, \text{ad}^3(A_1, B_1)x\}$ spans R^4 at every point $x \in R^4$, except at those points whose second component is zero. So $\Omega' = \{x \in R^4; b = 0\}$. Now we look for the largest set in Ω' invariant under $e^{A_1 t}$. If $(a, b, c, d) \in \Omega'$, and $e^{A_1 t}(a, b, c, d)$ remains in Ω' for all t , the system

$$\dot{x}_1 = x_3, \quad \dot{x}_2 = x_4, \quad \dot{x}_3 = -x_2, \quad \dot{x}_4 = \frac{1}{2}x_1 - \frac{5}{2}x_2$$

must be satisfied for all t , $-\infty < t < \infty$, with $x_1(0) = a$, $x_3(0) = c$, $x_4(0) = d$, and $x_2(t) = 0$. Inspection of the above system shows $x_1(t) = x_2(t) = x_3(t) = x_4(t) = 0$.

Again the hypotheses of Theorem 2 are satisfied. Thus for (PII), the zero equilibrium is globally asymptotically stable with feedback control (6).

Once again we note that $\{A_1x, B_1x, \text{ad}(A_1, B_1)x, \dots, \text{ad}^3(A_1, B_1)x\}$ does not span R^4 for every $x \in R^4 - \{0\}$. More computations will be needed to show the applicability of Jurdjevic and Quinn's theorem, if it is applicable at all.

5. Optimality. One further question may be posed regarding the feedback control (6); i.e., in what sense does (6) provide an optimal control? The answer has been provided by Jacobson [1] in the case when u is unconstrained. For completeness we derive a result analogous to Jacobson's for our constrained problem.

Define the cost functional

$$I(x^0, u) = \int_0^\infty \{q(x) + u^2\} dt$$

where $q(x)$ is positive definite function of x . For initial data x^0 , define

$$V(x^0) = \min_{u_1 \leq u \leq u_2} I(x^0, u).$$

Bellman's equation of dynamic programming asserts that V must satisfy

$$\min_{u_1 \leq u \leq u_2} [\nabla V(x) \cdot Ax + u \nabla V \cdot Bx + q(x) + u^2] = 0.$$

A solution of the Bellman equation is provided by

$$V(x) = x^T Qx$$

where $u = u^*$, where u^* is given by (6). From Theorem 2 we know that the choice $u = u^*$ yields the origin globally asymptotically stable. Also the Bellman equation shows that $q(x)$ satisfies

$$(12) \quad q(x) = \begin{cases} -2u_1(x^T QBx) - u_1^2 & \text{if } -x^T QBx < u_1, \\ (x^T QBx)^2 & \text{if } u_1 \leq -x^T QBx \leq u_2, \\ -2u_2(x^T QBx) - u_2^2 & \text{if } -x^T QBx > u_2. \end{cases}$$

We now show that this choice of u^* actually provides an optimal control. Let $u(t)$ be any control, $u_1 \leq u(t) \leq u_2$, which causes $x(t) \rightarrow 0$ as $t \rightarrow \infty$. For $V(x) = x^T Qx$ we see that

$$0 = 2u^*x^T QBx + q(x) + u^{*2} \leq 2ux^T QBx + q(x) + u^2.$$

Since

$$I(x^0, u) = V(x^0) + \int_0^\infty \{2ux^T QBx + q(x) + u^2\} dt$$

we find

$$I(x^0, u^*) \leq I(x^0, u),$$

so u^* is an optimal control. We summarize in the following theorem.

THEOREM 3 (Optimality). *If the hypotheses of Theorem 2 hold, then the feedback control (6) minimizes the cost functional*

$$I(x^0, u) = \int_0^\infty \{q(x) + u^2\} dt$$

where $q(x)$ is given by (12) and u is in the class of control functions, $u_1 \leq u \leq u_2$, which causes $x(t) \rightarrow 0$ as $t \rightarrow \infty$.

Appendix. In this Appendix a simple proof is given of the following theorem which was used in § 3.

THEOREM. *For J a diagonalizable $n \times n$ matrix with positive eigenvalues, there exists a $2n \times 2n$ positive definite matrix Q such that $A^T Q + QA = 0$ where*

$$A = \begin{bmatrix} 0 & J \\ -J & 0 \end{bmatrix}.$$

The theorem was originally given in [5]. The proof given here is due to an anonymous referee.

Proof. First observe that every symmetric solution Q of $A^T Q + QA = 0$ has the form

$$(A.1) \quad Q = \begin{bmatrix} XJ & 0 \\ 0 & X \end{bmatrix},$$

where X is a symmetric $n \times n$ matrix satisfying

$$(A.2) \quad XJ = J^T X.$$

By hypotheses we know $J = S^{-1}DS$, where D is a positive definite diagonal matrix. Hence (A.2) is equivalent to

$$(A.3) \quad YD = DY, \quad X = S^T YS.$$

(A.3) is satisfied for any positive definite diagonal matrix Y . Since $X = S^T YS$ and $XJ = S^T YDS$ are symmetric and positive definite, Q given by (A.1) is a symmetric positive definite solution of $A^T Q + QA = 0$.

Acknowledgments. I would like to thank Dr. D. H. Jacobson and Professor V. Jurdjevic for sending me preprints of their papers. Also, I thank Professors D. Drew and Z. Schuss for several stimulating discussions. Finally, I thank the referees for their valuable criticism which aided me in improving an earlier version of this paper.

REFERENCES

- [1] D. H. JACOBSON, *Stabilization and optimal control of nonlinear systems homogeneous-in-the-input*, Proceedings of Conference on Directions in Decentralized Control, Many-Person Optimization and Large-Scale Systems, (Boston, Massachusetts, September 1-3, 1975), Plenum Press, 1976.
- [2] V. JURDJEVIC AND J. QUINN, *Controllability and stability*, to appear.
- [3] H. ZIEGLER, *Principles of Structural Stability*, Blaisdell, Waltham, MA, 1968.
- [4] M. ZYCKOWSKI AND A. GAJEWSKI, *Optimal structural design in non-conservative problems of elastic stability*, Instability of Continuous Systems, H. Liepholz, ed., Springer-Verlag, New York, 1971, pp. 295-301.

- [5] J. WALKER, *On the application of Liapunov's direct method to linear lumped parameter elastic systems*, J. Appl. Mech., 41 (1974), pp. 278–284.
- [6] J. P. LASALLE, *Stability theory for ordinary differential equations*, J. Differential Equations, 4 (1968), pp. 57–65.
- [7] G. F. McDONOUGH, *Stability problems in the control of Saturn launch vehicles*, Dynamic Stability of Structures, G. Herrmann, ed., Pergamon Press, New York, 1967, pp. 113–128.

ON EXISTENCE OF A NASH EQUILIBRIUM POINT IN N -PERSON NONZERO SUM STOCHASTIC DIFFERENTIAL GAMES*

KENKO UCHIDA†

Abstract. Using the technique of Davis and Varaiya, we give an existence theorem for a Nash equilibrium point in N -person nonzero sum stochastic differential games. It is shown that if the Nash condition (generalized Isaacs condition) holds there is a Nash equilibrium point in feedback strategies. Also we discuss two special cases where the Nash condition holds.

1. Introduction. Using the technique of Davis and Varaiya [1], [2], we give an existence theorem for a Nash equilibrium point in N -person nonzero sum stochastic differential games. It is shown that if the Nash condition (see Definition 2) holds there is a Nash equilibrium point in feedback strategies. Also we discuss two special cases in which the Nash condition holds.

It is an essential point of the Davis and Varaiya technique that analogues of the time derivation and gradient of the value function are constructed using a martingale method. Consequently, we can obtain the optimal value directly by optimizing the Hamiltonian at each point. So it is not necessary to discuss the existence of the solution of the second order parabolic equations which are satisfied by the optimal value functions (cf. [3]). These properties are attractive from the differential game theoretic point of view. Recently, using this method, Elliott obtained the good existence theorem of a saddle point in a two-person zero sum stochastic differential game [4]. In this paper we discuss the nonzero sum case by the same approach. In the zero sum case, the concepts of upper and lower values play fundamental roles. On the other hand, in the nonzero sum game we can not use these useful concepts, that is, we must construct an equilibrium point. This makes the N -person nonzero sum game discussed here more difficult to solve.

Consider the following stochastic functional differential system:

$$(1) \quad dx_t = f(t, x, u_1, \dots, u_N) dt + \sigma(t, x) dB_t,$$

where $t \in [0, 1]$, B_t is a Brownian motion in R^m . Let C be the space of continuous functions from $[0, 1]$ to R^m . x denotes a member of C and x_t denotes the value of x at t . Initial value $x_0 \in R^m$ is given to (1). It is noted that the system function f depends on the past history of the process, i.e., $\{x_s: s \leq t\}$. The player i , $i = 1, \dots, N$, chooses a feedback control $u_i(t, x)$ with values in a compact metric space U_i . Then corresponding to this choice of control, player i incurs a cost of the form:

$$(2) \quad P_i(u_1, \dots, u_N) = E \left\{ g_i(x_1) + \int_0^1 h_i(t, x, u_1, \dots, u_N) dt \right\},$$

where

- (i) g_i and h_i are real-valued,
- (ii) $0 \leq g_i \leq k$ and $0 \leq h_i \leq k$ for some constant k ,

* Received by the editors September 9, 1976, and in revised form April 19, 1977.

† Department of Electrical Engineering, Waseda University, Tokyo, Japan.

- (iii) g_i and h_i satisfy the measurability properties described in the next section.

It is the objective of each player to minimize his own cost.

2. Preliminaries. The situation treated below is similar to that of Davis and Varaiya, and Elliott, so we continue their notations with slight modifications.

\mathcal{F}_t is the σ -field of C generated by $\{x_s: x \in C, s \leq t\}$. The Brownian motion B_t is separable and defined on an underlying probability space $(\Omega, \mathcal{A}, \mu)$. \mathcal{D} is the σ -field of subsets D of $[0, 1] \times C$ having the property that the section of D at time t is in \mathcal{F}_t for each t and the section of D at x is Lebesgue measurable for each x .

$\sigma(t, x)$ is an $m \times m$ matrix for each $(t, x) \in [0, 1] \times C$ whose elements $\sigma_{ij}(t, x)$ are \mathcal{D} -measurable and satisfy a uniform Lipschitz condition in x . The inverse matrix $\sigma^{-1}(t, x)$ is assumed to exist and be bounded for $(t, x) \in [0, 1] \times C$. Then the equation

$$dx_t = \sigma(t, x) dB_t, \quad x_0 \in R^m,$$

has a unique solution x_t and it induces a measure P_0 on its sample space (C, \mathcal{F}_1) according to the formula

$$P_0(A) = \mu\{\omega: x(\omega) \in A\}, \quad A \in \mathcal{F}_1.$$

Let Φ denote the set of \mathcal{D} -measurable functions $\phi: [0, 1] \times C \rightarrow R^m$ such that

$$\|\phi(t, x)\| \leq K(1 + \|x\|),$$

where $\|\cdot\|$ is the uniform norm in C . Write a_t for the matrix $\sigma(t, x)\sigma'(t, x)$ and for $\phi \in \Phi$ define

$$\zeta(\phi) = \int_0^1 \phi_t \cdot a_t^{-1} \cdot dx_t - \frac{1}{2} \int_0^1 \phi_t \cdot a_t^{-1} \phi_t dt,$$

where $\phi_t = \phi(t, x)$. Let the measure P_ϕ be defined by $P_\phi(A) = \int_A \exp(\zeta(\phi)) dP_0$, $A \in \mathcal{F}_1$.

LEMMA 1 [6].

- (i) P_ϕ is a probability measure;
- (ii) P_ϕ is mutually absolutely continuous with respect to P_0 ;
- (iii) $\{w_t, t \in [0, 1]\}$ is a Brownian motion under P_ϕ , where

$$dw_t = \sigma^{-1}(t, x)(dx_t - \phi(t, x) dt).$$

Let \mathcal{U}_i be the σ -field of Borel sets of U_i . An admissible feedback control for player i is a measurable function

$$u_i: ([0, 1] \times C, \mathcal{D}) \rightarrow (U_i, \mathcal{U}_i).$$

The set of such admissible controls for player i is denoted by \mathcal{M}_i . It is assumed that the Borel σ -field \mathcal{R}^m is always defined for each R^m .

The function f is assumed to satisfy

- (i) $f: [0, 1] \times C \times U_1 \times \cdots \times U_N \rightarrow R^m$ is measurable with respect to the σ -field $\mathcal{D} \otimes U_1 \otimes \cdots \otimes U_N$,
- (ii) for each $(t, x) \in [0, 1] \times C$, $f(t, x, \cdot, \cdot, \cdot, \cdot)$ is continuous on $U_1 \times \cdots \times U_N$,

- (iii) there exists a constant K such that for all $(t, x, u_1, \dots, u_N) \in [0, 1] \times C \times U_1 \times \dots \times U_N$,

$$|f(t, x, u_1, \dots, u_N)| \leq K(1 + \|x\|).$$

The cost rate $h_i: [0, 1] \times C \times U_1 \times \dots \times U_N \rightarrow R^+$ satisfies the same form of conditions as (i)–(ii).

For $(u_1, \dots, u_N) \in \mathcal{M}_1 \times \dots \times \mathcal{M}_N$ and $(t, x) \in [0, 1] \times C$, define

$$f^{u_1 \dots u_N}(t, x) = f(t, x, u_1(t, x), \dots, u_N(t, x)),$$

$$h_i^{u_1 \dots u_N}(t, x) = h_i(t, x, u_1(t, x), \dots, u_N(t, x)).$$

We see $f^{u_1 \dots u_N} \in \Phi$. Denoting $P_{u_1 \dots u_N} = P_{f^{u_1 \dots u_N}, P_{u_1 \dots u_N}}$ is the measure corresponding to the solution of (1) in the sense that under $P_{u_1 \dots u_N}$,

$$dx_t = f(t, x, u_1(t, x), \dots, u_N(t, x)) dt + \sigma(t, x) dB_t,$$

where B_t is the Brownian motion.

Let $E_{u_1 \dots u_N}$ denote the expectation with respect to $P_{u_1 \dots u_N}$. Then the cost for player i , corresponding to $(u_1, \dots, u_N) \in \mathcal{M}_1 \times \dots \times \mathcal{M}_N$, which was defined by (2) formally, is now

$$P_i(u_1, \dots, u_N) = E_{u_1 \dots u_N} \left\{ g_i(x_1) + \int_0^1 h_i(t, x, u_1(t, x), \dots, u_N(t, x)) dt \right\}.$$

DEFINITION 1. Admissible controls $(u_1^*, \dots, u_N^*) \in \mathcal{M}_1 \times \dots \times \mathcal{M}_N$ are said to form a *Nash equilibrium point* if for each $i = 1, \dots, N$ and for all $u_i \in \mathcal{M}_i$,

$$(3) \quad P_i(u_1^*, \dots, u_N^*) \leq P_i(u_1^*, \dots, u_{i-1}^*, u_i, u_{i+1}^*, \dots, u_N^*).$$

3. Nash equilibrium point. We start by quoting Theorem 3 of [5] in a form adapted to our problem.

THEOREM 1. $u^* = (u_1^*, \dots, u_N^*) \in \mathcal{M}_1 \times \dots \times \mathcal{M}_N$ is a *Nash equilibrium point* if for each i there exist a constant J_i^* and processes η_i^i and ξ_i^i with values in R and R^m respectively, adapted to \mathcal{F}_t , such that

- (i) $\int_0^1 |\xi_i^i|^2 dt < \infty$ a.s. (P_0),
- (ii) $E \int_0^1 \xi_i^i dx_t = 0$,
- (iii) $g_i(x_1) = J_i^* + \int_0^1 \eta_i^i dt + \int_0^1 \xi_i^i dx_t$ a.s. (P_0),

and

$$\begin{aligned} & \eta_i^i + \xi_i^i \cdot f(t, x, u_1^*(t, x), \dots, u_{i-1}^*(t, x), u_i(t, x), u_{i+1}^*(t, x), \dots, u_N^*(t, x)) \\ & + h_i(t, x, u_1^*(t, x), \dots, u_{i-1}^*(t, x), u_i(t, x), u_{i+1}^*(t, x), \dots, u_N^*(t, x)) \\ & \geq \eta_i^i + \xi_i^i \cdot f(t, x, u^*(t, x)) + h_i(t, x, u^*(t, x)) = 0, \end{aligned}$$

for almost all (t, x) and all $u_i \in \mathcal{M}_i$.

Consider the following auxiliary optimal control problem for player i . That is, for each $(u_1, \dots, u_{i-1}, u_{i+1}, \dots, u_N) \in \mathcal{M}_1 \times \dots \times \mathcal{M}_{i-1} \times \mathcal{M}_{i+1} \times \dots \times \mathcal{M}_N$,

$$P_i^* = \inf_{u_i \in \mathcal{M}_i} P_i(u_1, \dots, u_N).$$

For this optimal control problem, the following lemma can be deduced directly from the result of Davis and Varaiya [1], [2].

LEMMA 2. For each $(u_1, \dots, u_{i-1}, u_{i+1}, \dots, u_N) \in \mathcal{M}_1 \times \dots \times \mathcal{M}_{i-1} \times \mathcal{M}_{i+1} \times \dots \times \mathcal{M}_N$ there exist processes ΛV_t^i and ∇V_t^i with values in R and R^m respectively, adapted to \mathcal{F}_t such that

$$\int_0^1 |\nabla V_t^i|^2 dt < \infty \quad \text{a.s. } (P_0),$$

$$E \int_0^1 |\Lambda V_t^i| dt < \infty,$$

$$g_i(x_1) = P_i^* + \int_0^1 \Lambda V_t^i dt + \int_0^1 \nabla V_t^i dx_t \quad \text{a.s. } (P_0),$$

and for all $u_i \in \mu_i$,

$$\begin{aligned} & \Lambda V_t^i + \nabla V_t^i f(t, x, u_1(t, x), \dots, u_N(t, x)) \\ & \quad + h_i(t, x, u_1(t, x), \dots, u_N(t, x)) \geq 0, \quad \text{for almost all } (t, x). \end{aligned}$$

In particular, $E \int_0^1 \nabla V_t^i dx_t = 0$.

Now, for $(t, x, p_i) \in [0, 1] \times C \times R^m$, we introduce the Hamiltonian for each $i = 1, 2, \dots, N$:

$$\begin{aligned} H_i(t, x, p_i; u_1, \dots, u_N) \\ = p_i f(t, x, u_1, \dots, u_N) + h_i(t, x, u_1, \dots, u_N). \end{aligned}$$

DEFINITION 2. We say the Nash condition holds if there exists a function

$$u_i^*: ([0, 1] \times C \times R^{mN}, \mathcal{D} \otimes \mathcal{R}^{mN}) \rightarrow (U_i, \mathcal{U}_i),$$

for all $i = 1, 2, \dots, N$ such that for each $(t, x, u_1, \dots, u_N, p_1, \dots, p_N) \in [0, 1] \times C \times U_1 \times \dots \times U_N \times R^{mN}$,

$$\begin{aligned} H_i(t, x, p_i; u_1^*(t, x, p), \dots, u_N^*(t, x, p)) \\ \leq H_i(t, x, p_i; u_1^*(t, x, p), \dots, u_{i-1}^*(t, x, p), u_i, \\ u_{i+1}^*(t, x, p), \dots, u_N^*(t, x, p)), \end{aligned}$$

where $p = (p_1, \dots, p_N)$.

We can now state our main result.

THEOREM 2. If the Nash condition holds, there is a Nash equilibrium point.

Proof. Assume that there is no Nash equilibrium point; then, from Theorem 1, for each $u = (u_1, \dots, u_N) \in \mathcal{M}_1 \times \dots \times \mathcal{M}_N$ there exist $i \in \{1, 2, \dots, N\}$ and $u_i^0 \in \mathcal{M}_i$ such that

$$\begin{aligned} H_i(t, x, \xi_i^i; u(t, x)) \\ > H_i(t, x, \xi_i^i; u_1(t, x), \dots, u_{i-1}(t, x), u_i^0(t, x), u_{i+1}(t, x), \dots, u_N(t, x)), \end{aligned}$$

for a set of (t, x) of positive measure, and for each J_i^* and each process η_i^i and ξ_i^i satisfying (i), (ii) and (iii) of Theorem 1 and the inequality: $\eta_i^i + \xi_i^i f(t, x, u(t, x)) + h_i(t, x, u(t, x)) \geq 0$ for almost all (t, x) . Here note that the constant P_i^* and the processes ΛV_t^i and ∇V_t^i in Lemma 1 satisfy such hypotheses (i), (ii) and (iii) of Theorem 1 and the inequality just above. So, taking $J_i^* = P_i^*$, $\eta_i^i = \Lambda V_t^i$ and $\xi_i^i = \nabla V_t^i$,

we have the following statement (we call it *Statement A*); that is, for each $u \in \mathcal{M}_1 \times \cdots \times \mathcal{M}_N$ there exist $i \in \{1, 2, \cdots, N\}$, $u_i^0 \in \mathcal{M}_i$ and ∇V_t^i such that

$$H_i(t, x, \nabla V_t^i; u(t, x)) \\ > H_i(t, x, \nabla V_t^i; u_1(t, x), \cdots, u_{i-1}(t, x), u_i^0(t, x), u_{i+1}(t, x), \cdots, u_N(t, x)),$$

for a set of (t, x) of positive measure.

On the other hand, if the Nash condition holds, for each (t, x) taking $p_i = \tilde{p}_i(t, x)$, where \tilde{p}_i is an arbitrary \mathcal{D} -measurable function with value in R^m , and $u_i(t, x, \tilde{p}(t, x)) = u^+(t, x)$, $\tilde{p} = (\tilde{p}_1, \cdots, \tilde{p}_N)$ in the Nash condition, we have the following statement (we call it *Statement B*); that is, there exists $u^+ = (u_1^+, \cdots, u_N^+) \in \mathcal{M}_1 \times \cdots \times \mathcal{M}_N$ for each $i \in \{1, 2, \cdots, N\}$, each $u_i \in \mathcal{M}_i$ and each \mathcal{D} -measurable function \tilde{p}_i such that

$$H_i(t, x, \tilde{p}_i(t, x); u^+(t, x)) \\ \leq H_i(t, x, \tilde{p}_i(t, x); u_1^+(t, x), \cdots, u_{i-1}^+(t, x), u_i(t, x), u_{i+1}^+(t, x), \cdots, u_N^+(t, x)),$$

for all (t, x) .

Statement A is the negation of Statement B. Therefore, that Statement A holds implies that the Nash condition does not hold. Thus we obtain the theorem.

4. On the Nash condition. We show two cases in which the Nash condition holds. First, suppose

$$(H_1) \quad (i) \quad f(t, x, u_1, \cdots, u_N) = \sum_{j=1}^N f_j(t, x, u_j), \\ (ii) \quad h_i(t, x, u_1, \cdots, u_N) = \sum_{j=1}^N h_{ij}(t, x, u_j),$$

for each $i = 1, \cdots, N$. In this case the Hamiltonian can be written in a form:

$$(4) \quad H_i(t, x, p_i; u_1, \cdots, u_N) = \sum_{j=1}^N H_{ij}(t, x, p_i; u_j).$$

Consider $\inf_{u_i \in U_i} H_{ii}(t, x, p_i; u_i)$. For each (t, x, p_i) , H_{ii} is continuous on the compact metric space U_i , so the infimum is attained:

$$H_{ii}^*(t, x, p_i) = \min_{u_i \in U_i} H_{ii}(t, x, p_i; u_i).$$

Let S_i be a countable dense subset of U_i . H_{ii} is continuous in u_i for each (t, x, p_i) , so that for any $a \in R$,

$$\{(t, x, p_i): H_{ii}^*(t, x, p_i) < a\} \\ = \bigcup_{u_i \in S_i} \{(t, x, p_i): H_{ii}(t, x, p_i; u_i) < a\}.$$

Hence H_{ii}^* is measurable with respect to $\mathcal{D} \otimes \mathcal{R}^m$. An implicit function lemma of Beneš [7] shows that there is a measurable function $u_i^*: ([0, 1] \times C \times R^m,$

$\mathcal{D} \otimes \mathcal{R}^m) \rightarrow (U_i, \mathcal{U}_i)$ such that

$$(5) \quad \begin{aligned} H_{ii}^*(t, x, p_i) &= H_{ii}(t, x, p_i; u_i^*(t, x, p_i)) \\ &\leq H_{ii}(t, x, p_i; u_i), \end{aligned}$$

for all (t, x, p_i, u_i) . If we take the sum $\sum_{j \neq i}^N H_{ij}(t, x, p_j; u_j^*(t, x, p_j))$ by using the controls $u_j^*(t, x, p_j)$ for all $j = 1, \dots, N$, and add this sum to both sides of (5), then we obtain the Nash condition from (4).

Next, suppose the case with following convexity:

- (H₂) (i) U_i is a convex set for all $i = 1, \dots, N$,
 (ii) f is a convex function on U_j for fixed $(t, x, u_1, \dots, u_{j-1}, u_{j+1}, \dots, u_N)$ and all $j = 1, \dots, N$,
 (iii) h_i is a strictly convex function on U_i for fixed $(t, x, u_1, \dots, u_{i-1}, u_{i+1}, \dots, u_N)$ and all $i = 1, \dots, N$.

In this case, using the Theorem of Nikaido and Isoda [8], we can show that there is a point (u_1^0, \dots, u_N^0) in $U_1 \times \dots \times U_N$ such that for each $i = 1, \dots, N$ and each $(t, x, u_1, \dots, u_N, p) \in [0, 1] \times C \times U_1 \times \dots \times U_N \times R^{mN}$,

$$(6) \quad \begin{aligned} H_i(t, x, p_i; u_1^0, \dots, u_N^0) \\ \leq H_i(t, x, p_i; u_1^0, \dots, u_{i-1}^0, u_i, u_{i+1}^0, \dots, u_N^0). \end{aligned}$$

Now write $H_i^0(t, x, p_i) = H_i(t, x, p_i; u_1^0, \dots, u_N^0)$. (6) can be interpreted as follows:

$$H_i^0(t, x, p_i) = \min_{u_i \in U_i} H_i(t, x, p_i; u_1^0, \dots, u_{i-1}^0, u_i, u_{i+1}^0, \dots, u_N^0)$$

for each (t, x, p_i) . Repeating the same argument as before, the implicit function lemma of Beneš shows there exists a $\mathcal{D} \otimes \mathcal{R}^m$ -measurable function $u_i^*(t, x, p_i)$ such that

$$H_i^0(t, x, p_i) = H_i(t, x, p_i; u_1^0, \dots, u_{i-1}^0, u_i^*(t, x, p_i), u_{i+1}^0, \dots, u_N^0),$$

for each (t, x, p_i) . From the strictly convexity of h_i , $H_i(t, x, p_i; u_1^0, \dots, u_{i-1}^0, \cdot, u_{i+1}^0, \dots, u_N^0)$ is a strictly convex function on U_i and the minimum is attained by the unique point u_i^0 , so we obtain $u_i^0 = u_i^*(t, x, p_i)$ for fixed (t, x, p_i) . Thus the Nash condition holds.

The above arguments are now summarized as follows:

COROLLARY 1. *Under either assumption (H₁) or (H₂), there is a Nash equilibrium point.*

5. Extension to other solution conceptions.¹ The technique of the proof for Theorem 2 can be extended to show existence for the other solutions dealt with in [5]. According to [5], introduce the following solution conceptions.

DEFINITION 3. Admissible control $u^* = (u_1^*, \dots, u_N^*) \in \mathcal{M}_1 \times \dots \times \mathcal{M}_N$ is

(a) *efficient* if there is no $u = (u_1, \dots, u_N) \in \mathcal{M}_1 \times \dots \times \mathcal{M}_N$ such that

$$P_i(u_1, \dots, u_N) < P_i(u_1^*, \dots, u_N^*),$$

for all $i \in \{1, 2, \dots, N\}$,

¹ The results of this section were suggested by the reviewer.

(b) *in the core* if there is no $S \subset \{1, 2, \dots, N\}$ and no $u = (u_1, \dots, u_N) \in \mathcal{M}_1 \times \dots \times \mathcal{M}_N$ such that

$$P_i(u_S^*, u_S) < P_i(u_1^*, \dots, u_N^*),$$

for all $i \in S$, where $u_S^* = \{u_i^*, i \in \bar{S}\}$, $u_S = \{u_i, i \in S\}$, and \bar{S} is the complement of S .

THEOREM 3. *There is an efficient point if there exist $\lambda = (\lambda_1, \dots, \lambda_N) \in R^N$, $\lambda \geq 0$, $\lambda \neq 0$ and a function*

$$u_i^*: ([0, 1] \times C \times R^{mN}, \mathcal{D} \otimes \mathcal{R}^{mN}) \rightarrow (U_i, \mathcal{U}_i),$$

for all $i = 1, 2, \dots, N$ such that for each $(t, x, u_1, \dots, u_N, p_1, \dots, p_N) \in [0, 1] \times C \times U_1 \times \dots \times U_N \times R^{mN}$,

$$\begin{aligned} & \sum_{i=1}^N \lambda_i H_i(t, x, p_i: u_1^*(t, x, p), \dots, u_N^*(t, x, p)) \\ & \leq \sum_{i=1}^N \lambda_i H_i(t, x, p_i: u_1, \dots, u_N) \end{aligned}$$

where $p = (p_1, \dots, p_N)$.

Proof. Assume that there is no efficient point. Then using part (a) of Theorem 4 in [5], it can be shown that this assumption contradicts the condition above as in the proof of Theorem 2.

The following result is the combination of Theorem 2 and Theorem 3.

THEOREM 4. *Suppose that the Nash condition holds and that there exists $\lambda = (\lambda_1, \dots, \lambda_N) \in R^N$, $\lambda \geq 0$, $\lambda \neq 0$ such that*

$$\begin{aligned} & \sum_{i=1}^N \lambda_i H_i(t, x, p_i: u_1^*(t, x, p), \dots, u_N^*(t, x, p)) \\ & \leq \sum_{i=1}^N \lambda_i H_i(t, x, p_i: u_1, \dots, u_N), \end{aligned}$$

for each $(t, x, u_1, \dots, u_N, p_1, \dots, p_N) \in [0, 1] \times C \times U_1 \times \dots \times U_N \times R^{mN}$; then there is an efficient Nash equilibrium point.

Finally, using part (a) of Theorem 6 in [5], we can obtain the result for the core.

THEOREM 5. *Suppose that the Nash condition holds and that for each $S \subset \{1, 2, \dots, N\}$ there exist constant $\lambda_i^S \geq 0$, $i \in S$, not all zero, such that for each $(t, x, u_1, \dots, u_N, p_1, \dots, p_N) \in [0, 1] \times C \times U_1 \times \dots \times U_N \times R^{mN}$,*

$$\begin{aligned} & \sum_{i \in S} \lambda_i^S H_i(t, x, p_i: u_1^*(t, x, p), \dots, u_N^*(t, x, p)) \\ & \leq \sum_{i \in S} \lambda_i^S H_i(t, x, p_i: u_S^*(t, x, p), u_S); \end{aligned}$$

then there is an admissible point in the core.

Acknowledgments. The author is indebted to Professor Shimemura for valuable discussions and also thanks Associate Professor Matsumoto for pointing out the Elliott work [4].

The author is very grateful to the reviewer for pointing out a logical error in the previous version of the proof of the main theorem.

REFERENCES

- [1] M. H. A. DAVIS AND P. P. VARAIYA, *Dynamic programming conditions for partially observable stochastic systems*, this Journal, 11 (1973), pp. 226–261.
- [2] M. H. A. DAVIS, *On the existence of optimal policies in stochastic control*, this Journal, 11 (1973), pp. 587–594.
- [3] A. FRIEDMAN, *Stochastic differential games*, J. Differential Equations, 11 (1972), pp. 79–108.
- [4] R. ELLIOTT, *The existence of the value in stochastic differential games*, this Journal, 14 (1976), pp. 85–94.
- [5] P. P. VARAIYA, *N-player stochastic differential games*, this Journal, 14 (1976), pp. 538–545.
- [6] V. E. BENEŠ, *Existence of optimal stochastic control laws*, this Journal, 9 (1971), pp. 446–475.
- [7] ———, *Existence of optimal strategies based on specific information for a class of stochastic decision problems*, this Journal, 8 (1970), pp. 179–188.
- [8] H. NIKAIDO AND K. ISODA, *Note on non-cooperative convex games*, Pacific J. Math., 5 (1955), pp. 807–815.

RATES OF CONVERGENCE FOR SEQUENTIAL MONTE CARLO OPTIMIZATION METHODS

HAROLD J. KUSHNER†

Abstract. Sequential Monte Carlo methods of the stochastic approximation (SA) type, with and without constraints, are discussed. The rates of convergence are derived, and the quantities upon which the rates depend, are discussed. Let $\{X_n\}$ denote the SA sequence and define $U_n = (n+1)^\beta X_n$ for a suitable $\beta > 0$. The $\{U_n\}$ are interpolated into a natural continuous time process, and weak convergence theory is applied to develop the properties of the tails of the sequence. The technique has a number of advantages over past approaches—advantages which are discussed in the paper. It gives more insight (and is apparently more readily generalizable) than do other approaches—and suggests ways of improving the convergence. The particular “dynamical” nature of the approach allows one to say more about the “tail” process—and to do more “decision” (or “control”) analysis with it.

1. Introduction. The subject of stochastic approximation (SA) for unconstrained systems has been well developed in many respects over the past 25 years. See, e.g., the references in Wasan [1], and also Ljung [2], [3]. The treatment of SA under constraints is relatively recent; see Fabian [4], Kushner [6], Kushner and Gavin [5], Kushner and Sanvicente [7], [8], Kushner and Kelmanson [9], and Kushner [10]. The SA problem (with or without constraints) occurs when one wishes to choose a parameter $x \in R^r$ (Euclidean r -space) of a system which (at least locally) minimizes a scalar valued performance function $f(x)$ (without or under constraints), but where the form of $f(\cdot)$ is unknown (as it usually is in complex control problems), and where only noise corrupted measurements of the performance can be made, at various selected parameter settings. The algorithms give a sequence of parameter values $\{X_n\}$ which converges to a local minimum in some statistical sense. *The subject is a stochastic Monte-Carlo form of the general computational problem of nonlinear programming*, and has numerous applications to control theory and practice in the areas of optimization, identification and tracking.

All of the previous works on the constrained problem treat only the fact of convergence. Here we give results on rates of convergence, and obtain some new results for the unconstrained problem also. In particular, we will show that when suitably scaled and interpolated, the SA process can be approximated by a linear diffusion process, in the sense of weak convergence. The scaling and properties of the diffusion give the rates of convergence, and much interesting additional information as well. Some of the advantages will be made clear below.

In § 2, the unconstrained algorithm is introduced. Sections 3 and 4 introduce a Lagrangian and augmented penalty function algorithms. The unconstrained problem is further developed in § 5, and the theorem stated. Section 6 contains some background on weak convergence of a sequence of probability measures on certain metric spaces. This theory is a very natural tool for analyzing the

* Received by the editors August 17, 1976, and in revised form April 29, 1977.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. This research was supported in part by the Office of Naval Research under N000-14-76-C-0279, in part by the National Science Foundation under 73-03846-A01 and in part by the Air Force Office of Scientific Research under AFOSR 76-3063.

asymptotic properties of the interpolated process. In particular, it enables us to exploit the statistical structure of the SA process, to enhance the rate of convergence. The method of proof is new in SA, and seems to be more easily generalizable (to other types of noise sequences and to other types of constrained problems) than past approaches. relatively little is known concerning the statistical properties of SA sequences, considered as a process. Our approach seems to be a useful tool for dealing with such properties for it emphasizes the “process” aspects of the problem. See, also the remarks after the statement of Theorem 5.1.

2. The unconstrained case. Formulation. Let X_n (with components X_n^1, \dots, X_n^r) denote the n th estimate of the local minimum, and let e_i denote the unit vector in the i th coordinate direction, and let $\{a_n, c_n\}$ be null sequences, a_n being a positive definite matrix, and c_n a (finite difference interval) positive scalar. Define the “observation difference” $\delta Y_n = \{\delta Y_n^1, \dots, \delta Y_n^r\}$ and the *observation noises* $\xi_n^{i,1}, \xi_n^{i,2}$ by

$$\begin{aligned} \delta Y_n^i &= (\text{observation at parameter } (X_n + e_i c_n)) \\ &\quad - (\text{observation at parameter } (X_n - e_i c_n)) \\ &\equiv [f(X_n + e_i c_n) + \xi_n^{i,1}] - [f(X_n - e_i c_n) + \xi_n^{i,2}]. \end{aligned}$$

Let $\xi_n^i = \xi_n^{i,1} - \xi_n^{i,2}$ and $\xi_n = (\xi_n^1, \dots, \xi_n^r)$, and let \mathcal{B}_n denote the σ -algebra determined by $X_0, \dots, X_n, \xi_0, \dots, \xi_{n-1}$.

Define X_{n+1} by $(\delta f_n \equiv (\delta f_n^1, \dots, \delta f_n^r), \delta f_n^i \equiv f(X_n + e_i c_n) - f(X_n - e_i c_n))$

$$(2.1) \quad X_{n+1} = X_n - \frac{a_n}{2c_n} \delta Y_n = X_n - a_n \left\{ \frac{\delta f_n}{2c_n} + \frac{\xi_n}{2c_n} \right\}.$$

The w.p.1 convergence of $\{X_n\}$ (when there is only one stationary point of $f(\cdot)$) has been the subject of most of the references in Wasan [1], and we mention only some of the conditions usually assumed, namely:

$$(2.2) \quad E_{\mathcal{B}_n} \xi_n = 0 \quad \text{w.p.1,} \quad E_{\mathcal{B}_n} \xi_n \xi_n' \leq M, \quad \text{all } n, \text{ for some matrix } M.$$

$$(2.3) \quad \sum_n a_n = \infty.$$

$$(2.4) \quad \sum_n a_n c_n < \infty, \quad \sum_n |a_n|^2 / c_n^2 < \infty.$$

The conditions were considerably relaxed in Ljung [2] and in Kushner [10], although they required more smoothness on $f(\cdot)$ than the previous works did. Also, [10] proved convergence to a stationary point of $f(\cdot)$, even when not unique. The conditions required here will be given later.

By simple alterations in the calculations, it is possible to treat noncentral difference and continuous time forms.

3. The constrained problem. A Lagrangian method. Now, suppose that we wish to modify the problem so that $f(x)$ is minimized under constraints $q_i(x) \leq 0$, $i = 1, \dots, s$, where each $q_i(\cdot)$ is continuously differentiable. Let $Q(x)$ denote the matrix $\{q_{1,x}(x), \dots, q_{s,x}(x)\}$, where $q_{i,x}(\cdot)$ is the gradient of $q_i(\cdot)$, and let $\{b_n^i\}$, $i = 1, \dots, s$, denote positive null sequences and let $\lambda = (\lambda^1, \dots, \lambda^s)$, $\lambda^i \geq 0$.

Consider the Lagrangian algorithm.

$$(3.1) \quad \lambda_{n+1}^i = \max [0, \lambda_n^i + b_n^i q_i(X_n)], \quad i = 1, \dots, s,$$

$$(3.2) \quad X_{n+1} = X_n - a_n \left\{ \frac{\delta Y_n}{2c_n} + Q(X_n) \lambda_n \right\}.$$

Suppose that there is a known number M such that the constrained minimum θ , and corresponding multipliers $\bar{\lambda}$ satisfy $|\theta^i| \leq M$, $\bar{\lambda}^i \leq M$. Modify (3.1), (3.2) by projecting on to the sets $\{x: |x^i| < M\}$, $\{\lambda: \lambda^i < M\}$, whenever the bounds are exceeded. Then, under conditions (2.2)–(2.4), and convexity conditions on $f(\cdot)$ and $q(\cdot)$, [7] proved that $\{X_n\}$ converges w.p.1 to the constrained minimum θ .

It was not proved that λ_n converged to an optimal multiplier. Indeed, the optimal multiplier may not be unique. Yet, let us note for later use, that in very many of the examples which we simulated, it appeared that λ_n did converge to a $\bar{\lambda}$ such that the Kuhn–Tucker condition

$$(3.3) \quad f_x(\theta) + Q(\theta) \bar{\lambda} = 0$$

held.

4. Equality constraints. An augmented penalty function method. Suppose now that we wish to minimize $f(\cdot)$ subject to equality constraints $\phi_i(x) = 0$, $i = 1, \dots, s$, where $\phi_i(\cdot)$ are continuously differentiable. An SA version of Mieses' [11] augmented penalty function method was developed in [9]. Let $0 < k$ denote a real number, define $P(x) = \frac{1}{2} \sum_i |\phi_i(x)|^2$ and $\Phi(x) = \{\phi_{1,x}(x), \dots, \phi_{s,x}(x)\}'$. let $\pi(x)$ denote the operator: $(I - \pi(x))v$ is the projection of $v \in R^r$ onto the span of $\{\phi_{1,x}(x), \dots, \phi_{s,x}(x)\}$. The necessary condition of the calculus for a local stationary point at x is $\pi(x)f_x(x) = 0$.

Define the algorithm $(P_x(x) = \Phi'(x)\phi(x))$

$$(4.1) \quad X_{n+1} = X_n - a_n \left[\pi(X_n) \frac{\delta Y_n}{2c_n} + k \Phi'(X_n) \phi(X_n) \right].$$

Under essentially the conditions (2.2)–(2.4), bounded double differentiability of $f(\cdot)$, and that $\Phi'(x)\phi(x) = 0$ implies that $\Phi(x)$ is of full rank, reference [9] proved convergence w.p.1 to a θ such that $\pi(\theta)f_x(\theta) = 0$.

5. Unconstrained problem. Theorem statement. Return to the algorithm of § 2. let $a_n = A/(n+1)^\alpha$, $c_n = C/(n+1)^\gamma$, where C , α and γ are positive real numbers, $\alpha > \gamma$, and A is a positive definite matrix. let us list the following assumptions.

(A5.1) $f(\cdot)$ is continuous, and has bounded and continuous mixed second derivatives.

(A5.2) $\sum_n a_n^2 < \infty$ (i.e., $\alpha > 1/2$).

(A5.3) $\sum_n a_n = \infty$ (i.e., $\alpha \leq 1$).

(A5.4)¹ There is $\theta \in R^r$ such that $X_n \rightarrow \theta$ w.p.1.

(A5.5) $f(\cdot)$ has continuous third derivatives $f_{x_i x_j x_k}(x)$ at $x = \theta$. Define $B(\theta) =$ vector whose i th component is this third derivative divided by $3!$

¹ It is possible to treat the case where θ is a random variable. Assumption (A5.9) limits our consideration to rates for the sequences which converge to a strict local minimum.

$$(A5.6) \quad E_{\mathcal{B}_n} \xi_n = 0 \quad \text{w.p.1.}$$

$$(A5.7) \quad \text{There is a matrix } \Sigma(\theta) \text{ such that } E_{\mathcal{B}_n} \xi_n \xi_n' \rightarrow \Sigma(\theta) \text{ w.p.1, as } n \rightarrow \infty.$$

$$(A5.8) \quad \text{For some } \delta > 0, M_1 < \infty,$$

$$E_{\mathcal{B}_n} |\xi_n|^{2+\delta} \leq M_1 \quad \text{w.p.1, all } n.$$

$$(A5.9I) \quad \text{Let } \alpha = 1, \beta = \alpha/3 = 2\gamma. \text{ Define } F(\theta) = \text{Hessian matrix of } f(\cdot) \text{ at } \theta \\ \text{Let the eigenvalues of } AF(\theta) - \beta I \equiv \bar{K}_1 \text{ have positive real parts.}$$

or

$$(A5.9II) \quad \text{Let } \alpha < 1, \beta = 2\gamma = \alpha/3, \text{ and let the eigenvalues of } AF(\theta) \equiv \bar{F}_2 \text{ have positive real parts.}$$

$$(A5.10) \quad f_x(\theta) = 0.$$

Remark. The conditions are mostly self-explanatory. Assumption (A5.9) implies that θ is a strict local minimum. The w.p.1 convergence (A5.4) is assumed, because we are concerned with rates of convergence. The actual convergence is proved in the various cited references. Condition (A5.6) is essentially a classical condition in the subject. As discussed in § 11, the condition can be readily weakened provided that we can still show that $P\{|U_n| \geq N\} \rightarrow 0$ as $N \rightarrow \infty$, uniformly in n , where $\{U_n\}$ is defined below. Indeed, the possibility of such extensions is one of the advantages of our approach. See the note added in proof.

Let $\varepsilon_{i,n}$ and $\bar{\varepsilon}_{i,n}$ denote functions whose values may differ from usage to usage, but which depend on X_n and c_n , and tend to zero w.p.1, as $n \rightarrow \infty$. (In § 9, they may also depend on λ_n .) Define $\delta X_n = X_n - \theta$. Then, using (A5.4, 5, 10), (2.1) can be rewritten in the form

$$(5.1) \quad X_{n+1} = X_n - a_n [F(\theta) \delta X_n + B(\theta) c_n^2 + \varepsilon_{1,n} c_n^2 + \varepsilon_{2,n} \delta X_n] - a_n \xi_n / (2c_n).$$

The next step is to scale $\{\delta X_n\}$. For some $\beta > 0$ (to be selected below—we are obviously interested in the largest β for which the process $\{U_n\}$ makes sense) define $U_n = (n+1)^\beta \delta X_n$. Then, using $(n+2)^\beta = (n+1)^\beta (1 + \beta/(n+1) + O(1/n^2))$,

$$(5.2) \quad U_{n+1} = (I + \beta I/(n+1) - a_n F(\theta) - a_n \bar{\varepsilon}_{1,n}) U_n - a_n (n+1)^\beta c_n^2 B(\theta) \\ - a_n (n+1)^\beta \xi_n / (2c_n) + a_n \bar{\varepsilon}_n,$$

where

$$\bar{\varepsilon}_n = (n+1)^\beta \left[\bar{\varepsilon}_{2,n} c_n^2 + \frac{1}{2c_n} \xi_n O\left(\frac{1}{n}\right) \right].$$

It turns out that all limits of $\{U_n\}$ or of the interpolated $\{U_n\}$ introduced below do not depend on the (asymptotically negligible—for the β to be selected) $\{\bar{\varepsilon}_n\}$ sequence. To slightly simplify the development, we will drop the term henceforth, although its presence would not affect any of the subsequent arguments—except that an additional term would have to be carried.

Interpolation. Introduction. The next step in the formulation of the limit theorem involves an interpolation of $\{U_n\}$ into a continuous parameter process. The form of the interpolation is motivated by the following observation. Let $\{\psi_n\}$ denote a sequence of (zero mean) independent, identically distributed (for convenience here) random variables with unit variance and $E|\psi_n|^{2+\gamma} \leq M < \infty$ for some real $\gamma > 0$, $M > 0$, and D be a matrix whose eigenvalues have positive real

parts. For each small $\Delta > 0$, define the sequence $\{V_n^\Delta\}$ and function $V^\Delta(\cdot)$ by

$$V_{n+1}^\Delta = (I - \Delta D)V_n^\Delta + \sqrt{\Delta}\psi_n, \quad V_0^\Delta = x, \quad \text{fixed},$$

and $V^\Delta(t) = V_n^\Delta$ in $[n\Delta, (n+1)\Delta)$. Then $\{V^\Delta(\cdot)\}$ converges in several statistical senses to the process solving

$$dV = -DV dt + dW,$$

where $W(\cdot)$ is a Wiener process.

Interpolation. Let $D[0, \infty)$ denote the space of real valued functions on $[0, \infty)$ which are right continuous and have left hand limits at each t . Suppose that $D[0, \infty)$ and its products are endowed with the Skorokhod topology (see Billingsley [16] for $D[0, T]$, Lindvall [20] for $D[0, \infty)$.) We mention only that convergence of a sequence $\{x^n(\cdot)\}$ to a continuous $x(\cdot)$ in that topology is equivalent to uniform convergence on each finite interval, and that, under that topology, the space is equivalent to a complete separable metric space, in that there is a metric, generating the same topology, under which the space is complete and separable (which we suppose henceforth).

Define $\Delta t_n = (n+1)^{-\alpha}$, $t_n = \sum_{i=0}^{n-1} \Delta t_i$, $t_0 = 0$, $\delta W_n = (n+1)^{\beta+\gamma-\alpha} \xi_n = (n+1)^{\beta+\gamma-\alpha/2} (\xi_n \sqrt{\Delta t_n})$, $W_n = \sum_{i=0}^{n-1} \delta W_i$, $W_0 = 0$. For each integer n, N , define $W_n^N = W_{N+n} - W_N$, $\delta W_n^N = \delta W_{N+n}$, $U_0^N = U_N$, and define $U^N(\cdot)$, $W^N(\cdot)$ by:

$$U^n(t) = U_{N+n}, \quad W^N(t) = W_{N+n} - W_N = W_n^N \quad \text{on } [t_{N+n} - t_N, t_{N+n+1} - t_N).$$

Note that $a_n \xi_n (n+1)^\beta / (2c_n) = (A/(2C)) \delta W_n$ and $a_n (n+1)^\beta c_n^2 B(\theta) = (AB(\theta)C^2) \Delta t_n (n+1)^{\beta-2\gamma}$. Also, the paths of $W^N(\cdot)$ and $U^N(\cdot)$ are in $D^r[0, \infty)$.

Dropping the $a_n \bar{\varepsilon}_n$ term, we have

$$U_{n+1} = G_n U_n - \left(\frac{A}{2C} \right) \sqrt{\Delta t_n} \xi_n (n+1)^{\gamma+\beta-\alpha/2} - (AB(\theta)C^2) \Delta t_n (n+1)^{\beta-2\gamma}, \quad (5.3)$$

where

$$G_n = (I + \beta I / (n+1) - A \Delta t_n (F(\theta) + \bar{\varepsilon}_{1,n})).$$

It is clear from (5.3) that unless $\gamma + \beta - \alpha/2 \leq 0$, $\beta - 2\gamma \leq 0$, $E|U_n|^2$ will diverge. We use, henceforth, the maximum β , namely $\beta = 2\gamma = \alpha/3$, with which the exponents of $(n+1)$ in (5.3) are all zero.

Let $\bar{W}(\cdot)$ denote a standard R^r valued Wiener process and $\bar{U}(\cdot)$ the (stationary process) solution to

$$(5.4) \quad d\bar{U}(t) = -\bar{K}\bar{U}(t) dt - AB(\theta)C^2 dt - (A/(2C))\Sigma^{1/2}(\theta) d\bar{W}(t),$$

where $\bar{K} = \bar{K}_1$ or \bar{K}_2 (see (A5.9)).

The undefined terms (concerning weak convergence) in Theorem 5.1 will be defined in the next section, and the proof given in § 7.

THEOREM 5.1. *Under (A5.1)–(A5.10), $\{U^N(\cdot), W^N(\cdot)\}$ is tight on $D^{2r}[0, \infty)$, and $\{U^N(\cdot)\}$ converges weakly to the $\bar{U}(\cdot)$ of (5.4). (I.e., any weak limit has the probability law of $\bar{U}(\cdot)$ on $D^r[0, \infty)$ or on $C^r[0, \infty)$.)*

Remarks. Clearly, we have the fastest rate of convergence when $\alpha = 1$ and (A5.9I) holds. The optimal normalization scaling β , and asymptotic normalized variance ($\lim_{t \rightarrow \infty} E \bar{U}(t) \bar{U}'(t)$) are not new; see Sacks [13] and Fabian [12], at least for this unconstrained problem. Also, McLeish uses weak convergence methods on the Robbins–Munro process [23].

The technique emphasizes the behavior of the asymptotic part of $\{U_n\}$, considered as a dynamical process. The correlation structure of this process can sometimes be exploited to yield (at time N) a function of the $\{X_N, X_{N-1}, \dots\}$, which is a better estimate of θ . See § 8. In practice, we observe the path “dynamically”, and it is worthwhile to try to understand its dynamical behavior. In certain cases, e.g., the Lagrangian method, the procedure often inherently oscillates around $\theta, \bar{\lambda}$, as it converges. The properties of this oscillation can be deduced from the relevant results of § 9, and used to improve the estimate, or to design a more suitable process. The approach may facilitate the use of SA in control theory, where the interest is often inherently dynamical.

Extensions. By a slight change in the method of proof, we can also get the extensions:

I. Assume the conditions of Theorem 5.1, but set $\beta = \min[2\gamma, \alpha/3]$. If $2\gamma < \alpha/3$, (resp. $2\gamma > \alpha/3$) then noise (resp., bias) is relatively unimportant in the limit and the theorem holds with dw (resp., $B(\theta)$) set equal to zero. If $\alpha = 1$, and $0 < b < \beta$ and the eigenvalues of $[AF(\theta) - bI]$ have positive real parts, then the interpolation of $\{(n+1)^b(X_n - \theta)\}$ converges weakly to the zero process.

II. *One-sided differences.* Use the observation difference [observation at $(X_n + c_n e_i) - \text{observation at } X_n]/c_n$ instead of δY_n^i . Then the theorem holds if $A/(2C)$, $B_i(\theta)$ and $\beta = 2\gamma = \alpha/3$ (or $\beta = \min[2\gamma, \alpha/3]$ in I above) are replaced by A/C , $f_{xix_i}(\theta)/2$ and $\beta = \gamma = \alpha/4$ (or $\beta = \min[\gamma, \alpha/4]$ in I above), resp.

Theorems 9.1 and 10.1 can also be extended in the same way.

6. Weak convergence. The material is in Billingsley [16]. See, also Whitt [17], Kushner [18, Chap. 2], Iglehart [19] or Lindvall [20], who gives the extensions of weak convergence on $D[0, T]$ (for some real T) to that on $D[0, \infty)$. let $\{Z^n\}$ denote a sequence of random variables with values in a complete separable metric space S , with associated σ -algebra \mathcal{S} .

The $\{P^n\}$ sequence (and also $\{Z^n\}$, here) is said to be *tight* if for each $\varepsilon > 0$, there is a compact $K_\varepsilon \in \mathcal{S}$ such that $P^n(K_\varepsilon) \geq 1 - \varepsilon$, all n . If $\{P^n\}$ is tight, then any subsequence has a further subsequence which converges weakly to some measure P on (S, \mathcal{S}) . If $\{P^n\}$ converges weakly to P , then

$$\int f(x) P^n(dx) \rightarrow \int f(x) P(dx)$$

for every bounded measurable $f(\cdot)$ which is continuous on a measurable set $S_0 \subset \mathcal{S}$, such that $P(S_0) = 1$.

If $P^n \rightarrow P$ weakly, and if Z is a random variable whose values are in (S, \mathcal{S}) and with measure P , we say that $Z^n \rightarrow Z$ weakly also. In this sense, Theorem 5.1 is understood to mean that the measures of the $D'[0, \infty)$ valued random variables $U^N(\cdot)$ (or the measures that $U^N(s)$, $s < \infty$, induce on $D'[0, \infty)$) converge weakly to the measure that $\bar{U}(\cdot)$ induces on $D'[0, \infty)$.

Let $Z^n(\cdot)$ be a sequence of processes with paths in $D[0, \infty) = S$, w.p.1, and induced measures P^n on (S, \mathcal{S}) . Then, there is tightness of $\{P^n\}$ or $\{Z^n(\cdot)\}$ on $D(0, \infty)$, if the restrictions to $D[0, T_k]$ are tight for some sequence $T_k \rightarrow \infty$. We sometimes use (without explicit mention) the fact (and similar facts) that if $E \max_{t \leq T} |Z^n(t)| \rightarrow 0$ as $n \rightarrow \infty$ for each T , and $Z^n(\cdot) \in D[0, \infty)$ w.p.1, then $Z^n(\cdot)$ converges weakly to the zero element of $D[0, \infty)$. We note for future use, that if $h(\cdot, \cdot)$ is a bounded continuous function, and $Z^n(\cdot) \rightarrow Z(\cdot)$ weakly, where $Z(\cdot)$ has continuous paths, then the processes defined by $\int_0^t h(t, s) Z^n(s) ds$ converge weakly to the process defined by $\int_0^t h(t, s) Z(s) ds$.

7. Proof of Theorem 5.1. 1. Returning to (5.3) and using $\beta = 2\gamma = \alpha/3$, we first show that $\{U_n\}$ is tight on R' . Define (neglecting $a_n \bar{\varepsilon}_n$, as we can easily show is legitimate) $\{v_n\}$ and \tilde{U}_n by

$$v_{n+1} = G_n v_n - AB(\theta) C^2 \Delta t_n, \quad \tilde{U}_n = U_n - v_n.$$

Note that, if random functions Z_n and Z_n^ε take values in the same space for each $\varepsilon > 0$, and differ on at most an ω set of measure ε , and if $\{Z_n^\varepsilon\}$ is tight on some space for each $\varepsilon > 0$, then $\{Z_n\}$ is tight. Thus, since $\bar{\varepsilon}_{1,n} \rightarrow 0$ w.p.1, if tightness (and the theorem) is proved under the assumption that for each $\varepsilon > 0$, there is an integer $N_\varepsilon < \infty$ such that $|\bar{\varepsilon}_{1,n}| \leq \varepsilon$ for $n \geq N_\varepsilon$, and $|\bar{\varepsilon}_{1,n}|$ is uniformly bounded, then they will be true in general. We make the assumption on $\bar{\varepsilon}_{1,n}$. Under this assumption $\{v_n\}$ is bounded, hence tight on R' . We only need to prove that $\{\tilde{U}_n\}$ is tight. We have

$$\tilde{U}_{n+1} = G_n \tilde{U}_n - (A/(2C)) \sqrt{\Delta t_n} \xi_n.$$

By (A5.6), (A5.8), and under cases (A5.9I) or (A5.9II), there are positive constants M_1, M_2 such that, for large n ,

$$E_{\mathcal{B}_n} |\tilde{U}_{n+1}|^2 \leq |G_n|^2 |\tilde{U}_n|^2 + M_1 \Delta t_n \leq (1 - M_2 \Delta t_n) |\tilde{U}_n|^2 + M_1 \Delta t_n,$$

from which boundedness of $\{E|\tilde{U}_n|^2\}$, hence tightness of $\{\tilde{U}_n\}$ follows.²

2. A representation for $U^N(\cdot)$. Define $C_i^j = I$ for $i > j$ and $C_i^j = G_j \cdots G_i$ for $i \leq j$. Then (5.3) is solved to get

$$U_{N+n+1} = C_N^{N+n} U_N - \sum_{m=N}^{N+n} C_{m+1}^{N+n} [A/(2C) \delta W_m + AB(\theta) C^2 \Delta t_m],$$

or, equivalently, (a more convenient form for us)

$$(7.1) \quad U_{N+n+1} = C_N^{N+n} U_N - \sum_{m=N}^{N+n} C_{m+1}^{N+n} [(A/(2C))((W_{m+1} - W_N) - (W_m - W_N)) + AB(\theta) C^2 ((t_{m+1} - t_N) - (t_m - t_N))].$$

For the moment, let us consider *only* the *sum* in (7.1) involving the W_i . Denoting that sum by \tilde{I}_N^{N+n} and rewriting it by collecting the coefficients of each W_i

² It is precisely the difficulty of proving tightness of $\{U_n\}$ when (A5.6) is relaxed, that forces us to require (A5.6). See the last section, where relaxations of the condition are discussed. See the note added in proof.

yields

$$\tilde{I}_N^{N+n} = \sum_{m=N}^{N+n} [C_m^{N+n} - C_{m+1}^{N+n}](A/(2C))(W_m - W_N) + (A/(2C))(W_{N+n+1} - W_N).$$

Using $C_m^{N+n} - C_{m+1}^{N+n} = -C_{m+1}^{N+n}(\bar{K} \Delta t_m + \varepsilon_{3,m} \Delta t_m)$, where $\bar{K} = \bar{K}_1$ or \bar{K}_2 according to the case of (A5.9), yields $\tilde{I}_N^{N+n} = I_N^{N+n} + \hat{I}_N^{N+n} + (A/(2C))(W_{N+n+1} - W_N)$, where

$$I_N^{N+n} = \sum_{m=N}^{N+n} C_N^{N+n}(C_N^m)^{-1}(-\bar{K} \Delta t_m)(A/(2C))(W_m - W_N),$$

$$\hat{I}_N^{N+n} = \sum_{m=N}^{N+n} C_N^{N+n}(C_N^m)^{-1}(-\varepsilon_{3,m} \Delta t_m)(A/(2C))(W_m - W_N).$$

Now, for each N , define $C_N(\cdot)$ on $[0, \infty)$ by

$$C_N(t) = C_N^{N+n} \quad \text{on } [t_{N+n} - t_N, t_{N+n+1} - t_N).$$

Since $\sum_{n=N}^{\infty} (\Delta t_n)^2 \rightarrow 0$ as $N \rightarrow \infty$, we have that $C_N(t) \rightarrow \exp(-\bar{K}t)$, as $N \rightarrow \infty$, uniformly (w.p.1) on finite time intervals.

For each N , define the interpolations $I_N(\cdot)$, $\hat{I}_N(\cdot)$, $\tilde{I}_N(\cdot)$ of the sequences $\{I_N^{N+n}\}$, $\{\hat{I}_N^{N+n}\}$, $\{\tilde{I}_N^{N+n}\}$, by e.g.,

$$I_N(t) = I_N^{N+n} \quad \text{on } [t_{N+n} - t_N, t_{N+n+1} - t_N).$$

It is not hard to show that $\{\hat{I}_N(\cdot)\}$ is tight on $D'[0, \infty)$ and goes to the “zero” process weakly, as $N \rightarrow \infty$. Thus, we ignore it henceforth. Define

$$\tilde{J}_N(t) = - \int_0^t C_N(t^-) C_N^{-1}(s) \bar{K} (A/(2C)) W^N(s) ds + (A/(2C)) W^N(t),$$

$$J_N(t) = - \int_0^t \exp(-\bar{K}(t-s)) \bar{K} (A/(2C)) W^N(s) ds + (A/(2C)) W^N(t).$$

It is not hard to show that

$$\{\tilde{J}_N(t) - J_N(t)\}, \quad \{\tilde{J}_N(t) - I_N(t)\}$$

are tight on $D'[0, \infty)$ and tend to the “zero” processes weakly as $N \rightarrow \infty$. A similar development can be made for the sum in (7.1) which involves the $(t_m - t_N)$. Putting all the foregoing together and adding the neglected terms, we have

$$\begin{aligned} U^N(t) &= (\exp(-\bar{K}t)) U^N(0) + \int_0^t \exp(-\bar{K}(t-s)) \bar{K} (A/(2C)) W^N(s) ds \\ &\quad - (A/(2C)) W^N(t) \\ (7.2) \quad &+ \int_0^t \exp(-\bar{K}(t-s)) \bar{K} (AB(\theta)C^2)s ds - (AB(\theta)C^2)t + \bar{e}_N(t), \end{aligned}$$

where $\bar{e}_N(\cdot)$ is a process which tends to the zero process weakly, as $N \rightarrow \infty$. Note that if a subsequence of $\{U^N(0), W^N(\cdot)\}$ converges weakly, so does the subsequence of $\{U^N(\cdot)\}$, and the limit process has continuous paths w.p.1.

3. Suppose that $W^N(\cdot)$ were tight on $D'[0, \infty)$ and that any weak limit is a Wiener process with covariance $\Sigma(\theta)t$. Then the limit of $\{U^N(\cdot)\}$ corresponding to

any weakly convergent subsequence of $\{W^N(\cdot), U^N(0)\}$ is equivalent (in law on $D'[0, \infty)$) to the process given by (7.2) with $U^N(0)$, $W^N(\cdot)$ replaced by some $\tilde{U}(0)$, $\tilde{W}(\cdot)$, where $\tilde{W}(\cdot)$ is a Wiener process with covariance $\Sigma(\theta)t$. The theorem follows from this, since the resulting right hand side of (7.2) solves (5.4) (as an integration of (7.2) by parts will show).

4. Thus, we only need to prove the first sentence of 3. We use Theorem 2 of Scott [14] with an appropriate change of notation. Scott's result is the following.

Let $\{v_n^N\}$ denote an array of scalar valued random variables where v_i^N , $i < n$, are \mathcal{B}_n^N measurable, for some sequence of σ -algebras \mathcal{B}_n^N , which is nondecreasing in n , for each N . Let Σ be a positive real number. Let $E_{\mathcal{B}_n^N} v_i^N = 0$ w.p.1, and define $m_N(t) = \max \{i: E \sum_{j=0}^i (v_j^N)^2 \leq t\Sigma\}$, $t \in [0, \infty)$, all N . Define

$$(7.3) \quad V^N(t) = \sum_{i=0}^{m_N(t)} v_i^N.$$

Let (condition B of Scott's Theorem 2; the notation \xrightarrow{P} means convergence in probability)

$$(7.4) \quad \sum_{i=0}^{m_N(t)} E\{(v_i^N)^2 | \mathcal{B}_i^N\} \xrightarrow{P} \Sigma t,$$

$$(7.5) \quad \sum_{i=0}^{m_N(t)} E\{(v_i^N)^2 I_{\{|v_i^N| \geq \varepsilon\}} | \mathcal{B}_i^N\} \xrightarrow{P} 0, \quad \text{all } \varepsilon > 0,$$

as $N \rightarrow \infty$, for each $t < \infty$. Then $\{V^N(\cdot)\}$ is tight on $D[0, \infty)$, and the limit of any weakly convergent subsequence is a Wiener process with covariance Σt and mean zero. Actually Scott deals with $D[0, T]$ and $\Sigma = 1$, but his theorem is valid on $D[0, \infty)$ also (and for any $\Sigma > 0$, by a simple scaling).

Assume that $\Sigma(\theta) \neq 0$, for otherwise the result is trivially true. Let $\lambda \in R'$ be such that $\lambda' \Sigma(\theta) \lambda > 0$. Identify \mathcal{B}_i^N with \mathcal{B}_{N+i} , v_i^N with $\lambda' \delta W_i^N$, and Σ with $\lambda' \Sigma(\theta) \lambda$. Note that the convergence $X_n \rightarrow \theta$ and (A5.7) imply (7.4). Condition (A5.8) implies (7.5). Thus, by Scott's theorem $\{V^N(\cdot)\}$ is tight and converges to a Wiener process with mean zero and covariance $(\lambda' \Sigma(\theta) \lambda)t$, as $N \rightarrow \infty$. Note that

$$E \sum_{j=0}^i (v_j^N)^2 = E \sum_{j=0}^i (\lambda' E \xi_{N+j} \xi'_{N+j} \lambda) \Delta t_{N+j},$$

and $E \xi_{N+j} \xi'_{N+j} \rightarrow \Sigma(\theta)$ as $N, j \rightarrow \infty$. Thus, the result remains true if $E|\lambda' \xi_{N+j}|^2 = E(v_j^N)^2$ is replaced by $\lambda' \Sigma(\theta) \lambda$ in the definition of $m_N(t)$. With this change, the definition of $V^N(\cdot)$ is the same as that of $\lambda' W^N(\cdot)$; thus $\lambda' W^N(\cdot)$ converges weakly to a Wiener process with mean zero and covariance $\lambda' \Sigma(\theta) \lambda t$.

Since the result of the last paragraph is true for each λ (if $\lambda' \Sigma(\theta) \lambda = 0$, then the $W^N(\cdot)$ converge to the "zero" process), we can conclude that $\{W^N(\cdot)\}$ is tight on $D'[0, \infty)$, and that it converges weakly to the desired R' valued Wiener process, with covariance $\Sigma(\theta)$, degenerate or not. Q.E.D.

8. Exploitation of the "correlation structure" of the limit process. In order to illustrate a possible useful application of the correlation properties of (5.4), consider a simple scalar case with $B(\theta) = 0$, $2C = 1$, $\beta = \alpha/3$. Write $F = F(\theta)$,

$\Sigma(\theta) = \sigma^2$; let $\alpha = 1$ and let $\bar{K} \equiv AF - \beta > 0$. Then (in steady state)

$$(8.1) \quad \begin{aligned} E\bar{U}(t) &= 0, \quad E\bar{U}(t)\bar{U}(t+s) = V \exp(-\bar{K}s), \quad s > 0, \\ V &= A^2\sigma^2/(2(AF-\beta)). \end{aligned}$$

V is minimized by $A = 2\beta/F \equiv A_0$.

In practice, F is not usually known. In order to reduce the sensitivity of the iterate sequence $\{X_n\}$ to “large initial noises”, and to partially rectify the slow convergence that occurs when A is too small, it is common for an $A > A_0$ to be used (where, of course, F must be guessed). For similar reasons, an $\alpha < 1$ is sometimes used, and what follows also holds for the $\alpha < 1$ case (in which case the correlation (8.1) uses $\beta = 0$ and $\bar{K} = AF$). As A increases, \bar{K} increases, and the process $\{(n+1)^\beta(X_n - \theta)\} = \{U_n\}$ behaves more “wildly”. We will try to exploit this.

Let $b \in [0, 1)$, $0 < t < \infty$ and define $q_N(t) = \max \{i: \sum_{j=0}^i \Delta t_{j+N} \leq t\}$, $q_N(0) = 0$. Using the fact that $n^{1/3}(X_n - \theta) \xrightarrow{D} N(0, V)$, together with the correlation structure of (8.1) yields, for large N ,

$$(8.2) \quad \begin{aligned} E[b(X_N - \theta) + (1-b)(X_{N+q_N(t)} - \theta)]^2 \\ \approx V \left[\frac{b^2}{N^{2/3}} + \frac{2b(1-b) \exp(-\bar{K}t)}{N^{1/3}(N+q_N(t))^{1/3}} + \frac{(1-b)^2}{(N+q_N(t))^{2/3}} \right]. \end{aligned}$$

(Condition (8.2) holds as a limit relation if we multiply both sides by $N^{2/3}$, and let $N \rightarrow \infty$.)

By definition of $q_N(\cdot)$,

$$\sum_{i=N}^{N+q_N(t)} \Delta t_i \approx t \approx \log[N+q_N(t)] - \log N,$$

and $e^{-t}(N+q_N(t))/N \rightarrow 1$, as $N \rightarrow \infty$. Then (8.2) is approximately (the ratios tend to unity as $N \rightarrow \infty$)

$$(8.3) \quad \frac{V}{N^{2/3}} [b^2 + 2b(1-b) \exp(-(\bar{K} + 1/3)t) + (1-b)^2 \exp(-2t/3)].$$

At $b = 0$, the derivative of (8.3) with respect to b is

$$(8.4) \quad \frac{2V}{N^{2/3}} (\exp(-t/3)) [\exp(-\bar{K}t) - \exp(-t/3)].$$

At $A = A_0$, $\bar{K} = \beta = 1/3$, and (8.4) = 0. So, under “ideal” conditions (at least from an “asymptotic” point of view), the linear combination inside (8.2) does not yield an improved estimate. But if $A > A_0$, then (8.4) is < 0 , suggesting that we can improve the estimate of θ at iterate $N + q_N(t)$, by using some linear combination of past iterates. Such ideas have a natural appeal, and such smoothing is sometimes used in practice, irrespective of the value of A ; but, we see that it can be harmful, unless $A > A_0$ (or $\alpha < 1$) and the weights are carefully selected.

An open, and interesting, question in the general vector case is whether A can be selected (still guaranteeing $X_n \rightarrow \theta$ w.p.1), but such that \bar{K} has some complex eigenvalues—the $\{X_n\}$ sequence would then exhibit “oscillations” on the

average, which perhaps, could be exploited (via suitable smoothing of the $\{X_n\}$) to obtain better estimates of θ .

Calculations made on simple sample problems indicate that complex eigenvalues (often corresponding to the eigenvalues with the smallest real parts) occur quite frequently for the Lagrangian algorithm of §§ 3 and 9. Perhaps smoothing can be usefully applied there.

It should be clear from the foregoing, that our "interpolated process" analysis of $\{X_n\}$ and $\{U_n\}$ is rather natural and that it can yield much more insight into the properties of the sequences, and smoothed functionals of the sequence, than can the more standard analysis of only the random variables $\{X_n, U_n\}$.

9. Asymptotic rate for the Lagrangian method of § 3. The development is close to that for Theorem 5.1, and only an outline will be given.

Assumptions. Assume (A5.1)–(A5.8) and

(A9.1) There is a $\bar{\lambda}$ such that $\lambda_n \rightarrow \bar{\lambda}$ w.p.1.

(A9.2) If $q_i(\theta) = 0$, then suppose that $\bar{\lambda}^i > 0$. (Of course, if $q_i(\theta) < 0$, then $\bar{\lambda}^i = 0$.)

(A9.3) Each $q_i(\cdot)$ has continuous and bounded first and second derivatives at θ .

(A9.4) $Q(\theta)$ (see § 3) is of full rank.

Let a_n, c_n be as in § 5, and let $b_n = \text{diag}(b^1, \dots, b^s)/(n+1)^\alpha = \text{diag}(b_n^1, \dots, b_n^s)$, where $b^i > 0, i \leq s$. Define

$$Q_i(\theta) = \{\partial^2 q_i(x)/\partial x^k \partial x^j, k, j = 1, \dots, r\} \quad \text{at } x = 0, \quad i = 1, \dots, s.$$

Define $\bar{K}_0 = A(F(\theta) + \sum_i \bar{\lambda}^i Q_i(\theta))$ and

$$\bar{K}_2 = \begin{bmatrix} \bar{K}_0 & AQ(\theta) \\ -BQ'(\theta) & 0 \end{bmatrix}.$$

(A9.5) Let $\alpha < 1$ ($\alpha = 1$, resp.) and let the eigenvalues of \bar{K}_2 ($\bar{K}_1 = \bar{K}_2 - \beta I$, resp.) have positive real parts.

(A9.6) Let θ satisfy the Kuhn–Tucker condition $f_x(\theta) + \sum_i \bar{\lambda}^i q_{i,x}(\theta) = 0$.

Remark. The $\lambda_n \rightarrow \bar{\lambda}$ convergence was not proved in [7], but it appears from our simulations that convergence of $\{\lambda_n\}$ occurs quite frequently. Assumption (A9.2) is often assumed in the deterministic case; it simply says that the "economic price" of an "active" resource is positive at the optimal point.

If $q_i(\theta) < 0$, then $q_i(X_n) \rightarrow q_i(\theta) < 0$ and (3.1), the convergence $X_n \rightarrow \theta$, and divergence of $\sum_i b_n$ together imply that $\lambda_n^i = 0$ for all large n . We will henceforth ignore q_i if $q_i(\theta) < 0$. We can and will assume that all s constraints are active at θ —with no loss of generality. The linear independence of the $q_{i,x}(\theta)$ is also commonly assumed in the analysis of deterministic algorithms.

If \bar{K}_0 is positive definite and A and B are diagonal with the same constant diagonal elements, then (using (A9.4)) Polyak [21, proof of Thm. 1] implies (A9.5), for $\beta = 0$.

Development of the algorithm. Owing to the remarks above and since we are only concerned with large n , we can write (3.1) as $\lambda_{n+1}^i = \lambda_n^i + b_n^i q_i(X_n)$, $i =$

1, \dots , s . Define $\delta\lambda_n = \lambda_n - \bar{\lambda}$. Then, using (A9.6), we can write

$$(9.1) \quad \begin{pmatrix} \delta X_{n+1} \\ \delta\lambda_{n+1} \end{pmatrix} = \begin{pmatrix} I_r - a_n(\bar{K}_0 + \varepsilon_{1,n}) & -a_n(Q(\theta) + \varepsilon_{2,n}) \\ b_n(Q'(\theta) + \varepsilon_{3,n}) & I_s \end{pmatrix} \begin{pmatrix} \delta X_n \\ \delta\lambda_n \end{pmatrix} - \begin{pmatrix} a_n(B(\theta)c_n^2 + \varepsilon_{4,n}c_n^2) \\ 0 \end{pmatrix} - \frac{a_n}{2c_n} \begin{pmatrix} \xi_n \\ 0 \end{pmatrix},$$

where I_r = identity in R^r .

The terms $\varepsilon_{i,n}$ all depend only on X_n and λ_n and tend to 0 w.p.1 as $n \rightarrow \infty$, by the convergence (of $\{X_n, \lambda_n\}$) and smoothness (on $f(\cdot)$, $q(\cdot)$) assumptions.

Define $U_n = (n+1)^\beta \begin{pmatrix} \delta X_n \\ \delta\lambda_n \end{pmatrix}$. Following the method by which (5.2) was obtained from (5.1), we get (the $\bar{\varepsilon}_{i,n}$ have the properties of the $\varepsilon_{i,n}$ above)

$$(9.2) \quad \begin{aligned} U_{n+1} = & \left[I + \frac{\beta}{(n+1)} I - (n+1)^{-\alpha} (\bar{K}_2 + \bar{\varepsilon}_{1,n}) \right] U_n \\ & - (n+1)^{-\alpha} (n+1)^{\beta-2\gamma} \begin{pmatrix} AB(\theta)C^2 \\ 0 \end{pmatrix} \\ & - (n+1)^{-\alpha/2} (n+1)^{-\alpha/2+\beta+\gamma} (A/(2C)) \begin{pmatrix} \xi_n \\ 0 \end{pmatrix} \left(1 + O\left(\frac{1}{n}\right) \right). \end{aligned}$$

Define $\Delta t_n = (n+1)^{-\alpha}$, t_n , δW_n and $W^N(\cdot)$ as in § 5, let $U^N(\cdot)$ denote the function which equals U_{N+n} on $[t_{N+n} - t_N, t_{N+n+1} - t_N)$ (analogously to the definition of $W^N(\cdot)$). Set $\beta = 2\gamma = \alpha/3$. Then (9.2) can be rewritten as ($\bar{K} = \bar{K}_1$ or \bar{K}_2 according to whether $\alpha = 1$ or $\alpha < 1$)

$$(9.3) \quad U_{n+1} = [I - \Delta t_n (\bar{K} + \bar{\varepsilon}_{3,n})] U_n - \Delta t_n \begin{pmatrix} AB(\theta)C^2 \\ 0 \end{pmatrix} - (A/(2C)) \begin{pmatrix} \delta W_n \\ 0 \end{pmatrix} + \Delta t_n \bar{\varepsilon}_{4,n}.$$

Let $\bar{W}(\cdot)$ denote a standard R^r valued Wiener process, and let $\bar{U}(\cdot)$ be the (stationary process) solution to

$$(9.4) \quad d\bar{U}(t) = -\bar{K}\bar{U}(t) dt - \begin{pmatrix} AB(\theta)C^2 \\ 0 \end{pmatrix} dt - \begin{pmatrix} \frac{A}{2C} \Sigma^{1/2}(\theta) d\bar{W}(t) \\ 0 \end{pmatrix}.$$

THEOREM 9.1. Assume (A9.1)–(A9.6), (A5.1)–(A5.8), and let $a_n = A/(n+1)^\alpha$, $b_n = B/(n+1)^\alpha$, $c_n = C/(n+1)^\gamma$, where $C > 0$, B is diagonal with positive elements, and A is positive definite. let $\beta = 2\gamma = \alpha/3$. Then $\{U^N(\cdot), W^N(\cdot)\}$ is tight on $D^{2r+s}[0, \infty)$, and any weak limit of the $\{U^N(\cdot)\}$ has the law of the (stationary process solution) $\bar{U}(\cdot)$ in (9.4).

Remarks. The proof is almost the same as that of Theorem 5.1 and is omitted.

Owing to the presence of the “multiplier dynamics” in (9.4), it is more likely that \bar{K}_i will have some complex eigenvalues. The consequent oscillations (of the correlation function) should be exploitable, via ideas such as those in § 7, to yield a

smoothed sequence of estimators which are better than $\{X_n\}$. Indeed, such a possibility justifies our point of view concerning the advantages of studying weak convergence of the processes $U^N(\cdot)$ to $\bar{U}(\cdot)$, over the simpler convergence in distribution of the R^r valued sequence $\{U_n\}$.

10. Asymptotic rates for the equality constrained algorithm of § 4. We will need the assumptions:

(A10.1) The $\phi_i(\cdot)$ have two continuous derivatives at θ .

(A10.2) The $\phi_{i,x}(\theta)$, $i = 1, \dots, s$ (the rows of $\Phi(\theta)$) are linearly independent.

Let $v(y)$ (with components $v_1(y), \dots, v_r(y)$) denote the vector $\pi(y)f_x(y)$. let $(\pi f_x(y))_x$ denote the matrix

$$\begin{aligned} \begin{bmatrix} v'_{1,x}(y) \\ \vdots \\ v'_{r,x}(y) \end{bmatrix} &= \begin{bmatrix} \partial v_1(y)/\partial x_1 & \cdots & \partial v_1(y)/\partial x_r \\ \vdots & & \vdots \\ \partial v_r(y)/\partial x_1 & \cdots & \partial v_r(y)/\partial x_r \end{bmatrix} \equiv [w_1(y), \dots, w_r(y)] \\ &= \begin{bmatrix} w_{11}(y) & & w_{r1}(y) \\ \vdots & \cdots & \vdots \\ w_{1r}(y) & & w_{rr}(y) \end{bmatrix} \end{aligned}$$

Define

$$\bar{K}_0 = (\pi f_x(\theta))_x + k \Phi'(\theta) \Phi(\theta).$$

(A10.3I)³ Let $\alpha = 1$, $\beta = 2\gamma = 1/3$. Let the eigenvalues of $\bar{K}_1 = A\bar{K}_0 - \beta I$ have positive real parts.

or

(A10.3II) Let $\alpha < 1$, $\beta = 2\gamma = \alpha/3$. Let the eigenvalues of $\bar{K}_2 = A\bar{K}_0$ have positive real parts.

(A10.4) θ satisfies the necessary condition for a constrained minimum, $\pi(\theta)f_x(\theta) = 0$.

Let $\Sigma_\pi(X_n)$ denote the covariance (given \mathcal{B}_n) of the projection of ξ_n onto the orthogonal complement of the span of $\phi_{1,x}(X_n), \dots, \phi_{s,x}(X_n)$. We use the terminology of § 5, except that $\{X_n\}$ is given by (4.1). Under (A5.7), and the convergence (A5.4), there is a matrix $\Sigma_\pi(\theta)$ such that $\text{cov}[\pi(X_n)\xi_n | \mathcal{B}_n] \rightarrow \Sigma_\pi(\theta)$ w.p.1, as $n \rightarrow \infty$.

THEOREM 10.1. Assume (A5.1)–(A5.8) and (A10.1)–(A10.4). Define $U_n = (n+1)^\beta (X_n - \theta)$. Then $\{U^n(\cdot), W^N(\cdot)\}$ is tight on $D^{2r}[0, \infty)$, and there is a standard Wiener process $\bar{W}(\cdot)$ such that any weak limit of $\{U^N(\cdot)\}$ has the (stationary process solution) law of the $\bar{U}(\cdot)$ in (10.1), where $\bar{K} = \bar{K}_1$ or \bar{K}_2 , according to the case of (A10.3).

$$(10.1) \quad d\bar{U}(t) = -\bar{K}\bar{U}(t) dt - A\pi(\theta)B(\theta)C^2 dt - (A/(2C))\Sigma_\pi^{1/2}(\theta) d\bar{W}(t).$$

Remark. The proof is very close to that of Theorem 5.1 and is omitted. We remark only on the expansion of (4.1), and on the conditions. The $\bar{\varepsilon}_{i,n}, \varepsilon_{i,n}$ have the

³ See the discussion after Theorem 10.1, and, in particular, the representation (10.8) for \bar{K}_0 .

same meaning as in § 5. With $X_n - \theta = \delta X_n$, we can write

$$(10.2) \quad \begin{aligned} \delta X_{n+1} = & \delta X_n - a_n [\pi(X_n) \{f_x(X_n) + B(\theta)c_n^2 + \varepsilon_{1,n}c_n^2 + \varepsilon_{2,n}\delta X_n \\ & + \xi_n/(2c_n)\} + k\Phi'(\theta)\Phi(\theta)\delta X_n + \varepsilon_{3,n}\delta X_n]. \end{aligned}$$

using (A10.4) and the smoothness assumptions on $f(\cdot)$ and $\phi(\cdot)$, we get

$$(10.3) \quad \pi(X_n)f_x(X_n) = (\pi f_x(\theta))_x \delta X_n + \varepsilon_{4,n}\delta X_n.$$

Now, using (10.2), (10.3), the values of α, β, γ in (A10.3) (Case I or II), and a development of $(n+2)^\beta$ such as used in Theorem 5.1, we get (analogously to the unconstrained case)

$$(10.4) \quad \begin{aligned} U_{n+1} = & [I - \Delta t_n(\bar{K} + \bar{\varepsilon}_{1,n})]U_n - \Delta t_n A \pi(\theta) B(\theta) C^2 \\ & - \sqrt{\Delta t_n}(A/(2C))\pi(\theta)\xi_n - \Delta t_n \left(\bar{\varepsilon}_{2,n} + \xi_n O\left(\frac{1}{n}\right) \right). \end{aligned}$$

Using (10.4), the proof goes as it does for Theorem 5.1.

Remark on (A10.3). Without loss of generality, suppose that $\theta = 0$. Let $T(y)$ denote the tangent line or hyperplane to the curve or surface $\{x: \phi(x) = 0\}$ at y , and $T_0(y)$ the orthogonal complement to $T(y)$. In general $T_0(0) \supset \text{span}\{\phi_{i,x}(0), i \leq s\}$. We make the additional assumption that $T_0(0) = \text{span}\{\phi_{i,x}(0), i \leq s\}$. Let $x = (x_1, \dots, x_r)$ denote the generic point in R^r . With no loss of generality (and some gain in insight), we can assume that the basis is such that x_1, \dots, x_s and x_{s+1}, \dots, x_r form bases for $T_0(0)$ and $T(0)$, resp. Using the last three sentences and (A10.2), we have that there is a nonsingular $(s \times s)$ matrix $\tilde{\Phi}$ such that

$$(10.5) \quad \Phi'(0)\Phi(0) = \begin{bmatrix} \tilde{\Phi}'\tilde{\Phi} & 0 \\ 0 & 0 \end{bmatrix}.$$

Let N^ε denote an ε -neighborhood of $0 \in R^r$. There are differentiable functions $l_1(\cdot), \dots, l_s(\cdot)$ on R^{r-s} such that if $x \in N^\varepsilon \cap \{x: \phi(x) = 0\} \equiv N_\phi^\varepsilon$, and ε is small enough, then

$$x_i = l_i(x_{s+1}, \dots, x_r), \quad i = 1, \dots, s.$$

Henceforth, assume that ε is “sufficiently” small.

We will now develop a representation for $(\pi f_x(0))_x$. Let δ denote a small real number. By the definition of the tangent plane or line $T(0)$, $l_i(e_j\delta) = O(\delta^2)$, $j = s+1, \dots, r$. Consequently, for $j \geq s$, the smoothness of $\phi(\cdot)$, $\pi(\cdot)$ and $f_x(\cdot)$ implies that

$$(10.6) \quad \pi\left(e_j\delta + \sum_{i=1}^s e_i l_i(e_j\delta)\right) \cdot f_x\left(e_j\delta + \sum_{i=1}^s e_i l_i(e_j\delta)\right) = \pi(e_j\delta)f_x(e_j\delta) + O(\delta^2).$$

By the definition of $w_i(0)$,

$$(10.7) \quad \lim_{\delta \rightarrow 0} \pi(e_i\delta)f_x(e_i\delta)/\delta = w_i(0).$$

recall that $\pi(0)f_x(0) = 0$. Note that for each i , the vector

$$\pi(\delta e_i)f_x(\delta e_i) = [\text{projection of } f_x(\delta e_i) \text{ onto the tangent plane } T(\delta e_i)]$$

has components $O(\delta)$ on $T(0)$ and $O(\delta^2)$ on $T_0(0)$. Thus, by (10.7), $w_i(0)$ is in $T(0)$ and, hence, has the form $w_i(0) = \begin{bmatrix} 0 \\ g_i \end{bmatrix}$ for some $r-s$ vector g_i .

We will examine further the term $\bar{F}_\phi \equiv [g_{s+1}, \dots, g_r]$. Define the function $\bar{f}(\cdot)$ on $N^\varepsilon \cap T(0)$ by

$$\bar{f}(x_{s+1}, \dots, x_r) = f(l_1(x_{s+1}, \dots), \dots, l_s(x_{s+1}, \dots), x_{s+1}, \dots, x_r).$$

Note that the vector of the last $r-s$ components of the left side of (10.6) is the gradient of $\bar{f}(\cdot)$ at $e_j\delta$ (modulo a term of order $O(\delta^2)$). This fact, together with (10.6) and (10.7) imply that \bar{F}_ϕ is the Hessian matrix of $\bar{f}(\cdot)$ at 0. Finally, \bar{K}_0 takes the form

$$(10.8) \quad \bar{K}_0 = \begin{bmatrix} -\frac{k\tilde{\Phi}'\tilde{\Phi}}{g_1, \dots, g_s} & \begin{smallmatrix} 0 \\ \bar{F}_\phi \end{smallmatrix} \end{bmatrix}.$$

Thus (A10.3II) holds if the eigenvalues of the Hessian \bar{F}_ϕ have strictly positive real parts and $A = \text{diagonal}(a, a, \dots)$, $a > 0$. In any case, the representation (10.8) clarifies the meaning of the condition (A10.3).

Remark on the value of k . If $\alpha = 1$, we require that the real parts of the eigenvalues of

$$A\bar{K}_0 - \beta I$$

be positive. This requires that $k \geq$ some minimum positive value. Certain deterministic algorithms [15] also require a minimum value of k , for convergence. Here, convergence occurs for any k . But, if $\alpha = 1$, and k is too small, the rate (β) will not be $\alpha/3 = 1/3$, but something less, something which also seemed to hold in our experiments.

11. Extensions. Assumption (A5.6) was used, because we were not otherwise able to prove tightness of $\{U_n\}$ in any generality. Theorem 11.1 follows from the proof of the previous theorems. It is not too difficult to find reasonable conditions under which $W^n(\cdot)$ converges weakly to a Wiener process. After the theorem statement, we will comment on this. See the note added in proof.

THEOREM 11.1. *Assume the conditions of Theorem 5.1 (or of (9.1, 10.1), except for (A5.6), and suppose that $\{U_n, W^n(\cdot)\}$ are tight, and $\{W^n(\cdot)\}$ converges weakly to a Wiener process with mean 0, and covariance $\Sigma_0(\theta)$. Then the conclusions of the theorems still hold.*

Remarks. Our remarks will be confined to the unconstrained case, for the others are treated similarly. The $a_n \bar{\varepsilon}_n$ in (5.2) causes no problem; the difficulty in proving tightness of $\{U_n\}$ on R^r has been due to the $\varepsilon_{1,n}$ term, although it is "sure" to be eliminated eventually. Indeed, if tightness of $\{U_n\}$ can be shown, then even (A5.4) can be replaced by the weaker convergence in the theorem of Kushner [10].

Define $q_N(t) = 0$ on $[0, \Delta t_N)$, $q_N(t) = i$ on $[\Delta t_N + \dots + \Delta t_{N+i-1}, \Delta t_N + \dots + \Delta t_{N+i})$. Then

$$W^N(t) = \sum_{i=0}^{q_N(t)-1} \xi_{N+i}(\Delta t_{N+i})^{1/2}.$$

Tightness and convergence of $\{W^n(\cdot)\}$. Let us drop (A5.6)–(A5.8). We will replace them by the “reasonable” conditions (A11.1) to (A11.3). According to Billingsley [16, Thm. 15.5] if $\{W^N(\cdot)\}$ satisfies a criterion “similar” to that used to prove tightness⁴ on $C^r[0, \infty)$, then it will be tight on $D^r[0, \infty)$, and all limits will be continuous w.p.1. In particular, by Theorems 15.5, 12.3 and 12.2 of [16] and the definition of $W^N(\cdot)$, this holds if there is a real K such that

$$(11.1) \quad E|W_{n+m} - W_n|^4 \leq K \left| \sum_{i=n}^{n+m-1} \Delta t_i \right|^2, \quad \text{all } n, m > n.$$

Equation (11.2) is equivalent to (11.1), where $t_n^N = t_{N+n} - t_N$.

$$(11.2) \quad E|W^N(t_{n+m}^N) - W^N(t_n^N)|^4 \leq K \left| \sum_{i=n}^{n+m-1} \Delta t_i^N \right|^2, \quad \text{all } N, n, m > n.$$

Just to simplify the notation in the following development, we assume that the ξ_n are scalar valued.

Let \mathcal{B}_n be the σ -algebra determined by $X_0, \dots, X_{n+1}, \xi_0, \dots, \xi_n$. Assume
(A11.1) *There is an integer $k \geq 0$ such that for all $N, i, |E_{\mathcal{B}_N} \xi_{N+i}| \leq (1 + |\bar{\xi}_N|) \alpha_i^N$, where α_i^N are real quantities satisfying $\sum_{i=0}^{q_N(t)} \alpha_i^N (\Delta t_{N+i})^{1/2} \rightarrow 0$ as $N \rightarrow \infty$, for each t , where we define $\bar{\xi}_N = \sum_{l=0}^k |\xi_{N-l}|$.*

Let $R(\theta, l)$ and $\beta_{i,\Delta}^N$ be real quantities such that

$$\sum_l |R(\theta; l)| < \infty, \quad \sum_{i,j=1}^{q_N(t)} \beta_{i,j}^N (\Delta t_{N+i} \Delta t_{N+j})^{1/2} \rightarrow 0$$

as $N \rightarrow \infty$, for each t .

(A11.2) *$|E_{\mathcal{B}_N} \xi_{N+i} \xi_{N+i+l} - R(\theta, l)| \leq (1 + |\bar{\xi}_N|^2) \beta_{i,i+l}^N$, $l \geq 0$, for some integer $k \geq 0$, and all N, i, l .*

(A11.3) *There is a bounded function $R(\cdot, \cdot, \cdot, \cdot)$ and real number K such that $|E \xi_i \xi_j \xi_k \xi_l| \leq R(i, j, k, l)$ and (where $t, t+s$ are restricted to jump times of $W^N(\cdot)$)*

$$\sum_{i,j,k,l=q_N(t)}^{q_N(t+s)-1} R(N+i, N+j, N+k, N+l) (\Delta t_{N+i} \Delta t_{N+j} \Delta t_{N+k} \Delta t_{N+l})^{1/2} \leq K s^2.$$

Discussion of the conditions. The conditions can all be weakened, but we do not know their “best” form. Assumption (A11.1) replaces (A5.6). It describes the type of mixing or summability condition that holds if the $\{\xi_m\}$ were generated by the solution to an equation such as

$$(11.3) \quad \xi_{n+i+1} + a_0 \xi_{n+i} + a_1 \xi_{n+i-1} \cdots a_i \xi_n = \psi_n,$$

where the $\{\psi_n\}$ are independent, Gaussian, zero mean, and identically distributed, and the roots of $[\lambda^{i+1} + a_0 \lambda^i + \cdots + a_i = 0]$ are all strictly interior to the unit circle.

Similarly, (A11.2) holds for (11.3). If the noise $\{\xi_m\}$ were a stationary process, with $E \xi_i \xi_{i+l} = R(l)$, then (A11.2) would read

$$|E_{\mathcal{B}_N} \xi_{N+i} \xi_{N+i+l} - E \xi_{N+i} \xi_{N+i+l}| \leq (1 + |\bar{\xi}_N|^2) \beta_{i,i+l}^N.$$

⁴ $C[0, \infty)$ = space of continuous functions on $[0, \infty)$ with the metric of uniform convergence on finite intervals.

The condition (A11.2) is a type of “asymptotic” stationarity (as $X_n \rightarrow \theta$) combined with a mixing condition. Conditions (A11.1)–(A11.2) are used to show that the limit of $\{W^N(\cdot)\}$ is a Wiener process, given tightness, and (A11.3), which implies (11.2), also implies tightness. Also, it holds for (11.3).

Condition (A11.3) actually is not too restrictive. For example, it commonly occurs that there are functions $\bar{R}(\cdot, \cdot)$ such that $R(i, j, k, l) \leq \bar{R}(i, j)\bar{R}(k, l) + \bar{R}(i, l)\bar{R}(k, j) + \bar{R}(i, k)\bar{R}(j, l)$, and where $\bar{R}(i, j) \leq M\alpha^{|i-j|}$ for some $\alpha \in (0, 1)$, and real M . Then the sum in (A11.3) is bounded above by

$$(11.4) \quad 3M^2 \left(\sum_{i,j=q_N(t)}^{q_N(t+s)-1} \alpha^{|i-j|} (\Delta t_{N+i} \Delta t_{N+j})^{1/2} \right)^2$$

and, in turn, the bound in (A11.3) can be verified from (11.4). Note that (A11.3) implies that:

$$(11.5) \quad \text{For each } t, \text{ there is an } M_1(t) < \infty \text{ such that} \\ E|W^N(t)|^4 \leq M_1(t), \text{ all } N.$$

Under (A11.3), (11.2) holds and $\{W^N(\cdot)\}$ is tight on $D[0, \infty)$, and the paths of any limit process are continuous w.p.1. Assume that $\{U_N\}$ is tight on R^r . Then, $\{U^N(\cdot)\}$ is tight also.

We will prove that the limit of $\{W^N(\cdot)\}$ is a Wiener process with covariance $R_0(\theta) = R(\theta; 0) + 2 \sum_{i=1}^{\infty} R(\theta; i)$. First, some estimates are needed. let $\mathcal{B}_N(t)$ be the smallest σ -algebra which measures $\{W^N(s), U^N(s), s \leq t\}$, and, for notational simplicity, fix s, t , and set $m_N = N + q_N(t)$.

By (A11.1), and the definition of $W^N(t+s) - W^N(t)$,

$$(11.6) \quad |E_{\mathcal{B}_N(t)} W^N(t+s) - W^N(t)| \leq \sum_{i=q_N(t)}^{q_N(t+s)-1} (\Delta t_{N+i})^{1/2} \alpha_i^N (1 + |\bar{\xi}_{m_N}|).$$

We can write

$$(11.7) \quad \begin{aligned} E_{\mathcal{B}_N(t)} (W^N(t+s) - W^N(t))^2 &= \sum_{i,j=q_N(t)}^{q_N(t+s)-1} (\Delta t_{N+i} \Delta t_{N+j})^{1/2} E_{\mathcal{B}_N(t)} \xi_{N+i} \xi_{N+j} \\ &= \sum_{i=q_N(t)}^{q_N(t+s)-1} \Delta t_{N+i} E_{\mathcal{B}_N(t)} \xi_{N+i}^2 \\ &\quad + 2 \sum_{l=1}^{q_N(t+s)-l-1} \sum_{i=q_N(t)}^{q_N(t+s)-l-1} (\Delta t_{N+i} \Delta t_{N+i+l})^{1/2} E_{\mathcal{B}_N(t)} \xi_{N+i} \xi_{N+i+l}. \end{aligned}$$

By (A11.2), we can rewrite (11.7) as

$$(11.8) \quad \begin{aligned} R(\theta; 0) \sum_{i=q_N(t)}^{q_N(t+s)-1} \Delta t_{N+i} \\ + 2 \sum_{l=1}^{q_N(t+s)-l-1} R(\theta; l) \sum_{i=q_N(t)}^{q_N(t+s)-l-1} (\Delta t_{N+i} \Delta t_{N+i+l})^{1/2} + F_N(t, s), \end{aligned}$$

where the coefficients of the $R(\theta; 0)$ or $2R(\theta; l)$ tend to s , as $N \rightarrow \infty$, and where

$$(11.9) \quad |F_N(t, s)| \leq 2(1 + |\bar{\xi}_{m_N}|^2) \sum_{i,j=q_N(t)}^{q_N(t+s)-1} \beta_{i,j}^N (\Delta t_{N+i} \Delta t_{N+j})^{1/2}.$$

Let $h(\cdot)$ be a real valued bounded continuous function on some Euclidean space R^{2q} , with t_1, \dots, t_q arbitrary real numbers $\leq t$. Let $\{U^N(\cdot), W^N(\cdot)\}$ denote a weakly convergent subsequence. Let the limit process be denoted by $\bar{U}(\cdot), \bar{W}(\cdot)$. By (11.6) and the uniform integrability (11.5) and (A11.1),

$$(11.10) \quad Eh(W^N(t_i), U^N(t_i), i \leq q)(W^N(t+s) - W^N(t)) \rightarrow 0,$$

By weak convergence, and the uniform integrability (11.5), the left side of (11.10) also tends to

$$Eh(\bar{W}(t_i), \bar{U}(t_i), i \leq q)(\bar{W}(t+s) - \bar{W}(t)),$$

which must thus equal 0. Similarly, (11.8), (11.9), (A11.2) and the weak convergence and uniform integrability (11.5) yield that

$$Eh(\bar{W}(t_i), \bar{U}(t_i), i \leq q)[(\bar{W}(t+s) - \bar{W}(t))^2 - R_0(0)s] = 0.$$

The last paragraph, together with the arbitrariness of $h(\cdot)$, $t, s, t_i \leq t$, and the continuity of $\bar{W}(\cdot)$, w.p.1, imply that $\bar{W}(\cdot)$ is a continuous martingale with quadratic variation $R_0(\theta)s$; hence it is the asserted Wiener process.

Note added in proof. Tightness of $\{U_n\}$ and other recent extensions will appear in a monograph (in preparation) by the author and D. Clark.

REFERENCES

- [1] M. T. WASAN, *Stochastic Approximation*, Cambridge University Press, Cambridge, England, 1969.
- [2] L. LJUNG, *Convergence of recursive stochastic algorithms*, Rep. 7403, 1974, Lund Inst. of Tech., Div. of Automatic Control, Lund, Sweden.
- [3] L. LJUNG, T. SODERSTROM AND I. GUSTAVSSON, *Counterexamples to general convergence of a commonly used recursive identification method*, IEEE Trans. Automatic Control, AC-20 (1975), pp. 643-652.
- [4] V. FABIAN, *Stochastic approximation of constrained minima*, Proc. 4th Prague Conf. on Statistical Decision Theory and Information Theory, 1966, pp. 277-289.
- [5] H. J. KUSHNER AND H. T. GAVIN, *Stochastic approximation-like algorithms for constrained systems: Algorithms and numerical results*, IEEE Trans. Automatic Control, AC-19 (1971), pp. 349-357.
- [6] ———, *Stochastic approximation algorithms for constrained optimization problems*, Ann. Statist., 2 (1974), pp. 713-723.
- [7] H. J. KUSHNER AND E. SANVICENTE, *Stochastic approximation for constrained systems with observation noise on the system and constraints*, Automatica—J. IFAC, 11 (1975), pp. 375-380.
- [8] ———, *Penalty function methods for constrained stochastic approximation*, J. Math. Anal. Appl., 46 (1974), pp. 499-512.
- [9] H. J. KUSHNER AND M. L. KELMANSON, *Stochastic approximation algorithms of the multiplier type for the sequential Monte Carlo optimization of constrained systems*, this Journal, 14 (1976), pp. 827-842.
- [10] H. J. KUSHNER, *General convergence results for stochastic approximations via weak convergence theory*, J. Math. Anal. Appl., to appear.
- [11] A. MIELE, E. G. CRAGG, R. R. IYER AND A. V. LEVY, *Use of augmented penalty function in mathematical programming problems, Part I*, J. Optimization Theory Appl., 8 (1971), pp. 115-130.
- [12] V. FABIAN, *On asymptotic normality in stochastic approximation*, Ann. Math. Statist., 39 (1968), pp. 1327-1332.

- [13] J. SACKS, *Asymptotic distribution of stochastic approximation procedures*, Ibid., 29 (1958), pp. 273–405.
- [14] D. J. SCOTT, *Central limit theorems for martingales and for processes with stationary increments using a Skorokhod representation approach*, Advances in Appl. Probability, 5 (1973), pp. 119–137.
- [15] D. I. BERTSEKAS, *Multiplier methods: A survey*, Automatica—J. IFAC, 12 (1976), pp. 133–145.
- [16] P. BILLINGSLEY, *Convergence of probability Measures*, John Wiley, New York, 1968.
- [17] W. WHITT, *A guide to the application of limit theorems for sequences of stochastic processes*, Operations Res., 18 (1970), pp. 1207–1213.
- [18] H. J. KUSHNER, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Academic Press, New York, 1977.
- [19] D. E. IGLEHART, *Diffusion approximations in applied probability*, Math. of the Decision Sciences, Part II, Lectures in Appl. Math., 12, American Mathematical Society, Providence, RI, 1968.
- [20] T. LINDVALL, *Weak convergence of probability measures and random functions in the function space $D[0, \infty)$* , J. Appl. Probability, 10 (1973), pp. 109–121.
- [21] B. T. POLYAK, *Iteratives methods using Lagrange multipliers for solving extremal problems with constraints of the equation type*, Ž. Vycisl. Mat. i Mat. Fiz., 10 (1970), pp. 1098–1106; U.S.S.R. Computational Math. and Math. Phys., 10 (1970), pp. 42–52.
- [22] H. J. KUSHNER AND S. LAKSHMIVARAHAN, *Numerical studies for constrained stochastic approximation*, IEEE Trans. Automatic Control, to appear, July 1977.
- [23] D. L. MCLEISH, *Functional and random central limit theorems for the Robbins Munro process*, J. Appl. Probability, 13 (1976), pp. 148–154.

HEREDITARY CONTROL PROBLEMS: NUMERICAL METHODS BASED ON AVERAGING APPROXIMATIONS*

H. T. BANKS† AND J. A. BURNS‡

Abstract. An approximation scheme involving approximation of linear functional differential equations by systems of high order ordinary differential equations is formulated and convergence is established in the context of known results from linear semigroup theory. Applications to optimal control problems are discussed and a summary of numerical results is given. The paper is concluded with a brief survey of previous literature on this class of approximations for systems with delays.

1. Introduction. In this paper we consider a particular approximation scheme which can be employed to solve optimal control problems governed by linear hereditary systems (functional differential equations). The general idea may in actuality be clearly viewed in the context of classical Ritz techniques where the problem of minimizing a functional J over a space \mathcal{X} is approximated by a sequence of problems requiring minimization over approximating spaces \mathcal{X}^N . In § 2 we show that the original functional differential equation system can be equivalently formulated as a linear ordinary differential equation $\dot{z} = \mathcal{A}z + F$ in an appropriately chosen Hilbert space. We then in § 3 make use of classical approximation results from linear semigroup theory (essentially the analogue of the simple result: $\mathcal{A}^N \rightarrow \mathcal{A}$ implies $e^{\mathcal{A}^N t} \rightarrow e^{\mathcal{A} t}$) to establish in a concise manner convergence for a particular class of approximations. These approximations, which involve approximation of the original infinite dimensional hereditary system by a system of high order—but finite dimensional—ordinary differential equations, have been proposed frequently in the literature. We illustrate application of these ideas to a class of standard optimal control problems in § 4 and present a brief summary of our numerical experience (which has been documented in more detail elsewhere [4]) using these techniques. Finally in the concluding section we comment briefly on the numerous contributions to the mathematical and engineering literature which are related to the type of approximations under discussion here.

While we shall save for § 5 our more detailed comments concerning the literature, a remark or two here should suffice to put our presentation in this paper in perspective for the reader. Essentially, we carry out a particular case of the general ideas for approximating optimal controls via the Trotter type approximations suggested in an earlier paper by DeJulio [17]. Sasai and Shimemura [54] applied these ideas to obtain theoretical results for approximation of optimal controls for problems with partial differential equations. In a more recent paper, Sasai and Fukuda [53] attempted to apply the Trotter type approximation

* Received by the editors March 11, 1976, and in revised form May 2, 1977.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912, and Department of Mathematics, University of Colorado, Boulder, Colorado 80302. This research was supported in part by the U.S. Air Force under Contract AF-AFOSR-71-2078B and in part by the National Science Foundation under Grant NSF-GP-28931x1.

‡ Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061. This research was supported by the National Science Foundation under Grant MPS 75-0693.

techniques to develop results for approximate controllability of simple differential difference equations. However, their (formal, in part) presentation in [53] is neither complete nor correct (see our detailed comments in § 5 below). Use of the Trotter type ideas in computing approximate optimal controls for functional differential equation problems was first suggested by the present authors in [2] and an extensive numerical study was detailed in [4], a summary of which constitutes part of § 4 below.

The Hilbert spaces ($Z = R^n \times L_2$) chosen for our abstract formulation in § 2 have by now been widely recognized as appropriate settings for the investigation of certain problems involving hereditary systems. These spaces (or equivalents of them) have been used in previous functional differential equation studies by a number of authors including Krasovskii [20], [31], Coleman and Mizel [13], [14], Borisovič and Turbabin [12], Delfour and Mitter [18], [19], Webb [62], [63], as well as the present authors [2].

While our formulation and use of the Trotter–Kato approximation in § 3 differs somewhat from that in previous presentations such as [2], [17], [53], [54] (the present formulation, which emphasizes convergence of the sequence of approximating infinitesimal generators and *not* that of a sequence of certain associated projection operators, will encompass a number of other types of approximation schemes in addition to the one that is the focus of our attention in this paper—details on these ideas will appear in forthcoming papers), the *essence* of the ideas in this paper are not new. Indeed, as we point out in § 5, all of these various contributions to the approximation of optimal controls can be quite correctly viewed as a special case of the classical Ritz techniques. Furthermore, as we document in detail in § 5, the use of the particular approximations (high order ODE's for the FDE) discussed here has been widespread in the mathematical and engineering literature. However, in light of the often heuristic and/or incorrect use of both the Trotter ideas and the high order ODE approximation ideas in the literature, we feel that our presentations here, in addition to developing some new technical aspects of these ideas, will serve to clarify and unify a number of partially developed ideas in the literature. We offer a reasonably complete treatment (with concise and correct proofs) of both the theoretical (precise statements of convergence results are provided) and computational aspects of these very general approximation ideas as applied to a very specific class of problems. Indeed, in § 4 we report on our investigations of a number of test examples for which we have found and made a comparison of both analytic and numerical (with these approximations) solutions. To our knowledge, no such comprehensive treatment currently exists in the literature.

Throughout the manuscript we shall use $L_p^\nu(a, b) = L_p([a, b], R^\nu)$ to denote the customary Lebesgue spaces of R^ν -valued “functions” on $[a, b]$ whose components are integrable when raised to the p th power. Whenever $\nu = n$ (the dimension of the basic vector system under discussion) we shall write just $L_p(a, b)$. The usual Banach space $\mathcal{C}([a, b], R^\nu)$ of continuous functions on $[a, b]$ with values in R^ν will be denoted by $\mathcal{C}^\nu(a, b)$ and similarly, the measurable functions $\mathcal{M}([a, b], R^\nu)$ on $[a, b]$ into R^ν will be denoted simply by $\mathcal{M}^\nu(a, b)$. Again when $\nu = n$ we shall simply write $\mathcal{C}(a, b)$ and $\mathcal{M}(a, b)$. We shall also have occasion to use the Banach spaces $W_p^{(j)}(a, b) = W_p^{(j)}([a, b], R^n)$ of R^n -valued absolutely continuous functions

possessing $j-1$ absolutely continuous derivatives and j th derivatives [61] that are in L_p . For a measurable function $s \rightarrow x(s)$, the notation x_t will denote the function in $\mathcal{M}(-r, 0)$ given by $x_t(\theta) = x(t + \theta)$, $-r \leq \theta \leq 0$. The vector space of $n \times m$ matrices will be denoted by $\mathcal{L}_{n,m} = \mathcal{L}(R^m, R^n)$ and we shall not distinguish between a row vector and its transposed column form whenever it is clear from the context which is meant. Furthermore we will use the same symbol $|\cdot|$ to denote any one of several norms when it is perfectly clear from the usage which norm is intended. Finally $\langle \cdot, \cdot \rangle_{R^n}$ and $\langle \cdot, \cdot \rangle_{L_2}$ will denote the standard inner products in R^n and L_2 respectively while $\sigma(A)$, $\pi(A)$, and $\rho(A)$ will represent as usual the spectrum, point spectrum, and resolvent set in the complex plane \mathbb{C} of an operator A .

2. Equivalence of an FDE and an AEE. We consider the initial value problem

$$(2.1) \quad \dot{x}(t) = L(x_t) + f(t), \quad t \geq 0,$$

$$(2.2) \quad x(0) = \eta, \quad x_0 = \varphi$$

where $\eta \in R^n$, $\varphi \in L_2(-r, 0)$, $f \in L_2(0, t_1)$ for each $t_1 > 0$, and we make the following standing assumptions for the presentation in this paper.

The linear operator $L: \mathcal{M}(-r, 0) \rightarrow R^n$ is such that when restricted to $\mathcal{C}(-r, 0)$ it is continuous. Hence there exists a matrix function $H: (-\infty, \infty) \rightarrow \mathcal{L}_{n,n}$ which is of bounded variation, right continuous on $(-r, 0)$ with $H(s) = H(-r) = 0$ for $s \leq -r$, $H(0) = H(s)$ for $s \geq 0$, such that for each $\varphi \in \mathcal{C}(-r, 0)$

$$L(\varphi) = \int_{-r}^0 [d_s H(s)] \varphi(s).$$

We further assume specifically that the function $s \rightarrow H(s)$ consists of a saltus function with a finite number of jumps plus an absolutely continuous part which possesses an L_2 -derivative. That is, for $t_1 > 0$ and $x \in L_2(-r, t_1)$ the mapping (equivalence class) $t \rightarrow g(t) = L(x_t)$ can be written

$$L(x_t) = \sum_{i=0}^{\nu} A_i x(t - h_i) + \int_{-r}^0 D(\theta) x(t + \theta) d\theta$$

where $A_i \in \mathcal{L}_{n,n}$ and $D \in L_2([-r, 0], \mathcal{L}_{n,n})$.

Following Borisovič and Turbabin [12], we choose as our "state space" the product space $Z \equiv R^n \times L_2(-r, 0)$ with inner product

$$\langle (\xi, \psi), (\eta, \varphi) \rangle_Z \equiv \langle \xi, \eta \rangle_{R^n} + \langle \psi, \varphi \rangle_{L_2}.$$

For $(\eta, \varphi) \in Z$ a solution of (2.1)–(2.2) is a function x , in $L_2(-r, t_1)$ for each $t_1 > 0$, such that $t \rightarrow x(t)$ is absolutely continuous (A.C.) for $t \geq 0$, x satisfies (2.2), x satisfies (2.1) a.e. on $[0, \infty)$. It can be shown that for $f \in L_2(0, t_1)$ for all $t_1 > 0$ and $(\eta, \varphi) \in Z$, (2.1)–(2.2) possesses a unique solution which depends continuously on the initial data $(\eta, \varphi) \in Z$ (e.g., see [12]). We next define the underlying semigroup on Z associated with the homogeneous (2.1)–(2.2). Define for $t \geq 0$ $S(t): Z \rightarrow Z$ by $S(t)(\eta, \varphi) = (x(t), x_t)$ where x is the solution of (2.1)–(2.2) with $f \equiv 0$. Then $\{S(t)\}_{t \geq 0}$ is a C_0 -semigroup. Using rather standard arguments (these results are

discussed in [3], [12], [18], [62], [63] and, no doubt, other reports since they are by now well-known to investigators working on these problems) one can then establish:

- (2.3) If \mathcal{A} denotes the closed and densely defined infinitesimal generator of $\{S(t)\}$, then $\mathcal{D}(\mathcal{A}) = \{(\eta, \varphi) \in Z \mid \varphi \text{ is A.C., } \dot{\varphi} \in L_2(-r, 0), \text{ and } \eta = \varphi(0)\}$ and for $(\varphi(0), \varphi) \in \mathcal{D}(\mathcal{A})$ we have $\mathcal{A}(\varphi(0), \varphi) = (L(\varphi), \dot{\varphi})$.
- (2.4) $\sigma(\mathcal{A}) = \pi(\mathcal{A})$ and $\lambda \in \pi(\mathcal{A})$ iff $\det \Delta(\lambda) = 0$ where $\Delta(\lambda) \equiv \lambda I - \int_{-r}^0 [d_s H(s)] e^{\lambda s}$.
- (2.5) There exists a constant β such that $\sigma(\mathcal{A})$ lies in the left half plane $\{\lambda \in \mathbb{C} \mid \operatorname{Re}(\lambda) < \beta\}$ and for any $\varepsilon > 0$, there exists $M = M_{\varepsilon, \beta}$ finite such that

$$|S(t)| \leq M e^{(\beta + \varepsilon)t}, \quad t \geq 0.$$

Let us comment briefly on these results. The veracity of (2.4) follows from arguments very similar to those found in Hale's monograph [20, Lemma 20.1, pp. 99–100]. The distribution of the roots of $\det \Delta(\lambda) = 0$ is also discussed in [20] as well as [9] and the first part of (2.5) follows from those discussions. The second part of (2.5) follows from the result often used implicitly in linear semigroup theory, which we state precisely here as a lemma.

LEMMA 2.1. *Suppose $\{T(t)\}$ is a C_0 -semigroup with infinitesimal generator \mathcal{A} and suppose $T(\bar{t})$ is compact for some $\bar{t} > 0$. If $\sigma(\mathcal{A}) \subset \{\lambda \in \mathbb{C} \mid \operatorname{Re}(\lambda) < \omega\}$ then for any $\varepsilon > 0$, there exists $M = M_{\varepsilon, \omega}$ finite such that $|T(t)| \leq M e^{(\omega + \varepsilon)t}$, $t \geq 0$.*

Finally, returning to (2.3), direct calculations easily reveal that $\mathcal{S} \equiv \{(\eta, \varphi) \mid \varphi \text{ is A.C. with } \dot{\varphi} \in L_2 \text{ and } \varphi(0) = \eta\} \subset \mathcal{D}(\mathcal{A})$ and for $(\eta, \varphi) \in \mathcal{S}$, $\mathcal{A}(\eta, \varphi) = (L(\varphi), \dot{\varphi})$. The reverse of the above inclusion is equally straightforward to argue as follows:

$(\eta, \varphi) \in \mathcal{D}(\mathcal{A})$ implies there exists $(\mu, \psi) \in Z$ such that as $\varepsilon \rightarrow 0$, (i) $[x(\varepsilon) - x(0)]/\varepsilon \rightarrow \mu$ (which yields $\mu = L(\varphi)$) and (ii) $[x_\varepsilon - x_0]/\varepsilon \rightarrow \psi$ in $L_2(-r, 0)$, where x is the solution of (2.1)–(2.2) with $f \equiv 0$ corresponding to (η, φ) . For $\varepsilon > 0$, $\theta \in [-r, 0]$, define

$$\Phi_\varepsilon(\theta) = \frac{1}{\varepsilon} \int_\theta^{\theta + \varepsilon} x(s) ds.$$

Then from (ii) we have that $\dot{\Phi}_\varepsilon(\theta) = (x(\theta + \varepsilon) - x(\theta))/\varepsilon \rightarrow \psi(\theta)$ in $L_2(-r, 0)$ as $\varepsilon \rightarrow 0^+$ and, furthermore, $\Phi_\varepsilon(0) \rightarrow \eta$. Thus $\{\Phi_\varepsilon\}$ converges in the $W_2^{(1)}$ norm to some function $\Phi \in W_2^{(1)}$ (in fact, $\Phi(\theta) = \eta + \int_0^\theta \psi$) and consequently $\Phi_\varepsilon(\theta) \rightarrow \Phi(\theta)$ a.e. on $[-r, 0]$. But we also have $\Phi_\varepsilon \rightarrow \varphi$ a.e. on $[-r, 0]$. That is, $\varphi = \Phi$ a.e. on $[-r, 0]$ and Φ is an absolutely continuous representative for the L_2 “function” φ such that $(\eta, \varphi) = (\Phi(0), \Phi)$.

For $(\eta, \varphi) \in Z$ and $f \in L_2(0, t_1)$, we define $z(\cdot; \eta, \varphi, f): [0, t_1] \rightarrow Z$ (we shall often suppress notation and write just $z(\cdot; f)$ or even just z) by

$$(2.6) \quad z(t; f) = S(t)(\eta, \varphi) + \int_0^t S(t - \sigma)(f(\sigma), 0) d\sigma.$$

Whenever f is sufficiently smooth (i.e., $f \in C^1$) and $(\eta, \varphi) \in \mathcal{D}(\mathcal{A})$, this is equivalent

to the abstract evolution equation in Z in differentiated form

$$(2.7) \quad \begin{aligned} \dot{z}(t) &= \mathcal{A}z(t) + F(t), \quad t > 0, \\ z(0) &= (\eta, \varphi), \end{aligned}$$

where $F(t) = (f(t), 0) \in Z$ for each t . We shall, in the sequel, sometimes refer to (2.6) (or its sometimes equivalent form (2.7)) as the (AEE) while referring to (2.1)–(2.2) as the (FDE). The following theorem concerning the equivalence of solutions of the (FDE) and the (AEE) is established in [2] and improves and extends earlier results of Borisovič and Turbabin [12].

THEOREM 2.1. *Suppose $(\eta, \varphi) = (\varphi(0), \varphi) \in \mathcal{D}(\mathcal{A})$ and $f \in L_2(0, t_1)$ where $t_1 > 0$. Letting $x(\cdot; f)$ denote the solution of (2.1)–(2.2) corresponding to (η, φ) and f , we have*

$$z(t; f) = (x(t; f), x_t(\cdot; f))$$

for $t \geq 0$.

The proof of the above equivalence can be made quite concise and involves only observing that the result is valid for φ and f sufficiently smooth and then using elementary density and continuous dependence arguments. We remark here that most of the results given in this paper extend easily to the case where $Z = R^n \times L_p(-r, 0)$, $1 < p < \infty$ (see [2]).

3. Approximation results. Our concern in this paper is the approximation of solutions of (FDE) via use of approximate solutions of (AEE). We shall in this section carry out the details for a very specific class of approximations. These results can be considered a special case of the general theory developed in [2], although we shall give a slight reformulation of the approximation ideas here. In this formulation it will be clear that it is the convergence of the approximating infinitesimal generators to \mathcal{A} that is essential to the method while the convergence of certain associated projection operators is of less importance.

The underlying tool for the method discussed in this paper is a semigroup approximation result attributed to Trotter [28], [59], [60]. We shall employ here a version of this theorem given by Pazy [42, p. 90]. Using standard notation [28, p. 485], by $A \in G(M, \beta)$ we shall mean that A is the infinitesimal generator (i.g.) for a C_0 -semigroup $\{T(t)\}$ on a Banach space X satisfying

$$|T(t)| \leq M e^{\beta t}, \quad t \geq 0.$$

We state precisely the theorem fundamental to our discussions.

THEOREM 3.1. *Let $\mathcal{A} \in G(M, \beta)$ be the i.g. for a C_0 -semigroup $\{S(t)\}$ on a Hilbert space Z and suppose*

$$(3.1) \quad \mathcal{A}^N \in G(M, \beta) \quad \text{for } N = 1, 2, \dots,$$

$$(3.2) \quad \mathcal{A}^N z \rightarrow \mathcal{A}z \quad \text{for } z \in \mathcal{D}, \quad \mathcal{D} \text{ a dense subset of } Z,$$

$$(3.3) \quad \text{there exists } \lambda_0 \text{ with } \operatorname{Re}(\lambda_0) > \beta \text{ such that } (\mathcal{A} - \lambda_0 I)\mathcal{D} \text{ is dense in } Z.$$

Then if $\{S^N(t)\}$ denotes the C_0 -semigroup generated by \mathcal{A}^N , we have $S^N(t)z \rightarrow S(t)z$ for every $z \in Z$, $t \geq 0$, and, in fact the convergence is uniform in t on bounded intervals.

Our use of this theorem will entail beginning with the i.g. \mathcal{A} for $S(t)$ as defined in § 2 (see (2.3)) on the space $Z = R^n \times L_2(-r, 0)$ and finding approximating operators \mathcal{A}^N satisfying (3.1)–(3.3). We shall then be able to conclude the convergence properties stated in the theorem.

Let us suppose for the moment that we have a sequence of approximations $\mathcal{A}^N: \mathcal{D}(\mathcal{A}^N) \subset Z \rightarrow Z$ so that the conclusions of Theorem 3.1 obtain. Then for fixed $(\eta, \varphi) \in Z$ we define, for $f \in L_2(0, t_1)$ and $t \geq 0$,

$$(3.4) \quad \hat{z}^N(t; f) = S^N(t)(\eta, \varphi) + \int_0^t S^N(t-\sigma)(f(\sigma), 0) d\sigma,$$

and it follows directly that for $t \geq 0$, $\hat{z}^N(t; f) \rightarrow z(t; f)$ where $z(t; f)$ is defined in (2.6).

We shall be interested in making such approximations when $\mathcal{A}^N: Z \rightarrow Z^N$ is bounded linear and Z^N is a finite dimensional subspace of Z . Then $S^N(t) = e^{\mathcal{A}^N t}$ and the equation (3.4) can be equivalently written

$$(3.5) \quad \begin{aligned} \dot{\hat{z}}^N(t) &= \mathcal{A}^N \hat{z}^N(t) + (f(t), 0), \\ \hat{z}^N(0) &= (\eta, \varphi). \end{aligned}$$

If $(f(t), 0) \in Z^N$ for each t , then the differential equation in (3.5) is an ordinary differential equation (ODE) in the finite dimensional space Z^N . If, in addition, $(\eta, \varphi) \in Z^N$, then we see that (3.5) yields an ODE initial value problem in the finite dimensional subspace Z^N of Z and the solutions of such a sequence of problems are approximations for the solution of the initial value problem represented by (2.1)–(2.2) or equivalently (2.6).

As we shall see below, for the specific approximations considered in this paper, elements of Z of the form $(\xi, 0)$, $\xi \in R^n$, are in Z^N for each N . However, as one might expect, the initial data (η, φ) in general will not satisfy the requirement $(\eta, \varphi) \in Z^N$ for all N . Hence to effect our finite dimensional ODE approximations for (2.1)–(2.2), we use, in place of (3.4), the approximating system

$$(3.6) \quad z^N(t; f) = S^N(t)P^N(\eta, \varphi) + \int_0^t S^N(t-\sigma)(f(\sigma), 0) d\sigma$$

or, equivalently,

$$(3.7) \quad \begin{aligned} \dot{z}^N(t) &= \mathcal{A}^N z^N(t) + (f(t), 0), \\ z^N(0) &= P^N(\eta, \varphi), \end{aligned}$$

where $\{P^N\}$ is a sequence of linear operators $P^N: Z \rightarrow Z^N$. If one then assumes that in addition to (3.1), (3.2), (3.3), one also has $P^N(\eta, \varphi) \rightarrow (\eta, \varphi)$ for all (η, φ) in the desired initial data set [note that this is the *only* convergence requirement on the sequence $\{P^N\}$; specifically it is not in general necessary that $P^N z \rightarrow z$ for all $z \in Z$], then once again one has the desired convergence of the approximate solutions: $z^N(t; f) \rightarrow z(t; f)$ for $t \geq 0$ and $f \in L_2$.

We turn now to the particular case of the above quite general ideas that is the focus of our discussions in this paper. We consider the retarded n -vector system

$$(3.8) \quad \begin{aligned} \dot{x}(t) &= A_0 x(t) + A_1 x(t-r) + \int_{-r}^0 D(\theta) x(t+\theta) d\theta + f(t), \quad t > 0, \\ x(0) &= \varphi(0), \quad x_0 = \varphi, \end{aligned}$$

where $A_i \in \mathcal{L}_{n,n}$, $D \in L_2([-r, 0], \mathcal{L}_{n,n})$, $f \in L_2(0, t_1)$, and $r > 0$. For any positive integer N , we partition the interval $[-r, 0]$ into subintervals $[t_j^N, t_{j-1}^N]$, where $t_j^N = -jr/N$, $j = 0, 1, 2, \dots, N$. Let χ_j^N denote the characteristic function of $[t_j^N, t_{j-1}^N)$ for $j = 2, 3, \dots, N$, with χ_1^N the characteristic function of $[t_1^N, t_0^N] = [-r/N, 0]$. We define the finite dimensional subspace Z^N of Z by

$$(3.9) \quad Z^N \equiv \left\{ (\eta, \varphi) \in Z \mid \eta \in R^n, \varphi = \sum_{j=1}^N v_j^N \chi_j^N, v_j^N \in R^n \right\}.$$

Note that for $\xi \in R^n$, $(\xi, 0) \in Z^N$ for each N as was needed in our discussions above.

We next define the approximating operators $\mathcal{A}^N: Z \rightarrow Z^N$ by

$$(3.10) \quad \mathcal{A}^N(\eta, \varphi) \equiv \left(A_0 \varphi_0^N + A_1 \varphi_N^N + \sum_{j=1}^N \frac{r}{N} D_j^N \varphi_j^N, \sum_{j=1}^N \frac{N}{r} (\varphi_{j-1}^N - \varphi_j^N) \chi_j^N \right)$$

where

$$(3.11) \quad \begin{aligned} \varphi_0^N &\equiv \eta, \quad \varphi_j^N \equiv \frac{N}{r} \int_{t_j^N}^{t_{j-1}^N} \varphi(s) ds, \quad D_j^N \equiv \frac{N}{r} \int_{t_j^N}^{t_{j-1}^N} D(s) ds, \\ & \quad j = 1, 2, \dots, N. \end{aligned}$$

We remark at this point that the i.g. \mathcal{A} (see (2.3)) associated with (3.8) is given by

$$(3.12) \quad \mathcal{A}(\eta, \varphi) \equiv \left(A_0 \varphi(0) + A_1 \varphi(-r) + \int_{-r}^0 D(s) \varphi(s) ds, \dot{\varphi} \right)$$

for $(\eta, \varphi) = (\varphi(0), \varphi) \in \mathcal{D}(\mathcal{A})$.

Finally, we define the operators $P^N: Z \rightarrow Z^N$ by

$$(3.13) \quad P^N(\eta, \varphi) \equiv \left(\eta, \sum_{j=1}^N \varphi_j^N \chi_j^N \right)$$

where the φ_j^N are defined in (3.11). Observe that P^N is a continuous projection of Z onto the closed subspace Z^N .

From (3.11) and the definition of $|\cdot|_Z$ it follows easily that

$$(3.14) \quad |\varphi_j^N|^2 \leq \frac{N}{r} |(\eta, \varphi)|_Z^2, \quad j = 0, 1, \dots, N,$$

$$(3.15) \quad |D_j^N|^2 \leq \frac{N}{r} |D|_{L_2}^2, \quad j = 1, 2, \dots, N,$$

where for (3.14) we obviously require $N \geq r$ in case $j = 0$.

We next establish, via several lemmas, some interesting aspects of the approximations (3.10)–(3.11). That conditions (3.1)–(3.3) hold for these approximations will follow easily from these results and thus convergence of the associated solutions approximating the solution of (3.8) will be assured.

LEMMA 3.1. *The operators \mathcal{A}^N of (3.10) are bounded linear operators on Z and hence generate C_0 -semigroups $S^N(t) = e^{\mathcal{A}^N t}$ which are thus actually uniformly continuous semigroups.*

The proof of this lemma is rather straightforward in light of the estimates (3.14), (3.15) and will therefore be omitted.

We observe that indeed the operators $\mathcal{A}^N: Z \rightarrow Z^N$ are compact. Note that this precludes compactness of $S^N(t)$ for any $t \geq 0$ since otherwise one would then obtain that $I = S^N(t) - \sum_{k=1}^{\infty} (\mathcal{A}^N t)^k / k!$, being the difference of a compact operator and the limit of a uniformly convergent sequence of compact operators, is compact on Z . We further remark that while the \mathcal{A}^N are bounded, they are not uniformly bounded on Z (otherwise use of the convergence results of the next lemma along with the uniform boundedness principle would imply \mathcal{A} of (3.12) bounded on Z).

LEMMA 3.2 (Consistency). *Let $\mathcal{D} \equiv \{(\varphi(0), \varphi) \in Z \mid \varphi \text{ is continuously differentiable on } [-r, 0]\}$. Then \mathcal{D} is dense in Z and $\mathcal{A}^N z \rightarrow \mathcal{A}z$ for $z \in \mathcal{D}$.*

Proof. It suffices to show that for $\varphi \in C^{(1)}[-r, 0]$

$$(i) \quad A_0 \varphi_0^N + A_1 \varphi_N^N + \sum_{j=1}^N \frac{r}{N} D_j^N \varphi_j^N \rightarrow A_0 \dot{\varphi}(0) + A_1 \varphi(-r) + \int_{-r}^0 D(s) \varphi(s) ds$$

and

$$(ii) \quad \sum_{j=1}^N \frac{N}{r} (\varphi_{j-1}^N - \varphi_j^N) \chi_j^N \rightarrow \dot{\varphi} \quad \text{in } L_2(-r, 0).$$

(i) Since $\varphi_0^N = \varphi(0)$ and $\varphi_N^N = (N/r) \int_{-r}^{-(N-1)r/N} \varphi(s) ds$, it follows immediately that $A_0 \varphi_0^N + A_1 \varphi_N^N \rightarrow A_0 \dot{\varphi}(0) + A_1 \varphi(-r)$. Further, note that

$$\sum_{j=1}^N \frac{r}{N} D_j^N \varphi_j^N = \sum_{j=1}^N \frac{r}{N} \left(\frac{N}{r} \int_{t_j^N}^{t_{j-1}^N} D(s) ds \right) \varphi_j^N = \int_{-r}^0 D(s) \left(\sum_{j=1}^N \varphi_j^N \chi_j^N(s) \right) ds.$$

For φ continuous it is easily seen that $\sum_{j=1}^N \varphi_j^N \chi_j^N \rightarrow \varphi$ in $L_2(-r, 0)$ (we note that this implies $P^N z \rightarrow z$ for $z \in Z$ since the continuous functions are dense in L_2 and $|P^N| \leq 1$ —see (3.13)) and thus one finds

$$\begin{aligned} \left| \sum_{j=1}^N \frac{r}{N} D_j^N \varphi_j^N - \int_{-r}^0 D(s) \varphi(s) ds \right| &= \left| \int_{-r}^0 D(s) [\sum_{j=1}^N \varphi_j^N \chi_j^N(s) - \varphi(s)] ds \right| \\ &\leq \|D\|_{L_2} \|\sum_{j=1}^N \varphi_j^N \chi_j^N - \varphi\|_{L_2} \rightarrow 0 \quad \text{as } N \rightarrow \infty. \end{aligned}$$

(ii) Given $\varphi \in C^{(1)}[-r, 0]$ define for $j = 1, 2, \dots, N$

$$(3.16) \quad \mathcal{E}_j^N = \sup \{ |\dot{\varphi}(s) - \dot{\varphi}(\theta)| \mid s, \theta \in [t_j^N, t_{j-1}^N] \}$$

and let $K \equiv \sup \{ |\dot{\varphi}(s)| \mid s \in [-r, 0] \}$.

Writing, for $\tau \in [t_j^N, t_{j-1}^N]$,

$$\varphi(\tau) = \varphi(t_j^N) + \dot{\varphi}(t_j^N)(\tau - t_j^N) + \int_{t_j^N}^{\tau} [\dot{\varphi}(s) - \dot{\varphi}(t_j^N)] ds,$$

it is readily seen that

$$(3.17) \quad \varphi(\tau) = \varphi(t_j^N) + \dot{\varphi}(t_j^N)(\tau - t_j^N) + E_j^N(\tau)$$

where $|E_j^N(\tau)| \leq \mathcal{E}_j^N(\tau - t_j^N)$, $\tau \in [t_j^N, t_{j-1}^N]$. It then follows from (3.17) that for $j = 1, 2, \dots, N$

$$(3.18) \quad \varphi_j^N \equiv \frac{N}{r} \int_{t_j^N}^{t_{j-1}^N} \varphi(\tau) d\tau = \varphi(t_j^N) + \frac{1}{2} \dot{\varphi}(t_j^N) \frac{r}{N} + F_j^N$$

where $|F_j^N| \leq \mathcal{E}_j^N(r/2N)$.

Using (3.18) we then find

$$\begin{aligned} & \sum_{j=1}^N \int_{t_j^N}^{t_{j-1}^N} \left| \frac{N}{r} \{ \varphi_{j-1}^N - \varphi_j^N \} - \dot{\varphi}(s) \right|^2 ds \\ &= \int_{-r/N}^0 \left| \frac{N}{r} \{ \varphi(0) - \varphi_1^N \} - \dot{\varphi}(s) \right|^2 ds \\ &+ \sum_{j=2}^N \int_{t_j^N}^{t_{j-1}^N} \left| \frac{N}{r} \{ \varphi(t_{j-1}^N) - \varphi(t_j^N) \} + \frac{1}{2} \{ \dot{\varphi}(t_{j-1}^N) - \dot{\varphi}(t_j^N) \} + \frac{N}{r} \{ F_{j-1}^N - F_j^N \} - \dot{\varphi}(s) \right|^2 ds. \end{aligned}$$

The second term is bounded by (use the mean value theorem)

$$\begin{aligned} & \sum_{j=2}^N \int_{t_j^N}^{t_{j-1}^N} \left\{ \left| \frac{N}{r} (\varphi(t_{j-1}^N) - \varphi(t_j^N)) - \dot{\varphi}(s) \right| + \frac{1}{2} \mathcal{E}_j^N + \frac{N}{r} |F_j^N| + \frac{N}{r} |F_{j-1}^N| \right\}^2 ds \\ & \leq \sum_{j=2}^N \int_{t_j^N}^{t_{j-1}^N} \left\{ \mathcal{E}_j^N + \frac{1}{2} \mathcal{E}_j^N + \frac{1}{2} \mathcal{E}_j^N + \frac{1}{2} \mathcal{E}_{j-1}^N \right\}^2 ds \\ & \leq r \left\{ \sup_{1 \leq j \leq N} \frac{5}{2} \mathcal{E}_j^N \right\}^2, \end{aligned}$$

while estimates on the first term yield

$$\begin{aligned} & \int_{-r/N}^0 \left| \frac{N}{r} \left\{ \varphi(0) - \varphi\left(\frac{-r}{N}\right) - \frac{1}{2} \dot{\varphi}\left(\frac{-r}{N}\right) \frac{r}{N} - F_1^N \right\} - \dot{\varphi}(s) \right|^2 ds \\ & \leq \int_{-r/N}^0 \left\{ \left| \frac{N}{r} \left(\varphi(0) - \varphi\left(\frac{-r}{N}\right) \right) - \dot{\varphi}(s) \right| + \frac{1}{2} \left| \dot{\varphi}\left(\frac{-r}{N}\right) \right| + \frac{N}{r} |F_1^N| \right\}^2 ds \\ & \leq \int_{-r/N}^0 \left\{ \mathcal{E}_1^N + \frac{K}{2} + \frac{N}{r} |F_1^N| \right\}^2 ds \leq \frac{r}{N} \left\{ \frac{3}{2} \mathcal{E}_1^N + \frac{K}{2} \right\}^2. \end{aligned}$$

These estimates plus the observation that $\mathcal{E}_j^N \rightarrow 0$ uniformly in j as $N \rightarrow \infty$ (uniform continuity of $\dot{\varphi}$ on $[-r, 0]$) yield the desired results and the proof of the lemma is thus completed.

COROLLARY 3.1. *Let $z \in \tilde{\mathcal{D}} \equiv \{(\varphi(0), \varphi) | \varphi \in W_\infty^{(2)}\}$. Then*

$$|\mathcal{A}^N z - \mathcal{A} z| = O\left(\frac{1}{\sqrt{N}}\right),$$

where here the O -term depends, of course, also on z .

Proof. For $\varphi \in C^{(1)}$ and $s \in [t_j^N, t_{j-1}^N]$ one has

$$\begin{aligned} |\varphi_j^N - \varphi(s)| &= \left| \frac{N}{r} \int_{t_j^N}^{t_{j-1}^N} [\varphi(\xi) - \varphi(s)] d\xi \right| \\ &\leq K \frac{N}{r} \int_{t_j^N}^{t_{j-1}^N} |\xi - s| d\xi \leq K \frac{r}{N}, \end{aligned}$$

where $K = \sup \{|\dot{\varphi}(\xi)| \mid \xi \in [-r, 0]\}$. Thus one finds

$$\left| \sum \varphi_j^N \chi_j^N - \varphi \right|_{L_2}^2 = \sum \int_{t_j^N}^{t_{j-1}^N} |\varphi_j^N - \varphi(s)|^2 ds \leq K^2 \frac{r^3}{N^2}$$

and hence

$$\left| \sum \varphi_j^N \chi_j^N - \varphi \right|_{L_2} = O\left(\frac{1}{N}\right).$$

From the above estimates one sees immediately that the convergence in part (i) of the proof of Lemma 3.2 is $O(1/N)$. Next, for $\varphi \in W_\infty^{(2)}$, one observes that \mathcal{E}_j^N defined in (3.16) satisfies

$$|\mathcal{E}_j^N| \leq \tilde{K}(r/N)$$

where $\tilde{K} \equiv |\ddot{\varphi}|_{L_\infty}$. It then follows from the estimates in part (ii) of the proof of Lemma 3.2 that

$$\left| \sum \frac{N}{r} \{\varphi_{j-1}^N - \varphi_j^N\} - \dot{\varphi} \right|_{L_2}^2 \leq \frac{25}{4} \tilde{K}^2 \frac{r^3}{N^2} + \frac{r}{N} \left\{ \frac{3r}{2N} \tilde{K} + \frac{K}{2} \right\}^2$$

and thus the convergence in (ii) is $O(1/\sqrt{N})$. Combining these simple observations, one obtains the claimed order results.

LEMMA 3.3. *Let \mathcal{D} be as in Lemma 3.2 and \mathcal{A} as in (3.12). Then there exists a real number γ_0 such that for λ complex with $\operatorname{Re}(\lambda) > \gamma_0$, $(\mathcal{A} - \lambda I)\mathcal{D}$ is dense in Z .*

Proof. From (2.5) we have the existence of γ_0 such that $\operatorname{Re}(\lambda) > \gamma_0$ implies $\lambda \in \rho(\mathcal{A})$. That is, for such λ , given $w \in Z$, $(\mathcal{A} - \lambda I)z = w$ has a unique solution $z \in Z$. Fix λ with $\operatorname{Re}(\lambda) > \gamma_0$. Then for $(\xi, \psi) \in Z$

$$(3.19) \quad (L(\varphi) - \lambda\varphi(0), \dot{\varphi} - \lambda\varphi) = (\mathcal{A} - \lambda I)(\varphi(0), \varphi) = (\xi, \psi)$$

has a unique solution $(\varphi(0), \varphi) \in \mathcal{D}(\mathcal{A})$. Thus for $\psi \in L_2(-r, 0)$, the solution φ of (3.19) must be in $W_2^{(1)}(-r, 0)$. However if $(\xi, \psi) \in R^n \times \mathcal{C}(-r, 0)$ (which is dense in Z) we see that the solution $(\varphi(0), \varphi)$ of (3.19) must be such that $\dot{\varphi} = \lambda\varphi + \psi$ is actually continuous. Thus given any $(\xi, \psi) \in R^n \times \mathcal{C}(-r, 0)$, there exists $(\varphi(0), \varphi)$ in \mathcal{D} solving (3.19). It follows that $R^n \times \mathcal{C}(-r, 0) \subset (\mathcal{A} - \lambda I)\mathcal{D}$ and thus $(\mathcal{A} - \lambda I)\mathcal{D}$ is dense in Z .

Before stating the next lemma let us return momentarily to the approximate system (3.6) for (3.8) when the approximations are chosen as in (3.9)–(3.11), (3.13). The differentiated form (3.7)

$$\begin{aligned} (3.7) \quad \dot{z}^N(t) &= \mathcal{A}^N z^N(t) + (f(t), 0), \\ z^N(0) &= P^N(\varphi(0), \varphi) \end{aligned}$$

can be written, upon defining the functions $e_0^N, e_1^N, \dots, e_N^N$ by

$$(3.20) \quad e_0^N = (1, 0), \quad e_j^N = (0, \chi_j^N), \quad j = 1, 2, \dots, N,$$

and defining n -vector (column) coefficient functions $t \rightarrow w_j^N(t)$ by

$$(3.21) \quad z^N(t) = \sum_{j=0}^N w_j^N(t) e_j^N,$$

as the equivalent $n(N+1)$ dimensional vector system

$$(3.22) \quad \begin{aligned} \dot{w}^N(t) &= A^N w^N(t) + \text{col}(f(t), 0, \dots, 0), \\ w^N(0) &= \text{col}(\varphi(0), \varphi_1^N, \dots, \varphi_N^N), \end{aligned}$$

where $w^N = \text{col}(w_0^N, w_1^N, \dots, w_N^N)$ and the $n(N+1)$ square matrix A^N is defined by

$$(3.23) \quad A^N \equiv \begin{bmatrix} A_0 & \frac{r}{N} D_1^N & \cdot & \cdot & \frac{r}{N} D_{N-1}^N & A_1 + \frac{r}{N} D_N^N \\ \frac{N}{r} I & -\frac{N}{r} I & 0 & \cdot & \cdot & 0 \\ 0 & \frac{N}{r} I & -\frac{N}{r} I & \cdot & \cdot & \cdot \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdot & \cdot & 0 & \frac{N}{r} I & -\frac{N}{r} I \end{bmatrix}.$$

Here I is the $n \times n$ identity matrix.

The next lemma shows that, even though the approximations discussed in this paper are not eigenfunction expansions, the spectrum $\sigma(A^N) = \pi(A^N)$ of the approximating system ((3.7) with $f=0$) "converges" (loosely speaking; we make this precise in the remark below) to the spectrum $\sigma(\mathcal{A}) = \pi(\mathcal{A})$ of the original system ((3.8) with $f=0$). Recall (see (2.4)) that

$$\pi(\mathcal{A}) = \{\lambda \in \mathbb{C} | \det \Delta(\lambda) = 0\}$$

where for (3.12) one finds $\Delta(\lambda) = \lambda I - A_0 - A_1 e^{-\lambda r} - \int_{-r}^0 D(\theta) e^{\lambda \theta} d\theta$ while

$$\pi(A^N) = \{\lambda \in \mathbb{C} | \det \Delta^N(\lambda) = 0\}$$

with $\Delta^N(\lambda) \equiv \lambda I_N - A^N$, A^N being given in (3.23) and I_N being the $n(N+1)$ square identity matrix.

LEMMA 3.4. For fixed $\lambda \in \mathbb{C}$ the sequence of ratios

$$\frac{\det \Delta^N(\lambda)}{(\lambda + N/r)^{nN}}$$

converges as $N \rightarrow \infty$ to $\det \Delta(\lambda)$.

Proof. We fix $\lambda \in \mathbb{C}$. Then $\lambda \neq -N/r$ except possibly for at most one value of N and the quantities in the discussions below are well-defined for each value of N

(with at most one exception). Define, for $k = 1, 2, \dots, N$, the $n \times n$ matrices E_k by

$$(3.24) \quad E_k \equiv \sum_{j=k}^N \frac{r}{N} D_j^N \left(\frac{N}{r}\right)^{j-k} \left(\lambda + \frac{N}{r}\right)^{-j+k-1} + A_1 \left(\frac{N}{r}\right)^{N-k} \left(\lambda + \frac{N}{r}\right)^{-N+k-1}.$$

It is easily verified that for $k = 2, 3, \dots, N$

$$(3.25) \quad -E_k \frac{N}{r} + E_{k-1} \left(\lambda + \frac{N}{r}\right) - \frac{r}{N} D_{k-1}^N = 0.$$

We further note that

$$(3.26) \quad \left(\lambda + \frac{N}{r}\right) E_N - \frac{r}{N} D_N^N - A_1 = 0.$$

Next we define the $n(N+1)$ square matrix \mathcal{E} by

$$\mathcal{E} \equiv \begin{bmatrix} I & E_1 & \cdots & E_n \\ 0 & I & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdot & \cdot & 0 & I \end{bmatrix}$$

where again I is the $n \times n$ identity, and then consider the matrix $\mathcal{E}(\lambda I_N - A^N)$. From (3.25) and (3.26) we find immediately that

$$\mathcal{E}(\lambda I_N - A^N) = \begin{bmatrix} \lambda I - A_0 - \frac{N}{r} E_1 & 0 & \cdot & \cdot & 0 \\ -\frac{N}{r} I & \left(\lambda + \frac{N}{r}\right) I & \cdot & \cdot & \cdot \\ 0 & -\frac{N}{r} I & \cdot & \cdot & \cdot \\ \vdots & \vdots & & & 0 \\ 0 & \cdot & 0 & -\frac{N}{r} I & \left(\lambda + \frac{N}{r}\right) I \end{bmatrix}.$$

We thus have (e.g. see [43, p. 319]), using the fundamental properties of determinants,

$$\begin{aligned} \det \mathcal{E} \det (\lambda I_N - A^N) &= \det \{\mathcal{E}(\lambda I_N - A^N)\} \\ &= \det \left(\lambda I - A_0 - \frac{N}{r} E_1 \right) \left\{ \det \left[\left(\lambda + \frac{N}{r} \right) I \right] \right\}^N. \end{aligned}$$

But $\det \mathcal{E} = 1$ and we thus find

$$(3.27) \quad \begin{aligned} \det \Delta^N(\lambda) &= \det \left\{ \left(\lambda I - A_0 - \frac{N}{r} E_1 \right) \left(\lambda + \frac{N}{r} \right)^N \right\} \\ &= \det \left\{ \left[\lambda I - A_0 - \sum_{j=1}^N \frac{r}{N} D_j^N \left(\frac{N}{r} \right)^j \left(\lambda + \frac{N}{r} \right)^{-j} - A_1 \left(\frac{N}{r} \right)^N \left(\lambda + \frac{N}{r} \right)^{-N} \right] \left(\lambda + \frac{N}{r} \right)^N \right\}. \end{aligned}$$

We observe that the expression in (3.27) is the same that one obtains by formally expanding (using cofactors) along the first "row" of $\lambda I_N - A^N$. That derivation of (3.27) is, however, valid only if the n square matrices A_0 , A_1 , D are scalar (i.e. $n = 1$).

Rewriting (3.27) slightly, we have shown

$$\det \Delta^N(\lambda) = \det \left\{ \left[\lambda I - A_0 - A_1 \left(1 + \frac{\lambda r}{N} \right)^{-N} - \sum_{j=1}^N \frac{r}{N} D_j^N \left(1 + \frac{\lambda r}{N} \right)^{-j} \right] \left(\lambda + \frac{N}{r} \right)^N \right\}.$$

Consider the term involving the \sum , which can be rewritten

$$\begin{aligned} \sum_{j=1}^N \frac{r}{N} D_j^N \left(1 + \frac{\lambda r}{N} \right)^{-j} &= \sum_{j=1}^N \int_{t_j^N}^{t_{j-1}^N} D(\theta) \left(1 + \frac{\lambda r}{N} \right)^{-j} d\theta \\ &= \int_{-r}^0 D(\theta) \left\{ \sum_{j=1}^N \tilde{\chi}_j^N(\theta) \left(1 + \frac{\lambda r}{N} \right)^{-j} \right\} d\theta \\ &= \int_{-r}^0 d(\theta) f_N(\theta) d\theta \end{aligned}$$

where we now define $\tilde{\chi}_j^N$ as the characteristic function of $(-jr/N, -jr/N + r/N) = (t_j^N, t_{j-1}^N)$. It is not difficult to argue that $f_N(\theta) \rightarrow e^{\lambda\theta}$ as $N \rightarrow \infty$ for a.e. $\theta \in (-r, 0)$ and that this convergence is dominated. We thus find

$$\int_{-r}^0 D(\theta) f_N(\theta) d\theta \rightarrow \int_{-r}^0 D(\theta) e^{\lambda\theta} d\theta$$

as $N \rightarrow \infty$.

Recalling that $(1 + \lambda r/N)^{-N} \rightarrow e^{-\lambda r}$ as $N \rightarrow \infty$, we obtain

$$\begin{aligned} \lambda I - A_0 - A_1 \left(1 + \frac{\lambda r}{N} \right)^{-N} - \sum_{j=1}^N \frac{r}{N} D_j^N \left(1 + \frac{\lambda r}{N} \right)^{-j} \\ \rightarrow \lambda I - A_0 - A_1 e^{-\lambda r} - \int_{-r}^0 D(\theta) e^{\lambda\theta} d\theta. \end{aligned}$$

But for fixed $\lambda \in \mathbb{C}$ (recall then $\lambda = -N/r$ for at most one value of N) it follows from (3.27) that (for all but possibly one N)

$$\det \Delta^N(\lambda) = (\lambda + N/r)^{nN} \det \tilde{\Delta}^N(\lambda)$$

where

$$\tilde{\Delta}^N(\lambda) \equiv \left\{ \lambda I - A_0 - A_1 \left(1 + \frac{\lambda r}{N} \right)^{-N} - \sum_{j=1}^N \frac{r}{N} D_j^N \left(1 + \frac{\lambda r}{N} \right)^{-j} \right\}.$$

Thus, since $\lim \det \tilde{\Delta}^N(\lambda) = \det \Delta(\lambda)$, we have that the desired conclusion obtains for each fixed $\lambda \in \mathbb{C}$.

Remark. To be precise, we proved in the above lemma that $\det \tilde{\Delta}^N(\lambda) \rightarrow \det \Delta(\lambda)$ as $N \rightarrow \infty$. Given the definitions of $\pi(A^N)$ and $\pi(\mathcal{A})$, we see this essentially yields that the sequence of characteristic functions determining the eigenvalues of A^N (save possibly the roots $\lambda = -N/r$) has as its limit the characteristic function of the original hereditary system. Indeed, one can use the results developed here in § 3 to argue (we shall not give these since they involve rather standard ideas and results from elementary complex analysis) that given any $\lambda_0 \in \pi(\mathcal{A})$, there exists a sequence $\{\lambda_0^k\}$ of eigenvalues of the approximating systems such that $\lambda_0^k \rightarrow \lambda_0$. Conversely, if $\lambda_0 \notin \pi(\mathcal{A})$ and \mathcal{N} is any finite neighborhood of λ_0 with $\mathcal{N} \cap \pi(\mathcal{A}) = \emptyset$, then $\pi(\mathcal{A}^N) \cap \mathcal{N} = \emptyset$ for all N sufficiently large.

LEMMA 3.5. For Z^N , \mathcal{A}^N as defined in (3.9), (3.10), we have $(Z^N)^\perp \subset \ker(\mathcal{A}^N)$ and the projections P^N defined in (3.13) are (continuous) orthogonal projections along $(Z^N)^\perp$ onto Z^N .

Proof. Let $(\xi, \psi) \in (Z^N)^\perp$ so that

$$\xi \eta + \int_{-\tau}^0 \psi(\theta) \varphi(\theta) d\theta = 0$$

for all $(\eta, \varphi) \in Z^N$. This yields immediately that $\xi = 0$ and

$$\sum_{j=1}^N \int_{t_j^N}^{t_{j+1}^N} \psi(\theta) v_j^N d\theta = 0$$

for v_j^N arbitrary in R^n . This latter may be equivalently written $\sum_{j=1}^N \psi_j^N v_j^N = 0$, from which it follows that $\psi_j^N = 0$, $j = 1, 2, \dots, N$. Thus $(\xi, \psi) \in (Z^N)^\perp$ implies $\xi = 0$, $\psi_j^N = 0$, $j = 1, 2, \dots, N$, and from (3.10) one thus obtains $\mathcal{A}^N(\xi, \psi) = 0$. We note that this also implies immediately that $P^N = 0$ on $(Z^N)^\perp$. Since it is clear that P^N is a projection onto the closed subspace Z^N of Z we thus have $P^N = I$ on Z^N , $P^N = 0$ on $(Z^N)^\perp$ and P^N is the (canonical) orthogonal projection of Z onto Z^N along $(Z^N)^\perp$ (e.g., see [58, pp. 241, 333]).

Remark. Since Z^N , $(Z^N)^\perp$ are clearly invariant under \mathcal{A}^N we find that $(Z^N, (Z^N)^\perp)$ completely reduce [58, p. 268] the operators $S^N(t) = e^{\mathcal{A}^N t}$, $t \geq 0$. Indeed we have $S^N(t) = I$ on $(Z^N)^\perp$ while $S^N(t) = S^N(t)P^N$ on Z^N . Furthermore since the $(Z^N, (Z^N)^\perp)$ completely reduce $S^N(t)$ (see [58, p. 270]) standard arguments yield immediately that $\sigma(S^N(t)) = \exp\{\pi(A^N)t\} \cup \{1\}$. Note further that from Lemma 3.1 and standard results we have that

$$|S^N(t)| \leq M_N e^{\beta_N t}, \quad t \geq 0,$$

for constants β_N and M_N . That one can in actuality obtain such a bound with M and β independent of N is the conclusion of the next lemma.

LEMMA 3.6 (Stability). *There exist constants M, β such that $\mathcal{A}^N \in G(M, \beta)$ for all N .*

Proof. Since $Z = R^n \times L_2$ is a Hilbert space, it suffices [32, pp. 85–90], [42, pp. 14–18] to show that there exists β independent of N such that $\langle \mathcal{A}^N z, z \rangle \leq \beta \langle z, z \rangle$ for all $z \in Z$ and all N (i.e., $\mathcal{A}^N - \beta I$ is maximal dissipative for each N).

Let $z = (\eta, \varphi) \in Z$. Recalling (3.10), (3.11) we have

$$\begin{aligned} \langle \mathcal{A}^N z, z \rangle &= \left\langle A_0 \varphi_0^N + A_1 \varphi_1^N + \sum \frac{r}{N} D_j^N \varphi_j^N, \varphi_0^N \right\rangle_{R^n} + \sum_{j=1}^N \left\langle \frac{N}{r} \{\varphi_{j-1}^N - \varphi_j^N\} \chi_j^N, \varphi \right\rangle_{L_2} \\ &= T_1 + T_2. \end{aligned}$$

Considering first the second term in this sum, we obtain

$$T_2 = \sum_{j=1}^N \{ \langle \varphi_{j-1}^N, \varphi_j^N \rangle_{R^n} - \langle \varphi_j^N, \varphi_j^N \rangle_{R^n} \} \leq \sum_{j=1}^N \{ \frac{1}{2} (|\varphi_{j-1}^N|^2 + |\varphi_j^N|^2) - |\varphi_j^N|^2 \},$$

where we have used the fundamental inequality $2ab \leq a^2 + b^2$. Using this inequality once again we find

$$\begin{aligned} T_1 &\leq |A_0| |\varphi_0^N|^2 + |A_1| |\varphi_1^N| |\varphi_0^N| + \sum \frac{r}{N} |D_j^N| |\varphi_j^N| |\varphi_0^N| \\ &\leq |A_0| |\varphi_0^N|^2 + \frac{1}{2} \{ |A_1|^2 |\varphi_0^N|^2 + |\varphi_1^N|^2 \} + \sum \frac{r}{N} |D_j^N| |\varphi_j^N| |\varphi_0^N|. \end{aligned}$$

Combining these estimates for T_1 and T_2 we have

$$T_1 + T_2 \leq \{ |A_0| + \frac{1}{2} (|A_1|^2 + 1) \} |\varphi_0^N|^2 + \sum_{j=1}^N \frac{r}{N} |D_j^N| |\varphi_j^N| |\varphi_0^N|.$$

But with use again of the above mentioned inequality along with that of Cauchy-Schwarz we further argue

$$\begin{aligned} \sum_{j=1}^N \frac{r}{N} |D_j^N| |\varphi_j^N| |\varphi_0^N| &\leq \sum_j \frac{N}{r} \left\{ |\varphi_0^N| \int_{t_j^N}^{t_{j-1}^N} |D(s)| ds \right\} \left\{ \int_{t_j^N}^{t_{j-1}^N} |\varphi(s)| ds \right\} \\ &\leq \sum_j \frac{N}{r} \frac{1}{2} \left[\left\{ |\varphi_0^N| \int_{t_j^N}^{t_{j-1}^N} |D| \right\}^2 + \left\{ \int_{t_j^N}^{t_{j-1}^N} |\varphi| \right\}^2 \right] \\ &\leq \sum_j \frac{N}{r} \frac{1}{2} \left[|\varphi_0^N|^2 \frac{r}{N} \int_{t_j^N}^{t_{j-1}^N} |D|^2 + \frac{r}{N} \int_{t_j^N}^{t_{j-1}^N} |\varphi|^2 \right] \\ &= \frac{1}{2} |\varphi_0^N|^2 \int_{-r}^0 |D|^2 + \frac{1}{2} \int_{-r}^0 |\varphi|^2. \end{aligned}$$

Recalling that $\eta = \varphi_0^N$ we thus obtain

$$\langle \mathcal{A}^N z, z \rangle \leq \{ |A_0| + \frac{1}{2} (1 + |A_1|^2 + |D|_{L_2}^2) \} |\eta|^2 + \frac{1}{2} |\varphi|_{L_2}^2,$$

or, by choosing $\beta = |A_0| + \frac{1}{2} (1 + |A_1|^2 + |D|^2)$,

$$\langle \mathcal{A}^N z, z \rangle \leq \beta \{ |\eta|^2 + |\varphi|^2 \} = \beta \langle z, z \rangle.$$

Recalling Theorem 3.1 and in particular the conditions (3.1)–(3.3), we now conclude directly from Lemmas 3.2, 3.3 and 3.6 that for Z^N , \mathcal{A}^N as defined in (3.9), (3.10) one has $S^N(t)z = e^{\mathcal{A}^N t} z \rightarrow S(t)z$ for every $z \in Z$. (In fact one can use Corollary 3.1 to show that this convergence is $O(1/\sqrt{N})$ for sufficiently nice z .) Further we have seen that P^N defined in (3.13) satisfies $P^N z \rightarrow z$ for $z \in Z$ and our

previous remarks reveal that $z^N(t; f) \rightarrow z(t; f)$ where z^N is the solution of (3.6) (or (3.7)). We note that this yields $w_0^N(t; f) \rightarrow x(t; f)$ where w^N is the solution of (3.22) and x is the solution of (3.8). In fact this convergence ($z^N \rightarrow z$, $w_0^N \rightarrow x$) is actually uniform in f for f lying in a bounded subset of $L_2(0, t_1)$. The arguments for this claim are given in the proof of Theorem 3.1 of [2]. It also follows immediately (we shall use this result in § 4 below) that for each fixed $t \in [0, t_1]$, $z^N(t; f^K) \rightarrow z(t; f)$ as $N, K \rightarrow \infty$ whenever $f^K \rightarrow f$ weakly in $L_2(0, t_1)$.

To conclude this section, we turn briefly to the modifications needed to extend the above arguments and results (for (3.8)) to treat systems with multiple discrete delay terms. Let us consider the modifications required when the lags are given by $h_0 = 0$, $h_1 = r/2$, $h_2 = r$. (Similar considerations are valid in the case of a finite number of commensurate delays.) We thus consider in place of (3.8) the system

$$(3.28) \quad \begin{aligned} \dot{x}(t) &= A_0 x(t) + A_1 x(t - r/2) + A_2 x(t - r) + \int_{-r}^0 D(\theta) x(t + \theta) d\theta + f(t), \\ x(0) &= \varphi(0), \quad x_0 = \varphi. \end{aligned}$$

In this case we partition $[-r, 0]$ into $2N$ subintervals $[t_j^N, t_{j-1}^N]$ where now $t_j^N = -jr/2N$, $j = 0, 1, \dots, 2N$ and Z^N has the form

$$Z^N \equiv \left\{ (\eta, \varphi) \in Z \mid \eta \in R^n, \varphi = \sum_{j=1}^{2N} v_j^N \chi_j^N, v_j^N \in R^n \right\}$$

where now $\chi_j^N = \chi_{[t_j^N, t_{j-1}^N]}$, $j = 2, 3, \dots, 2N$ and $\chi_1^N = \chi_{[-r/(2N), 0]}$.

The approximating operators $\mathcal{A}^N: Z \rightarrow Z^N$ of (3.10) become

$$\mathcal{A}^N(\eta, \varphi) \equiv \left(A_0 \varphi_0^N + A_1 \varphi_N^N + A_2 \varphi_{2N}^N + \sum_{j=1}^{2N} \frac{r}{2N} D_j^N \varphi_j^N, \sum_{j=1}^{2N} \frac{2N}{r} (\varphi_{j-1}^N - \varphi_j^N) \chi_j^N \right)$$

where

$$\varphi_0^N \equiv \eta, \quad \varphi_j^N \equiv \frac{2N}{r} \int_{t_j^N}^{t_{j-1}^N} \varphi(s) ds, \quad D_j^N \equiv \frac{2N}{r} \int_{t_j^N}^{t_{j-1}^N} D(s) ds, \quad j = 1, 2, \dots, 2N.$$

Observing that $\varphi_N^N = 2N/r \int_{-r/2}^{-r/(2N)} \varphi(s) ds \rightarrow \varphi(-r/2)$ as $N \rightarrow \infty$, one sees that the arguments that $\mathcal{A}^N z \rightarrow \mathcal{A}z$ are essentially the same as those given in the proof of Lemma 3.2.

Letting $z^N(t) = \sum_{j=0}^{2N} w_j^N(t) e_j^N$ where the e_j^N are defined similar to (3.20), $j = 0, 1, \dots, 2N$, the approximate system (3.22) is an $n(2N+1)$ dimensional vector system with coefficient matrix A^N given by

$$A^N = \begin{bmatrix} A_0 & \frac{r}{2N} D_1^N & \cdots & \frac{r}{2N} D_{2N-1}^N & A_1 + \frac{r}{2N} D_N^N & \frac{r}{2N} D_{N+1}^N & \cdots & \frac{r}{2N} D_{2N-1}^N & A_2 + \frac{r}{2N} D_{2N}^N \\ \frac{2N}{r} I & -\frac{2N}{r} I & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & \frac{2N}{r} I & -\frac{2N}{r} I & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & \frac{2N}{r} I & -\frac{2N}{r} I \end{bmatrix}$$

and initial data

$$w^N(0) = \text{col}(\varphi(0), \varphi_1^N, \dots, \varphi_{2N}^N).$$

Furthermore, modifications of the arguments in the proof of Lemma 3.4 yield that

$$\det(\lambda I_N - A^N) = \det \left\{ \left[\lambda I - A_0 - A_1 \left(1 + \frac{\lambda r}{2N} \right)^{-N} - A_2 \left(1 + \frac{\lambda r}{2N} \right)^{-2N} - \sum_{j=1}^{2N} \frac{r}{2N} D_j^N \left(\frac{2N}{r} \right)^j \left(\lambda + \frac{2N}{r} \right)^{-j} \right] \left(\lambda + \frac{2N}{r} \right)^{2N} \right\},$$

and as $N \rightarrow \infty$, the expression in square brackets has the limit

$$\Delta(\lambda) = \lambda I - A_0 - A_1 e^{-\lambda r/2} - A_2 e^{-\lambda r} - \int_{-r}^0 D(\theta) e^{\lambda \theta} d\theta,$$

which is the characteristic function for (3.28) (with of course $f \equiv 0$).

An extension of the results of Lemma 3.5 to include systems (3.28) is trivial. However, the question of stability is slightly more difficult and we shall return to that momentarily.

For general multiple (not necessarily commensurate) delays the system equations have the form

$$(3.29) \quad \dot{x}(t) = \sum_{i=0}^{\nu} A_i x(t - h_i) + \int_{-r}^0 D(\theta) x(t + \theta) d\theta + f(t)$$

where $0 = h_0 < h_1 < \dots < h_{\nu} \leq r$. The approximating infinitesimal generators \mathcal{A}^N associated with this equation (i.e. the analogues of (3.10)) are given by

$$(3.30) \quad \mathcal{A}^N(\eta, \varphi) = \left(A_0 \eta + \sum_{i=1}^{\nu} A_i \varphi_{j_i}^N + \sum_{j=1}^N \frac{r}{N} D_j^N \varphi_j^N, \sum_{j=1}^N \frac{N}{r} (\varphi_{j-1}^N - \varphi_j^N) \chi_j^N \right)$$

where the index $j_i = l$ is chosen so that $h_i \in [t_l^N, t_{l-1}^N)$ for each $i = 1, 2, \dots, \nu$. The fact that the delays h_i are now allowed to be noncommensurate (and not necessarily a multiple of the basic partition increment) will greatly complicate some of the notation (e.g., in the analogues of (3.23) the matrices A_1, \dots, A_{ν} need not stay in the "same" columns as we change the partition index N); nonetheless, all the preceding arguments can with appropriate modifications be extended so that analogues of Lemmas 3.1–3.5 hold for the general systems (3.29).

To close this section, we return finally to the question of stability of the approximations for (3.29). While the technical details are slightly more complicated, the essence of the ideas given in the proof above of Lemma 3.6 combined with ideas due to Webb [63] (see also Reddien and Webb, *Numerical approximation of nonlinear FDE with L_2 initial functions*, to appear) concerning the use of weighting functions in L_2 type norms (we are grateful to both Professor Webb and Professor Franz Kappel who independently suggested to us that simplified stability arguments might be given via use of such weighted norms) can be

employed to give a proof of stability (Lemma 3.6) of the “averaging” approximations for the general systems (3.29). (In this case, however, the S^N are no longer contraction semigroups in Z but one does have $\mathcal{A}^N \in G(M, \beta)$ with $M > 1$ and M, β independent of N .) Thus, all the results of § 3 (in particular, convergence of the “averaging” approximations) obtain for general delay systems of the form (3.29).

We sketch briefly these stability arguments for the case $D \equiv 0$. Assume, without loss of generality, that N is sufficiently large so that the lags h_i , $i = 1, 2, \dots, \nu$, lie in distinct subintervals in the partition of $[-r, 0]$ by $\{t_j^N\}$. Let $J^N = \{j_1, j_2, \dots, j_\nu\}$ be that subset of the indices $\{1, 2, \dots, N\}$ such that $h_i \in [t_{j_i}^N, t_{j_i-1}^N)$ for $i = 1, 2, \dots, \nu$. Define the piecewise constant function τ_N (for each fixed N) on $[-r, 0)$ by

$$\tau_N(\theta) = a_j^N, \quad \theta \in [t_j^N, t_{j-1}^N), \quad j = 1, 2, \dots, N,$$

when the a_j^N 's are defined as follows:

$$a_{N+1}^N = 1$$

and for $j = N, N-1, \dots, 1$,

$$a_j^N = \begin{cases} a_{j+1}^N + 1 & \text{if } j \in J^N, \\ a_{j+1}^N & \text{if } j \notin J^N. \end{cases}$$

Let $|\cdot|_{\tau_N}$ and $\langle \cdot, \cdot \rangle_{\tau_N}$ denote respectively the norm and inner product in Z using the weighting function τ_N , i.e.

$$\langle (\eta, \varphi), (\zeta, \psi) \rangle_{\tau_N} = \langle \eta, \zeta \rangle_{R^n} + \int_{-r}^0 \varphi \psi \tau_N.$$

Using (3.30) (with $D = 0$) and inequalities similar to those employed in the proof of Lemma 3.6 above, we can argue that for any $z = (\eta, \varphi) \in Z$,

$$\begin{aligned} \langle \mathcal{A}^N z, z \rangle_{\tau_N} &\leq |A_0| |\eta|^2 + \sum_{i=1}^{\nu} \left\{ \frac{1}{2} |A_i|^2 |\eta|^2 + \frac{1}{2} |\varphi_{j_i}^N|^2 \right\} \\ &\quad + \sum_{j=1}^N \left\{ \frac{1}{2} |\varphi_{j-1}^N|^2 + \frac{1}{2} |\varphi_j^N|^2 - |\varphi_j^N|^2 \right\} a_j^N \\ &\leq \left(|A_0| + \frac{1}{2} \sum_{i=1}^{\nu} |A_i|^2 + \frac{a_1^N}{2} \right) |\eta|^2 \\ &\quad + \frac{1}{2} \sum_{i=1}^{\nu} |\varphi_{j_i}^N|^2 + \frac{1}{2} \sum_{j=1}^N (a_{j+1}^N - a_j^N) |\varphi_j^N|^2 \\ &= \left(|A_0| + \frac{1}{2} \sum_{i=1}^{\nu} |A_i|^2 + \frac{a_1^N}{2} \right) |\eta|^2. \end{aligned}$$

Noting further that $a_1^N \leq \nu + 1$ for all N (and indeed τ_N satisfies $1 \leq \tau_N(\theta) \leq \nu + 1$), we thus find for all $z \in Z$

$$\langle \mathcal{A}^N z, z \rangle_{\tau_N} \leq \beta |\eta|^2 \leq \beta \langle z, z \rangle_Z \leq \beta \langle z, z \rangle_{\tau_N}$$

since $|\cdot|_Z \leq |\cdot|_{\tau_N} \leq \gamma |\cdot|_Z$ where γ is independent of N . It follows that $\|S^N(t)\| \leq e^{\beta t}$

where $\|\cdot\|$ is the operator norm induced by the τ_N norm on Z . From the equivalency of the Z -norm and the τ_N norm on Z (independent of N) one then obtains $|S^N(t)| \leq M e^{\beta t}$ where $|\cdot|$ in this inequality is the operator norm induced by the original norm on Z and where M and β are independent of N .

4. Approximation in optimal control problems. We apply the approximation ideas discussed above to control problems governed by (3.8) where $f(t) = B(t)u(t)$, $t \rightarrow B(t)$ a bounded measurable $\mathcal{L}_{n,m}$ valued function and $u \in \mathcal{U}$, \mathcal{U} a closed convex subset of $L_2^m(0, t_1)$. Throughout we assume $(\eta, \varphi) = (\varphi(0), \varphi) \in \mathcal{D}(\mathcal{A})$.

Let $g_0: R^n \rightarrow R^1$, $g_1: L_2(-r, 0) \rightarrow R^1$, and $g_2: L_2^n(0, t_1) \times L_2^m(0, t_1) \rightarrow R^1$ be continuous and define

$$(4.1) \quad J(\xi, \gamma, q, v) \equiv g_0(\xi - \zeta(0)) + g_1(\gamma - \zeta) + g_2(q, v)$$

where $(\zeta(0), \zeta) \in Z$ is given.

Given $u \in \mathcal{U}$, let $x = x(\cdot; u) = x(u)$ denote the solution to (3.8) on $[0, t_1]$ corresponding to $f = Bu$. We then define

$$(4.2) \quad \Phi(u) \equiv J(x(t_1), x_{t_1}, x(u), u)$$

and our control problem is taken as:

(\mathcal{P}) Minimize Φ over \mathcal{U} .

A typical example to which our theory may be applied is one where the g_i are given by

$$\begin{aligned} g_0(\xi - \zeta(0)) &\equiv (\xi - \zeta(0))Q_0(\xi - \zeta(0)), \\ g_1(\gamma - \zeta) &\equiv \int_{-r}^0 (\gamma(\theta) - \zeta(\theta))Q_1(\gamma(\theta) - \zeta(\theta)) d\theta, \\ g_2(q, v) &\equiv \int_0^{t_1} \{q(s)\mathcal{W}q(s) + v(s)\mathcal{R}v(s)\} ds, \end{aligned}$$

with \mathcal{W} , $Q_i \in \mathcal{L}_{n,n}$, $\mathcal{R} \in \mathcal{L}_{m,m}$, and $Q_i \geq 0$, $\mathcal{W} \geq 0$, $\mathcal{R} > 0$. In this case one finds

$$\begin{aligned} (4.3) \quad \Phi(u) &= (x(t_1) - \zeta(0))Q_0(x(t_1) - \zeta(0)) \\ &+ \int_{-r}^0 (x_{t_1}(\theta) - \zeta(\theta))Q_1(x_{t_1}(\theta) - \zeta(\theta)) d\theta \\ &+ \int_0^{t_1} \{x(s)\mathcal{W}x(s) + u(s)\mathcal{R}u(s)\} ds, \end{aligned}$$

where x is the solution of (3.8).

Introducing the approximations of § 3 and writing z^N of (3.7) as $z^N(t; u) = (x^N(t; u), y^N(t; u))$ where $x^N(t; u) \in R^n$, $y^N(t; u) \in L_2(-r, 0)$ (i.e., in the notation of § 3, $x^N(t) = w_0^N(t)$ and $y^N(t) = \sum_{j=1}^N w_j^N(t)\chi_j^N$), we define the approximate problems for problem (\mathcal{P}) above as

(\mathcal{AP})_N Minimize Φ^N over \mathcal{U} ,

where

$$(4.4) \quad \Phi^N(u) \equiv J^N(x^N(t_1; u), y^N(t_1; u), x^N(u), u)$$

with

$$(4.5) \quad J^N(\xi, \gamma, q, v) \equiv g_0(\xi - \zeta(0)) + g_1(\gamma - \zeta^N) + g_2(q, v).$$

Here $\zeta^N \equiv \sum_{j=1}^N \zeta_j^N \chi_j^N$ (ζ_j^N defined as in (3.11)).

In the above mentioned example, the approximation to (4.3) becomes

$$(4.6) \quad \begin{aligned} \Phi^N(u) &= (w_0^N(t_1) - \zeta(0))Q_0(w_0^N(t_1) - \zeta(0)) \\ &+ \sum_{j=1}^N \frac{r}{N} (w_j^N(t_1) - \zeta_j^N)Q_1(w_j^N(t_1) - \zeta_j^N) \\ &+ \int_0^{t_1} \{w_0^N(s)\mathcal{W}w_0^N(s) + u(s)\mathcal{R}u(s)\} ds. \end{aligned}$$

From the results of § 3, we have immediately that for fixed $u \in \mathcal{U}$, $\Phi^N(u) \rightarrow \Phi(u)$ as $N \rightarrow \infty$ (Φ, Φ^N as in (4.2), (4.4)). In fact, we note that this convergence is uniform in u on bounded subsets of \mathcal{U} and further that $J^N(\xi, \gamma, q, v) \rightarrow J(\xi, \gamma, q, v)$ uniformly in $(\xi, \gamma, q, v) \in R^n \times L_2(-r, 0) \times L_2(0, t_1) \times L_2^m(0, t_1)$.

We make the further assumptions:

H(i) g_i is (continuous) convex, $i = 0, 1, 2$;

H(ii) $\Phi^N(v^N) \rightarrow +\infty$ if $|v^N| \rightarrow \infty$.

(This latter assumption obtains, for example, if $g_0, g_1 \geq 0$ and $g_2(x^N(v^N), v^N) \rightarrow +\infty$ as $|v^N| \rightarrow \infty$.) The assumptions regarding convexity and continuity that we make here are actually much stronger than necessary to carry out the rather standard arguments given below (e.g., see [8], [44]) and are made in this strong form only to simplify our presentation.

Under the hypotheses H(i), one finds that $u \rightarrow \Phi(u)$, $u \rightarrow \Phi^N(u)$ are convex and indeed so are J and J^N . Using well-known arguments one obtains that, under H(i), H(ii) solutions \bar{u} , \bar{u}^N for problems (\mathcal{P}) , $(\mathcal{AP})_N$ exist. If furthermore J (and hence J^N) is strictly convex, these solutions are unique.

From H(ii) we see readily that $\{\bar{u}^N\}$ must be bounded in $L_2^m(0, t_1)$ (if not, $|\bar{u}^{N_k}| \rightarrow \infty$ for some subsequence and then H(ii) contradicts the inequality $\Phi^{N_k}(\bar{u}^{N_k}) \leq \Phi^{N_k}(v) \rightarrow \Phi(v) < \infty$ for v arbitrary in \mathcal{U}).

Now let $\{\bar{u}^{N_k}\}$ be a subsequence of $\{\bar{u}^N\}$ such that $\bar{u}^{N_k} \rightharpoonup \tilde{u}$ for some $\tilde{u} \in \mathcal{U}$ (\mathcal{U} closed and convex is thus weakly closed). Then

$$z^{N_k}(t; \bar{u}^{N_k}) \rightarrow z(t; \tilde{u})$$

in Z for $t \in [0, t_1]$ and from the weak lower semicontinuity of J , we find for $v \in \mathcal{U}$

$$\begin{aligned} \Phi(\tilde{u}) &= J(x(t_1; \tilde{u}), y(t_1; \tilde{u}), x(\tilde{u}), \tilde{u}) \\ &\leq \liminf J(x^{N_k}(t_1; \bar{u}^{N_k}), y^{N_k}(t_1; \bar{u}^{N_k}), x^{N_k}(\bar{u}^{N_k}), \bar{u}^{N_k}) \\ &= \liminf J^{N_k}(x^{N_k}(t_1; \bar{u}^{N_k}), y^{N_k}(t_1; \bar{u}^{N_k}), x^{N_k}(\bar{u}^{N_k}), \bar{u}^{N_k}) \\ &= \liminf \Phi^{N_k}(\bar{u}^{N_k}) \leq \liminf \Phi^{N_k}(\bar{u}^{N_k}) \leq \liminf \Phi^{N_k}(v) = \Phi(v), \end{aligned}$$

which shows that \tilde{u} is in fact a solution of problem (\mathcal{P}) .

If one further has J strictly convex, standard arguments reveal that indeed the sequence $\{\bar{u}^N\}$ itself converges weakly to the unique solution \bar{u} of (\mathcal{P}) . We summarize these claims in the following theorem.

THEOREM 4.1. *Suppose H(i), H(ii) obtain. Then $\{\bar{u}^N\}$ has a subsequence $\{\bar{u}^{N_k}\}$ converging weakly to a control \bar{u} that is a solution of problem (\mathcal{P}) and furthermore $\Phi^{N_k}(\bar{u}^{N_k}) \rightarrow \Phi(\bar{u})$. If in fact J is strictly convex then the sequence $\{\bar{u}^N\}$ itself converges weakly in $L_2^m(0, t_1)$ to the unique solution \bar{u} of the problem (\mathcal{P}) and $\Phi^N(\bar{u}^N) \rightarrow \Phi(\bar{u})$.*

We remark that if Φ has the form given in (4.3), then under the above assumptions, one in fact finds that \bar{u}^N converges strongly in $L_2^m(0, t_1)$ to \bar{u} .

We turn next to report briefly on numerical results obtained using the ideas developed above. More details of these results (along with additional examples) may be found in [4]. We consider, for the sake of simplicity in demonstrating numerical aspects of the approximations discussed in this paper, a problem (\mathcal{P}) with Φ having the form (4.3) with $\zeta \equiv 0$, $Q_1 \equiv 0$, $Q_0 = \frac{1}{2}G$, $\mathcal{W} = \frac{1}{2}Q$, $\mathcal{R} = \frac{1}{2}R$ with $G \geq 0$, $Q \geq 0$, and $R > 0$. Thus we in fact consider the original problem (\mathcal{P}) with

$$(4.7) \quad \Phi(u) = \frac{1}{2}[x(t_1)Gx(t_1)] + \frac{1}{2} \int_0^{t_1} \{x(t)Qx(t) + u(t)Ru(t)\} dt.$$

For the system governing this problem we take (3.8) with $D \equiv 0$ and $f(t) = Bu(t)$ and $\mathcal{U} = L_2^m(0, t_1)$. The approximate problems $(\mathcal{AP})_N$ become ones of minimizing

$$(4.8) \quad \Phi^N(u) = \frac{1}{2}[w^N(t_1)G^Nw^N(t_1)] + \frac{1}{2} \int_0^{t_1} \{w^N(t)Q^Nw^N(t) + u(t)Ru(t)\} dt,$$

where w^N is the solution to (3.22) with $f(t) = Bu(t)$ and the $n(N+1)$ square matrices G^N , Q^N are defined by

$$G^N \equiv \begin{bmatrix} G & 0 & \cdots & 0 \\ 0 & \cdot & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & \cdot & \cdots & 0 \end{bmatrix}, \quad Q^N \equiv \begin{bmatrix} Q & 0 & \cdots & 0 \\ 0 & \cdot & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & \cdot & \cdots & 0 \end{bmatrix}.$$

In this case the problems $(\mathcal{AP})_N$ are nothing more than ODE linear regulator problems and standard techniques can be employed to readily solve these problems on the computer (see [4] for a complete discussion). We summarize now a comparison of the solutions of $(\mathcal{AP})_N$ with that of (\mathcal{P}) for several examples (again, complete details are found in [4]).

Example 4.1. Let $n = 1$, $t_1 = 3$, $G = 3$, $Q = 0$, and $R = 1$ so that (4.7) becomes

$$(4.9) \quad \Phi(u) = \frac{3}{2}[x(3)]^2 + \frac{1}{2} \int_0^3 [u(t)]^2 dt.$$

In (3.8) let $A_0 = 0$, $A_1 = 1$, $r = 1$, and $f(t) = u(t)$ (i.e., $B = 1$) while $\varphi \equiv 1$ so that the

governing system is

$$\dot{x}(t) = x(t-1) + u(t), \quad 0 \leq t \leq 3,$$

$$x_0 = 1.$$

Using the maximum principle for delay systems (e.g., see [7] and the references therein) one can argue (see [4]) that the optimal control for this problem is given by

$$\bar{u}(t) = \begin{cases} \delta\{-(t-2)^2/2 - 3/2\}, & 0 \leq t \leq 1, \\ \delta(t-3), & 1 \leq t \leq 2, \\ -\delta, & 2 \leq t \leq 3, \end{cases}$$

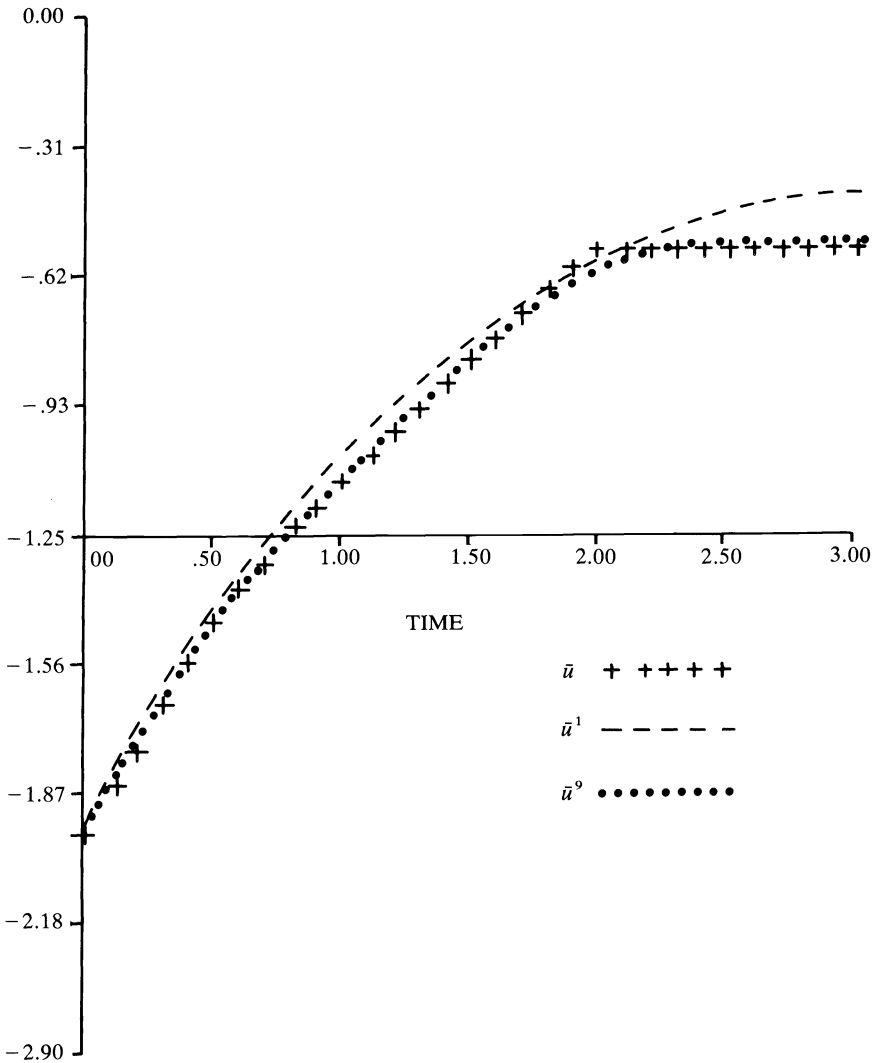


FIG. 4.1. Example 4.1.

TABLE 4.1
Example 4.1.

N	$\bar{\Phi}^N$
1	1.5885
5	1.7006
9	1.7149
13	1.7206
17	1.7237
20	1.7252

$$\bar{\Phi} \approx 1.7338$$

where δ is given (approximately) by

$$\delta \approx .56231.$$

The optimal value of the cost functional (4.9) is found to be

$$\Phi(\bar{u}) \approx 1.7338.$$

In Tables 4.1 and 4.2 we list selected numerical values found for $\bar{\Phi}^N \equiv \Phi^N(\bar{u}^N)$ and \bar{u}^N (along with values for $\bar{\Phi} \equiv \Phi(\bar{u})$ and \bar{u}) for several values of N when the method described above is applied to this example.

We note that the relative error between $\bar{\Phi}$ and $\bar{\Phi}^N$ at $N = 13$ (i.e. $\mathcal{E}^{13} \equiv |\bar{\Phi}^{13} - \bar{\Phi}|/|\bar{\Phi}|$) is less than one percent. This good approximation of $\bar{\Phi}^N$ to $\bar{\Phi}$ for N relatively small ($N < 50$ and in many cases $N < 20$) is typical of the rapid convergence attained using the above method on a variety of examples such as those detailed here and in [4].

Figure (4.1) contains graphs of \bar{u}^1 , \bar{u}^9 , and \bar{u} (although Table (4.2) shows a slight difference between \bar{u}^9 and \bar{u}^{20} , their graphs are essentially the same if one uses the scale employed in Fig. (4.1)).

TABLE 4.2
Example 4.1.

Time	\bar{u}^1	\bar{u}^9	\bar{u}^{20}	\bar{u}
0.0	-1.9650	-1.9667	-1.9663	-1.9681
0.30	-1.6283	-1.6517	-1.6518	-1.6656
0.61	-1.3501	-1.3882	-1.3898	-1.3867
0.91	-1.1204	-1.1670	-1.1731	-1.1775
1.20	-0.9446	-0.9914	-1.0016	-1.0122
1.50	-0.7876	-0.8240	-0.8309	-0.8435
1.80	-0.6608	-0.6823	-0.6813	-0.6748
2.10	-0.5611	-0.5850	-0.5761	-0.5623
2.41	-0.4873	-0.5464	-0.5534	-0.5623
2.72	-0.4404	-0.5418	-0.5529	-0.5623
3.00	-0.4249	-0.5417	-0.5528	-0.5623

TABLE 4.3
Example 4.2.

N	$\bar{\Phi}^N$
1	2.8025
5	3.0255
9	3.0596
13	3.0730
17	3.0799
20	3.0833

$$\bar{\Phi} \approx 3.1017$$

Example 4.2. Let $n = 1$, $t_1 = 2$, $G = 3$, $Q = 0$, and $R = 1$ while $A_0 = 1$, $A_1 = 1$, $r = 1$, $\varphi \equiv 1$, and $f(t) = u(t)$. Then (4.7) and (3.8) are respectively

$$\begin{aligned}\Phi(u) &= \frac{3}{2}[x(2)]^2 + \frac{1}{2} \int_0^2 [u(t)]^2 dt, \\ \dot{x}(t) &= x(t) + x(t-1) + u(t), \quad 0 \leq t \leq 2, \\ x_0 &= 1.\end{aligned}$$

The exact solution to (\mathcal{P}) with this choice of parameters can then be shown to be

$$\bar{u}(t) = \begin{cases} \delta \{e^{2-t} + (1-t)e^{1-t}\}, & 0 \leq t \leq 1, \\ \delta e^{2-t}, & 1 \leq t \leq 2, \end{cases}$$

where

$$\delta \approx -.3932.$$

One then finds

$$\Phi(\bar{u}) \approx 3.1017.$$

Numerical results for the associated problems $(\mathcal{AP})_N$ for this example are summarized in Tables 4.3 and 4.4.

TABLE 4.4
Example 4.2.

Time	\bar{u}^1	\bar{u}^9	\bar{u}^{20}	\bar{u}
0.0	-3.9685	-3.9712	-3.9734	-3.9737
0.19	-3.0455	-3.1118	-3.1201	-3.1182
0.39	-2.2869	-2.3875	-2.3973	-2.4083
0.59	-1.7183	-1.8306	-1.8376	-1.8533
0.80	-1.2925	-1.4054	-1.4082	-1.4014
1.00	-0.9741	-1.0858	-1.0884	-1.0687
1.20	-0.7366	-0.8510	-0.8604	-0.8750
1.41	-0.5602	-0.6803	-0.6961	-0.7093
1.61	-0.4305	-0.5523	-0.5678	-0.5807
1.81	-0.3365	-0.4505	-0.4635	-0.4755
2.00	-0.2747	-0.3735	-0.3842	-0.3932

The above two examples (along with several more presented in [4]) are quite simple. Indeed, they were chosen primarily because they are rather easily solved exactly using the maximum principle and thus can be used to investigate numerical convergence properties of the approximation methods that are the focus of this paper.

As has been known for some time [41], delay equations which are quite commonly encountered in physical problems often can be classified as belonging to one of two typical categories: equations with *retarded damping*

$$(4.10) \quad \ddot{y}(t) + K\dot{y}(t-r) + by(t) = g(t),$$

and those with *retarded restoring force*

$$(4.11) \quad \ddot{y}(t) + k\dot{y}(t) + \lambda y(t-r) = g(t),$$

where in each equation g represents some externally applied force. Equations (4.10) and (4.11) are special cases respectively of the more general equations

$$\ddot{y}(t) + k\dot{y}(t) + K\dot{y}(t-r) + by(t) = g(t),$$

(an artificially produced damping term $K\dot{y}(t-r)$ is added to help control or stabilize a system with insufficient natural damping $k\dot{y}(t)$; models such as these have been used in the study of antirolling stabilization systems in ships—see [39]–[41]), and

$$\ddot{y}(t) + k\dot{y}(t) + by(t) + \lambda y(t-r) = g(t)$$

(an artificially produced restoring force $\lambda y(t-r)$ is added to a dynamical system such as that associated with the automatic steering of high velocity aircraft—again see [41]).

The next two examples involve control problems with the typical systems (4.10) and (4.11). While the control problems themselves are not taken from any specific physical problems, the results reported below do demonstrate the efficacy of our method when it is applied to problems involving systems arising directly in certain areas of applications.

Example 4.3. Let $n = 2$, $t_1 = 2$, $G = \begin{pmatrix} 10 & 0 \\ 0 & 0 \end{pmatrix}$, $Q = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$, and $R = 1$ so that the payoff is given by

$$\Phi(u) = 5[x_1(2)]^2 + \frac{1}{2} \int_0^2 [u(t)]^2 dt.$$

In (3.8) we take $A_0 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$, $A_1 = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}$, $r = 1$, $\varphi = \begin{pmatrix} 10 \\ 0 \end{pmatrix}$ and $f(t) = \begin{pmatrix} 0 \\ u(t) \end{pmatrix}$.

The system is thus

$$\begin{aligned} \dot{x}_1(t) &= x_2(t), \\ \dot{x}_2(t) &= -x_1(t) - x_2(t-1) + u(t), \quad 0 \leq t \leq 2, \\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}_0 &= \begin{pmatrix} 10 \\ 0 \end{pmatrix}, \end{aligned}$$

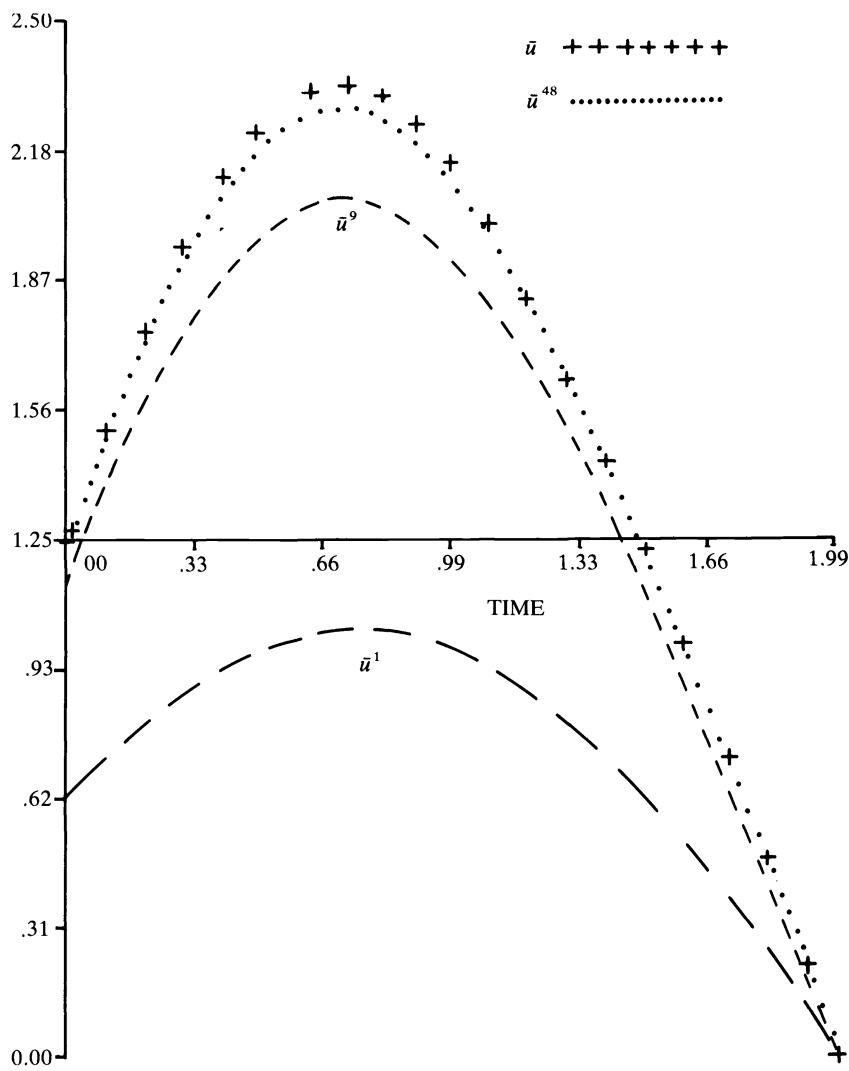


FIG. 4.2. Example 4.3.

which is the vector formulation of the scalar equation

$$\ddot{y}(t) + \dot{y}(t-1) + y(t) = u(t),$$

a controlled harmonic oscillator with retarded damping.

While the arguments for this example and Example 4.4 below are not as simple as those needed in Examples 4.1 and 4.2, one can still use the maximum principle for delay system control problems to obtain exact solutions. For this example the exact solution is given by

$$\bar{u}(t) = \begin{cases} \delta \sin(2-t) + (\delta/2)(1-t) \sin(t-1), & 0 \leq t \leq 1, \\ \delta \sin(2-t), & 1 \leq t \leq 2, \end{cases}$$

TABLE 4.5
Example 4.3.

N	$\bar{\Phi}^N$
1	.7379
5	2.3391
9	2.7390
13	2.9205
17	3.0238
20	3.0762
48	3.2587

$$\bar{\Phi} \approx 3.3991$$

where

$$\delta \approx 2.5599.$$

The corresponding optimal cost is

$$\Phi(\bar{u}) \approx 3.3991.$$

A representative sample of the numerical results for this example is presented in Tables 4.5 and 4.6. Graphs comparing \bar{u}^N and \bar{u} for several values of N are given in Fig. 4.2.

We observe that for this example the convergence of $\bar{\Phi}^N$ to $\bar{\Phi}$ (relative error 9% at $N = 20$) is not as fast as it is for the previously detailed examples. However we are still able to obtain a good numerical approximation by using a larger approximating ODE system (at $N = 48$, relative error in the payoff is only 4% and as Table 4.6 and Fig. 4.2 reveal, \bar{u}^{48} is a reasonably good approximation for \bar{u}).

TABLE 4.6
Example 4.3.

Time	\bar{u}^1	\bar{u}^9	\bar{u}^{20}	\bar{u}^{48}	\bar{u}
0.0	.6488	1.1501	1.2628	1.2295	1.2506
0.20	.8087	1.5745	1.6757	1.7261	1.7584
0.41	.9328	1.8830	2.0154	2.0828	2.1393
0.61	1.0080	2.0474	2.1931	2.2694	2.3284
0.81	1.0253	2.0562	2.1924	2.2684	2.3306
1.02	.9842	1.9207	2.0312	2.0889	2.1260
1.22	.8758	1.6537	1.7355	1.7746	1.8003
1.42	.7076	1.2919	1.3510	1.3785	1.4029
1.62	.4978	.8846	.9242	.9443	.9495
1.82	.2442	.4251	.4440	.4536	.4583
2.00	0.0	0.0	0.0	0.0	0.0

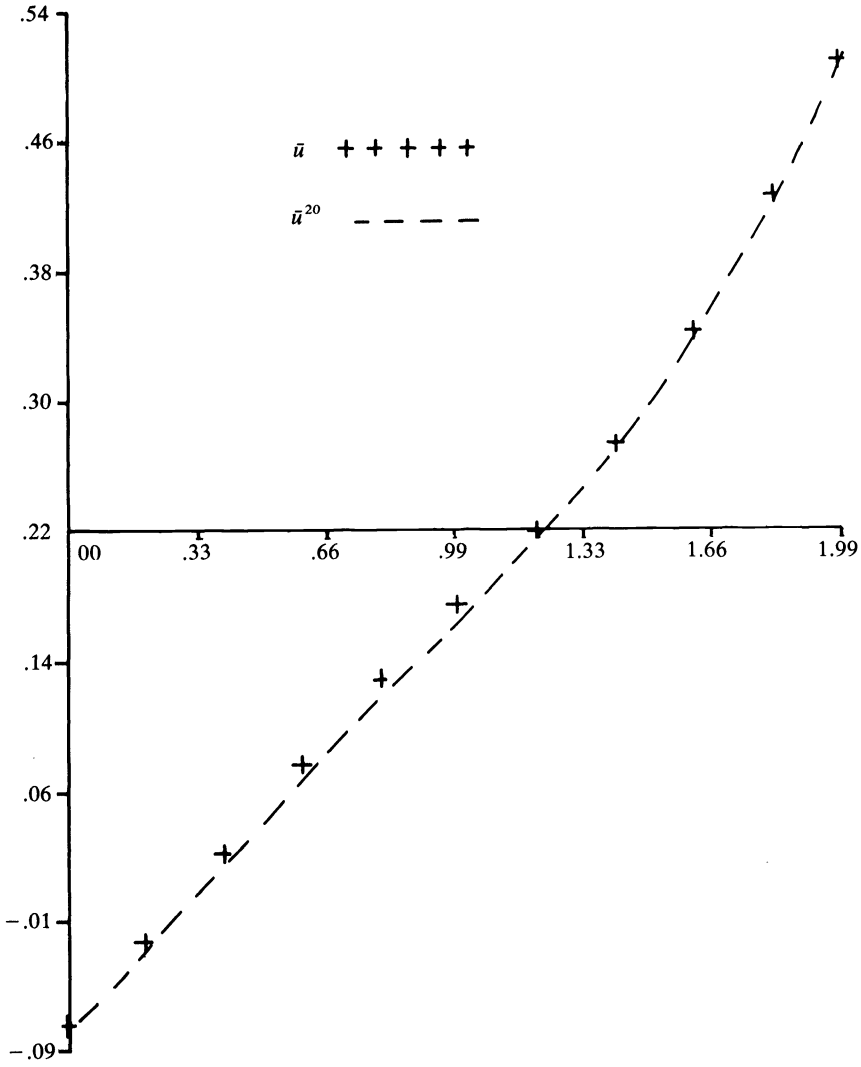


FIG. 4.3. Example 4.4.

Example 4.4. Let $n=2$, $t_1=2$, $G=\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $Q=\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$, $R=1$, $A_0=\begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix}$, $A_1=\begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix}$, $r=1$, $\varphi=\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $f(t)=\begin{pmatrix} 0 \\ u(t) \end{pmatrix}$. The problem (\mathcal{P}) thus becomes one of minimizing

$$\Phi(u) = \frac{1}{2} \{ [x_1(2)]^2 + [x_2(2)]^2 \} + \frac{1}{2} \int_0^2 [u(t)]^2 dt$$

TABLE 4.7
Example 4.4.

N	$\bar{\Phi}^N$
1	.1509
5	.1827
9	.1887
13	.1913
17	.1927
20	.1934

$$\bar{\Phi} \approx .1975$$

subject to

$$\dot{x}_1(t) = x_2(t),$$

$$\dot{x}_2(t) = -x_2(t) - x_1(t-1) + u(t), \quad 0 \leq t \leq 2,$$

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

The system here is the vector equivalent for the scalar equation

$$\ddot{y}(t) + \dot{y}(t) + y(t-1) = u(t),$$

which is a special case of (4.11).

The exact solution can be shown to be

$$\bar{u}(t) = \begin{cases} (\mu - \delta) e^{t-2} + [2\mu - 3\delta - (\mu - \delta)t] e^{t-1} + \delta(t+2) - \mu, & 0 \leq t \leq 1, \\ (\mu - \delta) e^{t-2} + \delta, & 1 \leq t \leq 2, \end{cases}$$

where

$$\mu \approx .5226194, \quad \delta \approx -.0259256.$$

The optimal payoff is

$$\Phi(\bar{u}) \approx .197478$$

and numerical results for the approximating solutions are summarized in Tables 4.7 and 4.8. The relative error when comparing $\bar{\Phi}$ and $\bar{\Phi}^{20}$ is found to be 2% and as Fig. 4.3 shows, \bar{u}^{20} is a very good approximation for \bar{u} .

TABLE 4.8
Example 4.4.

Time	\bar{u}^1	\bar{u}^9	\bar{u}^{20}	\bar{u}
0.0	-.1041	-.0877	-.0872	-.0870
0.20	-.0697	-.0379	-.0348	-.0336
0.41	-.0307	.0136	.0192	.0247
0.61	.0127	.0652	.0728	.0802
0.81	.0638	.1159	.1240	.1327
1.00	.1083	.1619	.1689	.1759
1.20	.1648	.2133	.2181	.2206
1.41	.2270	.2701	.2740	.2782
1.61	.2967	.3377	.3418	.3455
1.81	.3766	.4200	.4248	.4278
2.00	.4633	.5124	.5179	.5227

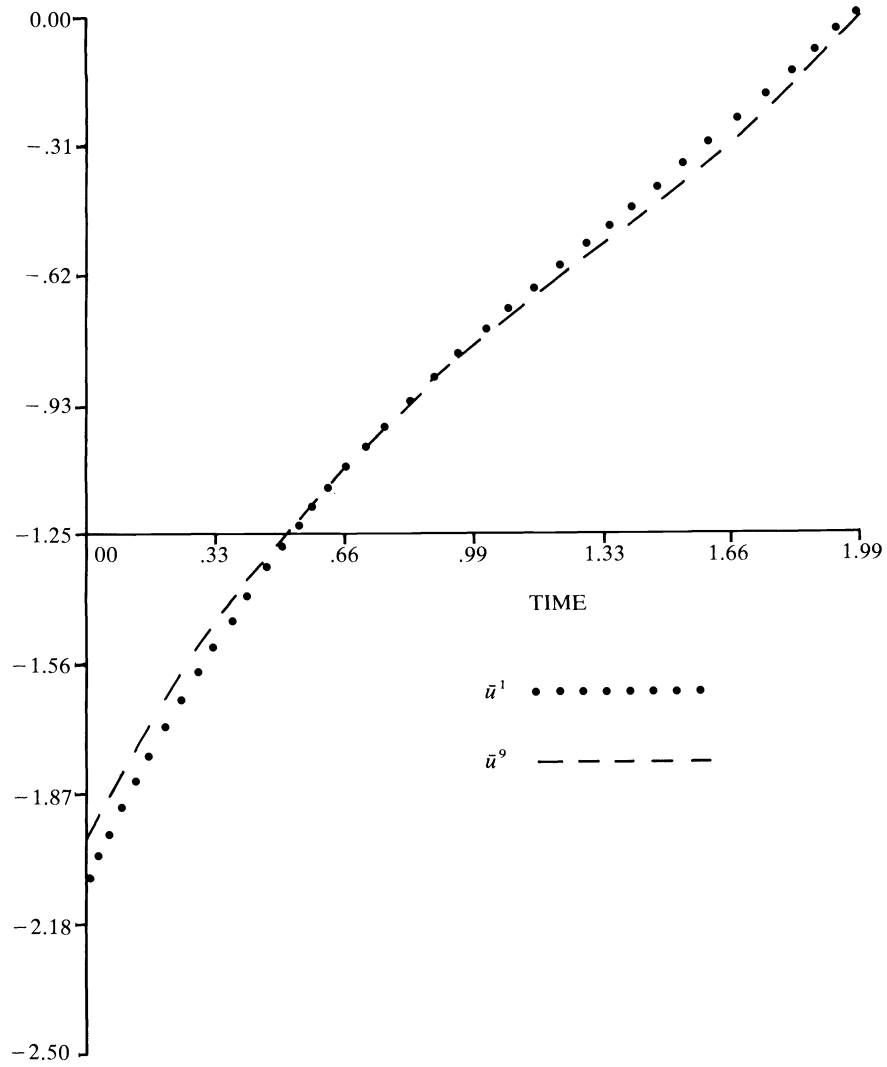


FIG. 4.4. Example 4.5.

Finally, we present numerical results only for a simple example for which $Q \neq 0$. Although we have not solved for the analytic solution to this problem, we present the numerical approximation results since they illustrate use of our method on an example for which $G = 0$ and $Q \neq 0$. Moreover, this example was also considered by Delfour [18] and hence the interested reader may compare our numerical results with those presented in [18]. If such a comparison is made, then it is found that the numerical solutions are in good agreement.

Example 4.5. Let $n = 1$, $t_1 = 2$, $G = 0$, $Q = 1$ and $R = 1$ so that the cost functional is

$$\Phi(u) = \frac{1}{2} \int_0^2 \{[x(t)]^2 + [u(t)]^2\} dt.$$

TABLE 4.9
Example 4.5.

N	$\bar{\Phi}^N$
1	1.5542
5	1.6243
9	1.6346
13	1.6387
17	1.6409
20	1.6419

TABLE 4.10
Example 4.5.

Time	\bar{u}^1	\bar{u}^9	\bar{u}^{20}
0.0	-2.0877	-1.9966	-1.9908
0.20	-1.7177	-1.6610	-1.6559
0.41	-1.4156	-1.3853	-1.3798
0.61	-1.1649	-1.1564	-1.1511
0.81	-0.9528	-0.9633	-0.9604
1.02	-0.7692	-0.7959	-0.7974
1.22	-0.6051	-0.6431	-0.6481
1.42	-0.4523	-0.4931	-0.4981
1.63	-0.3025	-0.3357	-0.3386
1.83	-0.1460	-0.1630	-0.1641
2.00	0.0	0.0	0.0

TABLE 4.11
Example 4.5.

Time	\bar{x}^1	\bar{x}^9	\bar{x}^{20}
0.0	1.0000	1.0000	1.0000
0.20	.8166	.8328	.8339
0.41	.6947	.7274	.7296
0.61	.6187	.6723	.6763
0.81	.5787	.6561	.6641
1.02	.5683	.6670	.6804
1.22	.5843	.6965	.7124
1.42	.6256	.7419	.7562
1.63	.6937	.8067	.8174
1.83	.7922	.8983	.9057
2.00	.9038	1.0033	1.0091

In (3.8) let $A_0 = 0$, $A_1 = 1$, $r = 1$, $\varphi \equiv 1$ and $f(t) = u(t)$ and the governing system is thus

$$\dot{x}(t) = x(t-1) + u(t), \quad 0 \leq t \leq 2,$$

$$x_0 = 1.$$

Selected numerical values for $\bar{\Phi}^N$, \bar{u}^N , and \bar{x}^N are given respectively in Tables 4.9, 4.10, and 4.11 while Fig. 4.4 compares the graphs of \bar{u}^1 and \bar{u}^9 (the graphs of \bar{u}^9 and \bar{u}^{20} are indistinguishable using the scale chosen in Fig. 4.4).

We conclude our presentation of numerical results by noting that while all the examples presented above involve only scalar controls, examples with vector controls are also discussed in detail in [4] and there it is demonstrated that the numerical approximations also agree well with exact solutions for these examples.

In closing this section on numerical results, we shall also make a few brief comments on CPU time and storage requirements for the approximation scheme discussed here. The numerical results presented in [4] and summarized above were generated by software packages developed on the IBM 360/67 (at Brown) and the IBM 30/158 system (at V.P.I.). In many cases, the calculations were checked by running the problems on both machines (on which different software packages were employed). With regard to storage, very important factors appear to be the approximation index N and the number of points at which one requires values for the solution of the Riccati equation for the approximating (of dimension $n(N+1)$) control problem. In our programs, we used, at various times, one of several integration schemes (modifications of standard predictor-corrector, Runge-Kutta, etc., schemes) and our comparisons indicate that choice of this scheme can also be extremely important with regard to storage and CPU time needed (in some cases, savings of up to 40%–50% in time can be effected by judicious choice of an integration technique). However, our goal in carrying out the numerical work was not so much to develop the most efficient software packages but rather to demonstrate the feasibility of the method under discussion. The reader should keep this in mind as we discuss storage and time requirements in the following paragraph.

For $n = 1$ there were, as one might expect, absolutely no problems with storage (the method can easily be implemented on a 256K machine with little care towards efficiency in use of storage). For systems with $n = 3$ and 2 or 3 dimensional control variables, one can carry out the computations necessary for approximation indices N up to a value of 10 with 256K storage. As N gets very large, one of course must expect to either work at developing efficiency of the software packages (through use of interpolation for values between stored values, etc.—we did only a little of this) or increase storage capabilities. For example, with $N = 48$ in Example 4.3 one has more than 4850 variables in the resulting Riccati equation but increased storage to 512K is sufficient to handle runs of this nature. This example (with $N = 48$) was by far the most costly (in CPU time) example of all those reported on in [4], taking approximately 3700 seconds ($\approx \$40$) of CPU time. The total amount of CPU time required for *all* the runs reported in Table 4.7 (Example 4.4 with $N = 1, 5, 9, 13, 17, 20$) was about 900 seconds ($\approx \$10$). For the scalar Example 4.1, the total CPU time needed for the runs reported in Table 4.1 ($N = 1, 5, 9, 13, 17, 20$) was approximately 260 seconds.

We have been involved in other computational efforts (using the scheme described in this manuscript as well as other approximation schemes which fit into the framework developed in § 3) since the results reported in [4] were obtained. While we do not claim that storage would be of no concern when implementing the method for large vector systems ($n \geq 5$), our experience has indicated that considerable savings can be attained (in both CPU time and storage required) by minor alterations in the software packages employed to generate the data reported above. Our efforts to date indicate that, with careful development of

software packages, implementation of the approximation scheme discussed in this paper is feasible (given the modern computational facilities available) for vector systems with regard to both storage and CPU time requirements.

5. Comments on previous literature and concluding remarks. The so-called Trotter–Kato type theorems (e.g., Theorem 3.1) were first developed [35], [59], [60] in connection with an operator-theoretic presentation of finite differencing techniques for partial differential equation initial value problems. DeJulio [17] several years ago suggested the use of Trotter type approximation results in the development of computational techniques for optimal control problems. Motivated by distributed parameter optimal control problems, DeJulio describes an approach to approximation of solutions for fairly general abstract minimization problems. While he doesn't discuss any specific problems, he does indicate how one might combine the Trotter approximation results (using "averaging" approximations such as those given in (3.9), (3.11) and (3.13)) with the well-known ε -technique advocated by Balakrishnan. Related, but somewhat different ideas for approximation of partial differential equation systems are presented, along with an extensive bibliography, by Aubin in [1].

In a subsequent paper, Sasai and Shimemura [54] employ the Trotter theorem to discuss theoretical aspects of approximations for solutions of optimal final value and time-optimal control problems governed by abstract systems. Slightly more general cost functions are treated in a similar fashion in [52] where it is also shown how one might formulate a finite difference approximation scheme for a controlled diffusion equation in the context of the Trotter type framework given in [54]. Approximate controllability of differential-difference control systems is discussed in [53] via use of the Trotter type approximation results and to our knowledge, these authors were the first to suggest use of the Trotter ideas in connection with differential-difference equation approximations. As we have already noted certain parts of their treatment (of a different problem—controllability—from the one treated here) are formal in nature and other aspects are incorrect. Specifically in [53] the relationship between solutions of the abstract equation and the differential-difference equation studied there is vague. (A concise argument to establish the equivalence of solutions of the AEE and FDE were first given, to our knowledge, in [2].) Furthermore, there are what appear to be nontrivial errors in [53] in the arguments that the Trotter approximations do converge. Two errors, in particular, are perhaps worthy of specific mention. First, the authors there incorrectly reduce the stability arguments to the question of obtaining bounds for matrix operators. The reduction from operators on Z^N (in our notation above) to operators on $R^{n(N+1)}$ carried out in [53] does *not* involve an *isometric* identification and the equivalence constants between operator norms in Z^N and $R^{n(N+1)}$ that one obtains using that identification are themselves dependent upon N . Thus, the arguments given in [53], even were they correct, are not sufficient to yield stability of the approximating scheme studied there. Even if the authors of [53] had used an isometric identification there is a second and more fundamental, we believe, error in their argument. Sasai and Fukuda argue that

$$|e^{\mathcal{A}Nt}| \leq e^{-(N/r)t} e^{(|\theta_2|+|\theta_3|)t}$$

where θ_2, θ_3 are specified matrices. A careful analysis of the matrix θ_2 given in [53] reveals that $|\theta_2| \cong \sqrt{2}(N/2)$ and thus the right side of the above inequality becomes unbounded as $N \rightarrow \infty$. While one can make an isometric identification between Z^N and $R^{n(N+1)}$ to correct the first abovementioned error, it is not clear to the authors of this paper that arguments can be made to extend the ideas of Sasai and Fukuda to give bounds for the resulting matrices that will ensure stability of the “averaging” approximations of § 3 of this paper.

There has been in the literature widespread use of the idea of approximating differential-difference equations (DDE) by higher-order ordinary differential equation (ODE) systems which are special cases of (3.22), (3.23). Much of the justification for these approximations has been discussed, if at all, in at best heuristic terms although Krasovskii [30] and Repin [46] did present convergence arguments in a rigorous if somewhat inelegant fashion. One popular approach (e.g., see [22], [45], [64]) to the heuristic arguments uses transfer function techniques. Briefly, consider the simple system

$$(5.1) \quad \dot{x}_0(t) = A_0 x_0(t) + A_1 x_0(t-r) + f(t),$$

which, upon use of Laplace transforms, can be represented by

$$(5.2) \quad \hat{x}_0(s)F(s) = \hat{f}(s),$$

where $F(s) = sI - A_0 - A_1 e^{-rs}$. An associated input-output diagram (see Fig. 5.1) involves the transcendental term e^{-rs} in one link. If this transcendental link is replaced by N successive links (see Fig. 5.2) with rational fraction transfer functions $(1 + (r/N)s)^{-1}$ the corresponding system becomes

$$(5.3) \quad \begin{aligned} \dot{x}_0^N(t) &= A_0 x_0^N(t) + A_1 x_N^N(t) + f(t), \\ \dot{x}_1^N(t) &= \frac{N}{r} \{x_0^N(t) - x_1^N(t)\}, \\ &\vdots \\ \dot{x}_N^N(t) &= \frac{N}{r} \{x_{N-1}^N(t) - x_N^N(t)\}. \end{aligned}$$

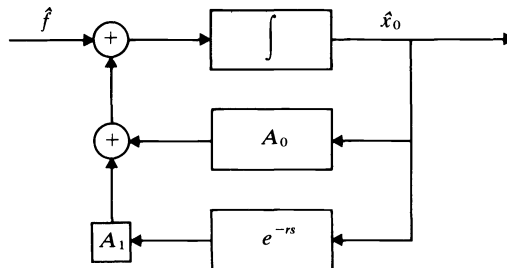


FIG. 5.1

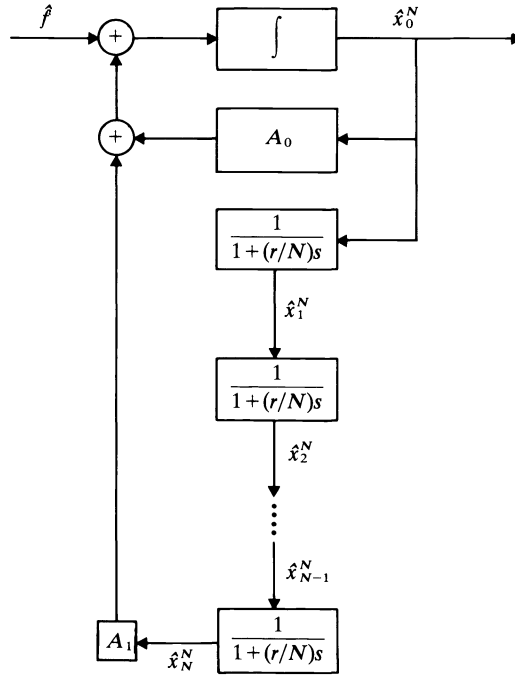


FIG. 5.2

Simple computations involving Laplace transforms yield

$$(5.4) \quad \hat{x}_0^N(s)F_N(s) = \hat{f}(s),$$

where $F_N(s) = sI - A_0 - A_1(1 + (r/N)s)^{-N}$. Thus the system (5.3) is an approximation to (5.1) since $F_N(s) \rightarrow F(s)$ as $N \rightarrow \infty$.

As Krasovskii [30] points out, these higher order ODE approximations were often (at least in their early periods of development and advocacy) based on truncated Taylor series expansion arguments (e.g., see [26], [50]). For example, in (5.1), identifying $x_1(t)$ and $x_0(t-r)$ and then using equality in the approximation

$$x_0(t) - x_0(t-r) \approx \dot{x}_0(t-r)r = \dot{x}_1(t)r,$$

one may replace (5.1) by

$$(5.5) \quad \begin{aligned} \dot{x}_0(t) &= A_0 x_0(t) + A_1 x_1(t) + f(t), \\ \dot{x}_1(t) &= \frac{1}{r} \{x_0(t) - x_1(t)\}, \end{aligned}$$

which is just (5.3) for $N = 1$.

While our results (and those of others discussed below) substantiate the validity of these approximations in certain situations, it has been known for some time (e.g., see [39], [40], [41]) that indiscriminate and imprecise use of such approximation ideas (as can be found frequently in the literature, especially in the *formulation* of mathematical models for certain physical and biological processes) can lead to substantial difficulties and, in some cases, to erroneous conclusions.

Turning next to a brief summary of previous uses of approximations such as (5.3), we remark that Jên-Wei [26] and Salukvadze [50] were among the first to propose such approximations. They formally replace the true state x_t in DDE control problems by states $(x_0(t), x_1(t)) = (x(t), x(t-r))$ and use ODE control methods (dynamic programming, linear regulator theory) for the replacement system (5.5). Neither of these investigators addressed the question of convergence but Krasovskii [30] later did. Considering closed loop optimal control problems and employing arguments that relied heavily on special features of the quadratic cost, closed loop (feedback) nature of his problems, he argued convergence of the approximate (high order ODE—as in (5.3)) system states, optimal controls and costs to those for the original DDE problem. At about the same time and independently, Repin [46] gave convergence arguments for a similar approximating system in the context of an investigation of the preservation of uniform asymptotic stability under the approximation of the DDE by the ODE's. Repin also gave error estimates that revealed that the convergence essentially was $O(1/\sqrt{N})$. In a later paper [34], Kurzhanskii showed that the convergence arguments for the higher order approximations of Repin–Krasovskii could be extended to treat nonautonomous variable-delay homogeneous differential-difference equations.

In [31] Krasovskii announced a generalization of his approximation ideas for an abstract class of approximations of which the results in [30] are a special case. While the results of that later note are in the same spirit as those in [2], it is apparent (even though no proofs were given in [31]) that Krasovskii's approach was quite different from that taken in [2]. We also note that Krasovskii in [30], [31] was among the first, to our knowledge, to treat optimal control of linear functional differential equation systems with quadratic cost in the context of the $R^n \times L_2$ norm, a norm which was subsequently employed by, among others, Coleman and Mizel [13], [14], Borisovič and Turbabin [12], Delfour and Mitter [19], (see also [18] for other references to their extensive efforts on the regulator problem), Webb [62], [63], and the present authors [2], [3] in their investigations of problems involving functional differential equations.

In [22] Ichikawa argued heuristically that one could derive necessary conditions (i.e., a maximum principle) for control of problems with DDE systems by first approximating the DDE by a higher order ODE (as in (5.3)), writing down the usual necessary conditions for the approximating ODE problems and then formally passing to the limit as $N \rightarrow \infty$. Similar heuristic ideas were employed in [23] to develop necessary conditions for problems governed by partial differential equation systems. A heuristic treatment (leading to formal and what appears to be incorrect results) of the minimum time settling problem (i.e., problems with terminal constraints $x_{t_1} = 0$; see [5], [6], [10], [25], [29], [33], and the summary in [7]) along the lines of the Ichikawa arguments was given by Westdal and Lehn in [64].

The Ichikawa approach was also employed by Soliman and Ray [55], [56] to formally derive optimal feedback control laws for problems with systems containing delays in both the state and the control variables. They observe that, while the rigor of their approach is certainly subject to challenge, most of their results agree in special cases with other rigorously obtained results and thus

conjecture that their results probably could be established with rigor. (Subsequent results of Delfour [18], and the present authors [2], in addition to the earlier work of Krasovskii [30], do indeed settle certain convergence questions raised by consideration of the Soliman-Ray discussions.)

Ross and Flügge-Lotz [48], [49], in their study of feedback control laws for DDE constrained problems, made use of differencing techniques on associated Riccati type equations. Ross and later Hess [21], who were apparently unaware of the earlier Krasovskii papers [30], [31], observed that these techniques were formally related to a high order ODE approximation for the original DDE.

Sannuti [51] and later Inoue, et al. [24] combined high order ODE approximation ideas with singular perturbation techniques to give a formal development of design of approximate optimal controls for DDE governed problems.

Delfour [18] has recently completed a rather thorough investigation of certain numerical schemes for an operator Riccati differential equation satisfied by the feedback gains in the solution of the linear regulator problem with functional differential equation systems. While his approach is somewhat different from that in [2], [4] and the present paper, his results are closely related to both our results and many of those cited above—in particular, those in [21], [30], [31], [46], [48], [49], [55], [56]. Using “averaging” approximations (i.e., high order ODE as in (5.3)) along with a simultaneous discretization in time (that is, a simultaneous time and state discretization as contrasted to a discretization only in the state employed in many of the above references), Delfour gives rigorous convergence arguments for the state, costate, and Riccati operator variables as well as for the optimal controls for the approximating discrete system control problems. Numerical results for a number of examples (via solution of the approximating Riccati equations) along with a comparison to the analytic solution of the original problem in one instance (a simple example very similar to Example 4.1 above) are also presented. Delfour remarks [18, p. 79] that a study of his numerical results seems to indicate that his overall approximation scheme is $O(1/N)$. An inspection of the numerical results of our approach (see [4], in which a number of examples are given in which the numerical solutions are compared to the analytic solutions) reveals similar convergence properties and it is not clear which, if either, of these related schemes (simultaneous discretization of time and state vs. initial approximation in the state only) is computationally superior when one restricts one’s attention to quadratic cost, linear system problems. From a theoretical point of view, Delfour’s treatment includes nonautonomous functional differential equation systems and yields somewhat more detailed information on the Riccati type variables in the problem while the approach in [2] and the present paper can be extended to more general cost functions and certain nonlinear systems and does not require a detailed analysis of an infinite dimensional Riccati equation. In addition, other approximation ideas [8], [36] can be formulated (see [3]) in the general framework first detailed in [2] and modified slightly in § 3 of the present paper.

The relationship between our approach and truncated Fourier type expansions (in terms of well-known special functions of mathematical physics) leading to differential equations for the coefficients should now be clear to

readers. Indeed, our approach in § 3 can be considered in some sense a special case of the general and well-known techniques discussed by Reeve in [45] if one roughly identifies our χ_j^N and w_j^N of § 3 with his p_n and a_n respectively. Furthermore it is easily recognized that these are really special instances of the fundamental idea which is the basis of classical Ritz–Galerkin techniques [11], [16], [27], [37], [38], [57]. Indeed, the results of § 3 when combined with § 4 of this paper clearly constitute a special case of general Ritz type methods.

Finally, we point out that the “averaging” approximations of § 3 (see (3.9), (3.11), (3.13)) are well-known to investigators in approximation theory and enjoy the classical least squares property. That is, given χ_j^N , the coefficients φ_j^N of (3.11) are exactly the solutions of problem of choosing a_j^N so as to minimize $|\varphi - \Sigma a_j^N \chi_j^N|_{L_2}$ (e.g., see [15, p. 92], [47, p. 31]).

Acknowledgment. The authors are grateful to referees whose questions and comments were most helpful.

REFERENCES

- [1] J-P. AUBIN, *Approximation of Elliptic Boundary-Value Problems*, Wiley-Interscience, New York, 1972.
- [2] H. T. BANKS AND J. A. BURNS, *An abstract framework for approximate solutions to optimal control problems governed by hereditary systems*, Proceedings, International Conference on Differential Equations (Univ. So. Calif., Sept., 1974), H. A. Antosiewicz, ed., Academic Press, New York, 1975, pp. 10–25.
- [3] ———, *Eigenmanifold decomposition for retarded functional differential equations in Hilbert space*, Math. Dept. Tech. Rep. TR-1, Virginia Polytech. Inst. and State Univ., Blacksburg, VA, 1974.
- [4] H. T. BANKS, J. A. BURNS, E. M. CLIFF AND P. R. THRIFT, *Numerical solutions of hereditary control problems via an approximation technique*, Brown Univ. LCDS Tech. Rep. 75-6, Providence, RI, 1975.
- [5] H. T. BANKS AND G. A. KENT, *Control of functional differential equations of retarded and neutral type to target sets in function space*, this Journal, 10 (1972), pp. 567–593.
- [6] 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.
- [7] H. T. BANKS AND A. MANITIUS, *Application of abstract variational theory to hereditary systems – A survey*, IEEE Trans. Automatic Control, AC-19 (1974), pp. 524–533.
- [8] ———, *Projection series for retarded functional differential equations with applications to optimal control problems*, J. Differential Equations, 18 (1975), pp. 296–332.
- [9] R. BELLMAN AND K. L. COOKE, *Differential Difference Equations*, Academic Press, New York, 1963.
- [10] Z. BIEN, *Optimal control of delay systems*, Ph.D. thesis, Dept. of Electrical Engineering, Univ. of Iowa, Ames, 1975.
- [11] G. BIRKHOFF, *The Numerical Solution of Elliptic Equations*, CBMS monograph, vol. 1, Society for Industrial and Applied Mathematics, Philadelphia, 1972.
- [12] J. G. BORISOVIČ AND A. S. TURBABIN, *On the Cauchy problem for linear nonhomogeneous differential equations with retarded argument*, Soviet Math. Dokl., 10 (1969), pp. 401–405.
- [13] B. D. COLEMAN AND V. J. MIZEL, *Norms and semigroups in the theory of fading memory*, Arch. Rational Mech. Anal., 23 (1966), pp. 87–123.
- [14] ———, *On the stability of solutions of functional differential equations*, Ibid., 30 (1968), pp. 173–196.
- [15] G. DAHLQUIST AND Å. BJÖRCK, *Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1974.

- [16] J. W. DANIEL, *The Approximate Minimization of Functionals*, Prentice-Hall, Englewood Cliffs, NJ, 1971.
- [17] S. DEJULIO, *Numerical solution of dynamical optimization problems*, this Journal, 8 (1970), pp. 135–147.
- [18] M. C. DELFOUR, *Numerical solution of the optimal control problem for linear hereditary differential systems with a linear-quadratic cost function and approximation of the Riccati differential equation*, Centre de Recherches Mathématiques, Université de Montréal, Tech. Rep. CRM-408, 1974.
- [19] M. C. DELFOUR AND S. K. MITTER, *Controllability, observability, and optimal feedback control of hereditary differential systems*, this Journal, 10 (1972), pp. 298–328.
- [20] J. K. HALE, *Functional Differential Equations*, Springer-Verlag, New York, 1971.
- [21] R. A. HESS, *Optimal control approximations for time delay systems*, AIAA J., 10 (1972), pp. 1536–1538.
- [22] K. ICHIKAWA, *Pontryagin's maximum principle in optimizing time-delay systems*, Elect. Engineering in Japan, 87 (1967), no. 12, pp. 75–83.
- [23] K. ICHIKAWA AND K. KANAI, *Application of Pontryagin's maximum principle to a distributed parameter system*, Ibid., 89 (1969), no. 10, pp. 19–29.
- [24] K. INOUE, H. AKASHI, K. OGINO AND Y. SAWARAGI, *Sensitivity approaches to optimization of linear systems with time delay*, Automatica, 7 (1971), pp. 671–679.
- [25] M. Q. JACOBS AND T. J. KAO, *An optimum settling problem for time-lag systems*, J. Math. Anal. Appl., 40 (1972), pp. 1–21.
- [26] C. JÊN-WEI, *The problem of synthesizing an optimal controller in systems with time delay*, Automat. Remote Control, 23 (1962), pp. 121–125.
- [27] L. V. KANTOROVICH AND V. I. KRYLOV, *Approximate Methods of Higher Analysis*, Noordhoff, Groningen, the Netherlands, 1964.
- [28] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
- [29] G. A. KENT, *A maximum principle for optimal control problems with neutral functional differential systems*, Bull. Amer. Math. Soc., 77 (1971), pp. 565–570.
- [30] N. N. KRASOVSKII, *The approximation of a problem of analytic design of controls in a system with time-lag*, J. Appl. Math. Mech., 28 (1964), pp. 876–885.
- [31] ———, *Approximation of an optimal control problem for a system with delay*, Soviet Phys. Dokl., 11 (1966), p. 219–221.
- [32] S. G. KREIN, *Linear Differential Equations in Banach Space*, Translations Math. Mono., vol. 29, American Mathematical Society, Providence, RI, 1971.
- [33] S. KURCYUSZ, *A local maximum principle for operator constraints and its application to systems with time lags*, Arch. Automat. i Telemach., 19 (1974).
- [34] A. B. KURZHANSKII, *The approximation of linear differential equations with retardation*, Differential Equations, 3 (1967), pp. 1088–1094.
- [35] P. D. LAX AND R. D. RICHTMYER, *Survey of the stability of linear finite difference equations*, Comm. Pure Appl. Math., 9 (1956), pp. 267–293.
- [36] E. M. MARKUSHIN AND S. N. SHIMANOV, *Approximate analytic design of controls for systems with time lag*, Automat. Remote Control, 29 (1968), pp. 367–374.
- [37] S. G. MIKHLIN, *The Numerical Performance of Variational Methods*, Wolters-Noordhoff, Groningen, the Netherlands, 1971.
- [38] S. G. MIKHLIN AND K. L. SMOLITSKIY, *Approximate Methods for Solution of Differential and Integral Equations*, American Elsevier, New York, 1967.
- [39] N. MINORSKY, *Experiments with activated tanks*, Trans. ASME, 69 (1941), pp. 735–747.
- [40] ———, *Self-excited oscillations in dynamical systems possessing retarded actions*, J. Appl. Mech., 9 (1942), A65–A71.
- [41] ———, *Nonlinear Oscillations*, Van Nostrand, New York, 1962.
- [42] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Math. Dept., Lecture Notes, vol. 10, Univ. of Maryland, College Park, 1974.
- [43] S. PERLIS, *Introduction to Algebra*, Blaisdell, Waltham, MA, 1966.
- [44] B. T. POLJAK, *Existence theorems and convergence of minimizing sequences in extremum problems with restrictions*, Dokl. Akad. Nauk SSSR, 166 (1966), pp. 287–290.
- [45] P. J. REEVE, *A method of approximating to pure time delay*, Int. J. Control, 8 (1968), pp. 53–63.

- [46] I. M. REPIN, *On the approximate replacement of systems with lag by ordinary differential equations*, J. Appl. Math. Mech., 29 (1965), pp. 254–264.
- [47] J. R. RICE, *The Approximation of Functions*, vol. I, Addison-Wesley, Reading, MA, 1964.
- [48] D. W. ROSS AND I. FLÜGGE-LOTZ, *An optimal control problem for systems with differential-difference equation dynamics*, this Journal, 7 (1969), pp. 609–623.
- [49] D. W. ROSS, *Controller design for time lag systems via a quadratic criterion*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 664–672.
- [50] M. E. SALUKVADZE, *Concerning the synthesis of an optimal controller in linear delay systems subjected to constantly acting perturbations*, Automat. Remote Control, 23 (1962), pp. 1495–1501.
- [51] P. SANNUTI, *Near optimum design of time-lag systems by singular perturbation method*, Proc. JACC, (1970), pp. 489–496.
- [52] H. SASAI, *On the convergence of difference approximations in distributed parameter optimal control problems*, Mem. Fac. Engrg. Nagoya Univ., 23 (1971), pp. 316–325.
- [53] H. SASAI AND T. FUKUDA, *Consideration of linear delay-differential systems by approximation systems*, Trans. Soc. Instrument Control Engr., 10 (1974), pp. 298–303.
- [54] H. SASAI AND E. SHIMEMURA, *On the convergence of approximating solutions for linear distributed parameter optimal control problems*, this Journal, 9 (1971), pp. 263–273.
- [55] 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.
- [56] ———, *Optimal control of multivariable systems with pure time delays*, Automatica, 7 (1971), pp. 681–689.
- [57] G. STRANG AND G. FIX, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
- [58] A. E. TAYLOR, *Introduction to Functional Analysis*, John Wiley, New York, 1958.
- [59] H. F. TROTTER, *Approximations of semigroups of operators*, Pacific J. Math., 8 (1958), pp. 887–919.
- [60] ———, *Approximation and perturbation of semigroups*, Linear Operators and Approximation II, P. L. Butzer and B. Sz. Nagy, eds., Birkhäuser Verlag, Basel, Switzerland, 1974, pp. 3–21.
- [61] R. S. VARGA, *Functional Analysis and Approximation Theory in Numerical Analysis*, Society for Industrial and Applied Mathematics, Philadelphia, 1971.
- [62] G. F. WEBB, *Linear functional differential equations with L_2 initial functions*, to appear.
- [63] ———, *Functional differential equations and nonlinear semigroups in L_p spaces*, J. Differential Equations, 20 (1976), pp. 71–89.
- [64] A. S. WESTDAL, AND W. H. LEHN, *Time optimal control of linear systems with delay*, Internat. J. Control, 11 (1970), pp. 599–610.

UN ALGORITHME DE MINIMISATION EN CHAÎNE EN OPTIMISATION CONVEXE*

C. CARASSO† ET P. J. LAURENT‡

Abstract. We consider the problem of minimization of a convex, nondifferentiable function on a linear manifold in \mathbb{R}^n . We give a general algorithm of "exchange" type which does not need any Haar assumption and can be continued indefinitely.

Introduction. En 1934, Rémès [27], [28] a donné un algorithme itératif pour le calcul de la meilleure approximation d'une fonction continue par un polynôme, au sens de la norme de Tchebycheff. Différentes généralisations de cet algorithme ont été successivement proposées. En 1959, Stiefel [30], [31] a énoncé un théorème d'échange et l'algorithme associé qui permet de résoudre le problème de Tchebycheff discret, c'est-à-dire, dans \mathbb{R}^N ; l'échange n'est plus lié au théorème d'alternance de Tchebycheff. Dans la même année, Cheney et Goldstein [13] décrivent un algorithme d'échange de principe identique pour la minimisation d'une fonctionnelle convexe sur un convexe de \mathbb{R}^N . En 1967, l'algorithme de Rémès a été décrit dans un espace vectoriel normé arbitraire [19], [20]. Cela permet notamment de traiter des problèmes d'approximation dans différents espaces courants (espaces de fonctions sommables, espaces de fonctions dérivables avec diverses normes, etc.). Dans tous ces travaux, la description de l'algorithme aussi bien que la démonstration de sa convergence restent toutefois liées à une hypothèse de type Haar sur le problème.

Il était bien senti que l'hypothèse de Haar était trop forte: dans de nombreux cas courants elle n'était jamais vérifiée et des expériences numériques ont montré que l'algorithme d'échange donnait encore d'excellents résultats lorsqu'elle n'était pas satisfaite. Une justification théorique de ce fait a été donnée en 1972: la convergence de l'algorithme de Rémès généralisé a été démontrée [7], [8] sans faire l'hypothèse de Haar mais en supposant que l'on ne rencontre pas de dégénérescences, c'est-à-dire de déterminant égal à zéro, au cours du déroulement effectif de l'algorithme sur un exemple donné et avec des conditions initiales données (hypothèse d'itérativité de l'algorithme). En 1973, la même étude est faite dans le cas de la minimisation d'une fonctionnelle convexe sur un sous-espace vectoriel de dimension finie [22].

Lorsque l'on admet que des dégénérescences (déterminants nuls) peuvent se produire, c'est-à-dire, si l'on renonce à l'hypothèse d'itérativité précédente, non seulement la démonstration de convergence n'est plus valable, mais l'algorithme lui-même doit être modifié. En 1965, Töpfer [33]–[35] a donné, pour le cas de l'approximation uniforme de fonctions continues, le principe d'un algorithme récursif par rapport à la dimension sans faire l'hypothèse de Haar. Il ne démontre toutefois pas sa convergence. L'idée de l'algorithme de Töpfer est reprise en 1973 pour la minimisation d'une fonctionnelle convexe et sa convergence est démontrée [10].

* Received by the editors April 2, 1976, and in revised form January 19, 1977.

† Mathématiques Appliquées, Université de Saint-Etienne, 42100 Saint Etienne, France.

‡ Mathématiques Appliquées, Université de Grenoble, 38041 Grenoble, France.

Par sa présentation même, de caractère récursif, cet algorithme restait toutefois purement théorique: en particulier l'arrêt effectif des sous-algorithmes qu'il mettait en oeuvre dans les différents niveaux de récursivité avait été mal étudié; de même le démarrage d'un sous-algorithme à l'apparition d'une dégénérescence était resté dans l'ombre.

L'algorithme nouveau que nous présentons est une extension de l'algorithme de Cheney-Goldstein qui s'inspire de l'algorithme de Töpfer mais en évitant les inconvénients énumérés ci-dessus. Tout d'abord il n'est pas récursif: la récursivité est en quelque sorte explicitée; la notion qui remplace la récursivité est celle de chaîne d'annulateurs minimaux introduite au § 3. Une version généralisée et complétée du théorème d'échange de Stiefel est donnée au § 4; elle permet de préciser exactement le mécanisme de création d'un nouveau niveau dans une chaîne (ce qui correspondait au démarrage d'un sous-algorithme dans la version récursive). En outre, la stratégie générale de l'algorithme est profondément modifiée: elle dépend de la précision avec laquelle on souhaite connaître la solution et de la précision avec laquelle les étapes intermédiaires sont pratiquées. On montre que la précision fixée a priori est atteinte après un nombre fini d'itérations.

Nous nous sommes limités ici à la minimisation d'une fonction convexe nondifférentiable (définie comme une borne supérieure d'une famille de fonctions affines) sur une variété linéaire (domaine défini par des contraintes linéaires de type égalité). Le même algorithme peut être étendu au cas où le domaine est un convexe fermé défini par un nombre infini d'inégalités linéaires (cf. [23]). Pour l'application de cet algorithme à des problèmes de recherche d'un meilleur approximant, nous renvoyons à [11].

L'algorithme proposé ici étant une extension des algorithmes de Cheney-Goldstein ou Rémès (il coïncide avec eux lorsqu'on ne rencontre pas de dégénérescences) son efficacité est comparable à celle bien connue de ces algorithmes, notamment lorsqu'il s'agit de problèmes d'approximation avec des normes non différentiables. L'avantage du nouvel algorithme réside dans le fait que l'on s'est libéré des hypothèses très fortes qui étaient nécessaires pour assurer le fonctionnement et la convergence des précédents.

Signalons que divers algorithmes ont été proposés par ailleurs pour la minimisation d'une fonction convexe non différentiable. Contrairement à l'algorithme présenté ici qui est de type dual (méthode de montée), la plupart d'entre eux sont des algorithmes directs (méthodes de descente). Citons les méthodes de type "plus profonde descente" (Demjanov [14], Bertsekas and Mitter [3]) et celles de type "gradient conjugué" proposées récemment par Lemaréchal [25] et Wolfe [37] pour la minimisation sans contrainte d'une fonction convexe non différentiable (avec des hypothèses différentes, par exemple d'uniforme convexité).

Ces auteurs n'ont pas comparé numériquement leurs méthodes à l'algorithme plus ancien de Cheney-Goldstein. On peut toutefois penser que les algorithmes directs mentionnés ci-dessus donnent leurs meilleurs résultats lorsque la non-différentiabilité de la fonction à minimiser est limitée (exemple: borne supérieure d'un petit nombre de fonctions quadratiques) alors que l'algorithme dual de Cheney-Goldstein et ses extensions conviennent bien pour des fonctions

convexes dont le caractère non différentiable est fort (un cas typique est celui des fonctions convexes que l'on rencontre dans l'approximation au sens de Tchebycheff; cf. [11], [18]–[20], [24]).

1. Énoncé du problème et hypothèses. On désigne par E l'espace Euclidien de dimension n et on note $\langle x, y \rangle$ le produit scalaire ordinaire de x et y dans E .

1.1. Le problème de minimisation. On désigne par b et c des applications bornées d'un ensemble quelconque T dans E et \mathbb{R} respectivement (i.e., $b(T)$ et $c(T)$ sont des ensembles bornés de E et \mathbb{R}). On définit la fonctionnelle f par

$$f(x) = \sup_{t \in T} (\langle x, b(t) \rangle - c(t)).$$

On montre facilement que f est une fonctionnelle convexe continue définie sur E à valeurs dans \mathbb{R} .

On définit une variété affine W par

$$W = \{x \in E \mid \langle x, \beta_j \rangle = \gamma_j, j = 1, \dots, k_0\},$$

où les $\beta_j, j = 1, \dots, k_0$, sont des éléments linéairement indépendants de E et les $\gamma_j, j = 1, \dots, k_0$, des nombres réels. La variété affine W est parallèle au sous-espace vectoriel

$$V = \{x \in E \mid \langle x, \beta_j \rangle = 0, j = 1, \dots, k_0\},$$

qui est de dimension $n_0 = n - k_0$.

On considère le problème de la minimisation de f sur W . Posons:

$$(P) \quad \alpha = \inf_{x \in W} f(x).$$

On suppose que α est fini et on s'intéresse à l'ensemble S des solutions:

$$S = \{\bar{x} \in W \mid f(\bar{x}) = \alpha\}.$$

On dira que $\tilde{x} \in W$ est une solution à ε près de (P) si l'on a $f(\tilde{x}) - \alpha \leq \varepsilon$.

1.2. Hypothèses. On fera l'hypothèse suivante:

L'ensemble

$$(H) \quad K = \{x \in V \mid \langle x, b(t) \rangle \leq 0, \text{ pour tout } t \in T\}$$

est un sous-espace vectoriel de E .

Remarquons tout d'abord que l'ensemble K précédent est égal au cône asymptote (cf. [21]) de tous les ensembles:

$$S_\lambda = \{x \in W \mid f(x) \leq \lambda\},$$

qui sont non vides. En particulier, s'il existe $\lambda \in \mathbb{R}$ tel que S_λ soit non vide et borné, alors K est réduit à $\{0\}$ et l'hypothèse (H) est donc vérifiée.

Soit f^* la polaire de f , épi $f^* = \{[y, \lambda] \in E \times \mathbb{R} \mid f^*(y) \leq \lambda\}$ son épigraphe et $\text{dom } f^* = \{y \in E \mid f^*(y) < \infty\}$ son domaine effectif. Si l'on note d l'application de T

dans $E \times \mathbb{R}$ définie par $d(t) = [b(t), c(t)]$, on montre facilement que

$$\text{épi } f^* = \overline{\text{co}} d(T) + \{0\} \times [0, +\infty],$$

où $\overline{\text{co}} \Omega$ désigne l'enveloppe convexe fermée de Ω , et l'on en déduit:

$$\text{dom } f^* = \overline{\text{co}} b(T).$$

On sait (cf. [21]) que α est fini si et seulement si $0 \in \text{dom } f^* + V^\perp$. On démontre alors, (en utilisant le théorème 7.7.3 de [21]) que l'hypothèse (H) est équivalente à:

$$(H') \quad 0 \in \text{ir co } b(T) + V^\perp,$$

où $\text{ir } \Omega$ désigne l'intérieur relatif de Ω .

L'hypothèse (H) (ou (H')) a les conséquences suivantes:

- (i) L'ensemble S des solutions est non vide.
- (ii) Il existe $k \in \mathbb{R}$ tel que pour tout $x \in V$ on ait:

$$\sup_{t \in T} |\langle x, b(t) \rangle| \leq k \sup_{t \in T} \langle x, b(t) \rangle.$$

- (iii) Il existe $\gamma \in \mathbb{R}$ et $\delta \in \mathbb{R}$ tels que pour tout $x \in W$ on ait:

$$\sup_{t \in T} |\langle x, b(t) \rangle - c(t)| \leq \gamma f(x) + \delta.$$

1.3. Exemples. (a) Soit X un espace vectoriel normé dont la norme est notée $\|\cdot\|$. Soient v_1, v_2, \dots, v_n, w des éléments linéairement indépendants de X . On cherche l'élément $x \in E$ vérifiant:

$$\langle x, \beta_j \rangle = \gamma_j, \quad j = 1, \dots, k_0,$$

(où $\beta_j \in E$ et $\gamma_j \in \mathbb{R}$, $j = 1, \dots, k_0$) qui minimise la fonctionnelle

$$f(x) = \left\| \sum_{i=1}^n x_i v_i - w \right\|.$$

Il s'agit donc du problème classique de la meilleure approximation dans une variété affine de dimension finie d'un espace vectoriel normé. On remarque que l'on peut toujours écrire:

$$f(x) = \max_{l \in B'} (\langle x, b(l) \rangle - c(l)),$$

où B' désigne la boule unité du dual X' de X et les applications b et c sont définies par

$$b(l) = [l(v_1), \dots, l(v_n)] \quad \text{et} \quad c(l) = l(w).$$

(b) On reprend l'exemple précédent mais avec

$$f(x) = \sup_{w \in C} \left\| \sum_{i=1}^n x_i v_i - w \right\|,$$

où C désigne une partie bornée de X . Ce problème a été étudié dans [24]. On remarque que la fonctionnelle f peut s'écrire:

$$f(x) = \sup_{[l, w] \in B' \times C} (\langle x, b(l, w) \rangle - c(l, w))$$

avec $b(l, w) = [l(v_1), \dots, l(v_n)]$, qui est en fait indépendant de w et $c(l, w) = l(w)$.

Pour ces deux exemples, toutes les hypothèses précédentes (y compris (H)) sont vérifiées.

2. Annulateur minimal. Soit \mathcal{V} un sous-espace vectoriel de dimension d de E et désignons par (v_1, \dots, v_d) une base arbitraire de \mathcal{V} .

2.1. Définitions. (i) Un sous-ensemble A de T sera appelé *annulateur* de \mathcal{V} si l'on a $\text{co}(b(A)) \cap \mathcal{V}^\perp \neq \emptyset$; autrement dit, s'il existe des éléments t_1, \dots, t_m de A et des poids correspondants $\rho_1 \geq 0, \dots, \rho_m \geq 0, \sum_{i=1}^m \rho_i = 1$ tels que l'élément $\sum_{i=1}^m \rho_i b(t_i)$ appartienne à \mathcal{V}^\perp .

Si l'on définit l'application a de T dans \mathbb{R}^d par

$$a(t) = [\langle v_1, b(t) \rangle, \dots, \langle v_d, b(t) \rangle],$$

alors A est un annulateur de \mathcal{V} si et seulement si $0 \in \text{co}(a(A))$.

(ii) Un annulateur A de \mathcal{V} sera dit *minimal* s'il n'existe pas d'annulateur de \mathcal{V} qui soit strictement contenu dans A .

2.2. Propriétés des annulateurs minimaux. Nous énonçons sans démonstration quelques propriétés des annulateurs minimaux (en abrégé a.m.) qui seront utiles pour la suite:

(i) En utilisant le théorème de Carathéodory [21, p. 74], on montre que tout a.m. de \mathcal{V} a au plus $d + 1$ éléments.

(ii) Un sous-ensemble $\{t_1, \dots, t_{k+1}\}$ comportant $k + 1$ éléments distincts est a.m. de \mathcal{V} si et seulement si les deux conditions suivantes sont vérifiées:

(a) il existe des coefficients $\rho_i > 0, \sum_{i=1}^{k+1} \rho_i = 1$ tels que $\sum_{i=1}^{k+1} \rho_i b(t_i) \in \mathcal{V}^\perp$, ce qui revient à dire que $0 \in \text{ir co}(a(A))$;

(b) la variété affine L engendrée par $a(A)$ est de dimension k .

(En tenant compte de (a), on remarque que cette variété affine L passe par l'origine et qu'elle est en fait égale au sous-espace vectoriel engendré par k quelconques des éléments de $a(A)$.)

(iii) De façon équivalente, un sous-ensemble A de T est un a.m. de \mathcal{V} si et seulement si $\text{co}(a(A))$ est un simplexe de \mathbb{R}^d (cf. [29, p. 12]) qui contient 0 dans son intérieur relatif.

(iv) Un sous-ensemble $A = \{t_1, \dots, t_{k+1}\}$ de T est un a.m. de \mathcal{V} si et seulement s'il existe des coefficients strictement positifs ρ_i uniques vérifiant $\sum_{i=1}^{k+1} \rho_i = 1$ tels que $\sum_{i=1}^{k+1} \rho_i b(t_i) \in \mathcal{V}^\perp$. Ces coefficients seront aussi notés $\rho_A(t)$, $t \in A$ (on a $\rho_A(t_i) = \rho_i$) et seront appelés les poids associés à A (et à \mathcal{V}).

(v) Etant donnés des éléments fixés t_1, \dots, t_m de T , considérons l'ensemble:

$$P = \{\rho \in \mathbb{R}^m \mid \rho_i \geq 0, i = 1, \dots, m; \sum_{i=1}^m \rho_i = 1; \sum_{i=1}^m \rho_i b(t_i) \in \mathcal{V}^\perp\}.$$

Un élément $\bar{\rho} \in P$ définit un annulateur $A = \{t_i | i = 1, \dots, m, \bar{\rho}_i > 0\}$ de \mathcal{V} qui est minimal si et seulement si $\bar{\rho}$ est un point extrémal de P . De la même façon, si $y \in \mathcal{V}^\perp$ est fixé, un élément $\bar{\rho}$ appartenant à l'ensemble

$$Q = \{\rho \in \mathbb{R}^m | \rho_i \geq 0, i = 1, \dots, m; \sum_{i=1}^m \rho_i = 1; \sum_{i=1}^m \rho_i b(t_i) = y\}$$

définit un annulateur $A = \{t_i | i = 1, \dots, m; \bar{\rho}_i > 0\}$ de \mathcal{V} qui est minimal si et seulement si $\bar{\rho}$ est un point extrémal de Q .

(vi) Si $A_0 = \{t_1, \dots, t_{k+1}\}$ est un a.m. de \mathcal{V} , le sous-espace vectoriel

$$\mathcal{V}_0 = \{x \in \mathcal{V} | \langle x, b(t_i) \rangle = 0, i = 1, \dots, k+1\} = \{x \in \mathcal{V} | \langle x, b(t_i) \rangle = 0, i = 1, \dots, k\}$$

que l'on appellera noyau de A_0 dans \mathcal{V} , est de dimension $d - k$.

Un sous-ensemble $A_1 = \{u_1, \dots, u_{h+1}\}$ de T est alors a.m. de \mathcal{V}_0 si et seulement si les deux conditions suivantes sont satisfaites:

(a) il existe des coefficients $\mu_i > 0, \sum_{i=1}^{h+1} \mu_i = 1$ tels que $\sum_{i=1}^{h+1} \mu_i a(u_i) \in L_0$, où L_0 est la variété affine de dimension k (qui est en fait un sous-espace vectoriel, cf. (ii)) engendrée par les $a(t_i), i = 1, \dots, k+1$;

(b) la variété affine (en fait sous-espace vectoriel) L_1 engendrée par les éléments $a(t_i), i = 1, \dots, k+1$ et $a(u_i), i = 1, \dots, h+1$ est de dimension $k+h$.

2.3. Minimisation de la fonctionnelle polyédrale associée à un annulateur minimal. Soit A un a.m. de \mathcal{V} . Considérons la fonctionnelle polyédrale f_A associée à A qui est définie par

$$f_A(x) = \max_{t \in A} (\langle x, b(t) \rangle - c(t)).$$

Il est clair que f_A est une minorante de f .

Soit \mathcal{W} une variété affine parallèle au sous-espace vectoriel \mathcal{V} (c'est-à-dire, s'en déduisant par translation) et étudions le problème de la minimisation de f_A sur \mathcal{W} . Posons:

$$(P_A) \quad \alpha_A = \inf_{x \in \mathcal{W}} f_A(x).$$

On démontre que α_A est fini et que l'on a:

$$\alpha_A = \sum_{t \in A} \rho_A(t) (\langle x_0, b(t) \rangle - c(t)),$$

où x_0 est un élément arbitraire de \mathcal{W} . De plus l'ensemble $\mathcal{W}_A = \{x \in \mathcal{W} | f_A(x) = \alpha_A\}$ des solutions de (P_A) est donné par

$$\mathcal{W}_A = \{x \in \mathcal{W} | \langle x, b(t) \rangle - c(t) = \alpha_A, \text{ pour tout } t \in A\}.$$

C'est une variété affine parallèle au sous-espace vectoriel

$$\mathcal{V}_A = \{x \in \mathcal{V} | \langle x, b(t) \rangle = 0, \text{ pour tout } t \in A\},$$

qui est le noyau de A dans \mathcal{V} . Si \mathcal{V} est de dimension d , et si A comporte $k+1$ éléments, \mathcal{V}_A est donc de dimension $d - k$.

La solution de (P_A) est donc unique si et seulement si $k = d$.

2.4. Problème dual associé au problème (P). Appelons \mathcal{A} l'ensemble des annulateurs minimaux de V . On a alors:

THEOREME. *Le montant α du problème (P) vérifie:*

$$\alpha = \sup_{A \in \mathcal{A}} \sum_{t \in A} \rho_A(t) (\langle x_0, b(t) \rangle - c(t)),$$

où x_0 est un élément arbitraire de W et $\rho_A(t)$, $t \in A$ sont les poids associés à A et à V .

Démonstration. Etant donné $\varepsilon > 0$, arbitraire, on va montrer qu'il existe $A \in \mathcal{A}$ tel que

$$\sum_{t \in A} \rho_A(t) (\langle x_0, b(t) \rangle - c(t)) > \alpha - \varepsilon.$$

On sait que l'on a (cf. [21]):

$$\alpha = \max_{y \in \text{dom}(f^*) \cap V^\perp} (\langle x_0, y \rangle - f^*(y)),$$

où x_0 est un élément arbitraire de W .

Le demi-espace ouvert de $E \times \mathbb{R}$:

$$D = \{[y, \lambda] \in E \times \mathbb{R} \mid \langle x_0, y \rangle - \lambda > \alpha - \varepsilon\}$$

coupé donc épi $(f^*) \cap (V^\perp \times \mathbb{R})$, donc aussi $\overline{\text{co}}(d(T)) \cap (V^\perp \times \mathbb{R})$. D'après l'hypothèse (H'), l'ensemble $(\text{ir co } b(T)) \cap V^\perp$ n'est pas vide. Il en résulte que $\overline{\text{co}}(d(T)) \cap (V^\perp \times \mathbb{R})$ est la fermeture de $\text{co}(d(T)) \cap (V^\perp \times \mathbb{R})$ et ainsi l'ensemble

$$D \cap \text{co}(d(T)) \cap (V^\perp \times \mathbb{R})$$

n'est pas vide. Soit $[y, \lambda]$ un élément de cet ensemble.

Il existe $t_1, \dots, t_m \in T$ et $\mu_i \geq 0$, $i = 1, \dots, m$, $\sum_{i=1}^m \mu_i = 1$ tels que

$$y = \sum_{i=1}^m \mu_i b(t_i), \quad \lambda = \sum_{i=1}^m \mu_i c(t_i)$$

et $\langle x_0, y \rangle - \lambda > \alpha - \varepsilon$.

Définissons l'ensemble:

$$Q = \left\{ \rho \in \mathbb{R}^m \mid \rho_i \geq 0, i = 1, \dots, m; \sum_{i=1}^m \rho_i = 1; \sum_{i=1}^m \rho_i b(t_i) = y \right\}.$$

Comme $\mu \in Q$, on a:

$$\min_{\rho \in Q} \sum_{i=1}^m \rho_i c(t_i) \leq \sum_{i=1}^m \mu_i c(t_i) = \lambda.$$

Comme Q est un convexe compact, il existe au moins un élément $\bar{\rho}$ extrémal de Q pour lequel le minimum est atteint. Soit $I = \{i \in \{1, \dots, m\} \mid \bar{\rho}_i > 0\}$ et $A = \{t_i \mid i \in I\}$. D'après § 2.2(v), A est un annulateur minimal de V et l'on a $\sum_{i \in I} \bar{\rho}_i c(t_i) \leq \lambda$, donc:

$$\langle x_0, \sum_{i \in I} \bar{\rho}_i b(t_i) \rangle - \sum_{i \in I} \bar{\rho}_i c(t_i) > \alpha - \varepsilon.$$

L'algorithme que nous allons décrire est dual dans le sens qu'il s'attache plutôt à résoudre le problème de maximisation qui est défini dans le théorème ci-dessus. Etant donné un nombre $\varepsilon > 0$, il fournira une séquence finie A^1, A^2, \dots, A^μ d'annulateurs minimaux de V et une séquence associée x^1, x^2, \dots, x^μ d'éléments de W telles que

$$\alpha^\nu = \sum_{t \in A^\nu} \rho_{A^\nu}(t)(\langle x_0, b(t) \rangle - c(t))$$

forme une séquence non décroissante, avec $f(x^\mu) - \alpha^\mu \leq \varepsilon$, ce qui entraîne simultanément $\alpha - \alpha^\mu \leq \varepsilon$ et $f(x^\mu) - \alpha \leq \varepsilon$, donc en particulier que x^μ est une solution à ε près de (P).

3. Chaîne d'annulateurs minimaux. Soit A_1 un annulateur minimal de V . Appelons $f_1 = f_{A_1}$ la fonctionnelle polyédrale qui lui est associée et considérons la minimisation de f_1 sur W . Notons α_1 le montant du minimum, W_1 l'ensemble des solutions et V_1 le sous-espace correspondant auquel W_1 est parallèle. Si A_1 comporte $k_1 + 1$ éléments, le sous-espace V_1 est de dimension $n_1 = n_0 - k_1 = n - k_0 - k_1$.

On refait la même construction mais relativement à V_1 : Si A_2 désigne un annulateur minimal de V_1 comportant $k_2 + 1$ éléments, on forme la fonctionnelle $f_2 = f_{A_2}$ et on note α_2 le montant de son minimum sur W_1 , W_2 l'ensemble des solutions et V_2 le sous-espace vectoriel de dimension $n_2 = n_1 - k_2 = n_0 - \sum_{i=0}^2 k_i$ auquel W_2 est parallèle.

On continue ainsi de proche en proche cette construction.

3.1. Chaîne d'annulateurs minimaux.

DEFINITION. On appelle *chaîne d'annulateurs minimaux* (en abrégé *chaîne*) une séquence finie d'annulateurs minimaux $\mathcal{C} = \{A_1, \dots, A_m\}$ obtenue comme ci-dessus pour laquelle on a $V_m = \{0\}$.

En résumé, si l'on pose $V_0 = V$, la séquence $\{A_1, \dots, A_m\}$ est une chaîne si l'on a:

A_i est un annulateur minimal de V_{i-1} ,

$V_i = \{x \in V_{i-1} | \langle x, b(t) \rangle = 0, \text{ pour tout } t \in A_i\}$,

$i = 1, \dots, m$,

$V_m = \{0\}$.

3.2. Solution associée à une chaîne. A une chaîne $\mathcal{C} = \{A_1, A_2, \dots, A_m\}$ sont associées de façon automatique, comme on l'a vu plus haut, la séquence des fonctionnelles polyédrales:

$$\{f_1, f_2, \dots, f_m\},$$

la séquence des variétés affines (ensembles de solutions successifs):

$$\{W_1, W_2, \dots, W_m\},$$

et des sous-espaces vectoriels correspondants:

$$\{V_1, V_2, \dots, V_m\} \quad (\text{avec } V_m = \{0\}),$$

la séquence des montants des minima de f_i sur W_{i-1} :

$$\{\alpha_1, \alpha_2, \dots, \alpha_m\}.$$

Comme $V_m = \{0\}$, la variété affine W_m est réduite à un point $x = x_{\mathcal{C}}$ qui est appelé *solution associée à la chaîne \mathcal{C}* .

Le calcul de $x = x_{\mathcal{C}}$ et des montants $\alpha_1, \alpha_2, \dots, \alpha_m$ se fait en résolvant le système linéaire suivant qui a $n + m$ équations et $n + m$ inconnues (et qui a une solution unique):

$$\begin{cases} \langle x, \beta_j \rangle = \gamma_j, & j = 1, \dots, k_0 & (k_0 \text{ équations}) \\ \langle x, b(t) \rangle - \alpha_i = c(t), & t \in A_i, & (k_i + 1 \text{ équations}), \\ & i = 1, \dots, m. \end{cases}$$

Donnons la structure de ce système dans le cas particulier où $n = 5$, $k_0 = 1$, $k_1 = 2$, $k_2 = 1$, $k_3 = 1$, en notant $A_i = \{t_{i,1}, \dots, t_{i,k_i+1}\}$:

β_1	0	0	0	\times	x_1	$=$	γ_1
$b(t_{11})$	-1	0	0		x_2		$c(t_{11})$
$b(t_{12})$	-1	0	0		x_3		$c(t_{12})$
$b(t_{13})$	-1	0	0		x_4		$c(t_{13})$
$b(t_{21})$	0	-1	0		x_5		$c(t_{21})$
$b(t_{22})$	0	-1	0		α_1		$c(t_{22})$
$b(t_{31})$	0	0	-1		α_2		$c(t_{31})$
$b(t_{32})$	0	0	-1		α_3		$c(t_{32})$

3.3. Chaîne régulière.

DEFINITION. On dira qu'une chaîne $\mathcal{C} = \{A_1, \dots, A_m\}$ est *régulière* si tous les annulateurs minimaux qui la composent sont constitués d'au moins deux éléments, c'est à dire si $k_i \geq 1$, $i = 1, \dots, m$.

Si la chaîne est régulière, V_i a donc une dimension strictement inférieure à celle de V_{i-1} . Ainsi la longueur m d'une chaîne régulière (c'est-à-dire, le nombre m d'annulateur minimaux qui la composent) est inférieure ou égale à la dimension n_0 de V_0 .

Etant donnée une chaîne quelconque \mathcal{C} , si l'on supprime tous les annulateurs minimaux réduits à un point, on obtient une nouvelle chaîne \mathcal{C}' qui est régulière. Cette opération ne change pas la solution x associée à la chaîne et les montants α_i qui n'ont pas été supprimés.

4. Théorème d'échange généralisé. L'algorithme sera basé sur le théorème suivant qui généralise et complète le théorème classique d'échange de Stiefel [30], [31], [21, pp. 117, 462].

4.1. THEOREME. Soit \mathcal{V} un sous-espace vectoriel quelconque de E . Si A_0 est un annulateur minimal de \mathcal{V} et si A_1 est un annulateur minimal du noyau $\mathcal{V}_0 = \{x \in \mathcal{V} | \langle x, b(t) \rangle = 0, \text{ pour tout } t \in A_0\}$, de A_0 dans \mathcal{V} , alors il existe une bipartition de A_0 en B_0 et $C_0 \neq \emptyset$ telle que $\tilde{A}_0 = B_0 \cup A_1$ soit un annulateur minimal de \mathcal{V} et $\tilde{A}_1 = C_0$ soit un annulateur minimal du noyau $\tilde{\mathcal{V}}_0 = \{x \in \mathcal{V} | \langle x, b(t) \rangle = 0, \text{ pour tout } t \in \tilde{A}_0\}$ de \tilde{A}_0 dans \mathcal{V} .

Démonstration. La démonstration que nous allons donner est constructive: elle pourra être utilisée dans l'algorithme pour pratiquer effectivement l'opération d'échange.

On suppose encore que \mathcal{V} est un sous-espace vectoriel dimension d engendré par v_1, \dots, v_d et on note $A_0 = \{t_1, \dots, t_{k+1}\}$, $A_1 = \{u_1, \dots, u_{h+1}\}$. On reprend toutes les notations utilisées dans le § 2.2(vi).

Comme A_0 est un a.m. de \mathcal{V} ,

(i) il existe des coefficients $\rho_i > 0$, $\sum_{i=1}^{k+1} \rho_i = 1$ tels que $\sum_{i=1}^{k+1} \rho_i a(t_i) = 0$,

(ii) le sous-espace vectoriel $L_0 = \mathcal{L}(a(t_i), i = 1, \dots, k+1)$ est de dimension k .

De même, comme A_1 est un a.m. de \mathcal{V}_0 ,

(iii) il existe des coefficients $\mu_i > 0$, $\sum_{i=1}^{h+1} \mu_i = 1$ tels que $\sum_{i=1}^{h+1} \mu_i a(u_i) \in L_0$,

(iv) le sous-espace $L_1 = \mathcal{L}(a(t_i), i = 1, \dots, k+1; a(u_i), i = 1, \dots, h+1)$ est de dimension $k+h$.

Posons $z_i = a(t_i)$, $i = 1, \dots, k+1$ et $z_0 = \sum_{i=1}^{h+1} \mu_i a(u_i)$. Comme z_0 appartient à L_0 , les éléments z_0, z_1, \dots, z_k sont linéairement dépendants: il existe donc des coefficients $\omega_0, \omega_1, \dots, \omega_k$ (avec $\omega_0 > 0$) tels que $\sum_{i=0}^k \omega_i z_i = 0$. On posera $\omega_{k+1} = 0$ et $\rho_0 = 0$. Quel que soit $\theta \in \mathbb{R}$, on a donc:

$$\sum_{i=0}^{k+1} (\omega_i - \theta \rho_i) z_i = \omega_0 z_0 + \sum_{i=1}^{k+1} \rho_i \left(\frac{\omega_i}{\rho_i} - \theta \right) = 0.$$

Comme ω_0 est positif, on voit donc que si l'on prend

$$(*) \quad \bar{\theta} = \min_{i=1, \dots, k+1} \left(\frac{\omega_i}{\rho_i} \right)$$

et si l'on pose $\lambda_i = \omega_i - \bar{\theta} \rho_i$, $i = 0, \dots, k+1$, on aura $\lambda_i \geq 0$, $i = 0, \dots, k+1$.

Notons $I = \{i \in \{1, \dots, k+1\} | \lambda_i = 0\}$. Cet ensemble n'est pas vide: c'est l'ensemble des indices i pour lesquels le minimum est atteint dans (*). Soit \tilde{I} le complémentaire de I dans $\{1, \dots, k+1\}$. On a:

$$\tilde{I} = \{i \in \{1, \dots, k+1\} | \lambda_i > 0\}.$$

On posera alors $\tilde{\rho}_i = \lambda_i / \lambda$, pour $i \in \tilde{I} \cup \{0\}$, avec $\lambda = \lambda_0 + \sum_{i \in \tilde{I}} \lambda_i$.

Décomposons l'ensemble A_0 en B_0 et C_0 de la façon suivante:

$$\begin{aligned} B_0 &= \{t_i | i \in \tilde{I}\}, \\ C_0 &= \{t_i | i \in I\} \neq \emptyset. \end{aligned}$$

Si l désigne le nombre d'éléments de B_0 , $0 \leq l \leq k$, alors l'ensemble C_0 comporte $\tilde{h} + 1$ éléments avec $\tilde{h} = k - l$.

On pose comme dans l'énoncé du théorème:

$$\tilde{A}_0 = A_1 \cup B_0,$$

$$\tilde{A}_1 = C_0,$$

$$\tilde{\mathcal{V}}_0 = \text{noyau de } \tilde{A}_0 \text{ dans } \mathcal{V}$$

$$= \{x \in \mathcal{V} \mid \langle x, b(u_i) \rangle = 0, i = 1, \dots, h+1; \langle x, b(t_i) \rangle = 0, i \in \tilde{I}\}.$$

Comme on a:

$$\tilde{\rho}_0 z_0 + \sum_{i \in \tilde{I}} \tilde{\rho}_i z_i = \sum_{i=1}^{h+1} \tilde{\rho}_0 \mu_i a(u_i) + \sum_{i \in \tilde{I}} \tilde{\rho}_i a(t_i) = 0,$$

\tilde{A}_0 est annulateur de \mathcal{V} . Comme il comporte $\tilde{k} + 1$ éléments (avec $\tilde{k} = l + h$) et que l'on peut vérifier que

$$\tilde{L}_0 = \mathcal{L}(a(u_i), i = 1, \dots, h+1; a(t_i), i \in \tilde{I})$$

est de dimension $\tilde{k} = l + h$, on en déduit que \tilde{A}_0 est minimal.

Par ailleurs, comme on a:

$$\sum_{i=1}^{k+1} \rho_i z_i = 0,$$

on peut écrire:

$$\sum_{i \in I} \left(\frac{\rho_i}{\sigma} \right) z_i = -\frac{1}{\sigma} \sum_{i \in \tilde{I}} \rho_i z_i \quad \text{avec } \sigma = \sum_{i \in I} \rho_i;$$

donc

$$\sum_{i \in I} \left(\frac{\rho_i}{\sigma} \right) a(t_i) \in \tilde{L}_0$$

et $\tilde{A}_1 = C_0$ est ainsi un annulateur de $\tilde{\mathcal{V}}_0$.

Comme \tilde{A}_0 est un a.m. qui comporte $\tilde{k} + 1$ éléments et que \tilde{A}_1 comporte $\tilde{h} + 1$ éléments, pour montrer que \tilde{A}_1 est un a.m. de $\tilde{\mathcal{V}}_0$, il nous reste à vérifier (d'après § 2.2(vi)) que

$$\tilde{L}_1 = \mathcal{L}(a(u_i), i = 1, \dots, h+1; a(t_i), i \in \tilde{I}; a(t_i), i \in I)$$

est de dimension $\tilde{k} + \tilde{h}$. Or on remarque que $\tilde{L}_1 = L_1$ et $\tilde{k} + \tilde{h} = (l + h) + (k - l) = k + h$, d'où le résultat.

4.2. Cas particuliers du théorème d'échange. L'appellation "théorème d'échange" est justifiée par le fait que l'on peut énoncer le théorème 4.1. de la façon suivante:

Si A_0 est un a.m. de \mathcal{V} et A_1 est un a.m. du noyau \mathcal{V}_0 de A_0 dans \mathcal{V} , alors on peut *échanger* l'ensemble des éléments de A_1 avec une partie C_0 des éléments de A_0 telle que le nouvel ensemble

$$\tilde{A}_0 = (A_0 \setminus C_0) \cup A_1$$

soit à nouveau un a.m. de \mathcal{V} et l'ensemble C_0 des éléments enlevés à A_0 forme un a.m. du noyau $\tilde{\mathcal{V}}_0$ de \tilde{A}_0 dans \mathcal{V} .

Supposons que \mathcal{V} soit de dimension d et que A_0 comporte exactement $d+1$ éléments. Alors son noyau dans \mathcal{V} est réduit à l'origine: $\mathcal{V}_0 = \{0\}$. Ainsi tout ensemble $A_1 = \{\hat{t}\}$ réduit à un seul élément $\hat{t} \in T$ est évidemment un a.m. de \mathcal{V}_0 . Le théorème 4.1. affirme alors que l'on peut échanger \hat{t} avec une partie C_0 des éléments de A_0 de sorte que

$$\tilde{A}_0 = (A_0 \setminus C_0) \cup \{\hat{t}\}$$

soit encore un a.m. de \mathcal{V} et C_0 soit un a.m. du noyau $\tilde{\mathcal{V}}_0$ de \tilde{A}_0 dans \mathcal{V} . Lorsque C_0 se compose d'un seul élément t_0 (ce qui signifie que le minimum dans l'équation (*) du théorème 4.1. est atteint pour un seul indice i_0), alors on a simplement échangé \hat{t} avec l'un des éléments t_0 de A_0 de sorte que le nouvel ensemble

$$\tilde{A}_0 = (A_0 \setminus \{t_0\}) \cup \{\hat{t}\}$$

forme un a.m. de \mathcal{V} . Dans ce cas on a le théorème classique de Stiefel.

4.3. Exemples. Nous allons illustrer le théorème 4.1. par quelques exemples simples mais typiques. Supposons que \mathcal{V} soit de dimension 2. Dans les figures suivantes nous représentons seulement les images par l'application a (définie au § 2.1.) des points de T qui sont considérés.

Exemple 1. Ce premier exemple correspond au théorème d'échange classique de Stiefel: $A_0 = \{t_1, t_2, t_3\}$, $A_1 = \{u_1\}$.

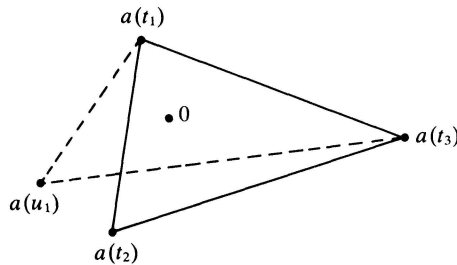


FIG. 1

Comme 0 appartient à l'intérieur de l'enveloppe convexe des trois points $a(t_1)$, $a(t_2)$, $a(t_3)$ (voir la figure 1) A_0 est un a.m. de \mathcal{V} . Comme \mathcal{V}_0 est ici égal à $\{0\}$, tout point u_1 est un a.m. de \mathcal{V}_0 . On voit que l'on peut échanger $a(u_1)$ avec $a(t_2)$ de sorte que $\tilde{A}_1 = \{t_1, t_3, u_1\}$ soit un a.m. de \mathcal{V} (et t_2 est un a.m. $\tilde{\mathcal{V}}_0$ réduit à $\{0\}$).

Exemple 2. $A_0 = \{t_1, t_2\}$, $A_1 = \{u_1, u_2\}$.

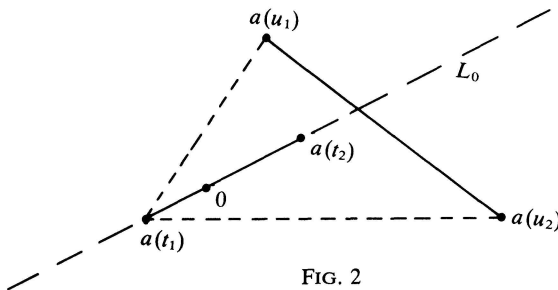


FIG. 2

Comme 0 appartient à l'intérieur relatif de l'enveloppe convexe des deux points $a(t_1)$, $a(t_2)$ (voir la figure 2) et que ces deux points engendrent une variété affine L_0 qui est un sous-espace vectoriel de dimension 1, on voit (cf. § 2.2(ii)) que A_0 est un a.m. de \mathcal{V} . En se basant sur § 2.2(vi), on vérifie que A_1 est un a.m. de \mathcal{V}_0 (l'intérieur relatif de l'enveloppe convexe des deux points $a(u_1)$, $a(u_2)$ coupe L_0 et la variété affine engendrée par les quatre points est égale à \mathbb{R}^2). On peut échanger A_1 avec t_2 de sorte que $\tilde{A}_0 = \{t_1, u_1, u_2\}$ soit un a.m. de \mathcal{V} et $\tilde{A}_1 = \{t_2\}$ est un a.m. de $\tilde{\mathcal{V}}_0$ réduit à $\{0\}$.

Exemple 3. $A_0 = \{t_1, t_2, t_3\}$, $A_1 = \{u_1\}$.

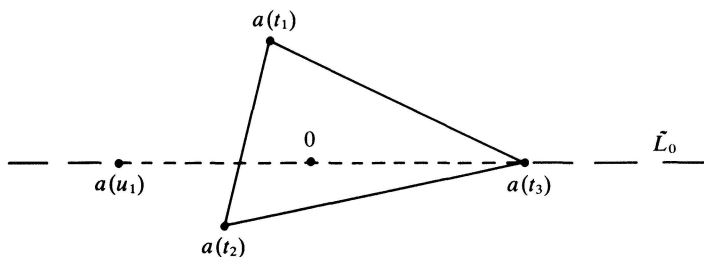


FIG. 3

Nous revenons à la situation initiale de l'exemple 1, où A_0 formé de trois points est un a.m. de \mathcal{V} et A_1 réduit à un point est un a.m. de $\mathcal{V}_0 = \{0\}$. Dans le cas présent (voir la figure 3), on échange u_1 avec les deux points t_1, t_2 : on obtient $\tilde{A}_0 = \{t_3, u_1\}$ qui est un a.m. de \mathcal{V} et $\tilde{A}_1 = \{t_1, t_2\}$ est un a.m. de $\tilde{\mathcal{V}}_0$ qui est de dimension 1 (on retrouve la situation initiale de l'exemple 2).

4.4. Opération d'échange sur une chaîne. L'algorithme sera principalement basé sur l'opération qui consiste à appliquer le théorème d'échange sur deux annulateurs successifs dans une chaîne.

DEFINITION. Etant donnée une chaîne $\mathcal{C} = \{A_1, \dots, A_m\}$, on dira que l'on *échange* A_{j-1} et A_j si l'on remplace ces deux annulateurs par \tilde{A}_{j-1} et \tilde{A}_j selon le théorème d'échange 4.1., de façon à obtenir une nouvelle chaîne.

En effet, A_{j-1} est un a.m. de $\mathcal{V} = V_{j-2}$ et A_j est un a.m. du noyau $\mathcal{V}_0 = V_{j-1}$ de A_{j-1} dans \mathcal{V} .

On peut donc trouver une bipartition de A_{j-1} en B_{j-1} et $C_{j-1} \neq \emptyset$ telle que, si l'on pose $\tilde{A}_{j-1} = B_{j-1} \cup A_j$ et $\tilde{A}_j = C_{j-1}$, alors

$$\tilde{\mathcal{C}} = \{A_1, A_2, \dots, A_{j-2}, \tilde{A}_{j-1}, \tilde{A}_j, \dots, A_m\}$$

constitue à nouveau une chaîne.

On remarque que si \mathcal{C} est une chaîne régulière, il n'en est pas forcément de même pour $\tilde{\mathcal{C}}$ (l'annulateur \tilde{A}_j peut être réduit à un point).

5. Algorithme. On suppose que pour tout $\varepsilon > 0$ et tout $x \in W$, il est possible de déterminer un $t \in T$ tel que

$$f(x) - [\langle x, b(t) \rangle - c(t)] \leq \varepsilon.$$

L'algorithme va consister en la construction d'une suite de chaînes régulières:

$$\mathcal{C} = \{A_1^\nu, \dots, A_m^\nu\},$$

telle que la suite des montants associés

$$\{\alpha_1^\nu, \dots, \alpha_m^\nu\}$$

soit lexicographiquement strictement croissante, c'est-à-dire, que pour tout ν , il existe un entier l^ν , $1 \leq l^\nu \leq m^\nu$, tel que

$$\begin{aligned} \alpha_i^{\nu+1} &= \alpha_i^\nu, & i &= 1, \dots, l^\nu - 1, \\ \alpha_{l^\nu}^{\nu+1} &> \alpha_{l^\nu}^\nu. \end{aligned}$$

Etant donné un nombre positif ε_1 , arbitrairement petit, on montrera qu'après un nombre fini μ d'itérations on obtiendra une chaîne \mathcal{C}^μ et une solution $x^\mu \in W$ telle que l'on ait:

$$f(x^\mu) - \alpha_1^\mu \leq \varepsilon_1$$

ce qui entraîne:

$$\alpha - \alpha_1^\mu \leq \varepsilon_1 \quad \text{et} \quad f(x^\mu) - \alpha \leq \varepsilon_1,$$

donc que x^μ est une solution à ε_1 près du problème (P).

Si ε_1 est la précision à atteindre, on se donne des nombres positifs $\varepsilon_2, \dots, \varepsilon_{n_0+1}$ tels que

$$(*) \quad \varepsilon_{i+1} < \varepsilon_i/2, \quad i = 1, \dots, n_0.$$

5.1. Description de l'algorithme. Supposons que, à l'itération ν , on ait une chaîne régulière $\mathcal{C}^\nu = \{A_1^\nu, \dots, A_m^\nu\}$, la solution x^ν , ainsi que les montants $\alpha_1^\nu, \dots, \alpha_m^\nu$ qui lui correspondent.

Déterminons un élément $t^\nu \in T$ tel que

$$f(x^\nu) - [\langle x^\nu, b(t^\nu) \rangle - c(t^\nu)] \leq \varepsilon_{m^\nu+1}$$

et posons:

$$A_{m^\nu+1}^\nu = \{t^\nu\},$$

$$\alpha_{m^\nu+1}^\nu = \langle x^\nu, b(t^\nu) \rangle - c(t^\nu)$$

(on remarque que $A_{m^\nu+1}^\nu$ peut être considéré comme un a.m. de $V_{m^\nu} = \{0\}$). Soit

$$J^\nu = \{j \in \{1, \dots, m^\nu + 1\} \mid \alpha_{m^\nu+1}^\nu + \varepsilon_{m^\nu+1} \leq \alpha_j^\nu + \varepsilon_j\}.$$

On voit que l'ensemble J^ν contient au moins l'indice $m^\nu + 1$. Notons

$$j^\nu = \min \{j \mid j \in J^\nu\}.$$

On distinguera trois cas suivant la valeur de j^ν :

1^{er} cas: $j^\nu = 1$. On a alors $f(x^\nu) \leq \alpha_{m^\nu+1}^\nu + \varepsilon_{m^\nu+1} \leq \alpha_1^\nu + \varepsilon_1$, donc $f(x^\nu) - \alpha_1^\nu \leq \varepsilon_1$, ce qui signifie que x^ν est solution à ε_1 près de (P); on arrête donc le calcul.

2^{ème} cas: $2 \leq j^\nu \leq m^\nu$. On échange alors $A_{j^\nu-1}^\nu$ et $A_{j^\nu}^\nu$ au sens qui a été précisé au § 4.3, c'est-à-dire, qu'on les remplace par $\tilde{A}_{j^\nu-1}^\nu$ et $\tilde{A}_{j^\nu}^\nu$ selon le théorème

d'échange. On obtient ainsi une nouvelle chaîne \mathcal{C}^ν :

$$\mathcal{C}^\nu = \{A_1^\nu, \dots, A_{j^\nu-2}^\nu, \tilde{A}_{j^\nu-1}^\nu, \tilde{A}_{j^\nu}^\nu, \dots, A_{m^\nu}^\nu\}$$

(l'annulateur $A_{m^\nu+1}^\nu = \{t^\nu\}$ n'est pas introduit dans \mathcal{C}^ν).

On remarque que $\tilde{A}_{j^\nu-1}^\nu$ ne peut pas être réduit à un point, car il contient $A_{j^\nu}^\nu$ et la chaîne \mathcal{C}^ν a été supposée régulière. Par contre $\tilde{A}_{j^\nu}^\nu$ peut être réduit à un point: si c'est le cas, on le supprime; on aboutit ainsi à une chaîne $\mathcal{C}^{\nu+1}$ qui est régulière, et dont le nombre de niveaux est égal, soit à m^ν , soit à $m^\nu - 1$.

3^{ème} cas: $j^\nu = m^\nu + 1$. Partant de la chaîne

$$\{A_1^\nu, A_2^\nu, \dots, A_{m^\nu}^\nu, \{t^\nu\}\}$$

on détermine le plus petit indice i^ν , $1 \leq i^\nu \leq m^\nu + 1$, tel que

$$\{A_1^\nu, A_2^\nu, \dots, A_{i^\nu-1}^\nu, \{t^\nu\}, A_{i^\nu}^\nu, \dots, A_{m^\nu}^\nu\}$$

soit encore une chaîne. Si l'on pose:

$$I^\nu = \{i \in \{1, \dots, m^\nu + 1\} | b(t^\nu) \in V_{i-1}^\perp\},$$

on a:

$$i^\nu = \min \{i | i \in I^\nu\}.$$

Si l'on a $i^\nu = 1$, cela signifie que $\{t^\nu\}$ est un annulateur de V et que le montant correspondant:

$$\alpha_{m^\nu+1}^\nu = \langle x^\nu, b(t^\nu) \rangle - c(t^\nu)$$

vérifie:

$$f(x^\nu) - \alpha_{m^\nu+1}^\nu \leq \varepsilon_{m^\nu+1} \leq \varepsilon_1,$$

donc que x^ν est une solution à ε_1 près de (P). On arrête donc le calcul.

Si l'on a $2 \leq i^\nu \leq m^\nu + 1$, on échange $A_{i^\nu-1}^\nu$ et $\{t^\nu\}$, ce qui donne $\tilde{A}_{i^\nu-1}^\nu$ et $\tilde{A}_{i^\nu}^\nu$. L'annulateur minimal $\tilde{A}_{i^\nu-1}^\nu$ ne peut être réduit à un point, car ce point serait t^ν et on a $b(t^\nu) \notin (V_{i^\nu-2}^\nu)^\perp$. Si l'annulateur minimal $\tilde{A}_{i^\nu}^\nu$ est réduit à un point, on le supprime comme dans le deuxième cas. On aboutit ainsi à une chaîne régulière $\mathcal{C}^{\nu+1}$ dont le nombre de niveaux est égal, soit à $m^\nu + 1$, soit à m^ν .

Dans les trois cas, on aboutit donc (sauf si le calcul s'arrête, la précision étant atteinte) à une nouvelle chaîne régulière $\mathcal{C}^{\nu+1}$. On calcule alors la nouvelle solution $x^{\nu+1}$ et les montants $\alpha_1^{\nu+1}, \dots, \alpha_{m^{\nu+1}}^{\nu+1}$ qui lui sont associés.

Remarque. Les opérations effectuées dans le troisième cas peuvent être remplacées par une succession d'opérations d'échange (ce sera le cas dans l'organigramme donné ci-dessous):

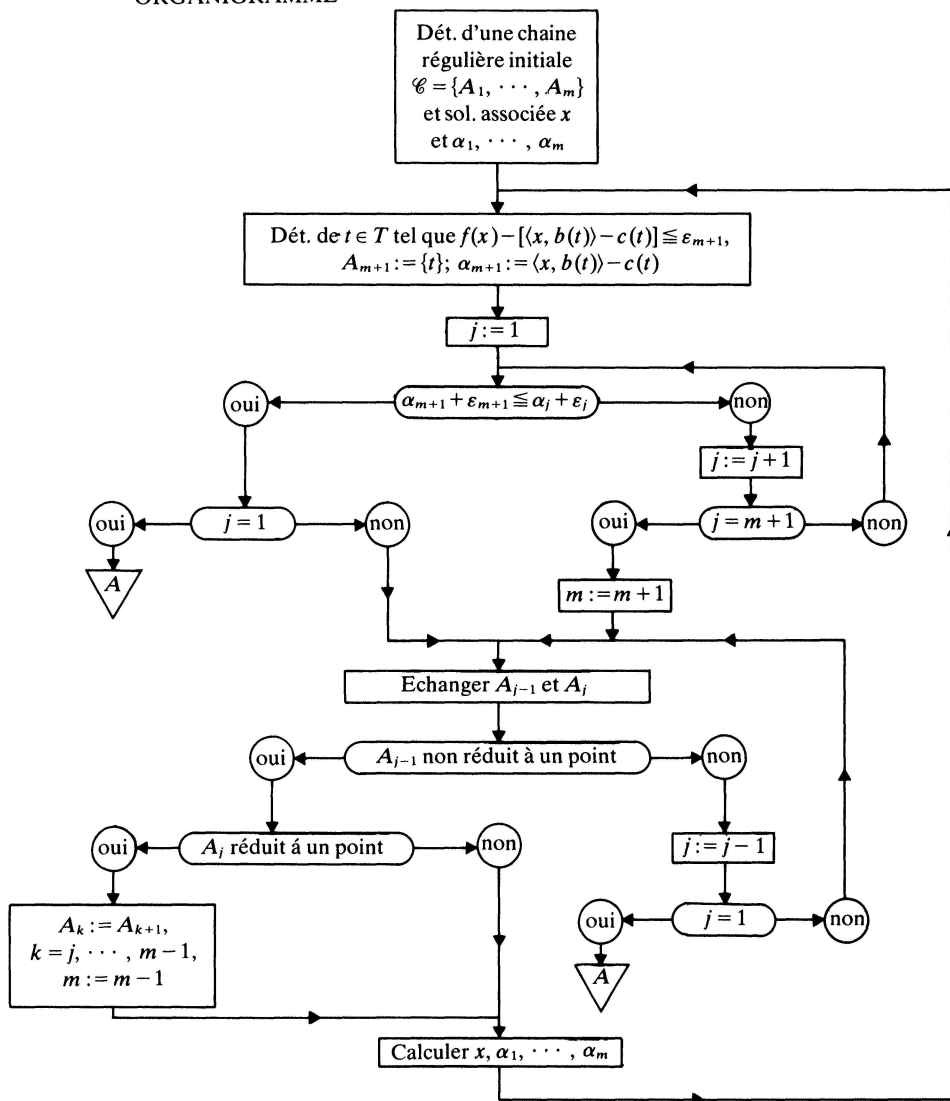
Partant de la chaîne:

$$\{A_1^\nu, A_2^\nu, \dots, A_{m^\nu}^\nu, \{t^\nu\}\},$$

on échange $A_{m^\nu}^\nu$ et $\{t^\nu\}$, ce qui donne $\tilde{A}_{m^\nu}^\nu$ et $\tilde{A}_{m^\nu+1}^\nu$. Si $\tilde{A}_{m^\nu}^\nu$ est réduit à un point, cela signifie que $\tilde{A}_{m^\nu}^\nu = \{t^\nu\}$ et que $\tilde{A}_{m^\nu+1}^\nu = A_{m^\nu}^\nu$, c'est-à-dire que l'on a la chaîne:

$$\{A_1^\nu, A_2^\nu, \dots, A_{m^\nu-1}^\nu, \{t^\nu\}, A_{m^\nu}^\nu\}.$$

ORGANIGRAMME



ORGANIGRAMME

On échange alors $A_{m^v-1}^v$ et $\{t^v\}$, et ainsi de suite jusqu'à ce qu'on obtienne:

soit $\{\{t^v\}, A_1^v, \dots, A_{m^v}^v\}$ (on arrête le calcul)

soit $\{A_1^v, A_2^v, \dots, \tilde{A}_{i^v-1}^v, \tilde{A}_{i^v}^v, A_{i^v}^v, \dots, A_{m^v}^v\}$,

où $\tilde{A}_{i^v-1}^v$ contient t^v et n'est pas réduit à ce point. Si $\tilde{A}_{i^v}^v$ est réduit à un point, on le supprime comme plus haut.

5.2. Propriétés de la suite de chaînes générées. Posons:

$$k^\nu = \begin{cases} j^\nu & \text{si } 1 \leq j^\nu \leq m^\nu, \\ i^\nu & \text{si } j^\nu = m^\nu + 1. \end{cases}$$

Lorsque $k^\nu = 1$, ce qui peut se produire, quand $j^\nu = 1$, ou quand $i^\nu = 1$ (j^ν étant alors égal à $m^\nu + 1$), alors on arrête le calcul, la précision ε_1 étant atteinte. Lorsque le calcul ne s'arrête pas à l'itération ν , on a les propriétés suivantes:

THEOREME. *Lorsque $k^\nu \geq 2$, on a:*

- (i) $A_k^{\nu+1} = A_k^\nu$ et $\alpha_k^{\nu+1} = \alpha_k^\nu$, pour $k = 1, \dots, k^\nu - 2$.
- (ii) Pour tout $t \in A_{k^{\nu-1}}^{\nu+1} \setminus A_{k^{\nu-1}}^\nu = \hat{A}^\nu$,

$$f(x^\nu) - [\langle x^\nu, b(t) \rangle - c(t)] \leq \varepsilon_{k^\nu}.$$

$$(iib) \alpha_{k^{\nu-1}}^{\nu+1} \geq \alpha_{k^{\nu-1}}^\nu + \sum_{t \in \hat{A}^\nu} \rho_{k^{\nu-1}}^{\nu+1}(t)(\varepsilon_{k^{\nu-1}} - \varepsilon_{k^\nu}).$$

Démonstration. Lorsque $k^\nu \geq 2$, l'itération concerne les niveaux $k^\nu - 1$ et k^ν de la chaîne: On échange $A_{k^{\nu-1}}^\nu$ avec, soit $A_{k^\nu}^\nu$ (cas où $1 \leq j^\nu \leq m^\nu$), soit $\{t^\nu\}$ (cas où $j^\nu = m^\nu + 1$). Les annulateurs $A_1^\nu, \dots, A_{k^{\nu-2}}^\nu$ sont donc inchangés et par conséquent les montants correspondants $\alpha_1^\nu, \dots, \alpha_{k^{\nu-2}}^\nu$ sont également inchangés.

Considérons d'abord le cas où $2 \leq j^\nu \leq m^\nu$ ($k^\nu = j^\nu$): On échange $A_{k^{\nu-1}}^\nu$ et $A_{k^\nu}^\nu$. On a donc:

$$A_{k^{\nu-1}}^{\nu+1} \setminus A_{k^{\nu-1}}^\nu = A_{k^\nu}^\nu$$

et pour tout $t \in A_{k^\nu}^\nu$, on a:

$$\langle x^\nu, b(t) \rangle - c(t) = \alpha_{k^\nu}^\nu.$$

D'après le choix même de j^ν (et comme $k^\nu = j^\nu$), on a:

$$\alpha_{k^\nu}^\nu + \varepsilon_{k^\nu} \geq \alpha_{m^{\nu+1}}^\nu + \varepsilon_{m^{\nu+1}}$$

et on sait que

$$\alpha_{m^{\nu+1}}^\nu + \varepsilon_{m^{\nu+1}} \geq f(x^\nu).$$

Il en résulte que pour tout $t \in A_{k^\nu}^\nu$, on a:

$$f(x^\nu) - [\langle x^\nu, b(t) \rangle - c(t)] \leq \varepsilon_{k^\nu}.$$

Formons:

$$\alpha_{k^{\nu-1}}^{\nu+1} = \sum_{t \in A_{k^{\nu-1}}^{\nu+1}} \rho_{k^{\nu-1}}^{\nu+1}(t) [\langle x^\nu, b(t) \rangle - c(t)].$$

Pour $t \in A_{k^{\nu-1}}^{\nu+1} \setminus A_{k^{\nu-1}}^\nu$, on a $\langle x^\nu, b(t) \rangle - c(t) = \alpha_{k^\nu}^\nu$. Pour $t \in A_{k^{\nu-1}}^{\nu+1} \cap A_{k^{\nu-1}}^\nu$, on a $\langle x^\nu, b(t) \rangle - c(t) = \alpha_{k^{\nu-1}}^\nu$. On en déduit:

$$\alpha_{k^{\nu-1}}^{\nu+1} = \alpha_{k^{\nu-1}}^\nu + \sum_{t \in A_{k^{\nu-1}}^{\nu+1}} \rho_{k^{\nu-1}}^{\nu+1}(t) (\alpha_{k^{\nu-1}}^\nu - \alpha_{k^\nu}^\nu).$$

Or d'après le choix de j^ν ($= k^\nu$), on a:

$$\alpha_{k^\nu}^\nu \geq \alpha_{m^{\nu+1}}^\nu + \varepsilon_{m^{\nu+1}} - \varepsilon_{k^\nu}$$

et

$$\alpha_{k^\nu-1}^\nu < \varepsilon_{m^\nu+1} + \varepsilon_{k^\nu+1} - \varepsilon_{k^\nu-1},$$

ce qui donne par différence:

$$\alpha_{k^\nu}^\nu - \alpha_{k^\nu-1}^\nu > \varepsilon_{k^\nu-1} - \varepsilon_{k^\nu},$$

d'où l'on déduit:

$$\alpha_{k^\nu-1}^{\nu+1} = \alpha_{k^\nu}^\nu + \left(\sum_{t \in A_{k^\nu}^{\nu+1}} \rho_{k^\nu-1}^{\nu+1}(t) \right) (\varepsilon_{k^\nu-1} - \varepsilon_{k^\nu}).$$

Considérons maintenant le cas où $j^\nu = m^\nu + 1$ ($k^\nu = i^\nu$): On échange $A_{k^\nu-1}^\nu$ et $\{t^\nu\}$. On a donc $A_{k^\nu-1}^{\nu+1} \setminus A_{k^\nu-1}^\nu = \{t^\nu\}$. Or on sait que l'on a:

$$\langle x^\nu, b(t^\nu) \rangle - c(t^\nu) = \alpha_{m^\nu+1}^\nu \geq f(x^\nu) - \varepsilon_{m^\nu+1} \geq f(x^\nu) - \varepsilon_{k^\nu}.$$

Pour $t \in A_{k^\nu-1}^{\nu+1} \setminus A_{k^\nu-1}^\nu = \{t^\nu\}$, on a $\langle x^\nu, b(t) \rangle - c(t) = \alpha_{m^\nu+1}^\nu$. Pour $t \in A_{k^\nu-1}^{\nu+1} \cap A_{k^\nu-1}^\nu$, on a $\langle x^\nu, b(t) \rangle - c(t) = \alpha_{k^\nu-1}^\nu$.

De la même façon que précédemment, on en déduit que

$$\alpha_{k^\nu-1}^{\nu+1} = \alpha_{k^\nu-1}^\nu + \rho_{k^\nu-1}^{\nu+1}(t^\nu)(\alpha_{m^\nu+1}^\nu - \alpha_{k^\nu-1}^\nu).$$

Comme $k^\nu - 1 \leq m^\nu$ et comme $j^\nu = m^\nu + 1$, on a $\alpha_{m^\nu+1}^\nu + \varepsilon_{m^\nu+1} > \alpha_{k^\nu-1}^\nu + \varepsilon_{k^\nu-1}$, ce qui entraîne $\alpha_{m^\nu+1}^\nu - \alpha_{k^\nu-1}^\nu > (\varepsilon_{k^\nu-1} - \varepsilon_{m^\nu+1}) > (\varepsilon_{k^\nu-1} - \varepsilon_{k^\nu})$, d'où le résultat.

6. Exemple numérique. L'exemple numérique suivant est seulement destiné à illustrer le mécanisme de l'algorithme. On considère la minimisation sur l'espace \mathbb{R}^2 de la fonction:

$$f(x) = \max_{i=0\text{à}8} (\langle x, b(i) \rangle - c(i))$$

dans laquelle b et c sont définis par le tableau 1.

TABLEAU 1

$i =$	0	1	2	3	4	5	6	7	8
$b_1(i) =$	4	$-4\sqrt{2}$	0	3	-6	-2	0	-0.5	0.5
$b_2(i) =$	4	0	$-4\sqrt{2}$	3	-6	0	-2	-0.5	0.5
$c(i) =$	0	0	0	-7	-1	-5.75	-5.75	-6	-6.3

En partant de l'annulateur minimal $A_1^0 = \{0, 1, 2\}$, on obtient la solution exacte à l'itération 7. Nous donnons la suite des systèmes linéaires qu l'on est amené à résoudre, les niveaux α_i correspondants et le point t^ν que l'on introduit éventuellement. L'évolution des chaines d'annulateurs minimaux est donnée dans la figure ci-dessous.

- *itération 0*: $m^0 = 1$

$$\begin{bmatrix} b_1(0) & b_2(0) & -1 \\ b_1(1) & b_2(1) & -1 \\ b_1(2) & b_2(2) & -1 \end{bmatrix} \begin{bmatrix} x_1^0 \\ x_2^0 \\ \alpha_1^0 \end{bmatrix} = \begin{bmatrix} c(0) \\ c(1) \\ c(2) \end{bmatrix} \quad \text{donne } \alpha_1^0 = 0,$$

$t^0 = 3$; $\alpha_2^0 = 7$. Le point t^0 est échangé avec 0 (voir la figure 4).

- *itération 1*: $m^1 = 1$

$$\begin{bmatrix} b_1(3) & b_2(3) & -1 \\ b_1(1) & b_2(1) & -1 \\ b_1(2) & b_2(2) & -1 \end{bmatrix} \begin{bmatrix} x_1^1 \\ x_2^1 \\ \alpha_1^1 \end{bmatrix} = \begin{bmatrix} c(3) \\ c(1) \\ c(2) \end{bmatrix} \quad \text{donne } \alpha_1^1 = 3.39695,$$

$t^1 = 4$; $\alpha_2^1 = 8.20588$. Une dégénérescence se produit. Le point t^1 est échangé avec $\{1, 2\}$, ce qui conduit à $A_1^1 = \{3, 4\}$ et $A_2^1 = \{1, 2\}$.

- *itération 2*: $m^2 = 2$

$$\begin{bmatrix} b_1(3) & b_2(3) & -1 & 0 \\ b_1(4) & b_2(4) & -1 & 0 \\ b_1(1) & b_2(1) & 0 & -1 \\ b_1(2) & b_2(2) & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1^2 \\ x_2^2 \\ \alpha_1^2 \\ \alpha_2^2 \end{bmatrix} = \begin{bmatrix} c(3) \\ c(4) \\ c(1) \\ c(2) \end{bmatrix} \quad \text{donne } \begin{cases} \alpha_1^2 = 5, \\ \alpha_2^2 = 1.88564, \end{cases}$$

$t^2 = 5$; $\alpha_3^2 = 6.41673$. Le point t^2 est échangé avec 1 dans A_2^2 , ce qui conduit à $A_1^2 = \{3, 4\}$, inchangé et $A_2^2 = \{5, 2\}$.

- *itération 3*: $m^3 = 2$

$$\begin{bmatrix} b_1(3) & b_2(3) & -1 & 0 \\ b_1(4) & b_2(4) & -1 & 0 \\ b_1(5) & b_2(5) & 0 & -1 \\ b_1(2) & b_2(2) & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1^3 \\ x_2^3 \\ \alpha_1^3 \\ \alpha_2^3 \end{bmatrix} = \begin{bmatrix} c(3) \\ c(4) \\ c(5) \\ c(2) \end{bmatrix} \quad \text{donne } \begin{cases} \alpha_1^3 = 5, \\ \alpha_2^3 = 5.25215, \end{cases}$$

$t^3 = 6$; $\alpha_3^3 = 7.60607$. Le point t^3 est échangé avec 2 dans A_2^3 , ce qui conduit à $A_1^3 = \{3, 4\}$ inchangé et $A_2^3 = \{5, 6\}$.

- *itération 4*: $m^4 = 2$

$$\begin{bmatrix} b_1(3) & b_2(3) & -1 & 0 \\ b_1(4) & b_2(4) & -1 & 0 \\ b_1(5) & b_2(5) & 0 & -1 \\ b_1(6) & b_2(6) & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1^4 \\ x_2^4 \\ \alpha_1^4 \\ \alpha_2^4 \end{bmatrix} = \begin{bmatrix} c(3) \\ c(4) \\ c(5) \\ c(6) \end{bmatrix} \quad \text{donne } \begin{cases} \alpha_1^4 = 5, \\ \alpha_2^4 = 6.41672, \end{cases}$$

$t^4 = 5$; $\alpha_3^4 = 6.41672$. Le point t^4 n'est pas introduit (il était d'ailleurs déjà présent dans A_2^4 , ce qui signifie qu'on atteint la solution exacte correspondant au niveau 2). On échange A_2^4 avec le point 4 de A_1^4 , ce qui conduit à $A_1^4 = \{3, 5, 6\}$. La dégénérescence est résorbée.

- *itération 5*: $m^5 = 1$

$$\begin{bmatrix} b_1(3) & b_2(3) & -1 \\ b_1(5) & b_2(5) & -1 \\ b_1(6) & b_2(6) & -1 \end{bmatrix} \begin{bmatrix} x_1^5 \\ x_2^5 \\ \alpha_1^5 \end{bmatrix} = \begin{bmatrix} c(3) \\ c(5) \\ c(6) \end{bmatrix} \quad \text{donne } \alpha_1^5 = 6.06246,$$

$t^5 = 7$; $\alpha_2^5 = 6.12$. Une nouvelle dégénérescence se produit. Le point t^5 est échangé avec $\{5, 6\}$, ce qui conduit à $A_1^5 = \{3, 7\}$ et $A_2^5 = \{5, 6\}$.

• *itération 6*: $m^6 = 2$

$$\begin{bmatrix} b_1(3) & b_2(3) & -1 & 0 \\ b_1(7) & b_2(7) & -1 & 0 \\ b_1(5) & b_2(5) & 0 & -1 \\ b_1(6) & b_2(6) & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1^6 \\ x_2^6 \\ \alpha_1^6 \\ \alpha_2^6 \end{bmatrix} = \begin{bmatrix} c(3) \\ c(7) \\ c(5) \\ c(6) \end{bmatrix} \quad \text{donne} \begin{cases} \alpha_1^6 = 6.14283, \\ \alpha_2^6 = 6.03567, \end{cases}$$

$i^6 = 8$; $\alpha_3^6 = 6.15711$. Le point i^6 est échangé directement avec le point 3 de A_1^6 (cf. troisième cas du § 5, avec $i^6 = 1$). On obtient donc $A_1^7 = \{8, 7\}$ et $A_2^7 = \{5, 6\}$ inchangé.

• *itération 7*: $m^7 = 2$

$$\begin{bmatrix} b_1(8) & b_2(8) & -1 & 0 \\ b_1(7) & b_2(7) & -1 & 0 \\ b_1(5) & b_2(5) & 0 & -1 \\ b_1(6) & b_2(6) & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1^7 \\ x_2^7 \\ \alpha_1^7 \\ \alpha_2^7 \end{bmatrix} = \begin{bmatrix} c(8) \\ c(7) \\ c(5) \\ c(6) \end{bmatrix} \quad \text{donne} \begin{cases} \alpha_1^7 = 6.15, \\ \alpha_2^7 = 6.05, \end{cases}$$

$i^7 = 8$; $\alpha_3^7 = 6.15$. On a atteint une solution $x_1^7 = -0.15$, $x_2^7 = -0.15$.

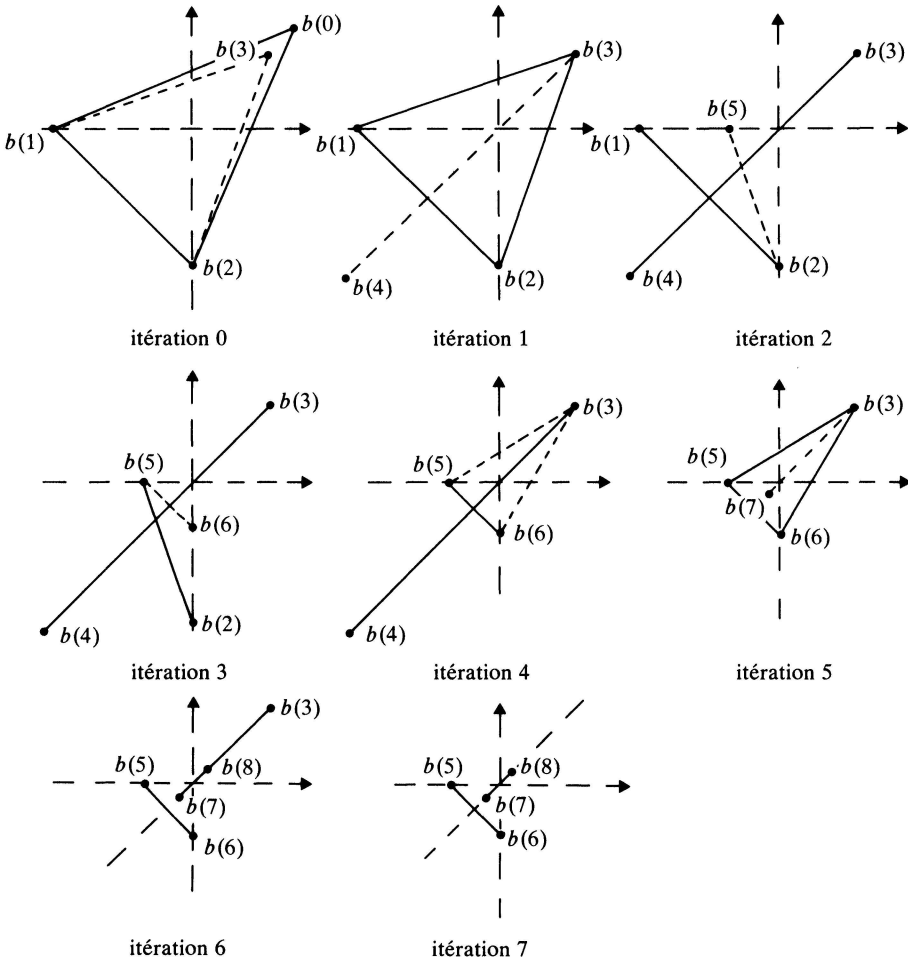


FIG. 4

7. Convergence. Avant de démontrer la convergence de l'algorithme, nous allons établir un théorème de convergence préparatoire dont l'énoncé est en liaison évidente avec les propriétés établies au § 5.2. Les principales difficultés pour montrer la convergence de l'algorithme sont en fait contenues dans ce théorème préparatoire.

En suivant les résultats du § 2, nous utiliserons les notations suivantes:

\mathcal{V} un sous-espace vectoriel de E contenu dans V ,

\mathcal{W} une variété affine parallèle à \mathcal{V} ,

$\{A^\nu\}$ une suite d'annulateurs minimaux de \mathcal{V} ,

$\{f^\nu\}$ la suite des fonctionnelles polyédrales associées à A^ν ,

$$\alpha^\nu = \min_{x \in \mathcal{W}} f^\nu(x),$$

$$\mathcal{W}^\nu = \{x \in \mathcal{W} \mid f^\nu(x) = \alpha^\nu\}.$$

On a alors le résultat suivant:

7.1. THEOREME. Soient ε et $\hat{\varepsilon}$ des nombres strictement positifs tels que $\hat{\varepsilon} < \varepsilon$. Notons $\hat{N} = \{\nu \in N \mid A^{\nu+1} \neq A^\nu\}$.

On suppose que \hat{N} est infini et que pour tout $\nu \in \hat{N}$ on a un élément $x^\nu \in \mathcal{W}^\nu$ tel que pour tout $t \in A^{\nu+1} \setminus A^\nu = \hat{A}^\nu$ on ait:

$$f(x^\nu) - [\langle x^\nu, b(t) \rangle - c(t)] \leq \hat{\varepsilon}.$$

Alors il existe $\mu \in \hat{N}$ tel que

$$f(x^\mu) - \alpha^\mu \leq \varepsilon.$$

Démonstration. Supposons que l'on ait $f(x^\nu) - \alpha^\nu > \varepsilon$ pour tout $\nu \in \hat{N}$ et montrons que cela conduit à une contradiction. Nous diviserons la démonstration en trois parties.

1^{ère} partie: Montrons d'abord que α^ν est alors une suite non décroissante, plus précisément que pour tout $\nu \in \hat{N}$, on a $\alpha^{\nu+1} > \alpha^\nu$ (pour $\nu \notin \hat{N}$, on a évidemment $\alpha^{\nu+1} = \alpha^\nu$).

Pour $\nu \in \hat{N}$, on a:

$$\alpha^{\nu+1} = \sum_{t \in A^{\nu+1}} \rho^{\nu+1}(t) (\langle x^\nu, b(t) \rangle - c(t)),$$

où $\rho^{\nu+1}(t)$, $t \in A^{\nu+1}$ sont les coefficients associés à $A^{\nu+1}$. Pour $t \in \hat{A}^\nu$, on a $\langle x^\nu, b(t) \rangle - c(t) \geq f(x^\nu) - \hat{\varepsilon}$. Pour $t \in A^{\nu+1} \cap A^\nu$, on a $\langle x^\nu, b(t) \rangle - c(t) = \alpha^\nu$. On en déduit que

$$\alpha^{\nu+1} \geq \sum_{t \in A^{\nu+1} \cap A^\nu} \rho^{\nu+1}(t) \alpha^\nu + \sum_{t \in \hat{A}^\nu} \rho^{\nu+1}(t) (f(x^\nu) - \hat{\varepsilon}),$$

donc

$$\alpha^{\nu+1} \geq \alpha^\nu + \left(\sum_{t \in \hat{A}^\nu} \rho^{\nu+1}(t) \right) (f(x^\nu) - \hat{\varepsilon} - \alpha^\nu).$$

Comme on a supposé $f(x^\nu) - \alpha^\nu > \varepsilon$, pour $\nu \in \hat{N}$, cela conduit à

$$\alpha^{\nu+1} \geq \alpha^\nu + \left(\sum_{t \in \hat{A}^\nu} \rho^{\nu+1}(t) \right) (\varepsilon - \hat{\varepsilon}),$$

ce qui montre bien que l'on a $\alpha^{\nu+1} > \alpha^\nu$ pour tout $\nu \in \hat{N}$.

2^{ème} partie: Montrons maintenant qu'il existe un sous-ensemble infini \tilde{N} de \hat{N} et une bipartition de A^ν en B^ν et $C^\nu \neq \emptyset$ telle que:

$$1^\circ. \lim_{\nu \in \tilde{N}} \left(\sum_{t \in B^\nu} \rho^\nu(t) \right) = 0.$$

$$2^\circ. \inf_{\nu \in \tilde{N}} \min_{t \in C^\nu} \rho^\nu(t) > 0.$$

3°. Pour tout $\nu \in \tilde{N}$, l'ensemble C^ν n'est pas contenu dans $A^{\nu-}$, ou ν désigne l'entier qui précède ν dans \tilde{N} .

Pour simplifier la démonstration, nous nous ramenons à une sous-suite telle que les annulateurs minimaux comportent le même nombre d'éléments. Soit donc N_0 un sous-ensemble infini de \hat{N} tel que pour tout $\nu \in N_0$, les annulateurs A^ν aient le même nombre d'éléments.

Si $\inf_{\nu \in N_0} \min_{t \in A^\nu} \rho^\nu(t) > 0$, alors on pose $B^\nu = \emptyset$ et $C^\nu = A^\nu$.

Sinon, il existe un sous-ensemble infini N_1 contenu dans N_0 tel que

$$\lim_{\nu \in N_1} \min_{t \in A^\nu} \rho^\nu(t) = 0.$$

On pose alors $B_1^\nu = \{t_1^\nu\}$, $\nu \in N_1$, où $t_1^\nu \in A^\nu$ vérifie

$$\rho^\nu(t_1^\nu) = \min_{t \in A^\nu} \rho^\nu(t)$$

et $C_1^\nu = A^\nu \setminus B_1^\nu$, $\nu \in N_1$.

Si l'on a à nouveau $\inf_{\nu \in N_1} \min_{t \in C_1^\nu} \rho^\nu(t) > 0$, alors il existe un sous-ensemble infini N_2 de N_1 tel que

$$\lim_{\nu \in N_2} \min_{t \in C_1^\nu} \rho^\nu(t) = 0.$$

On pose encore $B_2^\nu = B_1^\nu \cup \{t_2^\nu\}$, $\nu \in N_2$, où $t_2^\nu \in C_1^\nu$ vérifie

$$\rho^\nu(t_2^\nu) = \min_{t \in C_1^\nu} \rho^\nu(t)$$

et $C_2^\nu = A^\nu \setminus B_2^\nu$.

Comme $\sum_{t \in A^\nu} \rho^\nu(t) = 1$, après un nombre fini r de telles opérations ($r \leq \hat{n} - 1$, où \hat{n} est le nombre d'éléments des ensembles A^ν pour $\nu \in N_0$), on arrive à un sous-ensemble infini N_r de N_0 et à une bipartition de A^ν en B_r^ν et $C_r^\nu \neq \emptyset$ pour $\nu \in N_r$ telle que:

$$1^\circ. \lim_{\nu \in N_r} \sum_{t \in B_r^\nu} \rho^\nu(t) = 0.$$

$$2^\circ. \inf_{\nu \in N_r} \min_{t \in C_r^\nu} \rho^\nu(t) > 0.$$

Montrons que l'on peut extraire à nouveau une sous-suite de façon à satisfaire la condition 3°. Il suffit pour cela de montrer qu'il n'existe pas $\nu \in N_r$ tel que pour tout $\mu > \nu$, $\mu \in N_r$, on ait $C_r^\mu \subset A^\nu$. En effet, si un tel ν existait, comme A^ν contient un nombre fini d'éléments, il existerait un sous-ensemble C de A^ν et un sous-ensemble infini N_{r+1} de N_r tel que $C_r^\mu = C$, pour tout $\mu \in N_{r+1}$. D'autre part les coefficients $\rho^\nu(t)$, $t \in C$, $\nu \in N_{r+1}$ parcourent un compact. On pourrait donc trouver un sous-ensemble infini N_{r+2} de N_{r+1} tel que

$$\lim_{\nu \in N_{r+2}} \rho^\nu(t) = \rho(t) \quad \text{pour } t \in C.$$

Comme on a par ailleurs:

$$\lim_{\nu \in N_{r+2}} \sum_{t \in B_r^\nu} \rho^\nu(t) = 0$$

et comme $b(t)$ est borné pour $t \in T$, on aurait:

$$\lim_{\nu \in N_{r+2}} \sum_{t \in B_r^\nu} \rho^\nu(t) = 0.$$

On en déduirait donc que pour tout $x \in \mathcal{V}$:

$$\begin{aligned} \langle x, \sum_{t \in C} \rho(t)b(t) \rangle &= \lim_{\nu \in N_{r+2}} \langle x, \sum_{t \in C} \rho^\nu(t)b(t) \rangle \\ &= \lim_{\nu \in N_{r+2}} \langle x, \sum_{t \in C} \rho^\nu(t)b(t) + \sum_{t \in B_r^\nu} \rho^\nu(t)b(t) \rangle \\ &= \lim_{\nu \in N_{r+2}} \langle x, \sum_{t \in A^\nu} \rho^\nu(t)b(t) \rangle = 0 \end{aligned}$$

avec $\sum_{t \in C} \rho(t) = 1$. Par conséquent C serait un annulateur de \mathcal{V} contenu dans l'annulateur A^ν , ce qui entraîne $C = A^\nu$. Donc pour tout $\mu > \nu$, $\mu \in N_{r+2}$, $C_r^\mu = C$ serait un annulateur minimal de \mathcal{V} et comme C_r^μ est évidemment contenu dans l'annulateur minimal A^μ , on en déduirait $C_r^\mu = A^\mu$ donc $A^\nu = A^\mu$ pour tout $\mu > \nu$, $\mu \in N_{r+2}$, ce qui contredit la condition $A^{\nu+1} \neq A^\nu$, pour tout $\nu \in \hat{N}$.

3^{ème} partie: D'après le résultat démontré dans la deuxième partie, pour tout $\nu \in \tilde{N}$, l'ensemble C^ν n'est pas contenu dans $A^{\nu-}$, où $\nu-$ est l'entier qui précède ν dans \tilde{N} . Il existe donc des éléments de C^ν qui ont été introduits entre l'itération $\nu-$ et l'itération ν . Soit $\hat{\nu}$ la dernière itération vérifiant $\nu- \leq \hat{\nu} < \nu$ au cours de laquelle des éléments de C^ν ont été introduits. On remarque que $\hat{\nu}$ appartient forcément à \hat{N} . On forme:

$$\alpha^\nu = \sum_{t \in A^\nu} \rho^\nu(t) [\langle x^{\hat{\nu}}, b(t) \rangle - c(t)].$$

Décomposons la somme précédente suivant que t appartient à B^ν ou à C^ν :

(α) *Somme relative à B^ν :* On a:

$$\left| \sum_{t \in B^\nu} \rho^\nu(t) [\langle x^{\hat{\nu}}, b(t) \rangle - c(t)] \right| \leq \left(\sum_{t \in B^\nu} \rho^\nu(t) \right) (\gamma f(x^{\hat{\nu}}) + \delta).$$

Si l'on pose $\sum_{t \in B^\nu} \rho^\nu(t) = \eta^\nu$, on a $\lim_{\nu \in \tilde{N}} \eta^\nu = 0$.

(β) *Somme relative à C^ν* : D'après l'hypothèse du théorème, pour $t \in C^\nu \cap \hat{A}^{\hat{\nu}}$ (on sait que cet ensemble est non vide) on a:

$$\langle x^{\hat{\nu}}, b(t) \rangle - c(t) \geq f(x^{\hat{\nu}}) - \hat{\varepsilon}.$$

Pour $t \in C^\nu \setminus \hat{A}^{\hat{\nu}}$, on a:

$$\langle x^{\hat{\nu}}, b(t) \rangle - c(t) = \alpha^{\hat{\nu}}.$$

En posant $u^\nu = \sum_{t \in C^\nu} \rho^\nu(t)$ et $m = \inf_{\nu \in \tilde{N}} \min_{t \in C^\nu} \rho^\nu(t)$, d'après la deuxième partie de la démonstration, on a $\lim_{\nu \in \tilde{N}} u^\nu = 1$ et $m > 0$.

Avec ces notations on a:

$$\sum_{t \in C^\nu} \rho^\nu(t) [\langle x^{\hat{\nu}}, b(t) \rangle - c(t)] \geq u^\nu \alpha^{\hat{\nu}} + m(f(x^{\hat{\nu}}) - \hat{\varepsilon} - \alpha^{\hat{\nu}}).$$

En rassemblant les résultats de (α) et (β) ci-dessus, on obtient:

$$\alpha^\nu \geq u^\nu \alpha^{\hat{\nu}} + m(f(x^{\hat{\nu}}) - \hat{\varepsilon} - \alpha^{\hat{\nu}}) - \eta^\nu(\gamma f(x^{\hat{\nu}}) + \delta).$$

En posant $v^\nu = 1 - \eta^\nu \gamma / m$, on a:

$$m(v^\nu f(x^{\hat{\nu}}) - \alpha^{\hat{\nu}} - \hat{\varepsilon}) \leq \alpha^\nu - u^\nu \alpha^{\hat{\nu}} + \eta^\nu \delta.$$

En posant $b = \inf_{x \in \mathcal{W}} f(x)$, on a $\alpha^\nu \leq b$; on en déduit que $\lim_{\nu \in \tilde{N}} \alpha^\nu = \tilde{\alpha} \leq b$. Comme on a par ailleurs $\lim_{\nu \in \tilde{N}} u^\nu = 1$ et $\lim_{\nu \in \tilde{N}} \eta^\nu = 0$, on en déduit:

$$\limsup_{\nu \in \tilde{N}} (v^\nu f(x^{\hat{\nu}}) - \alpha^{\hat{\nu}} - \hat{\varepsilon}) \leq 0,$$

et par conséquent:

$$\limsup_{\nu \in \tilde{N}} v^\nu f(x^{\hat{\nu}}) \leq \tilde{\alpha} + \hat{\varepsilon}.$$

Si $\hat{\varepsilon}'$ désigne un réel strictement supérieur à $\hat{\varepsilon}$, à partir d'un certain rang ν_0 , on aura:

$$(*) \quad v^\nu > 0 \quad \text{et} \quad v^\nu f(x^{\hat{\nu}}) \leq \tilde{\alpha} + \hat{\varepsilon}', \quad \text{pour} \quad \nu \in \tilde{N}, \nu \geq \nu_0.$$

Par ailleurs, on a supposé que $f(x^{\hat{\nu}}) > \alpha^{\hat{\nu}} + \varepsilon$, pour $\nu \in \hat{N}$; donc si ε' désigne un réel strictement inférieur à ε , à partir d'un certain rang ν_1 , on aura:

$$(**) \quad f(x^{\hat{\nu}}) \geq \tilde{\alpha} + \varepsilon', \quad \text{pour} \quad \nu \in \tilde{N}, \nu \geq \nu_1.$$

Puisque l'on a $\hat{\varepsilon} < \varepsilon$, il est toujours possible de prendre $\hat{\varepsilon}'$ et ε' tels que $\hat{\varepsilon}' < \varepsilon'$. On voit alors que les inégalités (*) et (**) sont contradictoires: en effet, l'inégalité (**) entraîne (pour $\nu \geq \max(\nu_0, \nu_1)$) $v^\nu f(x^{\hat{\nu}}) \geq v^\nu(\tilde{\alpha} + \varepsilon')$ et comme $\lim_{\nu \in \tilde{N}} v^\nu = 1$, à partir d'un certain rang on a $v^\nu(\tilde{\alpha} + \varepsilon') > \tilde{\alpha} + \hat{\varepsilon}'$, donc $v^\nu f(x^{\hat{\nu}}) > \tilde{\alpha} + \hat{\varepsilon}'$, ce qui contredit bien (*).

7.2. Convergence de l'algorithme. On a le résultat suivant:

THEOREME. *Quel que soit ε_1 , l'algorithme décrit au § 5 conduit, après un nombre fini μ d'itérations, à un élément $x^\mu \in W$ qui est une solution à ε_1 près de (P).*

Plus précisément, on obtient un annulateur A_1^μ de V et un élément $x^\mu \in W$ tels que

$$f(x^\mu) - \alpha_1^\mu \leq \varepsilon_1$$

avec $\alpha_1^\mu = \sum_{t \in A_1^\mu} \rho_1^\mu(t) (\langle x^\mu, b(t) \rangle - c(t)) \leq \alpha$, ce qui entraîne $f(x^\mu) - \alpha \leq \varepsilon_1$.

Démonstration. Il faut montrer qu'il existe un entier μ tel que l'on ait $k^\mu = 1$, ce qui entraîne l'arrêt de l'algorithme, la précision ε_1 étant atteinte.

Supposons donc que l'on ait $k^\nu \geq 2$, pour tout $\nu \in N$ et montrons que cela conduit à une contradiction.

Comme $k^\nu \in [2, m^\nu + 1] \subset [2, n_0 + 1]$, l'indice k^ν prend une infinité de fois certaines des valeurs entières entre 2 et $n_0 + 1$. Soit \tilde{k} la plus petite de ces valeurs, $2 \leq \tilde{k} \leq n_0 + 1$. Il existe donc ν_0 tel que pour tout $\nu \geq \nu_0$, on ait $k^\nu \geq \tilde{k}$ et l'ensemble

$$\tilde{N} = \{\nu \mid \nu \geq \nu_0; k^\nu = \tilde{k}\}$$

est infini.

Pour tout $\nu \geq \nu_0$, on a $k^\nu \geq \tilde{k}$, donc $A_k^\nu = A_k$ et $V_k^\nu = V_k$ (indépendant de ν), $k = 1, \dots, \tilde{k} - 2$.

Si l'on pose $V_{\tilde{k}-2} = \mathcal{V}$, $\{A_{\tilde{k}-1}^\nu\}$ forme une suite d'annulateurs minimaux de \mathcal{V} et pour tout $\nu \in \tilde{N}$ et pour $t \in A_{\tilde{k}-1}^{\nu+1} \setminus A_{\tilde{k}-1}^\nu$, on a (cf. § 5.2):

$$f(x^\nu) - [\langle x^\nu, b(t) \rangle - c(t)] \leq \varepsilon_{\tilde{k}}.$$

D'après le théorème 6.1, quel que soit le nombre $\varepsilon > 0$ fixé, vérifiant $\varepsilon > \varepsilon_{\tilde{k}}$, il existe μ tel que

$$f(x^\mu) - \alpha_{\tilde{k}-1}^\mu \leq \varepsilon,$$

ce qui donne, en tenant compte du fait que $\alpha_{m^\mu+1}^\mu \leq f(x^\mu)$:

$$(i) \quad \alpha_{m^\mu+1}^\mu - \alpha_{\tilde{k}-1}^\mu \leq \varepsilon.$$

Or on a, d'après la définition de l'algorithme:

$$\alpha_{m^\mu+1}^\mu + \varepsilon_{m^\mu+1} > \alpha_{\tilde{k}-1}^\mu + \varepsilon_{\tilde{k}-1}.$$

Comme $\tilde{k} \leq m^\mu + 1$, on a $\varepsilon_{\tilde{k}} \geq \varepsilon_{m^\mu+1}$, et l'inégalité précédente entraîne:

$$(ii) \quad \alpha_{m^\mu+1}^\mu - \alpha_{\tilde{k}-1}^\mu > \varepsilon_{\tilde{k}-1} - \varepsilon_{\tilde{k}}.$$

Comme on a supposé (la condition (*), § 5) que $\varepsilon_{\tilde{k}} < \varepsilon_{\tilde{k}-1}/2$, il est possible de choisir ε tel que

$$\varepsilon_{\tilde{k}} < \varepsilon < \varepsilon_{\tilde{k}-1} - \varepsilon_{\tilde{k}},$$

et les inégalités (i) et (ii) deviennent alors contradictoires.

7.3. Remarque. On a présenté l'algorithme de sorte que, pour $\varepsilon_1 > 0$ fixé (arbitrairement petit), on aboutit, après un nombre fini d'itérations, à un élément $x^\mu \in W$ qui est une solution à ε_1 près de (P). Cette présentation correspond bien à l'emploi effectif de l'algorithme. Il est bien clair toutefois que l'utilisation répétée de cet algorithme fournit une suite (infinie) $x^\nu \in W$ telle que $\lim_{\nu \rightarrow \infty} f(x^\nu) = \alpha$ et une suite d'annulateurs minimaux A_1^ν de V telle que la suite α_1^ν des montants associés vérifie $\lim_{\nu \rightarrow \infty} \alpha_1^\nu = \alpha$. Pour cela on se donne une séquence de nombres

positifs $\varepsilon_1, \dots, \varepsilon_{n_0+1}$ vérifiant la condition (*), § 5. En utilisant l'algorithme décrit, on obtient une séquence finie de chaînes $\mathcal{C}^1, \dots, \mathcal{C}^\mu$ et de solutions associées x^1, \dots, x^μ telle que x^μ soit solution à ε_1 près du problème. En utilisant la nouvelle séquence de nombres $\varepsilon_1/2, \varepsilon_2/2, \dots, \varepsilon_{n_0+1}/2$ et en partant de \mathcal{C}^μ comme chaîne initiale, l'algorithme fournit $\mathcal{C}^{\mu+1}, \dots, \mathcal{C}^{\mu'}$ tel que $x^{\mu'}$ associé soit solution à $\varepsilon_1/2$ près. On continue de la même façon.

REFERENCES

- [1] A. AUSLENDER, *Méthodes numériques pour la décomposition et la minimisation de fonctions non-différentiables*, Numer. Math., 18 (1971), pp. 213–223.
- [2] E. M. L. BEALE, *On minimizing a convex function subject to linear inequalities*, J. Roy. Statist. Soc., B17 (1955), pp. 173–177.
- [3] D. P. BERTSEKAS AND S. MITTER, *A descent numerical method for optimization problems with nondifferentiable cost functionals*, this Journal, 11 (1973), pp. 637–652.
- [4] L. BITTNER, *Das Austauschverfahren der linearen Tschebyscheff Approximation bei nicht erfüllter Haarscher Bedingung*, Z. Angew. Math. Mech., 41 (1961), pp. 238–256.
- [5] B. BROSIOWSKI, *Über Tschebyscheffsche Approximationen mit linearen Nebenbedingungen*, Math. Zeit., 88 (1965), pp. 105–128.
- [6] G. BRUHN, *Lineare Approximation von Funktionen, die auf einer kompakten Menge stetig sind*, HMI-B38, Bericht des Hahn-Meitner Instituts Berlin, 1964.
- [7] C. CARASSO, *Etude de l'algorithme de Rémès en l'absence de conditions de Haar*, Numer. Math., 20 (1972), pp. 165–178.
- [8] ———, *Densité des hypothèses assurant la convergence de l'algorithme de Rémès*, R.A.I.R.O., R3 (1972), pp. 69–84.
- [9] ———, *Un algorithme de minimisation de fonctions convexes avec ou sans contraintes: "l'algorithme d'échange"*, Prépublication N° 3, Math., Univ. de Saint-Etienne, 7th IFIP Conference on Optimization Techniques, Springer-Verlag, Berlin, 1975.
- [10] C. CARASSO ET P. J. LAURENT, *Un algorithme pour la minimisation d'une fonctionnelle convexe sur une variété affine*, Séminaire d'analyse numérique, Grenoble, 18 octobre 1973.
- [11] ———, *Un algorithme général pour l'approximation au sens de Tchebycheff de fonctions bornées sur un ensemble quelconque*, Approximations-Kolloquium, Bonn (June 8–12, 1976), Lecture Notes in Math., no. 556, Springer-Verlag, Berlin, 1976.
- [12] E. W. CHENEY, *Introduction to Approximation Theory*, McGraw-Hill, New York, 1966.
- [13] E. W. CHENEY AND A. A. GOLDSTEIN, *Newton's method for convex programming and Tchebyscheff approximation*, Numer. Math., 1 (1959), pp. 253–268.
- [14] V. F. DEMJANOV, *Algorithms for some minimax problems*, J. Comput. System Sci., 2 (1968), pp. 342–380.
- [15] J. DESCLOUX, *Contribution au calcul des approximations de Tschebycheff*, Thèse, Eidgen. Techn. Hochschule, Zürich, 1960.
- [16] ———, *Dégénérescences dans les approximations de Tschebycheff linéaires et discrètes*, Numer. Math., 3 (1961), pp. 180–187.
- [17] A. A. GOLDSTEIN, *Constructive Real Analysis*, Series in Modern Mathematics, Harper and Row, New York, 1967.
- [18] P. J. LAURENT, *Approximation uniforme de fonctions continues sur un compact avec contraintes de type inégalité*, Rev. Franç. d'Inf. et de Rech. Opér., 5 (1967), pp. 81–95.
- [19] ———, *Théorèmes de caractérisation, d'une meilleure approximation dans un espace normé et généralisation de l'algorithme de Rémès*, Numer. Math., 10 (1967), pp. 190–208.
- [20] ———, *Charakterisierung und Gewinnung einer besten Approximation in einer Konvexen Teilmenge eines normierten Raumes*, ISNM 12, Birkhäuser Verlag, Basel, Switzerland, 1968, pp. 91–102.
- [21] ———, *Approximation et Optimisation*, Hermann, Paris, 1972.
- [22] ———, *Exchange algorithm in convex analysis*, Conf. on Approximation Theory, University of Texas, Austin, (1973), Academic Press, New York, 1974.

- [23] P. J. LAURENT AND C. CARASSO, *An algorithm of successive minimization in convex programming*, IX International Symposium on Mathematical Programming, Budapest, August, 1976 (to appear); Rapport de recherche n° 49 (Août 1976), Mathématiques Appliquées et Informatique, Université de Grenoble.
- [24] P. J. LAURENT AND PHAM DINH TUAN, *Global approximation of a compact set by elements of a convex set in a normed space*, Numer. Math., 15 (1970), pp. 137–150.
- [25] C. LEMARÉCHAL, *An extension of Davidon methods to nondifferentiable problems*, Math. Programming Study, 3 (1975), pp. 95–109.
- [26] J. J. MOREAU, *Fonctionnelles convexes*, Séminaire au Collège de France, Paris, 1966.
- [27] E. RÉMÈS, *Sur le calcul effectif des polynômes d'approximation de Tchebycheff*, C.R. Acad. Sci. Paris, 199 (1934), pp. 337–340.
- [28] ———, *General computational methods for Chebyshev approximation. Problems with real parameters entering linearly*, Izdat. Akad. Nauk. Ukrainsk. SSR, Kiev, 1957; Atomic Energy Commission Translations 4491.
- [29] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.
- [30] E. L. STIEFEL, *Über diskrete und lineare Tschebyscheff-Approximationen*, Numer. Math., 1 (1959), pp. 1–28.
- [31] ———, *Numerical methods of Tschebycheff approximation*, On Numerical Approximation, R. Langer, ed., University of Wisconsin, Madison, 1959, pp. 217–232.
- [32] ———, *Note on Jordan elimination, linear programming and Tchebycheff approximation*, Numer. Math., 2 (1960), pp. 1–17.
- [33] H. J. TÖPFER, *Über die Tschebyscheffsche Approximationsaufgabe bei nicht erfüllter Haarscher Bedingung*, HMI-B40, Berichte des Hahn-Meitner-Instituts Berlin, 1965.
- [34] ———, *Tschebyscheff-Approximation bei nicht erfüllter Haarscher Bedingung*, Z. Angew. Math. Mech., 45 (1965), pp. T81–T82.
- [35] ———, *Tschebyscheff-Approximation und Austauschverfahren bei nicht erfüllter Haarscher Bedingung*, Tagung, Oberwolfach (1965), ISNM 7, Birkhäuser Verlag, Basel, Switzerland, 1967, pp. 71–89.
- [36] CH. J. DE LA VALLÉE POUSSIN, *Sur la méthode de l'approximation minimum*, Soc. Scient. Bruxelles, Annales, deuxième partie, mémoire 35, 1911, pp. 1–16.
- [37] P. WOLFE, *A method of conjugate subgradients for minimizing nondifferentiable functions*, Math. Programming Study, 3 (1975), pp. 145–173.

ON THE STOCHASTIC MAXIMUM PRINCIPLE*

U. G. HAUSSMANN†

Abstract. A representation of the adjoint process, which appears in a general version of the maximum principle for control systems described by Girsanov solutions of stochastic differential equations, is given in terms of the linearization of the state equation. The result is only valid when the optimal control and the coefficients in the state equation are smooth; however two examples show that the result can nevertheless be applied to the nonsmooth case, solving in particular the linear regulator and the "predicted miss" problems.

1. Introduction. Consider the control problem

$$(1.1) \quad \min_u EL_0(z)$$

subject to

$$(1.2) \quad dz^{(1)} = g(t, z) dt, \quad 0 \leq t \leq 1,$$

$$dz^{(2)} = f(t, z, u) dt + \sigma(t, z) dw$$

$$z(0) = z_0$$

$$(1.3) \quad EL_i(z) \leq 0, \quad i = 1, 2, \dots, k_1,$$

$$EL_i(z) = 0, \quad i = k_1 + 1, \dots, k_2,$$

$$(1.4) \quad \int_0^1 \phi_i(t, z, u) dt \leq 0 \quad \text{w.p.1,} \quad i = 1, 2, \dots, k_3,$$

$$\psi_i(t, z) \leq 0 \quad \text{w.p.1,} \quad \text{all } t, \quad i = 1, \dots, k_4,$$

where $z = \begin{pmatrix} z^{(1)} \\ z^{(2)} \end{pmatrix} \in R^{n+m}$ and where w is an m -dimensional separable Brownian motion on (Ω, \mathcal{F}, P) . The admissible laws u will be defined shortly. In [1] a general maximum principle was derived for this problem without the constraints (1.4), using the Girsanov solution of (1.2) as opposed to the Itô solution (cf. [2] for a maximum principle for Itô solutions). Our first objective in this article is to give a maximum principle for the problem (1.1)–(1.4) under slightly weaker hypotheses than those of [1]. Since the proofs require only minor modifications we only indicate where changes are required in the proofs in [1] rather than rewrite the complete proof. This program is carried out in § 2.

In § 3 we come to the more substantial part of the work: we give a representation of the adjoint process arising in the maximum principle. This proof justifies the heuristic argument given in [1] for the claim that the adjoint process satisfies the same equation as given in [2]. This result rests upon the representation

* Received by the editors November 18, 1976, and in revised form February 28, 1977.

† Department of Mathematics, University of British Columbia, Vancouver, British Columbia, Canada V6T 1W5. This research was supported by the Canada Council under Grant W75001 and by the National Research Council of Canada under Grant A8051.

of functionals of Itô processes as stochastic integrals [3] and is restricted to the case where the optimal control law $\hat{u}(t, z)$ is Lipschitz in z . Such a result appears to be highly unsatisfactory; however we show that the linear regulator, as well as Beneš, "predicted miss" problem [4], can be "solved" by a combination of this representation, the general maximum principle, and appropriate smooth approximations.

We continue now with a precise statement of the problem. For each $t \in [0, 1]$ let $\{\mathcal{G}_t\}$ be an increasing sequence of σ -algebras in C^{n+m} and let \mathcal{Z}_t be the σ -algebra in C^{n+m} generated by

$$\{z \in C^{n+m} : z(s) \in B\}, \quad 0 \leq s \leq t, \quad B \text{ Borel in } R^{n+m}.$$

We designate the continuous functions mapping $[0, 1]$ into R^d under the sup norm by C^d .

We assume $\mathcal{G}_t \subset \mathcal{Z}_t$. Set \mathcal{B} for the Borel set of $[0, 1]$. Let Γ be a subset of R^l . We take U , the *admissible controls*, to be the $\mathcal{B} \otimes \mathcal{Z}_1$ measurable mappings

$$u : [0, 1] \times C^{n+m} \rightarrow \Gamma$$

for which

$$u(t, \cdot) : C^{n+m} \rightarrow \Gamma$$

is \mathcal{G}_t measurable. (It is assumed that the Borel algebra is on Γ .) In (1.1)–(1.4) we take $g : [0, 1] \times C^{n+m} \rightarrow R^n$, $f : [0, 1] \times C^{n+m} \times \Gamma \rightarrow R^m$, $\sigma : [0, 1] \times C^{n+m} \rightarrow R^{m \times m}$, where $R^{m \times m}$ is the set of real $m \times m$ matrices. Also $L : C^{n+m} \rightarrow R^{k_2+1}$, $\phi : [0, 1] \times C^{n+m} \times \Gamma \rightarrow R^{k_3}$ and $\psi : [0, 1] \times C^{n+m} \rightarrow R^{k_4}$. w is a separable Brownian motion on (Ω, \mathcal{F}, P) , so let \mathcal{F}_t be the sub σ -algebra of \mathcal{F} generated by $w(s, \cdot)$, $0 \leq s \leq t$ together with the P negligible sets of \mathcal{F} .

Since we are looking for necessary conditions we assume that there exists an \mathcal{F}_t adapted process \hat{z} on (Ω, \mathcal{F}, P) and a control $\hat{u} \in U$ such that (1.2)–(1.4) are satisfied for the pair (\hat{z}, \hat{u}) and such that for any \hat{z} -admissible control u , $EL_0(\hat{z}) \leq E\alpha^u(1)L_0(\hat{z})$, where u is \hat{z} -admissible if $u \in U$ and $E\alpha^u(1) = 1$ with

$$(1.5) \quad \begin{aligned} \alpha^u(t, \omega) = \exp \left\{ \int_0^t [\sigma(s, \hat{z}(\omega))^{-1} [f(s, \hat{z}(\omega), u(s, \hat{z}(\omega))) \right. \\ \left. - f(s, \hat{z}(\omega), \hat{u}(s, \hat{z}(\omega)))]' dw(s) \right. \\ \left. - \frac{1}{2} \int_0^t |\sigma(s, \hat{z}(\omega)) \right. \\ \left. \cdot [f(s, \hat{z}(\omega), u(s, \hat{z}(\omega))) - f(s, \hat{z}(\omega), \hat{u}(s, \hat{z}(\omega)))]^2 ds \right\}. \end{aligned}$$

Here a' denotes the transpose of a and E is expectation under P .

We assume

- A₁: For all $u \in U$, $g(\cdot, \cdot)$, $f(\cdot, \cdot, u(\cdot, \cdot))$, $\sigma(\cdot, \cdot)$ are $\mathcal{B} \otimes \mathcal{Z}_1$ measurable and (\mathcal{Z}_t) adapted; $\phi(\cdot, \cdot, u(\cdot, \cdot))$, $\psi(\cdot, \cdot)$ are $\mathcal{B} \otimes \mathcal{Z}_1$ measurable, and $L(\cdot)$ is \mathcal{Z}_1 measurable.
- A₂: $\sigma(t, z)^{-1}$ exists for each (t, z) .
- A₃: $E \exp(\lambda_0 \|\hat{z}\|^2) < \infty$ for some $\lambda_0 > 0$.
- A₄: $|L(z)| \leq K_0(1 + \|z\|^{p_0})$, $|\phi(t, z, u)| \leq K(t)(1 + \|z\|^{p_0})$, $\int_0^1 K(t) dt < \infty$ for some $p_0 < \infty$.

A₅: $\sigma(t, z)^{-1}f(t, z, u)$ and $\phi(t, z, u)$ are continuous in u uniformly in t for each z .

A₆: $|\sigma(t, z)^{-1}f(t, z, u)|^2 \leq K_0(1 + \|z\|^2)$ for all u in Γ , if $\|z\|_t = \sup_{0 \leq s \leq t} |z(s)|$.

This problem differs from the one treated in [1] in several ways. For one thing the controls are in what we call feedback form, i.e. functions of z . In fact, since we use the method of Girsanov to define the comparison solutions we should think of the underlying probability space as C^m , i.e. ω corresponds either to w or to z . If we take the former case i.e. $u(t, w)$, (open loop form), then the Girsanov solution becomes

$$(1.6) \quad d\hat{z} = f(t, \hat{z}, u(t, w)) dt + \sigma(t, \hat{z}) dw^u$$

where

$$w^u(t) = w(t) - \int_0^t \sigma(s, \hat{z})^{-1} [f(s, \hat{z}, u) - f(s, \hat{z}, w)] ds$$

is now a Wiener process under $dP^u = \alpha^u dP$. The difficulty is apparent in (1.6): the drift term ought to be $f(t, \hat{z}, u(t, w^u))$, but isn't! This problem does not arise if $u(t, z)$ is used. For the Girsanov solution the trajectories are invariant but not the Brownian motion so that admissible controls should be feedback; where for the Itô solution the trajectories change with the control but not the Brownian motion, so open loop controls avoid the difficulty of an implicit, or circular, definition.

Another difference in [1] is that A_3 is replaced by

A₃': $\int_0^1 |\sigma(t, z)|^2 dt \leq k < \infty$ for all z , $|g(t, z)|^2 \leq K_0(1 + \|z\|^2)$.

In fact A₃' implies A₃ and all $u \in U$ are \hat{z} -admissible, but, of course, A₃' is much easier to verify than A₃.

Finally, the addition of the "hard" constraints involving ϕ and ψ is important. The constraints

$$(1.7) \quad \psi_i(t, z) \leq 0 \quad \text{w.p.1}$$

can be disposed of very easily. With a Girsanov solution (\hat{z}, P^u) (1.7) translates to

$$(1.8) \quad \psi_i(t, \hat{z}) \leq 0 \quad \text{w.p.1} \quad P^u.$$

But (\hat{z}, P) is optimal, hence satisfies (1.7), and P^u is absolutely continuous with respect to P , so that (1.8) holds. For this reason constraints of this form, although they may be present, play no part in the derivation of a maximum principle, and we will henceforth assume they are absent.

If we now consider ϕ , Neustadt's theory cannot be applied immediately because the open, convex cone

$$\{y \in Y: y(\omega) < 0\}$$

is not open in any "reasonable" space, Y , e.g. $L_p(\Omega, \mathcal{F}, P)$. Moreover this approach would yield a multiplier in Y^* , not necessarily a nice space. However, for $u \in U$,

$$(1.9) \quad \Phi^u(\hat{z}) = \int_0^1 \phi(t, \hat{z}, u(t, z)) dt \leq 0 \quad \text{w.p.1} \quad P^u$$

if

$$(1.10) \quad \Phi''(\hat{z}) \leq 0 \quad \text{w.p.1 } P,$$

just as for ψ . Let

$$[\Phi''_+(z)]_i = \begin{cases} [\Phi''(z)]_i & \text{if } [\Phi''(z)]_i \geq 0, \\ 0 & \text{if } [\Phi''(z)]_i < 0, \end{cases}$$

where $[a]_i$ denotes the i th component. Then (1.10) is equivalent to

$$(1.11) \quad E\Phi''_+(z) = 0.$$

Hence the constraints (1.3)–(1.4) are rewritten as

$$(1.12) \quad \begin{aligned} EL_i(z) &\leq 0, & i = 1, 2, \dots, k_1, \\ EL_i(z) &= 0, & i = k_1 + 1, \dots, k_2, \\ E[\Phi''_+(z)]_i &= 0, & i = 1, 2, \dots, k_3. \end{aligned}$$

2. The stochastic maximum principle. We shall now derive the maximum principle along the lines used in §§ 3 and 4 of [1]. First we wish to represent $L(\hat{z})$ as

$$(2.1) \quad L(\hat{z}) = EL(\hat{z}) + \int_0^1 \chi dw$$

with χ in $L_2(dt \times dP)$. This follows from A_4 because A_3 implies

$$(2.2) \quad E\|\hat{z}\|^{q_0} \leq K(q_0) < \infty$$

for any $q_0 < \infty$, in particular for $q_0 = 2p_0$. Next we require an L_p bound on the densities α'' , i.e.

$$E(\alpha'')^p \leq K$$

for some $p > 1$, all u ; cf. [1, (3.4) ff.]. We obtain this bound by simply restricting u to U_p^K , the set of \hat{z} -admissible controls which do satisfy $E(\alpha'')^p \leq K$. If we set

$$(2.3) \quad \begin{aligned} \zeta''(t) &= E(L) + \int_0^t \alpha''_s \chi_s \sigma_s^{-1} \Delta f_s'' ds \\ \eta'' &= \Phi''_+(\hat{z}(\cdot, \omega)) - \Phi''_+(\hat{z}(\cdot, \omega)) \end{aligned}$$

where

$$\Delta f_t'' = f(t, \hat{z}(\cdot, \omega), u(t, \hat{z}(\cdot, \omega))) - f(t, \hat{z}(\cdot, \omega), \hat{u}(t, \hat{z}(\cdot, \omega)))$$

it now follows (cf. § 3 of [1]) for any p, K , that (ζ'', η'') solves:

$$(2.4) \quad \min_{u \in U_p^K} E\zeta''_0$$

subject to

$$(2.5) \quad \begin{aligned} E\zeta''_i(1) &\leq 0, & i = 1, 2, \dots, k_1, \\ E\zeta''_i(1) &= 0, & i = k_1 + 1, \dots, k_2, \\ E\eta''_i &= 0, & i = 1, 2, \dots, k_3, \end{aligned}$$

since $\Phi''_+(\hat{z}(\cdot, \omega)) = 0$ w.p.1.

Turning now to section four of [1], we set $\mathcal{F} = L_1(\Omega, \mathcal{F}, P) \times L_1(\Omega, \mathcal{F}, P)$ and $\mathcal{T}_1 = \{(\zeta^u(1), \eta^u): u \in U_p^K, \zeta^u(1) \in L_1(\Omega, \mathcal{F}, P)\}$. We remark that $\eta^u \in L_1(\Omega, \mathcal{F}, P)$ because of A_4 and (2.2). Now define

$$\delta\zeta_{t,u_0}(\omega) = \chi_t \sigma_t^{-1} \Delta f_t^{u_0}$$

$$\delta\eta_{t,u_0}(\omega) = \phi_+(t, \hat{z}(\cdot, \omega), \hat{u}_0(\hat{z}(\cdot, \omega))) - \phi_+(t, \hat{z}(\cdot, \omega), \hat{u}(t_0, \hat{z}(\cdot, \omega)))$$

for $u_0 \in L_1(C^{n+m}, \mathcal{G}, P_0 \hat{z}^{-1}; \Gamma) \equiv U_t$. Now define

$$\mathcal{M} = \text{co} \{(\delta\zeta_{t,u_0}, \delta\eta_{t,u_0}): 0 \leq t_0 < 1, t_0 \notin T_0, u_0 \in U_t\}$$

where $\text{co } A$ is the convex hull of the set A and T_0 is a null set to be defined later but containing those points t at which $K(t) = \infty$. Then $\mathcal{M} \subset \mathcal{T}$; cf. [1].

To assure that our perturbed controls are in U_p^K we rely not on [1, Lemma 4.1], but rather on the following lemma inspired by a result of Liptzer and Shiryaev as presented in [5].

LEMMA 2.1. $E(\alpha^u(1, \omega))^p \leq K$ for all \hat{z} -admissible u such that

$$(2.6) \quad \int_0^1 |\sigma(t, \hat{z}(\cdot, \omega))^{-1} [f(t, \hat{z}(\cdot, \omega), u(t, \omega)) - f(t, \hat{z}(\cdot, \omega), \hat{u}(t, \omega))]|^2 dt \leq \xi K_0(1 + \|z\|^2)$$

and all p such that

$$(2.7) \quad 2K_0 p^2 \xi \leq \lambda_0,$$

where ξ is any positive number.

Proof. We write $|\sigma^{-1} \Delta f^u|^2$ for the integrand above. We set $M_t^u = \int_0^t \sigma^{-1} \Delta f^u dw$, $A_t^u = \int_0^t |\sigma^{-1} \Delta f^u|^2 ds$, $Z_t^{u,\lambda} = \exp(\lambda M_t^u - (\lambda^2/2) A_t^u)$, $\tau_n^u = \inf \{t \geq 0: M_t^u \geq n \text{ or } |\sigma^{-1} \Delta f^u(t)|^2 \geq n\}$. Then for all $\lambda \geq 0$

$$E(Z_{t \wedge \tau_n^u}^{u,\lambda} | \mathcal{F}_{t'}) = Z_{t' \wedge \tau_n^u}^{u,\lambda} + \lambda E \int_{t' \wedge \tau_n^u}^{t \wedge \tau_n^u} Z_{s \wedge \tau_n^u}^{u,\lambda} \sigma^{-1} \Delta f^u dw(s) = Z_{t' \wedge \tau_n^u}^{u,\lambda}$$

because 1_A is the indicator function of A)

$$\int_{t'}^t E 1_{s \leq \tau_n^u} |Z_{s \wedge \tau_n^u}^{u,\lambda}|^2 |\sigma^{-1} \Delta f^u|^2 ds \leq \int_{t'}^t e^{2\lambda n} n^2 ds < \infty.$$

Hence $Z_{t \wedge \tau_n^u}^{u,\lambda}$ is a martingale for all $\lambda \geq 0$. Now

$$\begin{aligned} E(Z_{t \wedge \tau_n^u}^{u,\lambda})^p &= E\{[\exp(2pM_{t \wedge \tau_n^u}^u - 2p^2 A_{t \wedge \tau_n^u}^u)]^{1/2} \exp(p(p-1/2)A_{t \wedge \tau_n^u}^u)\} \\ &\leq [EZ_{t \wedge \tau_n^u}^{u,2p}]^{1/2} \cdot (E \exp(p(2p-1)A_{t \wedge \tau_n^u}^u))^{1/2} \\ &= [E \exp p(2p-1)A_{t \wedge \tau_n^u}^u]^{1/2} \\ &\leq [E \exp p(2p-1) \int_0^1 |\sigma^{-1} \Delta f^u|^2 ds]^{1/2} \\ &\leq [E \exp(p(2p-1)\xi K_0(1 + \|z\|^2))]^{1/2} \\ &\leq [E \exp \lambda_0 \|z\|^2]^{1/2} \exp(p(p-1/2)\xi K_0) \end{aligned}$$

if (2.6) holds. From Fatou's lemma it now follows that

$$\begin{aligned} E(\alpha^u(1, \omega))^p &= E(Z_1^u)^p \\ &\leq \exp(p(p-1/2)\xi K_0)[E \exp(\lambda_0 \|\dot{z}\|^2)]^{1/2} \\ &\leq \exp(\lambda_0/2)[E \exp(\lambda_0 \|\dot{z}\|^2)]^{1/2}. \end{aligned}$$

COROLLARY 2.2. *For any $p > 1$, $u_{\varepsilon\lambda}$ as defined in [1] is in U_p^k for ε sufficiently small, depending on p .*

Proof. With $u = u_{\varepsilon\lambda}$, $\sigma^{-1} \Delta f^u(t)$ has support on a set of measure ε . By A_6 , (2.7) is satisfied if $\xi = 2\varepsilon$.

With $(\zeta_\lambda, \eta_\lambda) = \sum_{i=1}^q \delta t_i (\delta \zeta_{t_i, u_i}, \delta \eta_{t_i, u_i})$ define $\theta(\zeta_\lambda, \eta_\lambda) = (\zeta^{\varepsilon\lambda}(1), \eta^{\varepsilon\lambda})$ where $(\zeta^{\varepsilon\lambda}(t), \eta^{\varepsilon\lambda})$ are given by (2.3) with $u = u_{\varepsilon\lambda}$ defined in [1]. Now Lemmas 4.3 to 4.6 of [1] establish certain properties of the first component of the function θ . Parallel but easier proofs give the same results for the second component of θ if we observe that when $(1/\varepsilon) \int_{I(\varepsilon\lambda)} \phi \, ds \rightarrow \phi(t) \, \delta t$, then $(1/\varepsilon) (\int_{I(\varepsilon\lambda)} \phi \, ds)_+ \rightarrow \phi_+(t) \, \delta t$.

Thus the following maximum principle has been established.

MAXIMUM PRINCIPLE. *There is a nonzero vector $(\rho, \theta) \in R^{k_3} \times R^{k_2+1}$ such that*

- (i) $\theta_i \leq 0$, $i = 0, 1, \dots, k_i$,
- (ii) $\theta' EL(\dot{z}) = \theta_0 EL_0(z)$,
- (iii) if $p(t, \omega) = [\chi(t, \omega) \sigma(t, \dot{z}(\cdot, \omega))^{-1}]' \theta$,

then for $t \notin T_0$, a set of Lebesgue measure zero,

$$(2.8) \quad \begin{aligned} &E\{p(t)'[f(t, \dot{z}, u(\dot{z})) - f(t, \dot{z}, \hat{u}(t, \dot{z}))] \\ &\quad + \rho'[\phi_+(t, \dot{z}, u(\dot{z})) - \phi_+(t, \dot{z}, \hat{u}(t, \dot{z}))]\} \leq 0 \end{aligned}$$

for any u in $\mathcal{U}_t = \{u: C^{n+m} \rightarrow \Gamma, \mathcal{G}_t \text{ measurable}, E|u| < \infty\}$.

COROLLARY. *If $E|\hat{u}(t)| < \infty$ for $t \notin T_0$, and if ϕ_+ is adapted, then (2.8) can be replaced by*

$$(2.8)' \quad \begin{aligned} &E\{p(t)'[f(t, \dot{z}, u(\dot{z})) - f(t, \dot{z}, \hat{u}(t, \dot{z}))] \\ &\quad + \rho'[\phi_+(t, \dot{z}, u(\dot{z})) - \phi_+(t, \dot{z}, \hat{u}(t, \dot{z}))]\} \dot{z}^{-1} \mathcal{G}_t \leq 0. \end{aligned}$$

Remarks. 1. Let us add that the condition that $\{\mathcal{G}_t\}$ be increasing can be replaced by

A_7 : for almost all $t \in [0, 1]$ there exists a $\delta(t) > 0$ and an increasing sequence of σ -algebras \mathcal{H}_s^t , $t - \delta(t) \leq s < t$, such that $\mathcal{H}_s^t \subset \mathcal{G}_s$ and $\lim_{s \uparrow t} \mathcal{H}_s^t = \mathcal{G}_t$.

Now the maximum principle still holds as stated. In the proof one changes the definition of $I_j(\varepsilon\lambda)$ and $u_{\varepsilon\lambda}$ to

$$\begin{aligned} I_j(\varepsilon\lambda) &= (t_j - \varepsilon\tau_j, t_j - \varepsilon\tau_j + \delta t_j], \\ u_{\varepsilon\lambda}(t) &= \begin{cases} E\{u_j^0 | \mathcal{H}_{t_j - \varepsilon}^t\}, & t \in I_j(\varepsilon\lambda), \\ \hat{u}(t), & t \notin \bigcup_j I_j(\varepsilon\lambda). \end{cases} \end{aligned}$$

Then

$$E\{u_j | \mathcal{H}_{t_j - \varepsilon}^t\} \rightarrow u_j \quad \text{w.p.1}$$

as $\varepsilon \rightarrow 0$ (if the correct version of conditional expectation is chosen). Using this fact and A_5 one completes the proof of [1, Lemma 4.6] by showing

$$E \left| \frac{1}{\varepsilon} \int_{I_j(\varepsilon\lambda)} \chi \sigma_t^{-1} (\Delta f_t^{u_{\varepsilon\lambda}} - \Delta f_t^{u_i}) dt \right| + E \left| \frac{1}{\varepsilon} \int_{I_j(\varepsilon\lambda)} (\Delta \phi^{u_{\varepsilon\lambda}} - \Delta \phi_t^{u_i}) dt \right| \rightarrow 0$$

as $\varepsilon \rightarrow 0$. This fact did not need proving in the previous case since it is obvious if $u_{\varepsilon\lambda}(t) = u_i$ on $I_j(\varepsilon\lambda)$.

Even if $\{\mathcal{G}_t\}$ is increasing, and left continuous (so that A_7 also holds with $\mathcal{H}_s^t = \mathcal{G}_s$) this last version of the maximum principle is slightly different: in the former case (i.e. without A_7) the exceptional null set T_0 contains $t = 1$ but not necessarily $t = 0$. In the case with A_7 the situation is reversed i.e. $\{0\} \in T_0$ but $\{1\}$ may not be.

Unfortunately this formulation still will not fit the Markov case, i.e. when all functions including the control u depend, at time t , only on $z(t)$, e.g. $u(t, z(t))$. However if we set

$$\bar{U} = \{u: R^{n+m} \rightarrow \Gamma \text{ continuous, } |u(x)| \leq K(1 + |x|^{q_0})\}$$

and

$$\mathcal{U}_t = \{u: C^{n+m} \rightarrow \Gamma, u(z) = v(z(t)), v \in \bar{U}_t\},$$

then the maximum principle still holds if

$$u_{\varepsilon\lambda}(t) \equiv u_i(z(t)), \quad t \in I_j(\varepsilon\lambda).$$

2. Next we remark that we can relax the condition $\hat{z}^{-1}\mathcal{L}_t = \mathcal{F}_t = w_t^{-1}\mathcal{L}_t$. Instead define $\mathcal{F}_t \equiv \hat{z}^{-1}\mathcal{L}_t$ directly and assume that (w_t, \mathcal{F}_t, P) is a separable Brownian motion. Then in (2.1) we have

$$L(\hat{z}) = EL(\hat{z}) + \int_0^1 \chi dw_t + N_1$$

where N_t is an $\{\mathcal{F}_t\}$ martingale, mean zero, which is orthogonal to w_t , hence to $\alpha^u(t)$. Since this implies that $\alpha^u(t)N_t$ is a martingale, then $E^u N_1 = E\alpha^u(0)N_0 = EN_0 = 0$. Hence we still have $E^u L = E\zeta^u(1)$ with ζ^u as in (2.3).

3. Finally we observe that Γ need not be constant; i.e. we can allow $u(t, \cdot)$ to map into Γ_t if for almost all t there is a number $\delta(t) > 0$ and a mapping m_t defined on $[t, t + \delta(t)] \times \Gamma_t$ such that $m_t(s, \cdot): \Gamma_t \rightarrow \Gamma_s$ and $m_t(\cdot, u)$ is continuous for each $u \in \Gamma_t$ and $m_t(t, u) = u$. (If A_7 is used then m is defined on $(t - \delta(t), t]$.) Now we define $u_{\varepsilon\lambda}(t) = m_{t_j}(t, u_{t_j})$ for $t \in I_j(\varepsilon\lambda)$. If Γ_t is increasing then we can let $m_t(s, \cdot)$ be the identity map.

3. The adjoint process—examples. In [2] Kushner derives a stochastic maximum principle for Itô equations. The adjoint process p is given as the solution to a linearized version of the Itô equation. In [1] we gave a formal argument to conclude that our multiplier p also satisfied this equation and hence our theorem reduces to Kushner's. We shall now give a proof using [3, Theorem 1] to obtain the desired representation of p . This represents a "rigorization" of the formal argument suggested in [1]. In addition to A_1 – A_6 we need to assume also:

H_1 : L is Fréchet differentiable with derivative $(d/dx)L(x)$ at x . Moreover

there are K, β, δ positive, finite, such that

$$\left| \frac{d}{dx} L(x) - \frac{d}{dx} L(y) \right| \leq K(1 + \|x\|^\beta)(1 + \|y\|^\beta) \|x - y\|^\delta$$

where, of course, the left side is the norm in C^* , the dual of C^{n+m} .

H₂: $|F(t, x)|^2 + |\sigma(t, x)|^2 \leq K(1 + \|x\|_t^2)$, $|F(t, x) - F(t, y)|^2 + |\sigma(t, x) - \sigma(t, y)|^2 \leq K\|x - y\|_t^2$ where $F(t, x)' = (f(t, x, \hat{u}(t, x))', g(t, x)')$.

H₃: $F(t, \cdot)$ and $\sigma(t, \cdot)$ are Fréchet differentiable for almost all t , with derivatives $F_x(t, x)$, $\sigma_x(t, x)$ which are Hölder continuous in x of order $\delta > 0$, uniformly in t . From H₂ it follows that the norms of F_x and σ_x are bounded by \sqrt{K} .

These assumptions are slightly stronger than Kushner's, particularly of course A₂ and A₃. This is the price we pay for using the Girsanov approach as opposed to the Itô approach when defining solutions, but there is a pay-off in the feedback control case; to obtain the representation of p we only require $\hat{u}(t, x)$ to be smooth in x , the comparison controls need not be. For now,

$$f_x \equiv \frac{\partial f}{\partial x} + \frac{\partial f}{\partial u} \frac{\partial \hat{u}}{\partial x}.$$

If

$$\frac{dL_i}{dz}(\hat{z})x = \int_0^1 \mu_i(dt; \hat{z})x(t)$$

then take $\mu(t; \hat{z})$ as the matrix with rows $\mu_i(t; \hat{z})$ and

$$(3.1) \quad \lambda(t, \omega) = \int_t^1 \mu(ds; \hat{z}(\omega)) \Phi(s, t, \omega) \begin{pmatrix} 0 \\ \sigma(t, \hat{z}(\omega)) \end{pmatrix}$$

where the columns of $\Phi(t, s)$ satisfy

$$(3.2) \quad dy = \begin{pmatrix} g_x(t, \hat{z}(\omega)) \\ f_x(t, \hat{z}(\omega), \hat{u}(t, \hat{z})) \end{pmatrix} y dt + \sum_{\rho=1}^m \begin{pmatrix} 0 \\ \sigma_x^\rho(t, \hat{z}(\omega)) \end{pmatrix} y dw^\rho$$

for $t \geq s$ with $\Phi(s, s) = I$, the $n + m$ dimensional identity matrix. According to [3, Thm. 1], it follows that

$$(3.3) \quad \chi(t, \omega) = E(\lambda(t) | \mathcal{F}_t) = E \left\{ \int_t^1 \mu(ds; \hat{z}) \Phi(s, t) | \mathcal{F}_t \right\} \begin{pmatrix} 0 \\ \sigma(t, \hat{z}(\omega)) \end{pmatrix}.$$

Hence

$$p(t, \omega)' = E \left\{ \int_t^1 \theta' \mu(ds; z) \Phi(s, t) | \mathcal{F}_t \right\} \begin{pmatrix} 0 \\ I \end{pmatrix}.$$

If we now define

$$\bar{p}(t, \omega) = \int_t^1 \theta' \mu(ds; \hat{z}) \Phi(s, t) \begin{pmatrix} 0 \\ I \end{pmatrix}$$

then (2.8) and (2.8)' can be rewritten with $\bar{p}(t, \omega)$ replacing $p(t, \omega)$.

For the special case treated in [2] where

$$L_i(z) = L_i(z(t_i))$$

with $t_0 = 1 > t_1 > t_2 > \dots > t_{k_2} \geq 0$, and with

$$\frac{dL_i}{dz}(\hat{z}(t_i)) = \mu_i(t_i; \hat{z}) - \mu_i(t_i -; \hat{z}) = l_i,$$

a $1 \times (n + m)$ matrix, one has

$$\begin{aligned} \bar{p}(1) &= \theta_0 l_0, \\ \bar{p}(t) &= \bar{p}(1) \Phi(1-t) \quad \text{if } 1 \geq t \geq t_1, \\ \bar{p}(t_i -) &= \bar{p}(t_i) + \theta_i l_i, \\ \bar{p}(t) &= \bar{p}(t_i -) \Phi(t_i, t) \quad \text{if } t_i > t \geq t_{i-1}. \end{aligned} \quad (3.4)$$

This is the desired representation of the adjoint process.

The above result is of course not very satisfactory because in general \hat{u} is not differentiable. However as the following examples show one can still obtain results even in the discontinuous case by combining smooth approximations and the general maximum principle.

Example 1.

$$\min EL[k'x(1)]$$

subject to $|u_i| \leq 1$ and

$$dx = [A(t)x + B(t)u] dt + C(t) dw, \quad 0 \leq t \leq 1, \quad (3.5)$$

where $x(t) \in R^n$, $u(t) \in R^d$, $w(t) \in R^n$, $k \in R^n$, and l is a differentiable positive even functional of one variable, such that H_1 holds and

$$y \frac{dl(y)}{dy} \geq 0 \quad (3.6)$$

and $l(y) \leq K(1 + |y|^{q_0})$ for some $q_0 < \infty$. Beneš [4] solved this problem without the differentiability of l , but this requirement can be relaxed by suitably approximating. If s is defined by

$$\frac{ds}{dt} = -A(t)'s, \quad s(1) = k$$

then engineers have guessed that the optimal law is

$$\hat{u}(t, x) = -\operatorname{sgn} B(t)'s(t)s(t)'x. \quad (3.7)$$

Beneš [4] has shown that this is correct using a complicated direct method. We shall show that this law satisfies our necessary conditions and hence is quite possibly an optimal control, (necessary conditions can never do any more).

Consider as the basic process on (Ω, \mathcal{F}, P)

$$dz = A_t z dt + C_t dw_t, \quad (3.8)$$

and set $u^\varepsilon(t, x) = -\operatorname{sgn}[B(t)'s(t)] \operatorname{sgn}^\varepsilon[s(t)'x]$ where $\operatorname{sgn}^\varepsilon y = \operatorname{sgn} y$ for $|y| \geq \varepsilon$, $\operatorname{sgn}^\varepsilon$ is odd, smooth and nondecreasing. As usual if x is a vector then $\operatorname{sgn} x$ has i th component $\operatorname{sgn} x_i$. So according to Girsanov [8] if

$$w_t^\varepsilon = w_t - \int_0^t C(\tau)^{-1} B(\tau) u^\varepsilon(\tau, z(\tau, \omega)) d\tau$$

then w^ε is a Wiener process under a certain probability law P^ε , and

$$(3.9) \quad dz = [A_t z_t + B_t u^\varepsilon(t, z_t)] dt + C_t dw^\varepsilon.$$

With $u = \hat{u}$ and

$$\hat{w}_t = w_t - \int_0^t C_\tau^{-1} B_\tau \hat{u} d\tau$$

this becomes

$$dz = (A_t z_t + B_t \hat{u}_t) dt + C_t d\hat{w}_t.$$

Moreover our representation theorem states that

$$L(z) = E^\varepsilon L(z) + \int_0^1 \chi^\varepsilon dw^\varepsilon,$$

$$\chi_t^\varepsilon = E^\varepsilon \left\{ \frac{dl}{dy} (k' z_1) k' \Phi^\varepsilon(1, t) | \mathcal{F}_t \right\} C_b,$$

$$\frac{d\Phi^\varepsilon}{dt}(t, \tau) = \left[A_t - B_t \operatorname{sgn}(B_t' s_t) s_t' \frac{d}{dy} \operatorname{sgn}^\varepsilon(s_t' z_t) \right] \Phi^\varepsilon(t, \tau)$$

$$\Phi(\tau, \tau) = I,$$

where E^ε is expectation under P^ε .

If we now set $\eta^\varepsilon(t) = \Phi^\varepsilon(t, \tau)' s(t)$, then

$$\frac{d\eta^\varepsilon}{dt} = -|s(t)' B(t)| \frac{d \operatorname{sgn}^\varepsilon}{dy} (s' z) \eta^\varepsilon$$

where

$$\begin{aligned} |s(t)' B(t)| &= s(t)' B(t) \operatorname{sgn}[B(t)' s(t)] \\ &= \sum_i \left| \sum_j B_{ji}(t) s_j(t) \right|. \end{aligned}$$

It follows that $\eta^\varepsilon(t) = \exp \left(\int_\tau^t f^\varepsilon(\sigma) d\sigma \right) s(\tau)$ where $f^\varepsilon(t)$ is a scalar depending on ε , t and $s(t)' z(t)$. If we now set $y(t) = s(t)' z(t)$, then

$$\chi_t^\varepsilon = E^\varepsilon \left\{ \frac{dl}{dy} (y(1)) \exp \left[\int_t^1 f^\varepsilon(\tau) d\tau \right] | \mathcal{F}_t \right\} s(t) C(t).$$

Hence with $u^\varepsilon(t, z(t)) = -\operatorname{sgn}[B(t)'s(t)] \operatorname{sgn}^\varepsilon[s(t)'z(t)]$,

$$(3.11) \quad \chi_t^\varepsilon C(t)^{-1} B(t)(u - u_t^\varepsilon) \geq 0, \quad u \in [-1, 1]^d$$

provided $|y(t)| \geq \varepsilon$ and also provided

$$(3.12) \quad \operatorname{sgn} y(t) = \operatorname{sgn} E^\varepsilon \left\{ \frac{dl}{dy}(y(1)) \exp \left[\int_t^1 f^\varepsilon(\tau) d\tau \right] \middle| \mathcal{F}_t \right\}.$$

Now $dy = s' C dw$, so that if $\ell(t) = \int_0^t |s'_\tau C_\tau|^2 d\tau$ then $y \circ \ell^{-1} = b$ is a one dimensional Brownian motion under P . (ℓ^{-1} exists since $CC' > 0$.) On the other hand if $u^\varepsilon(t, z(t)) \equiv v^\varepsilon(t, y(t))$ then

$$dy = s'_t B_t v^\varepsilon(t, y_t) dt + s'_t C_t dw_t^\varepsilon$$

and

$$\begin{aligned} b^\varepsilon(t) &\equiv \int_0^{\ell^{-1}(t)} s'_\tau C d w^\varepsilon \\ &= b(t) - \int_0^t \frac{s'(\ell^{-1}(\tau)) B(\ell^{-1}(\tau)) v^\varepsilon(\ell^{-1}(\tau), b(\tau)) d\tau}{|s'(\ell^{-1}(t)) C(\ell^{-1}(\tau))|^2} \end{aligned}$$

is a Wiener process under P^ε . We write the last equation as

$$b^\varepsilon(t) = b(t) - \int_0^t g(\tau)' \bar{v}^\varepsilon(\tau, b(\tau)) d\tau.$$

But according to Girsanov's theorem [8] it now follows that

$$\frac{dP^\varepsilon}{dP} = \alpha^\varepsilon = \exp \left\{ \int_0^{\ell(1)} g(t)' \bar{v}^\varepsilon(t, b_t) db_t - \frac{1}{2} \int_0^{\ell(1)} |g(t)' \bar{v}^\varepsilon(t, b_t)|^2 dt \right\}.$$

Note that $\ell(1) \equiv t_1$ is nonrandom. Now (3.12) is equivalent to

$$(3.13) \quad \operatorname{sgn} x = \operatorname{sgn} E \left\{ \frac{dl}{dy}(b(t_1)) \phi(t_1, t, b) | b(t) = x \right\}$$

where $dl(y)/dy$ is odd and nonnegative for $y > 0$, and where

$$\begin{aligned} \phi(t_1, t, b) &= \exp \left\{ - \int_t^{t_1} |g(\tau)| \frac{d}{dy} \operatorname{sgn}^\varepsilon(b(\tau)) d\tau \right. \\ &\quad \left. - \int_t^{t_1} |g(\tau)| \operatorname{sgn}^\varepsilon b(\tau) db_\tau - \frac{1}{2} \int_t^{t_1} |g(\tau)|^2 (\operatorname{sgn}^\varepsilon b(\tau))^2 d\tau \right\} \end{aligned}$$

is always positive and "even" in b . Moreover

$$\phi(t_1, t, b) = \phi(t_1, t_2, b) \phi(t_2, t, b)$$

for $t_1 \geq t_2 \geq t$. With $x > 0$ set $t_2 = \inf \{s > t: b(s) = 0\}$.

Then

$$\begin{aligned}
 & E\left\{\frac{dl}{dy}(b_{t_1})\phi(t_1, t, b)|b_t = x\right\} \\
 &= E\left\{E\left\{\frac{dl}{dy}(b_{t_1})\phi(t_1, t_2, b)|b_t = x, t_2\right\}\phi(t_2, t, b)|b_t = x\right\} \\
 &= E\left\{\frac{dl}{dy}(b_{t_1})\phi(t_1, t_2, b)|b_t = x, t_2 > t_1\right\}P(t_2 > t_1) \\
 &> 0.
 \end{aligned}$$

The last equality follows because if $t \leq t_1$ then the inner conditional expectation is zero since a Wiener process originating from 0 is symmetrically distributed and $(dl/dy)\phi$ is odd. The inequality follows because if $t_2 > t_1$ then on $[t, t_1]$ $b(\cdot)$ stays positive and $(dl/dy)(b_{t_1}) \geq 0$, but $\neq 0$. (We assume that l is not constant for in that case $\chi = 0$.) Hence $\chi_t^\varepsilon > 0$ if $b_t > 0$ and similarly $\chi_t^\varepsilon < 0$ if $b_t < 0$, so that (3.12) holds.

Now let $\varepsilon \rightarrow 0$. Clearly

$$\alpha^\varepsilon \rightarrow \exp\left\{\int_0^{t_1} |g(t)| \operatorname{sgn} b_t db_t - \frac{1}{2} \int_0^{t_1} |g(t)|^2 dt\right\} = \hat{\alpha}$$

in probability, and hence (by uniform integrability) in L_p for $p > 1$ sufficiently small. Since $E|l(b_{t_1})|^q < \infty$ for all $q < \infty$, since

$$\begin{aligned}
 l(b_{t_1}) &= E\alpha^\varepsilon l(b_{t_1}) + \int_0^1 \chi^\varepsilon dw^\varepsilon \\
 &= E\alpha^\varepsilon l(b_{t_1}) + \int_0^1 \chi^\varepsilon dw^0 \\
 &\quad + \int_0^{t_1} \left[\chi^\varepsilon \frac{C^{-1}B \operatorname{sgn}(B's)}{|s'C|^2} \right] \ell^{-1}(\tau) [\operatorname{sgn}^\varepsilon b_\tau - \operatorname{sgn} b_\tau] d\tau,
 \end{aligned}$$

and since

$$\begin{aligned}
 & E\left\{\int_0^{t_1} \left[\chi^\varepsilon \frac{C^{-1}B \operatorname{sgn}(B's)}{|s'C|^2} \right] \circ \ell^{-1}(\tau) [\operatorname{sgn}^\varepsilon b_\tau - \operatorname{sgn} b_\tau] d\tau\right\}^2 \\
 &\leq E\alpha^\varepsilon [l(b_{t_1}) - E\alpha^\varepsilon l(b_{t_1})]^2 \\
 &\quad \cdot E \int_0^{t_1} \left\{ \left[\frac{C^{-1}B \operatorname{sgn}(B's)}{|s'C|^2} \right] \circ \ell^{-1}(\tau) \right\}^2 [\operatorname{sgn}^\varepsilon b_\tau - \operatorname{sgn} b_\tau]^2 d\tau \\
 &\rightarrow 0
 \end{aligned}$$

as $\varepsilon \rightarrow 0$ (because $E(\alpha^\varepsilon)^p \leq K$), then $\int_0^1 \chi^\varepsilon d\hat{w}$ converges in L_2 . Hence by [9, Lemma 1], there exists $\hat{\chi}$ such that $\chi^\varepsilon \rightarrow \hat{\chi}$ a.e. and

$$l(k'z_1) = El(k'z_1) + \int_0^1 \hat{\chi} d\hat{w}.$$

Moreover taking limits in (3.11) yields that

$$\hat{\chi}_t C_t^{-1} B_t [u - \hat{u}(t, z_t)] \geq 0, \quad u \in [-1, 1]^d$$

for all $k'z_t \neq 0$, i.e. w.p.1, for all t .

Hence \hat{u} does in fact satisfy the necessary conditions of the general maximum principle. ($\theta_0 = -1$.)

Example 2. Let us consider a simple case of the linear regulator problem. We choose this simple scalar case only for ease of calculation.

$$dx = (a_t x + b_t u) dt + c_t dw, \quad \min_u E \int_0^1 (x^2 + u^2) dt.$$

Our optimal control candidate is to have the form

$$\hat{u}(t, x) = k(t)x.$$

The difficulty is that the system is not bounded in u and hence we must perform a transformation which requires the above form of u . We set $u = k(t)x + v$. Then the equivalent problem is

$$dx = [(a_t + b_t k_t)x + b_t v_t] dt + c_t dw,$$

$$\min_v E \int_0^1 [(1 + k_t^2)x^2 + 2k_t x_t v_t + v_t^2] dt;$$

moreover we now know that \hat{u} is optimal if and only if $v = 0$ is optimal, i.e. $\hat{v} = 0$. Hence instead of allowing v to assume all values we can restrict it to lie in $[-1, 1]$. Next set

$$dy = (2k_v x + v^2) dt + dw^0$$

where w_0 is an independent Brownian motion. The problem now is: find $k(t)$ such that $\hat{v} = 0$ minimizes

$$(3.14) \quad E \left\{ \int_0^1 x^2 (1 + k^2) dt + y(1) \right\}$$

subject to

$$(3.15) \quad dx = [(a + bk)x + bv] dt + c dw,$$

$$dy = [2kxv + v^2] dt + dw^0$$

in the class of measurable controls $v(t, x)$ with values in $[-1, 1]$. This is now a partially observable problem because v is to be independent of y . We apply the general necessary conditions to obtain (again $\theta = \theta_0 = -1$ and \mathcal{F}_t is generated by $w_s, s \leq t$)

$$E \left\{ \chi_t \begin{pmatrix} c_t & 0 \\ 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} b_t v - b_t \hat{v}_t \\ 2k_t \hat{x}_t v + v^2 - (2k_t \hat{x}_t \hat{v}_t + \hat{v}_t^2) \end{pmatrix} \middle| \mathcal{F}_t \right\} \cong 0$$

for all $v \in [-1, 1]$. Since $\hat{v} = 0$ it follows that

$$(3.16) \quad E(\chi_t | \mathcal{F}_t) \begin{pmatrix} c_t^{-1} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} b_t \\ 2k_t \hat{x}_t \end{pmatrix} = 0.$$

To compute χ we write (cf. (3.15))

$$\frac{d\Phi}{dt}(t, s) = \begin{pmatrix} a_t + b_t k_t & 0 \\ 0 & 0 \end{pmatrix} \Phi(t, s)$$

(since $\hat{v} \equiv 0$) so that

$$\Phi(t, s) = \begin{pmatrix} \exp \left[\int_t^s a_\tau + b_\tau k_\tau d\tau \right] & 0 \\ 0 & 1 \end{pmatrix}.$$

Also from (3.14)

$$d\mu(s; x, y) = (2(1 + k_s^2)\hat{x}_s ds, \delta_1(s) ds)$$

where $\delta_1(s)$ is the Dirac delta with mass at $s = 1$.

It now follows that

$$\begin{aligned} E(\chi_t | \mathcal{F}_t) \begin{pmatrix} c_t & 0 \\ 0 & 1 \end{pmatrix}^{-1} \\ &= E \left\{ \int_t^1 d\mu(s; \hat{x}, \hat{y}) \Phi(s, t) | \mathcal{F}_t \right\} \\ &= \left(E \left\{ \int_t^1 2(1 + k_s^2) \hat{x}_s \exp \left[\int_t^s a_\tau + b_\tau k_\tau d\tau \right] ds | \mathcal{F}_t \right\}, 1 \right) \\ &= \left(\int_t^1 2(1 + k_s^2) \exp \left[2 \int_t^s a_\tau + b_\tau k_\tau d\tau \right] \hat{x}_t ds, 1 \right), \end{aligned}$$

and hence from (3.16)

$$(3.17) \quad \hat{x}_t \left\{ b_t \int_t^1 2(1 + k_s^2) \exp \left[2 \int_t^s (a_\tau + b_\tau k_\tau) d\tau \right] ds + 2k_t \right\} = 0$$

so that

$$\begin{aligned} \frac{dk_t}{dt} + \frac{db_t}{dt} \int_t^1 (1 + k_s^2) \exp \left[2 \int_t^s a_\tau + b_\tau k_\tau d\tau \right] ds \\ - b_t(1 + k_t^2) - 2b_t \int_t^1 (1 + k_s^2) \exp \left[2 \int_t^s a_\tau + b_\tau k_\tau d\tau \right] (a_t + b_t k_t) ds = 0, \end{aligned}$$

i.e.

$$\frac{dk}{dt} - \frac{db_t}{dt} \frac{k_t}{b_t} - b_t(1 + k_t^2) + 2k_t(a_t + b_t k_t) = 0.$$

Now set $k(t) = -b(t)p(t)$. Then

$$\frac{dp}{dt} = -\frac{dk}{dt} b_t^{-1} + k_t b_t^{-2} \frac{db}{dt} = -(1 + k_t^2) - 2p_t(a_t + b_t k_t)$$

and $p(1) = 0$ since $k(1) = 0$. This is the well known Riccati equation and so the maximum principle does yield the optimal control.

Let us continue further by imposing the constraint $E|\hat{x}(1)| \leq 1$. (Note $\theta_0 = 0$ is impossible here.) Now $\theta = (\theta_0, \theta_1)$ with $\theta_0 \leq 0$, $\theta_1 \leq 0$. Let us rewrite this as $\theta = -(1, \rho)$ with $\rho \geq 0$. Then

$$d\mu(s) = \begin{pmatrix} 2(1 + k_s^2)\hat{x}_s ds & \delta_1(s) ds \\ 2\hat{x}_1 \delta_1(s) ds & 0 \end{pmatrix}$$

and (3.17) becomes

$$\hat{x}_t \left\{ b_t \int_t^1 2(1+k_s^2) \exp \left[2 \int_t^s (a_\tau + b_\tau k_\tau) d\tau \right] ds + 2k_t \right. \\ \left. + 2\rho \exp \int_t^1 (a_\tau + b_\tau k_\tau) d\tau \right\} = 0,$$

i.e.

$$\frac{dk_t}{dt} - \frac{db_t}{dt} \frac{k_t}{b_t} - b_t(1+k_t^2) + 2k_t(a_t + b_t k_t) \\ + \rho(a_t + b_t k_t) \cdot \exp \int_t^1 (a_\tau + b_\tau k_\tau) d\tau = 0$$

and $k_1 = -\rho$. Then with $k(t) = -b(t)p(t)$ we have

$$\frac{dp}{dt} = -1 - 2p_t a_t + b_t^2 p_t^2 + \rho(a_t - b_t^2 p_t) \exp \int_t^1 (a_\tau - b_\tau^2 p_\tau) d\tau, \\ (3.18) \quad p(1) = \rho/b(1).$$

The extra parameter ρ must be found from the condition $E|\hat{x}(1)|^2 \leq 1$. For the simple case $a = 0$, $b(t) = b$, we take $\rho = 0$ if

$$E|\hat{x}(1)|^2 = \int_0^1 c_t^2 \exp \left[2 \int_0^{b(1-t)} \tanh u du \right] dt \leq 1.$$

Otherwise one computes p hence k hence \hat{u} for various values of ρ and then presumably the smallest ρ for which $E|\hat{x}(1; \rho)|^2 \leq 1$ gives \hat{u} . Observe that if $q_t = p_t + \rho \exp \int_t^1 (a_\tau + b_\tau k_\tau) d\tau$ then (3.18) is equivalent to the system of Riccati equations

$$\dot{p} = -1 - pa - qa - b^2 pq, \\ \dot{q} = -1 - 2pa + b^2 p^2,$$

$$p(1) = \rho/b(1), \quad q(1) = \rho(b(1) + 1)/b(1).$$

REFERENCES

- [1] U. G. HAUSSMANN, *General necessary conditions for optimal control of stochastic systems*, Proceedings of the 1975 International Symposium on Stochastic System, Mathematical Programming Studies, 6 (1976), pp. 30-48; also in Modeling, Identification and Optimization, R. Wets, ed., North-Holland, Amsterdam, 1977.
- [2] H. J. KUSHNER, *Necessary conditions for continuous parameter stochastic optimization problems*, this Journal, 11 (1973), pp. 587-594.
- [3] U. G. HAUSSMANN, *Functionals of Itô processes as stochastic integrals*, this Journal, 16 (1978), pp. 252-269.
- [4] V. E. BENEŠ, *Full "bang" to reduced predicted miss is optimal*, this Journal, 14 (1976), pp. 62-84.
- [5] M. MEMIN, *Le théorème de Girsanov*, Séminaires de Mathématiques, Probabilité, U. de Rennes, France, 1972, pp. 173-187.

- [6] M. P. ERSHOV, *Extension of measures and stochastic equations*, Theory Probability Appl., 19 (1974), pp. 431–433.
- [7] K. A. YEN AND CH. YOEURP, *Représentations de martingales comme intégrales stochastiques de processus optionnels*, Séminaire de Probabilités X, Lecture Notes in Mathematics, vol. 511, Springer-Verlag, New York, 1976, pp. 422–432.
- [8] I. V. GIRSANOV, *On transforming a certain class of stochastic processes by absolutely continuous substitution of measures*, Theory Probability Appl., 5 (1960), pp. 285–301.
- [9] J. M. C. CLARK, *The representation of functionals of Brownian motion by stochastic integrals*, Ann. Math. Statist., 41 (1970), pp. 1282–1295.

FUNCTIONALS OF ITÔ PROCESSES AS STOCHASTIC INTEGRALS*

U. G. HAUSSMANN†

Abstract. Conditions are given under which a functional L of an Itô process $z(\cdot)$,

$$(1) \quad z(t) = z_0 + \int_0^t f(s, z) ds + \int_0^t \sigma(s, z) dw, \quad 0 \leq t \leq 1,$$

can be represented as

$$L(z(\cdot, w)) = \int_0^1 \chi(t, w) dw(t, w) \quad \text{w.p. 1,}$$

and an explicit formula for χ is given in terms of the Fréchet derivative of L and the solution of the linearized version of the Itô equation (1). The method of proof consists of applying a theorem of J. M. C. Clark to the Cauchy-Maruyama approximation of (1).

1. Introduction. On the probability space (Ω, \mathcal{F}, P) consider an r dimensional Wiener process $w(t)$ and a process $z(t)$ which satisfies

$$(1.1) \quad z(t) = z_0 + \int_0^t f(s, z) ds + \int_0^t \sigma(s, z) dw,$$

and on C^n , the space of continuous functions mapping $[0, 1]$ in R^n , Euclidean n space, consider a Frechet differentiable functional $L[x]$. Suppose that

$$(1.2) \quad L[z(\cdot, \omega)] = \int_0^1 \chi(t, \omega) dw(t) \quad \text{w.p. 1;}$$

then the problem is to find a "good" representation of χ . If \mathcal{F}_t is the family of σ -algebras generated by w then it is well known that $\chi(t) = d/dt \langle E(L[z]|\mathcal{F}_t), w \rangle_t$ with the usual notation $\langle \cdot \rangle_t$ for the increasing process associated with a martingale. Here $E(x|\mathcal{F}_t)$ denotes the conditional expectation of x given \mathcal{F}_t . For practical purposes one does not wish to talk about the martingale $E(L[z(\cdot)]|\mathcal{F}_t)$; rather one wishes to obtain χ in terms of the functional L , and the coefficients f and σ . This is what is meant by "good". Clark [1] has solved this problem for the case $z = w$, $z_0 = 0$, $f = 0$, $\sigma = I$. We extend his result by approximating z by piecewise smooth processes z_m as in the Maruyama approximation theorem, and then applying Clark's theorem to $L \circ z_m(w)$ before taking the limit in the mean on m . We stress that we always use only Theorem 1 of [1], and never the incorrect Theorem 4.

After stating various hypotheses about L , f , σ in this section, we continue in § 2 with the main result which states that χ can be represented essentially as the conditional expectation of a fundamental matrix solution of the linearization of (1.1). In the Appendix we supply a proof of Maruyama's theorem in the present setting.

Although this result may be of interest by itself, it was established in order to find a useful representation of the adjoint process which arises in the stochastic maximum principle [2]. This application is given in [3].

* Received by the editors November 18, 1976, and in revised form April 28, 1977.

† Department of Mathematics, University of British Columbia, Vancouver, British Columbia, Canada V6T 1W5. This research was supported by the Canada Council under Grant W75001 and by the National Research Council of Canada under Grant A8051.

Let \mathcal{F}_t be the σ -algebra generated by $\{w_s; 0 \leq s \leq t\}$ together with the negligible sets of \mathcal{F} , and let \mathcal{B}_c be the Borel algebra of C^n . Let $\|x\|_t = \sup_{0 \leq s \leq t} |x(s)|$, and set $\|x\| = \|x\|_1$. We shall require the following hypotheses.

(H₁) $L: C^n \rightarrow R^1$ is \mathcal{B}_c measurable, Fréchet differentiable with derivative $(d/dx)L(x)$ at x . Moreover there are K, β, δ positive, finite, such that

$$\left| \frac{d}{dx}L(x) - \frac{d}{dx}L(y) \right| \leq K(1 + \|x\|^\beta)(1 + \|y\|^\beta)\|x - y\|^\delta$$

where, of course, the left side is the norm in C^* , the dual of C^n .

(H₂) $f: [0, 1] \times C^n \rightarrow R^n$, $\sigma: [0, 1] \times C^n \rightarrow R^{n \times r}$ are measurable, and for any adapted process z , $f(t, z)$, $\sigma(t, z)$ are \mathcal{F}_t measurable.

(H₃) $|f(t, x)|^2 + |\sigma(t, x)|^2 \leq K(1 + \|x\|_t^2)$,

$$|f(t, x) - f(t, y)|^2 + |\sigma(t, x) - \sigma(t, y)|^2 \leq K\|x - y\|_t^2.$$

(H₄) $f(t, \cdot)$ and $\sigma(t, \cdot)$ are Frechet differentiable for almost all t , with derivatives $f_x(t, x)$, $\sigma_x(t, x)$ which are Hölder continuous in x of order $\delta > 0$, uniformly in t . From H₃ it follows that the norms of f_x and σ_x are bounded by \sqrt{K} .

Remark 1. From Gronwall's inequality it follows that

$$\|z\|_t^2 \leq K_1 \left(1 + \sup_{0 \leq s \leq t} \left| \int_0^s \sigma dw \right|^2 \right)$$

assuming that $z(0)$ is fixed. Hence for $q \geq 2$

$$\begin{aligned} E\|z\|_t^q &\leq K_2 \left(1 + E \sup_{s \leq t} \left| \int_0^s \sigma dw \right|^q \right) \leq K_3 \left[1 + E \left(\int_0^t |\sigma|^2 ds \right)^{q/2} \right] \\ &\leq K_4 \left(1 + \int_0^t E\|z\|_s^q ds \right) \leq K_5(q) \end{aligned}$$

for $t \in [0, 1]$, by Gronwall's inequality and Burkholder's inequality [4, § 9]. Hölder's inequality now yields

$$(1.3) \quad E\|z\|_t^q \leq K(q), \quad 1 \leq q < \infty.$$

Remark 2. $(d/dx)L(x)$ is a linear functional on C^n , so there exists a right continuous $1 \times n$ matrix valued function $\mu(\cdot, x)$ of bounded variation such that

$$(1.4) \quad \frac{d}{dx}L(x)y = \int_0^1 \mu(dt, x)y(t).$$

Moreover from (H₁) it follows that

$$(1.5) \quad \left| \frac{d}{dx}L(x) \right| \leq K_1(1 + \|x\|^{\beta+\delta}),$$

and that if $R_L(x, y) = L(x + y) - L(x) - (d/dx)L(x)y$ then

$$(1.6) \quad |R_L(x, y)| \leq \|y\|^{1+\delta} K_2(1 + \|x\|^{2\beta})(1 + \|y\|^\beta).$$

We shall also have occasion to apply $(d/dx)L(x)$ to functions $y(t)$ continuous on $[s, 1]$, but equal to zero on $[0, s)$. Such functions are in general not in C^n but we can define

$$(1.7) \quad \frac{d}{dx}L(x)y = \int_s^1 \mu(dt, x)y(t).$$

2. The representation theorem. Under (H_2) and (H_3) , z is the unique \mathcal{F}_t adapted solution of (1.1). If σ^ρ is the ρ th column of σ , define

$$(2.1) \quad dy = f_x(t, z)y \, dt + \sum_{\rho=1} \sigma_x^\rho(t, z)y \, dw^\rho$$

and for $t \geq s$ set $\Phi(t, s)$ to be the $n \times n$ matrix with i th column given by $y_i(t, s)$, the unique solution of (2.1) for $t \geq s$ with $y_i(s, s) = e_i$, the i th standard basis element in R^n , and with $y_i(t, s) = 0$ for $0 \leq t < s$. Note that the convention as in (1.7) is used for f_{xy} and σ_{xy}^ρ . Next define

$$\lambda(t, \omega) = \int_t^1 \mu(ds, z(\omega)) \Phi(s, t, \omega) \sigma(t, z(\omega)).$$

THEOREM 1. Under (H_1) – (H_4)

$$L[z(\cdot)] = \int_0^1 E(\lambda(t) | \mathcal{F}_t) \, dw(t) + EL[z(\cdot)].$$

The proof of this theorem consists of several steps and will occupy the remainder of this section.

Step 1. To apply Clark's result we need to compute the derivative dz/dw , as a continuous linear functional on C^n . Since this does not exist for z as given by (1.1) we approximate z by z_m where z_m is given by Maruyama's theorem (actually x_m^k below). To apply this last theorem, however, requires f and σ to be continuous almost everywhere in t . Hence we define

$$f_m(t, x) = m \int_0^t e^{m(s-t)} f(s, x) \, ds, \quad \sigma_m(t, x) = m \int_0^t e^{m(s-t)} \sigma(s, x) \, ds,$$

and we observe that f_m and σ_m are continuous in t and still satisfy (H_2) – (H_4) . Now let x_m be the unique solution of

$$(2.2) \quad dx = f_m(t, x) \, dt + \sigma_m(t, x) \, dw, \quad t \geq 0, \quad x(0) = z_0.$$

Next we take the Cauchy approximation of x_m as indicated in Maruyama's theorem (see Appendix). Fix k and let $t_j = j/k$, $j = 0, 1, \dots, k$. Then set

$$(2.3) \quad x_m^k(t) = x_m^k(t_j) + f_m(t_j, x_m^k)(t - t_j) + \sigma_m(t_j, x_m^k)(w(t) - w(t_j))$$

if $t_j < t \leq t_{j+1}$, and set

$$x_m^k(0) = z_0.$$

x_m^k is an approximation of x which explicitly exhibits its dependence on w .

Step 2 (Representation of $L[x_m^k]$: verification of Clark's hypotheses). We wish to write $L[x_m^k] = \int_0^1 \chi_m^k \, dw$ and to obtain a reasonable representation for χ_m^k . Clark's Theorem 1 [1] gives such a representation, but we must consider L as a functional of w and verify that it satisfies the hypothesis

$$(2.4) \quad L[x_m^k(w + \tilde{w})] - L[x_m^k(w)] = \frac{d}{dw} L[x_m^k(w)] \tilde{w} + R(w, \tilde{w})$$

where

$$(2.4') \quad |R(w, \tilde{w})| < K_0 \|\tilde{w}\|^{1+\delta} (1 + \|w\|^\alpha) (1 + \|\tilde{w}\|^\alpha)$$

for some finite constants K_0 and α . To this end consider w and \tilde{w} as two elements of C^n

with $w(0) = \tilde{w}(0) = 0$. Let $\tilde{x}_m^k(t) = x_m^k(t; w + \tilde{w}) - x_m^k(t; w)$ where of course both $x_m^k(t; w + \tilde{w})$ and $x_m^k(t; w)$ are defined by (2.3) but for the former w is replaced by $w + \tilde{w}$. Then

$$(2.5) \quad \begin{aligned} \tilde{x}_m^k(t) &= \tilde{x}_m^k(t_j) + (t - t_j)f_{mx}(t_j, x_m^k)\tilde{x}_m^k \\ &+ \sum_{\rho=1}^r (w^\rho(t) - w^\rho(t_j))\sigma_{m\rho}^\rho(t_j, x_m^k)\tilde{x}_m^k + \sigma_m(t_j, x_m^k)(\tilde{w}(t) - (\tilde{w}(t_j))) \\ &+ R_m^k(t, t_j) \quad \text{if } t_j < t \leq t_{j+1}, \\ x_m^k(0) &= 0, \end{aligned}$$

where

$$(2.6) \quad \begin{aligned} R_m^k(t, t_j) &= (t - t_j)\{f_m(t_j, x_m^k + \tilde{x}_m^k) - f_m(t_j, x_m^k) - f_{mx}(t_j, x_m^k)\tilde{x}_m^k\} \\ &+ \sum_{\rho} (w^\rho(t) - w^\rho(t_j))\{\sigma_m^\rho(t_j, x_m^k + \tilde{x}_m^k) - \sigma_m^\rho(t_j, x_m^k) - \sigma_{m\rho}^\rho(t_j, x_m^k)\tilde{x}_m^k\} \\ &+ \{\sigma_m(t, x_m^k + \tilde{x}_m^k) - \sigma_m(t, x_m^k)\}(\tilde{w}(t) - \tilde{w}(t_j)). \end{aligned}$$

We set $v_m^k(t) = R_m^k(t, t_j)$ for $t_j < t \leq t_{j+1}$ and similarly $b_m^k(t) = \sigma_m(t_j, x_m^k)(\tilde{w}(t) - \tilde{w}(t_j))$. Then define $c_m^k(t) = b_m^k(t) + v_m^k(t)$, and

$$\bar{x}_j(t) = \begin{cases} \tilde{x}_m^k(t) & \text{if } t \leq t_j, \\ \tilde{x}_m^k(t_j) & \text{if } t > t_j. \end{cases}$$

Similarly define $\bar{c}(t)$ from $c_m^k(t)$, $\bar{b}_j(t)$ from $b_m^k(t)$ and $\bar{v}_j(t)$ and from $v_m^k(t)$. Now (2.5) can be rewritten as

$$(2.7) \quad \bar{x}_{j+1} = M_{j+1}\bar{x}_j + \bar{c}_{j+1}$$

where M_j is a bounded linear operator on C^n . If $\Phi_{jj} = I$, $\Phi_{ji} = M_j M_{j-1} \cdots M_{i+1}$ for $j > i$, then from (2.7)

$$(2.8) \quad x_j = \sum_{i=1}^j \Phi_{ji} \bar{c}_i$$

because $\bar{x}_0 = 0$. Hence

$$\begin{aligned} L[x_m^k(w + \tilde{w})] - L[x_m^k(w)] &= \int_0^1 \mu(dt, x_m^k) \bar{x}_k(t) + R_L(x_m^k, \bar{x}_k) \\ &= \int_0^1 \mu(dt, x_m^k) \sum_{i=1}^k (\Phi_{ki} \bar{b}_i)(t) \\ &\quad + \int_0^1 \mu(dt, x_m^k) \sum_{i=1}^k (\Phi_{ki} \bar{v}_i)(t) + R_L(x_m^k, \bar{x}_k). \end{aligned}$$

We call the sum of the last two terms $R(w, \tilde{w})$. Since $\Phi_{ki} \bar{b}_i$ is linear in \tilde{w} , then (2.4) is satisfied if R satisfies (2.4)'. This requires an analysis of x_m^k and \bar{x}_k .

From (2.3) and (H₃) it follows that

$$\|x_m^k\|_{t_{j+1}} \leq \|x_m^k\|_{t_j} + [K(1 + \|x_m^k\|_{t_j}^2)]^{1/2}(k^{-1} + 2\|w\|)$$

or

$$(2.9) \quad \|x_m^k\| \leq k_1(1 + \|w\|^{k+1}),$$

and

$$\|\tilde{x}_m^k\|_{t_{j+1}} \leq \|\tilde{x}_m^k\|_{t_j} (1 + K(k^{-1} + 2\|w\|)) + 2K\|\tilde{w}\|(1 + \|x_m^k(w + \tilde{w})\|)$$

because

$$\begin{aligned} \tilde{x}_m^k(t) &= \tilde{x}_m^k(t_j) + (f_m(t_j, x_m^k + \tilde{x}_m^k) - f_m(t_j, x_m^k))(t - t_j) \\ &\quad + (\sigma_m(t_j, x_m^k + \tilde{x}_m^k) - \sigma_m(t_j, x_m^k))(w(t) - w(t_j)) \\ &\quad + \sigma_m(t_j, x_m^k + \tilde{x}_m^k)(\tilde{w}(t) - \tilde{w}(t_j)) \quad \text{if } t_j < t \leq t_{j+1}. \end{aligned}$$

Hence from (2.9) it follows that

$$\begin{aligned} \|\tilde{x}_m^k\| &\leq k_2\|\tilde{w}\|(1 + \|w\|^k)(1 + \|w + \tilde{w}\|^{k+1}) \\ (2.10) \quad &\leq k_3\|\tilde{w}\|(1 + \|w\|^{2k+1})(1 + \|\tilde{w}\|^{k+1}). \end{aligned}$$

Moreover from (2.7) it follows that if $t_i < t \leq t_{i+1}$, $i \leq j$, then

$$(M_{j+1}x)(t) = x(t_i) + (t - t_i)f_{mx}(t_i, x_m^k)x + \sum_{\rho} (w^{\rho}(t) - w^{\rho}(t_i))\sigma_{mx}^{\rho}(t_i, x_m^k)x,$$

so that (using (H₄))

$$|M_{j+1}| \leq k_4(1 + \|w\|)$$

and

$$(2.11) \quad |\Phi_{ki}| \leq k_5(1 + \|w\|^k).$$

From (2.6), (H₃) and (H₄), it follows that

$$\begin{aligned} \|v_m^k\| &\leq K\|\tilde{x}_m^k\|^{1+\delta}(k^{-1} + 2\|w\|) + 2K\|\tilde{x}_m^k\|\|\tilde{w}\| \\ &\leq k_6\|\tilde{w}\|^{1+\delta}(1 + \|w\|^{\alpha})(1 + \|\tilde{w}\|^{\alpha}). \end{aligned}$$

Now (2.4)' follows from (1.5), (1.6), (2.9), (2.10) and (2.11), and we have established

LEMMA 2.1. $L[x_m^k(w)]$ satisfies (2.4), (2.4)' with

$$(2.12) \quad \frac{dL}{dw}[x_m^k(w)]\tilde{w} = \int_0^1 \mu(dt, x_m^k) \sum_{i=1}^k (\Phi_{ki}\tilde{b}_i)(t).$$

Step 3 (Representation of $L[x_m^k]$: "finite" f_{mx}, σ_{mx}). Now we wish to write the right side of (2.12) as $\int_0^1 \lambda_m^k(dt, w)\tilde{w}(t)$, since Clark's theorem then tells us that

$$(2.13) \quad L[x_m^k(w)] = \int_0^1 E\{\lambda_m^k(1, w) - \lambda_m^k(t, w) | \mathcal{F}_t\} dw(t)$$

assuming as we always shall that $EL[x_m^k] = 0$. To identify λ_m^k we begin by considering the simpler case where f_{mx} and σ_{mx} act on \tilde{x} only at a finite number of times. Then we shall take limits. Assume for the moment that

$$\begin{aligned} f_{mx}(t_j, x_m^k) &= \sum_{i=0}^{j-1} \sum_{l=1}^{l(i)} \alpha_{il}^i \tilde{x}(t_i^l), \\ \sigma_{mx}^{\rho}(t_j, x_m^k) &= \sum_{i=0}^{j-1} \sum_{l=1}^{l(i)} \beta_{il}^{\rho} \tilde{x}(t_i^l) \end{aligned}$$

where $t_i = t_i^0 < t_i^1 < \dots < t_i^{l(i)} = t_{i+1}$. We shall denote the generic point of this partition by t^j , $j = 0, 1, \dots, l_0$ with $l_0 = \sum_{i=0}^{k-1} l(i)$. α_{il}^i and β_{il}^{ρ} are constant matrices. For the statement

of the next lemma we need some more notation. Consider the equations

$$\begin{aligned}
 \lambda_i(t) &= \lambda_i(t_j) + (t - t_j) f_{mx}(t_j, x_m^k) \lambda_i(\cdot) \\
 &\quad + \sum_{\rho} (w^{\rho}(t) - w^{\rho}(t_j)) \sigma_{mx}^{\rho}(t_j, x_m^k) \lambda_i(\cdot), \quad t^i \leq t_j < t \leq t_{j+1}, \\
 \lambda_i(t) &= \sigma_m(t_j, x_m^k), \quad t_j < t^i \leq t \leq t_{j+1}, \\
 \lambda_i(t) &= 0, \quad 0 \leq t < t^i.
 \end{aligned}
 \tag{2.14}$$

LEMMA 2.2. $\lambda_m^k(t) = \nu(t) + \bar{\lambda}(t)$ where

$$\begin{aligned}
 \nu(t) &= - \int_t^1 \mu(ds, x_m^k) \bar{\sigma}_m(s, x_m^k), \\
 \bar{\sigma}_m(t, x_m^k) &= \sigma_m(t_j, x_m^k) \quad \text{if } t_j < t \leq t_{j+1},
 \end{aligned}
 \tag{2.15}$$

and $\bar{\lambda}(t)$ is piecewise constant with jumps at $t = t^i$ of size

$$\begin{aligned}
 \bar{\lambda}(t^i) - \bar{\lambda}(t^i) &= - \int_{t^{i+1}}^1 \mu(dt, x_m^k) \lambda_{i+1}(t) \\
 &\quad + \int_{t^i}^1 \mu(dt, x_m^k) \lambda_i(t) - (\mu(t^{i+1}, x_m^k) \\
 &\quad - \mu(t^i, x_m^k)) \bar{\sigma}_m(t^i, x_m^k).
 \end{aligned}
 \tag{2.16}$$

Proof. For convenience we set $w(t) = w_t$, $w(t^l) = w_l$, $\tilde{w}(t^l) = \tilde{w}_l$, $(\sum_{i=1}^k \Phi_{ki} \bar{b}_i)(t) = y(t)$, $y(t^l) = y_l$, and $\sigma_l = \bar{\sigma}_m(t^l, x_m^k)$. Then from (2.5), (2.7) and (2.8) it follows that for $t^l < t \leq t^{l+1}$

$$y(t) = y_l + (t - t^l) \sum_{i=1}^l \bar{\alpha}_{li} y_i + \sum_{\rho} (w_t^{\rho} - w_l^{\rho}) \sum_{i=1}^l \bar{\beta}_{li}^{\rho} y_i + \sigma_l (\tilde{w}_t - \tilde{w}_l),
 \tag{2.17}$$

where $\bar{\alpha}$ and $\bar{\beta}$ are derived from α and β . Hence we can write

$$y_{l+1} = \sum_{i=1}^l \gamma_{l+1i} y_i + b_{l+1}$$

where

$$\begin{aligned}
 \gamma_{li} &= (t^l - t^{l-1}) \bar{\alpha}_{l-1i} + \sum_{\rho} (w_t^{\rho} - w_{l-1}^{\rho}) \bar{\beta}_{l-1i}^{\rho} + \delta_{l-1i}, \\
 b_l &= \sigma_{l-1} (\tilde{w}_l - \tilde{w}_{l-1}), \\
 \delta_{li} &= \begin{cases} I & \text{if } l = i, \\ 0 & \text{if } l \neq i. \end{cases}
 \end{aligned}$$

If we define

$$\begin{aligned}
 \Psi_{li} &= \sum_{l > k_1 > k_2 > \dots > k_{\nu} > i} \gamma_{lk_1} \gamma_{k_1 k_2} \gamma_{k_2 k_3} \dots \gamma_{k_{\nu} i}, \quad l > i, \\
 \Psi_{ii} &= I,
 \end{aligned}$$

then

$$y_l = \sum_{i=1}^l \Psi_{li} b_i,
 \tag{2.18}$$

and for $t^l < t \leq t^{l+1}$

$$\begin{aligned}
 y(t) &= \sum_{i=1}^l \Psi_{li} \sigma_{i-1} (\tilde{w}_i - \tilde{w}_{i-1}) + (t - t^l) \sum_{j=1}^l \tilde{\alpha}_{lj} \sum_{i=1}^j \Psi_{ji} \sigma_{i-1} (\tilde{w}_i - \tilde{w}_{i-1}) \\
 (2.19) \quad &+ \sum_{\rho} (w_t^{\rho} - w_l^{\rho}) \sum_{j=1}^l \tilde{\beta}_{lj}^{\rho} \sum_{i=1}^j \Psi_{ji} \sigma_{i-1} (\tilde{w}_i - \tilde{w}_{i-1}) + \sigma_l (\tilde{w}_t - \tilde{w}_l) \\
 &= \sum_{i=1}^l \left\{ \Psi_{li} \sigma_{i-1} + a_{li} (t - t^l) + \sum_{\rho} (w_t^{\rho} - w_l^{\rho}) b_{li}^{\rho} \right\} (\tilde{w}_i - \tilde{w}_{i-1}) + \sigma_l (\tilde{w}_t - \tilde{w}_l)
 \end{aligned}$$

where

$$a_{li} = \sum_{j=1}^l \tilde{\alpha}_{lj} \Psi_{ji} \sigma_{i-1}, \quad b_{li}^{\rho} = \sum_{j=1}^l \tilde{\beta}_{lj}^{\rho} \Psi_{ji} \sigma_{i-1}.$$

Again for convenience we set $\mu(t, x_m^k) = \mu(t)$, $\Delta_l \mu = \mu(t^{l+1}) - \mu(t^l)$, $\Delta_l \mu_0 = \int_{t^l}^{t^{l+1}} t d\mu(t)$, $\Delta_l \mu_{\rho} = \int_{t^l}^{t^{l+1}} d\mu(t) w^{\rho}(t)$. Then from (2.12) and (2.19)

$$\begin{aligned}
 \int_0^1 \lambda_m^k(dt, w) \tilde{w}(t) &= \int_0^1 d\mu(t) y(t) \\
 &= \sum_{l=0}^{l_0} \left[\Delta_l \mu \sum_{i=1}^l \left\{ \Psi_{li} \sigma_{i-1} - t^l a_{li} - \sum_{\rho} w_l^{\rho} b_{li}^{\rho} \right\} (\tilde{w}_i - \tilde{w}_{i-1}) \right. \\
 &\quad \left. + \Delta_l \mu_0 \sum_{i=1}^l a_{li} (\tilde{w}_i - \tilde{w}_{i-1}) + \sum_{\rho} \Delta_l \mu_{\rho} \sum_{i=1}^l b_{li}^{\rho} (\tilde{w}_i - \tilde{w}_{i-1}) \right. \\
 &\quad \left. + \int_{t^l}^{t^{l+1}} d\mu(t) \sigma_l (\tilde{w}_t - \tilde{w}_l) \right] \\
 &= \sum_{i=1}^{l_0} \left[- \sum_{l=i+1}^{l_0} \left\{ \Delta_l \mu \left(\Psi_{li+1} \sigma_i - t^l a_{li+1} - \sum_{\rho} w_l^{\rho} b_{li+1}^{\rho} \right) \right. \right. \\
 &\quad \left. \left. + \Delta_l \mu_0 a_{li+1} + \sum_{\rho} \Delta_l \mu_{\rho} b_{li+1}^{\rho} \right\} \tilde{w}_i \right. \\
 &\quad \left. + \sum_{l=i}^{l_0} \left\{ \Delta_l \mu (\Psi_{li} \sigma_{i-1} - t^l a_{li} - \sum_{\rho} w_l^{\rho} b_{li}^{\rho}) \right. \right. \\
 &\quad \left. \left. + \Delta_l \mu_0 a_{li} + \sum_{\rho} \Delta_l \mu_{\rho} b_{li}^{\rho} \right\} \tilde{w}_i - \Delta_i \mu \sigma_i \tilde{w}_i \right] \\
 &\quad + \int_0^1 d\mu(t) \bar{\sigma}_m(t, x_m^k) \tilde{w}(t).
 \end{aligned}$$

Clearly the last term of this last expression is $\int_0^1 d\nu(t) \tilde{w}(t)$, and the rest is the integral of \tilde{w} with respect to a piecewise constant integrator having jumps at t^i , $i = 1, 2, \dots, l_0$, of magnitude

$$(2.20) \quad - \int_{t^{i+1}}^1 d\mu(t) \xi_{i+1}(t) + \int_{t^i}^1 d\mu(t) \xi_i(t) - \Delta_i \mu \sigma_i.$$

Here, for $l \geq i$, $t^l < t \leq t^{l+1}$,

$$\begin{aligned} \xi_i(t) &= \Psi_{li}\sigma_{i-1} + (t - t^l)a_{li} + \sum_{\rho} (w_t^{\rho} - w_{t^l}^{\rho})b_{li}^{\rho} \\ (2.21) \quad &= \Psi_{li}\sigma_{i-1} + (t - t^l) \sum_{j=1}^l \bar{\alpha}_{lj} \Psi_{ji}\sigma_{i-1} + \sum_{\rho} (w_t^{\rho} - w_{t^l}^{\rho}) \sum_{j=i}^l \bar{\beta}_{lj}^{\rho} \Psi_{ji}\sigma_{i-1} \end{aligned}$$

upon substituting for a , b ; cf. (2.19). Moreover from (2.21) for $l \geq i$

$$(2.22) \quad \xi_i(t^{l+1}) = \sum_{j=i}^l \gamma_{l+1,j} \Psi_{ji}\sigma_{i-1} = \Psi_{l+1,i}\sigma_{i-1},$$

and we define $\xi_i(t^i) = \alpha_{i-1}$ to make ξ_i right continuous on $[t^i, 1]$. By comparing (2.17), (2.21) and (2.22) we conclude that ξ_i satisfies (2.14), if we observe that $\Psi_{li} = I$ for $t_j < t^i \leq t^l \leq t_{j+1}$. Hence $\xi_i(t) = \lambda_i(t)$, and the lemma is established.

Step 4 (Representation of $L[x_m^k]$: general f_{mx}, σ_{mx}). We can now consider the case of general f_{mx}, σ_{mx} . For any $s \in [0, 1]$ define $\lambda_{ms}^k(t)$ (similar to 2.14) by

$$(2.23) \quad \lambda_{ms}^k(t) = \begin{cases} 0 & \text{if } 0 \leq t < s, \\ \sigma_m(t_j, x_m^k) & \text{if } t_j < s \leq t \leq t_{j+1}, \\ \lambda_{ms}^k(t_j) + (t - t_j)f_{mx}(t_j, x_m^k)\lambda_{ms}^k(\cdot) \\ \quad + \sum_{\rho} (w^{\rho}(t) - w^{\rho}(t_j))\sigma_{mx}^{\rho}(t_j, x_m^k)\lambda_{ms}^k(\cdot) & \text{if } s \leq t_j < t \leq t_{j+1}. \end{cases}$$

LEMMA 2.3.

$$\frac{dL}{dw}[x_m^k(w)]\tilde{w} = \int_0^1 \lambda_m^k(dt, w)\tilde{w}(t)$$

where

$$(2.24) \quad \lambda_m^k(t, w) = - \int_t^1 \mu(dt, x_m^k(w))\lambda_{mi}^k(s).$$

Proof. For notational purposes we set $f_{mx} = \sigma_{mx}^0$, $t = w^0(t)$, and, as before, $\sum_{i=1}^k (\Phi_{ki}\bar{b}_i)(t) = y(t)$. For each (t_j, x_m^k) , $\rho = 0, 1, \dots, r$, there exists a right continuous function $\eta^{\rho j}(t)$ of bounded variation such that

$$\sigma_{mx}^{\rho}(t_j, x_m^k)y = \int_0^1 d\eta^{\rho j}(t)y(t).$$

Hence for any partition $\pi = \{t^0, t^1, t^2, \dots, t^{l_0}, 1\}$ with $t^0 = 0$, containing $\{0, t_1, t_2, \dots, 1\}$,

$$\int_0^1 d\eta^{\rho j}(t)y(t) = \sum_{i=0}^{l_0} (\eta^{\rho j}(t^{i+1}) - \eta^{\rho j}(t^i))y(t^i) + s^{\rho j}(y)$$

where $s^{\rho j}$ is an error term. Then it follows (cf. (2.17)) that $y(0) = 0$ and for $t_j \leq t^l < t \leq t^{l+1}$

$$(2.25) \quad \begin{aligned} y(t) &= y(t^l) + \sum_{\rho=0}^r (w^{\rho}(t) - w^{\rho}(t^l)) \sum_{i=1}^l \bar{\beta}_{li}^{\rho} y_i \\ &\quad + \sigma_l(\tilde{w}(t) - \tilde{w}(t^l)) + \sum_{\rho=0}^r (w^{\rho}(t) - w^{\rho}(t^l))s^{\rho j}(y) \end{aligned}$$

with the appropriate identification of $\bar{\beta} = (\bar{\alpha}, \bar{\beta}^1, \bar{\beta}^2, \dots, \bar{\beta}^\rho)$. We write $s_i(y)$ for the matrix with columns $s^{\rho j}(y)$, $\rho = 0, 1, \dots, r$, if $t_j \leq t^l < t_{j+1}$. Then (cf. (2.18))

$$y_{l+1} = \sum_{i=1}^l \Psi_l(\sigma_{i-1}(\tilde{w}_i - \tilde{w}_{i-1}) + s_{i-1}(y)(w_i - w_{i-1}))$$

where we recall that $y_i = y(t^i)$, $\tilde{w}_i = \tilde{w}(t^i)$, $w_i = w(t^i)$. Hence

$$\int_0^1 d\mu(t)y(t) = \int_0^1 d\lambda_\pi(t)\tilde{w}(t) + \int_0^1 d\lambda'_\pi(t)w(t)$$

where $\lambda_\pi(t) = \lambda_m^k(t)$ as given in Lemma 2.2 with $\bar{\beta}$ as in (2.25), and where $\lambda'_\pi(t) = \lambda_m^k(t)$ as given in Lemma 2.2 with $\bar{\beta}$ as in (2.25), and with $\sigma_m(t_j, x_m^k)$ replaced by $s_j(y)$ in (2.14)–(2.16). The proof now consists of showing that $\lambda'_\pi \rightarrow 0$ and $\lambda_\pi \rightarrow \lambda_m^k$.

Consider λ'_π . The part given by (2.15) is $\nu'(t) = -\int_t^1 d\mu(u)\bar{s}(u, y)$ if $\bar{s}(u, y) = s_i(y)$ for $t^l < u \leq t^{l+1}$. If $V(m)$ is the variation of m then

$$V(\nu') \leq \sup_{0 \leq j < k} |\bar{s}(t_j, y)| V(\mu) \rightarrow 0$$

as mesh $\pi \rightarrow 0$ because \bar{s} is constant on $(t_j, t_{j+1}]$. Moreover from (2.16)

$$V(\lambda') \leq 2 \sum_{i=0}^{l_0} \left| \int_{t^i}^1 d\mu(t)\lambda'_i(t) \right| + \sup_{0 \leq j < k} |\bar{s}(t_j, y)| V(\mu).$$

Since σ_{mx}^ρ is adapted, it follows that $\eta^{\rho j}(t) = \eta^{\rho j}(t_j)$ for $t \geq t_j$. Define $\Delta_i \eta^{\rho j} = \eta^{\rho j}(t^{i+1}) - \eta^{\rho j}(t^i)$, and $i(j) = l$ if $t^l = t_j$. Then from (2.14) it follows that

$$\lambda'_i(t) = \lambda'_i(t_j) + \sum_{\rho=0}^r (w^\rho(t) - w^\rho(t_j)) \sum_{l=0}^{i(j)-1} \Delta_l \eta^{\rho j} \lambda'_i(t^l)$$

if $t^i \leq t_j < t \leq t_{j+1}$, and

$$\lambda'_i(t) = \begin{cases} \bar{s}(t_j, y) & \text{if } t_j < t^i \leq t \leq t_{j+1}, \\ 0 & \text{if } 0 \leq t < t^i. \end{cases}$$

We define $V(\eta) = [\sum_{\rho=0}^r \sup_{0 \leq j < k} V(\eta^{\rho j})^2]^{1/2}$, and we define τ_i and $j(i)$ by $t_{j(i)} < t^i \leq t_{j(i)+1} = \tau_i$. Then

$$\sup_{\tau_i \leq t \leq \tau_{i+1}} |\lambda'_i(t)| \leq \sup_{\tau_i \leq t \leq t_j} |\lambda'_i(t)|(1 + 2\|w\|V(\eta))$$

and hence

$$\|\lambda'_i\| \leq \sup_{0 \leq j < k} |\bar{s}(t_j, y)| K(\omega)$$

where $K(\omega)$ is a random variable, finite w.p. 1, although it may change from equation to equation. We can now conclude that $V(\lambda') \rightarrow 0$ w.p. 1 as mesh $\pi \rightarrow 0$, and hence

$$\frac{dL}{dw} \tilde{w} = \int_0^1 d\mu(t)y(t) = \lim_{\pi \rightarrow 0} \int_0^1 d\lambda_\pi(t)\tilde{w}(t) \quad \text{w.p. 1.}$$

To compute this last limit, observe first that

$$\begin{aligned} & \int_0^1 d\mu(t) \bar{\sigma}_m(t) \tilde{w}(t) - \sum_{i=0}^{l_0} \Delta_i \mu \sigma_i \tilde{w}_i \\ &= \sum_{j=0}^{k-1} \left\{ \int_{t_j}^{t_{j+1}} d\mu(t) \sigma_m(t_j, x_m^k) \tilde{w}(t) - \sum_{i=i(j)}^{i(j+1)-1} \Delta_i \mu \sigma_m(t_j, x_m^k) \tilde{w}_i \right\} \\ & \rightarrow 0 \end{aligned}$$

as mesh $\pi \rightarrow 0$ because \tilde{w} is continuous. Hence

$$\begin{aligned} (2.26) \quad I^k &\equiv \lim_{\text{mesh } \pi \rightarrow 0} \left| \int_0^1 d(\lambda_\pi(t) - \lambda_m^k(t)) \tilde{w}(t) \right| \\ &= \lim_{\text{mesh } \pi \rightarrow 0} \left| \sum_{i=0}^{l_0} \left\{ - \int_{t^{i+1}}^1 d\mu(t) \lambda_{i+1}(t) + \int_{t^i}^1 d\mu(t) \lambda_i(t) \right. \right. \\ &\quad \left. \left. + \int_{t^{i+1}}^1 d\mu(t) \lambda_{m^{i+1}}^k(t) - \int_{t^i}^1 d\mu(t) \lambda_{m^i}^k(t) \right\} \tilde{w}(t^i) + \varepsilon_1 \right| \end{aligned}$$

where

$$\begin{aligned} \varepsilon_1 &= \int_0^1 d\lambda_m^k(t) \tilde{w}(t) - \sum_{i=0}^{l_0} (\lambda_m^k(t^{i+1}) - \lambda_m^k(t^i)) \tilde{w}(t^i) \\ &\rightarrow 0 \text{ as mesh } \pi \rightarrow 0 \end{aligned}$$

because $\tilde{w}(t)$ is continuous and $V(\lambda_m^k) < \infty$. To establish this last inequality we first need
CLAIM 1.

$$\sup_{t^{i+1} \leq t \leq 1} |\lambda_{m^i}^k(t) - \lambda_{m^{i+1}}^k(t)| \leq K(\omega) \left(\delta_j^i + \sum_{j=0}^r \sum_{\rho=0}^r \|w^\rho\| |\Delta_i \eta^{\rho j}| \right),$$

where $\delta_j^i = 0$ except when $t^i = t_{j(i)+1}$, when it is 1.

This claim is true because for $t_j \leq t^{i+1}$

$$\begin{aligned} & \sup_{t^{i+1} \leq t \leq t_{j+1}} |\lambda_{m^i}^k(t) - \lambda_{m^{i+1}}^k(t)| \\ & \leq \sup_{t^{i+1} \leq t \leq t_j} |\lambda_{m^i}^k(t) - \lambda_{m^{i+1}}^k(t)| \{1 + 2\|w\| V(\eta)\} \\ & \quad + |\sigma_m(t_{j(i)}, x_m^k)| 2 \sum_{\rho=0}^r \|w^\rho\| |\Delta_i \eta^{\rho j}|, \end{aligned}$$

and

$$|\lambda_{m^i}^k(t) - \lambda_{m^{i+1}}^k(t)| = \begin{cases} 0 & \text{if } t^i < t^{i+1} \leq t \leq t_{j(i)+1} \\ |\sigma_m(t_{j(i)}, x_m^k) - \sigma_m(t_{j(i)+1}, x_m^k)| & \text{if } t^i = t_{j(i)+1} < t^{i+1} \leq t \leq t_{j(i)+2}. \end{cases}$$

We now observe that

$$\begin{aligned} \sum_{i=0}^{l_0} |\lambda_m^k(t^{i+1}) - \lambda_m^k(t^i)| &= \sum_{i=0}^{l_0} \left| - \int_{t^{i+1}}^1 d\mu(t) \lambda_{m^{i+1}}^k(t) + \int_{t^i}^1 d\mu(t) \lambda_{m^i}^k(t) \right| \\ &\leq \sum_{i=0}^{l_0} \left| \int_{t^i}^{t^{i+1}} d\mu(t) \lambda_{m^i}^k(t) \right| + \sum_{i=0}^{l_0} \left| \int_{t^{i+1}}^1 d\mu(t) (\lambda_{m^i}^k(t) - \lambda_{m^{i+1}}^k(t)) \right| \\ &\leq V(\mu) K(\omega) + V(\mu) k(K(\omega) + \|w\| V(\eta)) \end{aligned}$$

where we have used Claim 1 and the fact that $\|\lambda_{m^i}^k\|$ can be bounded independently of i . Hence indeed $V(\lambda_m^k) < \infty$.

Returning to (2.26) we can write it as

$$(2.27) \quad I^k = \lim_{\text{mesh } \pi \rightarrow 0} \sum_{i=0}^{l_0} \left| \left\{ \int_{t^i}^{t^{i+1}} d\mu(t)(\lambda_i(t) - \lambda_{m^i}^k(t)) + \int_{t^{i+1}}^1 d\mu(t)(\lambda_i(t) - \lambda_{i+1}(t) - \lambda_{m^i}^k(t) + \lambda_{m^{i+1}}^k(t)) \right\} \tilde{w}(t^i) \right|.$$

CLAIM 2.

$$\lim_{\text{mesh } \pi \rightarrow 0} \sup_i \|\lambda_i - \lambda_{m^i}^k\| = 0.$$

Proof. For $t_j < s \leq t \leq t_{j+1}$, $\lambda_{ms}^k(t) = \sigma_m(t_j, x_m^k)$, and for $t_j < s \leq t_{j+1} < t \leq t_{j+2}$

$$\lambda_{ms}^k(t) = \sigma_m(t_j, x_m^k) + \sum_{\rho=0}^r (w^\rho(t) - w^\rho(t_{j+1})) \int_s^{t_{j+1}} d\eta^{\rho(j+1)} \sigma_m(t_j, x_m^k)$$

so that $\|\lambda_i - \lambda_{m^i}^k(t)\|_{t_j(i)+2} = 0$. Moreover for $t^i \leq t_j < t \leq t_{j+1}$,

$$\begin{aligned} |\lambda_i(t) - \lambda_{m^i}^k(t)| &\leq |\lambda_i(t_j) - \lambda_{m^i}^k(t_j)| \\ &+ \sum_{\rho=0}^r \left\{ |w^\rho(t) - w^\rho(t_j)| \left| \sum_{l=0}^{i(j)-1} \Delta_l \eta^{\rho j} (\lambda_i(t^l) - \lambda_{m^i}^k(t^l)) \right. \right. \\ &\quad \left. \left. + \sum_{l=0}^{i(j)-1} \Delta_l \eta^{\rho j} \lambda_{m^i}^k(t^l) - \sigma_{mx}^\rho(t_j, x_m^k) \lambda_{m^i}^k(t) \right| \right\}, \end{aligned}$$

from which it follows that

$$\begin{aligned} \|\lambda_i - \lambda_{m^i}^k\|_{t_{j+1}} &\leq \|\lambda_i - \lambda_{m^i}^k\|_{t_j} \{1 + 2\|w\|V(\eta)\} \\ &+ 2 \sum_{\rho=0}^r \|w^\rho\| |s^{\rho j}(\lambda_{m^i}^k)|, \end{aligned}$$

and hence

$$(2.28) \quad \|\lambda_i - \lambda_{m^i}^k\| < K(\omega) \sup_{0 \leq j < k} \sum_{\rho} |s^{\rho j}(\lambda_{m^i}^k)|.$$

Now

$$\begin{aligned} |s^{\rho j}(\lambda_{m^i}^k)| &= \left| \int_{t^i}^{t_j} d\eta^{\rho j}(t) \lambda_{m^i}^k(t) - \sum_{l=i}^{i(j)-1} \Delta_l \eta^{\rho j} \lambda_{m^i}^k(t^l) \right| \\ &\leq V(\eta^{\rho j}) \sup_{\substack{t^i \leq t, \tau \leq t_j \\ |t-\tau| < \text{mesh } \pi}} |\lambda_{m^i}^k(t) - \lambda_{m^i}^k(\tau)| \\ &\leq V(\eta^{\rho j}) 2 \sum_{\rho=0}^r m_\rho(\text{mesh } \pi) \sup_{0 \leq j < k} |\sigma_{mx}^\rho(t_j, x_m^k)(\lambda_{m^i}^k)| \\ &\leq K(\omega) \sum_{\rho=0}^r m_\rho(\text{mesh } \pi) \end{aligned}$$

where

$$m_\rho(h) = \sup_{\substack{0 \leq t, \tau \leq 1 \\ |t-\tau| < h}} |w^\rho(t) - w^\rho(\tau)|.$$

This establishes the claim.

From (2.27) we can now see that the proof of the lemma will be complete if

CLAIM 3.

$$\sum_{i=0}^{l_0} \sup_{t^{i+1} \leq t \leq 1} |\lambda_i(t) - \lambda_{i+1}(t) - \lambda_{m^i}^k(t) + \lambda_{m^{i+1}}^k(t)| \rightarrow 0$$

as mesh $\pi \rightarrow 0$.

Proof. As in the derivation of (2.28) we have

$$(2.29) \quad \begin{aligned} & \|\lambda_i - \lambda_{m^i}^k - \lambda_{i+1} + \lambda_{m^{i+1}}^k\|_{t_{i+1}} \\ & \leq \|\lambda_i - \lambda_{m^i}^k - \lambda_{i+1} + \lambda_{m^{i+1}}^k\|_{t_i} \{1 + 2\|w\|V(\eta)\} \\ & \quad + \sum_{\rho=0}^r \|w^\rho\| |s^{\rho j}(\lambda_{m^i}^k - \lambda_{m^{i+1}}^k)|, \end{aligned}$$

with $\|\lambda_i - \lambda_{m^i}^k - \lambda_{i+1} + \lambda_{m^{i+1}}^k\|_{t_{i(i)+2}} = 0$. We always have $t_{j(i)+2} \geq t^{i+1}$. Moreover

$$(2.30) \quad \begin{aligned} |s^{\rho j}(\lambda_{m^i}^k - \lambda_{m^{i+1}}^k)| &= \left| \int_{t^{i+1}}^{t_j} d\eta^{\rho j}(t)(\lambda_{m^i}^k(t) - \lambda_{m^{i+1}}^k(t)) \right. \\ & \quad \left. - \sum_{l=i+1}^{i(j)-1} \Delta_i \eta^{\rho j}(\lambda_{m^i}^k(t^l) - \lambda_{m^{i+1}}^k(t^l)) \right| \\ &\leq V(\eta^{\rho j}) \sup_{\substack{t^{i+1} \leq t, \tau \leq 1 \\ |t-\tau| < \text{mesh } \pi}} |\lambda_{m^i}^k(t) - \lambda_{m^{i+1}}^k(t) - \lambda_{m^i}^k(\tau) + \lambda_{m^{i+1}}^k(\tau)|. \end{aligned}$$

Note that this follows directly from the fact that $\lambda_{m^i}^k$ is constant on $[t^i, t_{j(i)+1}]$ if $t^{i+1} \leq t_{j(i)+1}$. In the case $t^{i+1} > t_{j(i)+1}$, we observe that $t^i = t_{j(i)+1}$, and then $\lambda_{m^i}^k$ is constant on $[t^i, t_{j(i)+2}]$ so that the equality (2.30) again holds. Now if $|t - \tau| < h$, then

$$\begin{aligned} & |\lambda_{m^i}^k(t) - \lambda_{m^{i+1}}^k(t) - \lambda_{m^i}^k(\tau) + \lambda_{m^{i+1}}^k(\tau)| \\ & \leq 2 \sum_{\rho=0}^r m_\rho(h) \sup_{0 \leq j < k} |\sigma_{mx}^\rho(t_j, x_m^k)(\lambda_{m^i}^k - \lambda_{m^{i+1}}^k)| \\ & \leq 2 \sum_{\rho=0}^r m_\rho(h) \sup_{0 \leq j < k} \left\{ V(\eta^{\rho j})K(\omega) \left(\delta_j^i + \sum_{\rho=0}^r \|w^\rho\| |\Delta_i \eta^{\rho j}| \right) \right. \\ & \quad \left. + |\Delta_i \eta^{\rho j}| \sigma_m(t_{j(i)}, x_m^k) \right\} \\ & \leq K'(\omega) \sum_{\rho=0}^r m_\rho(h) \left(\delta_j^i + \sup_{0 \leq j < k} |\Delta_i \eta^{\rho j}| + \sup_{0 \leq j < k} \sum_{\rho'=0}^r |\Delta_i \eta^{\rho' j}| \right) \end{aligned}$$

where we have used Claim 1. Hence

$$\begin{aligned} & \sum_{i=0}^{l_0} \sum_{\rho=0}^r \sup_{0 \leq j < k} |s^{\rho j}(\lambda_{m^i}^k - \lambda_{m^{i+1}}^k)| \leq V(\eta)K(\omega) \sum_{\rho=0}^r m_\rho(\text{mesh } \pi) \\ & \rightarrow 0 \quad \text{as mesh } \pi \rightarrow 0. \end{aligned}$$

Hence from (2.29) it follows that the claim is true and the proof is complete.

COROLLARY.

$$L[x_m^k] = - \int_0^1 E(\lambda_m^k(t) | \mathcal{F}_t) dw(t).$$

Proof. This follows from (2.13).

Step 5 ($k \rightarrow \infty$). The next step is to let $k \rightarrow \infty$. We define λ_{ms} by

$$(2.31) \quad \begin{aligned} \lambda_{ms}(t) &= 0 \quad \text{if } t < s, \\ \lambda_{ms}(s) &= \sigma_m(s, x_m), \end{aligned}$$

$$d\lambda_{ms}(t) = f_{mx}(t, x_m)\lambda_{ms} dt + \sum_{\rho=1}^r dw^\rho \sigma_{mx}^\rho(t, x_m)\lambda_{ms} \quad \text{if } s < t \leq 1,$$

where as usual the convention is that $f_{mx}\lambda_{ms}$ means the application of f_{mx} in $C^n[s, 1]$, the space of continuous functions on $[s, 1]$, i.e. $\int_s^1 d\eta^0(t)\lambda_{ms}(t)$.

LEMMA 2.4.

$$\lim_{k \rightarrow \infty} E\|\lambda_{ms}^k - \lambda_{ms}\| = 0.$$

Proof. If $t_{j_0} < s \leq t_{j_0+1}$, set

$$\tilde{\lambda}_{ms}^k(t) = \begin{cases} \lambda_{ms}^k(t) - \sigma_m(t_{j_0}, x_m^k) & \text{if } t \geq s, \\ 0 & \text{if } t < s. \end{cases}$$

Then in fact for $t_j < t \leq t_{j+1}$, $j = 0, 1, 2, \dots, k$,

$$\begin{aligned} \tilde{\lambda}_{ms}^k(t) &= \tilde{\lambda}_{ms}^k(t_j) + (t - t_j)H_s(t_j)f_{mx}(t_j, x_m^k)(\tilde{\lambda}_{ms}^k(\cdot) + \sigma_m(t_{j_0}, x_m^k)) \\ &\quad + \sum_{\rho=1}^r (w^\rho(t) - w^\rho(t_j))H_s(t_j)\sigma_{mx}^\rho(t_j, x_m^k)(\tilde{\lambda}_{ms}^k(\cdot) + \sigma_m(t_{j_0}, x_m^k)), \\ \tilde{\lambda}_{ms}^k(0) &= 0, \end{aligned}$$

where $H_s(t) = 0$ if $t < s$, $H_s(t) = 1$ if $t \geq s$. Now define $\bar{\lambda}_{ms}^k$ by

$$\begin{aligned} \bar{\lambda}_{ms}^k(t) &= \bar{\lambda}_{ms}^k(t_j) + (t - t_j)H_s(t_j)f_{mx}(t_j, x_m^k)(\bar{\lambda}_{ms}^k(\cdot) + \sigma_m(s, x_m^k)) \\ &\quad + \sum_{\rho=1}^r (w^\rho(t) - w^\rho(t_j))\sigma_{mx}^\rho(t_j, x_m^k)(\bar{\lambda}_{ms}^k(\cdot) + \sigma_m(s, x_m^k)) \quad \text{if } t_j < t \leq t_{j+1}, \\ \bar{\lambda}_{ms}^k(0) &= 0. \end{aligned}$$

Since $v(t) \equiv \bar{\lambda}_{ms}^k(t) - \tilde{\lambda}_{ms}^k(t)$ satisfies a stochastic differential equation, an application of (H_4) and Burkholder's inequality [4, § 9] yields

$$E\|v\|_t^2 \leq K_1 \int_0^t E\|v\|_\tau^2 d\tau + K_1 \int_0^t E|\Delta\sigma_m|^2 d\tau$$

where $\Delta\sigma_m = \sigma_m(s, x_m^k) - \sigma_m(t_j, x_m^k)$. Since

$$(2.32) \quad E|\Delta\sigma_m|^2 \leq 4\|\sigma(\cdot, x_m^k)\|^2 m^2 k^{-2}$$

then from Gronwall's inequality it follows that

$$(2.33) \quad \lim_{k \rightarrow \infty} E\|\tilde{\lambda}_{ms}^k - \bar{\lambda}_{ms}^k\|^2 = 0.$$

If we define $\bar{\lambda}_{ms}$ by

$$(2.34) \quad \begin{aligned} \bar{\lambda}_{ms}(t) &= 0 \quad \text{if } t \leq s, \\ d\bar{\lambda}_{ms}(t) &= f_{mx}(t, x_m)(\bar{\lambda}_{ms}(\cdot) + \sigma_m(s, x_m)) dt \\ &\quad + \sum_{\rho=1}^r dW^\rho \sigma_{m\rho}(t, x_m)(\bar{\lambda}_{ms}(\cdot) + \sigma_m(s, x_m)) \quad \text{if } t > s, \end{aligned}$$

then $\bar{\lambda}_{ms}$ satisfies a stochastic differential equation, with coefficients continuous in $(\bar{\lambda}_{ms}, x_m)$ except at $t = s$. Hence an application of Maruyama's theorem (see Appendix) yields

$$(2.35) \quad \lim_{k \rightarrow \infty} E \| (x_m^k, \bar{\lambda}_{ms}^k) - (x_m, \bar{\lambda}_{ms}) \|^2 = 0.$$

From (2.31) and (2.34) it follows that

$$\lambda_{ms}(t) = \begin{cases} \bar{\lambda}_{ms}(t) + \sigma_m(s, x_m) & \text{if } t \geq s, \\ 0 & \text{if } t < s, \end{cases}$$

so that

$$\begin{aligned} E \| \lambda_{ms}^k - \lambda_{ms} \|^2 &\leq \{ E \| \bar{\lambda}_{ms}^k - \bar{\lambda}_{ms} \|^2 + E \| \bar{\lambda}_{ms}^k - \bar{\lambda}_{ms} \|^2 \\ &\quad + E |\Delta \sigma_m|^2 + E |\sigma_m(s, x_m^k) - \sigma_m(s, x_m)|^2 \} 4 \\ &\rightarrow 0 \quad \text{as } k \rightarrow \infty \end{aligned}$$

by (2.33), (2.35), (2.32) and (H₃).

LEMMA 2.5.

$$L[x_m] = - \int_0^1 E(\lambda_m(t) | \mathcal{F}_t) dw(t) \quad \text{w.p. 1.}$$

where

$$\lambda_m(t) = - \int_t^1 d\mu(s; x_m) \lambda_{mt}(s).$$

Proof. From Remarks 1 and 2 and (2.35) it follows that

$$E | L[x_m^k] - L[x_m] |^2 \rightarrow 0,$$

and hence there exists a function Q_m , adapted and square integrable, such that

$$\int_0^1 E | Q_m(t) - E(\lambda_m^k(t) | \mathcal{F}_t) |^2 dt \rightarrow 0;$$

cf. [1, Lemma 1]. But from (H₁), ($L' = dL/dx$)

$$\begin{aligned} |\lambda_m^k(t) - \lambda_m(t)| &= \left| \int_t^1 d\mu(s; x_m) \lambda_{mt}(s) - \int_t^1 d\mu(s; x_m^k) \lambda_{mt}^k(s) \right| \\ &\leq \left| \int_t^1 d\mu(s; x_m) (\lambda_{mt}(s) - \lambda_{mt}^k(s)) \right| + \left| \int_t^1 d(\mu(s; x_m) - \mu(s; x_m^k)) \lambda_{mt}^k(s) \right| \\ &\leq |L'[x_m]| \|\lambda_{mt} - \lambda_{mt}^k\| + K(1 + \|x_m\|^\beta)(1 + \|x_m^k\|^\beta) \|\lambda_{mt}^k\| \|x_m - x_m^k\|^\delta, \end{aligned}$$

so from Lemma 2.4, Remark 1, and (2.35) it follows that for each t

$$E|E(\lambda_m^k(t) - \lambda_m(t)|\mathcal{F}_t)|^2 \leq E|\lambda_m^k(t) - \lambda_m(t)|^2 \rightarrow 0,$$

and $E\|\lambda_m^k - \lambda_m\|^2 < \infty$. Hence $E(\lambda_m^k(t)|\mathcal{F}_t)$ converges to $E(\lambda_m(t)|\mathcal{F}_t)$ in $L_2 dt dp$, so that $Q_m(t) = E(\lambda_m(t)|\mathcal{F}_t)$, and the result follows.

We observe that strictly speaking one ought to carry along $EL[x_m^k]$. However clearly this converges to $EL[x_m]$ as $k \rightarrow \infty$.

Step 6 ($m \rightarrow \infty$). The proof of the theorem can now be completed by letting $m \rightarrow \infty$. Again $EL[x_m] \rightarrow EL[z]$ because of the following lemma.

LEMMA 2.6.

$$\lim_{m \rightarrow \infty} E\|x_m - z\|^2 = 0.$$

Proof. If $y_m(t) \equiv x_m(t) - z(t)$, then

$$\begin{aligned} y_m(t) &= \int_0^t [f_m(s, x_m) - f_m(s, z)] ds + \int_0^t [f_m(s, z) - f(s, z)] ds \\ &\quad + \int_0^t [\sigma_m(s, x_m) - \sigma_m(s, z)] dw + \int_0^t [\sigma_m(s, z) - \sigma(s, z)] dw. \end{aligned}$$

Using (H_3) and (1.3) it follows that

$$E\|y_m\|_t^2 \leq 8K \int_0^t \|y_m\|_s^2 ds + 8\varepsilon$$

where m is chosen so large that

$$\begin{aligned} E \int_0^1 |f_m(s, z) - f(s, z)|^2 ds &< \varepsilon \\ E \int_0^1 |\sigma_m(s, z) - \sigma(s, z)|^2 ds &< \varepsilon. \end{aligned}$$

Now it follows from Gronwall's inequality that

$$E\|y_m\| \leq 8\varepsilon e^{8K}$$

and the proof is complete.

Now we define

$$\begin{aligned} \lambda_s(t) &= 0 \quad \text{if } t < s, \\ \lambda_s(s) &= \sigma(s, z), \\ d\lambda_s(t) &= f_x(t, z)\lambda_s dt + \sum_{\rho=1}^r dw^\rho \sigma_x^\rho(t, z)\lambda_s, \quad t \geq s. \end{aligned}$$

If we set $y_m = \lambda_{ms} - \lambda_s$, then $y_m(t) = 0$ for $s > t$, and

$$\begin{aligned} y_m(t) &= \sigma_m(s, x_m) - \sigma(s, z) + \int_s^t f_{mx}(\tau, x_m)y_m d\tau \\ &\quad + \int_s^t (f_{mx}(\tau, x_m) - f_x(\tau, z))\lambda_s d\tau + \int_s^t \sigma_{mx}(\tau, x_m)y_m dw \\ &\quad + \int_s^t (\sigma_{mx}(\tau, x_m) - \sigma_x(\tau, z))\lambda_s dw, \quad s \leq t. \end{aligned}$$

However

$$(2.36) \quad \begin{aligned} |\sigma_m(s, x_m) - \sigma(s, z)|^2 &\leq 2K\|x_m - z\|_s^2 + 2|\sigma_m(s, z) - \sigma(s, z)|^2, \\ |(f_{mx}(\tau, x_m) - f_x(\tau, z))\lambda_s|^2 &\leq 2K\|x_m - z\|_\tau^{2\delta}\|\lambda_s\|^2 + 2|f_{mx}(\tau, z)\lambda_s - f_x(\tau, z)\lambda_s|^2, \end{aligned}$$

and similarly for $|\sigma_{mx} - \sigma_x|^2$. Considering Remark 1, applied to x_m, z, λ_s , we observe that the terms on the left sides of (2.36) tend to zero in mean square $dt dP$ as $m \rightarrow \infty$. Hence as in the proof of Lemma 2.6

$$\lim_{m \rightarrow \infty} E\|\lambda_{ms} - \lambda_s\|^2 = 0 \quad \text{a.e. s.}$$

Finally the argument of Lemma 2.5 can be repeated to yield

$$(2.37) \quad L[z] = \int_0^1 E\left(\int_t^1 d\mu(s; z)\lambda_t(s)|\mathcal{F}_t\right) dw(t) \quad \text{w.p. 1.}$$

This completes the proof of the theorem.

We remark that the result can be extended. If $L = L(x, w)$ then dL/dw has two components; L_x is treated as above and L_w is treated as in [1] if L satisfies (H_1) in w as well as x . If, on the other hand, $f = f(t, x, w)$, $\sigma = \sigma(t, x, w)$ and (H_2) – (H_4) are satisfied in w as well as x , and if we represent

$$f_w(t, x, w)\tilde{w} = \int_0^t \nu_f(ds; t, x, w)\tilde{w}(s), \quad \sigma_w(t, x, w)w = \int_0^t \nu_\sigma(ds; t, x, w)\tilde{w}(s)$$

then $L[z]$ is still as given in (2.37) but

$$\begin{aligned} \lambda_t(s) &= \Phi(s, t)\sigma(t, z, w) - \int_t^s \Phi(s, \tau)\nu_f(t; \tau, z, w) d\tau \\ &\quad - \int_t^s \Phi(s, \tau)\nu_\sigma(t; \tau, z, w) dw_\tau, \end{aligned}$$

i.e.

$$\begin{aligned} d\lambda_t(s) &= (f_x(s, z, w)\lambda_t(s) - \nu_f(t; s, z, w)) ds \\ &\quad + \sum_\rho dw^\rho (\sigma_x^\rho(s, z, w)\lambda_t(s) - \nu_\sigma(t; s, z, w)), \\ \lambda_t(t) &= \sigma(t, z, w). \end{aligned}$$

Appendix. Consider functions $a(t, x)$ and $b(t, x)$, continuous on $[0, 1]$ except at s_0 , satisfying (H_2) and for some $\delta > 0$

$$(A.1) \quad |a(t, x)|^2 + |b(t, x)|^2 \leq K(1 + \|x\|_t^2),$$

$$(A.2) \quad |a(t, x) - a(t, y)|^2 + |b(t, x) - b(t, y)|^2 \leq K((1 + \|x\|_t^2)\|x - y\|_t^{2\delta} + \|x - y\|_t^2).$$

If x satisfies

$$(A.3) \quad x(t) = x_0 + \int_0^t a(s, x) ds + \int_0^t b(s, x) dw,$$

then as in Remark 1, for each $q < \infty$,

$$E\|x\|^q \leq K(q),$$

where $K(q)$ depends only on q and on the constant K in (A.1). For any partition $\pi = \{0 = t_0 < t_1 < \dots < t_k = 1\}$ of mesh $m(\pi)$ set

$$(A.4) \quad x^\pi(t) = x^\pi(t_j) + a(t_j, x^\pi)(t - t_j) + b(t_j, x^\pi)(w(t) - w(t_j)), \quad t_j < t \leq t_{j+1}, \\ x^\pi(0) = x_0.$$

THEOREM (Maruyama). $E\|x^\pi - x\|^2 \rightarrow 0$ as $m(\pi) \rightarrow 0$.

Proof. We follow essentially McShane's proof [5] with some modifications. Set

$$X(t_j) = x_0 + \sum_{i=0}^{j-1} a(t_i, x) \Delta_i t + \sum_{i=0}^{j-1} b(t_i, x) \Delta_i w,$$

$$\tilde{X}(t) = X(t_j), \quad t_j \leq t < t_{j+1},$$

$$\tilde{x}^\pi(t) = x^\pi(t_j), \quad t_j \leq t < t_{j+1},$$

$$\tilde{x}(t) = x(t_j), \quad t_j \leq t < t_{j+1}.$$

Then

$$(A.5) \quad E\|x^\pi - x\|_t^2 \leq 4E\|x^\pi - \tilde{x}^\pi\|_t^2 + 4E\|\tilde{x}^\pi - \tilde{X}\|_t^2 + 4E\|\tilde{X} - \tilde{x}\|_t^2 + 4E\|x - \tilde{x}\|_t^2.$$

Since $E\|x^\pi\|^2 \leq K$, independent of π , then $P(\|x^\pi\|^2 \geq M^2) \leq K/M^2$. Also

$$|x^\pi(t) - \tilde{x}^\pi(t)| \leq \sqrt{K} \sqrt{1 + \|x^\pi\|^2} \left(m(\pi) + \sup_{\substack{0 \leq t, \tau \leq 1 \\ |t - \tau| < m(\pi)}} |w(t) - w(\tau)| \right),$$

and $w(t)$ is uniformly continuous w.p. 1. Hence for any ε , $\eta > 0$

$$P\{\|x^\pi - \tilde{x}^\pi\| > \sqrt{K(1 + M^2)}\varepsilon\} < K/M^2 + \eta$$

if $m(\pi) < \varepsilon/2$ and $m(\pi)$ is also so small that $P(\sup |w(t) - w(\tau)| > \varepsilon/2) < \eta$. This implies that $\|x^\pi - \tilde{x}^\pi\| \rightarrow 0$ in probability as $m(\pi) \rightarrow 0$. Also $\|x^\pi - \tilde{x}^\pi\| \leq 2\|x^\pi\|$, $E\|x^\pi\|^{2+q} < K(2+q)$, independent of π , so by the Lebesgue theorem $E\|x^\pi - \tilde{x}^\pi\|^2 \rightarrow 0$.

If $t_j \leq t < t_{j+1}$, then

$$\begin{aligned} \|\tilde{x}^\pi - \tilde{X}\|_t^2 &= \max_{i \leq j} |x^\pi(t_i) - X(t_i)|^2 \\ &\leq 2 \max_{i \leq j} \left\{ \left| \sum_{l=0}^i (a(t_l, x^\pi) - a(t_l, x)) \Delta_l t \right|^2 + \left| \sum_{l=0}^i (b(t_l, x^\pi) - b(t_l, x)) \Delta_l w \right|^2 \right\} \\ &\leq 2 \int_0^t K((1 + \|x\|_s^2) \|x^\pi - x\|_s^{2\delta} + \|x^\pi - x\|_s^2) ds \\ &\quad + 2 \max_{i \leq j} \left| \sum_{l=0}^i (b(t_l, x^\pi) - b(t_l, x)) \Delta_l w \right|^2 \end{aligned}$$

Taking expectations and applying Burkholder's inequality [4, § 9] to the last sum considered as an integral yields

$$E\|\tilde{x}^\pi - \tilde{X}\|_t^2 \leq K \int_0^t E\{(1 + \|x\|_s^2) \|x^\pi - x\|_s^{2\delta}\} ds \leq K_0 \int_0^t E\|x^\pi - x\|_s^2 ds$$

if we take into account Remark 1. Note that as usual K is a constant but not necessarily the same one from equation to equation.

Next we define, for $t_j \leq t \leq t_{j+1}$

$$X(t) = X(t_j) + a(t_j, x)(t - t_j) + b(t_j, x)(w(t) - w(t_j))$$

so that

$$\begin{aligned} \|\tilde{X} - \tilde{x}\|^2 &= \max_i |X(t_i) - x(t_i)|^2 \\ &\leq \sup_{0 \leq t \leq 1} |X(t) - x(t)|^2 = \sup_{0 \leq t \leq 1} \left| \int_0^1 \tilde{a}(t, x) dt + \int_0^1 \tilde{b}(t, x) dw \right|^2 \end{aligned}$$

where $\tilde{a}(t, x) = a(t_j, x) - a(t, x)$ if $t_j < t \leq t_{j+1}$, and $\tilde{b}(t, x, \omega) = b(t_j, x) - b(t, x)$ if $t_j < t \leq t_{j+1}$. Hence

$$E\|\tilde{X} - \tilde{x}\|^2 < K \left\{ \int_0^1 E|\tilde{a}(t, x)|^2 dt + \int_0^1 E|\tilde{b}(t, x)|^2 dt \right\}$$

if Burkholder's inequality is again used. However a and b are continuous in t a.e. so that $|\tilde{a}(t, x(\omega))| + |\tilde{b}(t, x(\omega))| \rightarrow 0$ a.e. (t, ω) as $m(\pi) \rightarrow 0$. Moreover

$$\int_0^1 E(|\tilde{a}|^2 + |\tilde{b}|^2) dt \leq 2 \int_0^1 E(|a|^2 + |b|^2) dt < \infty$$

by (A.1) and Remark 1. Hence by dominated convergence $E\|\tilde{X} - \tilde{x}\|^2 \rightarrow 0$.

If we now choose $m(\pi)$ so small that

$$E\|\tilde{X} - \tilde{x}\|^2 < \varepsilon, \quad E\|x^\pi - \tilde{x}^\pi\|^2 < \varepsilon, \quad E\|x - \tilde{x}\|^2 < \varepsilon$$

then from (A.5)

$$E\|x^\pi - x\|_t^2 \leq 12\varepsilon + 4K_0 \int_0^t E\|x^\pi - x\|_s^2 ds.$$

An application of Gronwall's inequality now completes the proof. We observe that x_m , as defined in (2.2) satisfies (A.3), as does the pair $(x_m, \bar{\lambda}_{ms})$.

REFERENCES

- [1] J. M. C. CLARK, *The representation of functionals of Brownian motion by stochastic integrals*, Ann. Math. Statist., 41 (1970), pp. 1282–1295; 42 (1971), p. 1778.
- [2] U. G. HAUSSMANN, *General necessary conditions for optimal control of stochastic systems*, Proceedings of the 1975 International Symposium on Stochastic System; Mathematical Programming Studies, 6 (1976), pp. 30–48 and in Modeling, Identification and Optimization, North-Holland, Amsterdam, 1976.
- [3] ———, *On the stochastic maximum principle*, this Journal, 16 (1978), pp. 236–251.
- [4] D. L. BURKHOLDER, *Distribution function inequalities for martingales*, Ann. Probability, 1 (1973), pp. 19–42.
- [5] E. J. MCSHANE, *Stochastic Calculus and Stochastic Models*, Academic Press, New York, 1974.

ON POLE ASSIGNMENT FOR A CLASS OF INFINITE DIMENSIONAL LINEAR SYSTEMS*

AVRAHAM FEINTUCH† AND MOSHE ROSENFELD†

Abstract. Let $\{A, B\}$ be an infinite dimensional linear system where A is normal with compact resolvent and the input space is finite dimensional. Then the controllability of $\{A, B\}$ implies the possibility of assigning an arbitrary finite set of poles to the transfer function of the closed loop system formed by means of suitable linear state feedback. Thus such systems can always be stabilized through state feedback.

1. Introduction. Let \mathcal{H} be an infinite dimensional Hilbert space and consider the linear system

$$(1) \quad \dot{x}(t) = Ax(t) + Bu(t).$$

We assume that A is an unbounded linear normal operator with compact resolvent and that it is the infinitesimal generator for the C_0 semigroup $\{T(t) = e^{At}\}$. B will be an operator from a finite dimensional Hilbert space E_n into \mathcal{H} whose range will be denoted by \mathcal{M} .

Let F be a linear operator defined on \mathcal{H} with range in E_n and consider the closed loop system defined by

$$(2) \quad u = Fx + v.$$

Equation (2) is called a state feedback law and combining (1) and (2) gives

$$(3) \quad \dot{x}(t) = (A + BF)x(t) + Bv(t).$$

It is usually desirable to choose F such that stability is obtained in (2).

In the finite dimensional case the relationship between controllability and stabilization is well-known. Since in this case stability is determined purely by the location of the eigenvalues of the state operator, the feedback F is usually chosen so that the eigenvalues of $A + BF$ are contained in the open left half-plane $\{z: \operatorname{Re} z < 0\}$. The main result in this direction is that of Wonham [8].

THEOREM A. *If \mathcal{H} is finite dimensional then the pair $\{A, B\}$ is controllable if and only if, for every symmetric set Λ of n complex numbers, there exists a feedback operator F such that $\sigma(A + BF) = \Lambda$.*

In attempting to generalize this result to the infinite dimensional case one becomes immediately aware of a number of difficulties. The first of these is that the spectrum of an operator on infinite dimensional space consists, generally, of more than eigenvalues. It is well-known that by a perturbation of the form BF where B is finite dimensional, one can hope to move only the point spectrum and nothing else.

The second of these problems arises from the fact that position of the spectrum of A is not necessarily related to the stability of (1). There is the classical counterexample given in [11] of a semigroup with $\sigma(A)$ empty such that $\|T(t)\| = e^{\pi t/2}$.

The problem of stabilization for controllable linear systems in the infinite dimensional case was first considered by Slemrod [10] and later by Triggiani [9]. While Slemrod considered a more general problem than that considered here he was forced to introduce a controllability condition which, while in the finite dimensional case reduces to the usual condition, is much stronger than the one considered here. Also, he

* Received by the editors May 24, 1976, and in revised form May 16, 1977.

† Department of Mathematics, Ben Gurion University of the Negev, Beersheva 84 120, Israel.

considered only the problem of stabilization without being concerned with the placement of the poles of the transfer function of the closed loop feedback system. While Triggiani used essentially the same definition of controllability as ours and considered operators with compact resolvent (as we do) without the assumption of normality, he restricted himself to the case where only finitely many eigenvalues of A lie in the right half plane and was then able to reduce the problem to the finite dimensional case. Here we make no such assumption.

We point out that our hypothesis on A that it has compact resolvent removes both of the difficulties mentioned previously for infinite dimensional problems. Firstly, the spectrum of A in the finite plane consists only of eigenvalues and secondly the stabilization problem for (1) is equivalent to the question of the location of eigenvalues of $A + BF$ (see [9]). Thus this seems to be the natural type of situation to consider for a generalization of Wonham's result.

2. Preliminaries. We have assumed that A generates a C_0 semigroup. By the Hille–Yosida Theorem [11] this means that there exist positive real numbers M and β such that for every real $\lambda > \beta$, $\lambda \in \rho(A)$ (the resolvent set of A) and

$$\|(I\lambda - A)^{-n}\| \leq \frac{M}{(\lambda - \beta)^n}, \quad n = 1, 2, \dots$$

Then $(Is - A)^{-1}$ exists for all complex s with $\operatorname{Re} s > \beta$ and is given by

$$(Is - A)^{-1}x = \int_0^\infty e^{-st} e^{At} x \, dt.$$

If T is any bounded linear transformation, then

$$[(\lambda I - (A + T))^{-1} = (\lambda I - A - T)^{-1} = R(\lambda, A)[I - TR(\lambda, A)]^{-1}$$

where $R(\lambda, A) = (\lambda I - A)^{-1}$. Since for λ sufficiently large,

$$\|TR(\lambda, A)\| \leq \gamma < 1,$$

it follows that λ belongs to the resolvent set of $A + T$.

Following [1], the set of all states reachable in time interval t , starting with zero initial state, that is the set of elements:

$$x(t) = \int_0^t e^{A(t-s)} Bu(s) \, ds,$$

will be denoted by $\Omega(t)$.

DEFINITION. The system (1) is *controllable* if $\bigcup_t \Omega(t)$ is dense in \mathcal{H} .

As is well-known this is equivalent to the condition

$$\bigvee_{t \geq 0} e^{At} \mathcal{M} = \mathcal{H}$$

(i.e. the closed linear span of the subspaces $e^{At} \mathcal{M}$) where \mathcal{M} is the range of B .

3. Discrete operators.

DEFINITION. An operator A is *discrete* if there is a number $\lambda \in \rho(A)$ for which $R(\lambda, A)$ is compact.

Many important properties of discrete operators have been studied in detail in [2, Chap. 19]. We make use of a result proven there.

LEMMA 1. *If A is discrete, then:*

- (a) *its spectrum is a denumerable set of points with no finite limit point;*
- (b) *the resolvent $R(\lambda, A)$ is compact for every $\lambda \in \rho(A)$;*
- (c) *if B is a bounded operator, then $A + B$ is discrete.*

4. Controllability and feedback. Let \mathcal{M} denote the range of the input operator B and consider the system (1). Suppose we are free to modify (1) by setting

$$u(t) = Fx(t) + v(t)$$

where v is a new external input and $F: \mathcal{H} \rightarrow E_n$ (the input space). We refer to F as the state feedback. The obvious result of introducing state feedback is to change the pair $\{A, B\}$ in (1) into the pair $\{A + BF, B\}$.

The first step in our procedure for pole assignment will be to reduce the problem to the case where B is a rank one operator. This will allow us to replace the multi-input system (1) by a single input system while preserving controllability.

THEOREM 1. *Let A be a discrete normal operator all of whose eigenvalues are of multiplicity one. If $\{A, B\}$ is controllable, there is a vector $b \in \mathcal{M}$ such that $\{A, b\}$ is controllable. That is $\bigvee_{t \geq 0} e^{At}b = \mathcal{H}$.*

Proof. We will use an idea introduced by P. Fuhrmann in [6, Thm. 6.1]. Since there are no magnitude restrictions on the control variable, we can assume, without loss of generality, that $\{T(t) = e^{At}\}$ is a semigroup of contractions. Let $T = (A + 1)(A - I)^{-1}$ be the Cayley transform of A . Then T is compact, normal. Consider the discrete system

$$(4) \quad x_n = Tx_{n-1} + Bu_{n-1}.$$

Then (1) is controllable if and only if (4) is [6, Thm. 6.1].

The controllability condition for discrete systems takes the form

$$\bigvee_{n=0}^{\infty} T^n \mathcal{M} = \mathcal{H}.$$

Since T is compact and normal, there exists an orthonormal basis $\{e_n\}_{n=1}^{\infty}$ such that $Te_n = \lambda_n e_n$ for some sequence $\{\lambda_n\}$, and the assumption on the multiplicity of the eigenvalues of A implies $\lambda_n \neq \lambda_m$ for $n \neq m$.

Let $\{x_1, \dots, x_k\}$ be a basis for \mathcal{M} . Then $x_i = \sum \alpha_{in} e_n$ and the controllability condition implies that for each n , $\alpha_{in} \neq 0$ for at least one i . It follows that there exist numbers β_1, \dots, β_n such that if $b = \sum \beta_i x_i$ has the representation $b = \sum \alpha_n e_n$, then $\alpha_n \neq 0$ for all n . Thus b is a cyclic vector for T and $\{T, b\}$ is controllable. Again using Fuhrmann's result we obtain that $\{A, b\}$ is controllable.

Remark. The finite dimensional version of this theorem was originally proved by Wonham [8]. In [7] Fuhrmann extended this result to the infinite dimensional case when A is self-adjoint. An examination of Fuhrmann's argument shows that it also holds for A normal with every invariant subspace of A reducing. Since this is true for compact normal operators, Theorem 1 could actually be derived from Fuhrmann's theorem. It is natural to ask whether this is true without any assumptions on A . In [3] and [4] examples were given to show that this is not the case.

Now suppose A has eigenvalues of multiplicity greater than one. We note that since A is discrete the multiplicity of the eigenvalues of A is bounded by some integer K . By rearranging the basis we can decompose A into a finite direct sum of operators each of whose eigenvalues are of multiplicity one. Since the eigenvectors of A reduce A , the system can be broken up into a finite number of systems and each one can be considered separately. Thus it is no loss of generality to assume that the eigenvalues of

A have multiplicity one and we will do so from now on. Thus by Theorem (1) we can replace (1) by the system

$$(1') \quad \dot{x}(t) = Ax(t) + u(t)b$$

where $u(t)$ is a scalar valued function and b is a fixed vector in \mathcal{H} .

For vectors $b, c \in \mathcal{H}$, $b \otimes c$ will denote the rank one operator

$$(b \otimes c)x = (x, c)b.$$

Thus for (1') choosing a state feedback means choosing a vector c and replacing (1') by the system

$$\dot{x} = (A + b \otimes c)x + vb.$$

Our next result is a generalization of Lemma 4 of [8, 2.1].

THEOREM 2. *For any $c \in \mathcal{H}$, $\{A, b\}$ is controllable if and only if $\{A + b \otimes c, b\}$ is controllable.*

Proof. Again it is easier to work with discrete systems. By multiplying by a scalar, if, necessary, we can assume that A and $A + b \otimes c$ generate contraction semigroups.

Let T denote the Cayley transform of A . The Cayley transform of $A + b \otimes c$ is

$$\begin{aligned} (A + b \otimes c + I)(A + b \otimes c - I)^{-1} &= (A + I)(A + b \otimes c - I)^{-1} + b \otimes c(A + b \otimes c - I)^{-1} \\ &= -(A + I)(I - A)^{-1}V - b \otimes c(I - A)^{-1}V \end{aligned}$$

where $V = [I - b \otimes c(I - A)^{-1}]^{-1}$. Thus the Cayley transform of $A + b \otimes c$ is

$$-[TV + b \otimes c(I - A)^{-1}V].$$

We must show that if $\{T, b\}$ is controllable, then so is $\{TV + b \otimes c(I - A)^{-1}V, b\}$. If $Vb = y$, then

$$\begin{aligned} b &= V^{-1}y = [I - b \otimes c(I - A)^{-1}]y \\ &= y - ((I - A)^{-1}y, c)b. \end{aligned}$$

Thus $y = b + ((I - A)^{-1}y, c)b$ and $TVb = Ty = [I + ((I - A)^{-1}y, c)]Tb$. If $((I - A)^{-1}y, c) = -1$ then $y = 0$, which is impossible. Thus $TVb = \alpha Tb$ for some $\alpha \neq 0$.

A similar computation leads to the fact that $VTb = \alpha_1 b + \alpha_2 Tb$ for some $\alpha_1, \alpha_2 \neq 0$. Continuing in this manner we obtain that $\mathcal{H} = \bigvee_{n=0}^{\infty} T^n b \subset \bigvee_{n=0}^{\infty} (TV)^n b$, which implies that $\{TV, b\}$ is controllable. Since $(TV + b \otimes c(I - A)^{-1}V)^n b = (TV)^n b + \alpha_n b$ for some α_n , it follows that $\{TV + b \otimes c(I - A)^{-1}V, b\}$ is controllable. Thus $\{A, b\}$ controllable implies $\{A + b \otimes c, b\}$ controllable.

The other direction is trivial. Suppose $\{A + b \otimes c, b\}$ is controllable. By the previous argument, $\{[A + b \otimes c] + b \otimes (-c), b\}$ is controllable. But this is just $\{A, b\}$ and the proof is complete.

Remark. The same argument (with slightly messier computations) can be used for the multi-input case.

5. The main results. An examination of Wonham's proof of Theorem A shows that it is strongly geometric and depends very strongly on the finite dimensionality of the state space. A coordinate free proof of Wonham's theorem was given in [5]. This related the pole assignment procedure to the procedure of assigning the points where a rational function attains the value 1. This is the approach that we will generalize here.

DEFINITION. Let f be a meromorphic function in the complex plane. The 1-points of f are the points $\{z_n\}$ where $f(z) = 1$.

THEOREM 3. Suppose A is discrete, and $b, c \in \mathcal{H}$ such that $\sigma(A + b \otimes c) \cap \sigma(A) = \emptyset$. Then $\lambda \in \sigma(A + b \otimes c)$ if and only if it is a 1-point of the meromorphic function $((zI - A)^{-1}b, c)$.

Proof. Suppose $(A + b \otimes c)x = \lambda x$. This is equivalent to $Ax + (x, c)b = \lambda x$ or $(\lambda - A)^{-1}b = (1/(x, c))x$. Note that $(c, x) \neq 0$ since $\lambda \notin \sigma(A)$. Thus the above expression is well defined. Then

$$((\lambda - A)^{-1}b, c) = \left(\frac{1}{(x, c)}x, c \right) = 1.$$

Now suppose $((\lambda - A)^{-1}b, c) = 1$ and let $x = (\lambda - A)^{-1}b$. Just reversing the above procedure gives $\lambda \in \sigma(A + b \otimes c)$. This completes the proof.

THEOREM 4. Let A be a discrete normal operator with $\{A, b\}$ controllable. Let $\sigma(A) = \{\lambda_n\}$ and $\{\mu_k\}$ be any finite sequence. Then there exists a vector $c \in \mathcal{H}$ such that the spectrum of $A + b \otimes c$ in the finite plane is $\{\mu_k\}$.

Proof. Since A is a discrete, normal operator, by Lemma 1 and the Spectral theorem there is an orthonormal basis $\{e_n\}_{n=1}^{\infty}$ such that $Ae_n = \lambda_n e_n$ with $\lambda_n \rightarrow \infty$. Also $A + b \otimes c$ is discrete though not necessarily normal. If $b = \{\beta_n\}$, $c = \{\gamma_n\}$, a simple computation leads to the fact that

$$f(z) = ((zI - A)^{-1}b, c) = \sum_{n=1}^{\infty} \frac{\alpha_n}{z - \lambda_n}$$

where $\alpha_n = \beta_n \tilde{\gamma}_n$.

By Theorem 3 given a finite sequence $\{\mu_1, \dots, \mu_k\}$ we want to choose a vector $c \in \mathcal{H}$ such that $f(\mu_k) = 1$. We chose c to have exactly k nonzero coordinates which we will specify later. Then

$$f(z) = \sum_{n=1}^k \frac{\alpha_n}{z - \lambda_n}.$$

This is a rational function with numerator a polynomial of degree $k - 1$, denominator a polynomial of degree k , simple poles at $\lambda_1, \dots, \lambda_k$. Classical interpolation techniques allow us to choose $\alpha_1, \dots, \alpha_k$ such that $f(\mu_k) = 1$. Then just define $\gamma_n = (\tilde{\beta}_n/\alpha_n)$, for $n = 1, \dots, k$ and $c = (\gamma_1, \dots, \gamma_k, 0, \dots)$ is the required vector.

Remarks. The controllability of $\{A, b\}$ appeared in a rather subtle form. It assured us that $\beta_n \neq 0$ for all n .

6. Stabilization.

DEFINITION. The system (1) is *stable* if $\|T(t)x\| \rightarrow 0$ as $t \rightarrow \infty$ for all $x \in \mathcal{H}$. It is *exponentially stable* if $\|T(t)x\| \leq c_x e^{-\delta t}$, $\delta > 0$, $t \geq 0$.

In the finite dimensional case it is well-known that these are both equivalent to $\operatorname{Re} \lambda < 0$ for $\lambda \in \sigma(A)$. As pointed out in the Introduction, this is not the case for infinite dimensional systems. However in our case it does hold [9].

THEOREM B. If A is a discrete operator then exponential stability is equivalent to the condition: $\operatorname{Re} \lambda < 0$ for $\lambda \in \sigma(A)$.

This leads to our next result.

THEOREM 5. If A is normal, discrete and $\{A, b\}$ is controllable then there exists a vector $c \in \mathcal{H}$ such that $\{A + b \otimes c, b\}$ is exponentially stable.

Proof. Just pick c such that $\sigma(A + b \otimes c)$ is contained in the left half plane $\{z: \operatorname{Re} z < 0\}$ and apply Theorem B.

Example 1. Here we present a simple example to show how the pole assignment process works to achieve stabilization.

Let A be the diagonal operator with $\{1 + (n-1)i\}_{n=1}^{\infty}$ on the diagonal. Then A is a normal discrete operator which generates a C_0 semigroup. Let $b = \{1/n\}_{n=1}^{\infty}$. Then the system

$$\dot{x}(t) = Ax(t) + u(t)b$$

is controllable but not stable. In fact $\|T(t)\| = e^t$. We will stabilize it by finding a vector $c \in \mathcal{H}$ such that $\sigma(A + b \otimes c) = \{-1, -1+i, \infty\}$. Note that since $A + b \otimes c$ is discrete, $\infty \in \sigma(A + b \otimes c)$ for any c we choose.

Following the proof of Theorem 4 we will choose c with only its first two coordinates different from zero; i.e.

$$c = (\bar{\alpha}_1, \bar{\alpha}_2, 0, \dots)$$

with α_1, α_2 to be determined. Then

$$f(z) = ((zI - A)^{-1}b, c) = \frac{\alpha_1}{(z-1)} + \frac{\frac{1}{2}\alpha_2}{(z-1-i)} = \frac{(\alpha_1 + \frac{1}{2}\alpha_2)z - (\alpha_1 + \frac{1}{2}\alpha_2) - i\alpha_1}{(z-1)(z-1-i)}.$$

We will choose α_1 and α_2 so that $f(-1) = f(-1+i) = 1$. This leads to the system of equations

$$(2+i)\alpha_1 + \alpha_2 = -2(2+i), \quad -2\alpha_1 + \frac{1}{2}(i-2)\alpha_2 = -2(i-2).$$

Solving gives $\alpha_1 = (-5+2i)/2$, $\alpha_2 = 2i-1$. Thus $c = (-5-2i)/2, -1-2i, 0, \dots$.

Now consider the case where $\{A, b\}$ is not controllable. Let $C^- = \{z: \operatorname{Re} z < 0\}$, $C^+ = \{z: \operatorname{Re} z \geq 0\}$ and decompose $\sigma(A)$ into two sets σ^- and σ^+ such that

$$\sigma^- \subseteq C^-, \quad \sigma^+ \subseteq C^+.$$

Then A can be decomposed into $A^- \oplus A^+$. We thus obtain the following result.

THEOREM 6. *If $\{A^+, b^+\}$ is controllable, then $\{A, b\}$ is stabilizable.*

7. Problems. We have seen that by means of state feedback we can obtain any finite set as the poles of the transfer function of the new system. It is thus natural to ask what infinite sets can be obtained. The following example shows that there seems to be no simple answer to this question.

Example 2. Given: the system

$$\dot{x}(t) = Ax(t) + Bu(t)$$

where $\sigma(A) = \{\lambda_n\}_{n=1}^{\infty}$ with $\lambda_n \rightarrow \infty$. If $\{\lambda_n\}$ is not rather densely dispersed in the plane, it is easy to see that there are many sequences $\{\mu_n\}$ such that

$$\lim_{n \rightarrow \infty} \inf_m |\mu_n - \lambda_m| = \infty.$$

Assigning the poles to $\{\mu_n\}$ is equivalent to constructing a meromorphic function

$$f(z) = \sum_{m=1}^{\infty} \frac{\gamma_m}{z - \lambda_m}$$

such that $f(\mu_n) = 1$, $n = 1, 2, \dots$, and $\{\gamma_m\} \in l'$. We show that this can't happen. Note that

$$|f(\mu_n)| = \left| \sum_{m=1}^{\infty} \frac{\gamma_m}{\mu_n - \lambda_m} \right| \leq \|\gamma_m\| \left\| \frac{1}{\mu_n - \lambda_m} \right\|_{\infty} \rightarrow 0$$

as $n \rightarrow \infty$ by the choice of $\{z_n\}$. It is not hard to construct such examples where $\{\lambda_n\}$ and

$\{\mu_n\}$ have the same order of convergence. Thus it is not clear what sort of relationship must exist between $\{\lambda_n\}$ and $\{\mu_n\}$.

We feel that a more reasonable direction to extend the results given here would be to drop the condition that A be normal. Then, of course, the function $((zI - A)^{-1}b, c)$ would have a much more complicated form.

REFERENCES

- [1] A. V. BALAKRISHNAN, *Introduction to Optimization Theory in Hilbert Space*, Lecture Notes in Operations Research and Math. Systems, 42, Springer-Verlag, Berlin, 1971.
- [2] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators III*, Wiley-Interscience, New York, 1971.
- [3] A. FEINTUCH, *Strictly cyclic linear systems*, SIAM J. Appl. Math., to appear.
- [4] ———, *On single input controllability for infinite dimensional linear systems*, J. Math. Anal. and Appl., to appear.
- [5] ———, *One dimensional perturbations of cyclic operators*, Linear Algebra and Appl., to appear.
- [6] P. A. FUHRMANN, *On weak and strong reachability and controllability of infinite dimensional linear systems*, J. Optimization Theory Appl., 19 (1972), No. 2, pp. 77–89.
- [7] ———, *Some results on controllability*, Ricerche Automatica, (1974), no. 2–3, pp. 1–5.
- [8] W. M. WONHAM, *Linear Multivariable Control*, Lecture Notes in Economics and Math. Systems, 101, Springer-Verlag, Berlin, 1974.
- [9] R. TRIGGIANI, *On the stabilizability problem in Banach space*, J. Math. Anal. Appl., 52 (1975), pp. 383–403.
- [10] M. SLEMROD, *A note on complete controllability and stabilizability for linear control systems in Hilbert space*, this Journal, 12 (1974), pp. 500–508.
- [11] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, A. M. S. Colloquium Publication 31, American Mathematical Society, Providence, RI, 1957.

ON Φ -CONVEXITY IN EXTREMAL PROBLEMS*

SZYMON DOLECKI† AND STANISŁAW KURCYUSZ‡

Abstract. For a class of functions Φ on an arbitrary set X , Φ -convex subsets of X and functions on X are defined, the latter being least upper bounds of some functions from Φ . Also the generalized Fenchel transform and Φ -subgradients are determined and their properties investigated.

Φ -convexity and Φ -subdifferentiability of lower-semicontinuous functions on metric spaces are examined with respect to special important families Φ . Among related results, we present a theorem on the existence of minimizing points of nonlinear functions on Banach spaces and extensions of the notion of Hölder continuity. The relevance of the theory to perturbed extremal problems is indicated.

Introduction. An optimization problem with constraints is usually more troublesome than one without constraints. Therefore many attempts have been made to reduce constrained problems to “free” optimization questions. The development of computational techniques made this approach even more favorable.

This reduction motivates the main topic of the paper. To explain what we mean by reduction, let us consider the following example:

$$(1) \quad \text{minimize } Q(u) : C(u) = x$$

where U is a Banach space, C is an operator from U into another Banach space X and Q is a real function on U . We may proceed by the use of the Lagrange functional and formulate a derived problem

$$(2) \quad \text{minimize } (Q(u) + \varphi(C(u)) - \varphi(x)) : u \in U.$$

Here φ is a bounded linear form on X and the minimization is performed over the whole space U . The best situation is when we can find a φ such that (1) and (2) are equivalent in the sense that $\hat{u} \in U$ solves (1), if and only if it is a solution of (2).

A much weaker property of optimization problems holds, whenever $\sup_{\varphi \in X^*} \inf_{u \in U} (Q(u) + \varphi(C(u)) - \varphi(x)) = \inf_{Q(u)=x} Q(u)$. This property, when discovered for a given function, is a starting point for further investigations of which sequences $\{\varphi_n\}$ should be picked out and whether the corresponding minimizing vectors converge to the minimizing vector of (1).

There exists another approach to (1) as good as the one just described or even better. Instead of (2) we consider a substitute problem:

$$(3) \quad \text{minimize } (Q(u) + \rho \|C(u) - z\|^2 - \rho \|x - z\|^2) : u \in U$$

where $\rho > 0$ and $z \in X$. Also here the minimization is over the whole U . The same problems of convergence, existence etc. arise. We try to approach the original problem by taking the supremum over all ρ and z .

The approach with the linear Lagrange functional (2) is classical, but it often becomes ineffective due to so-called “duality gaps” [16]. It has turned out that the use of the Lagrangian (3)—often termed “augmented”, see [5], [28], [36]—yields much more effective computational techniques. The number of works devoted to these techniques (called “methods of multipliers” after Hestenes [20] or “shifted penalty techniques” after Powell [26], [36]) is now enormous. An excellent account of these topics is available in a recent survey paper of Bertsekas [5].

* Received by the editors April 26, 1976, and in revised form June 8, 1977.

† Institute of Mathematics, Polish Academy of Sciences, Sniadeckich 8, Warsaw, Poland.

‡ Institute of Automatic Control, Technical University of Warsaw, Nowowiejska 15/18, 00-665 Warsaw, Poland.

Both these Lagrangians lead to some duality theory and are examples of so-called generalized Lagrange functionals, defined and investigated by numerous authors [3], [6], [23], [25], [31], [32].

Since its very beginning convex analysis was very closely related to extremal convex problems, and the duality theory for convex problems and linear Lagrange functionals is now best expressed in its language [24], [27]. But only quite recently has it become evident [49], [35], [34], [32], [26] that for much more general extremal problems and Lagrangians the duality relations are expressible in terms very similar to those used in convex analysis. Following Ky Fan we use the name Φ -convexity. Ky Fan [17] defined Φ -convex sets to be intersections of level sets of functions belonging to a family of functions Φ . He was able to generalize the Krein–Milman theorem.

The aim of this paper is to discuss the notion and properties of Φ -convex sets and functions. The three following sections are devoted to the study of Φ -convexity in full generality (in relation to a more abstract concept of order convexity introduced by Kutateladze and Rubinov [47]).

The next three sections (4, 5, 6) discuss the important class of metric-like functions Φ . For such classes every l.s.c. function satisfying a mild growth condition is Φ -convex. Numerous metric-like families are presented. The results on dense Φ -subdifferentiability generalize some existence theorems of Bidaud [6] and are connected with works of Asplund [0], Edelstein [12], Baranger [1] and Ekeland [45]. The Φ -subdifferentiability is related to the notion of calmness of Clarke [8], [9].

In Section 7 we indicate briefly how the former results may be exploited in constrained optimization problems.

An outline of the theory (reporting major results of this paper) appeared in the summer of 1976 ([11]). Afterwards we got acquainted with works of Balder [40] and Lindberg [48] that overlap with some parts of this paper and contain other developments (for instance the stability and convergence results of Balder). We give a concise account of those works in Section 8. This section contains, as well, a hint of a stability theory (recently worked out in [42][43][44]) that determines the area of applicability of Φ -convex optimization.

1. A generalized concept of convexity. Let Φ be a family of real finite functions defined on a set X .

DEFINITION 1.1. An (extended) real function $f: X \rightarrow \bar{\mathbb{R}}$ is said to be Φ -convex, if

$$(1.2) \quad f(x) = \sup_{\varphi \in \Phi'} \varphi(x)$$

for some $\Phi' \subset \Phi$. For Φ' empty we set $f \equiv -\infty$.

A subset A of X is called Φ -convex, whenever

$$(1.3) \quad A = \bigcap_{\varphi \in \Phi'} \{x \in X: \varphi(x) \leq a_\varphi\}$$

for some $\Phi' \subset \Phi$ and some set $\{a_\varphi\}_{\varphi \in \Phi'}$ of reals. If Φ' is empty we set $A = X$.

DEFINITION 1.4. The topology $\tau(\Phi)$ induced by Φ is the coarsest topology of X for which Φ -convex sets are closed.

PROPOSITION 1.5. Let f be a Φ -convex function. Then the level sets $\{x: f(x) \leq a\}$ of f are Φ -convex and f is $\tau(\Phi)$ -lower semicontinuous.

Proof. Indeed, a level set of f is

$$\{x: f(x) \leq a\} = \left\{x: \sup_{\varphi \in \Phi'} \varphi(x) \leq a\right\} = \bigcap_{\varphi \in \Phi'} \{x: \varphi(x) \leq a\};$$

hence Φ -convex and consequently $\tau(\Phi)$ -closed. Therefore f is $\tau(\Phi)$ -lower semicontinuous.

For any function f we define its Φ -convex hull f^0 as the greatest Φ -convex function majorized by f . The Φ -convex hull $\text{co}_\Phi A$ of A is the intersection of all Φ -convex sets that contain A .

We shall now provide a simple but handy criterion for Φ -convexity of sets and functions

PROPOSITION 1.6 (separation property).

(i) A function $f: X \rightarrow \bar{\mathbb{R}}$ is Φ -convex, if and only if for each $x \in X$ and $r < f(x)$ there exists φ majorized by f and such that $\varphi(x) > r$.

(ii) A set A is Φ -convex, if and only if for each $x_0 \notin A$ there is a function $\varphi \in \Phi$, such that

$$(1.7) \quad \sup_{x \in A} \varphi(x) < \varphi(x_0).$$

Since elements of X may be viewed as finite real functions on the set Φ

$$(1.8) \quad x(\varphi) = \varphi(x),$$

we say that a function $g: \Phi \rightarrow \bar{\mathbb{R}}$ is X -convex, whenever $g(\varphi) = \sup_{x \in X'} \varphi(x)$, $\varphi \in \Phi$, for some $X' \subset X$.

Analogously we define X -convex subsets of Φ , the topology $\tau(X)$ of Φ , and so on Observe that the roles of X and Φ are fully symmetric.

DEFINITION 1.9. An arbitrary function $f: X \rightarrow \bar{\mathbb{R}}$ is called Φ -bounded, if it is minorized by an element φ of Φ .

It follows that the Φ -convex hull f^0 of a Φ -bounded function is always greater than $-\infty$.

Example 1.10. Let X be a topological vector space and let Φ stand for the set of all continuous affine functions on X . Convex functions f on X are those that fulfill $f(\lambda x) + (1 - \lambda)y \leq \lambda f(x) + (1 - \lambda)f(y)$ for each x, y and λ , $0 \leq \lambda \leq 1$; here by convention $(+\infty) + (-\infty) = (+\infty)$. A convex function is called proper if it is nowhere equal to $-\infty$ and if it is not $+\infty$ everywhere.

In view of the Hahn-Banach theorem the Φ -convex functions are precisely the convex, proper, lower semicontinuous functions, the function $\equiv +\infty$ and the function $\equiv -\infty$. Φ -convex sets are those which are closed and convex, $\tau(\Phi)$ is the weak topology of X . The topological dual X^* of X is a "layer" of Φ (a subset of those φ that vanish at zero). Its $\tau(X)$ is the weak* topology, X -convex sets are convex weakly* closed.

We see that Φ -convexity is a generalization of classical convexity combined with closedness. The essential fact about classical convex, proper, lower semicontinuous functions was used in the representation (1.2).

From now on we shall assume that

$$(1.11) \quad \varphi + c \in \Phi, \quad \text{whenever } \varphi \in \Phi \text{ and } c \in \mathbb{R}.$$

Define a map \mathbb{F} of real functions on X to real functions on Φ ($\mathbb{F}: \bar{\mathbb{R}}^X \rightarrow \bar{\mathbb{R}}^\Phi$) by

$$(1.12) \quad (\mathbb{F}f)(\varphi) = f^*(\varphi) = \sup_{x \in X} (\varphi(x) - f(x)).$$

In the same way we define $\hat{\mathbb{F}}: \bar{\mathbb{R}}^\Phi \rightarrow \bar{\mathbb{R}}^X$:

$$(1.13) \quad (\hat{\mathbb{F}}g)(x) = g^*(x) = \sup_{\varphi \in \Phi} (\varphi(x) - g(\varphi)).$$

DEFINITION 1.14. The map \mathbb{F} (1.12) is called the *Fenchel transform* and $\hat{\mathbb{F}}$ in (1.13) is said to be the *(inverse) Fenchel transform*. f^* is the Φ -conjugate of f and g^* is the X -conjugate of g .

It follows readily that functions of the form $\mathbb{F}f$, f arbitrary, are X -convex ($\hat{\mathbb{F}}g$ are Φ -convex). The second conjugate of a function f (defined by $f^{**}(x) = (\hat{\mathbb{F}}\mathbb{F}f)(x)$) is, of course, Φ -convex. Let us recall a geometrical interpretation of the Fenchel transform: $-f^*(\varphi) = \inf_{x \in X} (f(x) - \varphi(x))$ is the vertical distance between f and φ ; thus $f^*(\varphi) \leq 0$ exactly if $f(x) \geq \varphi(x)$ for all $x \in X$.

We note that $f^*(\varphi) = -\infty$, if and only if $f(x) \equiv +\infty$ while $f^*(\varphi) \equiv +\infty$ if and only if $f(x) = -\infty$ for some $x \in X$. Also $f^*(\varphi) < +\infty$ for some φ whenever f is Φ -bounded. Furthermore

$$(1.15) \quad f^* \geq g^* \quad \text{if } g \geq f,$$

$$(1.16) \quad (f+a)^*(\varphi) = f^*(\varphi - a) = f^*(\varphi) - a \quad \text{for } a \in \mathbb{R},$$

$$(1.17) \quad \text{if } \alpha > 0, \quad \varphi \in \Phi \text{ entail } \alpha\varphi \in \Phi, \text{ then}$$

$$(\alpha f)^*(\varphi) = \alpha f^*\left(\frac{1}{\alpha}\varphi\right) \quad \text{for } \alpha > 0,$$

$$(1.18) \quad f(x) + f^*(\varphi) \geq \varphi(x) \quad (\text{Fenchel inequality}).$$

The following theorem is an abstract formulation of the Moreau–Fenchel theorem.

THEOREM 1.19 [13]. *For an arbitrary function $f: X \rightarrow \bar{\mathbb{R}}$, its Φ -convex hull is equal to the second Fenchel conjugate:*

$$(1.20) \quad f^0(x) = f^{**}(x), \quad x \in X.$$

Proof. By definition

$$f^0(x) = \sup_{\varphi \leq f} \varphi(x) = \sup_{f^*(\varphi) \leq 0} \varphi(x).$$

Actually, we may consider only such φ for which $f^*(\varphi) = 0$, for if $\varphi \leq f$ and $f^*(\varphi) = -a < 0$, then $f^*(\varphi + a) = f^*(\varphi) + a = 0$. Hence

$$f^0(x) = \sup_{f^*(\varphi) = 0} \varphi(x).$$

On the other hand,

$$f^{**}(x) = \sup_{\varphi \in \Phi} (\varphi(x) - f^*(\varphi)) \geq \sup_{\substack{\varphi \\ f^*(\varphi) = 0}} (\varphi(x) - f^*(\varphi)) = \sup_{f^*(\varphi) = 0} \varphi(x).$$

Assume that $f^*(\varphi) = -a$. We have

$$\varphi(x) - f^*(\varphi) = (\varphi(x) + a) - f^*(\varphi + a)$$

with $f^*(\varphi + a) = 0$. This proves (1.20).

It is important to introduce another simple notion.

DEFINITION 1.21. A function $f: X \rightarrow \bar{\mathbb{R}}$ is called Φ -convex at x_0 , whenever $f(x_0) = f^0(x_0) (= f^{**}(x_0))$.

In other words, f is Φ -convex at x_0 if there exists a sequence $\{\varphi_n\} \subset \Phi$ of its minorants such that $\lim_{n \rightarrow \infty} \varphi_n(x_0) = f(x_0)$. Notice that such a function is $\tau(\Phi)$ -lower semicontinuous at x_0 (for each $\varepsilon > 0$ there is a neighborhood U of x_0 such that $f(x) \geq f(x_0) - \varepsilon$ for $x \in U$).

2. Possible extensions. The explicit definition of Φ -convex sets was first given by Ky Fan [17]. The transform \mathbb{F} has been defined by Moreau [49] and Weiss [35] and investigated by Vogel [34], Elster and Nehse [13], Seidler [32] and Kurcysz [23].

A reflection on the meaning of Φ -convexity leads to a more general “intersectional” formulation [50]. A pair $\{X, \mathcal{A}\}$, where X is a set and \mathcal{A} denotes a family of its subsets, is called a cyrtological space, whenever $X \in \mathcal{A}$ and $\bigcap_{A \in \mathcal{A}'} A \in \mathcal{A}$ for any subfamily \mathcal{A}' of \mathcal{A} . The set $\bigcap_{B \subset A, A \in \mathcal{A}} A$ is called the \mathcal{A} -convex hull of B and B is called \mathcal{A} -convex, if it is equal to its hull. A subfamily \mathcal{L} of \mathcal{A} is said to be a basis for \mathcal{A} if each \mathcal{A} -convex set may be represented as an intersection of some elements of \mathcal{L} . The topology $\tau(\mathcal{A})$ is the coarsest topology for which \mathcal{A} -convex sets are closed; A is \mathcal{A} -convex if and only if for each $x \notin A$, there is $L \in \mathcal{L}$ with $A \subset L$, $x \notin L$. For every topological space X , $\{X, \mathcal{F}\}$ is cyrtological for the family \mathcal{F} of its closed sets.

Our definition (1.3) of Φ -convex sets states, in terms of cyrtological spaces, that the sets

$$\{x: \varphi(x) \leq a\}, \quad \varphi \in \Phi, \quad a \in \mathbb{R},$$

form a basis for \mathcal{A} .

Φ -convex functions may be reduced to \mathcal{A} -convex sets by use of their epigraphs $\{(x, r): r \geq f(x)\}$ in $X \times \mathbb{R}$ with a basis composed of $\{(x, r): r \geq \varphi(x)\}$. Note that the latter is the 0-level of the function $\tilde{\varphi}(x, r) = \varphi(x) - r$.

A still more abstract notion of convexity was investigated by Kutateladze and Rubinov [47]. A brief presentation of their theory may help in understanding of the core idea of convexity. Let Z be a subset of a complete lattice (ordered set in which the supremum and the infimum exist for every subset) and let $\Omega \subset Z$ be such that the least element $-\infty$ belongs to Ω . An element z of Z is called Ω -convex, whenever

$$z = \sup \Omega'$$

for some $\Omega' \subset \Omega$. (The order character of convexity was noticed by Ky Fan [17].)

Example 2.1. Let $Z = 2^X$. The order structure is defined by the inclusion $A_1 \leq A_2$, whenever $A_1 \supset A_2$. Let Ω be a subset of Z , hence a family of subsets of X . The Ω -convex elements of Z are the \mathcal{A} -convex subsets of the cyrtological space $\{X, \mathcal{A}\}$ generated by Ω ; Ω constitutes a basis for \mathcal{A} .

For an element z of Z we define its support: $S(z) = \{y \in \Omega: y \leq z\}$. In order not to proliferate variants of “convex sets” we shall call a subset A of Ω a support, if A is the support of its supremum

$$A = \{y \in \Omega: y \leq \sup A\}.$$

The inclusion introduces the order in Ω for which the transform $S: z \rightarrow S(z)$ is an isomorphism of ordered sets. The transform S is named the Minkowski duality.

The Minkowski duality is useful for expressing Φ -convexity of subsets of X in terms of supports in $\bar{\mathbb{R}}^\Phi$.

We are given a set X and a family Φ of real finite functions on X . The formula

$$(Ix)(\varphi) = \varphi(x)$$

defines an embedding $I: X \rightarrow \bar{\mathbb{R}}^\Phi$. In $Z = \bar{\mathbb{R}}^\Phi$ we define the ordinary order structure: $f \geq g$, whenever $f(\varphi) \geq g(\varphi)$ for each $\varphi \in \Phi$. The set Ω is the image of X under I and may be identified with $\{x(\cdot): x \in X\}$. Φ separates points, if for each x_1, x_2 there is φ such that $\varphi(x_1) \neq \varphi(x_2)$.

THEOREM 2.2 [47]. *Let Φ separate points of X . A subset A of X is Φ -convex if and only if its image $I(A)$ is a support in $I(X)$.*

Proof. Define the function $\delta_A : \Phi \rightarrow \bar{\mathbb{R}}$ by

$$(2.3) \quad \delta_A(\varphi) = \sup_{x \in A} \varphi(x).$$

From (2.3) it follows that always $\delta_A = \sup I(A)$. Assume that A is Φ -convex. If for some $x_0 \in X$ one has $Ix_0 \not\leq \delta_A$ (there exists a φ such that $(Ix_0)(\varphi) > \delta_A(\varphi)$, or in other words $\varphi(x_0) > \sup_{x \in A} \varphi(x)$) then by the assumption $x_0 \notin A$; hence $Ix_0 \notin I(A)$ and $I(A)$ is a support in $I(X)$.

Assume now that $I(A)$ is a support and suppose that $x_0 \notin A$. Since I is injective $Ix_0 \notin I(A)$ and thus $Ix_0 \not\leq \delta_A$. There exists a $\varphi \in \Phi$ such that $(Ix_0)(\varphi) = \varphi(x_0) > \delta_A(\varphi)$. Since x_0 was an arbitrary element outside A , A is Φ -convex in view of Proposition 1.5.

Using the above mentioned symmetry of Φ - and X -convexity we obtain

COROLLARY 2.4. *A subset B of Φ is X -convex if and only if B is a support in Φ .*

The corollary means that B is X -convex if and only if $B = \{\varphi \in \Phi : \varphi(x) \leq b(x), x \in X\}$ for a certain function b .

The X -convex function δ_A is called the *support function* of A .

3. Φ -subdifferentiability.

DEFINITION 3.1. *A φ is a subgradient of $f : X \rightarrow \bar{\mathbb{R}}$ at x_0 , if $f(x_0) = \varphi(x_0)$ and*

$$(3.2) \quad f(x) \geq \varphi(x) \quad \text{for each } x.$$

The set of $\varphi + c$, where $c \in \mathbb{R}$ and φ are subgradients of f at x_0 is called the Φ -subdifferential of f at x_0 and is denoted $\partial_\Phi f(x_0)$. If $\partial_\Phi f(x_0) \neq \emptyset$ we say that f is Φ -subdifferentiable at x_0 .

Remark 3.3. We can equivalently define the Φ -subdifferential $\partial_\Phi f(x_0)$ as the set of all $\varphi \in \Phi$ for which

$$(3.4) \quad f(x) - f(x_0) \geq \varphi(x) - \varphi(x_0) \quad \text{for each } x \in X.$$

In a similar manner we may define X -subgradients and X -subdifferentials of a function $g : \Phi \rightarrow \bar{\mathbb{R}}$.

It follows that if f is Φ -subdifferentiable at x_0 , then it is Φ -convex at x_0 . Neither a subdifferential nor a set of all subgradients need be X -convex, as we easily deduce on using Corollary 2.4. Nevertheless we have

PROPOSITION 3.5. *The set of all subgradients of f at x_0 is X -convex in the layer $\{\varphi \in \Phi : \varphi(x_0) = f(x_0)\}$ of Φ .*

Proof. Let $\varphi_0 \in \partial_\Phi f(x_0)$ and $\varphi_0(x_0) = f(x_0)$. Then there is x_1 such that $\varphi_0(x_1) > f(x_1)$. But $f(x_1) \geq \varphi(x_1)$ for all subgradients of f (at x_0) and $\{\varphi \in \Phi : \varphi(x_0) = f(x_0)\} \cap \partial_\Phi f(x_0)$ is X -convex in $\{\varphi \in \Phi : \varphi(x_0) = f(x_0)\}$.

PROPOSITION 3.6. *The following statements are equivalent:*

- (i) $\varphi \in \partial_\Phi f(x)$,
- (ii) $f(x) - \varphi(x) = \inf_{y \in X} (f(y) - \varphi(y))$,
- (iii) $f(x) + f^*(\varphi) = \varphi(x)$.

Moreover, they imply

$$(iv) \quad x \in \partial_X f^*(\varphi)$$

and

$$(v) \quad f^*(\varphi) - \varphi(x) = \inf_{\psi \in \Phi} (f^*(\psi) - \psi(x)),$$

while if $f(x) = f^0(x)$, then each of (iv), (v) is equivalent to (i).

Proof. The equivalence of (i), (ii) and (iii) is trivial in view of (3.4) and (1.12).

(iii) \Leftrightarrow (iv), (v). By Proposition (iii), $f(x) = f^{**}(x)$ and we have

$$(3.7) \quad f^{**}(x) + f^*(\varphi) = \varphi(x).$$

Now we apply (iii) \Leftrightarrow (i) and (iii) \Leftrightarrow (ii) to obtain (iv), (v), respectively with f replaced by f^* . If $f(x) = f^0(x) = f^{**}(x)$ then (3.7) becomes (iii).

PROPOSITION 3.8. *Let σ be a topology of $X \times \Phi$ for which the mapping $(x, \varphi) \rightarrow \varphi(x)$ is continuous. Assume that $f: X \rightarrow \mathbb{R}$ is $\tau(\Phi)$ -lower semicontinuous. Then the multifunction $\partial_{\Phi} f$ is closed (in $X \times \Phi$).*

Proof. Observe that σ is stronger than the product topology $\tau(\Phi) \times \tau(X)$ and consequently the function

$$(x, \varphi) \rightarrow f(x) + f^*(\varphi) - \varphi(x)$$

is σ -lower semicontinuous. The graph of $\partial_{\Phi} f$ is

$$\{(x, \varphi) : \varphi \in \partial_{\Phi} f(x)\} = \{(x, \varphi) : f(x) + f^*(\varphi) - \varphi(x) \leq 0\}$$

(due to the Fenchel inequality). It is closed as a level set of a l.s.c. function.

PROPOSITION 3.9. (i) *Suppose X is an open subset of a normed space and functions $f: X \rightarrow \mathbb{R}$ and $\varphi \in \Phi$ are differentiable (Gâteaux or Fréchet) on X up to order $k \geq 1$. Assume that $\varphi_0 \in \partial_{\Phi} f(x_0)$. Then $f'(x_0) = \varphi'_0(x_0)$ and if $f^{(i)}(x_0) = \varphi^{(i)}_0(x_0)$, $1 \leq i < l \leq k$ then*

$$(3.10) \quad \begin{aligned} f^{(l)}(x_0) &= \varphi^{(l)}_0(x_0), \quad l \text{ odd}, \\ f^{(l)}(x_0) &\geq \varphi^{(l)}_0(x_0), \quad l \text{ even}. \end{aligned}$$

(ii) *Suppose X is a locally convex topological vector space, f is convex proper, φ_0 is concave and continuous. Then*

$$(3.11) \quad \varphi_0 \in \partial_{\Phi} f(x_0) \Rightarrow \partial \varphi_0(x_0) \cap \partial f(x_0) \neq \emptyset.$$

If, moreover, all functions from Φ are concave, proper and continuous then the implication in (3.11) may be inverted.

Proof. (i) follows immediately from the equivalence (i) \Leftrightarrow (ii) of Proposition (3.6).

(ii) From the theorem of Moreau–Rockafellar [21] it follows that

$$\partial(f - \varphi_0)(x_0) = \partial f(x_0) - \partial \varphi_0(x_0).$$

Moreover, $f - \varphi_0$ attains its minimum on X at x_0 by Proposition 3.6 (ii) and hence

$$0 \in \partial f(x_0) - \partial \varphi_0(x_0),$$

which is equivalent to $\partial \varphi_0(x_0) \cap \partial f(x_0) \neq \emptyset$.

In general there is no simple correspondence between Gâteaux or Fréchet derivatives and $\partial_{\Phi} f(x)$. First, an analytic function f on $X = \mathbb{R}$ may have an everywhere empty Φ -subgradient even if it is Φ -convex where all functions from Φ are also analytic. An example is furnished by an arbitrary constant function $f: X \rightarrow \mathbb{R}$ and $\Phi = \{\varphi(x) = e^{x-y} : y \in \mathbb{R}\}$. On the other hand there may be infinitely many functions φ in $\partial_{\Phi} f(x)$ even if f is differentiable. Take, for instance a constant function f on $X = \mathbb{R}$ and $\Phi^2 = \{\varphi(x) = \rho(x-y)^2 : \rho > 0, y \in \mathbb{R}\}$. Then $\partial_{\Phi} f(x_0)$ contains all functions $\varphi(x) = \rho(x-x_0)^2$ with $\rho > 0$. This pathological behavior is due to the unlimited freedom in choosing families Φ .

Both inequalities (3.2), (3.4) describe the global behavior of a nonlinear function (similar to ordinary subgradients in the convex case) while ordinary derivatives and known generalized subgradients of nonlinear functions [9], [21] possess local nature.

4. Φ -convexity of lower semicontinuous functions.

DEFINITION 4.1. Let (X, d) be a metric space. Consider a real function $\eta: X \times X \rightarrow \mathbb{R}$ satisfying the following properties:

- (i) η is nonnegative.
- (ii) $\eta(x_n, y_n) \rightarrow 0$ only if $d(x_n, y_n) \rightarrow 0$ and $d(x, y_n) \rightarrow 0$ implies $\eta(x, y_n) \rightarrow 0$ for each fixed x .
- (iii) There is a function $M: X \times X \times \mathbb{R}^+ \rightarrow \mathbb{R}^+$ such that

$$\frac{\eta(x, y_2)}{\eta(x, y_1)} \leq M(y_1, y_2, \delta)$$

whenever $d(x, y_1) \geq \delta$.

- (iv) For fixed δ and y_2 , $M(\cdot, y_2, \delta)$ is bounded on bounded sets.

A class

$$(v) \quad \Phi = \{r - \rho\eta(\cdot, y) : r \in \mathbb{R}, \rho > 0, y \in Y\}$$

where Y is a dense subset of X , is called *metric-like*, if η fulfils (i) through (iv).

THEOREM 4.2. Let Φ be metric-like. Then every lower semicontinuous Φ -bounded function $f: X \rightarrow \bar{\mathbb{R}}$ is Φ -convex.

COROLLARY 4.3. Each closed subset of X is Φ -convex, whenever Φ is metric-like.

Proof. Let F be a closed set. The function χ_F :

$$\chi_F(x) = \begin{cases} 0, & x \in F, \\ 1 & \text{otherwise} \end{cases}$$

is lower semicontinuous and bounded and thus Φ -convex. $F = \{x: \chi_F(x) \leq \frac{1}{2}\}$ is Φ -convex.

If we assume that for each $y \in Y$, $\eta(\cdot, y)$ is lower semicontinuous, then also each Φ -convex function is l.s.c. in the metric topology.

We encounter here that specific situation when the topology $\tau(\Phi)$ understood as the collection of all closed sets is equal to the family of all Φ -convex sets.

The proof of Theorem 4.2 will begin with the following simple lemma, useful also in the sequel.

LEMMA 4.4. Suppose B is a bounded subset of X and $\rho_0, \delta > 0$, $y_0 \in X$. Denote

$$\Delta(x, y, \rho) = \rho\eta(x, y) - \rho_0\eta(x, y_0)$$

where $\eta(\cdot, \cdot)$ satisfies the assumptions in Def. 4.1(i)–(iv), and

$$Z(B, \delta) = \{(x, y) \in X \times X : d(x, y) > \delta, y \in B\}.$$

Then there exists $\bar{\rho} \geq 0$ such that

$$(4.4a) \quad \Delta(x, y, \rho) \geq \frac{1}{2}\rho\eta(x, y)$$

whenever $\rho \geq \bar{\rho}$ and $(x, y) \in Z(B, \delta)$.

Proof.

$$\begin{aligned} \Delta(x, y, \rho) &= \rho\eta(x, y) \left(1 - \frac{\rho_0\eta(x, y_0)}{\rho\eta(x, y)}\right) \\ &\geq \rho\eta(x, y) \left(1 - \frac{\rho_0}{\rho} M(y, y_0, \delta)\right) \end{aligned}$$

and since $M(\cdot, y_0, \delta)$ is bounded on B , then for $\rho \geq \bar{\rho}$, say,

$$1 - \frac{\rho_0}{\rho} M(y, y_0, \delta) \geq \frac{1}{2}$$

and (4.4a) follows.

Proof of Theorem 4.2. Take an arbitrary $x \in X$ and $\bar{r} < f(\bar{x})$. By Proposition 1.6(ii) we must show that for some $\rho > 0$, $r \in R$ and $y \in Y$ we have

$$(4.5) \quad f(x) \geq r - \rho\eta(x, y) \quad \forall x \in X,$$

$$(4.6) \quad r - \rho\eta(\bar{x}, y) > \bar{r}.$$

Since f is lower semicontinuous, there is $\delta_1 > 0$ such that $f(x) \geq \bar{r} + \varepsilon$, $x \in K(\bar{x}, \delta_1)$,¹ where $\varepsilon = \frac{1}{2}(f(\bar{x}) - \bar{r}) > 0$. Since η is nonnegative, for all $\rho > 0$, $y \in X_0$ also

$$(4.7) \quad f(x) \geq \bar{r} + \varepsilon - \rho\eta(x, y) \quad \forall x \in K(\bar{x}, \delta_1).$$

By Φ boundedness of f ,

$$(4.8) \quad f(x) \geq r_0 - \rho_0\eta(x, y_0) \quad \forall x \in X$$

for some $\rho_0 > 0$, $r_0 \in R$ and $y_0 \in X_0$. Now take $x \notin K(\bar{x}, \delta_1)$ and $y \in B = K(\bar{x}, \delta_1/2)$ so that $d(x, y) \geq \delta_1/2$. Observe also that $\eta(x, y) \geq \gamma > 0$ for such (x, y) by Def. 4.1 (ii). We can now apply Lemma 4.4 with $\delta = \delta_1/2$ to obtain the existence of a ρ such that

$$\Delta(x, y, \rho) = \rho\eta(x, y) - \rho_0\eta(x, y_0) \geq \frac{1}{2}\rho\gamma$$

whenever $x \notin K(\bar{x}, \delta_1)$, $y \in B$ and $\rho \geq \bar{\rho}$. Taking $\rho > \bar{\rho}$ large enough we can now guarantee that

$$\Delta(x, y, \rho) \geq \bar{r} + \varepsilon - r_0$$

or

$$(4.9) \quad r_0 - \rho_0\eta(x, y_0) \geq \bar{r} + \varepsilon - \rho\eta(x, y) \quad \forall x \notin K(\bar{x}, \delta_1) \quad \forall y \in K(\bar{x}, \delta_1/2).$$

By Def. 4.1 (ii) there is $0 < \delta \leq \delta_1/2$ such that $\rho\eta(\bar{x}, y) < \varepsilon$ for $y \in K(\bar{x}, \delta)$. Take some $y \in K(\bar{x}, \delta) \cap Y$, $r = \bar{r} + \varepsilon$ and ρ as before. Then (4.7), (4.8), (4.9) give (4.5) while (4.6) results from the definition of ρ , y and r .

Remark 4.10. (i) It is evident from the final part of the proof that the assertion of the theorem remains valid when

$$\Phi = \{\varphi(x) = r - \rho\eta(x, y) : (\rho, y) \in \mathcal{N} \subset R_+ \times X\}$$

where \mathcal{N} has the property that for each $\rho > 0$ the set $X_\rho = \{y \in X : (\rho, y) \in \mathcal{N}\}$ is dense in X .

(ii) If, in the course of the above proof one takes for a fixed $\bar{x} \in X$ a sequence $\bar{r}_n = f(\bar{x}) - 1/n$ then one can construct a sequence $\{\rho_n, y_n, a_n\}$ with $\rho_n \rightarrow \infty$, $y_n \rightarrow \bar{x}$, $y_n \in X_{\rho_n}$, $a_n \in R$ such that $\varphi_n(x) = \rho_n\eta(x, y_n)$ satisfies

$$-\varphi(\bar{x}) + a_n > \bar{r}_n = f(\bar{x}) - \frac{1}{n},$$

$$f(x) \geq -\varphi_n(x) + a_n > -\varphi_n(x) + f(\bar{x}) + \varphi(\bar{x}) - \frac{1}{n}$$

or

$$f(x) - f(\bar{x}) \geq \varphi_n(\bar{x}) - \varphi_n(x) - \frac{1}{n}, \quad x \in X.$$

¹ In the sequel $K(\bar{x}, \delta)$ denotes a ball with radius δ around \bar{x} .

(iii) The same construction is applicable if we require f to be lower semicontinuous at \bar{x} and not on the whole space. Thus, Theorem 4.2 admits the following extension: if f is Φ -bounded and lower semicontinuous at \bar{x} , then f is Φ -convex at \bar{x} .

Next we shall give examples of functions $\eta(x, y)$ satisfying assumptions of Def. 4.1 (i)–(iv). We shall call a function $\psi: R^+ \rightarrow R^+$ *quietly increasing*, if ψ is nondecreasing and the quotient $\psi(t+t_0)/\psi(t)$ is bounded outside every neighborhood of zero for each $t_0 \geq 0$, that is

$$(4.11) \quad \forall \delta > 0 \quad \forall t_0 \geq 0 \quad \exists M(t_0, \delta) \quad \forall t \geq \delta \quad \frac{\psi(t+t_0)}{\psi(t)} \leq M(t_0, \delta).$$

Roughly speaking, these are the functions which do not grow more quickly than exponentials at infinity; e.g. e^{t^2} is not quietly increasing.

Let us also recall that a function $\psi: R^+ \rightarrow R^+$ is called *forcing* if

$$(4.12) \quad \psi(t) \rightarrow 0 \Leftrightarrow t \rightarrow 0.$$

PROPOSITION 4.13. *If $\psi: R^+ \rightarrow R^+$ is a quietly increasing forcing function then $\eta = \psi \circ d$ satisfies Def. 4.1 (i)–(iv).*

Proof. Def. 4.1 (i), (ii) are obvious. Take any $y_1, y_2 \in X$, $\delta > 0$ and $x \notin K(y_1, \delta)$. Put $t = d(x, y_1)$, $t_0 = d(y_1, y_2)$. Since

$$d(x, y_2) \leq d(x, y_1) + d(y_1, y_2) = t + t_0$$

and ψ is nonnegative nondecreasing, we have

$$\frac{\eta(x, y_2)}{\eta(x, y_1)} = \frac{\psi(d(x, y_2))}{\psi(d(x, y_1))} \leq \frac{\psi(t+t_0)}{\psi(t)} \leq M(d(y_1, y_2), \delta) = M(y_1, y_2, \delta).$$

This proves Def. 4.1 (iii). Moreover, since ψ is nondecreasing we can always modify $M(\cdot, \delta)$ to obtain a nondecreasing function. Hence for y_1 in a bounded set and y_2 fixed, $d(y_1, y_2) \leq c$ for some $c > 0$ and $M(y_1, y_2, \delta) \leq M(c, \delta)$.

Example 4.14. It is readily verified that such functions as t^α , $\alpha > 0$, $e^{\alpha t} - 1$, $\alpha > 0$, $\ln \alpha t - 1$, $\alpha > 0$ are quietly increasing forcing functions.

Example 4.15. Suppose X is a normed space. Letting $\Psi(t) = t^2$ one obtains

$$\eta(x, y) = \|x - y\|^2.$$

Theorem 4.2 asserts that every function f on X which is l.s.c. and fulfills

$$(4.16) \quad f(x) \geq r_0 - \rho_0 \|x - y_0\|^2, \quad x \in X$$

for some $\rho_0 > 0$, $y \in X$, $r_0 \in R$ is Φ -convex with Φ defined by Def. 4.1 (v), so that

$$f(x) = \sup_{(\rho, y, a) \in \Phi'} (-\rho \|x - y\|^2 + a), \quad x \in X, \quad \Phi' \subset \Phi.$$

Condition (4.16) is the quadratic growth condition of Rockafellar [28], [38] and we can recognize here Theorem 2 of Rockafellar [28].

We observe that the sum of quietly increasing forcing functions satisfies these properties, so that every polynomial with positive coefficients is a quietly increasing forcing function.

Another way of generating functions η satisfying Def. 4.1 (i)–(iv) is described below.

PROPOSITION 4.17. Assume $X = R^n$ with Euclidean topology. Suppose functions ψ_1, \dots, ψ_n are quietly increasing and forcing. Then the function

$$\eta(x, y) = \sum_{i=1}^n \psi_i(|x_i - y_i|)$$

satisfies Def. 4.1 (i)–(iv) with $d(x, y) = \|x - y\|$, $\|\cdot\|$ being any norm on R^n .

Proof. Def. 4.1 (i), (ii) are obvious. In order to prove 4.1 (iii), (iv) let us choose the norm $\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$. Take $y_1, y_2 \in R^n$, $\delta > 0$ and $x \notin K(y_1, \delta)$. Let $I_1(x) = \{i : |x_i - y_i| \geq \delta\}$ and $I_2(x) = \{1, \dots, n\} \setminus I_1(x)$. $I_1(x) \neq \emptyset$. Put $t_i = |x_i - y_i|$, $t_{0i} = |y_{1i} - y_{2i}|$. Clearly

$$|x_i - y_{2i}| \leq t_i + t_{0i}.$$

and hence

$$\begin{aligned} \frac{\eta(x, y_2)}{\eta(x, y)} &\leq \frac{\sum_{i=1}^n \psi_i(t_i + t_{0i})}{\sum_{j \in I_1(x)} \psi_j(|x_j - y_{1j}|)} \\ &\leq \sum_{i \in I_1(x)} \frac{\psi_i(t_i + t_{0i})}{\psi_i(t_i)} + \sum_{i \in I_2(x)} \frac{\psi_i(t_i + t_{0i})}{\sum_{j \in I_1(x)} \psi_j(t_j)}. \end{aligned}$$

Now, for $i \in I_1(x)$, $t_i \geq \delta$ and

$$\frac{\psi_i(t_i + t_{0i})}{\psi_i(t_i)} \leq M_i(t_{0i}, \delta).$$

For $i \in I_2(x)$

$$\psi_i(t_i + t_{0i}) \leq \psi_i(\delta + t_{0i}),$$

since ψ_i are nondecreasing. Finally,

$$\sum_{j \in I_1(x)} \psi_j(t_j + t_{0j}) \geq \inf_{j \in I_1(x)} \psi_j(\delta) \geq \inf_{1 \leq j \leq n} \psi_j(\delta) = c(\delta) > 0.$$

Therefore

$$\begin{aligned} \frac{\eta(x, y_2)}{\eta(x, y_1)} &\leq \sum_{i \in I_1(x)} M_i(t_{0i}, \delta) + \sum_{i \in I_2(x)} \frac{\psi_i(\delta + t_{0i})}{c(\delta)} \\ &\leq \sum_{i=1}^n \left[M_i(t_{0i}, \delta) + \frac{\psi_i(\delta + t_{0i})}{c(\delta)} \right] = M(y_1, y_2, \delta) < \infty. \end{aligned}$$

Arguing as in the proof of Proposition 4.13 we conclude that 4.1 (iv) also holds.

Suppose now that X is a normed space and define for $\alpha > 0$

$$(4.18) \quad \Phi^\alpha = \{\varphi(x) = -\rho\|x - y\|^\alpha + r : \rho > 0, y \in X, r \in R\}.$$

According to (4.12), (4.14) this family is metric-like. Therefore any (norm) l.s.c. and Φ^α -bounded function on X is Φ^α -convex.

5. Φ -subdifferentiability of Hölder functions. When Φ is metric-like then the Φ -subdifferentiability of a function f at x_0 amounts to requiring that

$$(5.1) \quad f(x) \geq f(x_0) - \rho\eta(x, y)$$

for some $\rho > 0$, y and all x .

An important situation is when the equality holds in (5.1) precisely for $y = x_0$. In

such a case, perturbation procedures for extremal problems behave particularly well (see § 7). The Φ -subdifferentiability for $\Phi_1 = \{r - \rho\| \cdot - y\|\}$ in normal spaces already entails this desirable behavior. Then it is possible to set $y = x_0$ in (5.1) to get

$$(5.2) \quad f(x) \geq f(x_0) - \rho\eta(x, x_0).$$

DEFINITION 5.3. We say that f is η -lower semicontinuous (η -calm) at x_0 if there exists a ball $K(x_0, \delta)$ around x_0 such that for $x \in K(x_0, \delta)$, (5.2) holds. Observe that η -calmness at x_0 is equivalent to

$$(5.4) \quad \liminf_{\substack{x \rightarrow x_0 \\ x \neq x_0}} \frac{f(x) - f(x_0)}{\eta(x, x_0)} > -\infty.$$

Here we recognize a generalization of a so-called calmness condition introduced by Clarke [8], [9]. (Clarke put $\eta(x, x_0) = \|x - x_0\|$.) The η -calmness is also related to the order β of stability of Rockafellar [28]. It is perhaps interesting to notice that (5.4) may be written as the "subderivative condition":

$$(5.5) \quad \liminf_{\substack{x \rightarrow x_0 \\ x \neq x_0}} \frac{f(x) - f(x_0) - \rho\eta(x, x_0)}{\eta(x, x_0)} \geq 0.$$

On the other hand, for $\eta(x, x_0) = \|x - x_0\|^\alpha$, $0 < \alpha \leq 1$, (5.2) is equivalent to the α -Hölder lower semicontinuity. If the constant ρ is independent of x_0 on an open set we obtain the classical notion of local Hölder continuity.

PROPOSITION 5.6. Let Φ be metric-like. Assume that $f: X \rightarrow \bar{\mathbb{R}}$ is Φ -bounded and η -calm at x_0 ; then it is Φ -subdifferentiable at x_0 .

Of particular interest are the families Φ :

$$(5.7) \quad \Phi^c = \{r - c\eta(x, y) : y \in X\}$$

where $c > 0$ is a fixed number. Here the situation is different. A function Φ -convex with Φ as in Def. 4.1(v) need not be Φ^c -convex for any positive c .

Now we return again to normed spaces and consider functions $\eta(x, y) = \|x - y\|^\alpha$, $\alpha > 0$. Uniform η -calmness of a finite function f means here precisely (global) Hölder continuity of f on X with exponent α .

Φ^c of (5.7) with $\eta(x, y) = \|x - y\|^\alpha$ will be denoted by $\Phi^{\alpha, c}$ and for a Hilbert space X we shall call $\Phi^{2, c} = Q^c$.

Suppose that a function f on a Hilbert space X is Q^c -convex. This means that

$$f(x) = \sup_{(y, a) \in \text{some set}} (-c\|x - y\|^2 + a)$$

or, since $\|x - y\|^2 = \|x\|^2 + \langle -2y, x \rangle + \|y\|^2$

$$f(x) + c\|x\|^2 = \sup_{(z, b) \in \text{some set}} (\langle z, x \rangle + b).$$

Thus a function f is Q^c -convex if and only if $f + c\|\cdot\|^2$ is convex in the ordinary sense, l.s.c. and nowhere $-\infty$. Similarly, f is Q^c -subdifferentiable if and only if $f + c\|\cdot\|^2$ is subdifferentiable in the ordinary sense. Hence the discussion reduces to investigating the convexity and subdifferentiability of $f + c\|\cdot\|^2$.

This suggests taking into consideration families $\Phi(\varphi_0)$ of the form

$$(5.8) \quad \Phi(\varphi_0) = -\varphi_0 + X^*$$

where X^* is the dual of the now arbitrary locally convex topological vector space X and the φ_0 is a fixed function on X . Clearly, a function f is $\Phi(\varphi_0)$ -convex if and only if $f + \varphi_0$ is convex, lower semicontinuous and either identically $-\infty$ or nowhere equal to $-\infty$.

If φ_0 is finite and continuous, then:

$$(5.9) \quad \text{dom}(f + \varphi_0) = \text{dom } f,$$

$$(5.10) \quad f + \varphi_0 \text{ is l.s.c. (proper) if and only if } f \text{ is l.s.c. (proper)}$$

Also, we have the following theorem.

THEOREM 5.11. *Suppose f is $\Phi(\varphi_0)$ -convex with φ_0 finite and continuous. If the interior of the epigraph of f is nonempty then f is $\Phi(\varphi_0)$ -subdifferentiable at any point of $\text{int dom } f$.*

This theorem is a generalization of a classical result [1], [21].

Proof. For f $\Phi(\varphi_0)$ -convex, $f + \varphi_0$ is a lower semicontinuous, convex proper function.

We shall show that the interior of $\text{epi}(f + \varphi_0)$ is not empty. We have $(g: X \rightarrow \bar{\mathbb{R}})$

$$\text{epi } g = \{(x, r) : r \geq g(x)\}.$$

Since $\text{int epi } g = (\text{cl}(\text{epi } g)^c)^c$ (where c denotes the complementary set and cl stands for the closure), we consider

$$\text{cl}(\text{epi } g)^c = \text{cl}\{(x, r) : r < g(x)\} = \{(x, r) : r \leq \limsup_{z \rightarrow x} g(z)\}.$$

Hence

$$(5.12) \quad \text{int epi } g = \{(x, r) : r > \limsup_{z \rightarrow x} g(z)\}.$$

From the assumption $\text{int epi } f$ is not empty and since φ_0 is continuous and finite $\text{int epi}(f + \varphi_0)$ remains nonempty. It follows [21] that $f + \varphi_0$ is subdifferentiable at every point of $\text{int dom } f$.

In order to conveniently discuss the convexity of $f + \varphi_0$ let us denote for a function $g: X \rightarrow \bar{\mathbb{R}}$

$$m_g(x, y, \lambda) = (1 - \lambda)g(x) + \lambda g(y) - g((1 - \lambda)x + \lambda y)$$

where $x, y \in X$ and $\lambda \in [0, 1]$. Clearly, g is convex if and only if $m_g \geq 0$ and strictly convex whenever $m_g(x, y, \lambda) > 0$ for $x \neq y$ and $\lambda \in (0, 1)$; also $m_{f+\varphi_0} = m_f + m_{\varphi_0}$. We close this section with some simple bounds on m_g .

PROPOSITION 5.13. (i) *Suppose g is Hölder continuous on a convex set $A \subset X$ with exponent α and constant c . Then*

$$|m_g(x, y, \lambda)| \leq c((1 - \lambda)^\alpha \lambda + (1 - \lambda)\lambda^\alpha) \|x - y\|^\alpha.$$

(ii) *Suppose g is Fréchet differentiable on an open convex set $A \subset X$ and g' is Hölder continuous on A with exponent β and constant c . Then*

$$|m_g(x, y, \lambda)| \leq c\lambda(1 - \lambda) \|x - y\|^{\beta+1}.$$

(iii) *Suppose g is twice Gâteaux differentiable on an open convex set $A \subset X$ and $g''(x) \geq m_0 I$ for $x \in A$. Then*

$$m_g(x, y, \lambda) \geq \frac{1}{2} m_0 \lambda(1 - \lambda) \|x - y\|^2.$$

Proof. Denote briefly

$$x_\lambda = (1-\lambda)x + \lambda y = x + \lambda(y-x) = y + (1-\lambda)(x-y).$$

Then

$$m_g(x, y, \lambda) = (1-\lambda)(g(x) - g(x_\lambda)) + \lambda(g(y) - g(x_\lambda)).$$

ad (i).

$$\begin{aligned} |g(x) - g(x_\lambda)| &\leq c\|x - x_\lambda\|^\alpha = c\lambda^\alpha\|x - y\|^\alpha, \\ |g(y) - g(x_\lambda)| &\leq c\|y - x_\lambda\|^\alpha = c(1-\lambda)^\alpha\|x - y\|^\alpha \end{aligned}$$

and the estimate follows.

ad (ii). By the mean value theorem

$$\begin{aligned} g(x) - g(x_\lambda) &= \langle g'(x_\theta), \lambda(x - y) \rangle, \\ g(y) - g(x_\lambda) &= \langle g'(x_\mu), (1-\lambda)(y - x) \rangle \end{aligned}$$

where $0 < \theta < \lambda < \mu < 1$. Thus

$$\begin{aligned} m_g(x, y, \lambda) &= \lambda(1-\lambda)\langle g'(x_\theta), x - y \rangle - \lambda(1-\lambda)\langle g'(x_\lambda), x - y \rangle \\ &= \lambda(1-\lambda)\langle g'(x_\theta) - g'(x_\mu), x - y \rangle \end{aligned}$$

and

$$\begin{aligned} |m_g(x, y, \lambda)| &\leq \lambda(1-\lambda)\|g'(x_\theta) - g'(x_\mu)\|\|x - y\| \\ &\leq \lambda(1-\lambda)c\|x_\theta - x_\mu\|^\beta\|x - y\| \\ &\leq \lambda(1-\lambda)c\|x - y\|^\beta + 1. \end{aligned}$$

ad (iii). Define the real function $G(\tau) = g(x_\tau)$. G is twice differentiable and

$$G''(\tau) = \langle g''(x_\tau)(y - x), y - x \rangle \geq m_0\|y - x\|^2 = m.$$

Therefore for $\tau_1 > \tau_2$

$$G'(\tau_1) = G'(\tau_2) + \int_{\tau_1}^{\tau_2} G''(\tau) d\tau \geq m(\tau_1 - \tau_2) + G'(\tau_2).$$

The remaining of the proof is classical (see e.g. [27]).

$$\begin{aligned} G(\lambda) &= G(0) + \int_0^\lambda G'(\tau) d\tau \leq G(0) + \lambda G'(\lambda) - \frac{1}{2}m\lambda^2, \\ G(1) &= G(\lambda) + \int_\lambda^1 G'(\tau) d\tau \geq G(\lambda) + (1-\lambda)G'(\lambda) + \frac{1}{2}m(1-\lambda)^2. \end{aligned}$$

Hence

$$\begin{aligned} m_g(x, y, \lambda) &= (1-\lambda)(G(0) - G(\lambda)) + \lambda(G(1) - G(\lambda)) \\ &\geq -(1-\lambda)\lambda G'(\lambda) + \frac{1}{2}m(1-\lambda)\lambda^2 + \lambda(1-\lambda)G'(\lambda) + \frac{1}{2}m(1-\lambda)^2 \\ &= \frac{1}{2}m\lambda(1-\lambda) = \frac{1}{2}m_0\lambda(1-\lambda)\|x - y\|^2. \end{aligned}$$

COROLLARY 5.14. *Suppose f is Fréchet differentiable and f' Lipschitz continuous on an open convex set A . If φ_0 is twice Gâteaux differentiable and φ_0'' is uniformly positive definite on A then there is ρ_0 such that $f + \rho_0\varphi_0$ is convex on A . When $A = X$, such f is $\Phi(\rho_0\varphi_0)$ -convex.*

6. Dense subdifferentiability of lower semicontinuous functions. Now, our objective is to study the subdifferentiability of l.s.c. functions defined on a Banach space X . Let $\Psi: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ be forcing and quietly increasing and define Φ to be

$$(6.1) \quad \Phi = \{\varphi(x) = r - \rho\Psi(\|x - y\|) : r \in \mathbb{R}, \rho > 0, y \in X\}.$$

By Proposition 4.13 and Def. 4.1, Φ is metric like, so that every l.s.c. and Φ -bounded function on X is Φ -convex. The special case of Φ with $\Psi(t) = t^\alpha$ is denoted by Φ^α .

THEOREM 6.2. *Suppose X is uniformly convex, $f: X \rightarrow \bar{\mathbb{R}}$ lower semicontinuous and Φ^α -bounded, with $\alpha \geq 1$. Then f is Φ^α -subdifferentiable on a dense subset of its domain.*

THEOREM 6.3 [46]. *Suppose X is reflexive, $f: X \rightarrow \bar{\mathbb{R}}$ weakly lower semicontinuous and Φ -bounded. Then f is Φ -subdifferentiable on a dense subset of its domain.*

These theorems should be viewed as analogues of the dense subdifferentiability of convex lower semicontinuous functions [41].

The proofs of the theorems lean on existence theorems for minimal points and the following auxiliary lemmas.

In the sequel we shall consider only functions $\varphi \in \Phi$ of the type $\varphi(x) = -\rho\Psi(\|x - y\|)$ and identify them with pairs $(\rho, y) \in \mathbb{R}^+ \times X$. We shall write $f^*(\rho, y)$ instead of $f^*(\varphi)$ and also abbreviate $\partial_\Phi f$ or $\partial_X f^*(\varphi)$ to ∂f , $\partial f^*(\rho, y)$ respectively.

LEMMA 6.4. *If $(\rho_n, y_n) \in \partial f(x_n)$ for $n = 1, 2, \dots$ and*

$$(6.5) \quad \rho_n \rightarrow \infty, \quad y_n \rightarrow x_0$$

then $x_n \rightarrow x_0$.

Proof. By definition,

$$\rho_n \Psi(\|x_n - y_n\|) - \rho_n \Psi(\|x - y_n\|) \leq f(x) - f(x_n) \quad \forall x \in X, \quad n = 1, 2, \dots$$

In particular, for $n = 1$ and $x = x_n$

$$f(x_n) \geq f(x_1) + \rho_1 \Psi(\|x_1 - y_1\|) - \rho_1 \Psi(\|x_n - y_1\|), \quad n = 1, 2, \dots$$

Taken into account both these inequalities, we have

$$\begin{aligned} & \rho_n \Psi(\|x_n - y_n\|) - \rho_n \Psi(\|x_0 - y_n\|) \\ & \leq f(x_0) - f(x_1) - \rho_1 \Psi(\|x_1 - y_1\|) + \rho_1 \Psi(\|x_n - y_1\|); \end{aligned}$$

therefore

$$\rho_n \Psi(\|x_n - y_n\|) - \rho_1 \Psi(\|x_n - y_1\|) \leq a + \rho_n \Psi(\|x_0 - y_n\|)$$

where $a = f(x_0) - f(x_1) - f(x_1) - \rho_1 \Psi(\|x_1 - y_1\|)$. Suppose now that for some $\delta > 0$ and subsequence $\{x_{n_k}\}$ of $\{x_n\}$ we have $\|x_{n_k} - y_{n_k}\| > \delta$, $k = 1, 2, \dots$. By (6.5), $\{y_{n_k}\}$ is contained in a ball B around x_0 . We may thus apply Lemma 4.4 to obtain for large k

$$\frac{1}{2}\rho_{n_k} \Psi(\|x_{n_k} - y_{n_k}\|) \leq a + \rho_{n_k} \Psi(\|x_0 - y_{n_k}\|)$$

and, because Ψ is nondecreasing

$$\Psi(\delta) \leq 2\left(\frac{a}{\rho_{n_k}} + \Psi(\|x_0 - y_{n_k}\|)\right).$$

But Ψ is forcing and the right hand side converges to zero with $k \rightarrow \infty$ by (6.5), and this contradicts that $\delta > 0$. Hence we must have $\|x_n - y_n\| \rightarrow 0$ and by (6.5) also $x_n \rightarrow x_0$.

Observe now that since f is Φ -convex then by Proposition 3.6, $(\rho, y) \in \partial f(x)$ if and only if $x \in \partial f^*(\rho, y)$. Now the assumptions of Lemma 6.4 can be written as (6.5) and $x_n \in \partial f^*(\rho_n, y_n)$.

Thus, in order to prove Theorems 6.2 and 6.3 we need only show that the set \mathcal{N} of all (ρ, y) at which f^* is subdifferentiable has the property that for any $x_0 \in X$ there exists a sequence $\{\rho_n, y_n\} \subset \mathcal{N}$ satisfying (6.5). This, however, is a trivial consequence of Proposition 3.6 and the following theorem.

THEOREM 6.6. *Suppose X, f satisfy the assumptions of Theorem 6.2 or 6.3. Let B be an open ball and let $B(\rho)$ denote the collection of all y in B such that $f(x) + \rho\Psi(\|x - y\|)$ attains a minimum on X . Then there exists $\rho_B > 0$ such that for all $\rho \geq \rho_B$, $B(\rho)$ is a G_δ -dense subset of B . If X is reflexive, f weakly l.s.c. then actually $B(\rho) = B$ for $\rho \geq \rho_B$.*

In order to apply this theorem to prove Theorems 6.2 and 6.3 it suffices to take B as some open ball around x_0 (x_0 representing any point of X). Then $\bigcup_{\rho \geq \rho_B} \{\rho\} \times B(\rho) \subset \mathcal{N}$ and \mathcal{N} possesses the above mentioned property. The aim of the following two lemmas is to reduce Theorem 6.6 to known existence theorems.

LEMMA 6.7. *Let f be Φ -bounded. For each ball B there are numbers $\bar{\rho} > 0$ and $b \in \mathbb{R}$ such that for any $\rho \geq \bar{\rho}$*

$$(6.8) \quad \inf_{y \in B} \inf_{x \in X} (f(x) + \rho\Psi(\|x - y\|)) \geq b.$$

Proof. Since f is Φ -bounded we have

$$f(x) + \rho\Psi(\|x - y\|) = f(x) + \rho_0\Psi(\|x - y_0\|) + \Delta(x, y, \rho) \geq r_0 + \Delta(x, y, \rho)$$

for some $\rho_0 > 0$, $y_0 \in X$, $r_0 \in \mathbb{R}$, where

$$\Delta(x, y, \rho) = \rho\Psi(\|x - y\|) - \rho_0\Psi(\|x - y_0\|).$$

$\Delta(\cdot, \cdot, \rho)$ is bounded on $X \times B$ for large ρ . Indeed, if $(x, y) \in B$ and $\|x - y\| \leq 1$, then

$$\Delta(x, y, \rho) \geq -\rho_0\Psi(\|x - y_0\|) \geq -\rho_0\Psi(1 + \gamma) = b$$

where $\gamma = \sup_{y \in B} \|y - y_0\|$. If $\|x - y\| > 1$ and $(x, y) \in X \times B$ then Lemma 4.4 yields

$$\Delta(x, y, \rho) \geq \frac{1}{2}\rho\Psi(\|x - y\|) \geq 0 \geq b \quad \text{for } \rho \geq \bar{\rho}.$$

LEMMA 6.9. *Consider a real function g on $X \times B$ such that*

$$\inf_{y \in B} \inf_{x \in X} g(x, y) \geq b > -\infty.$$

Let A be a subset of X and assume that for any $x \in A$ $h(x) = \sup_{y \in B} g(x, y)$ is finite. Then there is a closed ball $\bar{B}(y_0, r)$ such that for all $y \in B$

$$\inf_{x \in A} (g(x, y) + \Psi(\|x - y\|)) = \inf_{x \in A \cap \bar{B}(y_0, r)} (g(x, y) + \Psi(\|x - y\|)).$$

Proof. We may assume that A is nonempty. Define

$$r_1 = \inf_{x \in A} \|x - y_0\| + 1, \quad \gamma = \sup_{y \in B} \|y - y_0\|.$$

Take any \bar{x} from $B(y_0, r_1) \cap A$ and a number α satisfying $\alpha \geq b$ and

$$(6.10) \quad \alpha \geq h(\bar{x}) + \Psi(r_1 + \gamma).$$

Finally define for $y \in B$

$$(6.11) \quad L(y) = \{x \in A : g(x, y) + \Psi(\|x - y\|) \leq \alpha\}.$$

Observe that if $x \in L(y)$, then by (4.11)

$$\|x - y_0\| \leq \|x - y\| + \|y - y_0\| \leq \Psi^{-1}(\alpha - b) + \gamma = r$$

and consequently $L(y)$ is a subset of $B(y_0, r)$, for each $y \in B$. On the other hand for each $y \in B$,

$$L(y) \neq \emptyset.$$

This follows from the estimate based on (6.10):

$$\begin{aligned} g(\bar{x}, y) + \Psi(\|\bar{x} - y\|) &\leq h(\bar{x}) + \Psi(\|\bar{x} - y_0\| + \|y - y_0\|) \\ &\leq h(\bar{x}) + \Psi(r_1 + \gamma) \leq \alpha. \end{aligned}$$

The thesis of the lemma follows.

We shall apply a theorem of Bidaut [6] which generalizes the theorems of Edelstein [12] and Baranger [1]. It reads as follows:

THEOREM 6.12 [7, Thm. 4.2]. *Let X be a uniformly convex Banach space and let S be a closed subset of X . Assume that a function $F: S \rightarrow \bar{\mathbb{R}}$ is lower semicontinuous and bounded from below, and $\alpha \geq 1$. Then there exists a G_δ -dense subset D of X such that for any $y \in D$ the function $F(x) + \|x - y\|^\alpha$ attains its minimum over S .*

Proof of Theorem 6.6. Assume first that f is finite on X . Let B be an arbitrary ball. We use Lemma 6.7 to obtain formula (6.8) a fortiori valid for all $\rho \geq \rho_B = \bar{\rho} + 1$. Define

$$g(x, y) = f(x) + (\rho - 1)\Psi(\|x - y\|)$$

and apply Lemma 6.9 to obtain ($A = X$)

$$(6.13) \quad \inf_{x \in X} (f(x) + \rho\Psi(\|x - y\|)) = \inf_{x \in S} (f(x) + \rho\Psi(\|x - y\|))$$

where $S = \bar{B}(y_0, r)$ is a closed, convex, bounded set.

Case 1. X uniformly convex, f lower semicontinuous, $\Psi(t) = t^\alpha$, $\Phi = \Phi^\alpha$. Let $F(x) = (1/\rho)f(x)$; F is bounded from below on S . In fact, for $x \in S$

$$F(x) = \frac{1}{\rho}f(x) \geq \frac{r_0}{\rho} - \frac{\rho_0}{\rho}\|x - y_0\|^\alpha \geq \frac{r_0 - \rho_0 r^\alpha}{\rho}.$$

We conclude that F satisfies all the assumptions of Theorem 6.6. It suffices now to put $B(\rho) = B \cap D$ where D is defined in Theorem 6.12 and to observe that

$$\inf_{x \in X} (f(x) + \rho\|x - y\|^\alpha) = \rho \inf_{x \in S} (F(x) + \|x - y\|^\alpha).$$

Case 2. X reflexive, f weakly lower semicontinuous. The function

$$k_y(x) = f(x) + \rho\Psi(\|x - y\|)$$

is weakly lower semicontinuous and bounded from below, while the set S is weakly compact. The thesis follows from the theorems of Eberlein–Smulian and Weierstrass and from (6.13).

If f is not finite, take $A = \text{dom } f$ in Lemma 6.9 and substitute S by $A \cap S$ in the sequel.

Remark 6.14. Part of Theorem 6.2 for $\alpha = 1$ can be proved much more easily without assuming the uniform convexity of X , as pointed out by the referee. Fix $x_0 \in X$ and $\varepsilon > 0$. By Φ^1 -convexity, $f(x_0) = f^{**}(x_0)$ and there are $\rho > 0$, $y \in X$ such that

$$f(x_0) + \rho\|x_0 - y\| \leq \inf_{x \in X} (f(x) + \rho\|x - y\|) + \varepsilon^2.$$

The variational principle of Ekeland [45] now yields the existence of an $x_\varepsilon \in K(x_0, \varepsilon)$ such that $g(x) = f(x) + \rho\|x - y\| + \varepsilon\|x - x_\varepsilon\|$ attains a global minimum at x_ε . Therefore

$$\begin{aligned} f(x) - f(x_\varepsilon) &\geq -\rho(\|x - y\| - \|x_\varepsilon - y\| - \varepsilon\|x - x_\varepsilon\|) \\ &\geq -\rho_\varepsilon\|x - x_\varepsilon\| \end{aligned}$$

so that $(\rho_\varepsilon, x_\varepsilon) \in \partial f(x_\varepsilon)$.

7. Applications to extremal problems. Suppose that U, X are two abstract sets and consider a family $\{\Gamma x\}_{x \in X}$ of subsets of U indexed by X . Let Q be an (extended) real function on U . Define a family of optimization problems

$$(7.1) \quad Q(u) \rightarrow \inf, \quad u \in \Gamma x.$$

Such a formulation is very simple and encompasses problems with constraints of practically any nature. In spite of its generality the problem admits several nontrivial results [23].

The family $\{\Gamma x\}_{x \in X}$ may be interpreted as a multifunction from X to subsets of U . We shall be constantly using the inverse multifunction Γ^{-1} :

$$(7.2) \quad \Gamma^{-1}u = \{x : u \in \Gamma x\}.$$

Define the Lagrange function associated to $(7.1)_{x_0}$ with respect to a family Φ :

$$(7.3) \quad L(u, \varphi, x_0) = Q(u) - \sup_{x \in \Gamma^{-1}u} \varphi(x) + \varphi(x_0).$$

Define also the primal functional

$$(7.4) \quad \hat{Q}_\Gamma(x) = \inf_{u \in \Gamma x} Q(u).$$

It is always assumed that $\alpha\varphi + r \in \Phi$, whenever $\alpha > 0$, $r \in \mathbb{R}$ and $\varphi \in \Phi$.

PROPOSITION 7.5. Assume that Γ is nondegenerate:

$$(7.6) \quad \bigcap_{x \in X} \Gamma x = \emptyset$$

and that $\Gamma^{-1}u$ is Φ -convex for $u \in U$. Then

$$(7.7) \quad \inf_{u \in U} \sup_{\varphi \in \Phi} L(u, \varphi, x_0) = \hat{Q}_\Gamma(x_0).$$

Remark. It is always true that

$$(7.8) \quad \sup_{\varphi \in \Phi} \inf_{u \in U} L(u, \varphi, x_0) \leq \inf_{u \in U} \sup_{\varphi \in \Phi} L(u, \varphi, x_0).$$

Proof. Consider now $\sup_{\varphi \in \Phi} L(u, \varphi, x_0) = \sup_{\varphi \in \Phi} (Q(u) - \sup_{x \in \Gamma^{-1}u} \varphi(x) + \varphi(x_0))$ for $u \notin \Gamma x_0$. For such u one has that $x_0 \notin \Gamma^{-1}u$ and because $\Gamma^{-1}u$ is Φ -convex there exists a φ_0 such that $-\sup_{x \in \Gamma^{-1}u} \varphi_0(x) + \varphi_0(x_0) > 0$. Since $\alpha\varphi_0 \in \Phi$ for any $\alpha > 0$, it follows that $\sup_{\varphi \in \Phi} L(u, \varphi, x_0) = +\infty$, if $u \notin \Gamma x_0$.

If $u \in \Gamma x_0$, then $-\sup_{x \in \Gamma^{-1}u} \varphi(x) + \varphi(x_0) \leq 0$. To see that $-\sup_{x \in \Gamma^{-1}u} \varphi(x)$ is finite, observe that in view of (7.6) there is an $x_1 \notin \Gamma^{-1}u$ and thus for some φ_1 , $-\varphi_1(x_1) < -\sup_{x \in \Gamma^{-1}u} \varphi_1(x)$ because $\Gamma^{-1}u$ is Φ -convex. Therefore $\sup_{\varphi \in \Phi} L(u, \varphi, x_0) = Q(u)$ for $\alpha\varphi \in \Phi$ for each $\alpha > 0$.

We conclude that $\inf_{u \in U} \sup_{\varphi \in \Phi} L(u, \varphi, x_0) = \hat{Q}_\Gamma(x_0)$.

THEOREM 7.9 [23]. *The weak duality holds:*

$$(7.10) \quad \sup_{\varphi \in \Phi} \inf_{u \in U} L(u, \varphi, x_0) = \hat{Q}_\Gamma(x_0),$$

if and only if \hat{Q}_Γ is Φ -convex at x_0 .

The strong duality holds:

$$(7.11) \quad \inf_{u \in U} L(u, \varphi_0, x_0) = \hat{Q}_\Gamma(x_0),$$

for some $\varphi_0 \in \Phi$, if and only if \hat{Q}_Γ is Φ -subdifferentiable at x_0 .

Proof. The proof is based on straightforward computations [23]:

$$\begin{aligned} \inf_{u \in U} L(u, \varphi, x_0) &= \inf_{u \in U} (Q(u) + \inf_{x \in \Gamma^{-1}u} (-\varphi(x)) + \varphi(x_0)) \\ &= \inf_{\substack{u, x \\ u \in \Gamma x}} (Q(u) - \varphi(x)) + \varphi(x_0) \\ (7.12) \quad &= \inf_{x \in X} (\inf_{u \in \Gamma x} Q(u) - \varphi(x)) + \varphi(x_0) \\ &= -\hat{Q}_\Gamma^*(\varphi) + \varphi(x_0) \end{aligned}$$

and thus

$$(7.13) \quad \sup_{\varphi \in \Phi} \inf_{u \in U} L(u, \varphi, x_0) = \sup_{\varphi \in \Phi} (\varphi(x_0) - \hat{Q}_\Gamma^*(\varphi)) = \hat{Q}_\Gamma^{**}(x_0) = \hat{Q}_\Gamma^0(x_0)$$

in view of Def. 1.9. As $\hat{Q}_\Gamma^0(x_0) = \hat{Q}_\Gamma(x_0)$ means precisely that \hat{Q}_Γ is Φ -convex at x_0 the first assertion is proved. Equalities (7.11), (7.12) combined yield the formula

$$(7.14) \quad -\hat{Q}_\Gamma^*(\varphi_0) + \varphi_0(x_0) = \hat{Q}_\Gamma(x_0),$$

which in virtue of Proposition 3.5 is equivalent to

$$(7.15) \quad \varphi_0 \in \partial_\Phi \hat{Q}_\Gamma(x_0).$$

Theorem 7.9 shows how according to the regularity of the problem (that is, of the primal functional) we may replace the original constrained problem $(7.1)_{x_0}$ by a problem without constraints,

$$(7.16) \quad \text{compute } \inf_{u \in U} L(u, \varphi, x_0),$$

and by the dual (unconstrained) problem,

$$(7.17) \quad \text{maximize } \inf_{u \in U} L(u, \varphi, x_0), \quad \varphi \in \Phi.$$

In the case of (7.10) we shall be looking for appropriate (generalized) sequences $\{\varphi_\beta\}$ that approximate well the original problem. If we can establish (7.11), $\{\varphi_\beta\}$ may be replaced by one φ_0 .

The most favorable situation is when φ_0 supports \hat{Q}_Γ precisely at x_0 (x_0 is the only subgradient of \hat{Q}_Γ^* at φ_0). This property, described in § 5, guarantees that $(7.1)_{x_0}$ and the problem

$$(7.18) \quad L(u, 0, x_0) \rightarrow \inf, \quad u \in U$$

are equivalent, if $\{\Gamma_u^{-1}\}$ are one point sets (see [40], [42] for other conditions).

The applicability of the theory becomes clear, when we observe, as it was shown in [23], that the generalized Lagrange functional includes Lagrangians used by Mangasarian [25], Bertsekas [4], Buys [7], Rockafellar [28], Wierzbicki [37], [38] and others. We shall provide an example.

Example 7.19 (see [7], [28]).

$$\begin{aligned} & \text{minimize } f_0(y), \\ & f_i(y) \geq 0, \quad i = 1, 2, \dots, m \end{aligned}$$

where Y is an arbitrary set and f_i are real functions on Y . We insert this problem into the class of minimization questions of f_0 over

$$\{y: f_i(y) \leq x_i, \quad i = 1, 2, \dots, m\}.$$

Here the space of indices X may be identified with \mathbb{R}^m . The class $\Phi^2 = \{-\rho\|\cdot - z\|^2 + r: \rho > 0, z \in \mathbb{R}^m, r \in \mathbb{R}\}$ satisfies the assumptions of Proposition 7.5. In a Hilbert space we have:

$$\begin{aligned} \rho\|x - z\|^2 &= \rho\|x\|^2 - 2\rho\langle z, x \rangle + \rho\|z\|^2 \\ &= \rho\|x\|^2 + \langle \lambda, x \rangle + s. \end{aligned}$$

We observe that the minimum of this function over $D = \{x \in X: x_i \geq f_i(y), i = 1, 2, \dots, m\}$ is attained at the orthogonal projection of z on the cone D [38]

$$x_0 = \left\{ \max \left(-\frac{\lambda_i}{2\rho}, f_i(y) \right) \right\}.$$

Therefore (6.5) at $x_0 = 0$ becomes $(\varphi \sim (\rho, \lambda, s))$

$$L(y, \varphi, 0) = f_0(y) + \sum_{i=1}^n \left[\max^2 \left(-\frac{\lambda_i}{2\rho}, f_i(y) \right) + \lambda_i \max \left(-\frac{\lambda_i}{2\rho}, f_i(y) \right) \right]$$

and this is just the function of Buys [7] and Rockafellar [28].

Suppose from now on that X is a metric space and that Φ is metric-like.

Suppose moreover that U is a topological space and $S: U \times X \rightarrow X$ an operator continuous with respect to x for each u . Let D be a closed subset of X and put

$$\Gamma x = \{u \in Y: S(y, x) \in D\}.$$

Then, for $x_0 \notin \Gamma_u^{-1}$,

$$\inf_{x \in \Gamma_u^{-1}} \eta(x, x_0) > \eta(x_0, x_0) = 0$$

and hence family Φ satisfies the assumptions of Proposition 7.5. We can therefore define the augmented Lagrangian

$$(7.20) \quad L(y, \rho, z; x_0) = Q(y) + \rho \inf_{\{x: S(u, x) \in D\}} \eta(x, z) - \rho \eta(x_0, z).$$

for a family of optimization problems

$$\text{minimize } Q(y), \quad S(y, x) \in D.$$

Example 7.21. Suppose X is any of the spaces $C(\Omega)$, $L^p(\Omega)$, $1 \leq p \leq \infty$ where Ω is a closed subset of \mathbb{R}^n . Let D be the cone of functions from X nonnegative on Ω . Let $F: Y \rightarrow X$ and $S(x, y) = x - F(y)$. Then

$$S(y, x) \in D \Leftrightarrow F(y) \leq x.$$

Let Ψ be a quietly increasing forcing function, for instance equal to t^α , $\alpha > 0$, $e^{\alpha t} - 1$, $\alpha > 0$ or a weighted sum of these. Put $\eta(x, z) = \Psi(\|x - z\|)$. Then

$$\inf_{\{x: S(y, x) \in D\}} \eta(x, z) = \Psi(\|(F(y) - z)_+\|)$$

where $x_+(\omega) = \max(0, x(\omega))$, $x \in X$ for $\omega \in \Omega$ (compare [23]). Thus the Lagrangian (7.20) becomes

$$L(y, \rho, z; x_0) = Q(y) + \rho \Psi(\|(F(y) - z)_+\|) - \rho \Psi(\|x_0 - z\|).$$

With regard to Lagrangian (7.20) we can draw the following immediate consequences from the theorems of §§ 4 and 5.

- (i) If the primal functional is lower semicontinuous at x_0 and Φ -bounded then the weak duality holds.
- (ii) If the primal function is Φ -bounded and η -calm at x_0 then the strong duality holds.

Both these statements generalize respective theorems of [28] $X = \mathbb{R}^n$ and [38] for X being a Hilbert space, with $\Psi(t) = t^2$.

- (iii) If the primal functional is Φ -bounded and η -calm at x_0 then a sufficient and necessary condition for $u_0 \in \Gamma x_0$ to solve globally (7.1) $_{x_0}$ is the existence of a $\rho > 0$ such that \hat{y} minimizes $L(\cdot, \rho, x_0; x_0)$ over U .

This regards the theory of optimality conditions in absence of differentiability hypotheses, as developed for instance by Clarke [9], Ioffe and Tikhomirov [21] and also the theory of so-called exact penalty functionals.

- (iv) Suppose we face the situation as described in Example 7.21 with $X = L^p(\Omega)$, $1 < p < +\infty$ and $\Psi(t) = t^\alpha$, $\alpha \geq 1$. Then Theorem 6.2 on dense Φ^α -subdifferentiability implies that the set of all points $x_0 \in X$ at which strong duality holds is a dense subset of X , provided \hat{Q}_Γ is lower semicontinuous and Φ^α -bounded.

We may also apply Theorem 7.9 to a problem of moments (see [29], [33], [30], [11]). Given a proper convex function Q on a Banach space U , bounded linear operator $C: U \rightarrow X$ we ask when does

$$(7.22) \quad \inf_{Cu=x} Q(u) = \sup_{\varphi \in X^*} \inf_u \{Q(u): \varphi(Cu - x) = 0\}.$$

A necessary and sufficient condition is that \hat{Q}_Γ be l.s.c. at x and nowhere $-\infty$ (it is always convex). It is l.s.c., if and only if $\overline{C\{u: Q(u) \leq \alpha\}} \subset \bigcap_{\varepsilon > 0} C\{u: Q(u) \leq \alpha + \varepsilon\}$ [11].

8. Comments. A prototype of Theorem 4.2 may be found in a work of Rockafellar [28]. Theorem 4.2 specifies as a well-known fact [14] that in completely regular spaces each l.s.c. function bounded from below by a continuous function is the least upper bound of some continuous functions. Some special results of this kind are contained in [47], but there the support is assumed compact and thus much of the essence of Theorem 4.2 is dropped.

Balder [40] presents an abstract version of this theorem. X is supposed to be a topological space. The family Φ is said to be of *needle type* at x_0 , if for each neighborhood W of x_0 and for each $r \in \mathbb{R}$ and each $\varphi_0 \in \Phi$, there exists a $\varphi \in \Phi$ such that

$$(8.1) \quad \varphi(x) \leq \varphi(x_0) = r \quad \text{for } x \in W,$$

$$(8.2) \quad \varphi(x) \leq \varphi_0(x) \quad \text{for } x \notin W.$$

Using the axioms (8.1), (8.2) it is now easy to show that for Φ -being of needle type at x_0 , every Φ -bounded function l.s.c. at x_0 is Φ -convex at x_0 . A slightly weaker notion of *sharpness* of Φ was introduced by Lindberg [48] to prove an analogous result. In turn, examples furnished in [40, Lemma 1] are generated by metric-like families.

The main effect of Theorem 4.2 and of the subsequent considerations was to provide verifiable criteria for the needle type classes Φ . Another important notion introduced by Balder was that of flexibility. Φ is said to be flexible at x_0 , if for each neighborhood W of x_0 and for each $\varphi_1, \varphi_2 \in \Phi$ there is a φ and a neighborhood W_0 of x_0 , $W_0 \subset W$, such that

$$(8.3) \quad \varphi(x) \leq \varphi_1(x) \quad \text{for } x \notin W_0,$$

$$(8.4) \quad \varphi(x) \leq \varphi_2(x) \quad \text{for } x \in W_0.$$

The flexibility enables a reasonable approximation of $(7.1)_{x_0}$ by the problems (7.18) and, together with some weak additional assumptions, guarantees the equivalence of $(7.1)_{x_0}$ and (7.18). More research in this direction would be welcome.

In the case when $\alpha = 1$, Theorem 6.2 may be formulated for arbitrary Banach spaces and it possesses a simple proof based on the Ekeland theorem [45].

For our applications it is essential to know under what assumptions on a constraints multifunction Γ , the primal functional \hat{Q}_Γ is Φ -convex (Φ -subdifferentiable) for the functions Q from a predetermined family F .

The main difficulty in view of Theorem 4.2, Proposition 5.6 and Example 1.10 is to check the lower semicontinuity and the η -calmness of \hat{Q}_Γ . These questions were confronted in the recent papers [42], [43], [44]. For instance, it is shown in [44] that \hat{Q}_Γ is l.s.c. at x_0 for each $Q \in F$, if and only if Γ is F -stable at x_0 : for each $Q \in F$

$$(8.5) \quad y_0 \in \bigcup_{\varepsilon > 0} \{x : \Gamma x \subset \{u : Q(u) > r + \varepsilon\}\}$$

entails

$$(8.6) \quad y_0 \in \text{int} \{x : \Gamma x \subset \{u : Q(u) > r\}\}.$$

Special cases of this proposition were already observed in [11], [23].

F -stable multifunctions in case of F —the class of all l.s.c. functions—are precisely upper semicontinuous multifunctions. The η -calmness is related to the upper Hausdorff semicontinuity [42].

A thorough treatment of similar stability problems for global and local extremal problems is given in [43].

REFERENCES

- [0] E. ASPLUND, *Fréchet differentiability of convex functions*, Acta Math., 121 (1968), pp. 31–47.
- [1] J. BARANGER, *Existence de solutions pour des problèmes d'optimisation non convexe*, C. R. Acad. Sci. Paris Sér. A-B, 274 (1972), pp. A307–A309.
- [2] M. BELLMORE, H. J. GREENBERG AND J. J. JARVIS, *Generalized penalty-function concepts in mathematical optimization*, Operations Res., 18 (1970), no. 2.
- [3] C. BERGE, *Espaces Topologiques*, Dunod, Paris, 1966.
- [4] D. P. BERTSEKAS, *Combined primal-dual and penalty method for constrained minimization*, this Journal, 13 (1975), pp. 521–544.
- [5] ———, *Multiplier methods: A survey*, Automatica, 12 (1976), pp. 133–146.
- [6] M. D. BIDAUT, *Théorèmes d'existence et d'existence "en général" d'un contrôle optimal pour des systèmes régis par des équations aux dérivées partielles non linéaires*, Thesis, Université Paris VI, 1973.
- [7] J. D. BUYS, *Dual algorithms for constrained optimization*, Thesis, Rijkuniversiteit te Leiden, The Netherlands, 1972.

- [8] F. H. CLARKE, *Generalized gradients and applications*, Trans. AMS, 205 (1975), pp. 247–262.
- [9] ———, *A new approach to Lagrange multipliers*, Mathematics in Operations Research, 1 (1976), pp. 165–174.
- [10] S. DOLECKI, *Bounded controlling sequences, lower stability and certain penalty procedures*, Appl. Math. Optim., to appear.
- [11] S. DOLECKI AND S. KURCYSZ, *Convexité généralisée et optimisation*, C.R.Acad. Sci. Paris Ser. A-B, 283 (1976), pp. A-91-94.
- [12] M. EDELSTEIN, *On nearest points of sets in uniformly convex Banach spaces*, J. London Math. Soc., 48 (1968), pp. 375–377.
- [13] K. H. ELSTER AND R. NEHSE, *Zur Theorie der Polarfunktionale*, Math. Operationsforsch. Statist., 5 (1974), pp. 3–21.
- [14] R. ENGELKING, *Outline of General Topology*, North-Holland, Amsterdam, 1968.
- [15] J. P. EVANS AND F. J. GOULD, *Stability in nonlinear programming*, Operations Res., 18 (1970), pp. 107–118.
- [16] H. EVERETT, III, *Generalized Lagrange multipliers methods for solving problems of optimum allocation of resources*, Ibid., 11 (1963), pp. 399–417.
- [17] K. FAN, *On the Krein–Milman theorem*, Convexity, Proc. of Symposia in Pure Mathematics, vol. VII, American Mathematical Society, Providence, RI, 1963, pp. 211–220.
- [18] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming*, John Wiley, New York, 1968.
- [19] F. J. GOULD, *Extensions of Lagrange multipliers in nonlinear programming*, SIAM J. Appl. Math., 17 (1969), pp. 1280–1297.
- [20] M. R. HESTENES, *Multiplier and gradient methods*, Computing Methods in Optimization Problems, 2nd ed., L. A. Zadeh, L. W. Neustadt, A. V. Balakrishnan, eds., Academic Press, New York, 1969.
- [21] A. D. IOFFE AND W. M. TIKHOMIROV, *Theory of Extremal Problems*, Nauka, Moscow, 1974. (In Russian.)
- [22] J. L. JOLY AND P. J. LAURENT, *Stability and duality in convex minimization problems*, Rev. Française Informant. Recherche Opérationnelle, 5 (1971), R-2, pp. 3–42.
- [23] S. KURCYSZ, *Some remarks on generalized Lagrangians*, Proc. 7th IFIP Conference on Optimization Techniques, Nice, France, September 1975; Springer-Verlag, 1976.
- [24] P. J. LAURENT, *Approximation et Optimisation*, Hermann, Paris, 1972.
- [25] O. L. MANGASARIAN, *Unconstrained Lagrangians in nonlinear programming*, this Journal, 13 (1975), pp. 772–791.
- [26] M. J. D. POWELL, *A method for nonlinear constraints in minimization problems*, Optimization, R. Fletcher, ed., Academic Press, New York, 1969.
- [27] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.
- [28] ———, *Augmented Lagrange multiplier functions and duality in nonconvex programming*, this Journal, 12 (1974), pp. 268–285.
- [29] S. ROLEWICZ, *On a problem of moments*, Studia Math., 3 (1968), pp. 183–191.
- [30] ———, *Functional Analysis and Control Theory*, PWN, Warszawa, 1974. (In Polish.)
- [31] J. D. ROODE, *Generalized Lagrange functions and mathematical programming*, Optimization, R. Fletcher, ed., Academic Press, New York, 1969.
- [32] K. H. SEIDLER, *Zur Dualisierung in der nichtlinearen Optimierung*, Thesis, Technische Hochschule Ilmenau (GDR), 1972.
- [33] I. SINGER, *On a problem of moments of S. Rolewicz*, Studia Math., 48 (1973), pp. 95–98.
- [34] W. VOGEL, *Duale Optimierungsaufgaben und Sattelpunktsätze*, Unternehmensforschung, 1969, vol. 13, pp. 1–28.
- [35] E. A. WEISS, *Konjugierte Funktionen*, Arch. Math. (Brno), 20 (1969), pp. 538–545.
- [36] A. P. WIERZBICKI, *A penalty function shifting method in constrained static optimization and its convergence properties*, Arch. Automat. Telemech., 16 (1971), pp. 395–416.
- [37] A. P. WIERZBICKI AND A. HATKO, *Computational methods in Hilbert space for optimal control problems with delays*, Proc. of the 5th IFIP Conference on Optimization Techniques, Rome, 1973; Springer-Verlag, 1975.
- [38] A. P. WIERZBICKI AND S. KURCYSZ, *Projection on a cone, penalty functionals and duality theory for problems with inequality constraints in Hilbert space*, this Journal, 15 (1977), pp. 25–56.
- [39] W. I. ZANGWILL, *Nonlinear Programming*, Prentice-Hall, Englewood Cliffs, N.J., 1969.
- [40] E. J. BALDER, *An extension of duality-stability relations to nonconvex optimization problems*, this Journal, 15 (1977), pp. 329–343.
- [41] A. BRØNSTED AND R. T. ROCKAFELLAR, *On the subdifferentiability of convex functions*, Proc. Amer. Math. Soc., 16 (1965), pp. 605–611.

- [42] S. DOLECKI, *Constraints stability and moduli of semicontinuity*, 2nd IFAC Symp. on Distributed Parameter Systems, Warwick, England, June 1977, to appear.
- [43] ———, *Semicontinuity in constrained optimization, I and II*, Control and Cybernetics, to appear.
- [44] S. DOLECKI AND S. ROLEWICZ, *A characterization of semicontinuity-preserving multifunctions*, J. Math. Anal. Appl., to appear.
- [45] I. Ekeland, *A variational principle*, Ibid., 47 (1974), pp. 324–353.
- [46] S. KURCYUSZ, *On penalty methods in dynamic optimization*, presented at the 21 Internationales Wissenschaftliches Kolloquium TH Ilmenau, Ilmenau (GDR), November 1976.
- [47] S. S. KUTATELADZE AND A. M. RUBINOV, *Minkowski duality and applications*, Nauka, Novosibirsk, U.S.S.R., 1976. (In Russian.)
- [48] P. O. LINDBERG, *A generalization of Fenchel conjugation giving generalized Lagrangians and symmetric nonconvex duality*, Rep. TRITA-MAT-1976-12 (Aug), Dept. of Math. Royal Institute of Technology, Stockholm, 1976.
- [49] J. J. MOREAU, *Fonctionnelles convexes*, Séminaire sur les équations aux dérivées partielles, Collège de France, Paris 1966.
- [50] L. DANZER, B. GRÜNBAUM AND V. KLEE, *Helly's theorem and its relatives*, Convexity, Proc. of Symposia in Pure Mathematics, vol. VII, American Mathematical Society, Providence, RI, 1963, pp. 101–180.

GRADIENTS GENERALISES DE FONCTIONS MARGINALES*

J. B. HIRIART-URRUTY†

Abstract. In this paper, we give different evaluations of generalized gradients of functions defined by: $\varphi_F(x) = \inf \{f(x, y) | y \in F(x)\}$. In this expression, f is a locally Lipschitz function on $X \times Y$ and we examine successively the cases: $F(x) = Y$, $F(x) = F$ for all x , and F an arbitrary set-valued mapping with closed graph.

1. Introduction. X et Y étant deux espaces vectoriels euclidiens de dimension finie, considérons $f: X \times Y \rightarrow \mathbb{R}$ et F une multiapplication définie sur X et à valeurs dans les parties non vides de Y . On s'intéresse, de manière générale, à la fonction φ_F définie sur X par:

$$(A) \quad \varphi_F(x) = \inf \{f(x, y) | y \in F(x)\},$$

c'est-à-dire des *fonctions marginales à contraintes dépendantes*.

Outre les problèmes de programmation stochastique, la théorie des jeux avec transfert d'information, l'étude mathématique de modèles économiques, etc., conduisent à des problèmes d'optimisation où la fonction à minimiser a la forme indiquée en (A). De nombreux articles ont été consacrés à l'étude des propriétés de φ_F ; des conditions sur f et sur F ont été données assurant la continuité de φ_F [12]; de même l'existence et l'expression des dérivées directionnelles de φ_F ont été étudiées [2], [7], [8], [9]. Les hypothèses habituellement faites sur f sont des hypothèses de convexité (partielle ou totale) et de différentiabilité; les multiapplications F sont usuellement définies sous la forme

$$\forall x \in X, \quad F(x) = \{y \in Y | g_i(x, y) \leq 0, \quad \forall i \in \langle 1, m \rangle\} \cap F_0.$$

Dans notre étude, nous considérerons des fonctions f localement lipschitziennes. Pour de telles fonctions, F. H. Clarke [4] a introduit la notion de gradient généralisé et nous nous poserons le problème de la détermination (ou du moins d'une estimation) du gradient généralisé de fonctions du type φ_F décrit en (A).

Après avoir, dans un premier paragraphe, rappelé la définition et les propriétés essentielles du gradient généralisé, nous nous attacherons à donner des évaluations de gradients généralisés pour des fonctions marginales φ_F en allant de la situation la plus simple à la plus générale, c'est-à-dire que l'on examinera successivement les cas: $F(x) = Y$, $F(x) = F$ pour tout x , et F une multiapplication quelconque. Nous montrerons comment ces estimations généralisent les résultats (donnés sous forme d'égalité en général) connus dans le cas complètement (ou partiellement) convexe [12], [2] et dans le cas différentiable [2], [7], [8], [9].

Les principaux résultats de cet article ont été annoncés dans un Compte Rendu de l'Académie des Sciences de Paris (t. 283, Série A, 1976, pp. 333-335).

2. Notations et préliminaires. Dans toute la suite, lorsqu'on parlera d'espaces X , Y ou Z , il s'agira d'espaces vectoriels euclidiens de dimension finie (identifiés à un espace \mathbb{R}^p par exemple) munis du produit scalaire désigné par $\langle \cdot, \cdot \rangle$. Pour $x_0 \in X$, $\mathcal{V}(x_0)$ sera le filtre des voisinages de x_0 et on notera $L_{\text{loc}}^{\text{ip}}(X)$ l'espace des fonctions f localement lipschitziennes sur X , c'est-à-dire telles que:

$$\forall x_0 \in X, \quad \exists V_0 \in \mathcal{V}(x_0), \quad \exists k > 0, \quad \forall x, y \in V_0, \quad |f(x) - f(y)| \leq k \|x - y\|.$$

* Received by the editors October 20, 1976, and in revised form May 26, 1977.

† Université de Clermont, Complexe Scientifique des Cézeaux, Département de Mathématiques Appliquées, Boite Postale 45, 63170 Aubière, France.

Pour $f \in L_{\text{loc}}^{\text{ip}}(X)$ et $x_0 \in X$, le *gradient généralisé* de f en x_0 , défini par F. H. Clarke [4], est le convexe compact noté $\partial f(x_0)$ dont la fonction d'appui est:

$$(2.1) \quad \forall d \in X, \quad f^*(x_0; d) = \limsup_{\substack{x \rightarrow x_0 \\ \lambda \rightarrow 0^+}} \frac{f(x + \lambda d) - f(x)}{\lambda}.$$

Similairement:

$$(2.2) \quad \forall d \in X, \quad f_*(x_0; d) = -f^*(x_0; -d) = \liminf_{\substack{x \rightarrow x_0 \\ \lambda \rightarrow 0^+}} \frac{f(x + \lambda d) - f(x)}{\lambda}.$$

Le lien avec la notion de sous-différentiel d'une fonction convexe et la F -dérivée (dérivée au sens de Fréchet) d'une fonction différentiable est explicité dans la proposition suivante.

PROPOSITION 1. *Si $f: X \rightarrow \mathbb{R}$ est convexe, le gradient généralisé de f en x_0 est le sous-différentiel de f en x_0 .*

Si f est F -différentiable en x_0 de F -dérivée forte en x_0 [16, p. 71], alors $\partial f(x_0) = \{\nabla f(x_0)\}$.

Q étant un fermé non vide de X , F. H. Clarke définit la notion de cône normal à Q en $x_0 \in Q$ par l'intermédiaire du gradient généralisé de la fonction distance d_Q ; rappelons cette définition, ainsi qu'une propriété infinitésimale équivalente [4, Corollaire 1.20].

DEFINITION 1. Le *cône normal* à Q en $x_0 \in Q$ est l'ensemble noté $N(Q; x_0)$ (ou $N_Q(x_0)$) défini par l'une des propriétés équivalentes suivantes:

$$(2.3) \quad \begin{aligned} (i) \quad & N(Q; x_0) = \overline{\text{co}} \partial d_Q(x_0) \text{ (enveloppe cônica fermée de } \partial d_Q(x_0)), \\ (ii) \quad & N(Q; x_0) = \text{co} \{ \lim_{n \rightarrow \infty} \lambda_n (x_n - \bar{x}_n) \} \text{ avec } \lambda_n > 0, x_n \rightarrow x_0 \text{ et } \bar{x}_n \text{ un point de } Q \text{ à distance minimum de } x_n. \end{aligned}$$

Si $x_0 \notin Q$, nous conviendrons que $N(Q; x_0) = \emptyset$. Si Q est convexe, $N(Q; x_0)$ est le cône normal de l'analyse convexe [18, p. 15] et si Q est une sous-variété C^1 de X , $N(Q; x_0)$ coïncide avec la normale au sens de la géométrie différentielle.

Les propriétés algébriques et topologiques du gradient généralisé sont étudiées dans [4] et [5]; rappelons celles dont l'utilisation sera constante par la suite.

PROPOSITION 2. (i) ∂f est une multiapplication semi-continue supérieurement (s.c.s) et fermée sur X .

(ii) Si f_1 et f_2 sont localement lipschitziennes sur X et $x_0 \in X$,

$$\partial(f_1 + f_2)(x_0) \subset \partial f_1(x_0) + \partial f_2(x_0).$$

3. Gradients généralisés de fonctions marginales. Rappelons au préalable quelques notations et définitions relatives aux multiapplications.

Soit M une multiapplication de X dans Y à valeurs non vides sur un voisinage de $x_0 \in X$. Dans toute la suite, on utilisera les notations suivantes:

$$\text{LI}(M; x_0) = \liminf_{x \rightarrow x_0} M(x),$$

$$\text{LS}(M; x_0) = \limsup_{x \rightarrow x_0} M(x),$$

c'est-à-dire, $\text{LI}(M; x_0)$ et $\text{LS}(M; x_0)$ sont respectivement la limite inférieure et la limite

supérieure de la famille filtrée $(M(x)|x \in X; \mathcal{V}(x_0))$ [3, p. 126]. En d'autres termes:

$$LI(M; x_0) = \{u \in Y \mid \lim_{x \rightarrow x_0} d(u, M(x)) = 0\},$$

$$LS(M; x_0) = \{u \in Y \mid \liminf_{x \rightarrow x_0} d(u, M(x)) = 0\}.$$

Une application $\bar{y}: X \rightarrow Y$ sera appelée *sélection de M au voisinage de x_0* s'il existe $V_0 \in \mathcal{V}(x_0)$ tel que:

$$\forall x \in V_0, \quad \bar{y}(x) \in M(x).$$

Si de plus on a:

$$\sup \{\|\bar{y}(x)\| \mid x \in V_0\} < +\infty,$$

on dira que \bar{y} est une sélection de M bornée au voisinage de x_0 . Si \bar{y} est une sélection de M au voisinage de x_0 , on notera \bar{Y}_0 l'ensemble des valeurs d'adhérence de $\bar{y}(x)$ quand $x \rightarrow x_0$.

3.1. Considérons $f \in L_{loc}^{ip}(X \times Y)$; désignons par φ la *fonction marginale* définie par:

$$(3.1) \quad \forall x \in X, \quad \varphi(x) = \inf \{f(x, y) \mid y \in Y\}$$

et par M la multiapplication

$$(3.2) \quad M(x) = \{y \in Y \mid f(x, y) = \varphi(x)\}.$$

Des hypothèses seront faites par la suite pour qu'au voisinage du point x_0 , $M(x) \neq \emptyset$. Nous avons alors le théorème suivant:

THEOREME 1. Soit $f \in L_{loc}^{ip}(X \times Y)$ et $x_0 \in X$. Nous supposons que:

(H₁) Il existe une sélection \bar{y} de M , bornée au voisinage de x_0 .

Alors, φ est lipschitzienne au voisinage de x_0 et \bar{Y}_0 désignant l'ensemble des valeurs d'adhérence de $\bar{y}(x)$, quand $x \rightarrow x_0$, nous avons:

$$(3.3) \quad x^* \in \partial\varphi(x_0) \Rightarrow (x^*, 0) \in \text{co} \bigcup_{\bar{y}_0 \in \bar{Y}_0} \partial_{(x, y)} f(x_0, \bar{y}_0).$$

Démonstration. Considérons $\varepsilon > 0$ et $V_\varepsilon = \{x \in X \mid \|x - x_0\| \leq \varepsilon\}$ tel que:

$$\sup \{\|\bar{y}(x)\| \mid x \in V_\varepsilon\} < +\infty,$$

où \bar{y} est une sélection de M .

$$\forall x_1, x_2 \in V_\varepsilon, \quad \varphi(x_1) = f(x_1, \bar{y}(x_1)), \quad \varphi(x_2) = f(x_2, \bar{y}(x_2)).$$

Supposons par exemple que: $\varphi(x_1) \leq \varphi(x_2)$;

$$\varphi(x_2) - \varphi(x_1) = f(x_2, \bar{y}(x_2)) - f(x_1, \bar{y}(x_1)) \leq f(x_2, \bar{y}(x_1)) - f(x_1, \bar{y}(x_1)).$$

f étant localement lipschitzienne, on en déduit que:

$$\varphi(x_2) - \varphi(x_1) \leq k \|x_1 - x_2\|.$$

Par conséquent, φ est lipschitzienne au voisinage de x_0 .

Considérons $x^* \in \partial\varphi(x_0)$; on a:

$$(3.4) \quad \forall d_1 \in X, \quad \langle x^*, d_1 \rangle \leq \limsup_{\substack{x \rightarrow x_0 \\ \lambda \rightarrow 0^+}} \frac{\varphi(x + \lambda d_1) - \varphi(x)}{\lambda}.$$

Si $\bar{y}(x) \in M(x)$, on a

$$(3.5) \quad \forall d_1, d_2 \in X, \quad \langle x^*, d_1 \rangle \leq \limsup_{\substack{x \rightarrow x_0 \\ \lambda \rightarrow 0^+}} \frac{f(x + \lambda d_1, \bar{y}(x) + \lambda d_2) - f(x, \bar{y}(x))}{\lambda}.$$

Au voisinage de x_0 , il existe, par hypothèse, une sélection \bar{y} de M qui est bornée. Ainsi, l'ensemble \bar{Y}_0 des valeurs d'adhérence de $\bar{y}(x)$, quand $x \rightarrow x_0$, n'est pas vide.

Il résulte alors de (3.5) que:

$$\forall d_1, d_2 \in X, \quad \langle x^*, d_1 \rangle \leq \sup_{\bar{y}_0 \in \bar{Y}_0} \left[\limsup_{\substack{x \rightarrow x_0 \\ y \rightarrow \bar{y}_0 \\ \lambda \rightarrow 0^+}} \frac{f(x + \lambda d_1, y + \lambda d_2) - f(x, y)}{\lambda} \right].$$

En conséquence:

$$x^* \in \partial\varphi(x_0) \Rightarrow (x^*, 0) \in \overline{co} \bigcup_{\bar{y}_0 \in \bar{Y}_0} \partial_{(x,y)} f(x_0, \bar{y}_0).$$

\bar{y} étant bornée au voisinage de x_0 , \bar{Y}_0 est compact.

De plus, la multiapplication $(u, v) \rightarrow \partial_{(x,y)} f(u, v)$ est s.c.s. et à valeurs compactes. En conséquence:

$$\bigcup_{\bar{y}_0 \in \bar{Y}_0} \partial_{(x,y)} f(x_0, \bar{y}_0) \text{ est compact,}$$

d'où le résultat (3.3).

Remarques. 1. L'hypothèse (H_1) a assuré que φ est lipschitzienne au voisinage de x_0 et que $LS(M; x_0)$ n'est pas vide. De plus, la multiapplication M est *fermée* en x_0 [12, Théorème 8] et par conséquent, pour toute sélection \bar{y} de M définie au voisinage de x_0 et bornée, on a:

$$\limsup_{x \rightarrow x_0} \{\bar{y}(x)\} = \bar{Y}_0 \subset LS(M; x_0) \subset M(x_0)$$

et

$$x^* \in \partial\varphi(x_0) \Rightarrow (x^*, 0) \in co \bigcup_{\bar{y}_0 \in M(x_0)} \partial_{(x,y)} f(x_0, \bar{y}_0).$$

2. La multiapplication M étant à valeurs fermées, l'hypothèse (H_1) est équivalente à:

(H'_1) L'application $x \rightarrow d(0, M(x))$ est bornée au voisinage de x_0 .

Cette hypothèse est évidemment vérifiée si M est bornée au voisinage de x_0 . Si M admet une sélection \bar{y} *continue* en x_0 , la formule (3.3) prend une forme simplifiée et si on note $\bar{y}_0 = \lim_{x \rightarrow x_0} \bar{y}(x)$, on aura:

$$(3.3') \quad \partial\varphi(x_0) \subset \{x^* \in X | (x^*, 0) \in \partial_{(x,y)} f(x_0, \bar{y}_0)\}.$$

Une telle situation apparaît, par exemple, lorsque M est à valeurs convexes et s.c.i. [15] ou bien lorsque $M(x) = \{m(x)\}$ au voisinage de x_0 avec m localement bornée.

3. La relation (3.3') est analogue à celle obtenue en analyse convexe, à savoir:

$$\partial\varphi(x_0) = \{x^* \in X | (x^*, 0) \in \partial_{(x,y)} f(x_0, y_0)\},$$

où y_0 est un élément quelconque de $M(x_0)$. Cette propriété d'indépendance de

l'élément $y_0 \in M(x_0)$ est due à la convexité de f et n'est pas vérifiée dans le cas localement lipschitzien.

4. Si $M(x_0) = \{\bar{y}_0\}$ et si f est continûment différentiable en (x_0, \bar{y}_0) , alors, d'après (3.3'), φ est différentiable en x_0 et:

$$\nabla \varphi(x_0) = \nabla_x f(x_0, \bar{y}_0).$$

3.2. Le Théorème 1 va nous permettre d'établir une relation du même type que (3.3) lorsque la fonction marginale φ_F est définie comme suit:

$$(3.6) \quad \varphi_F(x) = \inf \{f(x, y) | y \in F\},$$

où F est un fermé non vide de Y . M_F désignera la multiapplication qui à x associe l'ensemble des éléments de F pour lesquels $f(x, y) = \varphi_F(x)$.

THEOREME 2. Soit $f \in L_{\text{loc}}^{\text{ip}}(X \times Y)$ et $x_0 \in X$. Nous supposons que:

(H₂) Il existe une sélection \bar{y} de M_F , bornée au voisinage de x_0 .

\bar{Y}_0 désignant l'ensemble des valeurs d'adhérence de $\bar{y}(x)$, quand $x \rightarrow x_0$, nous avons:

$$(3.7) \quad \exists k > 0 \quad \text{tel que} \quad x^* \in \partial \varphi_F(x_0) \Rightarrow (x^*, 0) \in \text{co} \bigcup_{\bar{y}_0 \in \bar{Y}_0} \{\partial_{(x,y)} f(x_0, \bar{y}_0) + \{0\} \times k \partial d_F(\bar{y}_0)\}.$$

Démonstration. Soit \bar{y} une sélection de M_F , définie et bornée sur un voisinage de x_0 . L'ensemble \bar{Y}_0 des valeurs d'adhérence de $\bar{y}(x)$, quand $x \rightarrow x_0$, est un compact de Y ; on peut trouver \mathcal{O}_1 et \mathcal{O}_2 des ouverts contenant respectivement x_0 et \bar{Y}_0 tels que:

- (i) $\forall x \in \mathcal{O}_1, \bar{y}(x) \in \mathcal{O}_2$,
- (ii) f soit lipschitzienne sur $\mathcal{O}_1 \times \mathcal{O}_2$.

Définissons sur $\mathcal{O}_1 \times \mathcal{O}_2$ la fonction suivante:

$$g(x, y) = f(x, y) + k d_F(y)$$

avec k , constante de Lipschitz de f sur $\mathcal{O}_1 \times \mathcal{O}_2$.

Si l'on désigne par Π_A la multiapplication qui à u associe $\Pi_A(u)$: ensemble des points de A à distance minimum de A , nous avons de façon évidente:

$$\Pi_{X \times F}(x, y) = \{x\} \times \Pi_F(y).$$

La multiapplication $\Pi_{X \times F}$ est s.c.s. [3, Théorème 2, p. 122] et $\Pi_{X \times F}(\{x_0\} \times \bar{Y}_0) = \{x_0\} \times \bar{Y}_0$. En conséquence, on peut prendre \mathcal{O}'_2 et \mathcal{O}'_2 des ouverts contenant respectivement x_0 et \bar{Y}_0 , vérifiant les conditions (i) et (ii) avec, en outre:

$$\Pi_F(\mathcal{O}'_2) \subset \mathcal{O}_2.$$

Considérons $(x, y) \in \mathcal{O}'_1 \times \mathcal{O}'_2$ et $\tilde{y} \in \Pi_F(y)$; nous avons:

$$f(x, \tilde{y}) - f(x, y) \leq k \|y - \tilde{y}\| = k d_F(y),$$

c'est-à-dire:

$$(3.8) \quad f(x, \tilde{y}) \leq g(x, y),$$

ce qui montre que la fonction f a été localement pénalisée; le calcul de φ_F revient à celui de φ :

$$(3.9) \quad \varphi(x) = \inf \{g(x, y) | y \in \mathcal{O}'_2\}.$$

Pour $x \in \mathcal{O}'_1$, $\bar{y}(x)$ est solution du problème (3.9) et par conséquent:

$$x^* \in \partial\varphi_F(x_0) \Rightarrow (x^*, 0) \in \text{co} \bigcup_{\bar{y}_0 \in \bar{Y}_0} \partial_{(x,y)} g(x_0, \bar{y}_0).$$

En (x_0, \bar{y}_0) : $\partial_{(x,y)} g(x_0, \bar{y}_0) \subset \partial_{(x,y)} f(x_0, \bar{y}_0) + k \partial_{(x,y)} d_{X \times F}(x_0, \bar{y}_0)$. Remarquons que: $\partial_{(x,y)} d_{X \times F}(x_0, \bar{y}_0) = \{0\} \times \partial d_F(\bar{y}_0)$, on a le résultat (3.7).

Remarques. 1. On peut écrire (3.7) en faisant apparaître le cône normal à F en x_0 :

$$(3.10) \quad x^* \in \partial\varphi_F(x_0) \Rightarrow (x^*, 0) \in \text{co} \bigcup_{\bar{y}_0 \in \bar{Y}_0} \{\partial_{(x,y)} f(x_0, \bar{y}_0) + \{0\} \times N(F; \bar{y}_0)\},$$

c'est-à-dire, dans le cas où l'on peut se ramener à $\bar{Y}_0 = \{\bar{y}_0\}$:

$$(3.11) \quad x^* \in \partial\varphi_F(x_0) \Rightarrow \exists u^* \text{ tel que } (x^*, u^*) \in \partial_{(x,y)} f(x_0, \bar{y}_0) \text{ et } -u^* \in N(F; x_0).$$

2. Il résulte, soit de (3.10), soit d'un calcul direct, une estimation de $\partial\varphi_F(x_0)$ indépendante de la contrainte F et qui est la suivante:

$$(3.12) \quad \partial\varphi_F(x_0) \subset \text{co} \bigcup_{\bar{y}_0 \in \bar{Y}_0} \Pi_X \partial_{(x,y)} f(x_0, \bar{y}_0),$$

Π_X désignant l'opérateur de projection sur X , parallèlement à Y .

Dans le Théorème 2, on a donné une évaluation de $\partial\varphi_F(x_0)$ en fonction du gradient généralisé de f par rapport aux *deux variables*, ce qui suppose $f \in L_{\text{loc}}^p(X \times Y)$. De façon plus classique, lorsque la fonction marginale φ_F s'écrit sous la forme (3.6) (ou bien sous la forme d'un supremum au lieu d'un infimum dans (3.6)), on a l'habitude d'exprimer $\partial\varphi_F(x_0)$ en fonction de $\partial_x f(x_0, \bar{y}_0)$. F. H. Clarke [5] puis L. Thibault [19] donnent des conditions permettant d'exprimer exactement $\partial\varphi_F(x_0)$ lorsque la famille $\{f(\cdot, y)\}_{y \in F}$ considérée en tant que fonctions de x est une famille de fonctions *quasi-différentiables* au sens de B. N. Pschenichnyi [17]. G. Lebourg [14] donne une évaluation différente de $\partial\varphi_F(x_0)$ en considérant $\partial_x f(x, y)$ pour $x \in V(V \in \mathcal{V}(x_0))$ et $y \in M_\varepsilon(x) = \{y \in F \mid \varphi_F(x) \cong f(x, y) - \varepsilon\}$.

Afin de compléter l'étude du gradient généralisé de φ_F , nous allons démontrer une inclusion de $\partial\varphi_F(x_0)$ dans $\text{co} \bigcup_{\bar{y}_0 \in \bar{Y}_0} \partial_x f(x_0, \bar{y}_0)$ sous des hypothèses voisines de celles du Théorème 2.

THEOREME 3. Soit $f: X \times Y \rightarrow \mathbb{R}$ et $x_0 \in X$. Supposons que:

(H_{3,a}) $\forall y \in Y$, la fonction $x \rightarrow f(x, y)$ est localement lipschitzienne sur X .

(H_{3,b}) Il existe au voisinage de x_0 une sélection bornée \bar{y} de M_F dont l'ensemble des valeurs d'adhérence, quand $x \rightarrow x_0$, est noté \bar{Y}_0 .

(H_{3,c}) La multiapplication: $(u, v) \rightarrow \partial_x f(u, v)$ est bornée au voisinage de $\{x_0\} \times \bar{Y}_0$ et fermée en tout point de $\{x_0\} \times \bar{Y}_0$.

Alors φ_F est lipschitzienne au voisinage de x_0 et:

$$(3.13) \quad \partial\varphi_F(x_0) \subset \text{co} \bigcup_{\bar{y}_0 \in \bar{Y}_0} \partial_x f(x_0, \bar{y}_0).$$

Démonstration. Pour x appartenant à un voisinage borné V_0 de x_0 , considérons $\bar{y}(x) \in M_F(x)$; $\forall x_1, x_2 \in V_0$, $f(x_1, \bar{y}(x_1)) = \varphi_F(x_1)$, $f(x_2, \bar{y}(x_2)) = \varphi_F(x_2)$.

Supposons par exemple que $\varphi_F(x_1) \leq \varphi_F(x_2)$:

$$(3.14) \quad \varphi_F(x_2) - \varphi_F(x_1) \leq f(x_2, \bar{y}(x_1)) - f(x_1, \bar{y}(x_1)).$$

Grâce au théorème de la valeur moyenne pour gradient généralisé [14, Théorème 2.1]: $\exists \tilde{x} \in]x_1, x_2[$ et $\tilde{u} \in \partial_x f(\tilde{x}, \bar{y}(x_1))$ tels que:

$$(3.15) \quad f(x_2, \bar{y}(x_1)) - f(x_1, \bar{y}(x_1)) = \langle \tilde{u}, x_2 - x_1 \rangle.$$

D'après (H_{3,b}), \bar{Y}_0 est borné. Soit V_2 un voisinage borné de \bar{Y}_0 ; on peut trouver un voisinage V_1 de x_0 tel que:

$$\forall x \in V_1, \quad \bar{y}(x) \in V_2.$$

La multiapplication $(u, v) \rightarrow \partial_x f(u, v)$ étant bornée au voisinage de $\{x_0\} \times \bar{Y}_0$, on déduit de (3.14) et (3.15) que φ_F est lipschitzienne au voisinage de x_0 .

Pour $d \in X$, considérons une suite $\{x_n\}$ de X qui converge vers x_0 et une suite $\{\lambda_n\}$ de \mathbb{R}_+^* qui converge vers 0, telles que:

$$(3.16) \quad \varphi_F(x_0; d) = \lim_{n \rightarrow \infty} \frac{\varphi_F(x_n + \lambda_n d) - \varphi_F(x_n)}{\lambda_n}.$$

En posant $\bar{y}_n = \bar{y}(x_n)$, nous avons:

$$(3.17) \quad \frac{\varphi_F(x_n + \lambda_n d) - \varphi_F(x_n)}{\lambda_n} \leq \frac{f(x_n + \lambda_n d, \bar{y}_n) - f(x_n, \bar{y}_n)}{\lambda_n},$$

et nous supposons, sans perte de généralité, que $\lim_{n \rightarrow \infty} \bar{y}_n = \bar{y}_0$.

En utilisant le théorème de la valeur moyenne précédemment cité, nous avons: $\exists \tilde{x}_n \in]x_n, x_n + \lambda_n d[$ et $\tilde{u}_n \in \partial_x f(\tilde{x}_n, \bar{y}_n)$ tels que

$$f(x_n + \lambda_n d, \bar{y}_n) - f(x_n, \bar{y}_n) = \lambda_n \langle \tilde{u}_n, d \rangle.$$

Par hypothèse, la multiapplication qui à (u, v) associe $\partial_x f(u, v)$ est bornée au voisinage de $\{x_0\} \times \bar{Y}_0$ et fermée en tout point de $\{x_0\} \in \bar{Y}_0$. En conséquence, d'après (3.16) et (3.17):

$$\varphi_F(x_0; d) \leq \langle \tilde{u}_0, d \rangle \quad \text{avec} \quad \tilde{u}_0 \in \partial_x f(x_0, \bar{y}_0) \quad \text{et} \quad \bar{y}_0 \in \bar{Y}_0$$

D'où:

$$\forall d \in X, \quad \varphi_F(x_0; d) \leq \sup_{\bar{y}_0 \in \bar{Y}_0} \delta_{\partial_x f(x_0, \bar{y}_0)}^*(d)$$

et

$$\partial \varphi_F(x_0) \subset \overline{\text{co}} \bigcup_{\bar{y}_0 \in \bar{Y}_0} \partial_x f(x_0, \bar{y}_0).$$

Pour tout $\bar{y}_0 \in \bar{Y}_0$, la multiapplication $\partial_x f$ est bornée au voisinage de (x_0, \bar{y}_0) et fermée en (x_0, \bar{y}_0) ; elle est donc s.c.s. en (x_0, \bar{y}_0) [1, p. 52, Théorème 1.1]. Elle est, de plus, à valeurs compactes; par conséquent:

$$\bigcup_{\bar{y}_0 \in \bar{Y}_0} \partial_x f(x_0, \bar{y}_0) \quad \text{est un compact,}$$

d'où le résultat.

Remarques. Dans ce théorème, la condition de Lipschitz locale ne porte que sur l'application partielle $f_y: x \rightarrow f(x, y)$ alors que la condition (H_{3,c}) porte sur $\partial_x f$ en tant que multiapplication des deux variables.

Signalons que si $f \in L_{\text{loc}}^{\text{ip}}(X \times Y)$ et $(x_0, y_0) \in X \times Y$, nous avons:

$$(3.18) \quad \partial_x f(x_0, y_0) \subset \Pi_X \partial_{(x,y)} f(x_0, y_0),$$

l'inclusion inverse n'ayant pas lieu, en général. De même:

$$(3.19) \quad \partial_{(x,y)} f(x_0, y_0) \not\subset \partial_x f(x_0, y_0) \times \partial_y f(x_0, y_0) \not\subset \partial_{(x,y)} f(x_0, y_0).$$

Ces différentes considérations seront illustrées sur des exemples construits à partir d'une famille de fonctions localement lipschitziennes sur \mathbb{R}^2 .

Pour $\sigma \geq 0$, $\tau \geq 0$, considérons $f_{\sigma,\tau}$ définie sur $X \times Y = \mathbb{R}^2$ de la façon suivante:

$$f_{\sigma,\tau}(x, y) = \begin{cases} |y - x| + \sigma(1 - x) & \text{si } x \leq 1, \\ |2x - y - 1| + \tau \log x & \text{si } x > 1. \end{cases}$$

$f_{\sigma,\tau}$ est localement lipschitzienne sur \mathbb{R}^2 et soit $\varphi_{\sigma,\tau}$ la fonction marginale, c'est-à-dire:

$$\varphi_{\sigma,\tau}(x) = \inf \{f_{\sigma,\tau}(x, y) | y \in Y\}.$$

Nous avons:

$$\varphi_{\sigma,\tau}(x) = \begin{cases} \sigma(1 - x) & \text{si } x \leq 1, \\ \tau \log x & \text{si } x > 1 \end{cases}$$

et

$$M(x) = \{\bar{y}(x)\} \quad \text{avec} \quad \bar{y}(x) = \begin{cases} x & \text{si } x \leq 1, \\ 2x - 1 & \text{si } x > 1. \end{cases}$$

Au point $x_0 = 1$, $\partial \varphi_{\sigma,\tau}(1) = [-\sigma, \tau]$ et en $(x_0, \bar{y}_0) = (1, 1)$, nous avons:

$$\partial_{(x,y)} f_{\sigma,\tau}(1, 1) = \text{co} \{(-1 - \sigma, 1), (1 - \sigma, -1), (2 + \tau, -1), (-2 + \tau, 1)\}.$$

(E₁) Afin d'illustrer la formule (3.3), faisons $\sigma = \frac{1}{2}$ et $\tau = 0$. Nous obtenons:

$$\partial \varphi_{\sigma,\tau}(1) = [-\frac{1}{2}, 0],$$

$$\{x^* | (x^*, 0) \in \partial_{(x,y)} f_{\sigma,\tau}(1, 1)\} = [-\frac{3}{4}, \frac{1}{4}],$$

et cela montre que l'on ne peut espérer, en général, avoir l'égalité dans (3.3).

Le gradient généralisé partiel par rapport à x en $(1, 1)$ est donné par:

$$\partial_x f_{\sigma,\tau}(1, 1) = [-(1 + \sigma), \tau + 2],$$

soit, dans l'exemple traité:

$$\partial_x f_{\sigma,\tau}(1, 1) = [-\frac{3}{2}, 2].$$

D'autre part, pour le même exemple:

$$\partial_y f_{\sigma,\tau}(1, 1) = [-1, +1]$$

et

$$\Pi_X \partial_{(x,y)} f_{\sigma,\tau}(1, 1) = [-2, +2].$$

Cela illustre les relations générales (3.18) et (3.19).

(E₂) Soit F le fermé de $Y = \mathbb{R}$ déterminé par: $F = \{y \in \mathbb{R} | y \geq 1\}$ et considérons φ_F :

$$\varphi_F(x) = \inf \{f_{\sigma,\tau}(x, y) | y \in F\},$$

c'est-à-dire,

$$\varphi_F(x) = \begin{cases} (1 + \sigma)(1 - x) & \text{si } x \leq 1, \\ \tau \log x & \text{si } x > 1. \end{cases}$$

Faisons $\sigma = \tau = \frac{1}{4}$; en appliquant la formule (3.7) du Théorème 2 (ou simplement (3.11)), nous avons:

$$\partial\varphi_F(1) \subset \{x^* | (x^*, 0) \in \partial_{(x,y)} f_{\sigma,\tau}(1, 1) + \{0\} \times N(F; 1)\} = [-\frac{7}{4}, \frac{1}{2}].$$

Si on applique la formule (3.13) de Théorème 3, nous avons:

$$\partial\varphi_F(1) \subset \partial_x f_{\sigma,\tau}(1, 1) = [-\frac{5}{4}, \frac{9}{4}]$$

ce qui donne deux estimations différentes de $\partial\varphi_F(1) = [-\frac{5}{4}, \frac{1}{4}]$.

3.3. Soit F une multiapplication définie sur X (notée $F: X \rightrightarrows Y$) et f une fonction localement lipschitzienne sur X ; considérons

$$(3.20) \quad \varphi_F(x) = \inf \{f(x, y) | y \in F(x)\}$$

et

$$(3.21) \quad M_F(x) = \{y \in F(x) | \varphi_F(x) = f(x, y)\}.$$

Une condition suffisante pour que φ_F soit continue en $x_0 \in X$ est que: $F(x_0)$ soit compact et F soit s.c.i., s.c.s. en x_0 [1, Théorème 1.3, p. 55]. D'une autre manière, s'il existe une sélection \bar{y} de M_F continue en x_0 , alors φ_F est continue en x_0 . Nous allons donner une condition *suffisante* pour que φ_F soit localement lipschitzienne.

LEMME 4. Soit $f \in L_{\text{loc}}^{\text{ip}}(X \times Y)$, $F: X \rightrightarrows Y$ et $x_0 \in X$. Sous l'hypothèse suivante:

(H₄) Il existe une sélection \bar{y} de M_F , lipschitzienne au voisinage de x_0 , la fonction φ_F définie par (3.20) est lipschitzienne au voisinage de x_0 .

Démonstration. Pour x_1 et x_2 dans un voisinage borné V_0 de x_0 , soient $\bar{y}_1 = \bar{y}(x_1)$ et $\bar{y}_2 = \bar{y}(x_2)$. Nous avons:

$$(3.22) \quad \begin{aligned} \varphi_F(x_1) &= f(x_1, \bar{y}_1) \quad \text{et} \quad \varphi_F(x_2) = f(x_2, \bar{y}_2), \\ \varphi_F(x_2) - \varphi_F(x_1) &= f(x_2, \bar{y}_2) - f(x_1, \bar{y}_2) + f(x_1, \bar{y}_2) - f(x_1, \bar{y}_1). \end{aligned}$$

La sélection \bar{y} est bornée au voisinage de x_0 ; f étant localement lipschitzienne:

$$f(x_2, \bar{y}_2) - f(x_1, \bar{y}_2) \leq k_0 \|x_1 - x_2\|;$$

V_0 étant choisi de sorte que \bar{y} soit lipschitzienne sur V_0 , nous avons de même:

$$f(x_1, \bar{y}_2) - f(x_1, \bar{y}_1) \leq k_0 \|\bar{y}_2 - \bar{y}_1\| \leq k_1 \|x_2 - x_1\|,$$

d'où le résultat d'après (3.22).

Remarque. L'hypothèse (H₄) est relative à f et F ; elle est équivalente dans le cas où M_F est à valeurs fermées, à la suivante:

(H'₄) $\delta(M_F(x_1), M_F(x_2)) \leq k \|x_1 - x_2\|$ au voisinage de x_0 , $\delta(A, B)$ désignant la distance entre A et B .

F étant une multiapplication de X dans Y , supposée fermée, on notera \mathcal{F} le graphe (fermé) de F dans $X \times Y$. Pour $(x_0, y_0) \in \mathcal{F}$, $N_{\mathcal{F}}(x_0, y_0)$ et $\mathcal{T}_{\mathcal{F}}(x_0, y_0)$ désigneront respectivement le cône normal à \mathcal{F} en (x_0, y_0) et le cône tangent à \mathcal{F} en (x_0, y_0) .

Dans le cas particulier où $F(x) = F$, c'est-à-dire, $\mathcal{F} = X \times F$, une évaluation de $\partial\varphi_F(x_0)$ est donnée par la formule (3.10) et l'on constate que:

$$\{0\} \times N(F; y_0) = N_{X \times F}(x_0, y_0).$$

Cela nous conduit à démontrer, dans le cas où F dépend de x , une formule analogue à (3.10) avec $N_{\mathcal{F}}(x_0, y_0)$.

La notion polaire de la notion de cône normal à un fermé est celle de *cône tangent* à ce fermé. Nous allons en rappeler la définition et en donner différentes caractérisations utiles dans les démonstrations.

DÉFINITION 5 [5, Proposition 3.6]. Soit Q un fermé non vide de Z et $z_0 \in Q$; on appelle cône tangent à Q en z_0 et on note $\mathcal{T}(Q; z_0)$ (ou $\mathcal{T}_Q(z_0)$) le cône polaire de $N(Q; z_0)$.

LEMME 6.

$$(3.23a) \quad \forall d \in X, \quad d \dot{\in} (Q; z_0; d) = \limsup_{\substack{z \rightarrow z_0 \\ z \in Q \\ \lambda \rightarrow 0^+}} \frac{d_Q(z + \lambda d)}{\lambda},$$

$$(3.23b) \quad d \in \mathcal{T}(Q; z_0) \Leftrightarrow \lim_{\substack{z \rightarrow z_0 \\ z \in Q \\ \lambda \rightarrow 0^+}} \frac{d_Q(z + \lambda d)}{\lambda} = 0.$$

Démonstration. Par définition,

$$d \dot{\in} (Q; z_0; d) = \limsup_{\substack{z \rightarrow z_0 \\ z \in Q \\ \lambda \rightarrow 0^+}} \frac{d_Q(z + \lambda d) - d_Q(z)}{\lambda}.$$

En considérant $z \rightarrow z_0$, z dans Q , on a immédiatement:

$$d \dot{\in} (Q; z_0; d) \geq \limsup_{\substack{z \rightarrow z_0 \\ z \in Q \\ \lambda \rightarrow 0^+}} \frac{d_Q(z + \lambda d)}{\lambda}.$$

Pour l'inégalité inverse, considérons $z \rightarrow z_0$ et $\lambda \rightarrow 0^+$. En désignant par \bar{z} une projection de z sur Q , nous avons:

$$|d_Q(z + \lambda d) - d_Q(\bar{z} + \lambda d)| \leq \|z - \bar{z}\| = d_Q(z),$$

soit

$$d_Q(z + \lambda d) - d_Q(z) \leq d_Q(\bar{z} + \lambda d),$$

d'où le résultat (3.23a).

Pour la caractérisation (3.23b), il suffit de remarquer que:

$$d \in \mathcal{T}(Q; z_0) \Leftrightarrow d \dot{\in} (Q; z_0; d) = 0$$

d'où le résultat (3.23a).

Outre la définition et la caractérisation (3.23b) données précédemment, on peut définir $\mathcal{T}(Q; z_0)$ à l'aide d'une caractérisation analogue à celles utilisées pour définir les différentes notions infinitésimales habituellement rattachées à Q et à z_0 .

LEMME 6'. $d \in \mathcal{T}(Q; z_0)$ si et seulement si la propriété suivante est vérifiée:

$$(3.24) \quad \left\{ \begin{array}{l} \forall \{z_n\}, \quad z_n \in Q, \quad \lim_{n \rightarrow \infty} z_n = z_0; \quad \forall \{\lambda_n\}, \quad \lambda_n > 0, \quad \lim_{n \rightarrow \infty} \lambda_n = 0, \\ \exists \{d_n\}, \quad \lim_{n \rightarrow \infty} d_n = d, \quad \exists \sigma: \mathbb{N} \rightarrow \mathbb{N} \text{ strictement croissante} \\ \text{tels que: } \forall n, \quad z_{\sigma(n)} + \lambda_{\sigma(n)} d_{\sigma(n)} \in Q. \end{array} \right.$$

Démonstration. Pour tout $d \in X$, nous avons, d'après (3.23a),

$$d_Q^\circ(z_0; d) = \limsup_{\substack{z \rightarrow z_0 \\ z \in Q \\ \lambda \rightarrow 0^+}} \frac{d_Q(z + \lambda d)}{\lambda}.$$

Considérons $\{z_n\}$ une suite d'éléments de Q qui converge vers z_0 et $\{\lambda_n\}$ une suite de scalaires positifs de limite 0 telles que:

$$(3.25) \quad d_Q^\circ(z_0; d) = \lim_{n \rightarrow \infty} \frac{d_Q(z_n + \lambda_n d)}{\lambda_n}.$$

Supposons que d vérifie la propriété (3.24); nous avons:

$$\lim_{n \rightarrow \infty} d_n = d \quad \text{et} \quad d_Q(z_{\sigma(n)} + \lambda_{\sigma(n)} d_{\sigma(n)}) = 0.$$

Ainsi:

$$\frac{d_Q(z_{\sigma(n)} + \lambda_{\sigma(n)} d)}{\lambda_{\sigma(n)}} \leq \|d_{\sigma(n)} - d\|,$$

ce qui implique, d'après (3.25), que: $d_Q^\circ(z_0; d) = 0$, c'est-à-dire, $d \in \mathcal{T}(Q; z_0)$.

Réciproquement, soit $d \in \mathcal{T}(Q; z_0)$. Considérons $\{z_n\}$ une suite d'éléments de Q qui converge vers z_0 et $\{\lambda_n\}$ une suite positive qui converge vers 0. D'après (3.23b):

$$(3.26) \quad \lim_{n \rightarrow \infty} \frac{d_Q(z_n + \lambda_n d)}{\lambda_n} = 0.$$

\bar{z}_n désignant une projection de $z_n + \lambda_n d$ sur Q , posons:

$$d_n = \frac{\bar{z}_n - z_n}{\lambda_n}.$$

Comme $d_n = (\bar{z}_n - (z_n + \lambda_n d))/\lambda_n + d$, $d_Q(z_n + \lambda_n d)/\lambda_n = \|d_n - d\|$, on a, d'après (3.26):

$$\lim_{n \rightarrow \infty} d_n = d.$$

et $z_n + \lambda_n d_n = \bar{z}_n \in Q$ pour tout n . La propriété (3.24) est donc vérifiée pour tout $d \in \mathcal{T}(Q; z_0)$.

Dans le but d'évaluer $\partial \varphi_F(x_0)$, nous allons dans un premier temps donner une majoration de la dérivée directionnelle généralisée de φ_F dans la direction d_1 .

THEOREME 7. Soit $f \in L_{\text{loc}}^{\text{ip}}(X \times Y)$, F une multiapplication fermée de X dans Y et $x_0 \in X$. On suppose que:

(H_{7.a}) φ_F est lipschitzienne au voisinage de x_0 ,

(H_{7.b}) il existe une sélection \bar{y} de M_F , bornée au voisinage de x_0 .

En désignant par \bar{Y}_0 l'ensemble des valeurs d'adhérence de $\bar{y}(x)$ quand $x \rightarrow x_0$, on pose:

$$\mathcal{U}(d_1) = \{d_2 \in Y \mid (d_1, d_2) \in \bigcap_{\bar{y}_0 \in \bar{Y}_0} \mathcal{T}_{\mathcal{F}}(x_0, \bar{y}_0)\}.$$

Alors:

$$(3.27) \quad \varphi_F^\circ(x_0; d_1) \leq \inf_{d_2 \in \mathcal{U}(d_1)} \sup_{\bar{y}_0 \in \bar{Y}_0} f_{(x,y)}^\circ(x_0, \bar{y}_0; d_1, d_2).$$

Démonstration. Si $\mathcal{U}(d_1) = \emptyset$, (3.27) est trivialement vérifiée avec la convention que l'infimum sur l'ensemble vide est $+\infty$. Supposons donc que $\mathcal{U}(d_1) \neq \emptyset$.

$d_1 \in X$ étant fixé, considérons une suite $\{x_n\}$ d'éléments de X et une suite $\{\lambda_n\}$ de scalaires positifs telles que:

$$(3.28) \quad \lim_{n \rightarrow \infty} x_n = x_0, \quad \lim_{n \rightarrow \infty} \lambda_n = 0,$$

$$\varphi_F(x_0; d_1) = \lim_{n \rightarrow \infty} \frac{\varphi_F(x_n + \lambda_n d_1) - \varphi_F(x_n)}{\lambda_n}.$$

On pose $\bar{y}_n = \bar{y}(x_n)$ où \bar{y} désigne une sélection de M_F , bornée au voisinage de x_0 . Considérons $\bar{y}_0 \in \bar{Y}_0$ une valeur d'adhérence de la suite $\{\bar{y}_n\}$; nous noterons comme la suite originelle la sous-suite de $\{\bar{y}_n\}$ qui converge vers \bar{y}_0 . D'après (3.28), on a:

$$(3.29) \quad \varphi_F(x_0; d_1) = \lim_{n \rightarrow \infty} \frac{\varphi_F(x_n + \lambda_n d_1) - f(x_n, \bar{y}_n)}{\lambda_n} \quad \text{et} \quad \lim_{n \rightarrow \infty} \bar{y}_n = \bar{y}_0.$$

Considérons $d_2 \in \mathcal{U}(d_1)$; pour tout $\bar{y}_0 \in \bar{Y}_0$, nous avons $(d_1, d_2) \in \mathcal{T}_{\mathcal{F}}(x_0, \bar{y}_0)$ et d'après la caractérisation (3.24), on en conclut l'existence d'une suite d'éléments $(d_1^{(n)}, d_2^{(n)})$ de $X \times Y$ telle que:

$$\lim_{n \rightarrow \infty} (d_1^{(n)}, d_2^{(n)}) = (d_1, d_2)$$

et

$$(x_n, \bar{y}_n) + \lambda_n (d_1^{(n)}, d_2^{(n)}) \in \mathcal{F} \quad \text{une infinité de fois.}$$

Désignons par $\{n_k\}$ la sous-suite telle que:

$$\forall k, \quad (x_{n_k} + \lambda_{n_k} d_1^{(n_k)}, \bar{y}_{n_k} + \lambda_{n_k} d_2^{(n_k)}) \in \mathcal{F},$$

autrement dit:

$$(3.30) \quad \forall k, \quad \bar{y}_{n_k} + \lambda_{n_k} d_2^{(n_k)} \in F(x_{n_k} + \lambda_{n_k} d_1^{(n_k)}).$$

Puisque $\lim_{k \rightarrow \infty} d_1^{(n_k)} = d_1$ et que φ_F est lipschitzienne au voisinage de x_0 ,

$$(3.31) \quad \forall k \geq k_0, \quad \varphi_F(x_{n_k} + \lambda_{n_k} d_1) \leq \varphi_F(x_{n_k} + \lambda_{n_k} d_1^{(n_k)}) + \lambda_{n_k} \varepsilon_k^{(1)}$$

avec $\lim_{k \rightarrow \infty} \varepsilon_k^{(1)} = 0^+$. D'après (3.30):

$$(3.32) \quad \forall k, \quad \varphi_F(x_{n_k} + \lambda_{n_k} d_1^{(n_k)}) \leq f(x_{n_k} + \lambda_{n_k} d_1^{(n_k)}, \bar{y}_{n_k} + \lambda_{n_k} d_2^{(n_k)}).$$

f étant localement lipschitzienne au voisinage de (x_0, \bar{y}_0) , on a:

$$(3.33) \quad \forall k \geq k_1, \quad f(x_{n_k} + \lambda_{n_k} d_1^{(n_k)}, \bar{y}_{n_k} + \lambda_{n_k} d_2^{(n_k)}) \leq f(x_{n_k} + \lambda_{n_k} d_1, \bar{y}_{n_k} + \lambda_{n_k} d_2) + \lambda_{n_k} \varepsilon_k^{(2)}$$

avec $\lim_{k \rightarrow \infty} \varepsilon_k^{(2)} = 0^+$.

Il résulte alors de (3.31), (3.32), (3.33), que:

$$(3.34) \quad \forall k \geq \bar{k},$$

$$\frac{\varphi_F(x_{n_k} + \lambda_{n_k} d_1) - f(x_{n_k}, \bar{y}_{n_k})}{\lambda_{n_k}} \leq \frac{f(x_{n_k} + \lambda_{n_k} d_1, \bar{y}_{n_k} + \lambda_{n_k} d_2) - f(x_{n_k}, \bar{y}_{n_k})}{\lambda_{n_k}} + \bar{\varepsilon}_k$$

avec $\lim_{k \rightarrow \infty} \bar{\varepsilon}_k = 0^+$.

En conséquence, d'après (3.29):

$$\varphi_F(x_0; d_1) \leq f_{(x,y)}^*(x_0, \bar{y}_0; d_1, d_2).$$

d_2 étant choisi tel que $(d_1, d_2) \in \mathcal{T}_{\mathcal{F}}(x_0, \bar{y}_0)$ pour l'élément $\bar{y}_0 \in \bar{Y}_0$ considéré, on en conclut que:

$$\varphi_F^*(x_0; d_1) \leq \inf_{d_2 \in \mathcal{U}(d_1)} \sup_{\bar{y}_0 \in \bar{Y}_0} f_{(x,y)}^*(x_0, \bar{y}_0; d_1, d_2).$$

Dans le cas où l'on peut se ramener à $\bar{Y}_0 = \{\bar{y}_0\}$ (sélection \bar{y} de M_F , continue en x_0), la formule (3.27) se simplifie et conduit à l'estimation suivante de $\partial\varphi_F(x_0)$:

THEOREME 8. Soit $f \in L_{\text{loc}}^1(X \times Y)$, $F: X \rightrightarrows Y$ fermée et $x_0 \in X$. On suppose que φ_F est lipschitzienne au voisinage de x_0 et qu'il existe une sélection \bar{y} de M_F , continue en x_0 . Alors:

$$(3.35) \quad x^* \in \partial\varphi_F(x_0) \Rightarrow (x^*, 0) \in \partial_{(x,y)} f(x_0, \bar{y}_0) + N_{\mathcal{F}}(x_0, \bar{y}_0).$$

Démonstration. Nous avons:

$$\mathcal{U}(d_1) = \{d_2 \in Y \mid (d_1, d_2) \in \mathcal{T}_{\mathcal{F}}(x_0, \bar{y}_0)\}$$

et

$$\varphi_F^*(x_0; d_1) \leq \inf_{d_2 \in \mathcal{U}(d_1)} f_{(x,y)}^*(x_0, \bar{y}_0; d_1, d_2).$$

Globalement:

$$\varphi_F^*(x_0; d_1) \leq f_{(x,y)}^*(x_0, \bar{y}_0; d_1, d_2) + \delta_{\mathcal{T}_{\mathcal{F}}(x_0, \bar{y}_0)}(d_1, d_2).$$

Or:

$$\delta_{\mathcal{T}_{\mathcal{F}}(x_0, \bar{y}_0)} = \delta_{N_{\mathcal{F}}(x_0, \bar{y}_0)}^*.$$

En conséquence:

$$\forall d_1 \in X, \quad \varphi_F^*(x_0; d_1) \leq \delta_{\Phi(x_0, \bar{y}_0)}^*(d_1, d_2)$$

avec $\Phi(x_0, \bar{y}_0) = \partial_{(x,y)} f(x_0, \bar{y}_0) + N_{\mathcal{F}}(x_0, \bar{y}_0)$, d'où l'inclusion (3.35).

Comme pour les formules précédemment établies, nous allons illustrer la formule (3.35) à l'aide d'un exemple; considérons la fonction $f_{\sigma, \tau}$ (introduite plus haut) pour $\sigma = \tau = \frac{1}{2}$, c'est à dire:

$$f_{\sigma, \tau}(x, y) = \begin{cases} |y - x| + \frac{1}{2}(1 - x) & \text{si } x \leq 1, \\ |2x - y - 1| + \frac{1}{2} \log x & \text{si } x > 1. \end{cases}$$

Comme multiapplication F , nous choisissons celle définie par:

$$F(x) = \begin{cases} \{y \mid 2e^{x-1} - 3 + y \leq 0\} & \text{si } x \leq 1, \\ \{y \mid -3x - y - 4 \leq 0\} & \text{si } x > 1. \end{cases}$$

La fonction $\varphi_{\sigma, \tau}$ définie par: $\varphi_{\sigma, \tau}(x) = \inf \{f_{\sigma, \tau}(x, y) \mid y \in F(x)\}$ est aisément déterminable et a pour expression:

$$\varphi_{\sigma, \tau}(x) = \begin{cases} -2e^{x-1} - \frac{3}{2}x + \frac{7}{2} & \text{si } x \leq 1, \\ \frac{1}{2} \log x & \text{si } x > 1. \end{cases}$$

Par ailleurs, \mathcal{F} étant le graphe de la multiapplication F , au point $(1, 1)$ nous avons:

$$N_{\mathcal{F}}(1, 1) = \left\{ (x, y) \mid x \leq 0, \frac{x}{3} \leq y \leq \frac{x}{2} \right\}.$$

En $x_0 = 1$, $\bar{y}_0 = 1$, on a:

$$\partial_{(x,y)} f_{\sigma, \tau}(1, 1) = \text{co} \left\{ \left(-\frac{3}{2}, 1\right), \left(\frac{1}{2}, -1\right), \left(\frac{5}{2}, -1\right) \right\}$$

ce qui fait que:

$$\{x^*|(x^*, 0) \in \partial_{(x,y)} f_{\sigma,\tau}(1, 1) + N_{\mathcal{F}}(1, 1)\} = [-\frac{9}{2}, \frac{1}{2}].$$

Comme $\partial\varphi_{\sigma,\tau}(1) = [-\frac{7}{2}, \frac{1}{2}]$, on a un exemple d'inclusion stricte dans la formule (3.35).

Lorsque la multiapplication F a la structure suivante:

$$F(x) = \{y \in Y | g_i(x, y) \leq 0, \quad \forall i \in \langle 1, m \rangle\},$$

c'est-à-dire:

$$\mathcal{F} = \{(x, y) \in X \times Y | g_i(x, y) \leq 0, \quad \forall i \in \langle 1, m \rangle\};$$

l'explicitation de $\partial\varphi_F(x_0)$ dans (3.35) à l'aide des gradients généralisés $\partial g_i(x_0, y_0)$ passe par la comparaison de $N_{\mathcal{F}}(x_0, y_0)$ et de $\overline{\text{cc}} \partial g_i(x_0, y_0)$ et cela est possible sous certaines conditions sur les fonctions g_i [11]. Dans ce cas, on peut donner des conditions d'optimalité pour la problème posé dans X sous la forme:

$$\text{Min} \{\varphi_F(x) | x \in Q_1\}, \quad Q_1 \subset X.$$

Il suffit d'appliquer les théorèmes de [11] ou [6] et la formule (3.35). La formule (3.35) a également d'autres applications intéressantes; nous l'exploiterons sur deux exemples.

Si l'on fait la remarque (triviale) que pour $F(x) = \{y(x)\}$ au voisinage de x_0 (i.e. un singleton), $\varphi_F(x)$ n'est autre que $f(x, y(x))$ et la formule (3.35) peut répondre à la question: comment évaluer $\partial\varphi_F(x) = \partial_x f(x, y(x))$?

Considérons $y: X \rightarrow Y$; on notera \mathcal{Y} le graphe de y dans $X \times Y$ et $\mathcal{J}_y(x_0)$ la matrice jacobienne de y en x_0 (lorsqu'elle existe). Nous avons alors la propriété suivante:

LEMME 9. Si $y: X \rightarrow Y$ est différentiable en $x_0 \in X$, de dérivée forte en x_0 , alors:

$$(3.36) \quad N_{\mathcal{Y}}(x_0, y(x_0)) \subset \{(u, v) | u + [\mathcal{J}_y(x_0)]^* v = 0\}.$$

Démonstration. Considérons une suite d'éléments (x_n, y_n) de $X \times Y$ et désignons par $(\bar{x}_n, y(\bar{x}_n))$ une projection de (x_n, y_n) sur le graphe \mathcal{Y} . On supposera que:

$$(3.37) \quad \exists \{\lambda_n\} \lambda_n > 0 \text{ telle que } \lambda_n(x_n - \bar{x}_n, y_n - y(\bar{x}_n)) \rightarrow (\bar{u}, \bar{v}),$$

$$x_n \rightarrow x_0, \quad y_n \rightarrow y(x_0).$$

Les éléments de $N_{\mathcal{Y}}(x_0, y(x_0))$ appartiennent précisément à l'enveloppe convexe fermée de l'ensemble des éléments vérifiant (3.37).

Soit $u \in X$; $(u, y(u)) \in \mathcal{Y}$ et puisque $(\bar{x}_n, y(\bar{x}_n))$ est un élément de \mathcal{Y} à distance minimum de (x_n, y_n) , nous avons:

$$\forall u \in X, \quad \|(x_n, y_n) - (\bar{x}_n, y(\bar{x}_n))\|^2 \leq \|(x_n, y_n) - (u, y(u))\|^2.$$

En écrivant $(x_n - u, y_n - y(u)) = (x_n - \bar{x}_n, y_n - y(\bar{x}_n)) + (\bar{x}_n - u, y(\bar{x}_n) - y(u))$, et en développant l'inégalité précédente, on a:

$$(3.38) \quad 2[\langle x_n - \bar{x}_n, u - \bar{x}_n \rangle + \langle y_n - y(\bar{x}_n), y(u) - y(\bar{x}_n) \rangle]$$

$$\leq \|u - \bar{x}_n\|^2 + \|y(u) - y(\bar{x}_n)\|^2 \quad \forall u \in X.$$

$\{\lambda_n\}$ étant la suite définie en (3.37) et d un élément quelconque de X , posons:

$$u_n = \bar{x}_n + \frac{1}{\lambda_n^2} \cdot d.$$

D'après (3.38), nous avons pour $u = u_n$:

$$(3.39) \quad \begin{aligned} & 2 \left[\frac{1}{\lambda_n^2} \langle x_n - \bar{x}_n, d \rangle + \langle y_n - y(\bar{x}_n), y \left(\bar{x}_n + \frac{1}{\lambda_n^2} d \right) - y(\bar{x}_n) \rangle \right] \\ & \leq \frac{1}{\lambda_n^4} \|d\|^2 + \left\| y \left(\bar{x}_n + \frac{1}{\lambda_n^2} d \right) - y(\bar{x}_n) \right\|^2. \end{aligned}$$

Par hypothèse, y est différentiable en x_0 , de dérivée (représentée par $\mathcal{J}_y(x_0)$) forte en x_0 . Cela implique [16, p. 71]:

$$\begin{aligned} \forall \varepsilon > 0, \quad \exists \delta(\varepsilon) \quad \forall x_i, \quad i = 1, 2, \quad \|x_i - x_0\| \leq \delta(\varepsilon), \\ \|y(x_1) - y(x_2) - \mathcal{J}_y(x_0)(x_1 - x_2)\| \leq \varepsilon \|x_1 - x_2\|. \end{aligned}$$

Comme $\bar{x}_n \rightarrow x_0$ et que $\lambda_n \rightarrow +\infty$, $u_n \rightarrow x_0$. Ainsi:

$$\begin{aligned} \forall \varepsilon > 0, \quad \exists n_0(\varepsilon), \\ \forall n \geq n_0(\varepsilon), \quad y \left(\bar{x}_n + \frac{1}{\lambda_n^2} d \right) - y(\bar{x}_n) = \frac{1}{\lambda_n^2} \mathcal{J}_y(x_0) \cdot d + \frac{\theta_n}{\lambda_n^2} \end{aligned}$$

avec $\|\theta_n\| \leq \varepsilon \|d\|$. En reportant dans (3.39):

$$\begin{aligned} & 2 \left[\frac{1}{\lambda_n^2} \langle x_n - \bar{x}_n, d \rangle + \frac{1}{\lambda_n^2} \langle y_n - y(\bar{x}_n), \mathcal{J}_y(x_0) \cdot d \rangle + \frac{1}{\lambda_n^2} \langle \theta_n, y_n - y(\bar{x}_n) \rangle \right] \\ & \leq \frac{1}{\lambda_n^4} \|d\|^2 + \frac{2}{\lambda_n^4} \|\mathcal{J}_y(x_0)d\|^2 + \frac{2\|\theta_n\|^2}{\lambda_n^4}, \end{aligned}$$

ou encore:

$$(3.40) \quad \langle \lambda_n(x_n - \bar{x}_n), d \rangle + \langle \lambda_n[\mathcal{J}_y(x_0)]^*(y_n - y(\bar{x}_n)), d \rangle \leq -\langle \lambda_n(y_n - y(\bar{x}_n)), \theta_n \rangle + \frac{K_1}{\lambda_n}.$$

En passant à la limite

$$\langle \bar{u}, d \rangle + \langle [\mathcal{J}_y(x_0)]^* \bar{v}, d \rangle \leq \varepsilon K_2,$$

d'où, finalement:

$$\forall d \in X, \quad \langle \bar{u}, d \rangle + \langle [\mathcal{J}_y(x_0)]^* \bar{v}, d \rangle \leq 0,$$

c'est-à-dire:

$$\bar{u} + [\mathcal{J}_y(x_0)]^* \bar{v} = 0.$$

PROPOSITION 10. Soit $f \in L_{\text{loc}}^{\text{ip}}(X \times Y)$ et $y: X \rightarrow Y$ différentiable en $x_0 \in X$, de dérivée forte en x_0 . Alors, la fonction $x \rightarrow \varphi(x) = f(x, y(x))$ est lipschitzienne au voisinage de x_0 et:

$$(3.41) \quad x^* \in \partial \varphi(x_0) \Rightarrow \begin{cases} \exists (x_1^*, x_2^*) \in \partial_{(x,y)} f(x_0, y(x_0)), \\ x^* = x_1^* + [\mathcal{J}_y(x_0)]^* x_2^*. \end{cases}$$

Démonstration. y ayant une dérivée forte en x_0 est lipschitzienne au voisinage de x_0 et d'après le lemme 4, φ est lipschitzienne au voisinage de x_0 .

D'après (3.35) et le résultat du lemme précédent:

$$x^* \in \partial \varphi(x_0) \Rightarrow (x^*, 0) \in \partial_{(x,y)} f(x_0, y(x_0)) + \{(u, v) | u + [\mathcal{J}_y(x_0)]^* v = 0\},$$

d'où le résultat annoncé.

Remarque. Dans le cas particulier où f ne dépend que de la variable y , nous avons:

$$(3.42) \quad \Psi(x) = f(y(x)).$$

Le problème des moindres carrés est un cas spécial du problème général de la minimisation de fonctions du type (3.42). En remarquant que:

$$\partial_{(x,y)} f(u_0, v_0) = \{0\} \times \partial_y f(v_0),$$

nous avons, sous les hypothèses de la Proposition 10:

$$(3.43) \quad \partial \Psi(x_0) \subset [\mathcal{J}_y(x_0)]^* \partial f(y(x_0)).$$

BIBLIOGRAPHIE

- [1] A. AUSLENDER, *Optimisation: Méthodes numériques*, Masson et Cie, Paris, 1976.
- [2] V. V. BERESNEV AND B. N. PSCHENICHNYI, *The differential properties of minimum functions*, U.S.S.R. Comput. Math. and Math. Phys., 14 (1973), no. 3, pp. 101–113.
- [3] C. BERGE, *Espaces topologiques, fonctions multivoques*, Dunod, Paris, 1966.
- [4] F. H. CLARKE, *Necessary conditions for nonsmooth problems in optimal control and the calculus of variations*, Ph.D. thesis, Dept. of Mathematics, University of Washington, Seattle, 1973.
- [5] ———, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.
- [6] ———, *A new approach to Lagrange multipliers*, Math. Operations Res., 2 (1976), pp. 165–174.
- [7] V. F. DEM'YANOV, *The minimax problem with dependant constraints*, U.S.S.R. Comput. Math. and Math. Phys., 12 (1972), no. 3, pp. 299–308.
- [8] V. F. DEM'YANOV AND V. N. MALOZEMOV, *Introduction to Minimax*, John Wiley, New York, 1974.
- [9] V. F. DEM'YANOV AND A. B. PEVNYI, *Expansion with respect to a parameter of the extremal values of games problems*, U.S.S.R. Comput. Math. and Math. Phys., 14 (1974), no. 5, pp. 33–45.
- [10] J. B. HIRIART-URRUTY, *Conditions nécessaires d'optimalité en programmation non différentiable*, C.R. Acad. Sci. Paris, Série A, 283 (1976), pp. 843–845.
- [11] ———, *Conditions nécessaires d'optimalité en programmation non différentiable*, Séminaire d'analyse numérique, Département de Mathématiques Appliquées, Université de Clermont, Aubière, (1975–1976).
- [12] W. W. HOGAN, *Point-to-set maps in mathematical programming*, SIAM Rev., 15 (1973), pp. 591–603.
- [13] ———, *Directional derivatives for extremal-value functions with applications to the completely convex case*, Operations Research, 21 (1973), pp. 188–209.
- [14] G. LEBOURG, *Valeur moyenne pour gradient généralisé*, C.R. Acad. Sci. Paris, Série A, (1975), pp. 795–797, Tome 281.
- [15] E. MICHAEL, *A survey of continuous selections*, Set-valued Mappings, Selections and Topological Properties of 2^X , Lecture Notes in Math. 171, Springer-Verlag, New York, 1970, pp. 54–59.
- [16] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
- [17] B. N. PSCHENICHNYI, *Necessary conditions for an Extremum*, Marcel Dekker, New York, 1971.
- [18] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1971.
- [19] L. THIBAUT, *Quelques propriétés des sous-différentiels de fonctions réelles localement lipschitziennes définies sur un espace de Banach séparable*, C.R. Acad. Sci. Paris, (1976), pp. 507–510.

CONDITIONS NECESSAIRES D'OPTIMALITE POUR UN PROGRAMME STOCHASTIQUE AVEC RECOURS*

J. B. HIRIART-URRUTY†

Abstract. Necessary conditions are derived for stochastic programs with locally Lipschitz objective and locally Lipschitz constraints. The problem consists in choosing $x \in X$ so as to satisfy the constraints $x \in F_1$, $f_{1,i}(x) \leq 0$, $i \in \langle 1, m_1 \rangle$ and minimize the cost function $f_{1,0}(x) + E\{\varphi_{2,0}(x; \omega)\}$, where

$$\varphi_{2,0}(x; \omega) = \inf \{f_{2,0}(x, y; \omega) | y \in Q_2(x; \omega)\},$$

$$Q_2(x; \omega) = F_2 \cap \{y \in Y | f_{2,i}(x, y; \omega) \leq 0 \quad i \in \langle 1, m_2 \rangle\}.$$

It is shown that under regularity conditions, optimality conditions similar to those obtained in the convex case may be derived using the concept of generalized gradients.

1. Introduction. Soient X, Y, Z des espaces vectoriels euclidiens de dimension finie et (Ω, \mathcal{A}, P) un espace probabilisé abstrait; dans la pratique, Ω pourra être un borélien de Z et P la loi d'un élément aléatoire auxiliaire. Dans cette étude, on considère le problème suivant:

$$\mathcal{P}_1 \quad \begin{cases} \text{Trouver } x \in X \text{ tel que:} \\ x \in F_1, f_{1,i}(x) \leq 0 \quad \forall i \in \langle 1, m_1 \rangle \\ \text{et minimisant } f_{1,0}(x) + E\{\varphi_{2,0}(x; \omega)\}. \end{cases}$$

Dans ce problème, $\varphi_{2,0}(x; \omega)$ est déjà le résultat d'un problème d'optimisation, formulé de la façon suivante:

$$\mathcal{P}_2(x; \omega) \quad \begin{cases} \varphi_{2,0}(x; \omega) = \inf \{f_{2,0}(x, y; \omega) | y \in Q_2(x; \omega)\} \\ \text{avec } Q_2(x; \omega) = F_2 \cap \{y \in Y | f_{2,i}(x, y; \omega) \leq 0 \quad \forall i \in \langle 1, m_2 \rangle\}. \end{cases}$$

$F_1 \subset X$ et $F_2 \subset Y$ sont des fermés non vides; pour tout $i \in \langle 0, m_1 \rangle$ $f_{1,i}$ est définie sur X et à valeurs réelles; de même, pour tout $i \in \langle 0, m_2 \rangle$ les fonctions $f_{2,i}$ sont définies sur $X \times Y \times \Omega$, à valeurs réelles, et telles que:

$$\forall (x, y) \in X \times Y, \quad \omega \rightarrow f_{2,i}(x, y; \omega)$$

soient mesurables.

Le problème \mathcal{P}_1 et le problème auxiliaire $\mathcal{P}_2(x; \omega)$ sont la modélisation de problèmes d'optimisation stochastique à deux étages (ou problèmes de programmation stochastique avec recours). Le processus de décision est explicité de la manière suivante: le choix de x dans \mathcal{P}_1 (1er étage) doit être fait avant l'observation des éléments aléatoires du problème modélisé, de façon à satisfaire certaines contraintes:

$$x \in Q_1 = F_1 \cap \{x \in X | f_{1,i}(x) \leq 0, \forall i \in \langle 1, m_1 \rangle\}$$

(contraintes du 1er étage)

et à minimiser une certaine fonction de coût. En ce sens, le problème \mathcal{P}_1 est un problème déterministe. La décision du 2ème étage est prise lorsqu'on a observé une réalisation ω de l'élément aléatoire du problème; celui qui prend la décision a la possibilité de rectifier une situation résultant d'une décision antérieure x et de la réalisation ω . Il faut

* Received by the editors October 20, 1976, and in revised form May 26, 1977.

† Université de Clermont, Complexe Scientifique des Cézeaux, Département de Mathématiques Appliquées, Boîte Postale 45, 63170 Aubière, France.

pour cela résoudre le *programme de recours* $\mathcal{P}_2(x; \omega)$, c'est-à-dire choisir un recours $y \in Y$ satisfaisant certaines contraintes:

$$y \in Q_2(x; \omega) \quad (\text{contraintes du 2ème étage})$$

et minimisant la pénalité encourue $f_{2,0}(x, y; \omega)$. Naturellement, on supposera que pour la décision x et presque sûrement (p.s.) en ω un tel recours est possible.

Dans le cadre de la *programmation linéaire stochastique*, le problème \mathcal{P}_1 prend une forme simplifiée du fait que l'on suppose:

$$\text{P.L.S.} \quad \begin{cases} F_1 \text{ et } F_2 \text{ sont des polytopes;} \\ f_{1,0}(x) = \langle c_1, x \rangle, \quad f_{2,0}(x, y; \omega) = \langle c_2(\omega), y \rangle, \\ \forall i \in \langle 1, m_1 \rangle, \forall j \in \langle 1, m_2 \rangle \quad f_{1,i} \text{ et } f_{2,j} \text{ sont des fonctions affines.} \end{cases}$$

L'étude de tels programmes est l'objet d'une abondante littérature et on trouvera dans R. J. B. Wets [18] une revue de résultats sur le programme déterministe équivalent, la région d'admissibilité, les caractérisations des différents recours possibles, la possibilité de dualiser le problème primal, etc.

Dans une série d'articles [12]–[15] R. T. Rockafellar et R. J. B. Wets ont considéré le problème de *programmation convexe stochastique*, c'est-à-dire que l'on suppose que:

$$\text{P.C.S.} \quad \begin{cases} F_1 \text{ et } F_2 \text{ sont convexes;} \\ \forall i \in \langle 0, m_1 \rangle \quad f_{1,i}: X \rightarrow \mathbb{R} \text{ sont convexes;} \\ \forall i \in \langle 0, m_2 \rangle, \forall \omega \in \Omega \text{ les fonctions } f_{2,i}(\cdot, \cdot; \omega) \text{ sont} \\ \text{convexes (et finies) sur } X \times Y. \end{cases}$$

Plus précisément, le même processus de décision que celui qui a été explicité plus haut a été modélisé par une approche différente. Le programme stochastique était formulé de la façon suivante:

$$\mathcal{P}_0 \quad \begin{cases} \text{Trouver } x \in X \text{ et } y \in \mathcal{L}_Y^p \text{ tels que:} \\ x \in Q_1, \quad y(\omega) \in Q_2(x; \omega) \quad \text{p.s.} \\ \text{et minimisant } f_{1,0}(x) + E\{f_{2,0}(x, y(\omega); \omega)\}. \end{cases}$$

Ainsi le problème \mathcal{P}_0 se pose en termes d'optimisation dans $X \times \mathcal{L}_Y^p$ et sous certaines conditions [12] R. T. Rockafellar et R. J. B. Wets ont montré que les solutions et valeurs optimales de \mathcal{P}_0 étaient reliées à celles de \mathcal{P}_1 de la manière suivante: si x_0 est solution de \mathcal{P}_1 , il existe $\bar{y}_0 \in \mathcal{L}_Y^p$ tel que $\bar{y}_0(\omega)$ soit p.s. solution de $\mathcal{P}_2(x_0; \omega)$ et telle que (x_0, \bar{y}_0) soit solution de \mathcal{P}_0 ; à l'inverse, un couple (x_0, \bar{y}_0) solution de \mathcal{P}_0 fournit une solution x_0 de \mathcal{P}_1 et $\bar{y}_0(\omega)$ solution p.s. de $\mathcal{P}_2(x_0; \omega)$. Ainsi, le programme stochastique tel qu'il est posé en \mathcal{P}_0 , c'est-à-dire sous la forme *statique*, est susceptible d'être traité du point de vue de l'analyse convexe. En considérant des lagrangiens L sur $(X \times \mathcal{L}_Y^p) \times U$, où U est un espace de perturbations approprié, la théorie de la dualité a en particulier permis de donner des conditions nécessaires et suffisantes pour que $(x_0, \bar{y}_0; u)$ soit un point-selle de L [12]–[14]. Moyennant certaines hypothèses de régularité sur le problème, ces conditions (du type Kuhn–Tucker) sont aussi nécessaires et suffisantes pour que (x_0, \bar{y}_0) soit solution de \mathcal{P}_0 et u solution du problème dual associé.

Dans notre approche du problème stochastique, nous considérons les programmes stochastiques modélisés sous la forme *dynamique*, c'est-à-dire \mathcal{P}_1 et $\mathcal{P}_2(x; \omega)$. On considérera $F_1 \subset X$ et $F_2 \subset Y$ des fermés non vides et on supposera que:

$$\forall i \in \langle 0, m_1 \rangle, \quad f_{1,i}: X \rightarrow \mathbb{R},$$

$$\forall i \in \langle 0, m_2 \rangle, \quad \forall \omega \in \Omega, f_{2,i}(\cdot, \cdot; \omega): X \times Y \rightarrow \mathbb{R}$$

sont des fonctions localement lipschitziennes. Pour de telles fonctions, F. H. Clarke [2] a introduit la notion de gradient généralisé et nous exprimerons les conditions nécessaires d'optimalité sous une forme analogue à celles établies dans le cas convexe dans [14] ou [19], avec des gradients généralisés au lieu de sous-différentiels et gradients. Ces conditions nécessaires dans le cas stochastique seront dérivées de l'étude des conditions nécessaires d'optimalité dans le cadre déterministe [4], [6]. Toutefois, la structure particulière de la fonction $\varphi_{2,0}(x; \omega)$ intervenant dans la résolution de $\mathcal{P}_2(x; \omega)$ pose le problème de la détermination (ou du moins d'une estimation) du gradient généralisé de fonctions φ_F de la forme

$$\varphi_F(x) = \inf \{f(x, y) | y \in F(x)\},$$

c'est à dire des *fonctions marginales à contraintes dépendantes*.

La considération des gradients généralisés de fonctions marginales φ_F a fait l'objet d'une étude séparée dans le cadre déterministe [7]; nous utiliserons de manière extensive les résultats qui y figurent. Pour les notations et les rappels concernant le gradient généralisé d'une fonction localement lipschitzienne, le cône tangent et le cône normal à un fermé, on se référera au § 2 de [7]. Les principaux résultats de cet article ont été annoncés dans un Compte-Rendu à l'Académie des Sciences de Paris (t. 283, Série A, 1976, pp. 943–946).

2. Conditions d'optimalité pour des programmes stochastiques avec recours.

2.1. Nous allons examiner successivement le cas des programmes stochastiques sous la forme d'un modèle général et le cas des programmes où les contraintes du premier étage et du second étage sont explicitées à l'aide de fonctions données. Dans le *modèle général*, les contraintes du premier étage seront définies par $Q_1 \subset X$ (sans autre précision sur la façon de définir Q_1); de même les contraintes du deuxième étage seront définies à partir de x , décision du premier étage, et de $\mathcal{G}(\omega) \subset X \times Y$.

Les données du problème sont donc:

$$\begin{aligned} \text{les fonctions de coût:} \quad & f_{1,0}: X \rightarrow \mathbb{R}, \\ & f_{2,0}: X \times Y \times \Omega \rightarrow \mathbb{R}; \\ \text{les définitions des contraintes:} \quad & Q_1 \subset X, \\ & \mathcal{G}: \Omega \rightarrow X \times Y. \end{aligned}$$

Le programme de recours se formule de la façon suivante: connaissant $x \in X$ et ayant observé la réalisation ω , il s'agit de choisir le recours $y \in Y$ satisfaisant certaines contraintes

$$(i) \quad (x, y) \in \mathcal{G}(\omega),$$

et minimisant le coût de pénalisation associé qui est: $f_{2,0}(x, y; \omega)$.

Désignons par Q_2 la multiapplication de $X \times \Omega$ dans Y dont le graphe dans $X \times Y$ est $\mathcal{G}(\omega)$, c'est à dire:

$$Q_2(x; \omega) = \{y \in Y | (x, y) \in \mathcal{G}(\omega)\}.$$

Le recours choisi dépend de x et ω ; nous le noterons $\bar{y}(x; \omega)$ (ou $\bar{y}(\omega)$ lorsqu'aucune ambiguïté n'est possible), il doit être tel que:

$$\mathcal{P}_2(x, \omega) \quad f_{2,0}(x, \bar{y}; \omega) = \varphi_{2,0}(x; \omega) = \inf \{f_{2,0}(x, y; \omega) | y \in Q_2(x; \omega)\}.$$

Nous désignerons par $M_{Q_2}(x; \omega)$ l'ensemble des solutions du problème $\mathcal{P}_2(x; \omega)$. Dans le programme du premier étage, on cherche $x \in X$ tel que:

$$(ii) \quad x \in Q_1,$$

et minimisant le coût $f_{1,0}(x) + E\{\varphi_{2,0}(x; \omega)\}$

$$\mathcal{P}_1 \quad \inf_{x \in Q_1} (f_{1,0}(x) + E\{\varphi_{2,0}(x; \omega)\}).$$

Concernant les contraintes et les fonctions de coût, nous ferons les hypothèses générales suivantes:

$$(S_1) \quad \begin{cases} Q_1 \text{ est un fermé non vide de } X, \\ \mathcal{G} \text{ est une multiapplication mesurable à valeurs fermées non vides de } X \times Y. \end{cases}$$

$$(S_2) \quad \begin{cases} f_{1,0} \text{ est localement lipschitzienne sur } X, \\ f_{2,0}: X \times Y \times \Omega \rightarrow \mathbb{R} \text{ est localement lipschitzienne sur } X \times Y \text{ pour } \omega \in \Omega \\ \text{et mesurable en } \omega \text{ pour tout } (x, y) \in X \times Y. \end{cases}$$

Nous supposons également que pour toute décision $x \in Q_1$, il existe p.s. la possibilité d'un recours $y(x; \omega)$, c'est-à-dire:

$$(R_1) \quad \forall x \in Q_1, \quad Q_2(x; \omega) \text{ est p.s. non vide}$$

(ou équivalamment $Q_1 \subset \Pi_X \mathcal{G}(\omega)$ p.s.). Cela signifie en clair qu'il y a la possibilité de rectifier p.s. une situation résultant d'un choix antérieur $x \in Q_1$ et de la réalisation ω .

Dans la mesure où pour tout $x \in Q_1$, il existe $y(x; \cdot) \in \mathcal{L}_Y^\infty$ tel que $y(x; \omega) \in Q_2(x; \omega)$ p.s., la condition (R_1) est la condition de *recours relativement complet* de R. T. Rockafellar et R. J. B. Wets [13].

LEMME 1. *Sous les hypothèses (S_1) , (S_2) , (R_1) , pour tout $x \in Q_1$, l'application $\omega \rightarrow \varphi_{2,0}(x; \omega)$ est mesurable à valeurs dans $\mathbb{R} \cup \{-\infty\}$ p.s. et la multiapplication $\omega \rightarrow M_{Q_2}(x; \omega)$ est mesurable.*

Démonstration. $x \in Q_1$ étant fixé, nous avons:

$$\{x\} \times Q_2(x; \omega) = (\{x\} \times Y) \cap \mathcal{G}(\omega),$$

et la mesurabilité de \mathcal{G} entraîne celle de la multiapplication $\omega \rightarrow Q_2(x; \omega)$ [11, Théorème 1M].

D'après l'hypothèse (R_1) , $\varphi_2(x; \omega) \in \mathbb{R} \cup \{-\infty\}$ p.s. L'application $(y; \omega) \rightarrow f_{2,0}(x, y; \omega)$ est un intégrande de Carathéodory sur $Y \times \Omega$, donc un intégrande normal sur $Y \times \Omega$ [11, p. 19]. La multiapplication $\omega \rightarrow Q_2(x; \omega)$ étant mesurable, la mesurabilité de l'application $\omega \rightarrow \varphi_{2,0}(x; \omega)$ et celle de la multiapplication $\omega \rightarrow M_{Q_2}(x; \omega)$ résultent du Théorème 2K dans [11].

Signalons également la propriété de mesurabilité des multiapplications: cône normal et cône tangent à $\mathcal{G}(\omega)$ en $(x(\omega), y(\omega))$.

LEMME 2. *Soit $(x, y): \Omega \rightarrow X \times Y$ une application mesurable telle que $(x(\omega), y(\omega)) \in \mathcal{G}$ p.s. Alors les multiapplications*

$$N_{\mathcal{G}}: \omega \rightarrow N_{\mathcal{G}(\omega)}(x(\omega), y(\omega)) \quad \text{et} \quad \mathcal{T}_{\mathcal{G}}: \omega \rightarrow \mathcal{T}_{\mathcal{G}(\omega)}(x(\omega), y(\omega))$$

sont mesurables.

Démonstration. Pour ω tel que $(x(\omega), y(\omega)) \notin \mathcal{G}(\omega)$, $N_{\mathcal{G}(\omega)}(x(\omega), y(\omega)) = \emptyset$ (voir la convention sur le cône normal [7, § 2]) et il est immédiat que:

$$\text{dom } N_{\mathcal{G}} = \{\omega \in \Omega \mid (x(\omega), y(\omega)) \in \mathcal{G}(\omega)\} \quad \text{est mesurable.}$$

La multiapplication \mathcal{G} étant mesurable, l'application $\omega \rightarrow d_{\mathcal{G}(\omega)}(x, y)$ est mesurable pour x, y fixés [11, Proposition 1A]. Ainsi, la multiapplication $\omega \rightarrow \partial_{(x,y)} d_{\mathcal{G}(\omega)}(x(\omega), y(\omega))$ est mesurable [2, Corollaire 1.17]. Or, $N_{\mathcal{G}(\omega)}(x(\omega), y(\omega))$ est le plus petit cône convexe fermé contenant $\partial_{(x,y)} d_{\mathcal{G}(\omega)}(x(\omega), y(\omega))$; la multiapplication $N_{\mathcal{G}}$ est donc mesurable [11, Proposition 1H]. La multiapplication $\mathcal{T}_{\mathcal{G}}$ qui n'est autre que la polaire de la précédente est également mesurable [11, Corollaire 2T].

Sous les hypothèses du Lemme 1, il se peut que le programme de recours $\mathcal{P}_2(x; \omega)$ ne donne pas de valeur optimale finie (c'est à dire $\varphi_{2,0}(x; \omega) = -\infty$ et par conséquent $M_{Q_2}(x; \omega) = \emptyset$). Il est plus habituel (et plus raisonnable) de supposer que le programme de recours est résoluble (du moins au voisinage de la décision optimale x_0), c'est-à-dire:

$$(R_2) \quad \forall x \in Q_1, \quad M_{Q_2}(x; \omega) \text{ est p.s. non vide.}$$

Naturellement, cette hypothèse est plus forte que la condition (R_1) .

Le lemme suivant donne une condition suffisante pour qu'une fonctionnelle moyenne $E\varphi$ (dont $E\varphi_{2,0}$ est un exemple) soit lipschitzienne au voisinage d'un point x_0 . En outre, il permet de comparer le gradient généralisé de $E\varphi$ en x_0 à l'espérance de la multiapplication: $\omega \rightarrow \partial\varphi(x_0; \omega)$. Rappelons, à ce sujet, que:

$$E\{\partial\varphi(x_0; \omega)\} = \{E(X^*)|X^* \in \mathcal{L}_X^1, X^*(\omega) \in \partial\varphi(x_0; \omega) \text{ p.s.}\}.$$

Le résultat qui suit a été démontré dans le cas où X est un Banach séparable dans [17]. La démonstration en est aisée en pensant à la caractérisation de $E\{\partial\varphi(x_0; \omega)\}$ à l'aide des fonctions d'appui et en appliquant l'inégalité de Fatou-Lebesgue aux fonctions d'appui.

LEMME 3. Soit $\varphi: X \times \Omega \rightarrow \mathbb{R}$ un intégrande et $x_0 \in X$. On suppose que:

$$(S_3) \quad \begin{cases} \text{(i)} & E\{|\varphi(x_0; \omega)|\} < +\infty, \\ \text{(ii)} & \text{il existe un voisinage } V_0 \text{ de } x_0 \text{ et une fonction } k \in \mathcal{L}_+^1 \text{ tels que:} \\ & \forall x, y \in V_0, \quad |\varphi(x; \omega) - \varphi(y; \omega)| \leq k(\omega)\|x - y\| \text{ p.s.} \end{cases}$$

Alors, $E\varphi$ est lipschitzienne au voisinage de x_0 et on a:

$$(2.1) \quad \partial(E\varphi)(x_0) \subset E\{\partial\varphi(x_0; \omega)\}.$$

Ce lemme permet aussi de donner une condition suffisante pour que $\partial E\varphi(x_0)$ soit réduit à un seul élément (qui est nécessairement $\nabla E\varphi(x_0)$). Pour $x_0 \in X$, désignons par D_{x_0} les éléments $\omega \in \Omega$ pour lesquels $\partial\varphi(x_0; \omega)$ n'est pas réduit à un seul élément. Nous avons alors le résultat suivant:

LEMME 4. Soit φ et $x_0 \in X$ vérifiant les hypothèses du Lemme 3; alors:

$$(2.2) \quad P(D_{x_0}) = 0 \Rightarrow \begin{cases} \text{(iii)} & E\varphi \text{ est différentiable en } x_0 \text{ et } \nabla E\varphi \text{ est continue} \\ & \text{en } x_0 \text{ relativement à l'ensemble où } \nabla E\varphi \text{ existe,} \\ \text{(iv)} & \nabla E\varphi(x_0) = \int_{D_{x_0}^c} \nabla \varphi(x_0; \omega) dP(\omega). \end{cases}$$

Démonstration. La multiapplication $\omega \rightarrow \partial\varphi(x_0; \omega)$ étant mesurable, pour tout $d \in X$, la fonction d'appui dans la direction d : $\omega \rightarrow \varphi^*(x_0; d; \omega)$ est mesurable. Ainsi l'épaisseur de $\partial\varphi(x_0; \omega)$ dans la direction $d \neq 0$ qui est définie par:

$$\omega \rightarrow e_d(\omega) = \varphi^*\left(x_0; \frac{d}{\|d\|}; \omega\right) - \varphi_*\left(x_0; \frac{d}{\|d\|}; \omega\right)$$

est mesurable.

De même, la diamètre $\Delta(\omega)$ de $\partial\varphi(x_0; \omega)$ est une fonction mesurable de ω car:

$$\forall \rho > 0, \quad \Delta(\omega) = \max_{\|d\|=\rho} e_d(\omega).$$

En conséquence:

$$D_{x_0} = \{\omega \in \Omega \mid \Delta(\omega) > 0\} \quad \text{est probabilisable.}$$

Comme

$$\forall \omega \in D_{x_0}^c, \quad \partial\varphi(x_0; \omega) = \{\nabla\varphi(x_0; \omega)\},$$

il résulte de (2.1) que $\partial E\varphi(x_0)$ est réduit à un seul élément, d'où (iii) [2, Proposition 1.10] et (iv).

Remarque. Contrairement au cas convexe [1], on ne peut affirmer qu'une condition suffisante que (iii) et (iv) aient lieu est que $P(D'_{x_0}) = 0$, où D'_{x_0} désigne les éléments $\omega \in \Omega$ pour lesquels $\varphi(\cdot; \omega)$ n'est pas différentiable en x_0 .

Nous sommes en mesure à présent de donner une condition nécessaire pour que $x_0 \in X$ soit solution du problème \mathcal{P}_1 . Nous serons amenés à faire l'hypothèse suivante sur la multiapplication M_{Q_2} :

$$(S_4) \quad \begin{aligned} &\exists \bar{y}: X \times \Omega \rightarrow Y \text{ mesurable en } \omega \text{ tel que } \bar{y}(x; \omega) \in M_{Q_2}(x; \omega) \text{ p.s.} \\ &\text{au voisinage de } x_0 \text{ et tel que } \bar{y}(\cdot; \omega) \text{ soit p.s. continue en } x_0. \end{aligned}$$

La mesurabilité de $\bar{y}(x; \cdot)$ peut être déduite des hypothèses précédentes. La condition de continuité en x_0 est une hypothèse simplificatrice qui peut être affaiblie (voir la Remarque 1 suivant le Théorème 5).

THEOREME 5. *Soit $x_0 \in X$; on suppose que $\varphi_{2,0}$ vérifie l'hypothèse (S_3) et que l'hypothèse (S_4) est vérifiée. Une condition nécessaire pour que x_0 soit solution du problème \mathcal{P}_1 est que:*

$$(2.3) \quad \left\{ \begin{aligned} &(a) \quad x_0 \in Q_1, \\ &(b) \quad \text{il existe } y_0: \Omega \rightarrow Y, \text{ mesurable tel que } (x_0, y_0(\omega)) \in \mathcal{G}(\omega) \text{ p.s.;} \\ &\quad \text{de plus, il existe } \rho \in \mathcal{L}_X^1 \text{ tel que:} \\ &(c) \quad v - E\rho \in \partial f_{1,0}(x_0) \text{ et } -v \text{ est normal à } Q_1 \text{ en } x_0, \\ &(d) \quad (\rho(\omega), 0) \in \partial_{(x,y)} f_{2,0}(x_0, y_0(\omega); \omega) + N_{\mathcal{G}(\omega)}(x_0; y_0(\omega)) \text{ p.s.} \end{aligned} \right.$$

Démonstration. Soit f définie au voisinage de x_0 par:

$$\forall x \in X, \quad f(x) = f_{1,0}(x) + E\{\varphi_{2,0}(x; \omega)\}.$$

On écrira simplement:

$$\bar{f} = f_{1,0} + E\varphi_{2,0}.$$

x_0 étant solution du problème \mathcal{P}_1 , on a nécessairement [4], [6]

$$(2.4) \quad 0 \in \partial f(x_0) + N(Q_1; x_0).$$

Les hypothèses faites assurent que $f_{1,0}$ et $E\varphi_{2,0}$ sont lipschitziennes au voisinage de x_0 et par suite:

$$\partial f(x_0) \subset \partial f_{1,0}(x_0) + \partial E\varphi_{2,0}(x_0).$$

Ainsi, d'après (2.4):

$$(2.5) \quad 0 \in \partial f_{1,0}(x_0) + \partial E\varphi_{2,0}(x_0) + N(Q_1; x_0).$$

Soient $v_1, v_2, v \in X$ tels que:

$$\begin{aligned} v_1 &\in \partial f_{1,0}(x_0), & v_2 &\in \partial E\varphi_{2,0}(x_0), \\ v_1 + v_2 &= v & \text{et} & \quad -v \in N(Q_1; x_0). \end{aligned}$$

D'après le lemme 3, $\partial E\varphi_{2,0}(x_0) \subset E\partial\varphi_{2,0}(x_0)$. Par conséquent, il existe $\rho \in \mathcal{L}_X^1$ tel que:

$$v_2 = E\rho, \quad \rho(\omega) \in \partial\varphi_{2,0}(x_0; \omega) \quad \text{p.s.}$$

Par application du Théorème 8 de [7],

$$\rho(\omega) \in \partial\varphi_{2,0}(x_0; \omega) \Rightarrow (\rho(\omega), 0) \in \partial_{(x,y)} f_{2,0}(x_0, y_0(\omega); \omega) + N_{\mathcal{G}(\omega)}(x_0; y_0(\omega))$$

avec $y_0(\omega) = \bar{y}(x_0; \omega)$ p.s.; d'où le résultat annoncé.

Remarques. 1. Les hypothèses du théorème sont des conditions suffisantes permettant d'assurer que $E\varphi_{2,0}$ existe et est lipschitzienne au voisinage de x_0 . L'hypothèse de continuité p.s. de l'intégrande sélection \bar{y} est faite *par souci de simplification*; dans le cas où \bar{y} serait seulement borné au voisinage de x_0 , une formule analogue à (2.3) peut être établie avec $\bar{Y}_0(\omega) = \limsup_{x \rightarrow x_0} \{\bar{y}(x; \omega)\}$. Dans le cas linéaire, des conditions sur les moments du second ordre des éléments aléatoires du problème suffisent dans certains cas à avoir une condition de Lipschitz sur $E\varphi_{2,0}$ [8]. De la même manière, il est aisé de voir que, sous les hypothèses suivantes:

$$(S_5) \quad \left\{ \begin{array}{l} \text{l'application sélection } \bar{y} \text{ de } M_{Q_2} \text{ lipschitzienne en } x \text{ au voisinage de } x_0, \\ \text{de coefficient de Lipschitz } \bar{k}(\omega), \\ f_{2,0}(\cdot, \cdot; \omega) \text{ lipschitzienne au voisinage de } (x_0, \bar{y}(x_0; \omega)), \text{ de coefficient } k(\omega), \\ E\{|\varphi_{2,0}(x_0; \omega)|\} < +\infty, \\ k \text{ et } \bar{k} \in \mathcal{L}_+^2, \end{array} \right.$$

les hypothèses du théorème précédent sont satisfaites.

2. La multiapplication $\omega \rightarrow N_{\mathcal{G}(\omega)}(x_0; y_0(\omega))$ étant mesurable (Lemme 2), la partie (d) des conditions d'optimalité peut prendre la forme décomposée suivante:

$$(d') \quad \left\{ \begin{array}{l} \exists \rho_1, \rho_2: \Omega \rightarrow X, \rho_3: \Omega \rightarrow Y, \text{ mesurables tels que:} \\ (\rho_1(\omega), \rho_3(\omega)) \in \partial_{(x,y)} f_{2,0}(x_0, y_0(\omega); \omega) \quad \text{p.s.} \\ (\rho_2(\omega), -\rho_3(\omega)) \text{ normal à } \mathcal{G}(\omega) \text{ en } (x_0, y_0(\omega)) \quad \text{p.s.} \\ \rho(\omega) = \rho_1(\omega) + \rho_2(\omega) \quad \text{p.s.} \end{array} \right.$$

Au voisinage de x_0 , on a:

$$\varphi_{2,0}(x; \omega) = f_{2,0}(x, \bar{y}(x; \omega); \omega)$$

avec \bar{y} application sélection de M_{Q_2} . Les propriétés de $\bar{y}(\cdot, \omega)$ en tant que fonction de x ne sont pas, dans le cas général, facilement discernables. Cela dépend essentiellement de la structure de M_{Q_2} ; un cas particulièrement intéressant est le cas où le programme de recours a une seule solution au voisinage de la solution x_0 , c'est-à-dire:

$$M_{Q_2}(x; \omega) = \{\bar{y}(x; \omega)\} \quad \text{au voisinage de } x_0.$$

Si $\bar{y}(\cdot; \omega)$ est différentiable en x_0 , nous désignerons par $J_{\bar{y}}(x_0; \omega)$ la matrice jacobienne de $\bar{y}(\cdot; \omega)$ en x_0 ; il est clair que l'application $\omega \rightarrow J_{\bar{y}}(x_0; \omega)$ (ou $[J_{\bar{y}}(x_0; \omega)]^*$) est mesurable. Comme précédemment, posons $\bar{y}(x_0; \omega) = y_0(\omega)$. Sous des hypothèses de différentiabilité de $\bar{y}(\cdot, \omega)$ en x_0 , la condition nécessaire d'optimalité prend la forme suivante:

THEOREME 6. Soit $x_0 \in X$; on suppose que $\varphi_{2,0}$ vérifie (S_3) et qu'il existe une application sélection \bar{y} de M_{Q_2} , mesurable en ω , telle que:

(S₆) p.s. $\bar{y}(\cdot; \omega)$ est différentiable en x_0 , de dérivée forte en x_0 [9, p. 71].

Alors, une condition nécessaire pour que x_0 soit solution de \mathcal{P}_1 est que:

- (a) $x_0 \in Q_1$,
- (b) $\exists \rho \in \mathcal{L}_X^1$, $v \in X$ tels que: $v - E\rho \in \partial f_{1,0}(x_0)$ et $-v$ normal à Q_1 en x_0 ,
- (c) $\exists (\rho_1, \rho_2): \Omega \rightarrow X \times Y$ mesurable tels que:

$$(\rho_1(\omega), \rho_2(\omega)) \in \partial_{(x,y)} f_{2,0}(x_0, y_0(\omega); \omega) \quad \text{p.s.}$$

$$\rho(\omega) = \rho_1(\omega) + [J_{y_0}(x_0; \omega)]^* \rho_2(\omega).$$

Démonstration. Au voisinage de x_0 , nous avons:

$$\varphi_{2,0}(x; \omega) = f_{2,0}(x, \bar{y}(x; \omega); \omega) \quad \text{p.s.}$$

On connaît, d'autre part, l'existence de $\rho \in \mathcal{L}_X^1$ et de $v \in X$ tels que:

$$v - E\rho \in \partial f_{1,0}(x_0), \quad -v \in N_{Q_1}(x_0).$$

De plus, (voir la démonstration de la Proposition 10 dans [7])

$$(2.6) \quad \text{p.s.} \quad (\rho(\omega), 0) \in \partial_{(x,y)} f_{2,0}(x_0, \bar{y}(x_0; \omega); \omega) + \{(u, v) | u + [J_{\bar{y}}(x_0; \omega)]^* v = 0\}.$$

$\bar{y}(x_0; \cdot)$ étant mesurable, la multiapplication $\omega \rightarrow \partial_{(x,y)} f_{2,0}(x_0, \bar{y}(x_0; \omega); \omega)$ est mesurable. Il est facile de voir que la multiapplication Δ définie par:

$$\Delta(\omega) = \{(u, v) | u + [J_{\bar{y}}(x_0; \omega)]^* v = 0\}$$

est aussi mesurable. Considérons la multiapplication:

$$\Gamma: \omega \rightarrow \Gamma(\omega) = [\partial_{(x,y)} f_{2,0}(x_0, \bar{y}(x_0; \omega); \omega) - (\rho(\omega), 0)] \cap [-\Delta(\omega)].$$

Γ est mesurable [11, Théorème 1M] et d'après (2.6):

$$\Gamma(\omega) \neq \emptyset \quad \text{p.s.}$$

Soit (σ_1, σ_2) une sélection mesurable de Γ [11, Corollaire 1C], nous avons:

$$\sigma_1(\omega) + [J_{\bar{y}}(x_0; \omega)]^* \sigma_2(\omega) = 0 \quad \text{p.s.}$$

$$(\sigma_1(\omega) + \rho(\omega), \sigma_2(\omega)) \in \partial_{(x,y)} f_{2,0}(x_0, \bar{y}(x_0; \omega); \omega) \quad \text{p.s.}$$

En posant:

$$\rho_2(\omega) = \sigma_2(\omega), \quad \rho_1(\omega) = \rho(\omega) + \sigma_1(\omega),$$

et sachant que $\bar{y}(x_0; \omega) = y_0(\omega)$, on a le résultat annoncé.

2.2. Nous allons examiner à présent le modèle de programmation stochastique avec recours où les contraintes du premier étage (contraintes explicites) et les contraintes du deuxième étage (contraintes induites) sont exprimées à l'aide respectivement de fonctions déterministes et de fonctions aléatoires. Ainsi:

$$(2.7) \quad Q_1 = F_1 \cap \{x \in X | f_{1,i}(x) \leq 0, i \in \langle 1, m_1 \rangle\}$$

et pour le programme de recours:

$$(2.8) \quad Q_2(x; \omega) = F_2 \cap \{y \in Y | f_{2,i}(x, y; \omega) \leq 0, i \in \langle 1, m_2 \rangle\}.$$

F_1 est un fermé non vide de X et F_2 un fermé non vide de Y ; on fait sur $f_{1,i}$ (resp. $f_{2,i}$) les mêmes hypothèses générales que sur $f_{1,0}$ (resp. $f_{2,0}$), c'est à dire (S_2) .

Explicitons les conditions de Kuhn-Tucker de base lorsque Q_1 et $Q_2(x; \omega)$ ont la forme indiquée en (2.7) et (2.8).

CONDITIONS DE KUHN-TUCKER (forme du gradient généralisé)

(a) $x_0 \in F_1$, on a $\{\lambda_{1,i}\}_{i \in \langle 1, m_1 \rangle}$ tel que:

$$\forall i \in \langle 1, m_1 \rangle, \quad \lambda_{1,i} \geq 0, \quad f_{1,i}(x_0) \leq 0, \quad \lambda_{1,i} \cdot f_{1,i}(x_0) = 0.$$

(b) $\rho \in \mathcal{L}_X^1$, $v \in X$ tels que:

$$v - E\rho \in \partial f_{1,0}(x_0) + \sum_{i=1}^{m_1} \lambda_{1,i} \cdot \partial f_{1,i}(x_0),$$

$$-v \text{ normal à } F_1 \text{ en } x_0.$$

(c) $\exists w: \Omega \rightarrow Y$ mesurable, $\lambda_{2,i}: \Omega \rightarrow \mathbb{R} (i \in \langle 1, m_2 \rangle)$ mesurables tels que:

$$\text{p.s. } (\rho(\omega), w(\omega)) \in \partial_{(x,y)} f_{2,0}(x_0, y_0(\omega); \omega) + \sum_{i=1}^{m_2} \lambda_{2,i}(\omega) \partial_{(x,y)} f_{2,i}(x_0, y_0(\omega); \omega),$$

$$\forall i \in \langle 1, m_2 \rangle, \quad \lambda_{2,i}(\omega) \geq 0, \quad \lambda_{2,i}(\omega) \cdot f_{2,i}(x_0, y_0(\omega); \omega) = 0,$$

$$-w(\omega) \text{ est normal à } F_2 \text{ en } y_0(\omega).$$

On posera $y_0(\omega) = \bar{y}(x_0; \omega)$; on désignera par $I_1(x_0)$ (resp. $I_2(x_0, y_0(\omega))$) l'ensemble des contraintes actives en x_0 (resp. en $(x_0, y_0(\omega))$).

THEOREME 7. *On suppose que $\varphi_{2,0}$ vérifie (S_3) et que l'hypothèse (S_4) est vérifiée. Il est supposé de plus que*

$$(S_7) \quad \exists \bar{d}_1 \in \mathcal{T}(F_1; x_0) \quad \forall i \in I_1(x_0), \quad f_{1,i}^*(x_0; \bar{d}_1) < 0,$$

$$(S_8) \quad \exists (d_1, d_2): \Omega \rightarrow X \times Y \text{ mesurable tel que:}$$

$$\text{p.s. } d_2(\omega) \in \mathcal{T}(F_2; y_0(\omega)), \quad f_{2,i}^*(x_0, y_0(\omega); d_1(\omega), d_2(\omega); \omega) < 0 \quad \forall i \in I_2(x_0, y_0(\omega)).$$

Alors une condition nécessaire pour que x_0 soit solution de \mathcal{P}_1 est que les conditions de Kuhn-Tucker soient vérifiées.

Démonstration. x_0 étant une solution optimale de \mathcal{P}_1 , on a:

$$x_0 \in F_1, \quad f_{1,i}(x_0) \leq 0 \quad \forall i \in \langle 1, m_1 \rangle,$$

$$0 \in \partial f_{1,0}(x_0) + \partial E\varphi_{2,0}(x_0) + N(Q_1; x_0).$$

De par l'hypothèse (S_7) et les résultats de [6], on a:

$$(2.9) \quad N(Q_1; x_0) \subset N(F_1; x_0) + \mathbb{R}^+ \partial \left(\text{Max}_{i \in \langle 1, m_1 \rangle} f_{1,i} \right)(x_0).$$

Soient v, v_1, v_2, v_3 tels que:

$$v = v_1 + v_2 + v_3, \quad -v \in N(F_1; x_0),$$

$$v_1 \in \partial f_{1,0}(x_0), \quad v_2 \in \partial E\varphi_{2,0}(x_0), \quad v_3 \in \mathbb{R}^+ \partial \left(\text{Max}_{i \in \langle 1, m_1 \rangle} f_{1,i} \right)(x_0).$$

$v_2 \in \partial E\varphi_{2,0}(x_0)$ et d'après le résultat du lemme 3, il existe $\rho \in \mathcal{L}_X^1$ tel que:

$$v_2 = E\rho, \quad \rho(\omega) \in \partial \varphi_{2,0}(x_0; \omega) \quad \text{p.s.}$$

D'autre part [4, Proposition 9], [6, Lemme 11]

$$v_3 = \sum_{i=1}^{m_1} \lambda_{1,i} \cdot \partial f_{1,i}(x_0)$$

avec

$$\forall i \in \langle 1, m_1 \rangle, \quad \lambda_{1,i} \geq 0 \quad \text{et} \quad \lambda_{1,i} \cdot f_{1,i}(x_0) = 0.$$

En résumé:

$$v - E\rho = v_1 + v_3 \in \partial f_{1,0}(x_0) + \sum_{i=1}^{m_1} \lambda_{1,i} \cdot \partial f_{1,i}(x_0),$$

d'où les conditions (a) et (b) de Kuhn-Tucker.

Soit $\mathcal{G}(\omega)$ le graphe dans $X \times Y$ de la multiapplication $x \rightarrow Q_2(x; \omega)$ défini en (2.8), c'est à dire:

$$(2.10) \quad \mathcal{G}(\omega) = (X \times F_2) \cap \{(x, y) \in X \times Y \mid f_{2,i}(x, y; \omega) \leq 0 \quad \forall i \in \langle 1, m_2 \rangle\}.$$

Rappelant que $\bar{y}(x_0; \omega) = y_0(\omega)$, nous avons grâce à l'hypothèse de régularité (S₈):

$$(2.11) \quad \text{p.s.} \quad N_{\mathcal{G}(\omega)} \subset \{0\} \times N(F_2; y_0(\omega)) + \mathbb{R}^+ \left(\text{Max}_{i \in \langle 1, m_2 \rangle} f_{2,i} \right) (x_0, y_0(\omega); \omega).$$

Nous savons d'autre part que:

$$(2.12) \quad \text{p.s.} \quad (\rho(\omega), 0) \in \partial_{(x,y)} f_{2,0}(x_0, y_0(\omega); \omega) + N_{\mathcal{G}(\omega)}(x_0, y_0(\omega); \omega)$$

soit

$$(2.13) \quad \text{p.s.} \quad (\rho(\omega), 0) \in \{0\} \times N(F_2; y_0(\omega)) + \partial_{(x,y)} f_{2,0}(x_0, y_0(\omega); \omega) + \mathbb{R}^+ \partial \left(\text{Max}_{i \in \langle 1, m_2 \rangle} f_{2,i} \right) (x_0, y_0(\omega); \omega).$$

Les multiapplications:

$$\begin{aligned} \omega &\rightarrow \{0\} \times N(F_2; y_0(\omega)), \\ \omega &\rightarrow \partial_{(x,y)} f_{2,0}(x_0, y_0(\omega); \omega), \\ \omega &\rightarrow \mathbb{R}^+ \partial \left(\text{Max}_{i \in \langle 1, m_2 \rangle} f_{2,i} \right) (x_0, y_0(\omega); \omega) \end{aligned}$$

sont mesurables et on déduit aisément de (2.13) l'existence de $w: \Omega \rightarrow Y$ mesurable tel que:

$$(2.14) \quad \begin{aligned} \text{p.s.} \quad (0, -w(\omega)) &\in \{0\} \times N(F_2; y_0(\omega)), \\ (\rho(\omega), w(\omega)) &\in \partial_{(x,y)} f_{2,0}(x_0, y_0(\omega); \omega) + \mathbb{R}^+ \partial \left(\text{Max}_{i \in \langle 1, m_2 \rangle} f_{2,i} \right) (x_0, y_0(\omega); \omega). \end{aligned}$$

Considérons $(\rho_1, w_1), (\rho_2, w_2)$ mesurables tels que:

$$(2.15) \quad \begin{aligned} (\rho_1, w_1) + (\rho_2, w_2) &= (\rho, w), \\ \text{p.s.} \quad (\rho_2(\omega), w_2(\omega)) &\in \mathbb{R}^+ \partial \left(\text{Max}_{i \in \langle 1, m_2 \rangle} f_{2,i} \right) (x_0, y_0(\omega)). \end{aligned}$$

Soit Γ la multiapplication définie par:

$$\Gamma(\omega) = \left\{ (\lambda_1, \dots, \lambda_{m_2}) \forall i \in \langle 1, m_2 \rangle, \lambda_i \geq 0, \lambda_i \cdot f_{2,i}(x_0, y_0(\omega); \omega) = 0, \right. \\ \left. (\rho_2(\omega), w_2(\omega)) \in \sum_{i=1}^{m_2} \lambda_i \partial_{(x,y)} f_{2,i}(x_0, y_0(\omega); \omega) \text{ p.s.} \right\}.$$

Γ est une multiapplication mesurable, à valeurs fermées et d'après (2.15)

$$\Gamma(\omega) \neq \emptyset \quad \text{p.s.}$$

On peut ainsi choisir $(\lambda_1, \dots, \lambda_{m_2})$ mesurables tels que:

$$\text{p.s.} \quad (\rho_2(\omega), w_2(\omega)) \in \sum_{i=1}^{m_2} \lambda_{2,i}(\omega) \partial_{(x,y)} f_{2,i}(x_0, y_0(\omega); \omega) \\ \forall i \in \langle 1, m_2 \rangle, \quad \lambda_{2,i}(\omega) \geq 0, \quad \lambda_{2,i}(\omega) \cdot f_{2,i}(x_0, y_0(\omega); \omega) = 0,$$

d'où la condition (c) de Kuhn–Tucker.

Remarques. 1. Les hypothèses (S₇) et (S₈) sont des hypothèses de régularité relatives à la forme de représentation de Q_1 et de $Q_2(x; \omega)$; elles peuvent être affaiblies comme dans le cas déterministe [6].

2. Les conditions nécessaires du type Kuhn–Tucker (forme du gradient généralisé) sont semblables aux conditions nécessaires (et suffisantes) établies dans le cas convexe (forme du sous-différentiel, [14]) et dans le cas convexe différentiable (forme du gradient, [19]). Les conditions (b) et (c) peuvent être particularisées dans le cas où les fonctions $f_{1,i}$ et $f_{2,i}$ sont C^1 et dans le cas où F_1 et F_2 ont des structures particulières. Nous allons en donner un exemple.

Soient $F_1 \subset X$ et $F_2 \subset Y$ définis par:

$$F_1 = \{x \in X | h_1(x) = 0\},$$

$$F_2 = \{y \in Y | h_2(y) = 0\}.$$

THEOREME 8. *On suppose que $\varphi_{2,0}$ vérifie (S₃) et que l'hypothèse (S₄) est vérifiée. Il est supposé de plus que:*

(S'₇) h_1 est différentiable en x_0 , de dérivée forte $\nabla h_1(x_0)$ vérifiant:

$$\forall i \in I_1(x_0), \quad f_{1,i}^*(x_0; \nabla h_1(x_0)) < 0,$$

(S'₈) h_2 est différentiable en $y_0(\omega)$ p.s., de dérivée forte $\nabla h_2(y_0(\omega))$ et il existe $d_1: \Omega \rightarrow Y$ mesurable tel que:

$$\forall i \in I_2(x_0, y_0(\omega)), \quad f_{2,i}^*(x_0, y_0(\omega); d_1(\omega), \nabla h_2(y_0(\omega)); \omega) < 0 \quad \text{p.s.}$$

Alors une condition nécessaire pour que x_0 soit solution de \mathcal{P}_1 est que:

(a) $h_1(x_0) = 0$; on a $\{\lambda_{1,i}\}_{i \in \langle 1, m_1 \rangle}$ tel que:

$$\forall i \in \langle 1, m_1 \rangle, \quad \lambda_{1,i} \geq 0, \quad f_{1,i}(x_0) \leq 0, \quad \lambda_{1,i} \cdot f_{1,i}(x_0) = 0,$$

(b) $\exists \rho \in \mathcal{L}_X^1, v \in X$ tels que:

$$\langle v, \nabla h_1(x_0) \rangle = 0, \quad v - E\rho \in \partial f_{1,0}(x_0) + \sum_{i=1}^{m_1} \lambda_{1,i} \cdot \partial f_{1,i}(x_0),$$

(c) $\exists w: \Omega \rightarrow Y$ mesurable, $\lambda_{2,i}: \Omega \rightarrow \mathbb{R} (i \in \langle 1, m_2 \rangle)$ mesurables tels que:

$$p.s. \quad \langle \nabla h_2(y_0(\omega)), w(\omega) \rangle = 0,$$

$$(\rho(\omega), w(\omega)) \in \partial_{(x,y)} f_{2,0}(x_0, y_0(\omega); \omega) + \sum_{i=1}^{m_2} \lambda_{2,i}(\omega) \partial_{(x,y)} f_{2,i}(x_0, y_0(\omega); \omega)$$

$$\forall i \in \langle 1, m_2 \rangle, \quad \lambda_{2,i}(\omega) \geq 0, \quad \lambda_{2,i}(\omega) \cdot f_{2,i}(x_0, y_0(\omega); \omega) = 0.$$

Démonstration. Elle est immédiate en remarquant que, sous les hypothèses qui ont été faites:

$$\mathcal{T}(F_1; x_0) = \{\lambda \nabla h_1(x_0) | \lambda \geq 0\},$$

$$\mathcal{T}(F_2; y_0(\omega)) = \{\lambda \nabla h_2(y_0(\omega)) | \lambda \geq 0\}.$$

Il suffit alors de transcrire les conditions de Kuhn–Tucker écrites sous la forme générale.

L'apparition, dans les conditions de Kuhn–Tucker, des multiplicateurs $\{\lambda_{1,i}\}_{i \in \langle 1, m_1 \rangle}$ et $\{\lambda_{2,i}(\cdot)\}_{i \in \langle 1, m_2 \rangle}$ était attendue du fait de la forme des contraintes du premier et deuxième étage; ils sont interprétés usuellement comme des "prix d'équilibre" associés aux contraintes. L'apparition du multiplicateur aléatoire ρ (qui ne correspond pas à une contrainte spécifique) a été mise en évidence dans le cas convexe par R. T. Rockafellar et R. J. B. Wets [14]; son interprétation explicite en quelque sorte le fait que la décision x_0 est déterministe pure: elle doit être prise avant l'observation des éléments aléatoires du problème.

BIBLIOGRAPHIE

- [1] D. P. BERTSEKAS, *Stochastic optimization problems with non differentiable cost functionals*, J. Optimization Theory Appl., 12 (1973), pp. 218–231.
- [2] F. H. CLARKE, *Necessary conditions for nonsmooth problems in optimal control and the calculus of variations*. Ph.D. thesis, Dept. of Mathematics, University of Washington, Seattle, 1973.
- [3] ———, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.
- [4] ———, *A new approach to Lagrange multipliers*, Math. Operations Res., 2 (1976), pp. 165–174.
- [5] J. B. HIRIART-URRUTY, *About properties of the mean value functional and of the continuous infimal convolution in stochastic convex analysis*, Proc. 7th IFIP Conference, Lecture Notes in Computer Science, Springer-Verlag, New York, 1976.
- [6] ———, *Conditions nécessaires d'optimalité en programmation non différentiable*, Séminaire d'Analyse Numérique, Département de Mathématiques Appliquées, Université de Clermont, Aubière, 1975–1976.
- [7] ———, *Gradients généralisés de fonctions marginales*, this Journal, 15 (1977), pp. 301–316.
- [8] P. KALL, *Stochastic linear programming*, Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, New York, 1976.
- [9] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
- [10] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.
- [11] ———, *Integral functionals, normal integrands and measurable selections in nonlinear operators and the calculus of variations*, L. Waelbroeck, ed., Lecture Notes in Math., Springer-Verlag, New York, 1976.
- [12] R. T. ROCKAFELLAR AND R. J. B. WETS, *Stochastic convex programming: Basic duality*, Pacific J. Math., 62 (1976), pp. 173–195.
- [13] ———, *Stochastic convex programming: Relatively complete recourse and induced feasibility*, this Journal, 14 (1976), pp. 574–589.
- [14] ———, *Stochastic convex programming: Kuhn–Tucker conditions*, J. Mathematical Economics, 2 (1975), pp. 349–370.
- [15] ———, *Stochastic convex programming: Extended duality and singular multipliers*, Pacific J. Math., to appear.

- [16] ———, *Continuous versus measurable recourse in N -stage stochastic programming*, J. Math. Anal. Appl., 48 (1974), pp. 836–859.
- [17] L. THIBAUT, *Quelques propriétés des sous-différentiels de fonctions réelles localement lipschitziennes définies sur un espace de Banach séparable*, C.R. Acad. Sci. Paris, 282 (1976), pp. 507–510.
- [18] R. J. B. WETS, *Stochastic programs with fixed recourse: The equivalent deterministic program*, SIAM Rev., 6 (1974), pp. 309–339.
- [19] ———, *Duality relations in stochastic programming*, Symposia Mathematica, Academic Press, New York, 1975.
- [20] W. ZIEMBA, *Stochastic programs with simple recourse*, Mathematical Programming in Theory and Practice, P. Hammer and G. Zoutendijk, eds., North-Holland, Amsterdam, 1974.

OPTIMALITY CONDITIONS FOR THE AVERAGE COST PER UNIT TIME PROBLEM WITH A DIFFUSION MODEL*

H. J. KUSHNER†

Abstract. Defining the solution to a stochastic differential equation to be the solution to the martingale problem of Strook and Varadhan, we obtain results on the existence of an optimal stationary control for the average cost per unit time problem, a necessary and sufficient condition for optimality of a control, and a number of other related results.

1. Introduction. The purpose of the paper is the development of a necessary and sufficient "dynamic programming" like condition, for the average cost per unit time problem. The condition is similar to those developed for other problems by Davis and Varaiya [1] and Bismut [2]. In addition to its intrinsic interest, the criterion appears to be useful for the problem of approximation and computation (see the corollary and remark in § 6). The main Theorems are 3.1 (characterizing the invariant measure), 4.4 (existence of an optimal stationary control), 5.1 (characterizing the auxiliary $V^u(\cdot)$ function), and 6.1 (necessary and sufficient condition for optimality). Also, a number of auxiliary results are obtained.

We will use conditions (A1)–(A5).

(A1) Let $\sigma(\cdot)$ denote a bounded uniformly continuous and uniformly positive definite $r \times r$ matrix valued function on the Euclidean space R^r .

Let \mathcal{U} denote a compact convex set in some Euclidean space and which contains the origin.

(A2) $f(\cdot)$, $b(\cdot, \cdot)$, $k(\cdot, \cdot)$ are measurable R^r , R^r , and R^1 valued functions on R^r , $R^r \times \mathcal{U}$ and $R^r \times \mathcal{U}$, respectively; b and k are bounded, and are continuous in their second argument for each value of the first argument, and $b(x, 0) = 0$. $f(\cdot)$ is bounded on bounded sets.

(A3) The set $\{b(x, \alpha), k(x, \alpha), \alpha \in \mathcal{U}\} \equiv (b(x, \mathcal{U}), k(x, \mathcal{U}))$ is convex and compact for each $x \in R^r$.

Any measurable \mathcal{U} valued function $u(\cdot)$ on R^r is called an *admissible control*. Functions $b(\cdot, \cdot)$ and $k(\cdot, \cdot)$ are said to be *admissible* if they satisfy (A2), (A3) and have the form $b(x, u(x))$, $k(x, u(x))$ for admissible $u(\cdot)$. We will often write $b^u(\cdot) = b(\cdot, u(\cdot))$, $k^u(\cdot) = k(\cdot, u(\cdot))$. Our systems model is the stochastic differential equation

$$(1.1) \quad dx(t) = [f(x(t)) + b^u(x(t))] dt + \sigma(x(t)) dw(t), \quad x(0) = x,$$

where $w(\cdot)$ is a standard Wiener process. In particular, the process $x(\cdot)$ will be defined to be the solution to the martingale problem of Strook and Varadhan [3]; hence $w(\cdot)$ may be defined implicitly in terms of $x(\cdot)$. As pointed out by Bismut [2], there are a number of advantages to using the "martingale problem solution" definition of (1.1), particularly when questions of existence are of interest. Here, only feedback (Markov) controls are considered.

* Received by the editors April 25, 1977.

† Lefschetz Center for Dynamical Systems, Divisions of Applied Mathematics and Engineering, Brown University, Providence, Rhode Island 02912. This work was supported in part by the Air Force Office of Scientific Research under Grant AF-AFOSR 76-3063, by the National Science Foundation under Grant 73-03846-A01 and by the Office of Naval Research under Grant NONR N000 14-76-C-0279.

The cost functional is

$$(1.2) \quad \theta(u) = \lim_{T \rightarrow \infty} \frac{1}{T} E_x^u \int_0^T k(x(s), u(x(s))) ds = \int k(x, u(x)) \mu_u(dx),$$

where $\mu_u(\cdot)$ is the unique invariant measure for (1.1), which will exist under conditions to be imposed. Also, E_x^u denotes expectation under control $u(\cdot)$, and initial condition x .

In order for the problem to be well defined, we need some sort of recurrence for each control. In a sense, we will assume ((A4), (A5)) that the effects of $f(\cdot)$ dominate those of $b(\cdot, \cdot)$ for large $|x|$ and all $u(\cdot)$. Assumption (A4) will be convenient, and (A5), while avoidable, does provide a relatively simple method for obtaining some required estimates. Both are satisfied by a large number of problems.

(A4) *There is a nonnegative twice continuously differentiable real valued function $W_1(\cdot)$ such that $W_1(x) \rightarrow \infty$ as $|x| \rightarrow \infty$ and for some $\varepsilon > 0$ and compact K_1 ,*

$$(1.3) \quad \mathcal{L}^u W_1(x) \leq -\varepsilon, \quad x \notin K_1, \quad \text{all admissible } u(\cdot),$$

where

$$\mathcal{L}^u = \sum_{i,j} a_{ij}(x) \frac{\partial^2}{\partial x_i \partial x_j} + \sum_i [f_i(x) + b_i^u(x)] \frac{\partial}{\partial x_i},$$

the differential generator of (1.1), and $a(\cdot) = \sigma(\cdot)\sigma'(\cdot)/2$.

(A5) *Let $W_2(x) = W_1^2(x)$. There are constants $c_2 > 0$, $\alpha > 0$, such that for all admissible $u(\cdot)$,*

$$(1.4) \quad \mathcal{L}^u W_2(x) \leq c_2 - q_2(x), \quad \text{where } q_2(x) \geq 0,$$

and $q_2(x)/W_1(x) \geq \alpha > 0$. Let K_2 denote a compact set such that $q_2(x) \geq c_2$ for $x \notin K_2$.

Remark. Suppose that $f(x) = Ax$ and $\dot{x} = f(x)$ is asymptotically stable. Then we may use $x'Px = W_1(x)$, where P satisfies the Lyapunov equation $A'P + PA = -Q < 0$. Also, (A5) holds.

For some additional motivation, let us consider a dynamic programming approach. Suppose that there is a smooth function $V(\cdot)$ and a constant γ such that

$$(1.5) \quad \inf_{u(x) \in \mathcal{U}} [\mathcal{L}^u V(x) + k(x, u(x)) - \gamma] = 0$$

$$\equiv A(x) + \inf_{u(x) \in \mathcal{U}} [V'_x(x)b(x, u(x)) + k(x, u(x)) - \gamma], \quad \text{each } x.$$

If the solution to (1.1) is well defined for $u(x) = \bar{u}(x)$, the minimizer in (1.5), and if

$$E_x^{\bar{u}} V(x(t))/t \rightarrow 0 \quad \text{as } t \rightarrow \infty,$$

then

$$\lim_{T \rightarrow \infty} \frac{1}{T} E_x^{\bar{u}} \int_0^T k^u(x(s)) ds = \gamma.$$

If, in addition, $E_x^u V(x(t))/t \rightarrow 0$ as $t \rightarrow \infty$, then¹ $\gamma \leq \theta(u)$. If $P^u(x, t, \cdot) \rightarrow \mu_u(\cdot)$ ($P^u(x, t, \Gamma) \equiv P_x^u(x(t) \in \Gamma)$) strongly (in variation) as $t \rightarrow \infty$, then $\theta(u) = \int k(x, u(x)) \mu_u(dx)$.

In general, we do not know whether such a smooth $V(\cdot)$ or an optimal control exists. Part of our aim is to replace (1.5) by a local maximum-principle which does for

¹ See Kushner [4] for a formal discussion.

our problem what the work of Davis and Varaiya [1] or Bismut [2] did for the control problem on a finite time interval or for the discounted control problem.

A one dimensional version—on a finite interval with reflection—was treated by Mandl [5], and an incomplete development of the r -dimensional version of Mandl's result is given in [6].

In § 2, the solution to (1.1) is defined, and some properties listed. Sections 3, 4, 5, 6 deal with the existence of an invariant measure for each $u(\cdot)$, with certain continuity properties of the measure with respect to $b''(\cdot)$ and with the existence of an optimal control, with the existence and properties of an auxiliary $V(\cdot)$ function, and with the maximum principle, respectively.

2. The solution to (1.1). Let C , \mathcal{C}_t and \mathcal{C} denote the space of R^r valued continuous functions on $[0, \infty)$, and the σ -algebras induced by the coordinate projections $x(s)$, $s \leq t$, and $x(s)$, $s < \infty$, respectively, where $x(\cdot)$ is the generic element of C . Define $f^N(\cdot)$: $f^N(x) = f(x)$ if $|x| \leq N$, and is zero otherwise. Then, for each $x \in R^r$, there is a unique measure Q_x^N on (C, \mathcal{C}) such that $\{Q_x^N, x \in R^r\}$ solves the martingale problem of Strook and Varadhan [3]. There is also a standard Wiener process $W^{x,N}(\cdot)$ defined on (C, \mathcal{C}, Q_x^N) , and adapted to the (completed with respect to Q_x^N) $\{\mathcal{C}_t\}$ and such that

$$dx(t) = f^N(x(t)) dt + \sigma(x(t)) dW^{x,N}(t), \quad x(0) = x, \quad \text{w.p.1 } Q_x^N.$$

Define $S_M = \{x: |x| \leq M\}$. We will show that the solution is also well defined for $N = \infty$. The stability condition (A4) yields the following result. In the lemma, suppose (without loss of generality (w.l.o.g.)) that K_1 is in the interior of S_M .

LEMMA 2.1. Assume (A1), (A2), (A4). Let $x \in S_M$ and $N \geq M$. Define $\sigma_M = \inf \{t: x(t) \notin S_M - K_1\}$, $\sigma(K_1) = \inf \{t: x(t) \in K_1\}$. Then (P_x^N, E_x^N) correspond to Q_x^N

$$(2.1) \quad P_x^N\{x(t) \text{ hits } K_1 \text{ before hitting } \partial S_M\} \geq 1 - \frac{W_1(x)}{k_M},$$

where $k_M = \inf_{|x|=M} W_1(x)$,

$$(2.2) \quad E_x^N \sigma_M \leq W_1(x) / \varepsilon.$$

Proof. By Itô's lemma, for any $t < \infty$ (\mathcal{L}^0 corresponds to $u(\cdot) \equiv 0$)

$$\begin{aligned} E_x^N W_1(x(t \cap \sigma_M)) &= W_1(x) + E_x^N \int_0^{t \cap \sigma_M} \mathcal{L}^0 W_1(x(s)) ds \\ &\leq W_1(x) - \varepsilon E_x^N(t \cap \sigma_M). \end{aligned}$$

This inequality implies $E_x^N W_1(x(\sigma_M)) \leq W_1(x)$, $W_1(x) \geq \varepsilon E_x^N \sigma_M$, from which both (2.1) and (2.2) follow. Q.E.D.

Since $k_M \rightarrow \infty$ as $M \rightarrow \infty$, it can be shown that (2.1) implies that

$$(2.3) \quad \lim_{M \rightarrow \infty} \sup_N P_x^N \left\{ \sup_{0 \leq t \leq T} |x(t)| \geq M \right\} = 0, \quad \text{for each } x \text{ and } T.$$

By virtue of (2.3), there is a unique solution (P_x^0) to the martingale problem for coefficients (f, σ) , and each $x \in R^r$. Similarly, since neither the r.h.s (right hand side) of (1.3) nor K_1 depend on $u(\cdot)$, there is a unique solution (P_x^u) to the martingale problem

for all $x \in R'$ and coefficients $(f + b^u, \sigma)$ where $u(\cdot)$ is admissible. Furthermore,

$$(2.4a) \quad E_x^u \sigma(K_1) \leq W_1(x)/\varepsilon, \quad E_x^u \text{ corresponding to } P_x^u,$$

$$(2.4b) \quad P_x^u \left\{ \sup_{t \leq T} |x(t)| \geq N \right\} \rightarrow 0 \quad \text{as } N \rightarrow \infty, \text{ uniformly in } u(\cdot),$$

and in x in bounded sets,

$$(2.4c) \quad P_x^u \{x(t) \text{ hits } K_1 \text{ before hitting } \partial S_N\} \geq 1 - W_1(x)/k_N,$$

if $S_N \supset K_1$ and $x \in S_N$.

There is a Wiener process $W^{x,u}(\cdot)$, defined on (C, \mathcal{C}, P_x^u) and adapted to $\{\mathcal{C}_t\}$ (completed with respect to P_x^u) and such that

$$(2.5) \quad dx(t) = [f(x(t)) + b^u(x(t))] dt + \sigma(x(t)) dW^{x,u}(t), \quad \text{w.p.1 } (P_x^u).$$

For each real $0 < T < \infty$, define

$$\zeta_0^T(u) = \int_0^T [\sigma^{-1}(x(s))b^u(x(s))]^T dW^{x,0}(s) - \frac{1}{2} \int_0^T \|\sigma^{-1}(x(s))b^u(x(s))\|^2 ds.$$

Then (the proof of (2.6) is the same as that of Theorem 6.2 in [3], where $f(\cdot) \equiv 0$; see also Girsanov [7])

$$(2.6) \quad dP_x^u = \exp \zeta_0^T(u) \cdot dP_x^o \quad \text{on } (C, \mathcal{C}_T).$$

Thus, (at least on each (C, \mathcal{C}_T)), for each x all the measures P_x^u are mutually absolutely continuous, so that a.s. statements with respect to one are also a.s. statements with respect to the others. Also² (a.s. P_x^o)

$$(2.7) \quad dW^{x,0}(t) - \sigma^{-1}(x(t))b^u(x(t)) dt = dW^{x,u}(t).$$

(See Girsanov [7] or Davis and Varaiya [1]).

By [3], the solution to the martingale problem with coefficients (f^N, σ) or $(f^N + b^u, \sigma)$, for admissible $b^u(\cdot)$, is a strong Markov and a strong Feller process and in each of these cases the measures of $x(t)$ have densities with respect to Lebesgue measure for all $x = x(0)$ and all $t > 0$. Furthermore, these densities are positive almost everywhere. By the stability condition (A4) and (2.3), (2.4b), these facts are also true for the solution with coefficients $(f + b^u, \sigma)$ for admissible $b^u(\cdot)$. Define $P_x^u\{x(t) \in \Gamma\} \equiv P^u(x, t, \Gamma)$ and denote its density at y by $p^u(x, t, y)$.

The following lemma will be useful in the following sections.

LEMMA 2.2. Assume (A1), (A2) and (A4), (A5). Define

$$\sigma(K_2) = \inf \{t: x(t) \in K_2 \cup K_1\}.$$

Then

$$(2.8) \quad E_x^u \sigma^2(K_2) \leq \frac{2[W_2(x) + c_2 W_1(x)/\varepsilon]}{\varepsilon \alpha}, \quad x \notin K_1 \cup K_2.$$

Proof. By Itô's lemma and (A5)

$$0 \leq E_x^u W_2(x(t \wedge \sigma(K_2))) \leq W_2(x) + E_x^u \int_0^{t \wedge \sigma(K_2)} [c_2 - \alpha W_1(x(s))] ds,$$

² Whenever such differentials are equated, we mean to equate the corresponding integrals.

from which we get (use $E_x^u \sigma(K_2) < \infty$)

$$0 \leq W_2(x) + E_x^u \int_0^{\sigma(K_2)} [c_2 - \alpha W_1(x(s))] ds.$$

The last inequality, (2.4a) and $\sigma(K_2) \leq \sigma(K_1)$ imply that

$$0 \leq W_2(x) + c_2 E_x^u \sigma(K_1) - \alpha \varepsilon \int_0^\infty E_x^u I_{\{\sigma(K_2) \geq s\}} E_{x(s)}^u \sigma(K_2) ds.$$

The integrand equals

$$E_x^u I_{\{\sigma(K_2) \geq s\}} (\sigma(K_2) - s).$$

Hence, the integral equals $E_x^u \sigma^2(K_2)/2$, from which (2.8) follows. Q.E.D.

3. The invariant measure. Let G and G_1 be spheres in R^r , centered at the origin, with radii γ and γ_1 , respectively, $\gamma < \gamma_1$, and boundaries Γ and Γ_1 respectively, and with $G \supset K_1 \cup K_2$. Define $\tau' = \inf \{t: x(t) \in \Gamma_1\}$, $\tau_1 = \inf \{t: x(t) \in \Gamma\}$, $\tau'_1 = \inf \{t: t > \tau_1, x(t) \in \Gamma_1\}$, and define τ_n and τ'_n , $n > 1$, recursively by $\tau_n = \inf \{t: t > \tau'_{n-1}, x(t) \in \Gamma\}$, $\tau'_n = \inf \{t: t > \tau_n, x(t) \in \Gamma_1\}$. τ will be used for $\tau_2 - \tau_1 = \tau_2$, when $x \in \Gamma$. Define $\tilde{X}_n = x(\tau_n)$. Then, if $x \in \Gamma$, $\{\tilde{X}_n\}$ is a (homogeneous) Markov chain on the state space Γ , and Khazminskii [8] uses it to construct the invariant measure for $\{x(t)\}$. Let $\tau(A)$ denote the amount of time ($\int_0^\tau I_{\{x(t) \in A\}} dt$) that $x(t)$ spends in a Borel set A during $[0, \tau_2] = [0, \tau]$, when $x(0) = x \in \Gamma$ (if $x(0) \in \Gamma$, then $\tau_1 = 0$).

THEOREM 3.1. Assume (A1), (A2), (A4). Then there is a constant c_3 :

$$(3.1) \quad \sup_{\substack{x \in \Gamma \\ u}} E_x^u \tau \leq c_3 < \infty.$$

Both $\{\tilde{X}_n\}$ and $x(\cdot)$ have unique finite invariant measures (for each $u(\cdot)$) $\tilde{\mu}_u$ and μ_u , respectively, where for each Borel set A (note that $\mu_u(R^r) = 1$)

$$(3.2) \quad \mu_u(A) = \tilde{\mu}_u(A) / \tilde{\mu}_u(R^r), \quad \tilde{\mu}_u(A) = \int_\Gamma \tilde{\mu}_u(dx) E_x^u \tau(A).$$

The measure μ_u has a density (with respect to Lebesgue measure) which is positive almost everywhere and the value at the point y is given by

$$(3.3) \quad \int p^u(x, t, y) \mu_u(dx).$$

For any bounded Borel function $F(\cdot)$,

$$(3.4) \quad \int F(x) \tilde{\mu}_u(dx) = \int_\Gamma \tilde{\mu}_u(dx) E_x^u \int_0^\tau F(x(s)) ds.$$

Also

$$(3.5) \quad \sup_u E_x^u \tau_1 \leq W_1(x) / \varepsilon, \quad x \notin G.$$

For each Borel set A and bounded measurable function $F(\cdot)$,

$$(3.6) \quad P^u(x, t, A) \rightarrow \mu_u(A), \quad E_x^u F(x(t)) \rightarrow \int F(x) \mu_u(dx), \quad \text{as } t \rightarrow \infty.$$

Proof. Set $\tau'_0 = \inf \{t: x(t) \notin G_1\}$. To prove (3.1), we first show that, for fixed $t > 0$ and some real $c < 1$

$$(3.7) \quad \inf_{x,u} P_x^u \{\tau'_0 \leq t\} \geq 1 - c.$$

Inequality (3.7) follows from the fact that there is a $c < 1$ such that

$$\begin{aligned} \inf_{x,u} P_x^u \left\{ \sup_{s \leq t} |x(s)| \geq \gamma_1 \right\} &= \inf_{x,u} P_x^u \left\{ \sup_{s \leq t} \left| x + \int_0^s (f(x(s)) + b^u(x(s))) ds \right. \right. \\ &\quad \left. \left. + \int_0^s \sigma(x(s)) dW^{x,u}(s) \right| \geq \gamma_1 \right\} \\ &\geq \inf_{u, x \in G_1} P_x^u \left\{ \sup_{s \leq t} \left| \int_0^s \sigma(x(s)) dW^{x,u}(s) \right| \geq \gamma_1 + |x| + Kt \right\} \\ &\geq 1 - c, \end{aligned}$$

where K is a bound on $|f + b^u|$ in G_1 . Now

$$\begin{aligned} P_x^u \{\tau'_0 > nt\} &= E_x^u I_{\{\tau'_0 > (n-1)t\}} I_{\{\tau'_0 > nt\}} \\ &= E_x^u I_{\{\tau'_0 > (n-1)t\}} E_{x(nt-t)}^u I_{\{\tau'_0 > t\}} \leq E_x^u I_{\{\tau'_0 > (n-1)t\}} c \\ &\leq \dots \leq c^n \end{aligned}$$

which implies that $\sum_n nt c^{n-1} = c_4$ is an upper bound to $E_x^u \tau'_0$. Hence, $E_x^u \tau' \leq c_4$ for $x \in \Gamma$. Indeed, (to be used later)

$$(3.8) \quad E_x^u (\tau')^\alpha \leq \sum_n (nt)^\alpha c^{n-1} < \infty, \quad x \in G.$$

Inequality (3.5) follows from (2.4a), since $G \supset K_1 \cup K_2$. Thus, for $x \in \Gamma$,

$$\begin{aligned} E_x^u \tau &= E_x^u \int_0^{\tau'} ds + E_x^u \int_{\tau'}^{\tau} ds \\ &= E_x^u \tau' + E_x^u E_{x(\tau')}^u \tau_1 \leq c_4 + E_x^u W_1(x(\tau'))/\varepsilon \leq c_3 \end{aligned}$$

for some real c_3 , which gives (3.1).

In [8], Khazminskii proves that there is a unique invariant measure $\tilde{\mu}_u$ under the conditions (i): that $P^u(x, t, A) > 0$, all open A , all x and all $t > 0$, and (ii): that $x(\cdot)$ be recurrent (Khazminskii's definition of recurrence is implied by (3.5)) and (iii): that $x(\cdot)$ is a strong Feller and a strong Markov process. Under the additional condition (3.1) (for fixed $u(\cdot)$) there is a unique finite invariant measure μ_u given by (3.2) ([8, Thms. 2.1, 3.2 and 3.3]). Equations (3.3) and (3.4) follow by simple calculations.

Since μ_u and $P^u(x, t, \cdot)$ all have densities which are positive almost everywhere (all x , and all $t > 0$), they are mutually absolutely continuous. Condition (3.6) then follows from [8, Thm. 3.4] or Doob [9, Thm. 5] (let his Φ equal our μ_u and his $P_1 = 0$), since μ_u and $P^u(x, t, \cdot)$ are mutually absolutely continuous for $t > 0$. Q.E.D.

4. Existence of an optimal control. Theorems 4.1 and 4.2 provide some preliminary results. Theorem 4.3 proves that, in a sense, the invariant measure is continuous in the function $b^u(\cdot)$. This leads directly to the existence Theorem 4.4. Define $g(\cdot, \cdot) = (b(\cdot, \cdot), k(\cdot, \cdot))$.

THEOREM 4.1. Assume (A1)–(A3). Let $\{u^n(\cdot)\}$ denote a sequence of admissible controls, and write $g^n(\cdot) = (b^{u^n}(\cdot), k^{u^n}(\cdot))$. If there is a bounded measurable function

$\bar{g}(\cdot)$ such that

$$\int_A g^n(x) dx \rightarrow \int_A \bar{g}(x) dx, \quad \text{all Borel } A,$$

then $\bar{g}(\cdot)$ is admissible in the sense that there is an admissible $u(\cdot)$ such that $\bar{g}(\cdot) = (b^u(\cdot), k^u(\cdot))$, for almost all x .

Proof. The theorem is a standard existence theorem. See Roxin [10] or McShane and Warfield [11]. By an argument such as that used by Roxin [10], $\bar{g}(x) \in \{\gamma: \gamma = g(x, \alpha), \text{ for some } \alpha \in \mathcal{U}\} \equiv g(x, \mathcal{U})$ for almost all x . We can assume that the conclusion holds for all x . Then the theorem follows by the implicit function theorem in [11]. Q.E.D.

A family $\{\phi_\alpha\}$ of measures on R^r is said to be *tight* if for each $\varepsilon > 0$, there is a compact K_ε such that $\phi_\alpha(R^r - K_\varepsilon) \leq \varepsilon$, all α . Let $l(\cdot)$ denote *Lebesgue measure*.

THEOREM 4.2. Assume (A1), (A2), (A4), (A5). Then $\{\mu_u, u(\cdot) \text{ admissible}\}$ is tight. Also, $\mu_u(A) \rightarrow 0$ as $l(A) \rightarrow 0$, uniformly in $u(\cdot), A$.

Remark. The theorem is true, but harder to prove, without (A5). Since (A5) will be used later anyway, we use it now to simplify the proof.

Proof. Since $\bar{\mu}_u(R^r)$ is bounded uniformly in $u(\cdot)$ (by virtue of (3.1)) to show tightness, we only need to show that (refer to the definition of μ_u in (3.2))

$$\sup_{\substack{x \in \Gamma \\ u}} E_x^u \tau(S'_N) \rightarrow 0 \quad \text{as } N \rightarrow \infty,$$

where $S'_N = \{y: |y| \geq N\}$. Recall that $G_1 \supset K_1 \cup K_2$, and assume that $\{y: |y| \leq N\} \supset G_1$. (See (A4), (A5) for the definition of K_1, K_2 .) Then, for $x \in \Gamma$,

$$\begin{aligned} (E_x^u \tau(S'_N))^2 &= [E_x^u \tau(S'_N) I_{\{\tau(S'_N) > 0\}}]^2 \\ &\leq \sup_{\substack{x \in \Gamma_1 \\ u}} E_x^u \tau_1^2 \cdot \sup_{\substack{x \in \Gamma \\ u}} P_x^u \{\tau(S'_N) > 0\} \equiv M_1 M_2^N. \end{aligned}$$

By (2.4c), $M_2^N \leq \sup_{x \in \Gamma_1} W_1(x)/k_N$ and by Lemma 2.2, $M_1 < \infty$. The first assertion of the theorem now follows, since $k_N \rightarrow \infty$ as $N \rightarrow \infty$.

Fix ε and (by tightness) choose compact K_ε such that $\mu_u(K_\varepsilon) \geq 1 - \varepsilon$, all $u(\cdot)$. Then, for $t > 0$,

$$\begin{aligned} \mu_u(A) &= \int_{R^r - K_\varepsilon} \mu_u(dx) P^u(x, t, A) + \int_{K_\varepsilon} \mu_u(dx) P^u(x, t, A) \\ &\leq \varepsilon + \sup_{x \in K_\varepsilon} P^u(x, t, A). \end{aligned}$$

Also, by (2.6) (recall that the superscript o corresponds to $u = b = 0$),

$$\begin{aligned} [P^u(x, t, A)]^2 &= [E_x^o I_{\{x(t) \in A\}} \exp \zeta_0^t(u)]^2 \\ &\leq E_x^o I_{\{x(t) \in A\}} E_x^o \exp 2\zeta_0^t(u) \leq \text{const.} \cdot P^o(x, t, A). \end{aligned}$$

The last assertion of the theorem follows from the last two inequalities since ε is arbitrary and $P^o(x, t, A) \rightarrow 0$ uniformly in $x \in K_\varepsilon$ and in A , as $l(A) \rightarrow 0$. Q.E.D.

THEOREM 4.3. Assume (A1)–(A5). Let $\bar{b}(\cdot)$ be a bounded measurable function, and $\{b^{u_n}(\cdot)\}$ admissible, such that (write sub- or superscript n for u_n)

$$(4.1) \quad \int_A b^n(x) dx \rightarrow \int_A \bar{b}(x) dx, \quad \text{all Borel } A.$$

Then (Theorem 4.1) there is an admissible $u(\cdot)$ such that $\bar{b}(\cdot) = b^u(\cdot)$ a.e. Also,

$$(4.2) \quad \exp \zeta_0^t(u_n) \rightarrow \exp \zeta_0^t(u)$$

weakly in L_1 (with respect to P_x^0) as $n \rightarrow \infty$, for each x and $t > 0$. In particular,

$$(4.3) \quad P^n(x, t, A) \rightarrow P^u(x, t, A),$$

$$(4.4) \quad E_x^n F(x(t)) \rightarrow E_x^u F(x(t)), \quad \text{each } x, t > 0, \text{ Borel } A, \text{ bounded measurable } F(\cdot),$$

$$(4.5) \quad \int \mu_n(dx) F(x) \rightarrow \int \mu_u(dx) F(x), \quad F(\cdot) \text{ bounded and measurable.}$$

Proof. The convergence (4.2) is proved by³ Bismut [2, Thm. IV-3], and (4.3), (4.4) follow from that convergence and (2.6).

Since $\{\mu_n\}$ is tight, it is weakly sequentially compact, (Billingsley [12], p. 37)]. That is, each subsequence contains a further subsequence $\{\mu_{n_i}\}$ such that, for some probability measure $\hat{\mu}$,

$$\int F(x) \mu_{n_i}(dx) \rightarrow \int F(x) \hat{\mu}(dx), \quad \text{all bounded continuous } F(\cdot)$$

(Billingsley [12, pp. 35–37]). Let n index such a convergent subsequence, with (weak) limit $\hat{\mu}$.

Let $F(\cdot)$ be bounded and continuous. Let $\varepsilon > 0$ and define K_ε as in Theorem 4.2 and write

$$\begin{aligned} \int \mu_n(dx) F(x) &= \int \mu_n(dx) E_x^n F(x(t)) \\ &= \int_{K_\varepsilon} \mu_n(dx) E_x^n F(x(t)) + \int_{R' - K_\varepsilon} \mu_n(dx) E_x^n F(x(t)). \end{aligned}$$

The second term is $\leq \varepsilon \sup_x |F(x)|$. Since $x(\cdot)$ is a Feller process under each control, $E_x^v F(x(t))$ is continuous in x , for each $v(\cdot)$. Then, the function $E_x^n F(x(t))$ is continuous in x and converges to the continuous function $E_x^u F(x(t))$, by (4.4). This convergence implies that, for each $\delta \geq 0$, there is a Borel set $A_\delta \subset K_\varepsilon$, with $l(A_\delta) \leq \delta$, such that $|E_x^n F(x(t)) - E_x^u F(x(t))| \leq \varepsilon$ for large n , and $x \notin A_\delta$, $x \in K_\varepsilon$. Those estimates, Theorem 4.2, the arbitrariness of δ and the weak convergence imply that the first term on the right goes to $\int_{K_\varepsilon} \hat{\mu}(dx) E_x^u F(x(t))$, as $n \rightarrow \infty$. Since the l.h.s. (left hand side) converges to $\int \hat{\mu}(dx) F(x)$, and $\varepsilon > 0$ is arbitrary, we conclude that

$$\int \hat{\mu}(dx) F(x) = \int \hat{\mu}(dx) E_x^u F(x(t)).$$

This equation together with the arbitrariness of $t > 0$ and $F(\cdot)$, implies that $\hat{\mu}$ is an invariant measure—under control $u(\cdot)$. Thus, the uniqueness Theorem 3.1 implies that $\hat{\mu} = \mu_u$. Since the result does not depend on the selected subsequence, we have that $\mu_n \rightarrow \mu_u$ weakly, as $n \rightarrow \infty$, and (4.5) holds for bounded and continuous $F(\cdot)$.

Let $F(\cdot)$ be bounded and measurable. Then, for $t > 0$, the invariance of μ_n implies

$$\int \mu_n(dx) F(x) = \int \mu_n(dx) F_t^n(x), \quad F_t^n(x) = E_x^n F(x(t)).$$

³ In [2], $f = 0$, but the proof is exactly the same.

By the strong Feller property, $E_x^u(F(x(t)))$ is continuous in x . Now, as in the proof of (4.5) for continuous $F(\cdot)$, the almost uniform convergence of $F_t^n(x)$ to $E_x^u F(x(t))$, Theorem 4.2 and the tightness and weak convergence of $\{\mu_n\}$ imply

$$\int \mu_n(dx) F_t^n(x) \rightarrow \int \mu_u(dx) E_x^u F(x(t)),$$

which must also equal the limit of $\int \mu_n(dx) F(x)$. This implies (4.5), since by the invariance of μ_u , the above r.h.s. equals $\int \mu_u(dx) F(x)$. Q.E.D.

THEOREM 4.4. Assume (A1)–(A5). Then there is an optimal admissible control.

Proof. Let $\{u_n(\cdot)\}$ denote a minimizing sequence. Then $\theta \equiv \lim \theta(u_n) = \inf_{u(\cdot)} \theta(u)$. Let u also index a weak* ($\sigma(L_\infty, L_1)$ topology) convergent subsequence of $\{b^n(\cdot), k^n(\cdot)\}$ with limit $(\bar{b}(\cdot), \bar{k}(\cdot))$, where we let n replace the index u_n . There is an admissible control $u(\cdot)$ such that (Theorem 4.1) $(\bar{b}(\cdot), \bar{k}(\cdot)) = (b^u(\cdot), k^u(\cdot))$. If $k(\cdot)$ does not depend on the control, then $\theta(u_n) \rightarrow \theta(u)$ by (4.5). Hence, in this case there is an optimal control.

Now let $k(\cdot)$ depend on the control. Let $F_n(\cdot)$ be a sequence of bounded measurable functions which converges to a function $F(\cdot)$ in the weak* topology. Then⁴ (Bismut [2, Proposition IV-4, p. 48])

$$\int_0^t F_n(x(s)) ds \rightarrow \int_0^t F(x(s)) ds$$

in probability (P_x° , each x), as $n \rightarrow \infty$. Note that $E_x^\circ \exp 2\zeta_0'(v)$ is bounded uniformly in x and in the control $v(\cdot)$. Let $F_n(\cdot)$ be defined by $F_n(x) = k(x, u_n(x))$, and set $F(x) = k(x, u(x))$. By the convergence in probability, the convergence (4.2) and the boundedness of $E_x^\circ \exp 2\zeta_0'(u_n)$ uniformly in n, x , we have

$$\begin{aligned} E_x^{u_n} \int_0^t F_n(x(s)) ds &= E_x^\circ \exp \zeta_0'(u_n) \int_0^t F_n(x(s)) ds \\ (4.6) \quad &\rightarrow E_x^\circ \exp \zeta_0'(u) \int_0^t F(x(s)) ds = E_x^u \int_0^t F(x(s)) ds \\ &\quad \text{(a continuous function of } x \text{).} \end{aligned}$$

Integrating the left and right sides, respectively, of (4.6) with respect to μ_n and μ_u , respectively, and using the invariance of these measures yields the two equations

$$(4.7a) \quad t\theta(u_n) = \int \mu_n(dx) \int_0^t E_x^{u_n} k(x(s), u_n(x(s))) ds,$$

$$(4.7b) \quad t\theta(u) = \int \mu_u(dx) \int_0^t E_x^u k(x(s), u(x(s))) ds.$$

Now, (4.6) implies that the right hand integral in (4.7a) converges to that in (4.7b) for each x . This, together with the tightness of $\{\mu_n\}$, the last part of Theorem 4.2, and an argument like that used in the proof of Theorem 4.3 to show (4.5) yields that

$$\theta(u_n) \rightarrow \theta(u). \quad \text{Q.E.D.}$$

5. The auxiliary $V^u(\cdot)$ function. Our aim is to get a replacement for the $V(\cdot)$ function in (1.5), which will play an important role in the sequel.

⁴ In [2], $f = 0$, but the proof is exactly the same for our case.

In this section the control $u(\cdot)$ is fixed, and we return to the Markov chain $\{\tilde{X}_n\}$ of § 3. For a measurable set $\gamma \in \Gamma$, let $\pi_u(x, \gamma) = P_x^u\{\tilde{X}_1 \in \gamma\}$, $x \in \Gamma$, and recall that the unique invariant measure for $\{\tilde{X}_n\}$ is denoted by $\tilde{\mu}_u$. Let ϕ be a finite measure on Γ . The chain $\{\tilde{X}_n\}$ is said to be *uniformly ϕ recurrent* if for each measurable $\gamma \in \Gamma$ such that $\phi(\gamma) > 0$,

$$(5.1) \quad \sum_{m=1}^n P_x^u\{\tilde{X}_m \in \gamma, \tilde{X}_i \notin \gamma, i < m\} \rightarrow 1 \quad \text{uniformly in } x \in \Gamma, \quad \text{as } n \rightarrow \infty$$

(Orey [13, p. 26]). A sufficient condition for (5.1) is (Orey [13, p. 29]) that if $\phi(\gamma) > 0$ then there is an $n < \infty$ and $\varepsilon > 0$ (perhaps depending on γ) such that

$$(5.2) \quad \sum_{m=1}^n P_x^u\{\tilde{X}_m \in \gamma, \tilde{X}_i \notin \gamma, i < m\} \geq \varepsilon$$

for all $x \in \Gamma$. If the chain is uniformly ϕ recurrent and a-periodic then there are constants C and $\rho \in (0, 1)$ such that

$$(5.3) \quad |P_x^u\{\tilde{X}_n \in \gamma\} - \tilde{\mu}_u(\gamma)| \leq C\rho^n$$

uniformly in γ and $x \in \Gamma$ (a consequence of equation (6.2) in [13, p. 26], and the invariance of $\tilde{\mu}_u$). Thus, the n -step transition probability $\pi_u^{(n)}(x, \cdot)$ converges to $\tilde{\mu}_u$ in variation, at an exponential rate.

Define, for $x \in R'$ (see § 3 for the definition of τ_i , τ and $\bar{\mu}_u$),

$$(5.4) \quad \begin{aligned} \tilde{V}^u(x) &= E_x^u \int_0^{\tau_1} [k^u(x(s)) - \theta(u)] ds \\ &+ \lim_n \sum_{m=1}^n [E_x^u \int_{\tau_m}^{\tau_{m+1}} k^u(x(s)) ds - \int \bar{\mu}_u(dx) k^u(x)], \end{aligned}$$

$$(5.5) \quad \begin{aligned} V^u(x) &= E_x^u \int_0^{\tau_1} \tilde{k}^u(x(s)) ds + \lim_n E_x^u \int_{\tau_1}^{\tau_n} \tilde{k}^u(x(s)) ds, \\ &\text{where } \tilde{k}^u(x) \equiv k(x, u(x)) - \theta(u). \end{aligned}$$

THEOREM 5.1. Assume (A1), (A2), (A4), (A5). Then $\tilde{V}^u(x)$ and $V^u(x)$ are well defined. There are constants C_0, C_1 such that

$$(5.6) \quad \begin{aligned} |\tilde{V}^u(x)| &\leq C_0 + C_1 E_x^u \tau_1, \\ |V^u(x)| &\leq C_0 + C_1 E_x^u \tau_1. \end{aligned}$$

The tail of (5.5) $(E_x^u \int_{\tau_n}^{\tau_m} \cdot)$ goes to zero as $n, m \rightarrow \infty$, uniformly in x , and $E_x^u \int_0^{\tau_n} \tilde{k}^u(x(s)) ds$ is bounded uniformly in n and $x \in \Gamma$.

Proof. Let $\phi = l =$ Lebesgue measure on Γ . $\pi_u(x, \gamma) > 0$ if γ is open in Γ . Then $\pi_u(x, \gamma) > 0$ if $l(\gamma) > 0$. Since $\pi_u(\cdot, \gamma)$ is continuous (by the strong Feller property—it also follows from the assertion 6° of Khazminskii in [8], with a suitable definition of U, Γ there), $\inf_{x \in \Gamma} \pi_u(x, \gamma) > 0$. Thus, by the criterion (5.2), with $n = 1$, $\{\tilde{X}_n\}$ is uniformly l -recurrent.

Let $\hat{F}(\cdot)$ be a bounded measurable function on Γ . By virtue of (5.3),

$$(5.7) \quad \sum_n \left| E_x^u \hat{F}(\tilde{X}_n) - \int_{\Gamma} \tilde{\mu}_u(dx) \hat{F}(x) \right| \leq \text{const.}, \quad x \in \Gamma.$$

Let $\hat{F}(x) = E_x^u \int_0^\tau k^u(x(s)) ds$, which is bounded on Γ by (3.1). Note that

$$E_x^u \hat{F}(\tilde{X}_n) = E_x^u E_{\tilde{X}_n}^u \int_0^\tau k^u(x(s)) ds = E_x^u \int_{\tau_n}^{\tau_{n+1}} k^u(x(s)) ds.$$

Then, by using (3.4) and (5.7), we get

$$(5.8) \quad \sum_{n=1}^{\infty} \left| E_x^u \int_{\tau_n}^{\tau_{n+1}} k^u(x(s)) ds - \int \mu_u(dx) k^u(x) \bar{\mu}_u(R') \right| \leq \text{const.}, \quad x \in \Gamma.$$

This, together with $E_x^u \int_0^\tau |k^u(x(s))| ds \leq \sup_x |k^u(x)| E_x^u \tau_1$, implies both that $\tilde{V}^u(\cdot)$ is well defined and also that the first line of (5.6) holds.

Now, redo the above argument with $\hat{F}(x) = E_x^u \tau$, $x \in \Gamma$. Then

$$(5.9) \quad \sum_{n=0}^{\infty} \left| E_x^u E_{\tilde{X}_n}^u \tau - \int \tilde{\mu}_u(dx) E_x^u \tau \right| \leq \text{const.}, \quad x \in \Gamma.$$

Since⁵ $\int \tilde{\mu}_u(dx) E_x^u \tau = \bar{\mu}_u(R')$, the convergence in (5.8) and (5.9) allows us to replace $\bar{\mu}_u(R')$ in (5.8) by $E_x^u E_{\tilde{X}_n}^u \tau = E_x^u \int_{\tau_n}^{\tau_{n+1}} ds$, and still to get convergence. From this, we get both that $V^u(\cdot)$ is well defined, and that the last bound in (5.6) holds. The last two assertions of the theorem follow from (5.8) and (5.9). Q.E.D.

The next lemma gives some useful estimates. The constant C may have different values in each usage.

LEMMA 5.1. Assume (A1), (A2), (A4), (A5). Then, for all admissible $u(\cdot)$, $v(\cdot)$ and all $x, s \geq 0$, $t \geq 0$, and Markov times $\alpha \leq t$.

$$(5.10) \quad E_x^v |V^u(x(t))| \leq C[1 + W_1(x) + t],$$

$$(5.11a) \quad E_{x(t)}^v |V^u(x(t+s))| \leq C[1 + W_1(x(t)) + s],$$

$$(5.11b) \quad E_{x(t)}^v |V^u(x((s+t) \cap \rho))| \leq C[1 + W_1(x(t)) + s], \quad w.p.1$$

for any Markov time $\rho \geq t$,

$$(5.11c) \quad E_{x(t)}^v W_i(x(t+s)) \leq C[1 + W_i(x(t)) + s] \quad w.p.1,$$

$$(5.12) \quad E_x^v |V^u(x(\beta))|^2 \leq C[1 + t + W_2(x)], \quad \text{each } u(\cdot), v(\cdot), \quad \beta = \alpha \cap t,$$

$$(5.13) \quad \int \mu_v(dx) |V^u(x)| < \infty, \quad \int \mu_v(dx) |W_1(x)| < \infty, \quad \text{each } u, v.$$

Proof. By (5.6),

$$E_x^v |V^u(x(t))| \leq \text{const.} [1 + E_x^v E_{x(t)}^u \tau_1].$$

By (A4), there is a constant ε_1 such that $\mathcal{L}^v W_1(x) \leq \varepsilon_1$, all x and $v(\cdot)$. Thus, by an application of Itô's lemma,

$$E_x^v W_1(x(t)) \leq W_1(x) + \varepsilon_1 t.$$

The last two inequalities and the bounds (3.5) and (3.8) imply (5.10). Inequality (5.11) follows by similar calculations.

We prove (5.12) only for $\alpha = t$. The general proof is similar. Write

$$V^u(x(t)) \leq C[1 + E_{x(t)}^u \tau_1] \leq C[1 + W_1(x(t))].$$

Continuing, and using (5.11c) and (A5), we get

$$E_x^v |V^u(x(t))|^2 \leq C E_x^v [1 + W_2(x(t))] \leq C[1 + W_2(x) + t].$$

⁵ See (3.4), with $F(x) \equiv 1$.

To prove (5.13), define $\sigma_N = \inf \{t: |x(t)| \geq N\}$, and for each integer M define $W_1^M(x) = \min [W_1(x), M]$, and note that (by (A5) and Itô's lemma)

$$E_x^v W_2(x(t \cap \sigma_N)) \leq W_2(x) + E_x^v \int_0^{t \cap \sigma_N} [c_2 - q_2(x(s))] ds.$$

Thus, by bounding $q_2(\cdot)$ and letting $N \rightarrow \infty$, we obtain

$$0 \leq W_2(x) + E_x^v \int_0^t [c_2 - \alpha W_1^M(x(s))] ds.$$

Divide the last equation by t , let $t \rightarrow \infty$, and get (using (3.6), the convergence of $P^v(x, t, \cdot)$ to the invariant measure $\mu_v(\cdot)$)

$$c_2 \geq \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \alpha E_x^v W_1^M(x(s)) ds = \alpha \int \mu_v(dx) W_1^M(x).$$

Hence, $c_2 \geq \alpha \int \mu_v(dx) W_1(x)$. This, together with $V^u(x) \leq C \cdot [1 + W_1(x)]$, implies (5.13). Q.E.D.

Theorem 5.2 will be used to obtain the optimality criterion in Theorem 6.1.

THEOREM 5.2. Assume (A1), (A2) and (A4), (A5). Then $V^u(\cdot)$ is continuous and the process M_t^u given by

$$M_t^u = V^u(x(t)) - V^u(x) + \int_0^t \tilde{k}^u(x(s)) ds$$

is a continuous square integrable martingale, adapted to $\{\mathcal{G}_t\}$ and under P_x^u , each x .

Proof. First we note several facts. $V^u(\cdot)$ does not depend on G_1 . If $\gamma < \gamma' < \gamma_1$, and G'_1 is a sphere with radius γ' , then several cycles ($\Gamma \rightarrow \Gamma'$, etc.) of the process for the G'_1 case may be included in one cycle for the G_1 case, but the values of $V^u(\cdot)$ are the same. Also, as $\gamma_1 \downarrow \gamma$, $\sup_{x \in \Gamma} E_x^u \tau \rightarrow 0$. Note also that

$$\inf_x P_x^u \{\tau_n \geq T\} \rightarrow 1, \text{ as } n \rightarrow \infty, \text{ each } T < \infty.$$

Let n_T denote the largest integer i such that $\tau_i \leq T$. Write $\tilde{k}^u(x(s))$ as $\hat{k}(s)$. It will be shown that

$$(5.14) \quad \lim_{T \rightarrow \infty} \lim_{m, n \rightarrow \infty} E_x^u \int_{T \cap \tau_n}^{\tau_m} \hat{k}(s) ds = 0, \text{ uniformly in bounded } x \text{ sets.}$$

Equation (5.14) implies that

$$V^u(x) = E_x^u \int_0^T \hat{k}(s) ds + \varepsilon(T, x),$$

where $\varepsilon(T, x) \rightarrow 0$ as $T \rightarrow \infty$, uniformly in bounded x sets. This and the strong Feller property imply that $V^u(\cdot)$ is continuous.

Now, for some constants C_i , and any integer Q ,

$$\begin{aligned} \lim_{m, n} E_x^u \left| \int_{T \cap \tau_n}^{\tau_m} \hat{k}(s) ds \right| &\leq E_x^u \left| \int_T^{\tau_{N_T+1}} \hat{k}(s) ds \right| + E_x^u \sum_{i > n_T} \left| \int_{\tau_i}^{\tau_{i+1}} \hat{k}(s) ds \right| \\ &\leq C_1 E_x^u |\tau_{N_T+1} - \tau_{N_T}| + \sum_{i \geq Q} E_x^u \left| \int_{\tau_i}^{\tau_{i+1}} \hat{k}(s) ds \right| \\ &\quad + C_2 E_x^u \tau I_{\{Q \geq N_T\}}. \end{aligned}$$

The first and third terms on the right can be made arbitrarily small (uniformly in bounded x sets) by selecting large T and small $(\gamma_1 - \gamma)$, and large T , respectively. The central term can be made small, uniformly in x , by choosing Q large. This implies (5.14).

It can be shown that

$$E_{x(t)}^u[V^u(x(t+s)) - V^u(x(t)) + \int_t^{s+t} \tilde{k}^u(x(v)) dv] = 0, \quad \text{all } s \geq 0, t \geq 0,$$

where the conditional expectation above exists by Lemma 5.1. The martingale property follows from this. The continuity and square integrability follow from the continuity of $V^u(\cdot)$ and (5.12), respectively. Q.E.D.

6. The maximum principle. Let $u(\cdot)$ be admissible. It will be seen in Theorem 6.2 that there is a Borel function (an r -row vector) $\psi^u(\cdot)$ such that for each $x \in R^r$,

$$(6.1) \quad \begin{aligned} M_t^u &= \int_0^t \psi^u(x(s)) \sigma(x(s)) dW^{x,u}(s) \quad \text{w.p.1 } P_x^u, \\ E_x^u \int_0^t |\psi^u(x(s))|^2 ds &< \infty. \end{aligned}$$

Let $v(\cdot)$ be admissible. Then $x(\cdot)$ satisfies (w.p.1 P_x^v)

$$dx(t) = [f(x(t)) + b^v(x(t))] dt + \sigma(x(t)) dW^{x,v}.$$

By (2.7), we can suppose that (w.p.1 P_x^v)

$$(6.2) \quad \begin{aligned} dW^{x,v}(t) &= dW^{x,0}(t) - \sigma^{-1}(x(t)) b^v(x(t)) dt \\ &= dW^{x,u}(t) + \sigma^{-1}(x(t)) (b^u(x(t)) - b^v(x(t))) dt. \end{aligned}$$

Since P_x^0, P_x^u and P_x^v are mutually absolutely continuous, all a.s. statements with respect to one also a.s. statements with respect to the others.

Theorem 6.1 is the "maximum" or "Hamilton-Jacobi" principle, a natural development for our problem, of some of the ideas in [1] and [2].

THEOREM 6.1. Assume (A1), (A2), (A4), (A5), and let $u(\cdot), v(\cdot)$ be admissible. If

$$(6.3) \quad e^{u,v}(x) \equiv (k^u(x) - k^v(x)) + \psi^u(x)(b^u(x) - b^v(x)) > 0$$

on a set A of positive Lebesgue measure, then there is an admissible control $\bar{v}(\cdot)$ such that $\theta(\bar{v}) < \theta(u)$. The condition $e^{u,v}(x) \leq 0$ a.e. for each admissible $v(\cdot)$ is necessary and sufficient for $u(\cdot)$ to be optimal.

Proof. First, we derive the basic formula (6.5). Using (6.1), (6.2) and the definition of M_t^u yields (a.s. P_x^u)

$$\begin{aligned} 0 &= V^u(x(t)) - V^u(x) + \int_0^t \tilde{k}^u(x(s)) ds \\ &\quad - \int_0^t \psi^u(x(s)) \sigma(x(s)) [dW^{x,v}(s) - \sigma^{-1}(x(s)) (b^u(x(s)) - b^v(x(s))) ds]. \end{aligned}$$

Define

$$\sigma_N = \min \left\{ t: \int_0^t |\psi^u(x(s)) \sigma(x(s))|^2 ds = N \right\}.$$

Then

$$(6.4) \quad \begin{aligned} 0 = & E_x^v V^u(x(t \cap \sigma_N)) - V^u(x) + E_x^v \int_0^{t \cap \sigma_N} \tilde{k}^u(x(s)) ds \\ & + E_x^v \int_0^{t \cap \sigma_N} \psi^u(x(s)) [b^u(x(s)) - b^v(x(s))] ds \end{aligned}$$

where the expectations exist by Lemma 5.1.

By the uniform integrability implied by (5.12), the first term on the r.h.s. of (6.4) tends to $E_x^v V^u(x(t))$ as $N \rightarrow \infty$. Also, by (2.6) (use $W^{x,u}$ in $\zeta_0^t(v-u)$ not $W^{x,0}$),

$$\begin{aligned} E_x^v \int_{t \cap \sigma_N}^t |\psi^u(x(s))| ds &= E_x^u \exp \zeta_0^t(v-u) \cdot \int_{t \cap \sigma_N}^t |\psi^u(x(s))| ds \\ &\leq [E_x^u \exp 2\zeta_0^t(v-u)]^{1/2} \left[E_x^u \left(\int_{t \cap \sigma_N}^t |\psi^u(x(s))|^2 ds \right) \right]^{1/2} \\ &\equiv A_1 A_{2N}. \end{aligned}$$

$A_1 < \infty$, and $A_{2N} \rightarrow 0$ as $N \rightarrow \infty$ by (6.1). Thus, we can replace $t \cap \sigma_N$ by t throughout (6.4). Setting $\sigma_N = 0$ in the above equation yields

$$E_x^v \int_0^t |\psi^u(x(s))| ds \leq \text{const.} \left[E_x^u \int_0^t |\psi^u(x(s))|^2 ds \right]^{1/2}.$$

But

$$\begin{aligned} E_x^u \int_0^t |\psi^u(x(s))|^2 ds &\leq \text{const.} \{E_x^u |V^u(x(t)) - V^u(x)|^2 + 1 + t^2\} \\ &\leq \text{const.} \{1 + t + W_1(x)\}^2. \end{aligned}$$

Then by the last inequality, Schwarz's inequality, (5.11c), (A5) and (5.13), $E_x^v \int_0^t \psi^u(x(s)) [b^u(x(s)) - b^v(x(s))] ds$ is integrable with respect to μ_v .

Furthermore, with $\sigma_N \cap t$ set equal to t , the E_x^v can be put inside all the integral signs in (6.4). Doing this and integrating each term with respect to μ_v , and using the invariance of μ_v (under control $v(\cdot)$) yields $\int \mu_v(dx) [E_x^v V^u(x(t)) - V^u(x)] = 0$ and

$$0 = \int_0^t ds \int \mu_v(dx) E_x^v \{ \tilde{k}^u(x(s)) + \psi^u(x(s)) [b^u(x(s)) - b^v(x(s))] \}.$$

Now subtract the zero quantity $\int \mu_v(dx) \tilde{k}^v(x)$ from the above equation, and use the invariance of μ_v (under control $v(\cdot)$) to get

$$0 = t \left\{ \int \mu_v(dx) [(\tilde{k}^u(x) - \tilde{k}^v(x)) + \psi^u(x)(b^u(x) - b^v(x))] \right\},$$

or, equivalently

$$(6.5) \quad 0 = \int \mu_v(dx) [e^{u,v}(x) + \theta(v) - \theta(u)].$$

Next, let $A = \{x: e^{u,v}(x) > 0\}$ and let $I(A) > 0$. Define the admissible control $\bar{v}(\cdot)$ by: $\bar{v}(x) = u(x)$ on $R^r - A$, $\bar{v}(x) = v(x)$ on A . Since (6.5) holds for all $u(\cdot)$, $v(\cdot)$, $e^{u,\bar{v}}(x) \geq 0$ and

$$0 = \int \mu_{\bar{v}}(dx) [e^{u,\bar{v}}(x) + \theta(\bar{v}) - \theta(u)].$$

But $e^{u,\bar{v}}(x) > 0$ on A , and $\mu_{\bar{v}}(A) > 0$ by Theorem 3.1. Thus, $\theta(\bar{v}) < \theta(u)$, proving the first assertion of the theorem. The second assertion follows easily by the same type of argument on (6.5). Q.E.D.

Remark. The reason for inserting the corollary is discussed after the proof.

COROLLARY. Assume (A1)–(A5). Let $u(\cdot)$ be optimal and $v(\cdot)$ an admissible control. Then

$$(6.6) \quad \theta(v) = \theta(u) - \int \mu_v(dx) e^{u,v}(x).$$

Let $\psi^u(\cdot)$ be bounded on bounded x -sets. For each $\varepsilon > 0$, let $\psi_\varepsilon(\cdot)$ denote a (r -row vector) Borel function such that

$$(6.7) \quad \int_K |\psi_\varepsilon(x) - \psi^u(x)| dx \rightarrow 0, \quad \text{each bounded set } K, \text{ as } \varepsilon \rightarrow 0,$$

$$\sup_{\varepsilon, x \in K} |\psi_\varepsilon(x)| < \infty, \quad \text{each bounded set } K.$$

Let K denote a fixed compact set in R' , which is the closure of its interior. Suppose that the function $v_\varepsilon(\cdot)$ is calculated by

$$(6.8) \quad v_\varepsilon(x) = \arg \inf_{\alpha \in \mathcal{U}} [k(x, \alpha) + \psi_\varepsilon(x)b(x, \alpha)], \quad \text{for almost all } x \in K,$$

$$v_\varepsilon(x) = 0, \quad x \notin K.$$

Then $v_\varepsilon(\cdot)$ can be assumed to be admissible, and

$$(6.9) \quad \overline{\lim}_{\varepsilon \rightarrow 0} \theta(v_\varepsilon) \leq \theta(u) - \lim_{\varepsilon \rightarrow 0} \int_{R'-K} \mu_{v_\varepsilon}(dx) [(k^u(x) - k^o(x)) + \psi^u(x)b^u(x)].$$

Proof. Equation (6.6) is just (6.5). By the complete lattice property [14, p. 302] of $L_1(K)$, the inf in (6.8) can be assumed to be in $L_1(K)$. Then, by the properties of \mathcal{U} , $b(\cdot)$, $k(\cdot)$ in (A2) and (A3), we get that the inf (evaluated at x) is in the set $k(x, \mathcal{U}) + \psi_\varepsilon(x)b(x, \mathcal{U})$, for almost all $x \in K$. Then, the implicit function theorem cited in Theorem 4.1 can be used to show that there is an admissible control which attains the inf almost everywhere. We call this control $v_\varepsilon(\cdot)$.

Note that, if $\psi_\varepsilon(x) \rightarrow \psi^u(x)$ for a fixed $x \in K$, then

$$(6.10) \quad \inf_{\alpha \in \mathcal{U}} [k(x, \alpha) + \psi_\varepsilon(x)b(x, \alpha)] \rightarrow \inf_{\alpha \in \mathcal{U}} [k(x, \alpha) + \psi^u(x)b(x, \alpha)], \quad \text{as } \varepsilon \rightarrow 0.$$

Also, the L_1 convergence (6.7) implies that for each $\delta > 0$, there is an $\varepsilon_\delta > 0$ and a set $A_{\varepsilon_\delta} \in K$ with $l(A_{\varepsilon_\delta}) < \delta$ for $\varepsilon < \varepsilon_\delta$ and such that $|\psi_\varepsilon(x) - \psi^u(x)| \leq \delta$ for $x \in A_{\varepsilon_\delta}$, $\varepsilon < \varepsilon_\delta$. This, together with (6.10), implies that the difference between the sides in (6.10) converges in $L_1(K)$ as $\varepsilon \rightarrow 0$. Note that the r.h.s. of (6.10) equals $k^u(x) + \psi^u(x)b^u(x)$ (almost everywhere) by optimality of $u(\cdot)$, and the theorem. Now, by (6.6),

$$(6.11) \quad \theta(v_\varepsilon) = \theta(u) - \int_K \mu_{v_\varepsilon}(dx) [(k^u(x) - k^{v_\varepsilon}(x)) + \psi^u(x)(b^u(x) - b^{v_\varepsilon}(x))]$$

$$- \int_{R'-K} \mu_{v_\varepsilon}(dx) [(k^u(x) - k^o(x)) + \psi^u(x)b^u(x)].$$

The integrand of the first integral of (6.11) equals $[k^u(x) + \psi^u(x)b^u(x)] - [k^{v_\varepsilon}(x) + \psi_\varepsilon(x)b^{v_\varepsilon}(x)] - (\psi^u(x) - \psi_\varepsilon(x))b^{v_\varepsilon}(x)$. The remarks below (6.10), and the fact

that $\mu_v(A) \rightarrow 0$ as $l(A) \rightarrow 0$ uniformly in $v(\cdot)$ and A (Theorem 4.2) imply both that the first integral on the r.h.s. of (6.11) goes to zero, as $\varepsilon \rightarrow 0$, and the theorem. Q.E.D.

Remark on the corollary. The corollary was given because it will probably be useful when used in conjunction with a procedure for computing or estimating ψ^o). Usually, we would not be able to calculate $\psi^u(\cdot)$ exactly, and the corollary asserts that, even if the computation is approximate, its use to get a control may yield good results, since the cost is "continuous in $\psi_\varepsilon(\cdot)$ ", in a sense, provided that $\int_{R'-S_N} \mu_v(dx) |\psi^u(x)| \rightarrow 0$ as $N \rightarrow \infty$, uniformly in $v(\cdot)$. We would expect that this latter condition would hold quite often.

THEOREM 6.2. Assume (A1), (A2), (A4), (A5). Then M_t^u has the representation (6.1).

Proof. By Theorem 2.3 of Davis and Variaya [1], and the square integrability of M_t^u , there is a process $\xi^{u,x}(\cdot)$ such that $E_x^u \int_0^t |\xi^{u,x}(s)|^2 ds < \infty$ for each t and such that

$$M^u(s) = \int_0^s \xi^{x,u}(s) \sigma(x(s)) dW^{x,u}(s), \quad \text{w.p.1 } P_x^u.$$

Let \mathcal{A}^u denote the class of continuous random functions that are square integrable martingales under P_x^u , each x , and are also homogeneous additive functions of the Markov process $x(\cdot)$, and which are adapted to $\{\mathcal{G}_t\}$. If $N(\cdot) \in \mathcal{A}^u$, then the quadratic variation $\langle N, N \rangle_t$ has a representation which is a homogeneous additive nondecreasing function of $x(\cdot)$. It does not otherwise depend on $x(0) = x$. (See, for example, Meyer [15, Thm. 3, p. 126]. The result is also implied by Kunita and Watanabe [16, Appendix].) The processes $M^u(\cdot)$ and

$$W^{x,u}(t) = \int_0^t \sigma^{-1}(x(s)) [dx(s) - (f(x(s)) + b^u(x(s)))] ds$$

are in \mathcal{A}^u . Also $W^{x,u}(\cdot) \pm M^u(\cdot)$ are both in \mathcal{A}^u . Then

$$\langle W^{x,u} + M^u, W^{x,u} + M^u \rangle_t - \langle W^{x,u} - M^u, W^{x,u} - M^u \rangle_t = 4 \langle W^{x,u}, M^u \rangle_t$$

is a homogeneous additive process. But

$$\langle W^{x,u}, M^u \rangle_t = \int_0^t \xi^{x,u}(s) \sigma(x(s)) ds,$$

which must also have a representation as a homogeneous additive function of $x(\cdot)$, and which does not otherwise depend on $x(0) = x$. Thus, there is a Borel function $\psi^u(\cdot)$, not depending on $x = x(0)$, such that $\xi^{u,x}(s) \sigma(x(s)) = \psi^u(x(s)) \sigma(x(s))$, for almost all s w.p.1 P_x^u , each x . Q.E.D.

REFERENCES

- [1] M. H. A. DAVIS AND P. P. VARAIYA, *Dynamic programming conditions for partially observable stochastic systems*, this Journal, 11 (1973), pp. 226–261.
- [2] J. M. BISMUT, *Théorie probabiliste du contrôle des diffusions*, Mem. Amer. Math. Soc., 4 (1976), no. 167.
- [3] W. W. STROOK AND S. R. S. VARADHAN, *Diffusion processes with continuous coefficients*, Comm. Pure and Appl. Math., 22 (1969), pp. 345–400, 479–530.
- [4] H. J. KUSHNER, *Introduction to Stochastic Control Theory*, Holt, Rinehart and Winston, New York, 1971.
- [5] P. MANDL, *On the control of non-terminating diffusions*, Theor. Probability Appl., 9 (1964), pp. 591–603.

- [6] YA. A. KOGAN, *On optimal control of a non-terminating diffusion process with reflection*, Ibid., 14 (1969), pp. 496–502.
- [7] I. V. GIRSANOV, *On transforming a certain class of stochastic processes by absolutely continuous transformation of measures*, Ibid., 5 (1960), pp. 285–301.
- [8] R. Z. KHAZMINSKII, *Ergodic properties of recurrent diffusion processes, and stabilization of the solution to the Cauchy problem for parabolic systems*, Ibid., 5 (1960), pp. 179–196.
- [9] J. L. DOOB, *Asymptotic properties of Markov transition probabilities*, Trans. Amer. Math. Soc., 63 (1948), pp. 393–421.
- [10] E. ROXIN, *The existence of optimal controls*, Michigan Math. J., 9 (1962), pp. 109–119.
- [11] E. J. MCSHANE AND R. B. WARFIELD, *On Fillipov's implicit function lemma*, Proc. Amer. Math. Soc., 18 (1967), pp. 41–47.
- [12] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.
- [13] S. OREY, *Lecture Notes on Limit Theorems for Markov Chain Transition Probabilities*, Van Nostrand Reinhold, London, 1971.
- [14] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, John Wiley, New York, 1958.
- [15] P. A. MEYER, *Intégrals stochastiques*, Séminaire de Probabilités no. 1, Lecture Notes in Math 39, Springer-Verlag, Berlin, pp. 72–141.
- [16] H. KUNITA AND S. WATANABE, *On square integrable martingales*, Nagoya Math. J., 30 (1967), pp. 209–245.

THE FREE BOUNDARY FOR VARIATIONAL INEQUALITIES WITH NONLOCAL OPERATORS*

AVNER FRIEDMAN† AND MAURICE ROBIN‡

Abstract. It is well known that the solution of a stationary optimal stopping time problem corresponding to a diffusion process is a solution of an elliptic variational inequality. In this paper we study the corresponding situation for some other Markov processes whose paths are right continuous and whose generator is a nonlocal operator. An existence and uniqueness of a solution of the variational inequality is proved. Special attention is then paid to the shape of the free boundary. It is shown that the free boundary is a curve $y = \varphi(x)$ and $\varphi(x)$ is piecewise monotone.

Introduction. In this paper we consider the variational inequality

$$\begin{aligned} -Au + \alpha u &\leq f, \\ u &\leq 0, \\ (-Au + \alpha u - f)u &= 0 \end{aligned}$$

where A is a nonlocal operator (the generator of a Markov process). We shall be particularly interested in the properties of the free boundary. The variational inequality will correspond to an optimal stopping problem for some noncontinuous Markov process. We shall first study (§§ 1–6) the case where

$$Au = \frac{\partial u(x, y)}{\partial y} + \lambda(x, y) \left[\int_0^1 g(x, y, dz) u(z, 0) - u(x, y) \right];$$

this corresponds to a 2-dimensional process (x_t, y_t) where x_t is called a semi-Markov process. In §§ 7, 8 we shall extend the study to a process which is the continuous analogue of the process arising in the $M/G/1$ queuing system. Both processes are often involved in applications to replacement problems and queuing control; see, for instance, [1], [10].

The shape and smoothness of the free boundary for elliptic and parabolic variational inequalities have been studied by various authors (see, for instance, [3], [4], [5], [6] and the references given there). However, it seems that results on the free boundary for nonlocal operators have not appeared so far in the literature.

In § 1 we introduce the semi-Markov process and the corresponding optimal stopping time problem:

$$u(x) = \inf_{\tau} J_{xy}(\tau),$$

where

$$J_{xy}(\tau) = E_{xy} \int_0^{\tau} e^{-\alpha s} f(x_s, y_s) ds,$$

and τ is any stopping time for the process (x_t, y_t) .

In § 2 we establish the existence and uniqueness of the solution of the corresponding variational inequality. In § 3 we introduce an approximation procedure by discretizing the x -variable. This procedure is needed in § 5.

* Received by the editors June 6, 1977. This work was partially supported by National Science Foundation under Grant MCS75-21416 A01.

† Department of Mathematics, Northwestern University, Evanston, Illinois 60201.

‡ IRIA, Domaine de Voluceau, Rocquencourt, France.

In § 4 it is proved that the free boundary is a y -graph, that is, $u(x, y) < 0$ if and only if $y < \varphi(x)$, for some bounded function $\varphi(x)$.

In § 5 we study the monotonicity of $\varphi(x)$; it is proved, under suitable assumptions, that $\varphi(x)$ is piecewise monotone. The continuity of $\varphi(x)$ is established in § 6.

In § 7 we construct a Markov process which is the continuous analogue of a queuing process. Finally, in § 8 we extend some of the results of §§ 3–6 to this process.

1. A stopping time problem for semi-Markov process. Roughly we say that $(x_t, t \geq 0)$ is a semi-Markov process if it is a right continuous, left limited jump process and if, defining

$$(1.1) \quad y_t = t - \sup \{s; s < t, x_s \neq x_t\}$$

we have that $\{(x_t, y_t), t \geq 0\}$ is a Markov process. The process y_t is called the *age* (at time t) of the last jump of the component x .

It is clear that a semi-Markov process is a generalization of the pure jump Markov process, where the x component alone is already Markovian.

We shall now give the precise definition of a semi-Markov process. We are given

$$E = [0, 1] \times R^+$$

and functions $\lambda(x, y)$, $q(x, y, \Gamma)$ defined for $(x, y) \in E$ and for any Borel set $\Gamma \subset [0, 1]$, such that

$$(1.2) \quad \lambda(x, y) \text{ is continuous on } E \quad \text{and} \quad 0 \leq \lambda \leq M \quad (M \text{ constant});$$

$q(x, y, \Gamma)$ is a transition probability from E into $[0, 1]$ such that

$$(1.3) \quad \int_0^1 q(x, y, dz) q(z) \in C_b^0(E) \quad \text{for any } g \in C^0[0, 1],$$

where $C_b^0(E)$ is the space of bounded continuous functions on E .

It is well known (see [8], [12]) that one can associate with (λ, q) a Markov process

$$X = (\Omega, \mathcal{F}_t, \theta_t, (x_t, y_t), P_{xy}),$$

such that X is a right continuous homogeneous process, x_t is a jump process, and if ν is the time of the first jump then

$$(1.4) \quad \begin{aligned} P_{xy}(\nu \leq t) &= 1 - \exp \left[- \int_y^{y+t} \lambda(x, \sigma) d\sigma \right], \\ P_{xy}(x_\nu \in \Gamma, \nu \leq t) &= \int_0^t \lambda(x, y + \sigma) \exp \left[- \int_y^{y+\sigma} \lambda(x, \eta) d\eta \right] q(x, y + \sigma, \Gamma) d\sigma. \end{aligned}$$

Moreover, when $[0, 1]$ is endowed with the discrete topology then X is a Feller process and, therefore, it has the strong Markov property. From the construction given in [8] follows the property (1.1).

The process $(x_t, t \geq 0)$ alone is called a *semi-Markov process*.

If λ and q do not depend on y then the semi-Markov process reduces to a pure jump Markov process. This can be seen either from the construction of X or from the infinitesimal generator A of X , which is computed below.

LEMMA 1.1. *Under the conditions (1.2), (1.3), the semi-group $\Phi(t)$ of X satisfies the following properties:*

$$(1.5) \quad \Phi(t)g \in C_b^0(E) \quad \text{for any } g \in C_b^0(E),$$

$$(1.6) \quad \lim_{t \downarrow 0} (\Phi(t)g)(x, y) = g(x, y) \quad \text{for any } g \in C_b^0(E)$$

(that is, Φ is a Feller semi-group when E is endowed with the Euclidean topology). Moreover, if g and $\partial g / \partial y$ are bounded and continuous in y , then g belongs to the domain of the weak infinitesimal generator A of Φ , and

$$(1.7) \quad Ag(x, y) = \frac{\partial g(x, y)}{\partial y} + \lambda(x, y) \left[\int_0^1 q(x, y, dz) g(z, 0) - g(x, y) \right].$$

Recall [5, p. 50] that g belongs to the domain D_A of the weak infinitesimal generator A if the limit

$$(1.8) \quad Ag(x, y) \equiv \lim_{t \downarrow 0} \frac{(\Phi(t)g)(x, y) - g(x, y)}{t}$$

exists in the sense of pointwise and bounded convergence.

Proof. Let g be bounded and Borel measurable on E . Then

$$(1.9) \quad (\Phi(t)g)(x, y) = E_{xy}g(x_t, y_t) = E_{xy}I_{\nu > t}g(x_t, y_t) + E_{xy}I_{\nu \leq t}g(x_t, y_t).$$

On $\{\nu > t\}$, $x_t = x$ and $y_t = y + t$ P_{xy} a.s. Also, by the strong Markov property,

$$E_{xy}I_{\nu \leq t}g(x_t, y_t) = E_{xy}I_{\nu \leq t}\Phi(t - \nu)g(x_\nu, y_\nu);$$

$y_\nu = 0$ P_{xy} a.s. and the distribution of (x_ν, ν) is given by (1.4). Therefore, (1.9) becomes

$$(1.10) \quad \begin{aligned} (\Phi(t)g)(x, y) = & \exp \left[- \int_y^{y+t} \lambda(x, \sigma) d\sigma \right] g(x, y+t) \\ & + \int_0^t \lambda(x, y+s) \exp \left[- \int_y^{y+s} \lambda(x, \sigma) d\sigma \right] \\ & \cdot \int_0^1 q(x, y+s, dz) (\Phi(t-s)g)(z, 0) ds. \end{aligned}$$

From this formula (1.6) follows immediately. In order to prove (1.5), we view (1.10) as an integral equation for the function

$$\psi(t, x, y) = (\Phi(t)g)(x, y).$$

By the standard technique of successively iterating the equation one can represent ψ as a uniformly convergent series. By (1.2), (1.3), each term in this series is a continuous function. Hence the same is true of ψ .

Finally, the assertion (1.7) follows easily from (1.10), with the convergence in (1.8) being pointwise and bounded convergence.

Remark. If g belongs to D_A then Dynkin's formula [5, p. 133] is valid. Therefore, as a consequence of Lemma 1.1 we conclude that if g and $\partial g / \partial y$ are bounded and continuous in y , then

$$(1.11) \quad E_{xy} e^{-\alpha\tau} g(x_\tau, y_\tau) - g(x, y) = E_{xy} \int_0^\tau e^{-\alpha s} (Ag - \alpha g)(x_s, y_s) ds$$

for any stopping time τ .

Let $f \in C_b^0(E)$ and $\alpha > 0$, and define a cost function

$$(1.12) \quad J_{xy}(\tau) = E_{xy} \int_0^\tau e^{-\alpha s} f(x_s, y_s) ds$$

for any stopping time τ . We consider the problem of minimizing $J_{xy}(\tau)$ in the set \mathcal{A} of all stopping times τ . Let

$$(1.13) \quad u(x, y) = \inf_{\tau \in \mathcal{A}} J_{xy}(\tau).$$

By the standard formalization of dynamic programming one derives for u the variational inequality

$$(1.14) \quad \begin{aligned} -Au + \alpha u &\leq f, \\ u &\leq 0, \\ (-Au + \alpha u - f)u &= 0 \end{aligned}$$

in E , where A is given by (1.7).

This variational inequality was investigated in Chapter 1 of [11] in case $E = [0, 1] \times [0 \leq y \leq \bar{y}]$ where $\bar{y} < \infty$ and $[0, 1]$ is provided with the discrete topology. It was proved that there exists a unique solution of (1.14) in some class of functions and that an optimal stopping time $\hat{\tau}$ exists, and it is given by

$$\hat{\tau} = \inf \{s: u(x_s, y_s) = 0\}.$$

For our purposes we shall need to work with $E = [0, 1] \times [0, \infty)$ and with $[0, 1]$ provided with the Euclidean topology. The method we shall use to solve (1.14) is very similar to the method of [11], [2].

2. Existence of a solution of the variational inequality. We begin with the following “penalized problem” for a function u_ε :

$$(2.1) \quad -\frac{\partial u_\varepsilon}{\partial y} - \lambda(x, y) \left[\int_0^1 q(x, y, dz) u_\varepsilon(z, 0) - u_\varepsilon(x, y) \right] + \alpha u_\varepsilon + \frac{1}{\varepsilon} u_\varepsilon^+ = f \quad \text{in } E,$$

for any $\varepsilon > 0$.

THEOREM 2.1. *Let the assumptions (1.2), (1.3) hold and let $f \in C_b^0(E)$. Then there exists a unique solution $u_\varepsilon(x, y)$ of (2.1) satisfying:*

$$u_\varepsilon, \frac{\partial u_\varepsilon}{\partial y} \text{ belong to } C_b^0(E).$$

Proof. We shall use the notation

$$(2.2) \quad (Q\varphi)(x, y) = \int_0^1 q(x, y, dz) \varphi(z, 0).$$

Consider the equation

$$(2.3) \quad -Aw + \alpha w = f.$$

By the Feller property and [5, p. 24], this equation has a unique solution in $D_A \cap C_b^0(E)$, given by

$$(2.4) \quad v^0(x, y) = E_{xy} \int_0^\infty e^{-\alpha s} f(x_s, y_s) ds.$$

We are interested however in showing more, namely, that the solution satisfies:

$$(2.5) \quad w, \frac{\partial w}{\partial y} \text{ belong to } C_b^0(E),$$

so that Aw has the explicit form given by (1.7).

To prove this we introduce a sufficiently large positive number μ (to be determined below) and consider the equation

$$(2.6) \quad -\frac{\partial v}{\partial y} + (\lambda + \alpha + \mu)v - \lambda Qv = f + \mu v^0 \equiv g \quad \text{in } E.$$

Since $v^0 \in C_b^0(E)$, also $g \in C_b^0(E)$.

We define inductively a sequence of functions v^k by

$$-\frac{\partial v^k}{\partial y} + (\lambda + \alpha + \mu)v^k = g + \lambda Qv^{k-1}.$$

It is clear that

$$v^k, \frac{\partial v^k}{\partial y} \text{ belong to } C_b^0(E)$$

and

$$\|v^k - v^{k-1}\| \leq \frac{M}{\alpha + \mu} \|v^{k-1} - v^{k-2}\|,$$

where $\|z\|$ denotes the supremum of $|z|$ over E . Taking μ so that $\alpha + \mu > 2M$, we obtain a solution $v = \lim v^k$ of (2.6) satisfying:

$$v, \frac{\partial v}{\partial y} \text{ belong to } C_b^0(E).$$

Since $v \in D_A$ we can rewrite (2.6) in the form

$$-Av + (\alpha + \mu)v = g = f + \mu v^0.$$

Therefore

$$-A(v^0 - v) + (\alpha + \mu)(v^0 - v) = 0, \quad (v^0 - v) \in D_A.$$

It follows that $v^0 = v$. We have thus proved:

LEMMA 2.2. *For any $f \in C_b^0(E)$ there exists a unique solution w of (2.3) satisfying (2.5).*

Consider now the equation

$$(2.7) \quad u_\varepsilon(x, y) = \int_0^\infty e^{-\alpha t} \Phi(t) \left[f - \frac{1}{\varepsilon} u_\varepsilon^+ \right] (x, y) dt.$$

Note that if $u_\varepsilon \in C_b^0(E)$ then, by [5, p. 24], $u_\varepsilon \in D_A$ and therefore u_ε is a solution of

$$(2.8) \quad -Au_\varepsilon + \alpha u_\varepsilon + \frac{1}{\varepsilon} u_\varepsilon^+ = f,$$

which can be written

$$-Au_\varepsilon + \alpha u_\varepsilon = f - \frac{1}{\varepsilon} u_\varepsilon^+ \equiv g_\varepsilon \in C_b^0(E).$$

Hence, by Lemma 2.2, $\partial u_\varepsilon / \partial y$ belongs to $C_b^0(E)$. Thus in order to complete the proof of Theorem 2.1 it remains to show that (2.7) has a unique solution in $C_b^0(E)$.

Proof of uniqueness for (2.7). From (2.7) and the Markov property we get

$$u_\varepsilon(x, y) = e^{-\alpha t} \Phi(t) u_\varepsilon + \int_0^t e^{-\alpha s} \Phi(s) \left[f - \frac{1}{\varepsilon} u_\varepsilon^+ \right] ds.$$

It is now easy to check, using again the Markov property, that

$$\rho_s = e^{-\alpha s} u_\varepsilon(x_s, y_s) + \int_0^s e^{-\alpha t} \left(f - \frac{1}{\varepsilon} u_\varepsilon^+ \right)(x_t, y_t) dt$$

is a P_{xy} martingale. But then (by [15, Lemma 4.1]) the process

$$\rho_t \psi_t - \int_0^t \rho_s \eta_s ds \quad \text{where } \psi_t = \int_0^t \eta_s ds$$

is also a martingale for any nonanticipative integrable η_s , and, in particular,

$$\begin{aligned} \tilde{\rho}_s &= e^{-\alpha s} \exp \left[-\frac{1}{\varepsilon} \int_0^s v_r dr \right] u_\varepsilon(x_s, y_s) \\ &\quad + \int_0^s e^{-\alpha t} \exp \left[-\frac{1}{\varepsilon} \int_0^t v_r dr \right] \left[f - \frac{1}{\varepsilon} u_\varepsilon^+ + \frac{1}{\varepsilon} v_t u_\varepsilon \right] dt \end{aligned}$$

is a martingale for any v adapted to t with values in $[0, 1]$. Therefore,

$$u_\varepsilon(x, y) = E_{xy} \int_0^\infty e^{-\alpha t} \exp \left[-\frac{1}{\varepsilon} \int_0^t v_r dr \right] \left[f - \frac{1}{\varepsilon} u_\varepsilon^+ + \frac{1}{\varepsilon} v_t u_\varepsilon \right] dt.$$

Using the fact that

$$-\frac{1}{\varepsilon} u_\varepsilon^+ = \inf_v \left[-\frac{1}{\varepsilon} v u_\varepsilon \right],$$

we then get

$$(2.9) \quad u_\varepsilon(x, y) = \inf_v J_{xy}^\varepsilon(v)$$

where

$$(2.10) \quad J_{xy}^\varepsilon(v) = E_{xy} \int_0^\infty e^{-\alpha t} \exp \left[-\frac{1}{\varepsilon} \int_0^t v_r dr \right] f(x_t, y_t) dt.$$

Therefore, if there exists a solution $u_\varepsilon \in C_b^0(E)$ of (2.7), then it is given by (2.9); this proves uniqueness.

Proof of existence for (2.7). We introduce the nonstationary problem for u_ε^T :

$$(2.11) \quad u_\varepsilon^T(t, x, y) = \int_t^T e^{-\alpha(s-t)} \Phi(s-t) \left[f - \frac{1}{\varepsilon} (u_\varepsilon^T(s))^+ \right] ds.$$

Define a mapping Π by

$$(\Pi w)(t, x, y) = \int_t^T e^{-\alpha(s-t)} \Phi(s-t) \left[f - \frac{1}{\varepsilon} w^+(s) \right] ds$$

for any $w(s) = w(s, x, y)$ in $Y = C^0([0, T]; C_b^0(E))$. By the Feller property, Πw

belongs to the same space Y . It is easy to check that Π^k is a contraction if k is sufficiently large. Therefore, (2.11) has a unique solution u_ε^T in Y .

Let $\mathcal{F}_t^{t+h} = \theta_t^{-1} \mathcal{F}_h$, $\mathcal{F}_t = \theta_t^{-1} \mathcal{F}$ and let P_{txy} be the probability in \mathcal{F}_t defined by

$$E_{txy} \varphi \cdot \theta_t = E_{xy} \varphi \quad \text{for any } \varphi \text{ bounded and } \mathcal{F} \text{ measurable.}$$

For any process v adapted to \mathcal{F}_t^s with values in $[0, 1]$, define

$$J_{txy}^{\varepsilon, T}(v) = E_{txy} \int_t^T e^{-\alpha(s-t)} \exp \left[-\frac{1}{\varepsilon} \int_t^s v_r dr \right] f(x_s, y_s) ds.$$

Analogously to the proof of (2.9) one can show (see [11, Chap. 1]) that

$$(2.12) \quad u_\varepsilon^T(t, x, y) = \inf_v J_{txy}^{\varepsilon, T}(v).$$

We introduce the function

$$(2.13) \quad w(x, y) = \inf_v J_{xy}^\varepsilon(v)$$

where $J_{xy}^\varepsilon(v)$ is defined in (2.10) and v varies in the set of all v adapted to \mathcal{F}_t with values in $[0, 1]$, and the function

$$(2.14) \quad w(t, x, y) = \inf_v J_{txy}^\varepsilon(v)$$

where v_r is adapted to \mathcal{F}_t^r with values in $[0, 1]$ and

$$(2.15) \quad J_{txy}^\varepsilon(v) = E_{txy} \int_t^\infty e^{-\alpha(s-t)} \exp \left[-\frac{1}{\varepsilon} \int_t^s v_r dr \right] f(x_s, y_s) ds.$$

Then clearly

$$(2.16) \quad w(x, y) = w(t, x, y).$$

Now, from the stochastic interpretation of u_ε^T and $w(t, x, y)$ we get (cf. [11, Chap. 1] or [2])

$$(2.17) \quad |u_\varepsilon^T(t, x, y) - w(t, x, y)| \leq e^{-\alpha(T-t)} \|f\|.$$

Therefore,

$$(2.18) \quad u_\varepsilon^T(t, x, y) \rightarrow w(t, x, y) \quad \text{uniformly in } E \quad \text{as } T \uparrow \infty.$$

Recalling (2.16) we conclude that $w(x, y)$ is continuous and bounded in E .

A simple application of the Markov property gives

$$u_\varepsilon^T(t, x, y) = e^{-\alpha h} \Phi(h) u_\varepsilon^T(t+h, x, y) + \int_t^{t+h} e^{-\alpha(s-t)} \Phi(s-t) \left[f - \frac{1}{\varepsilon} (u_\varepsilon^T(s))^+ \right] ds$$

for $t < t+h < T$. Taking $T \uparrow \infty$ and using (2.17) and (2.16), we find that

$$w(x, y) = e^{-\alpha h} \Phi(h) w(x, y) + \int_0^h e^{-\alpha s} \Phi(s) \left[f - \frac{1}{\varepsilon} w^+ \right] ds.$$

Taking $h \uparrow \infty$ we obtain the relation (2.7). Thus $u_\varepsilon = w(x, y)$ is the solution of (2.7) in $C_b^0(E)$. This completes the proof of Theorem 2.1.

THEOREM 2.3. *Under the assumptions of Theorem 2.1,*

$$(2.19) \quad u_\varepsilon \rightarrow u \quad \text{uniformly in } E \quad \text{as } \varepsilon \rightarrow 0,$$

where $u(x, y)$ is defined by (1.13).

Proof. We first show that

$$(2.20) \quad u_\varepsilon \geq u.$$

Define

$$\hat{\tau}_\varepsilon = \inf \{s: u_\varepsilon(x_s, y_s) \geq 0\}.$$

Then, by Dynkin's formula and (2.1),

$$u_\varepsilon(x, y) = E_{xy} \left[\int_0^{\hat{\tau}_\varepsilon} e^{-\alpha s} \left[f - \frac{1}{\varepsilon} u_\varepsilon^+ \right] ds + e^{-\alpha \hat{\tau}_\varepsilon} u_\varepsilon(x_{\hat{\tau}_\varepsilon}, y_{\hat{\tau}_\varepsilon}) \right].$$

But since $u_\varepsilon(x_s, y_s) \leq 0$ if $0 \leq s \leq \hat{\tau}_\varepsilon$, u_ε^+ vanishes (in the integrand). Since also $u_\varepsilon(x_{\hat{\tau}_\varepsilon}, y_{\hat{\tau}_\varepsilon}) \geq 0$, we get

$$u_\varepsilon(x, y) \geq J_{xy}(\hat{\tau}_\varepsilon) \geq u(x, y).$$

In order to estimate $u_\varepsilon - u$ from above, take any stopping time τ and define

$$v_\tau(t) = \begin{cases} 0 & \text{if } t < \tau, \\ 1 & \text{if } t \geq \tau. \end{cases}$$

Then

$$J_{xy}^\varepsilon(v_\tau) - J_{xy}(\tau) = E_{xy} \int_\tau^\infty e^{-\alpha s} \exp \left[-\frac{1}{\varepsilon}(s - \tau) \right] f(x_s, y_s) ds.$$

It follows that

$$J_{xy}^\varepsilon(v_\tau) - J_{xy}(\tau) \leq C \|f\| \varepsilon \quad (C \text{ constant}).$$

Since τ is arbitrary, we obtain the inequality

$$(2.21) \quad u_\varepsilon(x, y) - u(x, y) \leq C \|f\| \varepsilon.$$

This completes the proof of (2.19).

We shall need the following condition:

$$(2.22) \quad \int \int_E e^{-2\delta y} \lambda^2(x, y) \left[\int_0^1 q(x, y, dz) g(z) \right]^2 dx dy \leq C \int_0^1 g^2(z) dz \quad (C > 0, \delta > 0)$$

for some constants C, δ and for any continuous function g .

THEOREM 2.4. *Let (1.2), (1.3), (2.22) hold. Then, for any $f \in C_b^0(E)$ there exists a unique solution $w \in C_b^0(E)$, uniformly Lipschitz with respect to y , of the variational inequality*

$$(2.23) \quad \begin{aligned} & -\frac{\partial w}{\partial y} + (\lambda + \alpha)w - \lambda Qw \leq f \quad \text{for each } x \in [0, 1] \quad \text{a.e. in } y, y \geq 0, \\ & w \leq 0 \quad \text{in } E, \\ & \left(-\frac{\partial w}{\partial y} + (\lambda + \alpha)w - \lambda Qw - f \right) w = 0 \quad \text{for each } x \in [0, 1] \quad \text{a.e. in } y, y \geq 0, \end{aligned}$$

and $w(x, y) \equiv u(x, y)$.

Proof. From the stochastic interpretation of u_ε ,

$$u_\varepsilon(x, y) \leq J_{xy}^\varepsilon(1) = E_{xy} \int_0^\infty e^{-\alpha s} \exp \left[-\frac{1}{\varepsilon}s \right] f ds.$$

Therefore

$$u_\varepsilon \leq \frac{\varepsilon}{1 + \alpha\varepsilon} \|f\|,$$

so that

$$(2.24) \quad \frac{1}{\varepsilon} u_\varepsilon^+ \leq \|f\|.$$

From (2.1) we then conclude that

$$(2.25) \quad \left\| \frac{\partial u_\varepsilon}{\partial y} \right\| \leq C, \quad C \text{ a constant independent of } \varepsilon.$$

We also have

$$(2.26) \quad -\frac{\partial u_\varepsilon}{\partial y} + (\lambda + \alpha)u_\varepsilon - \lambda Q_\varepsilon = f - \frac{1}{\varepsilon} u_\varepsilon^+ \leq f.$$

We can take a subsequence of u_ε such that

$$(2.27) \quad \begin{aligned} u_\varepsilon &\rightarrow u \quad \text{uniformly (by Theorem 2.3),} \\ \frac{\partial u_\varepsilon}{\partial y} &\rightarrow \frac{\partial u}{\partial y} \quad \text{in } L^\infty \text{ weak* (by (2.25)).} \end{aligned}$$

From (2.26) we then obtain the first inequality of (2.23). Obviously, $u \leq 0$. We shall now prove that for each $x \in [0, 1]$,

$$(2.28) \quad (-Au + \alpha u - f)u = 0 \quad \text{a.e. in } y.$$

If $u < 0$ in some bounded open subset V of E then, by (2.19), $u_\varepsilon < 0$ in any open subset W of V with closure in V , provided ε is sufficiently small. Hence

$$-Au_\varepsilon + \alpha u_\varepsilon - f = 0 \quad \text{in } W.$$

From (2.27) it then follows that, for any x_0 ,

$$-Au + \alpha u - f = 0 \quad \text{for a.a. } (x_0, y) \in W,$$

hence for a.a. $(x_0, y) \in V$. This completes the proof of (2.28). We have thus proved that $w = u$ is a solution of (2.23) with the asserted regularity properties. It remains to prove uniqueness. The proof, which uses the idea of Laetsch [9], is the same as in [11] with minor changes. For completeness we describe it briefly.

Let $\bar{\varphi}$ be the bounded solution of $-A\bar{\varphi} + \alpha\bar{\varphi} = f - f_0$ where f_0 is a positive constant. If v_1, v_2 are two bounded solutions of (2.23), then set

$$\bar{u}_i = v_i - \bar{\varphi}, \quad \mu = \min \{ \inf \bar{u}_1, \inf \bar{u}_2 \}, \quad \varphi = \bar{\varphi} - \mu.$$

The functions $u_i = \bar{u}_i - \mu$ are then nonnegative solutions of the variational inequality

$$(-Aw + \alpha w, v - w) \geq (f_0, v - w) \quad \text{for any } v \leq \varphi,$$

$$w \leq \varphi,$$

$$w \in L_\delta^2(E) = \{z(x, y); z e^{-\delta y} \in L^2(E)\}$$

for any fixed $\delta > 0$, where (\cdot, \cdot) denote the scalar product in $L_\delta^2(E)$.

Let γ be the largest number in $[0, 1]$ such that $\gamma u_1 \leq u_2$. Assume that $\gamma < 1$. Let $\beta \in (\gamma, 1)$ be defined by

$$\beta = \frac{f_0 + \lambda_0 \gamma \|u_1\|}{f_0 + \lambda_0 \|u_1\|}$$

where λ_0 is positive and sufficiently large. Then

$$\beta f_0 + \lambda_0 \beta u_1(x, y) \leq f_0 + \lambda_0 u_2(x, y).$$

The function βu_1 satisfies the variational inequality

$$(-A(\beta u_1) + \alpha(\beta u_1) + \lambda_0(\beta u_1), v - \beta u_1) \geq (\beta f_0 + \lambda_0 \beta u_1, v - \beta u_1) \\ \text{for any } v \leq \beta \varphi; \quad \beta u_1 \leq \beta \varphi,$$

and the function u_2 satisfies the variational inequality

$$(-A u_2 + \alpha u_2 + \lambda_0 u_2, v - u_2) \geq (f_0 + \lambda_0 u_2, v - u_2) \\ \text{for any } v \leq \varphi; \quad u_2 \leq \varphi.$$

If

$$(2.29) \quad \text{the comparison lemma for variational} \\ \text{inequalities holds for } -A + (\alpha + \lambda_0)I,$$

thus we conclude $\beta u_1 \leq u_2$; a contradiction. It follows that $\gamma = 1$, that is, $u_1 \leq u_2$. Similarly $u_2 \leq u_1$, and the proof is complete.

To prove (2.29), it suffices to establish the coercive inequality

$$(2.30) \quad (-A w + (\alpha + \lambda_0)w, w) \geq \mu(w, w) \quad (\mu > 0).$$

Since

$$(-A w + (\alpha + \lambda_0)w, w) \geq \frac{1}{2}|w(0)|^2 + (\alpha + \lambda_0)|w|^2 - (\lambda Q w, w),$$

and

$$|(\lambda Q w, w)| \leq \sqrt{C} \cdot |w(0)|_{L^2(0,1)},$$

where C is the constant appearing in (2.22), (2.30) follows provided $\lambda_0 > C$.

3. Discretization of the x -component. Later on we shall need an approximation scheme for the x -component. For any positive integer n , let

$$h = \frac{1}{n}, \quad S_n = \{0, h, 2h, \dots, (n-1)h\}.$$

We shall assume the following:

$$(3.1) \quad \begin{aligned} &\text{For any } n \text{ we are given a transition probability} \\ &q^n(i, y, j) \text{ from } S_n \times R^+ \text{ into } S_n \text{ such that, for any} \\ &g \in C^0[0, 1], (Q^n g)(i, y) \equiv \sum_{j \in S_n} q^n(i, y, j)g(jh) \text{ belongs to} \\ &C_b^0(R^+); (\tilde{Q}^n g)(i, y) \equiv \sum_{j \in S_n} q^n(i, y, j)g(jh) \text{ on} \\ &x \in [ih, (i+1)h] \text{ (if } i = n-1 \text{ then we take} \\ &x \in [(n-1)h, nh] = [(n-1)h, 1]) \text{ converges to} \\ &(Qg)(x, y) = \int_0^1 q(x, y, dz)g(z) \text{ uniformly in compact} \\ &\text{subsets of } E. \end{aligned}$$

Set

$$\begin{aligned}\lambda^n(i, y) &= \lambda(ih, y), \\ \lambda^n(x, y) &= \lambda^n(i, y) \quad \text{if } x \in [ih, (i+1)h)\end{aligned}$$

and define $\Phi^n(t)$ to be the semi-group of the Markov process (x_s, y_t) corresponding to λ^n, q^n in the state space $S_n \times R^+$. We denote the corresponding probabilities by P_{iy}^n .

If g^j, g are functions in $C_b^0(E)$ and if $g^j \rightarrow g$ uniformly in compact subsets of E , then we write:

$$g^j \rightarrow g \quad \text{in } C_K.$$

For any function $h(i, y)$ defined on $S_n \times R^+$, we denote by $h(x, y)$ the function on E defined by

$$h(x, y) = h(i, y) \quad \text{if } x \in [ih, (i+1)h),$$

with $h(x, y) = h(n-1, y)$ if $x = 1$.

LEMMA 3.1. Let $g^n \in C_b^0(E)$, $g \in C_b^0(E)$, $g^n \rightarrow g$ in C_K . Then

$$(3.2) \quad \Phi^n(t)g^n \rightarrow \Phi(t)g \quad \text{in } C_K.$$

Proof. Since $g^n(x, 0) \rightarrow g(x, 0)$ uniformly, $Q^n g^n \rightarrow Qg$ uniformly in compact sets. Now, by formula (1.10) and the paragraph following it we have that $\Phi^n(t)g^n$ is the unique solution of the integral equation

$$\begin{aligned}v_n &\equiv \Phi^n(t)g^n(i, y) \\ &= \exp \left[- \int_y^{y+t} \lambda^n(i, \sigma) d\sigma \right] g^n(i, y+t) \\ &\quad + \int_0^t \lambda^n(i, y+s) \exp \left[- \int_y^{y+s} \lambda^n(i, \sigma) d\sigma \right] Q^n(\Phi^n(t-s)g^n)(i, y+s) ds,\end{aligned}$$

where

$$(Q^n \varphi)(i, y) = \sum_{j \in S_n} q^n(i, y, j) \varphi(j, 0).$$

The solution is obtained by iterations of the form

$$\begin{aligned}v_n^{k+1}(t, i, y) &= \exp \left[- \int_y^{y+t} \lambda^n(i, \sigma) d\sigma \right] g^n(i, y+t) \\ &\quad + \int_0^t \lambda^n(i, y+s) \exp \left[- \int_y^{y+s} \lambda^n(i, \sigma) d\sigma \right] (Q^n v_n^k)(t-s, i, y+s) ds \\ &\equiv \Pi^n v_n^k,\end{aligned}$$

and

$$v_n(t, i, y) = \lim_{k \rightarrow \infty} \left[v_n^0 + \sum_{j=1}^k (v_n^j - v_n^{j-1}) \right],$$

where $v_n^0 = v^0$ is any initial function in $C_b^0(E)$, and

$$\|v_n^j - v_n^{j-1}\| \leq C \frac{T^{j-1}}{(j-1)!} \|v^0 - \Pi^n v^0\| \quad (0 \leq j \leq T).$$

Therefore

$$(3.3) \quad \|v_n - v_n^k\| \leq C_T \beta_k, \quad \beta_k \rightarrow 0 \quad \text{if } k \rightarrow \infty, \quad \text{uniformly with respect to } n.$$

We next show that, for fixed k ,

$$(3.4) \quad v_n^k \rightarrow v^k \quad \text{in } C_K, \quad \text{if } n \rightarrow \infty$$

where v^k is the k th iterate of v^0 for the operator $\Phi(t)$ and the function g . Using (3.1) and $g^n \rightarrow g$ in C_K , $\lambda^n \rightarrow \lambda$ in C_K , we readily obtain the conclusion (3.4) for $k = 1$. The proof for general k is obtained similarly by induction. The assertion (3.2) now follows by combining (3.3) (and the corresponding estimate for $v - v^k$) with (3.4).

THEOREM 3.2. *Let*

$$\begin{aligned} u^n(i, y) &= \inf_{\tau} E_{iy}^n \int_0^{\tau} e^{-\alpha s} f(x_s, y_s) ds, \\ u^n(x, y) &= u^n(i, y) \quad \text{if } x \in [ih, (i+1)h), \\ u(x, y) &= \inf_{\tau} E_{xy} \int_0^{\tau} e^{-\alpha s} f(x_s, y_s) ds. \end{aligned}$$

Then $u^n \rightarrow u$ in C_K , as $n \rightarrow \infty$.

Proof. We begin with the proof that

$$u_\varepsilon^n(x, y) \rightarrow u_\varepsilon(x, y) \quad \text{in } C_K, \quad \text{if } n \rightarrow \infty,$$

where u_ε^n is the unique solution of

$$u_\varepsilon^n(x, y) = \int_0^\infty e^{-\alpha s} \Phi^n(s) \left[f - \frac{1}{\varepsilon} (u_\varepsilon^n)^+ \right] ds, \quad x \in S_n.$$

We recall that u_ε is obtained as a limit of u_ε^T ($T \rightarrow \infty$). Similarly we can obtain u_ε^n as a limit of $u_\varepsilon^{n,T}$ where

$$u_\varepsilon^{n,T}(t, x, y) = E_{txy}^n \int_t^T e^{-\alpha(s-t)} \Phi^n(s-t) \left[f - \frac{1}{\varepsilon} (u_\varepsilon^{n,T})^+ \right](s) ds.$$

We can now proceed by the same method as is the proof of Lemma 3.1. We write, for any compact set $K \subset E$,

$$\begin{aligned} \|u_\varepsilon^n - u_\varepsilon\|_K &\leq \|u_\varepsilon^n - u_\varepsilon^{n,T}(t)\| + \|u_\varepsilon^{n,T}(t) - v_n^{k,T}(t)\| + \|v_n^{k,T}(t) - v^{k,T}(t)\|_K \\ &\quad + \|v^{k,T}(t) - u_\varepsilon^T(t)\| + \|u_\varepsilon^T(t) - u_\varepsilon\| \\ &\equiv J_1 + J_2 + J_3 + J_4 + J_5. \end{aligned}$$

where $v_n^{k,T}$ ($v^{k,T}$) is the k th iteration in the successive approximation for $u_\varepsilon^{n,T}$ (u_ε^T). We have

$$\begin{aligned} J_1 + J_5 &\leq C e^{-\alpha(T-t)} \|f\|, \\ J_4 &\leq C_T \beta_k, \quad \beta_k \rightarrow 0, \end{aligned}$$

by the method of proof of (3.3); also

$$J_2 \leq C_{T,\varepsilon} \beta_k,$$

and

$$J_3 \rightarrow 0 \quad \text{if } n \rightarrow \infty \quad \text{for fixed } \varepsilon, k,$$

by the method of proof of (3.4). The constants C , C_T , $C_{T,\varepsilon}$ are independent of n . Taking $n \rightarrow \infty$ and then $k \rightarrow \infty$, we obtain, since T can be made arbitrarily large,

$$(3.5) \quad \lim_{n \rightarrow \infty} \|u_\varepsilon^n - u_\varepsilon\|_K = 0.$$

In the proof of Theorem 2.3 we have shown that

$$\|u_\varepsilon - u\| \leq C\varepsilon \|f\|.$$

Similarly

$$\|u_\varepsilon^n - u^n\| \leq C\varepsilon \|f\|$$

where C is a constant independent of n . Combining this with (3.5), we have the assertion of Theorem 3.2.

Example. Let $q(x, y, z) = (1/x)I_{z < x} dz$. Then we take

$$q^n(i, j) = \frac{1}{i} I_{j < i} \quad \text{if } i > 0,$$

$$q^n(0, j) = \delta_{0j}.$$

If we take $\lambda(x, y)$ such that $\lambda(x, y)/x$ is bounded then also the condition (2.22) is satisfied. (In this example we have an absorbing state $x = 0$.)

4. The free boundary is a y -graph. In this and the following two sections we shall always assume that the conditions (1.2), (1.3) and (2.22) are satisfied and that $f \in C^0(E)$, but f is not necessarily bounded. We shall need the following assumption: there exist positive constants y_0 , m , ε such that

$$(4.1) \quad \begin{aligned} f(x, y) &\geq -m \quad \text{in } E, \\ f(x, y) &\geq \varepsilon + M(m/\alpha) \quad \text{if } 0 \leq x \leq 1, \quad y \geq y_0. \end{aligned}$$

This assumption is satisfied, for instance, if

$$(4.2) \quad \inf_{0 \leq x \leq 1} f(x, y) \rightarrow \infty \quad \text{as } y \rightarrow \infty.$$

LEMMA 4.1. *If (4.1) is satisfied then there exists a solution u of the variational inequality (2.23) which is bounded, continuous and uniformly Lipschitz continuous in y . Further, (i) the solution is unique if either f is bounded or (4.2) holds; (ii) the solution u has compact support.*

Proof. For any $R > 0$, let

$$f_R(x, y) = \begin{cases} f(x, y) & \text{if } y < R, \\ f(x, R) & \text{if } y \geq R. \end{cases}$$

By Theorem 2.4 there exists a unique solution u_R of (2.23) with $f = f_R$. From the stochastic interpretation (1.13) and from (4.1) it follows that

$$(4.3) \quad u_R(x, y) \geq \inf_{\tau} E_{xy} \int_0^{\tau} e^{-\alpha s} (-m) dx \geq -\frac{m}{\alpha}.$$

Set

$$\tilde{f} = f_R + \lambda Q u_R.$$

The function $w = u_R$ satisfies the variational inequality

$$(4.4) \quad \begin{aligned} -\frac{\partial w}{\partial y} + (\alpha + \lambda)w &\leq \tilde{f}, & w &\leq 0, \\ \left(-\frac{\partial w}{\partial y} + (\alpha + \lambda)w - \tilde{f}\right)w &= 0, \end{aligned}$$

and, by (4.1) and (4.3),

$$(4.5) \quad \tilde{f}(x, y) \geq \varepsilon \quad \text{if } y \geq y_0.$$

For any $y^* \geq y_0$, consider the function

$$\bar{w}(y) = \varepsilon(y^* - y) \quad \text{for } y^* < y < y^* + N, \quad N = m/(\alpha\varepsilon).$$

It satisfies

$$\bar{w}(y^* + N) = -N\varepsilon = -m/\alpha \leq w(x, y^* + N) \quad (\text{by (4.3)}).$$

Further,

$$-\frac{\partial \bar{w}}{\partial y} + (\alpha + \lambda)\bar{w} = \varepsilon + (\alpha + \lambda)\bar{w} \leq \varepsilon \leq \tilde{f} \quad \text{if } y^* < y < N.$$

By comparison with w we then get

$$\bar{w}(y) \leq w(x, y) \quad \text{if } y^* \leq y \leq y^* + N.$$

In particular, $w(x, y^*) \geq \bar{w}(y^*) = 0$, so that $w(x, y^*) = 0$. Since y^* is any number $\geq y_0$,

$$u_R(x, y) = 0 \quad \text{if } y \geq y_0.$$

If we now take $R > y_0$ then $u = u_R$ is a solution of (2.23) having all the properties asserted in Lemma 4.1, including the compact support property. If further, f is bounded, then uniqueness follows from Theorem 2.4. Thus it remains to prove uniqueness in case (4.2) holds.

Suppose w is another bounded solution. Then w satisfies (4.4) with

$$\tilde{f} = f + \lambda Qw.$$

In view of (4.2), there exists a $\bar{y} > 0$ such that $\tilde{f} \geq \varepsilon > 0$ if $y \geq \bar{y}$. Using the comparison function \bar{w} as before (with $y^* \geq \bar{y}$) we find that $w(x, y) = 0$ if $y \geq \bar{y}$. Therefore w is a solution of (2.23) with $f = f_R$ provided R is sufficiently large. By the uniqueness part of Theorem 2.4 for the bounded function f_R it then follows (when $R > y_0$) that $w = u$.

Let

$$G = \{(x, y) \in E; u(x, y) < 0\},$$

$$\Gamma_0 = \partial G \cap \{0 < x < 1\},$$

$$\Gamma = \text{closure of } \Gamma_0.$$

Γ is called the *free boundary*.

Recall that the optimal stopping occurs when (x_t, y_t) hits the set $E \setminus G$.

In this and the following two sections we shall study the shape and smoothness of the free boundary.

We shall need the additional assumptions:

$$(4.6) \quad \lambda_y \text{ exists and is continuous in } E, \text{ and } \lambda_y(x, y) = 0 \text{ at any point where } \lambda(x, y) = 0; \lambda_y/\lambda \text{ is bounded on the set where } \lambda \neq 0;$$

$$(4.7) \quad \text{for any Borel set } \Gamma \subset [0, 1], q_y(x, y, \Gamma) \text{ exists and is continuous in } E, \text{ and } q_y(x, y, \Gamma) - kq(x, y, \Gamma) \leq 0 \text{ (} k \text{ nonnegative constant);}$$

$$(4.8) \quad (\alpha - k)\lambda_y \leq 0;$$

$$(4.9) \quad f_y \text{ exists and is continuous in } E \text{ and } f_y - kf \geq 0 \text{ at any point where } \lambda = 0, f_y - kf - \lambda_y/\lambda \geq 0 \text{ at any point when } \lambda \neq 0.$$

THEOREM 4.2. *Let the assumptions (4.2) and (4.6)–(4.9) be satisfied. Then there exists a finite valued function $y = \varphi(x)$, $0 \leq x \leq 1$, such that*

$$(4.10) \quad \begin{aligned} u(x, y) &< 0 & \text{if } y < \varphi(x), \\ u(x, y) &= 0 & \text{if } y \geq \varphi(x). \end{aligned}$$

Proof. For a fixed $x \in [0, 1]$, the set

$$I = \{y: y > 0, u(x, y) < 0\}$$

is an open set, which is bounded, by Lemma 4.1. Hence I is a disjoint countable union of bounded intervals (a_i, b_i) . If we show that for any such interval (a_i, b_i) , $a_i = 0$, then the assertion of the theorem follows. Set $a = a_i$, $b = b_i$.

We have

$$(4.11) \quad -\frac{\partial u}{\partial y} + (\alpha + \lambda)u - \lambda Qu = f \quad \text{a.e. in } (a, b).$$

Hence $\partial u/\partial y$ is continuous in (a, b) , (4.11) holds for all y and (by the regularity assumptions in (4.6), (4.7), (4.9)) $\partial^2 u/\partial y^2$ is a bounded function.

Define a function w by $u = e^{ky}w$. Then

$$(4.12) \quad -\frac{\partial w}{\partial y} + (\alpha + \lambda - k)w - \lambda e^{-ky}Qw = f e^{-ky} \quad \text{in } (a, b).$$

Differentiating this equation with respect to y and then substituting for $(w - e^{-ky}Qw)$ its expression from (4.12), we obtain a differential equation for $\zeta = \partial w/\partial y$. For the function $\eta = e^{\gamma y}\zeta$ (γ constant), the equation becomes

$$\begin{aligned} -\frac{\partial \eta}{\partial y} + (\alpha + \lambda - k + \lambda_y/\lambda + \gamma)\eta &= \lambda e^{-ky+\gamma y}(-kQw + Q_y w) + e^{-ky+\gamma y}(-kf + f_y - \lambda_y/\lambda)f \\ &\quad + e^{\gamma y}(\lambda_y/\lambda)(\alpha - k)w \equiv \Phi; \end{aligned}$$

at the points where $\lambda = 0$, the terms with λ_y/λ do not appear. Taking γ sufficiently large and using (4.6)–(4.9), we see that $\Phi \geq 0$. Since also

$$w(x, y) < 0 \quad \text{if } y < b, \quad w(x, b) = 0,$$

we have

$$\frac{\partial w(x, b-0)}{\partial y} \geq 0; \quad \text{hence } \eta(x, b-0) \geq 0.$$

It follows that $\eta(x, y) \geq 0$ if $a < y < b$, so that $e^{-ky}u(x, y)$ is monotone increasing in y , $a < y < b$. Therefore $u(x, a) < 0$, and this implies that $a = 0$.

Remark. If the conditions (4.6)–(4.9) are satisfied for x in a subset J of $[0, 1]$ then the assertion of the theorem is also valid for each x in J .

5. The free boundary is an x -graph. We shall need the following assumptions:

$$(5.1) \quad \lambda_x \text{ exists and is continuous in } E;$$

$$(5.2) \quad \text{for any Borel set } \Gamma \subset [0, 1], q_x(x, y, \Gamma) \text{ exists and is continuous in } E, \text{ and } \lambda_x q + \lambda q_x - \lambda q \left(\int_0^y \lambda_x(x, \eta) d\eta \right) \geq 0;$$

$$(5.3) \quad f_x \text{ exists and is continuous in } E, \text{ and } f_x - f \left(\int_0^y \lambda_x(x, \eta) d\eta \right) \leq 0.$$

THEOREM 5.1. *Let the assumptions (5.1)–(5.3) hold. Then there exists a function $x = \psi(y)$ such that*

$$(5.4) \quad \begin{aligned} u(x, y) &< 0 & \text{if } x > \psi(y), & y \geq 0, \\ u(x, y) &= 0 & \text{if } x < \psi(y), & y \geq 0. \end{aligned}$$

Proof. We introduce the discretized problem with

$$q^n(i, y, j) = \int_{j/n}^{(j+1)/n} q\left(\frac{i}{n}, y, dz\right).$$

The condition (3.1) is easily verified. Therefore, if the $u^n(i, y)$ are the solutions of the variational inequalities

$$(5.5) \quad \begin{aligned} L_{ni}u^n &\equiv -\frac{\partial u^n(i, y)}{\partial y} + (\alpha + \lambda(i, y))u^n(i, y) \\ &\quad - \lambda(i, y) \sum_{j=0}^{n-1} q(i, y, j)u^n(j, 0) \leq f(i, y), \end{aligned}$$

$$\begin{aligned} u^n(i, y) &\leq 0, \\ (L_{ni}u^n - f(i, y))u^n(i, y) &= 0 \end{aligned}$$

and if \tilde{u}^n is defined by

$$(5.6) \quad \tilde{u}^n(x, y) = u^n(i, y) \quad \text{when } \frac{i}{n} \leq x < \frac{i+1}{n},$$

then, by Theorem 3.1,

$$(5.7) \quad \tilde{u}^n \rightarrow u \quad \text{uniformly on compact subsets of } E.$$

Set

$$(5.8) \quad \zeta_i(y) = u^n(i, y) \exp \left[- \int_0^y \lambda_i(\eta) d\eta \right], \quad \lambda_i(\eta) = \lambda(i, \eta).$$

Then the ζ_i satisfy

$$(5.9) \quad \begin{aligned} M_i \zeta &\equiv -\frac{\partial \zeta_i}{\partial y} + \alpha \zeta_i - \lambda_i(Q_i \zeta) \exp \left[- \int_0^y \lambda_i \right] \leq f_i \exp \left[- \int_0^y \lambda_i \right], \\ \zeta_i &\leq 0, \quad \left(M_i \zeta - f_i \exp \left[- \int_0^y \lambda_i \right] \right) \zeta_i = 0 \end{aligned}$$

where

$$Q_i \zeta = \sum_{j=0}^{n-1} q^n(i, y, j) u^n(j, 0).$$

We want to show that

$$(5.10) \quad \zeta_{i+1} \leq \zeta_i.$$

By comparing the variational inequalities for ζ_i and ζ_{i+1} we see that if

$$(5.11) \quad f_i \exp \left[- \int_0^y \lambda_i \right] \geq f_{i+1} \exp \left[- \int_0^y \lambda_{i+1} \right],$$

$$(5.12) \quad \lambda_i(Q_i \zeta) \exp \left[- \int_0^y \lambda_i \right] \geq \lambda_{i+1}(Q_{i+1} \zeta) \exp \left[- \int_0^y \lambda_{i+1} \right],$$

then (5.10) follows. We now only have to observe that (5.11) is a consequence of (5.1) and (5.3) and (since $\zeta_j \leq 0$) (5.12) is a consequence of (5.1) and (5.2).

From (5.10) and (5.7) we deduce that the function

$$u(x, y) \exp \left[- \int_0^y \lambda(x, \eta) d\eta \right]$$

is monotone increasing in x , and this yields the assertion of the theorem.

When the conditions of both Theorem 4.2 and Theorem 5.1 are satisfied, the curve $y = \varphi(x)$ is monotone increasing.

The proof of Theorem 5.1 can be used also to prove, under suitable conditions, that the curve $y = \varphi(x)$ is not monotone, but piecewise monotone. To prove such a result, we impose the following conditions: Let $0 = x_0 < x_1 < x_2 < \cdots < x_m = 1$, and let

$$J_0 = \bigcup_{i=0}^k (x_{2i} < x < x_{2i+1}),$$

$$J_1 = \bigcup_{i=1}^k (x_{2i-1} < x < x_{2i})$$

where $m = 2k + 1$. Then

$$(5.13) \quad (-1)^i \left(f_x - f \int_0^y \lambda_x(x, \eta) d\eta \right) \leq 0 \quad \text{if } x \in J_i \quad (i = 0, 1);$$

$$(5.14) \quad q_x(x, y, \Gamma) \text{ exists and is continuous, for any Borel set } \Gamma, \text{ and} \\ (-1)^i (\lambda_x q + \lambda q_x - \lambda q (\int_0^y \lambda_x(x, \eta) d\eta)) \geq 0 \quad \text{if } x \in J_i \quad (i = 0, 1).$$

THEOREM 5.2. *Let the assumptions of Theorem 4.2 hold and assume that (5.1) and (5.13), (5.14) are satisfied. Then*

$$(5.15) \quad \begin{aligned} \varphi(x) &\text{ is increasing for } x \in J_0, \\ \varphi(x) &\text{ is decreasing for } x \in J_1. \end{aligned}$$

The proof is the same as for Theorem 5.1, except that when we take a partition of $[0, 1]$ into n intervals we choose x_1, \dots, x_m as points of the partition.

Example. If $q(x, y, z) = dz$, $\lambda(x, y) = G'(y)/(1 - G(y))$ where $G(y)$ is strictly increasing, $G(-\infty) = 0$, $G(+\infty) \leq 1$, then all the conditions of Theorem 5.2 are satisfied if

$$(\alpha - k)\lambda_y \leq 0, \quad f_y - kf - \frac{\lambda_y}{y} \geq 0 \quad (f_y - kf \geq 0 \text{ when } \lambda = 0)$$

for some nonnegative constant k , and

$$(-1)^i f_x(x, y) \geq 0 \quad \text{for } x \in J_i \quad (i = 0, 1).$$

6. Further properties of the free boundary. We shall need the following condition for a point $x_0 \in [0, 1]$: for any continuous function $\theta(z) \geq 0$, $0 \leq t \leq 1$,

$$(6.1) \quad f(x_0, y) - \lambda(x_0, y) \int_0^1 q(x_0, y, dz) \theta(z) \neq 0 \quad \text{in any } y\text{-interval.}$$

THEOREM 6.1. *Let (4.1) or (4.2) be satisfied and denote by u the unique solution of the variational inequality (2.23) established in Lemma 4.1. Assume that there exists a function $\varphi(x)$, $0 \leq x \leq 1$, such that (4.10) is satisfied. Then, for any $x = x_0$ for which (6.1) holds, the function $\varphi(x)$ is continuous at $x = x_0$.*

Proof. Suppose $\varphi(x)$ is discontinuous at $x = x_0$. Then there exists a sequence x_n such that

$$x_n \rightarrow x_0, \quad \varphi(x_n) \rightarrow y_0, \quad y_0 \neq \varphi(x_0).$$

Since $u(x_0, y_0) = \lim u(x_n, \varphi(x_n)) = 0$, it follows that $y_0 > \varphi(x_0)$.

For any $y_2 < y_0$ we have $\varphi(x_n) > y_2$ if n is sufficiently large. Consequently $u(x_n, y) < 0$ if $y \leq y_2$, so that

$$-\frac{\partial u}{\partial y} + (\alpha + \lambda)u - \lambda Qu = f \quad \text{if } x = x_n, \quad y \leq y_2.$$

Integrating this equation with respect to y , $y_1 < y < y_2$, where $\varphi(x_0) < y_1 < y_2$, and noting that

$$\begin{aligned} -\int_{y_1}^{y_2} \frac{\partial u}{\partial y} dy &= -u(x_n, y_2) + u(x_n, y_1) \rightarrow 0, \\ \int_{y_1}^{y_2} (\alpha + \lambda)u dy &\rightarrow \int_{y_1}^{y_2} (\alpha + \lambda(x_0, y))u(x_0, y) dy = 0, \end{aligned}$$

we obtain the relation

$$\int_{y_1}^{y_2} [f(x_0, y) - \lambda(x_0, y) \int_0^1 q(x_0, y, dz) \theta(z)] dy = 0$$

where $\theta(z) = -u(z, 0) \geq 0$. Differentiating with respect to y_2 , we obtain a contradiction to the assumption (6.1).

Example. Assumption (6.1) is satisfied if λ, q, f are analytic in y and $f(x, y) \rightarrow \infty$ as $y \rightarrow \infty$, for each x .

Counterexample. We shall show, by a counterexample, that the condition (6.1) is essential. Let $h(x)$ be any continuous positive function for $0 \leq x \leq 1$. Let A be any closed subset of $[0, 1]$ and let $g(x)$ be a continuous function on $[0, 1]$ such that $g(x) = 0$ if $x \in A$, $g(x) > 0$ if $x \in A^c$. Define

$$u(x, y) = \begin{cases} -g(x)(y - h(x))^2 & \text{if } y < h(x), \\ 0 & \text{if } y \geq h(x). \end{cases}$$

Then u and $\partial u / \partial y$ belong to $C_b^0(E)$ and the function

$$f = -\partial u / \partial y + (\lambda + \alpha)u - \lambda Qu$$

belongs to $C_b^0(E)$. It is clear that u is the solution of the variational inequality (2.23)

and (4.10) holds with

$$\varphi(x) = \begin{cases} h(x) & \text{if } x \in A^c, \\ 0 & \text{if } x \in A. \end{cases}$$

Thus $\varphi(x)$ is discontinuous at any boundary point of A^c . In this example, the condition (6.1) is not satisfied.

Our next result is concerned with the starting point of the free boundary on $y = 0$. We shall assume:

$$(6.2) \quad q(x, y, dz) = 0 \quad \text{if } z > x.$$

$$(6.3) \quad \begin{aligned} f(x, 0) &> 0 \quad \text{if } 0 \leq x < \bar{x}, \quad f(x, 0) < 0 \quad \text{if } \bar{x} < x \leq 1, \\ f_x(\bar{x}, 0) &\text{ exists,} \end{aligned}$$

$$(6.4) \quad f(x, y) \text{ is monotone increasing in } y.$$

THEOREM 6.2. *Let the conditions of Theorem 4.2 hold and let (6.2)–(6.4) hold. Then*

$$(6.5) \quad \begin{aligned} \varphi(x) &= 0 \quad \text{if } 0 \leq x < \bar{x}, \\ \varphi(x) &> 0 \quad \text{if } \bar{x} < x \leq 1. \end{aligned}$$

Proof. From (6.3) it follows that for any $x > \bar{x}$ there is a $\delta = \delta_x > 0$ such that

$$f(x, y) < 0 \quad \text{if } 0 < y < \delta.$$

From the probabilistic interpretation of u ,

$$u(x, 0) \leq J_{x0}(\delta \wedge \nu) = E_{x0} \int_0^{\delta \wedge \nu} e^{-\alpha s} f(x_s, y_s) ds$$

where ν is the first jump of x_t . Since $x_s = x$, $y_s = s$ if $s < \nu$, it follows that $u(x, 0) < 0$, so that $\varphi(x, 0) > 0$.

To prove that $u(x, 0) = 0$ if $x < \bar{x}$, we employ the comparison function

$$w(x, y) = \begin{cases} -\gamma(x - \bar{x}) & \text{if } x > \bar{x} \\ 0 & \text{if } x \leq \bar{x}. \end{cases} \quad (\gamma > 0),$$

We want to verify that

$$(6.6) \quad -\frac{\partial w}{\partial y} + (\lambda + \alpha)w - \lambda Qw \leq f \quad \text{if } x > \bar{x},$$

$$(6.7) \quad -\lambda Qw \leq f \quad \text{if } x \leq \bar{x};$$

for then it would follow that $w \leq u$, so that $u(x, 0) = 0$ if $x < \bar{x}$ and, therefore, $\varphi(x) = 0$.

Now, (6.7) is satisfied since the left hand side is equal to zero whereas, by (6.4) the right hand side is ≥ 0 . As for (6.6), using (6.2) we see that it is sufficient to show that

$$(6.8) \quad 0 \leq f/\gamma + \alpha(x - \bar{x}) \quad \text{if } x > \bar{x}.$$

Since

$$f(x, y) \geq f(x, 0) \geq -c(x - \bar{x}) \quad (c \text{ positive constant}),$$

(6.8) is valid if we take $\gamma > \alpha/c$.

We can establish a result as in Theorem 5.1 under the assumption that (6.2) holds. (In this case q_x generally does not exist, so that Theorem 5.1 cannot be applied.)

Let us first assume that (6.3), (6.4) are satisfied, so that $u(x, y) = 0$ if $x \leq \bar{x}$. We also assume that

$$(6.9) \quad q(x, y, dz) = q(x, y, z) dz, \quad q(x, y, z) = 0 \quad \text{if } z > x;$$

for any continuous function $\theta(z) \geq 0$ and $x \in (\bar{x}, 1]$,

$$(6.10) \quad \frac{\partial}{\partial x} \int_0^x q(x, y, z) \theta(z) dz = q(x, y, x-0) \theta(x) + \int_0^x q_x(x, y, z) \theta(z) dz,$$

and

$$(6.11) \quad \begin{aligned} & \lambda_x(x, y) - \lambda(x, y) \int_0^y \lambda_x(x, \eta) d\eta \\ & + \lambda(x, y) \left[q(x, y, x-0) + \frac{\int_0^x q_x(x, y, z) \theta(z) dz}{\int_0^x q(x, y, z) \theta(z) dz} \right] \geq 0, \end{aligned}$$

provided the denominator does not vanish.

For example, if

$$(6.12) \quad \int_0^x q(x, y, z) \theta(z) dz = \frac{\alpha}{x^\alpha} \int_0^x z^{\alpha-1} \theta(z) dz \quad (\alpha \geq 1),$$

then (6.10) is satisfied, and (6.11) reduces to

$$(6.13) \quad \lambda_x(x, y) - \lambda(x, y) \int_0^y \lambda_x(x, \eta) d\eta \geq 0, \quad x > \bar{x},$$

which is certainly true if $\lambda(x, y) = \lambda(y)$ for $x > \bar{x}$.

THEOREM 6.3. *Let the assumptions (5.1), (5.3), (6.3), (6.4) and (6.9)–(6.11) hold. Then there exists a function $x = \psi(y)$ such that (5.4) is valid.*

Proof. We proceed as in the proof of Theorem 5.1. However, in contrast to the preceding proof, we shall now establish (5.10) by induction on i , where $i = 0$ corresponds to $x = \bar{x}$. Since $\zeta_0 = 0$, $\zeta_1 \leq 0$, we have $\zeta_1 \leq \zeta_0$.

Suppose next that (5.10) is true for all $i \leq j-1$. We shall prove that

$$\zeta_{j+1} \leq \zeta_j.$$

It suffices to show that (5.11), (5.12) hold with $i = j$. As before, (5.11) is a consequence of (5.1), (5.3). As for (5.12) (with $i = j$), in view of the inductive assumption it is sufficient to show that

$$(6.14) \quad \frac{\partial}{\partial x} \left\{ \lambda(x, y) \left(\int_{\bar{x}}^x q(x, y, z) \theta(z) dz \right) \exp \left[- \int_0^y \lambda(x, \eta) d\eta \right] \right\} \geq 0$$

for any function $\theta(t)$ satisfying

$$\theta(t) \geq 0 \quad \theta(t) \text{ is monotone increasing for } \bar{x} \leq t \leq x.$$

The left hand side of (6.14) is equal to

$$\begin{aligned} & \left(\lambda_x - \lambda \int_0^y \lambda_x d\eta \right) \int_{\bar{x}}^x q(x, y, z) \theta(z) dz \\ & + \lambda(x, y) \left[q(x, y, x-0) \theta(x) + \int_{\bar{x}}^x q_x(x, y, z) \theta(z) dz \right], \end{aligned}$$

and since

$$\theta(x) \geq \int_{\bar{x}}^x q(x, y, z) \theta(z) dz \geq \int_{\bar{x}}^x q(x, y, z) \theta(z) dz,$$

the inequality (6.14) is a consequence of (6.11).

Remark. Theorem 6.3 is valid when $\bar{x} = 0$; in this case we may drop the conditions (6.3), (6.4) and assume, instead, that $\lambda(0, y) \equiv 0$ (so as to ensure that $\zeta_1 \leq \zeta_0$).

7. Another Markov process motivated by queuing. Let us consider the queuing system $M/G/1$ with limit capacity N . This means that interarrival times are independent random variables, exponentially distributed with parameter λ . The service times are i.i.d. random variables with a general distribution G . There is only one server and if the total number of customers in the system reaches N then all the arrivals are rejected.

Let N_t = number of customers in the system at time t , Y_t = the elapsed time from the beginning of the service time of the customers being served. It is known that N_t is not a Markov process, but (N_t, Y_t) is a Markov process and its generator has the form

$$\begin{aligned} Ag(n, y) = & I_{n>0} \left[\frac{\partial g}{\partial y} + \mu(y)(g(n-1, 0) - g(n, 0)) \right] \\ & + I_{n \leq N} \lambda [g(n+1, y) - g(n, y)] \end{aligned}$$

if

$$\mu(y) = \frac{G'(y)}{1 - G(y)}.$$

We are going to consider a similar generator with the first component in $[0, 1]$ instead of the discrete set $\{0, 1, \dots, N\}$, namely

$$\begin{aligned} (7.1) \quad Ag(x, y) = & \frac{\partial g}{\partial y} + \lambda(x) \left[\int_x^1 q_1(x, dz) g(z, y) - g(x, y) \right] \\ & + \mu(x, y) \left[\int_0^x q_2(x, dz) g(z, 0) - g(x, y) \right]. \end{aligned}$$

We begin with a lemma proving that one can associate a Markov process with the generator (7.1). The following assumptions will be needed:

$$(7.2) \quad \begin{aligned} 0 \leq \lambda(x) \leq M, \lambda \text{ is continuous on } [0, 1], \lambda(1) = 0, \\ 0 \leq \mu(x, y) \leq M, \mu \text{ is continuous on } E, \mu(0, y) \equiv 0, \end{aligned}$$

$$(7.3) \quad \begin{aligned} q_1(x, \Gamma) \text{ is a transition probability from } [0, 1] \text{ into } [x, 1], \\ q_2(x, \Gamma) \text{ is a transition probability from } [0, 1] \text{ into } [0, x], \\ q_1(1, \Gamma) = I_\Gamma(1), q_2(0, \Gamma) = I_\Gamma(0) \text{ for any } \Gamma; \\ \int_x^1 q_1(x, dz) g(z) \text{ and } \int_0^x q_2(x, dz) g(z) \text{ are continuous on } [0, 1], \end{aligned}$$

and

$$\begin{aligned} (7.4) \quad & \int_0^1 \int_0^\infty e^{-\gamma y} \mu^2(x, y) \left(\int_0^x q_2(x, dz) g(z) \right)^2 dx dy \leq C \int_0^1 g^2(z) dz, \\ & \int_0^1 \lambda^2(x) \left(\int_x^1 q_1(x, dz) g(z) \right)^2 dx \leq C \int_0^1 g^2(z) dz \\ & \text{for some } C > 0, \gamma > 0. \end{aligned}$$

LEMMA 7.1. *Under the assumptions (7.2)–(7.4) there exists a right continuous, left limited strong Markov process such that its semi-group is given by*

$$\begin{aligned}
 \Phi(t)g(x, y) = & e^{-\lambda(x)t} \exp \left[- \int_y^{y+t} \mu(x, r) dr \right] g(x, y+t) \\
 & + \int_0^t \lambda(x) e^{-\lambda(x)\sigma} \exp \left[- \int_y^{y+\sigma} \mu(x, \eta) d\eta \right] \\
 & \cdot \int_x^1 q_1(x, dz) \Phi(t-\sigma)g(z, y+\sigma) d\sigma \\
 & + \int_0^t \mu(x, y+\sigma) e^{-\lambda(x)\sigma} \exp \left[- \int_y^{y+\sigma} \mu(x, \eta) d\eta \right] \\
 & \cdot \int_0^x q_2(x, dz) \Phi(t-\sigma)g(z, 0) d\sigma
 \end{aligned}
 \tag{7.5}$$

and is Feller in the Euclidean topology. Further, if g and $\partial g / \partial y$ are bounded continuous functions on E then g belongs to the domain of the weak infinitesimal generator A and Ag is given by (7.1).

Proof. Formally, the process we want to obtain is such that on a small interval $[0, h]$ the probability of an upward jump is $\lambda(x)h$ given that the present state is (x, y) , $q_1(x, \Gamma)$ giving the distribution of x_h , and the probability of a downward jump is $\mu(x, y)h$, $q_2(x, \Gamma)$ giving the distribution of x_h . The y -component increases like $y+t$ between downward jumps of x_t .

Now, such a process can be constructed by solving a martingale problem as in Stroock [13] (who used the technique of [14]) on $\Omega = D([0, \infty); E)$, the space of right-continuous, left limited functions. The only difference here is that instead of the recursive definition of $P^{(n+1)}$ given in [13, p. 310] we now have the recursive formula

$$P_{xy}^{(n+1)}(A) = \int_{\Omega \times R^+ \times E} (\delta_\omega \otimes_t P_{z\eta}^{(n)})(A) \cdot \tilde{Q}_{xy}(d\omega dt dz d\eta) \quad (n \geq 1)$$

where

$$\begin{aligned}
 \tilde{Q}_{xy} = & \delta_{\varphi_{xy}}(\omega) \cdot \left[\lambda(x) e^{-\lambda(x)t} \exp \left[- \int_y^{y+t} \mu(x, r) dr \right] dt \cdot q_1(x, dz) \cdot \delta_{y+t}(\eta) \right. \\
 & \left. + \mu(x, y+t) \exp \left[- \int_y^{y+t} \mu(x, r) dr \right] e^{-\lambda(x)t} dt \cdot q_2(x, dz) \delta_0(\eta) \right],
 \end{aligned}$$

with $\delta_{\varphi_{xy}}$ = mass one on the trajectory (in Ω) $\varphi_{xy}(t) = (x, y+t)$ for $t \geq 0$, and $\delta_{y+t}(\delta_0)$ means mass one on the point $y+t$ (0) in R^+ , and

$P_{xy}^{(1)}$ is the probability on Ω induced by the projection on Ω of the probability (on $\Omega \times R^+$),

$$Q_{xy} = \delta_{\varphi_{xy}} \times F_{xy}, \quad F_{xy}([t, \infty)) = \exp \left[- \int_0^t (\lambda(x) + \mu(x, y+\sigma)) d\sigma \right].$$

Then one can construct a strong Markov process which is, of course, right continuous and left limited.

As in § 1 we can obtain the formula (7.5) and then complete the proof of Lemma 7.1.

Remark. One can show that for any compact subset K of E ,

$$\lim_{t \rightarrow 0} \sup_{(x,y) \in K} P(t, x, y, \Gamma) = 0$$

where Γ is the complement of any neighborhood of (x, y) . Therefore, by [5, Thm. 3.13], the process is quasi left continuous. It follows, by [11, Chap. 1] that there exists an optimal stopping time for the problem (1.13), given by

$$\hat{\tau} = \inf \{s: u(x_s, y_s) = 0\}.$$

8. The stopping time problem. We consider the stopping time problem (1.13) where $J_{xy}(\tau)$ is defined by (1.12) and (x_t, y_t) is the process constructed in § 7; the conditions (7.2)–(7.4) are always assumed in this section.

Introduce

$$Q_1 g(x, y) = \int_x^1 q_1(x, dz) g(z, y),$$

$$Q_2 g(x) = \int_0^x q_2(x, dz) g(z, 0)$$

and consider the variational inequality:

$$-\frac{\partial u}{\partial y} + (\lambda + \mu + \alpha)u - \lambda Q_1 u - \mu Q_2 u \leq f$$

a.e. in y for each x ,

$$(8.1) \quad u \leq 0 \quad \text{in } E,$$

$$\left(-\frac{\partial u}{\partial y} + (\lambda + \mu + \alpha)u - \lambda Q_1 u - \mu Q_2 u - f \right) u = 0$$

a.e. in y for each x .

The following theorem can be proved exactly as Theorem 2.4.

THEOREM 8.1. *For any $f \in C_b^0(E)$ there exists a unique solution $u \in C_b^0(E)$ of (8.1) which is uniformly Lipschitz continuous in y ; $u(x, y)$ is given by (1.13) (for the present process (x_t, y_t)).*

The results of § 4 can also be extended. Analogously to Lemma 4.1 we have (with minor changes in the proof):

LEMMA 8.2. *Assume that for some positive constants ε, m, y_0 ,*

$$(8.2) \quad \begin{aligned} f(x, y) &\geq -m \quad \text{in } E, \\ f(x, y) &\geq \varepsilon + (2m/\alpha)M \quad \text{if } y \geq y_0. \end{aligned}$$

Then there exists a solution u of (8.1) which belongs to $C_b^0(E)$ and which is uniformly Lipschitz continuous in y . Further,

(i) *the solution is unique if either f is bounded or if*

$$(8.3) \quad \inf_{0 \leq x \leq 1} f(x, y) \rightarrow \infty \quad \text{as } y \rightarrow \infty;$$

(ii) the solution has a compact support.

We now assume:

$$\begin{aligned}
 (8.4) \quad & \mu_y \text{ and } f_y \text{ exist and are continuous;} \\
 & \text{at any point where } \mu = 0 \text{ also } \mu_y = 0; \\
 & \mu_y \geq 0, \text{ and } \mu_y/\mu \text{ is bounded on the set above } \mu \neq 0; \\
 & f_y - (2M + \alpha)f - Q_1f - (\mu_y/\mu)f \geq 0 \text{ on the set where } \mu \neq 0; \\
 & f_y - (2M + \alpha)f - Q_1f \geq 0 \text{ on the set where } \mu = 0.
 \end{aligned}$$

THEOREM 8.3. Assume that (8.3) and (8.4) hold. Then there exists a bounded function $\varphi(x)$ such that

$$\begin{aligned}
 (8.5) \quad & u(x, y) < 0 \quad \text{if } y < \varphi(x), \\
 & u(x, y) = 0 \quad \text{if } y \geq \varphi(x).
 \end{aligned}$$

Proof. We begin as in the proof of Theorem 4.2. Setting $u = e^{ky}w$ we get, analogously to (4.12),

$$(8.6) \quad -\frac{\partial w}{\partial y} + (\lambda + \alpha + \mu - k)w - \lambda Q_1w - \mu e^{-ky}Q_2w = f e^{-ky}.$$

Let $\zeta = \partial w / \partial y$ and differentiate (8.6) with respect to y . Then,

$$\begin{aligned}
 (8.7) \quad & -\frac{\partial \zeta}{\partial y} + (\lambda + \alpha + \mu - k)w + \mu_y(w - e^{-ky}Q_2w) \\
 & = \lambda Q_1\zeta - k\mu e^{-ky}Q_2w - k e^{-ky}f + e^{-ky}f_y.
 \end{aligned}$$

When $\mu = 0$ we have $\mu_y = 0$, and when $\mu \neq 0$ one can substitute $(w - e^{-ky}Q_2w)$ from (8.6) into (8.7). We obtain

$$\begin{aligned}
 (8.8) \quad & -\frac{\partial \zeta}{\partial y} + \left(\lambda + \alpha - k + \mu + \frac{\mu_y}{\mu} \right) \zeta = \lambda Q_1\zeta + \frac{\mu_y}{\mu} (\lambda + \alpha - k)w \\
 & - \frac{\mu_y}{\mu} Q_1w + e^{-ky} \left[f_y - k e^{-ky}f - \frac{\mu_y}{\mu} f \right] \\
 & - k\mu e^{-ky}Q_2w \equiv \tilde{f},
 \end{aligned}$$

where the terms with μ_y/μ drop out when $\mu = 0$.

Notice that

$$Q_1\zeta(x, y) = \int_x^1 q_1(x, dz)\zeta(z, y)$$

and that the variational inequality for $u(z, y)$ gives

$$\zeta(z, y) \geq (\lambda + \alpha - k + \mu)w - \lambda Q_1w - \mu e^{-ky}Q_2w - f(z, y) e^{-ky}.$$

Applying Q_1 to this inequality and substituting the result in the $\lambda Q_1\zeta$ occurring in (8.8), and then choosing

$$k = 2M + \alpha$$

we deduce, using (8.4), that $\tilde{f} \geq 0$. We can now complete the proof as before.

We shall next require the additional assumptions:

$$\begin{aligned}
 & q_i(x, dz) = q^i(x, z) dz \quad (i = 1, 2); \\
 & \lambda_x, \mu_x, f_x \text{ and } q_x^1, q_x^2 \text{ are continuous;} \\
 & \text{for any continuous and positive function } \theta, \\
 & \lambda_x Q_1 \theta + \lambda Q_{1x} \theta + \lambda q^1(x, x+0) \theta(x) \geq 0, \\
 (8.9) \quad & \mu_x Q_2 \theta + \mu Q_{2x} \theta + \mu q^2(x, x-0) \theta(x) - \mu \left[\int_0^y \mu_x(\eta) d\eta + \lambda_x y \right] Q_2 \theta \geq 0
 \end{aligned}$$

where

$$\begin{aligned}
 Q_{1x} \theta &= \int_0^x q_x^1(x, z) \theta(z) dz, \quad Q_{2x} \theta = \int_x^1 q_x^2(x, z) \theta(z) dz; \\
 f_x - f \left[\int_0^y \mu_x(\eta) d\eta + \lambda_x y \right] &\leq 0.
 \end{aligned}$$

THEOREM 8.4. Assume that (8.3) and (8.9) hold. Then there exists a function $\psi(y)$ such that

$$\begin{aligned}
 (8.10) \quad & u(x, y) < 0 \quad \text{if } x > \psi(y), \\
 & u(x, y) = 0 \quad \text{if } x \leq \psi(y).
 \end{aligned}$$

Thus, if the condition (8.4) is also satisfied, then $y = \varphi(x)$ is a monotone increasing curve.

The proof of Theorem 8.4 is similar to the proof of Theorem 5.1. It is based on the approximation procedure of § 3 which is valid also for the present case, with minor changes.

An analogue of Theorem 5.2 can also be established, so that (under suitable conditions) $\varphi(x)$ changes the direction of monotonicity a given number of times.

Remark 1. The conditions made on f, λ, q in Theorem 4.2 and throughout §§ 5, 6 are actually only needed for $y \leq y_0$, where y_0 is such that $u(x, y) \equiv 0$ if $y \geq y_0$. A similar remark applies to § 8.

Remark 2. In this paper we have considered only the constraint $u \leq 0$. However, the more general case of $u \leq \varphi$ with φ, φ_y in $C_b^0(E)$ can be immediately reduced to the previous case with f replaced by $f + A\varphi - \alpha\varphi$ and u replaced by $u - \varphi$.

Remark 3. The results of this paper (except for the x -graph properties) can be extended to the case where x varies in any fixed compact subset $K \subset R^n$. The x -graph properties can be extended in case K is rectangular.

REFERENCES

- [1] M. ALAM AND V. V. S. SARMA, *Optimal maintenance and replacement via semi Markov decision model*, Internat. J. on Systems Sci., 6 (1975), pp. 809–818.
- [2] A. BENSOUSSAN AND J. L. LIONS, *Temps d'arrêt et contrôle impulsionnel: Inéquations variationnelles et quasi variationnelles d'évolution*, Cahier de Math. de la Décision, no. 7523, Univ. Paris IX, 1975.
- [3] H. BREZIS AND A. FRIEDMAN, *Estimates on the support of solutions of parabolic variational inequalities*, Illinois J. Math., 20 (1976), pp. 82–87.
- [4] L. A. CAFFARELLI, *The regularity of the free boundaries in higher dimensions*, Acta Math., to appear.
- [5] E. B. DYNKIN, *Markov Processes*, vol. 1, Springer-Verlag, Berlin, 1965.
- [6] A. FRIEDMAN, *Parabolic variational inequalities in one-space dimension and smoothness of the free boundary*, J. Functional Analysis, 18 (1975), pp. 151–176.

- [7] A. FRIEDMAN AND D. KINDERLEHRER, *A one-phase Stefan problem*, Indiana Univ. Math. J., 24 (1975), pp. 1005–1035.
- [8] I. I. GIKHMAN AND A. V. SKOROKHOD, *The Theory of Stochastic Processes*, vol. 2, Springer-Verlag, Berlin, 1975.
- [9] T. LAETSCH, *A uniqueness result of elliptic quasi variational inequalities*, J. Functional Analysis, 18 (1975), pp. 286–287.
- [10] U. PRABLAU AND S. STIDHAM, *Optimal control of queueing systems*, Mathematical Methods in Queueing Theory, Lecture Notes in Economics and Math. Systems, no. 98, Springer-Verlag, Berlin, 1974.
- [11] M. ROBIN, *Contrôle impulsionnel des processus de Markov*, Thesis, IRIA, Rocquencourt, France, 1977.
- [12] L. STONE, *Necessary and sufficient conditions for optimal control of semi-Markov jump process*, this Journal, 11 (1973), pp. 187–201.
- [13] D. W. STROOCK, *Some stochastic processes which arise from a model of a motion of a bacterium*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 28 (1974), pp. 305–315.
- [14] D. W. STROOCK AND S. R. S. VARADHAM, *Diffusion processes with continuous coefficients*, Comm. Pure Appl. Math., 22 (1969), pp. 345–400.
- [15] ———, *On degenerate elliptic-parabolic operators of second order and their associated diffusions*, Ibid., 25 (1972), pp. 651–713.

A NOTE ON WEAK STABILIZABILITY OF CONTRACTION SEMIGROUPS*

CLAUDE D. BENCHIMOL†

Abstract. A recent result on weak stabilizability is that the system $\dot{x} = Ax + Bu$, where A is the infinitesimal generator of a contraction semigroup over a Hilbert space H , and B is linear bounded, is weakly stabilizable if: (i) A has a compact resolvent and (ii) (A, B) is (approximately) controllable. In this note, we show that condition (i) is superfluous and (ii) can be weakened to (iii), the weakly unstable states are (approximately) controllable, which actually turns out to be a necessary condition. Indeed, if (i) is verified, (iii) is necessary and sufficient for strong stabilizability. Moreover, we give a simple, direct proof, using semigroup theoretic techniques, in particular obviating the need to invoke the “LaSalle invariance principle”. The main tool is a decomposition applicable to all contraction semigroups which is derived from results of Sz.-Nagy, C. Foias and S. R. Foguel.

1. Introduction. A standard result (see Wonham [9]) in finite dimension, is that the time invariant system

$$(1.1) \quad \dot{x} = Ax + Bu$$

is stabilizable by feedback

$$u = Kx$$

if and only if the unstable modes of the system are controllable. The extension of this result to infinite dimensional systems, i.e., where A is the infinitesimal generator of a C_0 semigroup, and B is linear and bounded has been the subject of many investigations recently [6], [8], [10], [11]. As may be expected, there are many nonequivalent notions of stability, depending on the topology used. The notion of “weak stability” would appear to be the weakest. Thus

DEFINITION 1.1. A C_0 semigroup $T(t)$ over a Hilbert space H is *weakly stable* if $\forall x, y \in H, (T(t)x, y) \rightarrow 0$ as $t \rightarrow +\infty$.

Slemrod [6, Thm. 3.5] shows that if A generates a C_0 contraction semigroup $T(t)$ over a Hilbert space H and B is a linear bounded transformation mapping a Hilbert space H_i into H , the semigroup generated by $A - BB^*$ is weakly stable provided:

- (i) A has a compact resolvent (note that for some reason this condition is stated in terms of A^* in [6], although of course the two are equivalent);
- (ii) (A, B) is (approximately) controllable.

In his proof, he uses the “LaSalle invariance principle”. We shall show (Theorem 3.1) that the assumption (i) is superfluous (and in fact is sufficient to yield strong stability) and (ii) can be considerably weakened. Moreover, our techniques are simpler and more directly semigroup theoretic, relying on a functional decomposition of contraction semigroups, following Sz.-Nagy–Foias. We also incidentally indicate the relevance of the Sz.-Nagy–Foias theory [7] to the whole problem.

We begin with some results of interest on their own.

2. Canonical decomposition for contraction semigroups. In this section, we state two decomposition theorems for C_0 contraction semigroups, and merge them into one corollary. First, we recall some definitions:

* Received by the editors April 5, 1977, and in revised form June 29, 1977.

† System Science Department, University of California at Los Angeles, California 90024. This research was supported in part by the United States Air Force Office of Scientific Research, Applied Mathematics Division, under Grant 73-2492.

DEFINITION 2.1. Let H be a Hilbert space, and V be a bounded operator in H . We say that a subspace K *reduces* V if and only if

$$(2.1) \quad VK \subseteq K \quad \text{and} \quad V^*K \subseteq K.$$

DEFINITION 2.2. A bounded operator V in H is

(i) *Unitary* if

$$V^*V = VV^* = I,$$

(ii) *Completely non unitary* (c.n.u.) if there exists no subspace other than $\{0\}$ reducing V to a unitary operator.

Remark. It follows from (2.1) that both K and K^\perp reduce V and V^* .

THEOREM 2.1 [Sz.-Nagy-Foias]. Let $T(t)$ be a C_0 contraction semigroup in a Hilbert space H . Then H can be decomposed into an orthogonal sum $H = H_u \oplus H_{\text{cnu}}$ where H_u and H_{cnu} are reducing subspaces for $T(t)$, such that

- (i) The restriction $T_u(t) = T(t)|_{H_u}$ of $T(t)$ to H_u is a unitary group.
- (ii) The restriction $T_{\text{cnu}}(t) = T(t)|_{H_{\text{cnu}}}$ of $T(t)$ to H_{cnu} is a c.n.u. semigroup.
- (iii) This decomposition (where of course H_u or H_{cnu} can be trivial) is unique and H_u can be characterized by

$$(2.2) \quad H_u = \{x \in H; \forall t \geq 0, \|T(t)x\| = \|T^*(t)x\| = \|x\|\}.$$

Moreover

$$(2.3) \quad H_u = \bar{K}_u,$$

where $K_u = \mathcal{D}(A) \cap H_u$, and A denotes the infinitesimal generator of $T(t)$.

Proof. For the sake of completeness we sketch a proof. For more details see Sz.-Nagy and Foias [7, pp. 9–10 and 136].

If we denote $D_{T(t)} = I - T^*(t)T(t)$ and $D_{T^*(t)} = I - T(t)T^*(t)$, then $\|T(t)x\| = \|T^*(t)x\| = \|x\|$ is equivalent to

$$(2.4) \quad (D_{T(t)}x, x) = (D_{T^*(t)}x, x) = 0.$$

Since $T(t)$ is a contraction, $D_{T(t)}$ and $D_{T^*(t)}$ are both self adjoint nonnegative definite. It follows that $(2.4) \Leftrightarrow x \in \mathcal{N}(D_{T(t)}) \cap \mathcal{N}(D_{T^*(t)})$ where $\mathcal{N}(\cdot)$ stands for the null space of an operator. Therefore $H_u = \bigcap_{t>0} [\mathcal{N}(D_{T(t)}) \cap \mathcal{N}(D_{T^*(t)})]$ which shows that it is closed. Using (2.2), it is easy to see that H_u is left invariant under $T(s)$ and $T^*(s)$ for any s . To show (2.3) (which is *not* specifically contained in [7]), we first note that since H_u is closed,

$$(2.5) \quad \bar{K}_u = \overline{\mathcal{D}(A) \cap H_u} \subseteq H_u.$$

Then, for any x in H_u , $R(\lambda, A)x = \int_0^\infty e^{-\lambda s} T(s)x ds$ is also in H_u , as easily proved by checking that

$$T(t)T^*(t)R(\lambda, A)x = T^*(t)T(t)R(\lambda, A)x = R(\lambda, A)x.$$

But $R(\lambda, A)x \in \mathcal{D}(A)$. Therefore $\forall x \in H_u$, $\lambda R(\lambda, A)x \in K_u$. But we know that $\lambda R(\lambda, A)x \rightarrow x$ as $\text{Re } \lambda \rightarrow +\infty$ [1, p. 169]. Hence $H_u \subseteq \bar{K}_u$, which finally implies $H_u = \bar{K}_u$, associated with (2.5).

The following decomposition theorem is due to Foguel [4]. Here, we give a simple proof, using mainly elementary properties of contraction semigroups; this in turn shows the relevance of the contraction assumption to the stability problem.

THEOREM 2.2 [Foguel]. Let $T(t)$ be a C_0 contraction semigroup in a Hilbert space H . Let $W = \{x \in H; T(t)x \rightarrow 0 \text{ (weakly) as } t \rightarrow +\infty\}$. Then

- (i) W reduces $T(s)$ for any s .
- (ii) On W^\perp , $T(s)$ is reduced to a unitary group, that is $W^\perp \subseteq H_u$.
- (iii) W coincides with the subspace $\{x \in H; T^*(t)x \rightarrow 0 \text{ (weakly) as } t \rightarrow \infty\}$.

Proof. First, note that W is a closed subspace of H .

Proof of (i). Let $x \in W$. Then, for any $s \geq 0$,

$$(2.6) \quad T(t)T(s)x = T(s+t)x \rightarrow 0 \text{ (weakly) as } t \rightarrow +\infty,$$

and hence $T(s)x \in W$. In order to prove that $T^*(s)x \in W$, we need an intermediate result proved below.

$T^*(t)$ being a contraction, we have $\forall x \in H, \forall t_2 \geq t_1$

$$\|T^*(t_2)x\|^2 = \|T^*(t_2 - t_1)T^*(t_1)x\|^2.$$

Therefore, for any x , $\|T^*(t)x\|^2$ is a nonincreasing function of t , bounded from below by 0. Hence, it converges as $t \rightarrow +\infty$. Therefore, for any fixed s

$$Z(t) = \|T^*(t)x\|^2 - \|T^*(t+s)x\|^2 \rightarrow 0 \text{ as } t \rightarrow +\infty.$$

But

$$\begin{aligned} Z(t) &= (T^*(t)x, T^*(t)x) - (T(s)T^*(t)x, T^*(t)x) \\ &= ([I - T(s)T^*(s)]T^*(t)x, T^*(t)x) \\ &= \|[I - T(s)T^*(s)]^{1/2}T^*(t)x\|^2. \end{aligned}$$

Hence $\forall x \in H, \forall s \geq 0$

$$(2.7) \quad [I - T(s)T^*(s)]T^*(t)x \rightarrow 0 \text{ as } t \rightarrow +\infty.$$

Multiplying to the left by $T^*(s)$, we have

$$T^*(s)[I - T(s)T^*(s)]T^*(t)x = [I - T^*(s)T(s)]T^*(t+s)x.$$

Therefore, it follows that $\forall x \in H, \forall s \geq 0$

$$(2.8) \quad [I - T^*(s)T(s)]T^*(t)x \rightarrow 0 \text{ as } t \rightarrow \infty.$$

Next, we use the fact that if $Y(t)$ is a bounded linear operator such that $Y(t)x \rightarrow 0$ for any x in H , then $Y^*(t)x \rightarrow 0$ (weakly), for any x in H . Applying this to (2.7) and (2.8), we get $\forall x \in H, \forall s \geq 0$

$$(2.9) \quad T(t)[I - T(s)T^*(s)]x \rightarrow 0 \text{ (weakly) as } t \rightarrow +\infty,$$

$$(2.10) \quad T(t)[I - T^*(s)T(s)]x \rightarrow 0 \text{ (weakly) as } t \rightarrow +\infty.$$

Now, we are ready to complete the proof of (i). For, if we take x in W , we have $T(t)x \rightarrow 0$ (weakly) as $t \rightarrow +\infty$, and subtracting it from (2.9), we get

$$(2.11) \quad \begin{aligned} -T(t)T(s)T^*(s)x &= -T(t+s)T^*(s)x \rightarrow 0 \text{ (weakly) as } t \rightarrow +\infty \\ &\Rightarrow T(t)T^*(s)x \rightarrow 0 \text{ (weakly) as } t \rightarrow +\infty. \end{aligned}$$

Grouping (2.6) and (2.11), we have

$$x \in W \Rightarrow \begin{cases} T(s)x \in W, \\ T^*(s)x \in W. \end{cases}$$

Therefore W and W^\perp reduce $T(s)$, for any s .

Proof of (ii). Statements (2.9) and (2.10) can be interpreted as:

$$(2.12) \quad \text{For any } s \begin{cases} \mathcal{R}(D_{T^*(s)}) = \text{Range}[I - T(s)T^*(s)] \subseteq W, \\ \mathcal{R}(D_{T(s)}) = \text{Range}[I - T^*(s)T(s)] \subseteq W. \end{cases}$$

$D_{T^*(s)}$ and $D_{T(s)}$ being self adjoint, we have

$$\begin{aligned} \mathcal{R}(D_{T^*(s)})^\perp &= \mathcal{N}(D_{T^*(s)}), \\ \mathcal{R}(D_{T(s)})^\perp &= \mathcal{N}(D_{T(s)}). \end{aligned}$$

Therefore (2.12) is equivalent to: $\forall s \geq 0$

$$\begin{aligned} \text{or} \quad W^\perp &\subseteq \mathcal{N}(D_{T(s)}) \cap \mathcal{N}(D_{T^*(s)}), \\ W^\perp &\subseteq \bigcap_{s \geq 0} [\mathcal{N}(D_{T(s)}) \cap \mathcal{N}(D_{T^*(s)})] = H_u. \end{aligned}$$

This completes the proof of (ii).

Proof of (iii). Let $x \in W$. Any y in H can be uniquely decomposed as $y = y_W + y_{W^\perp}$ where $y_W \in W$ and $y_{W^\perp} \in W^\perp$.

Then $(T^*(t)x, y) = (T^*(t)x, y_W + y_{W^\perp}) = (x, T(t)y_W) + (x, T(t)y_{W^\perp})$. Since W^\perp reduces $T(t)$, $T(t)y_{W^\perp} \in W^\perp$ and $(x, T(t)y_{W^\perp}) = 0$. But since $y_W \in W$, $T(t)y_W \rightarrow 0$ (weakly) as $t \rightarrow +\infty$ and $(T^*(t)x, y) = (x, T(t)y_W) \rightarrow 0$ as $t \rightarrow +\infty$ and $T^*(t)x \rightarrow 0$ (weakly) as $t \rightarrow +\infty$. Reversing the role of $T(t)$ and $T^*(t)$, we can show that $T^*(t)x \rightarrow 0$ (weakly) $\Rightarrow T(t)x \rightarrow 0$ (weakly), which completes the proof. We can unite the two theorems into the following corollary.

COROLLARY 2.1. *Let H be a Hilbert space, and $T(t)$ a C_0 contraction semigroup in H . Then H can be decomposed into three orthogonal subspaces H_{cnu} , W_u and W^\perp , all reducing $T(t)$ and $T^*(t)$, such that*

$$\begin{aligned} W_u \oplus W^\perp &= H_u, \\ W_u \oplus H_{\text{cnu}} &= W \quad (\text{with the above notations}). \end{aligned}$$

It follows that

- On H_{cnu} , $T(t)$ is completely nonunitary, and weakly stable;
- On W_u , $T(t)$ is unitary and weakly stable;
- On W^\perp , $T(t)$ is unitary, and $\forall x \in W^\perp$, $T(t)x \not\rightarrow 0$ and $T^*(t)x \not\rightarrow 0$ as $t \rightarrow +\infty$.

Proof. The proof follows immediately from the two theorems.

The above result motivates the following definition:

DEFINITION 2.3. Let $T(t)$ be a C_0 contraction semigroup over a Hilbert space H . Then $W = \{x; T(t)x \rightarrow 0 \text{ (weakly) as } t \rightarrow +\infty\}$ is called the *weakly stable subspace*. W^\perp is called the “*weakly unstable subspace*” and elements of W^\perp are called “*weakly unstable states*”. [Of course $T(t)x \not\rightarrow 0$ as $t \rightarrow +\infty$ does not imply that $x \in W^\perp$.]

3. Necessary and sufficient condition for weak stabilizability of C_0 contraction semigroups. In order to prove the main theorem of this section, we need some preliminary results. First, we recall what is meant by “controllability”.

DEFINITION 3.1. Consider the system

$$(3.1) \quad \dot{x} = Ax + Bu$$

where A generates a C_0 semigroup $T(t)$ over a Hilbert space H and B is a linear bounded operator mapping another Hilbert space H_i into H . The set C of x in H , for

which given any $\varepsilon > 0$, there exist a $t > 0$ and $u(\cdot)$ in $L_2[(0, t); H_i]$ such that

$$(3.2) \quad \left\| x - \int_0^t T(t-\sigma)Bu(\sigma) d\sigma \right\| < \varepsilon$$

is called the set of (approximately) controllable states. If $C = H$, the system is approximately controllable. See [1].

LEMMA 3.1. *With the above notations, C is a closed subspace and can be characterized by*

$$(3.3) \quad C = \overline{\bigcup_{t \geq 0} \text{Range}[T(t)B]}.$$

It follows that

$$(3.4) \quad C^\perp = \bigcap_{t \geq 0} \mathcal{N}[B^*T^*(t)].$$

C (resp. C^\perp) is called the (approximately) controllable (resp. uncontrollable) subspace.

Proof. See [1, pp. 207–210].

Next, we state two perturbation results.

LEMMA 3.2. *Let A be the infinitesimal operator of a C_0 semigroup $T(t)$ in a Hilbert space H , and D be a bounded operator in H . Then $A + D$ generates a C_0 semigroup $S(t)$ in H . Furthermore,*

- (i) *If A and D are self adjoint, so is $S(t)$, for any $t \geq 0$.*
- (ii) *If A and D are dissipative, $S(t)$ is a contraction semigroup.*
- (iii) *If A has a compact resolvent, so does $A + D$.*
- (iv) *If $T(t)$ is compact, for any $t > 0$, so is $S(t)$.*

Proof. See [1, pp. 220–225].

LEMMA 3.3. *Let K be any bounded operator mapping a Hilbert space H_i into H . Let $S(t)$ denote the semigroup generated by $A + BK$. Then $\forall t \geq 0$, $B^*T^*(t)x = 0$ if and only if $\forall t \geq 0$, $B^*S^*(t)x = 0$. (The (approximately) controllable subspace of (A, B) coincides with the one of $(A + BK, B)$.)*

Proof. Follows immediately from the identities

$$(3.5) \quad S^*(t)x = T^*(t)x + \int_0^t T^*(t-\sigma)K^*B^*S^*(\sigma)x d\sigma$$

and

$$(3.6) \quad T^*(t)x = S^*(t)x - \int_0^t S^*(t-\sigma)K^*B^*T^*(\sigma)x d\sigma$$

THEOREM 3.1. *Let A be the infinitesimal generator of a C_0 contraction semigroup $T(t)$ in a Hilbert space H , and B a bounded operator mapping another Hilbert space H_i into H . Then, the system $\dot{x} = Ax + Bu$ is weakly stabilizable if and only if the “weakly unstable states” of $T(t)$ are (approximately) controllable, and $K = -B^*$ is a stabilizing feedback gain.*

Proof. Let C be the controllable subspace of (A, B) , as defined above. Let W be the weakly stable subspace of $T(t)$, as defined in § 2. Then the theorem can be expressed as

$$(A, B) \text{ is weakly stabilizable} \Leftrightarrow W^\perp \subseteq C \Leftrightarrow C^\perp \subseteq W.$$

(i) *Necessity.* Suppose there exists a bounded operator K such that $A + BK$ generates a weakly stable semigroup $S(t)$. Then, let $x \in C^\perp$. By definition of C^\perp , we

have $\forall t \geq 0, B^*T^*(t)x = 0$. Therefore, from (3.6), we get $\forall y \in H$

$$(T^*(t)x, y) = (S^*(t)x, y) = (x, S(t)y) \rightarrow 0 \quad \text{as } t \rightarrow +\infty,$$

by assumption. Therefore $T^*(t)x \rightarrow 0$ (weakly) as $t \rightarrow +\infty$, and since $T^*(t)$ is a contraction, we can use (iii) of Theorem 2.2 to prove that $T(t)x \rightarrow 0$ (weakly) as $t \rightarrow +\infty \Rightarrow x \in W$. So $C^\perp \subseteq W$. Q.E.D.

(ii) *Sufficiency*. Assume $C^\perp \subseteq W$. Let $K = -B^*$ be the feedback gain. Then $-BB^*$ is obviously a bounded dissipative operator, and by (ii) of Lemma 3.2, $A - BB^*$ generates a contraction semigroup $S(t)$. Then, applying Theorem 2.1 to $S(t)$, we obtain a decomposition of H into two orthogonal subspaces H_u^s , reducing $S(t)$ to a unitary group, and H_{cnu}^s , reducing $S(t)$ to a c.n.u. semigroup. Then, by Corollary 2.1, we have $\forall x \in H_{\text{cnu}}^s$,

$$S(t)x \rightarrow 0 \text{ (weakly) as } t \rightarrow +\infty.$$

Therefore, it only remains to prove that $S(t)$ is weakly stable on H_u^s . Define K_u^s as in Theorem 2.1. Then, for any x in $K_u^s \subseteq \mathcal{D}(A)$ we have $\forall t \geq 0$

$$\frac{d}{dt} \|S^*(t)x\|^2 = ((A^* - BB^*)S^*(t)x, S^*(t)x) + (S^*(t)x, (A^* - BB^*)S^*(t)x) = 0.$$

Since A^* and $-BB^*$ are dissipative, the above equation implies that $\forall t \geq 0, B^*S^*(t)x = 0$. But, by Lemma 3.3, this implies that $\forall t, B^*T^*(t)x = 0$ or equivalently $x \in C^\perp$. So

$$(3.7) \quad x \in K_u^s \Rightarrow x \in C^\perp.$$

But by assumption $C^\perp \subseteq W$. Therefore

$$(3.8) \quad x \in K_u^s \Rightarrow x \in W.$$

Using (3.6) and (3.7), we get:

$$\forall t \geq 0, \quad \forall x \in K_u^s, \quad S^*(t)x = T^*(t)x.$$

Since $x \in W, T^*(t)x \rightarrow 0$ (weakly) as $t \rightarrow +\infty$, by (iii) of Theorem 2.2. So does $S^*(t)x$, and so does $S(t)x$ by the same argument. Therefore $\forall x \in K_u^s, S(t)x \rightarrow 0$ (weakly) as $t \rightarrow +\infty$. Since K_u^s is dense in H_u^s (Theorem 2.1), and $\|S(t)\| \leq 1$, then, for any x in $H_u^s, S(t)x \rightarrow 0$ (weakly) as $t \rightarrow +\infty$, by the triangle inequality. This completes the proof.

COROLLARY 3.1. *If A generates a C_0 contraction semigroup and has a compact resolvent, the condition of Theorem 3.1 is necessary and sufficient for the strong stabilizability of (A, B) . In particular $A - BB^*$ generates a strongly stable semigroup.*

Proof.

(i) *Necessity*. Follows from the fact that strong stability \Rightarrow weak stability.

(ii) *Sufficiency*. From (iii) of Lemma 3.2, $A - BB^*$ has a compact resolvent $R(\lambda, A - BB^*)$ and generates a contraction semigroup $S(t)$, which is weakly stable by Theorem 3.1. Let λ_0 be a point in the resolvent set of $A - BB^*$. Then, for any x in $\mathcal{D}(A)$, there exist a y in H such that $x = R(\lambda_0, A - BB^*)y$. Then $S(t)x = S(t)R(\lambda_0, A - BB^*)y = R(\lambda_0, A - BB^*)S(t)y$. Since $\forall y \in H, S(t)y \rightarrow 0$ (weakly) as $t \rightarrow +\infty$; and since $R(\lambda_0, A - BB^*)$ is compact,

$$\forall x \in \mathcal{D}(A), \quad S(t)x \rightarrow 0 \quad \text{as } t \rightarrow +\infty.$$

Since $\mathcal{D}(A)$ is dense in H , and $\|S(t)\| \leq 1$,

$$\forall x \in H, \quad S(t)x \rightarrow 0 \quad \text{as } t \rightarrow +\infty \quad \text{Q.E.D.}$$

COROLLARY 3.2. *If A generates a C_0 contraction self adjoint semigroup, the conditions of Theorem 3.1 are necessary and sufficient for the strong stabilizability of (A, B) . In particular $A - BB^*$ generates a strongly stable self adjoint contraction semigroup.*

Proof. (i) This is in Corollary (3.1).

(ii) *Sufficiency.* By (i) and (ii) of Lemma (3.2), $A - BB^*$ generates a self adjoint contraction semigroup $S(t)$ which is weakly stable, by Theorem 3.1. Therefore $\forall x \in H$, $(S(2t)x, x) \rightarrow 0$ as $t \rightarrow +\infty$; but $(S(2t)x, x) = (S(t)S(t)x, x) = \|S(t)x\|^2 \rightarrow 0$ as $t \rightarrow +\infty$. Q.E.D.

COROLLARY 3.3. *If A generates a compact contraction semigroup, the conditions of Theorem 3.1 are necessary and sufficient for the exponential stabilizability of (A, B) . In particular, $A - BB^*$ generates an exponentially stable semigroup.*

Proof. Necessity is as before. Sufficiency follows from the fact that for a compact semigroup, weak stability \Rightarrow exponential stability. See [3].

This corollary is also a consequence of the sufficient condition proven in [8].

4. Conclusion and remarks. Triggiani [8] has given a number of counterexamples of systems which are (approximately) controllable (A.C.) but not strongly stabilizable (S.S.). This paper shows that the A.C. of the weakly unstable states implies the weak stabilizability (W.S.) of the system, provided $T(t)$ is a *contraction* semigroup.¹ In particular, wave equations, which generate unitary groups in general, can be weakly stabilized if (A, B) is A.C. and strongly stabilized if in addition, the domain happens to be compact (thus insuring the compactness of the resolvent).

For further results involving semigroups other than contractions, we refer to [2].

Acknowledgment. The author wishes to thank Dr. N. Levan for illuminating discussions on Sz.-Nagy and Foias' theory.

REFERENCES

- [1] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer-Verlag, New York, 1976.
- [2] C. D. BENCHIMOL, *The stabilizability of infinite dimensional linear time-invariant systems*, Ph.D. thesis, University of California at Los Angeles, 1977.
- [3] R. DATKO, *Uniform asymptotic stability of evolutionary processes in a Banach space*, SIAM J. Math. Anal., 3 (1972), pp. 428-445.
- [4] S. R. FOGUEL, *Powers of a contraction in Hilbert space*, Pacific J. Math., 13 (1963), pp. 551-562.
- [5] R. E. O'BRIEN, *Controllability, stabilization and mean ergodic theorems*, George Washington University and Goddard Space Flight Center, 1976.
- [6] M. SLEMROD, *A note on complete controllability and stabilizability for linear control systems in Hilbert space*, this Journal, 12 (1974), pp. 500-508.
- [7] B. SZ.-NAGY AND C. FOIAS, *Analyse Harmonique des Opérateurs de l'Espace de Hilbert*, Masson & Cie, Akadémiai Kiadó, Budapest, 1967. (French edition, Masson & Cie.)
- [8] R. TRIGGIANI, *On the stabilizability problem in Banach space*, J. Math. Anal. Appl., 52 (1975), pp. 383-403.
- [9] W. M. WONHAM, *On pole assignment in multi-input controllable linear systems*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 660-665.
- [10] J. ZABCZYK, *Remarks on the algebraic Riccati equation in Hilbert space*, Appl. Math. Optim., 2 (1976), pp. 251-258.
- [11] ———, *Complete stabilizability implies exact controllability*, Seminarul de Ecuatii Functionale, Universitatea din Timisoara, Roumania, 1976.

¹ The author was informed by one of the referees that a *sufficient* condition for weak stabilizability (A.C. \Rightarrow W.S.) was independently and simultaneously obtained by R. E. O'Brien [5]. Our result is a *necessary and sufficient* condition which shows that the system need not be A.C. on the whole space, in order to be weakly stabilized.

NONPARAMETRIC IDENTIFICATION FOR DIFFUSION PROCESSES*

G. BANON†

Abstract. It is proved that under a specific condition (so-called condition G_2) on the transition probability operator of a measurable stationary Markov process, a recursive kernel estimate of the initial density is convergent in quadratic mean.

Assumptions on the differential stochastic equations driven by Brownian motion are derived under which the stationary solution satisfies condition G_2 .

The above results are applied to solve a class of nonlinear identification problems.

1. Introduction. An important problem in control engineering is the identification of dynamical systems. In this paper we focus our interest on systems represented by the Itô stochastic differential equation. Thus far, primary emphasis has been on solving the identification problem for linear systems, and many techniques have been proposed (e.g.: Kalman and Bucy (1961), Banon (1971), Aguilar-Martin (1974), Alengrin (1974), Salut (1976)). In contrast, our interest is to develop an approach for the identification of a class of non-linear systems.

More precisely, let $\{X_t, t \in [0, \infty)\}$ be a stochastic process defined on a probability space $(\Omega, \mathcal{A}, \mathcal{P})$ and satisfying the Itô stochastic differential equation:

$$(1.1a) \quad dX_t = m(X_t) dt + \sigma(X_t) dW_t,$$

with initial condition

$$(1.1b) \quad X_0 = X.$$

In (1.1a) W_t is the Brownian motion defined on the same probability space $(\Omega, \mathcal{A}, \mathcal{P})$.

Henceforth we shall make the following assumptions relative to the problem (1.1):

A_1 : The initial random variable X is defined on $(\Omega, \mathcal{A}, \mathcal{P})$ and is of second order: $EX^2 < \infty$.

A_2 : $m(\cdot)$ and $\sigma(\cdot)$ are Borel measurable functions on R satisfying, for $x, y \in R$, the uniform Lipschitz condition:

$$(1.2) \quad \begin{aligned} |m(x) - m(y)| &\leq K|x - y|, \\ |\sigma(x) - \sigma(y)| &\leq K|x - y| \end{aligned}$$

and the linear growth condition:

$$(1.3) \quad \begin{aligned} |m(x)| &\leq K\sqrt{1 + x^2}, \\ |\sigma(x)| &\leq K\sqrt{1 + x^2} \end{aligned}$$

where K is a positive constant.

Under A_1 and A_2 we know (Wong (1971, Prop. 4.1, P_5 and P_6 , p. 150)) that the solution of problem (1.1) is unique with probability one and is a measurable Markov process.

Under some additional conditions the unique solution of problem (1.1) must have a stationary transition density, say $p_{X_t|X_0=a}(\cdot)$ satisfying the forward equation of

* Received by the editors November 29, 1976, and in revised form June 24, 1977.

† Laboratoire d'Automatique et d'Analyse des Systèmes du Centre National de la Recherche Scientifique, 31400 Toulouse, France. This research was supported by the National Science Foundation under Grant 01P75-04371 and by the "Centre National de la Recherche Scientifique."

Kolmogorov and $p_{X_t|X_0=a}(\cdot)$ must tend to a limiting density, say $p(\cdot)$ as t goes to infinity.

A class of such X_t processes for which $m(\cdot)$ and $\sigma^2(\cdot)$ are polynomials has been constructed and some specific processes of this class are given in Wong (1964, examples B and E).

For the identification of coefficients of diffusion processes, Lipcer and Šyriaev (1974) use maximum likelihood estimates.

In this paper, we propose a nonparametric procedure to estimate point by point the function $m(\cdot)$ when the function $\sigma(\cdot)$ is known or when $\sigma(\cdot)$ is unknown but takes a constant value.

Considering the properties of the transition density of the X_t process, its limiting density $p(\cdot)$ can be explicitly related to the pair $(m(\cdot), \sigma(\cdot))$ in the following way:

$$\frac{1}{2}(\sigma^2 p)' = mp.$$

So the techniques of nonparametric estimation used here to estimate point by point the density function $p(\cdot)$ appears to be a powerful tool in solving our initial problem of nonlinear identification.

For the case of a stationary Markov process having an initial density $p(\cdot)$, in § 2, we give a local (or point by point) recursive estimate of $p(\cdot)$ and its derivative $p'(\cdot)$, for which we show quadratic mean convergence under a specific assumption on the process itself. For the case of stochastic processes defined by problem (1.1) in § 3, we derive sufficient conditions on the pair $(m(\cdot), \sigma(\cdot))$ which imply the assumptions made in the previous section.

Independently of the estimation of the density we give, in § 4, a quadratic mean convergent recursive estimate of σ^2 (assuming that the function $\sigma(\cdot)$ is constant).

Finally, in § 5, applying the previous results, we give the solution to a class of nonlinear identification problems by suggesting a local estimate of the function $m(\cdot)$.

2. Estimation of density probability for Markov processes. In the case of a sequence X_1, X_2, \dots, X_n of independent and identically distributed random variables whose distribution is absolutely continuous, many nonparametric estimates of the density function have been proposed in the past. We may recall the Rosenblatt-Parzen estimate (Rosenblatt (1956), Parzen (1962)), the Yamato estimate (Yamato (1972)) and the Deheuvels estimate (Deheuvels (1973)). As far as we know, in the case of a dependent sequence, relatively few results have been obtained. Under specific conditions on the nature of the sequence it has been shown that the Rosenblatt-Parzen estimate is still convergent (Roussas (1969), Rosenblatt (1970)).

In this paper, instead of a sequence of random variables, we have to deal with a stochastic process $\{X_t, t \in [0, \infty)\}$. In this section we assume that the X_t process is a measurable stationary Markov process which has an initial density $p_X(\cdot)$ or simply $p(\cdot)$, and we introduce an estimate of $p(x_0)$, where x_0 is any point of R , which has the same structure as the Deheuvels estimate. Such structure is "recursive" and seems to be well adapted to our problem of estimating the function $m(\cdot)$ (see § 5). In order to prove the convergence to zero of the variance of our estimate (see Theorem 2.1) we have to impose a specific condition on the transition density of the stationary Markov process.

For each $t \in [0, \infty)$, we define the transition probability operator H_t of a stationary process X_t by:

$$(H_t f)(a) = E(f(X_t) | X_0 = a) \quad a \in R,$$

where $f(\cdot)$ is any Borel measurable bounded function on R .

A stationary process X_t with transition operator H_t is said to satisfy the condition $G_2(s, \alpha)$ (Rosenblatt (1970, p. 202) for the case of a sequence of random variables): if there is some $s > 0$ such that

$$|H_s|_2 = \sup_{\{f: E f(X) = 0\}} \frac{E^{1/2}(H_s f)^2(X)}{E^{1/2} f^2(X)} \leq \alpha < 1.$$

The H_t operator is in fact a contraction (for any $t \in [0, \infty)$: $|H_t|_2 \leq 1$).

For stationary Markov processes, the transition probability operator verifies the semigroup property, i.e. for $s, t \geq 0$: $H_{s+t} = H_s H_t = H_t H_s$ (Wong (1971, p. 183)). As a consequence of the semigroup and contraction properties, the condition $G_2(s, \alpha)$ implies (Banon (1977, p. 79)) for $t \in [0, \infty)$:

$$(2.1) \quad |H_t|_2 < \beta^t / \alpha \quad \text{with } \beta = \alpha^{1/s} < 1.$$

THEOREM 2.1 (quadratic mean convergence of $p_t(x_0)$). *Let $\{X_t, t \in [0, \infty)\}$ be a measurable stationary Markov process having a continuous and bounded initial density $p(\cdot)$ on R and satisfying condition G_2 .*

Let $K(\cdot)$ be a probability density function (i.e., nonnegative, Borel measurable function such that $\int_R K(y) dy = 1$), and be bounded on R ; and let h_t be a bounded and strictly positive function on R^+ such that, $t \in [0, \infty)$:

$$(2.2a) \quad h_t \downarrow 0, \quad t \rightarrow \infty,$$

$$(2.2b) \quad b_t = \int_0^t h_s ds < \infty,$$

and

$$(2.3) \quad b_t \rightarrow \infty, \quad t \rightarrow \infty.$$

For $t > 0$, let

$$(2.4) \quad p_t(x_0) = \frac{1}{b_t} \int_0^t K\left(\frac{x_0 - X_s}{h_s}\right) ds,$$

then

$$p_t(x_0) \xrightarrow[t \rightarrow \infty]{\text{q.m.}} p(x_0).$$

Proof. To prove the convergence in quadratic mean it is necessary and sufficient to show that:

$$(2.5) \quad E p_t(x_0) \xrightarrow[t \rightarrow \infty]{} p(x_0)$$

and

$$(2.6) \quad \text{Var } p_t(x_0) \xrightarrow[t \rightarrow \infty]{} 0.$$

Because we may write

$$E p_t(x_0) - p(x_0) = \frac{1}{b_t} \int_0^t h_s \left(E \frac{1}{h_s} K\left(\frac{x_0 - X}{h_s}\right) - p(x_0) \right) ds,$$

and because $b_t \rightarrow \infty$ as $t \rightarrow \infty$, a sufficient condition to show the asymptotic unbiasedness (2.5) is that $E(1/h_s)K((x_0 - X)/h_s) - p(x_0)$ converges to zero as s goes to infinity.

Using the same procedure as in Bochner (1955, Thm. 1.1.1., p. 2) or Rosenblatt (1971, p. 1816) we have:

$$\begin{aligned} \left| E \frac{1}{h_s} K\left(\frac{x_0 - X}{h_s}\right) - p(x_0) \right| &= \left| \int_R \frac{1}{h_s} K\left(\frac{x_0 - x}{h_s}\right) p(x) dx - p(x_0) \right| \\ &\leq \int_R K(y) |p(x_0 - h_s y) - p(x_0)| dy. \end{aligned}$$

We now split the region of integration in two:

$$\begin{aligned} (2.7) \quad \int_R K(y) |p(x_0 - h_s y) - p(x_0)| dy &\leq 2 \sup_{x \in R} p(x) \int_{\{y: h_s |y| > \varepsilon\}} K(y) dy \\ &+ \sup_{\{y: h_s |y| \leq \varepsilon\}} |p(x_0 - h_s y) - p(x_0)|. \end{aligned}$$

Because $p(\cdot)$ is continuous on R , there exists an $\varepsilon > 0$ such that the last term in (2.7) is as small as desired; because $p(\cdot)$ is bounded on R and h_s is decreasing to zero (condition (2.2a)), for every ε (and in particular for the one chosen above), there exists an $s > 0$ such that the first term is as small as desired, which proves (2.5).

Now we show the convergence to zero of the variance (statement (2.6)). We denote:

$$(2.8) \quad f_s(x) = K\left(\frac{x_0 - x}{h_s}\right) - EK\left(\frac{x_0 - X}{h_s}\right) \quad \text{for } x \in R$$

and

$$(2.9) \quad C(s_1, s_2) = Ef_{s_1}(X_{s_1})f_{s_2}(X_{s_2}) \quad \text{for } s_1, s_2 \in [0, \infty);$$

we may write the variance of $p_t(x_0)$ as:

$$(2.10) \quad \text{Var } p_t(x_0) = \frac{1}{b_t^2} \int_0^t \int_0^t C(s_1, s_2) ds_1 ds_2.$$

With the use of the stationarity property of the X_t process, (2.9) becomes:

$$(2.11) \quad C(s_1, s_2) = Ef_{s_1}(X_0)f_{s_2}(X_{|s_1 - s_2|}).$$

In order to use later on condition G_2 we introduce in (2.11) the transition probability operator:

$$C(s_1, s_2) = E(f_{s_1}(X_0)(H_{|s_1 - s_2|}f_{s_2})(X_0)).$$

By using the Schwarz inequality we get:

$$(2.12) \quad C(s_1, s_2) \leq E^{1/2}f_{s_1}^2(X)E^{1/2}(H_{|s_1 - s_2|}f_{s_2})^2(X).$$

Because X_t is a Markov process and $Ef_{s_2}(X) = 0$, and more specifically from (2.1), (2.12) becomes

$$(2.13) \quad C(s_1, s_2) < \frac{\beta^{|s_1 - s_2|}}{\alpha} E^{1/2}f_{s_1}^2(X)E^{1/2}f_{s_2}^2(X).$$

By construction of $f_s(\cdot)$, expression (2.8), we have for any $s \in [0, \infty)$:

$$\begin{aligned} \frac{1}{h_s} E f_s^2(X) &\leq \frac{1}{h_s} E K^2\left(\frac{x_0 - X}{h_s}\right) = \int_R K^2(y) p(x_0 - h_s y) dy \\ &\leq \sup_{x \in R} p(x) \int_R K^2(y) dy, \end{aligned}$$

which is bounded since $K(\cdot)$ and $\int_R K(y) dy$ are bounded. If we denote $C = \sup_{x \in R} p(x) \int_R K^2(y) dy$, (2.13) becomes:

$$C(s_1, s_2) < (C/\alpha) \sqrt{h_{s_1} h_{s_2}} \beta^{|s_1 - s_2|},$$

and (2.10) becomes:

$$\text{Var } p_t(x_0) < \frac{C}{\alpha b_t^2} \int_0^t \int_0^t \sqrt{h_{s_1} h_{s_2}} \beta^{|s_1 - s_2|} ds_1 ds_2,$$

which may be written:

$$\begin{aligned} &\frac{2C}{\alpha b_t^2} \int_0^t \int_{s_2}^t \sqrt{h_{s_1} h_{s_2}} \beta^{(s_1 - s_2)} ds_1 ds_2 \\ &\leq \frac{2C}{\alpha b_t^2} \int_0^t h_{s_2} \int_{s_2}^t \beta^{(s_1 - s_2)} ds_1 ds_2 \quad (\text{because } h_t \text{ is a decreasing function}) \\ &= \frac{2C}{\alpha b_t^2} \int_0^t h_{s_2} \int_0^{t-s_2} \beta^s ds ds_2 \quad (\text{by change of variable}) \\ &\leq \frac{2C}{\alpha b_t} \int_0^t \beta^s ds \\ &< \frac{2C}{\alpha b_t \ln(1/\beta)}. \end{aligned}$$

Since $\beta < 1$ (see (2.1)), we must have $\ln(1/\beta) > 0$ and therefore the variance of $p_t(x_0)$ must tend to zero as b_t goes to infinity; this completes the proof of the theorem. \square

We now study the properties of $p'_t(x_0)$ as an estimate of $p'(x_0)$:

THEOREM 2.2 (quadratic mean convergence of $p'_t(x_0)$). *Let $\{X_t, t \in [0, \infty)\}$ be a measurable stationary Markov process having a continuous and bounded initial density $p(\cdot)$ on R and satisfying condition G_2 . Let $K(\cdot)$ be a continuous probability density function of bounded variation on R and such that $K'(\cdot)$ is bounded on R . Let h be a bounded and strictly positive function on R^+ such that condition (2.2) is verified and so is*

$$(2.14) \quad h_t^2 b_t \xrightarrow[t \rightarrow \infty]{} \infty.$$

For $t > 0$, let

$$(2.15) \quad p'_t(x_0) = \frac{1}{b_t} \int_0^t \frac{1}{h_s} K'\left(\frac{x_0 - X_s}{h_s}\right) ds.$$

If $p'(\cdot)$ is continuous and bounded on R , then

$$p'_t(x_0) \xrightarrow[t \rightarrow \infty]{\text{q.m.}} p'(x_0).$$

Proof. As in the proof of Theorem 2.1 we first show that $p'_t(x_0)$ is asymptotically unbiased; a sufficient condition is that:

$$E(1/h_s^2)K'((x_0 - X)/h_s) - p'(x_0)$$

converges to zero as s goes to infinity. To show that point we use an argument similar to that in Bhattacharya (1967) or Shuster (1969). Because $K(\cdot)$ is of bounded variation, $\lim_{|y| \rightarrow \infty} K(y)$ exist (Natanson (1955, p. 239)), and these limits must be zero since $\int_{\mathbb{R}} K(y) dy = 1$; therefore, integrating by parts, we get:

$$(2.16) \quad E \frac{1}{h_s^2} K' \left(\frac{x_0 - X}{h_s} \right) = \frac{1}{h_s} \int_{\mathbb{R}} K \left(\frac{x_0 - x}{h_s} \right) p'(x) dx.$$

From (2.16), the abovementioned convergence to zero that we need follows by the same argument as in the proof of Theorem 2.1 (see inequality (2.7)).

The convergence to zero of $\text{Var } p'_t(x_0)$ as $t \rightarrow \infty$ follows in the same way as we have proved the convergence of $\text{Var } p_t(x_0)$ but now under the stronger condition (2.14)

$$h_t^2 b_t \rightarrow \infty, \quad t \rightarrow \infty.$$

The only point which remains to be shown is that $\int_{\mathbb{R}} K'^2(y) dy$ is bounded.

This property follows from the fact that $K'(\cdot)$ and $\int_{\mathbb{R}} |K'(y)| dy$ are bounded, this last integral being bounded since $K(\cdot)$ is of bounded variation (Natanson (1955, p. 259)). \square

Remark 2.1. Both estimates $p_t(x_0)$ and $p'_t(x_0)$ defined in Theorems 2.1 and 2.2 respectively are recursive, i.e., are respectively the solutions to ($t > 0$):

$$\frac{dp_t(x_0)}{dt} = -\frac{h_t}{b_t} p_t(x_0) + \frac{1}{b_t} K \left(\frac{x_0 - X_t}{h_t} \right)$$

and

$$\frac{dp'_t(x_0)}{dt} = -\frac{h_t}{b_t} p'_t(x_0) + \frac{1}{b_t h_t} K' \left(\frac{x_0 - X_t}{h_t} \right).$$

The initial conditions of these two differential equations can be arbitrary (when no a priori information is available); they do not affect the final value of both estimates.

3. Estimation of density probability for diffusion processes. In the previous section we have shown the quadratic mean convergence of $p_t(x_0)$ and $p'_t(x_0)$ under the assumptions that the X_t Markov process is stationary, measurable and satisfies condition G_2 .

Indeed, the stationarity assumption is not essential here, since we are dealing with asymptotic properties. A sufficient condition should be the existence of the limit of $p_{X_t}(\cdot)$ as t goes to infinity.

We now assume that the X_t process to be estimated is defined by problem (1.1) under A_1 - A_2 and has a transition density $p_{X_t|X_0=a}(\cdot)$ which converges, for all $a \in \mathbb{R}$, to a bounded limiting density $p(\cdot)$ as t goes to infinity.

As we have seen in the Introduction, such process is a measurable Markov process. More, the limit of $p_{X_t}(\cdot)$ must be equal to $p(\cdot)$ since we have for all $x \in R$:

$$\begin{aligned} \lim_{t \rightarrow \infty} p_{X_t}(x) &= \lim_{t \rightarrow \infty} \int_R p_{X_t|X_0=a}(x) p_{X_0}(a) da \\ &= \int_R \lim_{t \rightarrow \infty} p_{X_t|X_0=a}(x) p_{X_0}(a) da \\ &= \int_R p(x) p_{X_0}(a) da = p(x). \end{aligned}$$

Hence, to have the stationarity of the X_t process defined by (1.1) under A_1 - A_2 and the condition that the limiting density $p(\cdot)$ exists, it is necessary and sufficient to choose the initial density $p_{X_0}(\cdot)$ equal to $p(\cdot)$.

For the sake of simplicity, from now on we assume that the above X_t process is stationary.

In this section we derive sufficient conditions on the pair $(m(\cdot), \sigma(\cdot))$ to have the transition density convergence and condition G_2 satisfied.

Let us denote $P(x, t|a, s) = \mathcal{P}(X_t < x | X_s = a)$ the transition function of the unique X_t process solution of problem (1.1) where $a, x \in R$ and $t > s$.

Under A_1, A_2 (see § 1) and the additional condition:

A_3 : $x \in R, \sigma(x) \geq \sigma_0 > 0$,

we know (Wong (1971, Prop. 7.1, (a) and (e), p. 173)) that $P(x, t|\cdot, \cdot)$ is the unique solution of the backward equation of Kolmogorov:

$$(3.1a) \quad \frac{1}{2} \sigma^2(a) \frac{\partial^2 P(x, t|a, s)}{\partial a^2} + m(a) \frac{\partial P(x, t|a, s)}{\partial a} = - \frac{\partial P(x, t|a, s)}{\partial s}$$

with the terminal condition:

$$(3.1b) \quad \lim_{s \uparrow t} P(x, t|a, s) = \begin{cases} 1 & \text{if } x > a, \\ 0 & \text{otherwise} \end{cases}$$

and is absolutely continuous; that is, $P(x, t|a, s)$ can be written as:

$$P(x, t|a, s) = \int_{-\infty}^x p(y, t|a, s) dy, \quad a, x \in R, \quad t > s.$$

Because the functions $m(\cdot)$ and $\sigma(\cdot)$ do not depend on time, we see from (3.1a) that $p(x, t|a, s)$ depends only on $t - s$ and not on t and s separately. Let $p_a(\cdot, \cdot)$ $a \in R$ denote the transition density on $R \times R^+$ of the X_t process satisfying (1.1). (We will use the shorter notation $p_a(x, t)$ instead of $p_{X_{t+s}|X_s=a}(x)$; we can drop the s because of the stationarity of the transition density.)

If in addition to A_1, A_2 and A_3 , we assume

A_4 : $m'(\cdot), \sigma'(\cdot)$ and $\sigma''(\cdot)$ satisfy the conditions of type (1.2) and (1.3), then (Wong (1971, Prop. 7.1, (d), p. 173)) $p_a(\cdot, \cdot)$ is the unique fundamental solution of the forward equation of Kolmogorov (Fokker-Planck equation):

$$(3.2a) \quad \frac{1}{2} \frac{\partial^2}{\partial x^2} (\sigma^2(x) p_a(x, t)) - \frac{\partial}{\partial x} (m(x) p_a(x, t)) = \frac{\partial}{\partial t} p_a(x, t)$$

with the initial condition

$$(3.2b) \quad \lim_{t \downarrow 0} \int_R f(x) p_a(x, t) dx = f(a)$$

(for any Schwartz function of rapid descent $f(\cdot)$).

We now want to show that under sufficient conditions on the functions $m(\cdot)$ and $\sigma(\cdot)$, $p_a(\cdot, \cdot)$, the unique solution of problem (3.2), converges to a bounded limiting density as t goes to infinity. In other words we want to find stability conditions of problem (3.2). To find out these conditions we need the following lemma:

LEMMA 3.1 (expansion formula). *Let $p_a(\cdot, \cdot)$ be the unique solution of problem (3.2); then $p_a(\cdot, \cdot)$ can be written in the form for $a, x \in R, t \in [0, \infty)$*

$$(3.3) \quad p_a(x, t) = \pi(x) \int_R e^{-\lambda t} \sum_{j,k=1}^2 \phi_j(x, \lambda) \phi_k(a, \lambda) d\rho_{jk}(\lambda),$$

where $\pi(\cdot)$ is any nonnegative solution of the equation

$$(3.4) \quad \frac{1}{2} \frac{d}{dx} (\sigma^2(x) \pi(x)) = m(x) \pi(x), \quad x \in R,$$

$\phi_1(\cdot, \lambda)$ and $\phi_2(\cdot, \lambda)$ are solutions of the Sturm–Liouville equation

$$(3.5a) \quad \frac{1}{2} \frac{d}{dx} \left(\sigma^2(x) \pi(x) \frac{du}{dx}(x) \right) + \lambda \pi(x) u(x) = 0, \quad \lambda \in R, \quad x \in R,$$

and satisfy the conditions

$$\begin{aligned} \phi_1(0, \lambda) &= 1, & \phi_1'(0, \lambda) &= 0, \\ \phi_2(0, \lambda) &= 0, & \phi_2'(0, \lambda) &= 1, \end{aligned}$$

($\rho_{jk}(\lambda)$) is the limiting matrix of the spectral matrix ($\rho_{\gamma,jk}(\lambda)$) as γ goes to ∞ , associated with equation (3.5a) together with the boundary conditions (corresponding to the reflecting barriers in the regular case of problem (3.2))

$$(3.5b) \quad u'(-\gamma) = u'(\gamma) = 0 \quad \gamma \in R$$

once $\phi_1(\cdot, \lambda)$ and $\phi_2(\cdot, \lambda)$ are chosen as basis for the solutions of (3.5a).

Proof. We shall give a proof by constructing a solution (see also solution of Problem 9, Chapter 4, in Wong (1971, p. 178)). We can verify that any function $f_a(\cdot, t)$ of the form

$$\pi(\cdot) \int_R e^{-\lambda t} \sum_{j,k=1}^2 \phi_j(\cdot, \lambda) \psi_k(a, \lambda) d\rho_{jk}(\lambda)$$

is a solution of equation (3.2a) where $\pi(\cdot)$ and $\phi_j(\cdot, \lambda)$, $j = 1, 2$, are defined in the same way as in Lemma 3.1, and $\psi_j(\cdot, \lambda)$, $j = 1, 2$, are any function of the same class as the ϕ 's.

By setting $t = 0$ in the above expression and using the expansion theorem (Coddington and Levinson (1955, Thm. 5.2, p. 251)) the ψ 's may be regarded as the transform of $f_a(x, 0)$ by the means of the ϕ 's, i.e.:

$$\psi_j(a, \lambda) = \int_R f_a(x, 0) \phi_j(x, \lambda) dx, \quad j = 1, 2, \quad a \in R.$$

If we now assume that $f_a(x, 0)$ satisfies the initial condition of problem (3.2), then we get $\psi_j(\cdot, \lambda) \equiv \phi_j(\cdot, \lambda)$, $j = 1, 2$, which completes the proof of the lemma. \square

Remark 3.1. We know (Coddington and Levinson (1955, Thm. 5.1, p. 251)) that the limiting matrix $(\rho_{jk}(\lambda))$ defined in Lemma 3.1 always exists but could be not unique (in the so called limit-circle case); actually we are not going to introduce more conditions to have the uniqueness, because we only need here the existence property.

The expansion formula (3.3) for the fundamental solution $p_a(x, t)$ seems to be more convenient for our purpose than the approximate formula for appropriate solutions of equation (3.1a) given in Cohen and Lewis (1967) or in Ludwig (1975).

Let Λ be the set of nonconstancy points of $(\rho_{jk}(\lambda))$.

The set Λ is called the *spectrum* of the problem (3.5) with $\gamma \rightarrow \infty$. Despite the fact that the spectrum could not be completely defined (since ρ could be not unique) we can say something about the nature of the spectrum.

We know that there exists an increasing sequence of eigenvalues $\{\lambda_{\gamma,n}\}$ and a complete orthonormal set of corresponding eigenfunctions $\{\theta_{\gamma,n}(\cdot)\}$ associated with the Sturm–Liouville problem (3.5), $n = 0, 1, 2, \dots$. Using the boundary conditions (3.5b) we have:

$$\lambda_{\gamma,n} = \int_{-\gamma}^{\gamma} \frac{\sigma^2(z)}{2} \pi(z) \left(\frac{d\theta_{\gamma,n}(z)}{dz} \right)^2 dz,$$

which shows, since the integrand is nonnegative, that

$$\lambda_{\gamma,n} \geq 0, \quad n = 0, 1, 2, \dots$$

Letting $\gamma \rightarrow \infty$, we see that the spectrum Λ cannot lie on the negative part of the real line.

Further, we can verify directly that for every $\gamma \in R$, $\lambda_{\gamma,0} = 0$ and $\theta_{\gamma,0}(\cdot)$ is a constant, say θ_γ , such that:

$$\int_{-\gamma}^{\gamma} \theta_\gamma^2 \pi(z) dz = 1.$$

To say something about $\lambda_{\gamma,0}$ as γ goes to infinity, we must consider the integrability of $\pi(\cdot)$. If $\pi(\cdot)$ is not integrable on R then zero cannot remain an eigenvalue as $\gamma \rightarrow \infty$ since the square integrability with respect to $\pi(\cdot)$ of the corresponding solution of problem (3.5) cannot be maintained any more.

We can now state the following assumption under which we shall show the convergence of the transition density to a limiting density:

A_5 : The pair $(m(\cdot), \sigma(\cdot))$ is such that the solutions of equation (3.4), are bounded and integrable on R .

For example, $m(x) = Ax$ together with $\sigma(x) = B$, where A and B are two constants such that $A < 0$ and $B \neq 0$, satisfy A_5 .

LEMMA 3.2 (convergence to a limiting density). *Let $\{X_t, t \in [0, \infty)\}$ be the process defined by problem (1.1) under A_1 – A_4 . Then under the additional assumption A_5 , the transition density $p_a(\cdot, t)$ of the X_t process converges as $t \rightarrow \infty$ for all $a \in R$ to a bounded and continuous limiting density on R , $p(\cdot)$, which is a solution of (3.4).*

Proof. Under A_5 , we have seen that zero is an eigenvalue ($\lambda_{\infty,0} = 0$); in letting $t \rightarrow \infty$ in expression (3.3) we have for all $a, x \in R$:

$$\lim_{t \rightarrow \infty} p_a(x, t) = \pi(x) \left(\sum_{j,k=1}^2 \phi_j(x, 0) \phi_k(a, 0) r_j r_k \right)$$

where $r_j r_k$ is the jump of $\rho_{jk}(0)$.

Since, by construction, the corresponding eigenfunction $\theta_{\infty,0}(\cdot)$ can be written for $x \in R$:

$$\theta_{\infty,0}(x) = r_1\phi_1(x, 0) + r_2\phi_2(x, 0),$$

and because $\theta_{\infty,0}(\cdot)$ is constant on R (equals θ_{∞} as previously noted), taking into account the conditions on the ϕ 's at $x = 0$, we must have successively: $r_1 \neq 0$, $r_2 = 0$, $\phi_1(x) = 1$ for every $x \in R$ and finally $r_1 = \theta_{\infty}$. The above limit can now be written:

$$\lim_{t \rightarrow \infty} p_a(x, t) = \theta_{\infty}^2 \pi(x).$$

Since $\int_R \theta_{\infty}^2 \pi(x) dx = 1$, the above statement shows that the limit is a density. Because $\pi(\cdot)$ is a solution of equation (3.4), the limiting density can only be the density $p(\cdot)$ of the lemma and must be bounded on R .

Finally, since $p(\cdot)$ is a solution of equation (3.4), $p(\cdot)$ must be continuous on R under A_2 . \square

Indeed, under specific assumptions on $m(\cdot)$ and $\sigma(\cdot)$ we can say more about the nature of the spectrum.

LEMMA 3.3 (nature of the spectrum). *Let, for $x \in R$:*

$$(3.6) \quad \mu(x) = \frac{m^2(x)}{2\sigma^2(x)} + \frac{m'(x)}{2} - \frac{m(x)\sigma'(x)}{\sigma(x)} + \frac{\sigma'^2(x)}{8} - \frac{\sigma(x)\sigma''(x)}{4}.$$

If $\min(\lim_{x \rightarrow -\infty} \mu(x), \lim_{x \rightarrow \infty} \mu(x)) = \mu$ for some $\mu \in (-\infty, \infty]$, then Λ can only be discrete in the interval $(-\infty, \mu)$.

Proof. We shall give an outline of the proof. By using the standard transformation (Birkhoff and Rota (1969, p. 296) or Titchmarsh (1946, p. 22)) we can rewrite the equation (3.5a) in the form of the Schroedinger equation:

$$(3.7) \quad d^2v(y)/dy^2 + (\lambda - q(y))v(y) = 0$$

with

$$(3.8) \quad y(x) = \int_0^x \frac{\sqrt{2}}{\sigma(z)} dz, \quad x \in R,$$

$$v(y(x)) = \left(\frac{\sigma(x)\pi(x)}{\sqrt{2}} \right)^{1/2} u(x),$$

and

$$q(y(x)) = \frac{1}{\sqrt{\sigma(x)\pi(x)}} \frac{d^2\sqrt{\sigma(x)\pi(x)}}{dy^2}.$$

Since the spectrum is unchanged in the transformation, we may study the nature of the spectrum from equation (3.7). We know (Coddington and Levinson (1955, Prob. 2, p. 255), Titchmarsh (1946, p. 113) or Schiff (1955, Chap. II, § 8)) that if the potential function $q(y)$ is bounded from below say by μ , as y tends to either endpoint of its domain then the spectrum is discrete in the interval bounded above by μ .

More precisely, if $\min(\lim_{x \rightarrow -\infty} q(y(x)), \lim_{x \rightarrow \infty} q(y(x))) = \mu$ for some $\mu \in [-\infty, \infty)$, then Λ can only be discrete in the interval $(-\infty, \mu)$. Using (3.4), (3.7) and (3.8), we obtain (Banon (1977, pp. 145–148)) $q(y(x)) = \mu(x)$ $x \in R$, where $\mu(x)$ is given by (3.6), which completes the proof of the lemma. \square

We now state the last assumption on $m(\cdot)$ and $\sigma(\cdot)$.

A_6 : The pair $(m(\cdot), \sigma(\cdot))$ is such that μ of the Lemma 3.3 is strictly positive.

As an example of functions $m(\cdot)$ and $\sigma(\cdot)$ satisfying assumption A_6 we can mention the class of functions such that $m(x) \sim Ax^\alpha$ and $\sigma(x) \sim Bx^\beta$ as $|x| \rightarrow \infty$ with $\alpha, \beta = 0, 1$, $A < 0$ and $B \neq 0$. For this class of functions we may simplify the study of $\mu(x)$ at the infinity by noting from (3.6) that: if $\alpha = 0, 1; \beta = 0$, then

$$\mu(x) \sim \frac{A^2}{2B^2} x^{2\alpha};$$

if $\alpha = \beta = 1$, then

$$\mu(x) \sim (1/(2B^2))(A - B^2/2)^2;$$

if $\alpha = 0, \beta = 1$, then

$$\mu(x) \sim B^2/8.$$

We notice that the exponent in the first expression is even and the coefficients in the three cases are always strictly positive and so is μ of Lemma 3.3.

Assumption A_6 , as it can be seen from Lemma 3.3, implies that the spectrum Λ can only be discrete at the beginning of the interval $[0, \infty)$. Such a result will allow us to prove the following lemma:

LEMMA 3.4 (the condition G_2). *Let $\{X_t, t \in [0, \infty)\}$ be the process defined by problem (1.1) under A_1 - A_5 (with $p_X(\cdot) \equiv p(\cdot)$); then under the additional assumption A_6 , the X_t process satisfies condition G_2 (see § 2).*

Proof. Using expression (3.3) (with $p(x)$ instead of $\pi(x)$) of Lemma 3.1 we may write, for any function $f(\cdot)$ on R which is Borel measurable, bounded and such that $Ef(X) = 0$, and any $s > 0$:

$$(H_s f)(a) = \int_R \sum_{j,k=1}^2 \phi_k(a, \lambda) e^{-\lambda s} \int_R f(x) \phi_j(x, \lambda) p(x) dx d\rho_{jk}(\lambda),$$

where H_s is the transition probability operator defined in § 2.

Using the Parseval equality we get:

$$E(H_s f)^2(X) = \int_R e^{-2\lambda s} \sum_{j,k=1}^2 g_j(\lambda) g_k(\lambda) d\rho_{jk}(\lambda),$$

with

$$(3.9) \quad g_j(\lambda) = \int_R f(x) \phi_j(x, \lambda) p(x) dx \quad j = 1, 2.$$

Since, under A_5 , zero is an eigenvalue and the corresponding jumps of $\rho_{jk}(0)$ are 1 for $j = k = 1$ and zero otherwise, we have:

$$E(H_s f)^2(X) = \left(\int_R f(x) p(x) dx \right)^2 + \int_{R-\{0\}} e^{-2\lambda s} \sum_{j,k=1}^2 g_j(\lambda) g_k(\lambda) d\rho_{jk}(\lambda).$$

Since $Ef(X) = 0$, we have:

$$E(H_s f)^2(X) = \int_{R-\{0\}} e^{-2\lambda s} \sum_{j,k=1}^2 g_j(\lambda) g_k(\lambda) d\rho_{jk}(\lambda).$$

Let λ_0 be the lower bound of $\Lambda - \{0\}$; then

$$E(H_s f)^2(X) \leq e^{-2\lambda_0 s} \int_R \sum_{j,k=1}^2 g_j(\lambda) g_k(\lambda) d\rho_{jk}(\lambda).$$

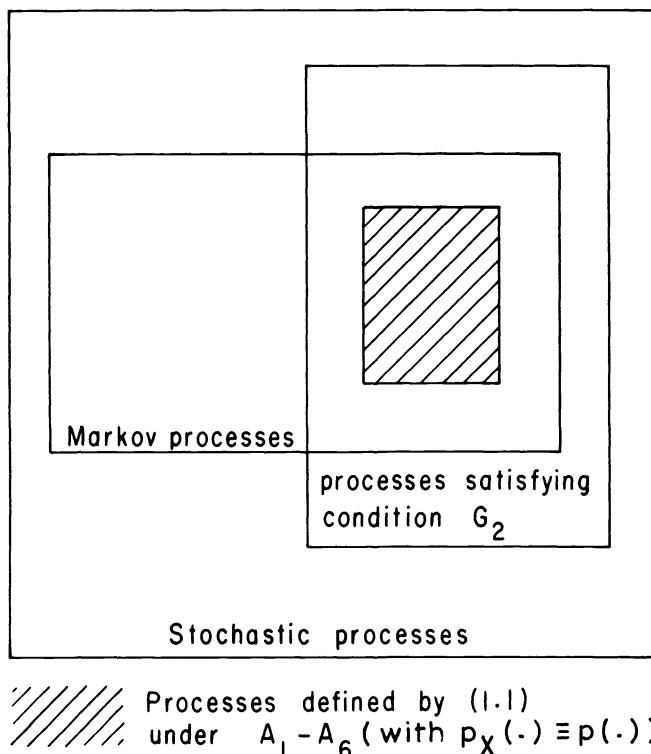


FIG. 1. Consequence of Lemma 3.4.

Recalling expression (3.9) and using once more the Parseval equality we get the following bound:

$$(3.10) \quad E(H_s f)^2(X) \leq e^{-2\lambda_0 s} \int_{\mathbb{R}} f^2(x) p(x) dx = e^{-2\lambda_0 s} E f^2(X).$$

Inequality (3.10) implies:

$$(3.11) \quad \|H_s\|_2 \leq e^{-\lambda_0 s}$$

Under A_6 , we have seen that the spectrum Λ can only be discrete at the beginning of $[0, \infty)$, therefore the lower bound λ_0 of $\Lambda - \{0\}$ must be strictly positive, so that from (3.11) condition G_2 is certainly satisfied. \square

Finally, using Lemma 3.4, we may rewrite Theorems 2.1 and 2.2 of § 2 for the special case of the X_t Markov process defined by problem (1.1).

THEOREM 3.1 (quadratic mean convergence of $p_t(x_0)$). *Let $\{X_t, t \in [0, \infty)\}$ be the process defined by problem (1.1) under $A_1 - A_6$ (with $p_X(\cdot) \equiv p(\cdot)$). Let $K(\cdot)$ and h be the functions defined in Theorem 2.1. Let $p_t(x_0)$ be the estimate defined by (2.4).*

Then:

$$p_t(x_0) \xrightarrow[t \rightarrow \infty]{\text{q.m.}} p(x_0).$$

THEOREM 3.2. (quadratic mean convergence of $p'_t(x_0)$). *Let $\{X_t, t \in [0, \infty)\}$ be the process defined by problem (1.1) under $A_1 - A_6$ (with $p_X(\cdot) \equiv p(\cdot)$). Let $K(\cdot)$ and h be the function defined in Theorem 2.2. Let $p'_t(x_0)$ be the estimate defined by (2.15).*

If $p'(\cdot)$ is continuous and bounded on R , then:

$$p'_t(x_0) \xrightarrow[t \rightarrow \infty]{\text{q.m.}} p'(x_0).$$

Remark 3.2. Assumptions A_2 - A_6 are only sufficient conditions. The case $m(x) = -\text{sgn}(x)$ and $\sigma(x) = 1$, for example, which does not satisfy the A_2 condition still works on (see solution of Problem 12, Chapter 4 in Wong (1971, p. 179)).

From the above results (more specifically from Lemma 3.4) we may draw Fig. 1.

4. Estimation of σ^2 . We now assume that the function $\sigma(\cdot)$ in (1.1a) takes a constant value σ .

Recalling that $\sigma^2(\cdot)$ may be characterized as the conditional expectation:

$$\sigma^2(x) = \lim_{t \rightarrow 0} \frac{1}{t} E((X_{t+s} - X_s)^2 | X_s = x), \quad x \in R, \quad s \in [0, \infty),$$

we suggest a recursive estimate of σ^2 .

THEOREM 4.1 (quadratic mean convergence of σ_n^2). *Let $\{X_t, t \in [0, \infty)\}$ be the solution to the stochastic differential equation (1.1a) where $m(\cdot)$ satisfies (1.2) and (1.3), and $\sigma(x) = \sigma \forall x \in R$. Let the initial condition satisfy $EX^4 < \infty$.*

Let $\{\tau_i\}_{i=1}^\infty$ be a bounded sequence of positive numbers such that:

$$(4.1) \quad \tau_i \rightarrow 0, \quad i \rightarrow \infty,$$

and $\{t_i\}_{i=1}^\infty$ a sequence such that $0 \leq t_1$ and $t_i + \tau_i \leq t_{i+1}$ $i = 1, 2, \dots$.

Let, for $n = 1, 2, \dots$,

$$(4.2) \quad \sigma_n^2 = \frac{1}{n} \sum_{i=1}^n \frac{1}{\tau_i} (X_{t_i+\tau_i} - X_{t_i})^2.$$

Then:

$$\sigma_n^2 \xrightarrow[n \rightarrow \infty]{\text{q.m.}} \sigma^2.$$

Because the proof would use the same arguments as those used by Wong and Zakai (1965) to prove the property of quadratic variation of diffusion processes, it will be omitted.

Remark 4.1. The estimate σ_n^2 defined in Theorem 4.1 is recursive, i.e., is the solution to:

$$\sigma_n^2 = \frac{n-1}{n} \sigma_{n-1}^2 + \frac{1}{n\tau_n} (X_{t_n+\tau_n} - X_{t_n})^2$$

with $n = 1, 2, \dots$. The choice of the sequence $\{t_i\}$ can be determined by practical consideration (e.g. $t_{i+1} - t_i$ is equal to the computational time of the above difference equation).

5. Estimation of $m(x_0)$. Recall that $m(\cdot)$ can be interpreted as the conditional expectation:

$$m(x) = \lim_{t \rightarrow 0} \frac{1}{t} E(X_{t+s} - X_s | X_s = x), \quad x \in R, \quad s \in [0, \infty);$$

one possible way to estimate $m(\cdot)$ (as we did for $\sigma^2(\cdot)$ constant) could be to estimate directly the above conditional mean.

We propose here another approach using nonparametric density estimation.

In § 3 we have seen that the limiting density $p(\cdot)$, when it exists, must verify equation (3.4). So, we are able to express the function $m(\cdot)$ in terms of $p(\cdot)$. At any point x_0 of R such that $p(x_0) \neq 0$ we have:

$$(5.1a) \quad m(x_0) = \frac{1}{2}(\sigma^{2'}(x_0) + \sigma^2(x_0)q(x_0)),$$

where

$$(5.1b) \quad q(x_0) = p'(x_0)/p(x_0).$$

Using the result of § 3, we may find a procedure to estimate $q(x_0)$.

THEOREM 5.1 (convergence in probability of $q_t(x_0)$). *Let $\{X_t, t \in [0, \infty)\}$ be the process defined by problem (1.1) under A_1 - A_6 (with $p_X(\cdot) \equiv p(\cdot)$). Let $\varepsilon > 0$. Let $K_1(\cdot)$ and $K_2(\cdot)$ be the functions $K(\cdot)$ defined in Theorems 2.1 and 2.2 respectively. Let h_t be a positive function on R^+ satisfying conditions (2.2) and (2.14). For $t > 0$, let*

$$(5.2) \quad q_t(x_0) = \frac{\int_0^t (1/h_s) K_2'((x_0 - X_s)/h_s) ds}{\int_0^t K_1((x_0 - X_s)/h_s) ds + \varepsilon}.$$

If $p'(\cdot)$ is continuous and bounded on R , then:

$$q_t(x_0) \xrightarrow[t \rightarrow \infty]{p.} q(x_0).$$

Proof. Multiplying the numerator and denominator of (5.2) by $1/(b_t)$ $t > 0$ and using the estimates given by (2.4) and (2.15) we get:

$$q_t(x_0) = \frac{p_t'(x_0)}{p_t(x_0) + \varepsilon/b_t}.$$

Because the mean square convergence implies the convergence in probability, we may apply the property of the latter relative to continuous functions (here to the quotient) (Lukacs (1975, p. 43)).

Therefore, using the results of Theorems 3.1 and 3.2, we have:

$$q_t(x_0) \xrightarrow[t \rightarrow \infty]{p.} p'(x_0)/p(x_0),$$

which completes the proof by considering (5.1b). \square

Finally we can state our last result which is relative to a local estimate of the function $m(\cdot)$:

COROLLARY 5.1 (convergence in probability of $m_t(x_0)$). *Let $\sigma(\cdot)$ be the function defined in § 1 (which we assume known). Let, for $t > 0$,*

$$(5.3) \quad m_t(x_0) = (1/2)(\sigma^{2'}(x_0) + \sigma^2(x_0)q_t(x_0)),$$

where $q_t(x_0)$ is given by (5.2).

Then under the conditions of Theorem 5.1

$$m_t(x_0) \xrightarrow[t \rightarrow \infty]{p.} m(x_0).$$

Proof. The convergence in probability of $m_t(x_0)$ follows from Theorem 5.1 and from considering expressions (5.1a) and (5.3). \square

In the case when the function $\sigma(\cdot)$ is unknown but takes a constant value σ , we may use the result of the previous section.

COROLLARY 5.2 (convergence in probability of $m_{t,n}(x_0)$). Let $\sigma(x) = \sigma \forall x \in R$, (but σ unknown). Let $EX^4 < \infty$, and let σ_n^2 be the estimate defined by (4.2). Let, for $t > 0$ and $n = 1, 2, \dots$,

$$(5.4) \quad m_{t,n}(x_0) = (1/2)\sigma_n^2 q_t(x_0),$$

where $q_t(x_0)$ is given by (5.2).

Then under the conditions of Theorem 5.1:

$$m_{t,n}(x_0) \xrightarrow[t,n \rightarrow \infty]{P.} m(x_0).$$

Proof. The convergence in probability of $m_{t,n}(x_0)$ follows by using the Theorems 4.1 and 5.1 and considering expressions (5.1a) and (5.4). \square

Remark 5.1. The previous results are true under the specific condition that $p'(\cdot)$ is continuous and bounded (see Theorem 5.1). Actually we should introduce a last assumption, say A_7 , on the pair $(m(\cdot), \sigma(\cdot))$ such that the above condition is satisfied.

When $\sigma(\cdot)$ takes a constant value on R we can verify that under the following condition:

$$\min \left(\lim_{x \rightarrow -\infty} m(x), -\lim_{x \rightarrow \infty} m(x) \right) > 0,$$

assumptions A_5 and A_6 are satisfied. Moreover, under A_2 $p'(\cdot)$ must be continuous and bounded on R .

This can be seen in writing explicitly the solutions of equation (3.4):

$$\pi(x) = C \exp \int_0^x \frac{2m(z)}{\sigma^2} dz.$$

Many functions $K(\cdot)$ satisfying conditions of Theorem 2.1 and 2.2 have been proposed in the past (Parzen (1962), Rosenblatt (1971)). Perhaps the simplest choices for $K_1(\cdot)$ and $K_2(\cdot)$ in (5.1) would be, for $y \in R$

$$K_1(y) = \frac{1}{2} I_{(-1,1]}(y),$$

and

$$K_2(y) = (1 - |y|) I_{(-1,1]}(y);$$

i.e. $K_2'(y) = -\text{sgn}(y) I_{(-1,1]}(y) = -2 \text{sgn}(y) K_1(y)$ where

$$I_A(y) = \begin{cases} 1 & \text{if } y \in A, \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \text{sgn}(y) = \begin{cases} 1 & \text{if } y \geq 0, \\ -1 & \text{otherwise.} \end{cases}$$

6. Conclusion. The main results in this paper are the introduction of a continuous and recursive density estimate, its convergence for a subclass of stationary Markov processes satisfying an ergodicity condition (G_2), and the derivation of assumptions under which diffusion processes belong to such a subclass.

Actually, many questions remain unanswered. Among them, how the identification procedure proposed here could be extended to multidimensional stochastic processes.

Could assumptions A_5 and A_6 be related to one another?

Under which additional assumptions can we prove the convergence to zero of the integrated mean square error or the convergence with probability one of our estimates? Can we say something about the rate of convergence to zero of the mean square error?

Acknowledgment. This work is the result of a collaboration with Professor Eugene Wong. His suggestion of using expansion formula for transition density representation in the density estimate convergence proof has been very helpful.

REFERENCES

- J. AGUILAR-MARTIN (1974), *Estimation et filtrage dans la modélisation stochastique en automatique*, Thèse, Université Paul Sabatier de Toulouse, France.
- G. ALENGRIN (1974), *La théorie du filtrage non linéaire et ses applications au traitement du signal et à l'identification en automatique*, Thèse, Université Paul Sabatier de Toulouse, France.
- G. BANON (1971), *Etude d'algorithmes d'estimation des paramètres pour l'identification adaptative en temps réel des processus linéaires perturbés par un bruit corrélé*, Thèse, Université Paul Sabatier de Toulouse, France.
- (1977), *Estimation non paramétrique de densité de probabilité pour les processus de Markov*, Thèse, Université Paul Sabatier de Toulouse, France.
- P. K. BHATTACHARYA (1967), *Estimation of a probability density function and its derivatives*, Sankhyā, Ser. A, 29, pp. 373–382.
- G. BIRKHOFF AND G. C. ROTA (1969), *Ordinary Differential Equations*, Blaisdell, New York.
- S. BOCHNER (1955), *Harmonic Analysis and the Theory of Probability*, University of California, Berkeley.
- E. A. CODDINGTON AND N. LEVINSON (1955), *Theory of Ordinary Differential Equations*, McGraw-Hill, New York.
- J. K. COHEN AND R. M. LEWIS (1967), *A ray method for the asymptotic solution of the diffusion equation*, J. Inst. Math. Appl., 3, pp. 266–290.
- P. DEHEUVELS (1973), *Sur l'estimation séquentielle de la densité*, C. R. Acad. Sci. Paris Sér. A, 276, 16 Avril, pp. 1119–1121.
- R. E. KALMAN AND R. S. BUCY (1961), *New results in linear filtering and prediction theory*, Trans. ASME Ser. D, J. Basic Engrg., 83, pp. 95–108.
- R. Š. LIPČER AND A. N. ŠYRIAĖV (1974), *Statistic of Random Processes, Non Linear Filtering and Related Problems*, Moscow. (In Russian.)
- D. LUDWIG (1975), *Persistence of dynamical systems under random perturbations*, SIAM Rev. 17, pp. 605–640.
- E. LUKACS (1975), *Stochastic Convergence*, Academic Press, second ed., New York.
- I. P. NATANSON (1955), *Theory of Functions of a Real Variable*, Frederick Ungar, New York.
- E. PARZEN (1962), *On estimation of a probability density function and mode*, Ann. Math. Statist., 33, pp. 1065–1076.
- M. ROSENBLATT (1956), *Remarks on some non-parametric estimates of a density function*, Ibid., 27, pp. 832–837.
- (1970), *Density estimates and Markov sequences*, Nonparametric Techniques in Statistical Inference, M. Puri, ed., pp. 199–210.
- (1971), *Curve estimates*, Ann. Math. Statist., 42, pp. 1815–1842.
- G. ROUSSAS (1969), *Nonparametric estimation in Markov processes*, Ann. Inst. Statist. Math., 21, pp. 73–87.
- G. SALUT (1976), *Identification optimale des systèmes linéaires stochastiques*, Thèse, Université Paul Sabatier de Toulouse, France.
- L. I. SCHIFF (1955), *Quantum Mechanics*, McGraw-Hill, New York.
- E. F. SCHUSTER (1969), *Estimation of a probability density function and its derivatives*, Ann. Math. Statist., 40, pp. 1187–1195.
- E. C. TITCHMARSH (1946), *Eigenfunction Expansions Associated with Second-Order Differential Equations*, Oxford University Press, London, 1946.
- E. WONG AND M. ZAKAI (1965), *The oscillation of stochastic integrals*, Z. Wahrscheinlichkeitstheorie vol. 16, American Mathematical Society, Providence, RI, pp. 264–276.
- (1971), *Stochastic Processes in Information and Dynamical Systems*, McGraw-Hill, New York.
- E. WONG AND M. ZAKAI (1965), *The oscillation of stochastic integrals*, Z. Wahrscheinlichkeitstheorie Verw. Geb., 4, pp. 103–112.
- H. YAMATO (1972), *Sequential estimation of a continuous probability density function and mode*, Bull. Math. Statist., 14, pp. 1–12.

PERTURBED STOCHASTIC LINEAR REGULATOR PROBLEMS*

CHUN-PING TSAI†

Abstract. This paper is concerned with the approximate solution of stochastic optimal control problems which arise by perturbing the stochastic linear regulator problem, through an additive term with a small parameter δ in the drift coefficient of the unperturbed dynamical equations. The system states are assumed completely observable. Our main results concern expansions of solutions of the perturbed equation in powers $\delta, \delta^2, \delta^3, \dots$ of the small parameter δ .

1. Introduction. The problem of optimal control of Markov diffusion processes has been the subject of a great deal of research over the past several years. See for instance [5], [1]. However, it is a difficult matter to calculate optimal feedback control laws, except for the linear regulator problem and a few other special cases. In this paper, we consider a nonlinear perturbation of the stochastic linear regulator, which takes the form of a small quantity δ times a certain function, and develop a technique for computing approximately the optimal feedback control. The system states are assumed completely observable. Our main results concern expansions of solution of the perturbed problem in powers $\delta, \delta^2, \delta^3, \dots$ and the validity of these expansions. Part of this problem in a special case has been considered by Kolmanovskii and Nishikawa et al. See [9], [10].

Consider a stochastic system whose state $\xi^\delta(t)$ is an n dimensional vector, which satisfies a stochastic differential equation

$$(1.1^\delta) \quad d\xi^\delta(t) = (A(t)\xi^\delta(t) + \delta g(\xi^\delta(t)) + B(t)u(t)) dt + \sigma(t) dw(t)$$

with initial data

$$(1.2^\delta) \quad \xi^\delta(s) = x.$$

Here w is a Brownian motion process of some dimension d . The system state $\xi(t)$ is assumed known to the controller. The control $u(t)$ at time t is a vector, of some dimension k , chosen using a feedback control law Y :

$$(1.3) \quad u(t) = Y(t, \xi(t)).$$

The problem is to find among all $Y \in \mathcal{Y}(R^k)$, to be defined in § 2, one for which the following quadratic criterion of expected system performance is minimum.

$$(1.4^\delta) \quad J^\delta(s, x; u) = E_{s,x} \int_s^T L(t, \xi^\delta(t), u(t)) dt$$

where T denotes the finite terminal time and $L(t, x, u) = x'M(t)x + u'N(t)u$. For convenience, we use the notation

$$f^{\delta,u}(t, x) = A(t)x + \delta g(x) + B(t)u.$$

When $\delta = 0$, this is the well-known linear regulator problem, for which the optimal feedback control is a linear function of state. See [5, § 6.5].

$$(1.5) \quad Y^{0*}(s, x) = -N^{-1}(s)B'(s)K(s)x$$

* Received by the editors November 30, 1976, and in revised form August 3, 1977.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. This research was supported by the Air Force Office of Scientific Research under AFOSR 76-3063.

and u^* is given by (1.3) and (1.5). Here $K(s)$ is the unique symmetric, nonnegative definite matrix solution of the following matrix Riccati equation

$$\frac{dK(s)}{ds} = -K(s)A(s) - A'(s)K(s) + K(s)B(s)N^{-1}(s)B'(s)K(s) - M(s)$$

with $K(T) = 0$. It has been proved that $K(s)$ is bounded on any finite time interval.

Let $\phi^\delta(s, x)$ denote the minimum cost function and consider it as a function of the initial data

$$(1.6^\delta) \quad \phi^\delta(s, x) = \inf_{Y \in \mathcal{Y}(R^k)} J^\delta(s, x; Y).$$

We want to show under certain conditions that ϕ^δ satisfies the partial differential equation for all $(s, x) \in [0, T] \times R^n$

$$(1.7^\delta) \quad \phi_s^\delta + \frac{1}{2} \text{tr} \{ \sigma \sigma' \phi_{xx}^\delta \} + H^\delta(s, x, \phi_x^\delta) = 0$$

together with the data

$$(1.8^\delta) \quad \phi^\delta(T, x) = 0$$

where $x \in R^n$ and tr is the trace of a square matrix, i.e.,

$$\text{tr} \{ \sigma \sigma' \phi_{xx}^\delta \} = \sum_{i,j=1}^n (\sigma \sigma')_{ij} \frac{\partial^2 \phi^\delta}{\partial x_i \partial x_j},$$

ϕ_x denotes the gradient of ϕ in the variables $x = (x_1, \dots, x_n)'$, regarded as a row vector

$$(1.9^\delta) \quad H^\delta(s, x, P) = \min_{u \in K = R^k} [L(s, x, u) + P' \cdot f^{\delta, u}(s, x)]$$

and the optimal feedback control $Y^{\delta*}$ satisfies

$$(1.10) \quad y'N(s)y + \phi_x^\delta(s, x) \cdot B(s)y = \min \quad \text{on } R^k$$

when $y = Y^{\delta*}(s, x)$.

Thus, the completely observable optimal problem is in principle reduced to solving the Cauchy problem (1.7^δ)–(1.8^δ) for ϕ^δ and then minimizing the left-hand side of (1.10) over R^k for each $(s, x) \in [0, T] \times R^n$. This is usually difficult to do in practice. But for $\delta = 0$, it is well known that the solution is (see [5, Chap. VI.5])

$$\phi^0(s, x) = x'K(s)x + q(s), \quad 0 \leq s \leq T,$$

where $q(s) = \int_s^T \text{tr} \{ \sigma \sigma' K \} dt$ and the corresponding solution of (1.10) is just $Y^{0*}(s, x)$ as in (1.5).

We wish to find ϕ^δ , ϕ_x^δ (and hence also $Y^{\delta*}$) approximately in terms of quantities computable from ϕ^0 , ϕ_x^0 . We show that the following type of expansions hold uniformly for (s, x) in any compact set:

$$(1.11) \quad \phi^\delta = \phi^0 + \delta \theta_1 + \delta^2 \theta_2 + \dots + \delta^k \theta_k + o(\delta^k),$$

$$(1.12) \quad \phi_x^\delta = \phi_x^0 + \delta \theta_{1x} + \delta^2 \theta_{2x} + \dots + \delta^k \theta_{kx} + o(\delta^k).$$

The coefficients in (1.11) have the property that $k! \theta_k$ is the k th derivative of ϕ^δ with respect to δ when $\delta = 0$. Hence they satisfy the equations found by formally differentiating (1.7^δ) repeatedly with respect to δ and setting $\delta = 0$. These equations involve

the partial derivatives of H^δ and ϕ^0 of corresponding orders. Whether such expansions are available will depend on smoothness properties of H^δ which will be guaranteed by the assumptions in § 2. Suppose we use the optimal unperturbed policy Y^{0*} in the perturbed problem. We would like to know how close to the optimum is the performance $J^\delta(s, x; Y^{0*})$ in perturbed problems. Our method will also answer this question.

We begin in § 2 by discussing assumptions on $f, L, \mathcal{Y}(k)$ and then get some preliminary results about the existence, uniqueness, boundedness of the moments of ξ^δ and some properties of H^δ . In § 3, the existence and uniqueness of solutions of dynamic programming equations are proved. Then we use a verification theorem to show that the solution is ϕ^δ . In § 4, we give the approximation method and prove that it is valid and finally we discuss the goodness of Y^{0*} in the perturbed problem.

2. Assumptions and preliminary results. If Γ is an open set, we write $g \in C^l(\Gamma)$ to mean that the function g together with its partial derivatives of orders $j = 1, \dots, l$ are continuous on Γ . If Γ is not open, then $g \in C^l(\Gamma)$ means that g agrees on Γ with a function $h \in C^l(\Gamma')$ where Γ' is open and $\Gamma \subset \Gamma'$. Similarly, let $C^{1,2}(Q)$ be defined as above except g is twice continuously differentiable with respect to x and continuously differentiable with respect to t . Let $C_p(Q)$ denote the class of all continuous functions ψ which satisfy a polynomial growth condition on Q , i.e., for some positive constants c, m , $|\psi(t, x)| \leq c(1 + |x|^m)$ when $(t, x) \in Q$. Let us also denote by $C_p^{(l)}(Q)$ the class of functions in $C^l(Q)$ which satisfy a polynomial growth condition on Q .

The following assumptions are made.

(AI) $A(t), B(t), M(t)$ and $N(t)$ are bounded C^∞ matrix-valued functions with size $n \times n, n \times k, k \times k$, respectively. $M(t)$ is symmetric semi-positive and $N(t)$ is symmetric positive definite.

(AII) $g \in C_p^{(l)}(R^n)$ and $g_x(x)$ is a matrix-valued function with diagonal elements bounded above and off-diagonal elements bounded.

(AIII) There exist positive constants M_1, M_2, M_3 , constants α_1, α_2 independent of δ and a positive C^∞ function $V(x)$ such that

- (a) $\frac{1}{2} \text{tr} \{ \sigma(t) \sigma'(t) V_{xx}(x) \} + V_x(x) \cdot (A(t)x + \delta g(x)) \leq M_1(1 + V(x))$,
- (b) $(1 + |x|) |V_x(x)| \leq M_1(1 + V(x))$,
- (c) $V(x) \rightarrow \infty$ as $|x| \rightarrow \infty$,
- (d) $\alpha_1 + M_2|x|^2 \leq V(x) \leq \alpha_2 + M_3|x|^2$.

(AIV) There exist positive constants c_1, c_2 such that

- (a) $|\sigma(t)| \leq c_1(1 + |x|)$

and for all $\nu \in R^n$

- (b) $\sum_{i,j=1}^n (\sigma(t) \sigma'(t))_{ij} \nu_i \nu_j \geq c_2 |\nu|^2$.

(AV) $K = R^k$.

Condition (AIV) (b) means that noise enters directly into each component of the system. The corresponding dynamic programming equation is uniformly parabolic. This enables us to apply results about parabolic partial differential equations. However, uniform parabolicity does not hold for many systems of interest. In some cases, we can still get the results by the method which involves approximating the noise coefficient σ by σ^ε which has an inverse $(\sigma^\varepsilon)^{-1}$ for each $\varepsilon > 0$.

Under the assumptions of (AIII) and (AV), we have the existence and uniqueness of the solution $\eta(t)$ of the free-control system, i.e., let $u = 0$ in (1.1^δ). Moreover, for any positive integer m , there exists a positive constant C_m depending on t, M_1, M_2, M_3 such that for all $\delta > 0$ (see [4], [7], [8])

$$(2.1) \quad E_{s,x} |\eta(t)|^m \leq C_m (1 + |x|^m).$$

If K is a compact subset of R^k , then H^δ is clearly well defined by (1.9 $^\delta$). Let $u = V(s, x, p)$ be the unique vector in K at which H^δ is a minimum on K for every (s, x, p) belonging to an open set Γ of R^{2n+1} where p is a vector of dimension n . Then if Γ is a set such that $V \in C^l(\Gamma)$, we have $H^\delta \in C^{l+1}(\Gamma)$. Moreover, $H_s = L_s + Pf_s$, $H_x = L_x + Pf_x$, $H_p = f$; see [2]. When $K = R^k$, we can see easily that H^δ is a C^∞ function and

$$(2.2) \quad V(s, x, p) = -\frac{1}{2}N^{-1}B'p'.$$

Let K be any closed convex subset of R^k containing 0 as an interior point. Let

$$Q = [0, T] \times R^n, \quad Q_r = [0, T] \times \{|x| \leq r\}, \quad r = 1, 2, \dots.$$

We define $\mathcal{Y}(K)$ as the set of all functions Y such that

- (a) $Y(s, x) \in K$ for all $(s, x) \in Q$,
- (b) when $(s, x), (s, y) \in Q_r$ with $0 \leq s \leq T' < T$

$$(2.3) \quad |Y(s, x) - Y(s, y)| \leq \alpha_r |x - y|,$$

- (c) for all $(s, x) \in Q$, $|Y(s, x)| \leq \beta(1 + |x|)$.

The positive constants α_r, β may be different for different functions Y , and α_r may also depend on T' .

LEMMA 2.1. *Conditions (2.3) insure the existence and uniqueness of the process $\xi^\delta(t)$ in (1.1 $^\delta$), given the control Y and the initial data (1.2 $^\delta$).*

Proof. Using the same V as in (AIII), we have

$$\begin{aligned} & \frac{1}{2} \operatorname{tr} \{ \sigma(t) \sigma'(t) V_{xx}(x) \} + V_x(x) \cdot (A(t)x + \delta g(x) + B(t)Y(t, x)) \\ & \leq M_1(1 + V(x)) + \beta(1 + |x|) |V_x(x)| \\ & \leq M_1(1 + \beta)(1 + V). \end{aligned}$$

Hence the solution ξ^δ of (1.1 $^\delta$) with (1.2 $^\delta$) exists. This follows a result of [4]. Uniqueness comes from (2.3) (b). Q.E.D.

Furthermore, there exists a positive constant C'_m such that

$$(2.4) \quad E_{s,x} \max_{s \leq t \leq 1} |\xi^\delta(t)|^m \leq C'_m(1 + |x|^m).$$

The next lemma is an estimate on $E_{sx} |\xi^\delta(t)|$. We utilize here the assumption on g_x . This lemma and the result about W following this lemma will be used in the proof of Lemma 3.2.

LEMMA 2.2. $E_{s,x} |\xi^\delta(t)| \leq C''_m(1 + |x|)$ where C''_m is a positive constant independent of K .

Proof. Since $0 \in K$, by (1.6 $^\delta$) we have

$$\phi^\delta(s, x) \leq E_{s,x} \int_s^T \eta'(t) M(t) \eta(t) dt$$

where $\eta(t)$ is the solution of free control system. Because $M(t)$ is bounded, there exists a constant $\beta_1 > 0$ such that

$$\phi^\delta(s, x) \leq \beta_1 E_{s,x} \int_s^T |\eta(t)|^2 dt.$$

By (2.1) we have

$$\phi^\delta(s, x) \leq \beta_2(1 + |x|^2)$$

for a positive constant β_2 . Let $u^{\delta*}(t)$ be the optimal control and $\xi^{\delta*}$ the corresponding trajectory. From (1.6 $^\delta$) and the positive definite of $N(t)$ we have

$$(2.5) \quad E_{s,x} \int_s^T |u^{\delta*}(t)|^2 dt \leq \frac{\beta_2}{\gamma} (1 + |x|^2)$$

where the positive constant γ satisfies

$$\mu' N(t) \mu \geq \gamma |\mu|^2$$

for all $\mu \in R^k$.

Now subtract the equation governing the free-control system from (1.1 $^\delta$) with $u = u^{\delta*}$. Using the mean value theorem, we obtain

$$d(\xi^{\delta*} - \eta)(t) = \left(A(t) + \delta \int_0^1 g_x(\eta + \lambda(\xi^{\delta*} - \eta)) d\lambda \right) (\xi^{\delta*} - \eta)(t) dt + B(t) u^{\delta*}(t) dt$$

with $(\xi^{\delta*} - \eta)(s) = 0$. Let

$$g_x(x) = G_1(x) + G_2(x)$$

where $G_1(x)$ is a diagonal matrix-valued function whose diagonal elements are those of $g_x(x)$ and $G_2(x)$ is the same as g_x except diagonal elements are all zero. Let

$$A_1(t) = A(t) + \delta \int_0^1 G_2(\eta + \lambda(\xi^{\delta*} - \eta)) d\lambda.$$

By assumption, $A_1(t)$ is still a bounded matrix-valued function. Let $X_2(t, v)$ be the principal matrix solution of the following equation at initial time v ,

$$dX_1(t, v) = \delta \int_0^1 G_1(\eta + \lambda(\xi^{\delta*} - \eta)) d\lambda X_1(t, v) dt.$$

Since the elements of G_1 are bounded above, $X_1(t, u)$ is bounded for $s \leq v \leq t$. Using the variation of constants formula, we get

$$(\xi^{\delta*} - \eta)(t) = \int_s^T X_1(t, v) [A_1(v)(\xi^{\delta*} - \eta)(v) + B(v) u^{\delta*}(v)] dv.$$

Then

$$\begin{aligned} |\xi^{\delta*}(t) - \eta(t)|^2 &\leq 2 \left| \int_s^T X_1(t, v) A_1(v) (\xi^{\delta*}(v) - \eta(v)) dv \right|^2 \\ &\quad + 2 \left| \int_s^T X_1(t, v) B(v) u^{\delta*}(v) dv \right|^2. \end{aligned}$$

Taking expectation $E_{s,x}$ and by (2.5), Cauchy-Schwarz and Gronwall's inequalities, we have

$$E_{s,x} |\xi^{\delta*}(t) - \eta(t)|^2 \leq \beta_3 (1 + |x|^2)$$

where β_3 is a positive constant. Hence

$$\begin{aligned} E_{s,x} |\xi^{\delta*}(t)|^2 &\leq 2E_{s,x} |\xi^{\delta*}(t) - \eta(t)|^2 + 2E_{s,x} |\eta(t)|^2 \\ &\leq \beta_4 (1 + |x|^2) \end{aligned}$$

where β_4 is a positive constant. Since $(E_{s,x}|\xi^{\delta*}(t)|^p)^{1/p}$ is nondecreasing as p increases, we obtain

$$E_{s,x}|\xi^{\delta*}(t)| \leq C_m^\alpha(1+|x|)$$

where C_m^α is positive constant not depending on K . Q.E.D.

Let $X_2(t, v)$ be the principal matrix solution at initial time v of

$$dX_2(t, v) = \delta G_1(\xi^{\delta*}(t))X_2(t, v) dt.$$

By assumption on G_1 , $X_1(t, v)$ is bounded for $s \leq v \leq t$. Let W be the solution of

$$dW(t) = (A(t) + \delta g_x(\xi^{\delta*}(t)))W(t) dt$$

with $W(s) =$ identity matrix. Again, using the same technique as before, we can show that $W(t)$ is bounded and the bound does not depend on x .

This next lemma, a modification of Lemma V.5.2 of [5], is concerned with the probabilistic representation for solution $\psi(s, x)$ of a linear partial differential equation

$$\psi_s + \frac{1}{2} \text{tr} \{ \sigma \sigma' \psi_{xx} \} + \psi_x \cdot f + g(s, x)\psi + h(s, x) = 0.$$

LEMMA 2.3. *Let ψ be a solution of the above equation in $[0, T] \times R^n$ with $\psi(T, x) = \Psi(x)$, suppose that ψ, h, Ψ belong to $C_p^{1,2}(Q)$ and g is bounded and continuous on Q ; then*

$$\psi(s, x) = E_{s,x} \int_s^T D(u)h(u, \xi(u)) du + E_{s,x}D(T)\Psi(\xi(T))$$

where

$$D(u) = \exp \int_s^u g(v, \xi(v)) dv$$

Proof. Consider ψD in the proof of the cited lemma. Q.E.D.

3. Dynamic programming equation. The goal of this section is to establish Theorem 3.1 below. Let \mathcal{F}_0 denote the set of all nonnegative real-valued functions ψ on Q such that

(i) the partial derivatives $\psi_s, \psi_{x_i}, \psi_{x_i x_j}, i, j = 1, \dots, n$ are continuous on Q and satisfy a Hölder condition on each compact subset of Q ,

(ii) $\psi \in C_p(Q)$,

(iii) $\psi(T, x) = 0$ for all x .

We seek a solution in \mathcal{F}_0 of

$$(3.1) \quad \psi_s + \frac{1}{2} \text{tr} \{ \sigma \sigma' \psi_{xx} \} + H^\delta(s, x, \psi_x) = 0$$

with the Cauchy data $\psi(T, x) = 0$. If ϕ^δ is such a solution, let $Y^{\delta*}$ be defined by (1.10). Since the first term on the left-hand side of (1.10) is quadratic in y and the second term is linear in y , (1.10) uniquely determines $Y^{\delta*}$. The function $Y^{\delta*}$ clearly satisfies (2.3) (a). By a similar proof as for Theorem 2.2 of [3] it satisfies (2.3) (b), we shall prove later that (2.3) (c) holds.

The proof of Theorem 3.1 will proceed by approximation of the domain. The following theorem is quoted from [5]. It tells us that the existence of each ϕ^δ and $Y^{\delta*}$ implies a solution to the minimum problem. Let Γ be an open subset of Q and $\delta^*\Gamma$ be a

closed subset of the boundary of Γ such that $(\tau, \xi^\delta(\tau)) \in \partial^* \Gamma$ with probability 1, for every choice of initial data $(s, x) \in \Gamma$ and every admissible control, where τ is the first exit time. Let

$$J(s, x; Y) = E_{s,x} \int_s^\tau L(t, \xi^\delta(t), u(t)) dt.$$

Verification theorem. Let $\psi(s, x)$ be a solution of (3.1) with the boundary data $\psi(s, x) = 0$ for $(s, x) \in \partial^* \Gamma$ such that ψ is in $C_p^{1,2}(\Gamma)$ and continuous on the closure $\bar{\Gamma}$, then

(a) $\psi(s, x) \leq J(s, x; Y)$ for any admissible feedback control Y and any initial data $(s, x) \in \Gamma$;

(b) if Y^* is an admissible feedback control such that (1.10) is satisfied when $Y = Y^*(s, x)$, then $\psi(s, x) = J(s, x; Y^*)$.

This Y^* is optimal among all admissible feedback control laws, for all choices of initial data $(s, x) \in \Gamma$.

Let us now show that there is a unique solution in \mathcal{F}_0 of (3.1). This will be done by approximation in two stages. In the first step we assume K is a compact set containing zero as an interior point. Let τ_r denote the first time $t \leq T$ when $|\xi^\delta(t)| = r$; if $|\xi^\delta(t)| < r$ for $s \leq t \leq T$, we set $\tau_r = T$, where $\xi^\delta(s) = x$ and $|x| < r$. Let

$$(3.2) \quad J_r^\delta(s, x; Y) = E_{s,x} \int_s^{\tau_r} L(t, \xi^\delta(t), u(t)) dt.$$

As $r \rightarrow \infty$, τ_r increases to T . Since L is nonnegative, the monotone convergence theorem implies that $J_r(s, x; Y)$ tends to $J^\delta(s, x; Y)$. For $r = 1, 2, \dots$, let

$$\phi_r^\delta(s, x) = \min_{Y \in \mathcal{U}(K)} J_r(s, x; Y);$$

then $0 \leq \phi_1^\delta \leq \phi_2^\delta \leq \dots$; since $0 \in K$, we have $\phi_r^\delta \leq J_r^\delta(s, x; 0)$. From (2.1) and $J_r^\delta(s, x; 0) \leq J(s, x; 0)$ we have

$$(3.3) \quad \phi_r^\delta(s, x) \leq \nu_1(1 + |x|^2)$$

for some positive constants ν_1 . Let

$$(3.4) \quad \phi^\delta(s, x) = \lim_{r \rightarrow \infty} \phi_r^\delta(s, x).$$

Clearly, ϕ^δ satisfies (3.3) too. By Theorem VI 6.1 of [5] and the verification theorem, for each r , ϕ_r^δ is a solution of (3.1) with the boundary data $\phi_r^\delta = 0$ on

$$\Sigma_r = ([0, T] \times \{|x| = r\}) \cup (\{T\} \times \{|x| \leq r\}).$$

In order to show that ϕ^δ also belongs to \mathcal{F}_0 , we need to establish a uniform bound on any compact set for the gradients $(\phi_r^\delta)_x$.

LEMMA 3.1. *Let B be a compact subset of Q_{r_0} ; then $(\phi_r^\delta)_x$ is bounded on B uniformly with respect to $r > r_0$.*

Proof. With $\phi = \phi_r^\delta$ in Lemma 5.3, of [2, p. 494], we have

$$(3.5) \quad (\phi_r^\delta)_x(s, x) = E_{s,x} \int_s^{\tau_r} L_x W dt + E_{s,x} (\phi_r^\delta)_x(\tau_r, \xi^\delta(\tau_r)) W(\tau_r)$$

where τ_r is the exit time from Q_r with $Y = Y^*$, the optimal control law corresponding to ϕ_r^δ . Since $(\phi_r^\delta)_x(T; \xi^\delta(T)) = 0$, $|L_x| \leq \alpha_1 L + \alpha_2$ for suitable α_1, α_2 and W is bounded, we have

$$|(\phi_r)_x(s, x)| \leq \alpha_1 \phi_r^\delta(s, x) + \alpha_2(T-s) + \max_{|x|=r} |(\phi_r^\delta)_x| P\{\tau_r < T\}.$$

Let N_r be a number such that

$$\phi_r^\delta(s, x) \leq N_r(r - |x|)$$

whenever $|x| < r$. Then

$$(3.6) \quad |(\phi_r^\delta)_x(s, x)| \leq \alpha_1 \phi_r^\delta(s, x) + \alpha_2(T-s) + N_r P\{\tau_r < T\}.$$

In order to show that $N_r P\{\tau_r < T\}$ is uniformly bounded with respect to $r > r_0$, we have to estimate N_r . Given x take x^0 with $|x^0| = r$, $|x - x^0| = r - |x|$. Let $\nu = -x^0/r$; we construct a barrier θ at (s, x^0) as follows:

$$\theta(s, x) = e^{T-s} (1 - e^{-k_r \nu \cdot (x - x^0)})$$

where k_r is the positive root of

$$c_r k^2 - M_r k - 1 = 0$$

where M_r is a bound of $|At + \delta g(x) + BY^*|$ whenever $|x| \leq r$ and c_2 is defined in (AIV). By straightforward calculation, we have

$$\theta_s + \frac{1}{2} \text{tr} \{\sigma \sigma' \theta_{xx}\} + \theta_x \cdot (Ax + \delta g + BY^*) \leq -1$$

and $\theta \geq 0$ on Q_r . By the maximum principle,

$$\phi_r^\delta(s, x) \leq \left(\max_{Q_r} L^{Y^*} \right) \theta.$$

Moreover, since $\theta(s, x^0) = 0$ and $r - |x| = |x - x^0|$,

$$\theta(s, x) \leq \left(\max_{Q_r} |\theta_x| \right) (r - |x|),$$

$$\theta(s, x) \leq e^T k_r (r - |x|).$$

Since K is compact, $k_r \leq D_1(1+r)^l$ for some positive constant D_1 and $L^Y \leq D_2(1+r)^2$, therefore

$$\phi_r^\delta(s, x) \leq D_3(1+r)^{2+l}$$

for some $D_3 > 0$. We take $N_r = D_3(1+r)^{2+l}$. Finally,

$$P\{\tau_r < T\} \leq r^{-\lambda} E_{s,x} \max_{s \leq t \leq T} |\xi^\delta(t)|^\lambda.$$

By (2.4), we have

$$P\{\tau_r < T\} \leq r^{-\lambda} C_\lambda'' (1 + |x|^\lambda),$$

where C_λ'' does not depend on r . If we take $\lambda \geq 2+l$ and recall (3.3), this proves the lemma.

Standard estimates for second order parabolic equations (see [5, pp. 60, 65, 191]) and passages to the limit then imply the desired properties of ϕ^δ . The technical details of the argument are similar to the proof of Theorem VI.6.2, of [5].

We have shown that $\phi^\delta \in \mathcal{F}_0$ and $Y^* \in \mathcal{Y}(K)$. Hence by the verification theorem, $\phi^\delta(s, x)$ is the minimum values of $J^\delta(s, x; Y)$.

The following lemma is a probabilistic representation of ϕ_x^δ .

LEMMA 3.2. $\phi_x^\delta(s, x) = E_{s,x} \int_s^T L_x(t, \xi^{\delta*}(t), u^*(t)) W(t) dt$, where $W(t)$ is defined in § 2.

Proof. By (3.6), $N_r P_r(\tau_r < T) \rightarrow 0$ as $r \rightarrow \infty$ and $\phi_r^\delta = \lim_{r \rightarrow \infty} \phi_{rx}^\delta$, we have

$$|\phi_x^\delta| \leq \alpha_1 \phi^\delta + \alpha_2 (T - S).$$

Thus, using the boundedness of W , which was proved in § 2, we obtain

$$E_{s,x} |\phi_x^\delta(\tau^r, \xi^\delta(\tau^r)) W(\tau^r)| \leq \text{const.} (1+r)^2 P(\tau_r < T).$$

Hence

$$E_{s,x} |\phi_x^\delta(\tau^r, \xi^\delta(\tau^r)) W(\tau^r)| \rightarrow 0 \quad \text{as } r \rightarrow \infty.$$

By Lemma 5.3 of [2] on ϕ^δ and $\tau^r \rightarrow T$ as $r \rightarrow \infty$, we have

$$\phi_x^\delta(s, x) = E_{s,x} \int_s^T L_x(t, \xi^{\delta*}(t), u^*(t)) W(t) dt. \quad \text{Q.E.D.}$$

From Lemma 3.2 and Lemma 2.2, there exists a positive constant ν independent of K such that

$$(3.7) \quad |\phi_x^\delta(s, x)| \leq \nu(1 + |x|).$$

Let us now consider the case $K = R^k$. For $m = 1, 2, \dots$, let

$$K^m = \{Y \mid |Y| \leq m\}.$$

Consider the corresponding $H^{\delta m}$ in (3.1) and solution $\phi^{\delta m}$ found by the previous method. Then $\phi^{\delta 1} \geq \phi^{\delta 2} \geq \dots \geq 0$. Let

$$\phi^\delta = \lim_{m \rightarrow \infty} \phi^{\delta m}.$$

By (3.3) and (3.7), $\phi^{\delta m}$ and $\phi_x^{\delta m}$ are uniformly bounded on each compact set. Since $H^{\delta m}$ tends to H^δ as $m \rightarrow \infty$, the same reasoning indicated right after the proof of Lemma 3.1 shows that $\phi^\delta \in \mathcal{F}_0$ and satisfies (3.1). It remains to show the corresponding optimal control policy $Y^{\delta*}$ satisfies (2.3) (c) and hence belongs to $\mathcal{Y}(R^k)$. Let $Y^m = Y^{\delta m*}$ be the optimal control function corresponding to $\phi^{\delta m}$. Then Y^m tends to $Y^{\delta*}$ as $m \rightarrow \infty$. We want to estimate $|Y^m|$. Given (s, x) , let $M(y) = y' N y + \phi_x^\delta \cdot B y$. Then M is minimum of K^m for $Y = Y^m = Y^m(s, x)$. Since 0 is an interior point of K , we have

$$M_y(Y^m) \cdot Y^m = \frac{d}{dz} M(z Y^m) \Big|_{z=1} \leq 0.$$

Therefore,

$$2 Y^m' N(t) Y^m + \phi_x^\delta \cdot B Y^m \leq 0.$$

Since $B(t)$ is bounded, we have for some positive constants ν_2

$$|Y^m| \leq \nu_2 |\phi_x^\delta|$$

By (3.5) then

$$|Y^m| \leq \nu_3 (1 + |x|)$$

where ν_3 does not depend on K^m . Therefore we obtain

THEOREM 3.1. *The function $\phi^\delta(s, x)$ defined by (1.6 $^\delta$) belongs to \mathcal{F}_0 and satisfies (3.1). The function $Y^{\delta*}(s, x)$ defined by (1.10) belongs to $\mathcal{Y}(R^k)$. Thus $Y^{\delta*}$ is optimal.*

Actually ϕ^δ is as smooth as we want (C^∞), since H^δ is C^∞ and we also have the Cauchy data; see [6].

4. Asymptotic formulas for $\phi^\delta, \phi_x^\delta$. We are now ready to consider the expansions of solution of the perturbed problem in terms of the solution of unperturbed problem. At the end of this section we also indicate how the methods tell the goodness of the policy Y^{0*} in the perturbed problem. Since $A(t), B(t), M(t)$ and $N(t)$ are C^∞ functions, then $\phi^\delta, \phi^0, Y^{\delta*}$ and Y^{0*} are C^∞ functions too.

LEMMA 4.1. $\phi^\delta(s, x) \rightarrow \phi^0(s, x)$.

Proof. Let $Y^{\delta*}, Y^{0*}$ be the controls corresponding to ϕ^δ, ϕ^0 , respectively, and $\xi^{\delta*}, \xi^{0*}$ be the corresponding Markov processes, respectively (given the same initial data (s, x)). Let ξ, ζ be the solutions of

$$\begin{aligned} d\xi(t) &= (A(t)\xi(t) + \delta g(\xi(t)) + B(t)Y^{0*}(t, \xi^{0*}(t))) dt + \sigma(t) dw(t), \\ d\zeta(t) &= (A(t)\zeta(t) + B(t)Y^{\delta*}(t, \xi^{\delta*}(t))) dt + \sigma(t) dw(t) \end{aligned}$$

with initial data $\xi(s) = \zeta(s) = x$. Suppose $X(t, v)$ is the principal matrix solution at initial time v of $dX/dt = A(t)X$; then we have

$$\xi^{\delta*} - \zeta = \delta \int_s^t X(t, v) g(\xi^{\delta*}(v)) dv$$

and it is easy to see $\xi^{\delta*} \rightarrow \zeta$ in probability as $\delta \rightarrow 0$. Similarly we have $\xi \rightarrow \xi^{0*}$ in probability as $\delta \rightarrow 0$. By definition of ϕ^δ, ϕ^0 we have

$$\begin{aligned} \phi^\delta(s, x) &= J^\delta(s, x; Y^{\delta*}) \leq J^\delta(s, x; Y^{0*}), \\ \phi^0(s, x) &= J^0(s, x; Y^{0*}) \leq J^0(s, x; Y^{\delta*}). \end{aligned}$$

Here we use Theorem 3.1 in [3] that the nonanticipative control does not do better than the optimal feedback control. Thus

$$J^\delta(s, x; Y^{\delta*}) - J^0(s, x; Y^{\delta*}) \leq \phi^\delta - \phi^0 \leq J^\delta(s, x; Y^{0*}) - J^0(s, x; Y^{0*}),$$

i.e.,

$$\begin{aligned} (4.1) \quad E_{s,x} \int_s^T [\xi^{\delta*'}(t)M(t)\xi^{\delta*}(t) - \zeta'(t)M(t)\zeta(t)] dt &\leq \phi^\delta - \phi^0 \\ &\leq E_{s,x} \int_s^T [\xi'(t)M(t)\xi(t) - \xi^{0*'}(t)M(t)\xi^{0*}(t)] dt. \end{aligned}$$

Since $E_{s,x}(\xi^{\delta*'}M\xi^{\delta*} - \zeta'M\zeta)^2$ and $E_{s,x}(\xi'M\xi - \xi^{0*'}M\xi^{0*})^2$ are bounded and the bounds do not depend on δ , we use Lebesgue dominated convergence theorem to get the result, since we have convergence in probability of $\xi^{\delta*} \rightarrow \zeta, \xi \rightarrow \xi^{0*}$. Q.E.D.

LEMMA 4.2. $\phi_x^\delta(s, x) \rightarrow \phi_x^0(s, x)$ uniformly on any compact set.

Proof. Since in (3.7), ν is independent of δ , (3.7) implies that ϕ_x^δ is uniformly bounded on any compact set. By Theorem 3.1, $\phi^\delta \in \mathcal{F}_0$. Moreover, we know $\phi^\delta \in C^\infty$. Hence ϕ_x^δ is equicontinuous on any compact set. By Ascoli's theorem, there exists a subsequence $\phi_x^{\delta_n}$ which converges uniformly to a limit ζ . Let us show that $\zeta = \phi_x^0$. Since

$$\int_{x_{0i}}^{x_i} \phi_{x_i}^{\delta_n} dx_i \rightarrow \int_{x_{0i}}^{x_i} \zeta_i dx_i$$

and using Lemma 4.1, we obtain

$$(4.2) \quad \int_{x_{0i}}^{x_i} \phi_{x_i}^{\delta_n} dx_i = \phi^{\delta_n}(s, x_i) - \phi^{\delta_n}(s, x_{0i}) \rightarrow \phi^0(s, x_i) - \phi^0(s, x_{0i}).$$

Then using fundamental theorem of calculus, we have $\zeta_{x_i} = \phi_{x_i}^0$ and hence the lemma. Q.E.D.

LEMMA 4.3. $\xi^{\delta*} \rightarrow \xi^{0*}$ in probability as $\delta \rightarrow 0$.

Proof. From (1.1 $^\delta$), (1.1 0) we have

$$d(\xi^{\delta*} - \xi^{0*}) = (A(\xi^{\delta*} - \xi^{0*}) + \delta g(\xi^{\delta*}) + B(Y^{\delta*}(t, \xi^{\delta*}) - Y^{0*}(t, \xi^{\delta*}) + Y^{0*}(t, \xi^{\delta*}) - Y^{0*}(t, \xi^{0*}))) dt$$

with $\xi^{\delta*}(s) - \xi^{0*}(s) = 0$. By (1.5)

$$Y^{0*}(t, \xi^{\delta*}) - Y^{0*}(t, \xi^{0*}) = N^{-1}(t)B'(t)K(t)(\xi^{\delta*}(t) - \xi^{0*}(t)).$$

Let $X(t, v)$ be the principal matrix solution at initial time v of

$$dX = (A(t) - N^{-1}(t)B'(t)K(t))X dt.$$

Then

$$\xi^{\delta*} - \xi^{0*} = \int_s^t X(t, v)(\delta g(\xi^{\delta*}(v)) + B(Y^{\delta*}(v, \xi^{\delta*}(v)) - Y^{0*}(v, \xi^{\delta*}(v)) + Y^{0*}(v, \xi^{\delta*}(v)) - Y^{0*}(v, \xi^{0*}(v)))) dv.$$

Since $E_{s,x}|g(\xi^{\delta*}(v))|$, $E_{s,x}|Y^{\delta*}(v, \xi^{\delta*}(v)) - Y^{0*}(v, \xi^{\delta*}(v))|^2$ are bounded independent of δ and $Y^{\delta*}(v, \xi^{\delta*}(v))$ approaches $Y^{0*}(v, \xi^{\delta*}(v))$ almost surely,

$$E|\xi^{\delta*} - \xi^{0*}| \rightarrow 0 \quad \text{as } \delta \rightarrow 0.$$

Thus

$$\xi^{\delta*} \rightarrow \xi^{0*} \quad \text{in probability as } \delta \rightarrow 0. \quad \text{Q.E.D.}$$

We now consider formula (1.11) with $k = 1$.

LEMMA 4.4. $\phi^\delta = \phi^0 + \delta\theta_1 + o(\delta)$ where $\delta^{-1}o(\delta) \rightarrow 0$ as $\delta \rightarrow 0$ uniformly on any compact set and θ_1 satisfies

$$(4.3^0) \quad (\theta_1)_s + \frac{1}{2} \text{tr} \{\sigma\sigma'(\theta_1)_{xx}\} + (\theta_1)_x \cdot f^{0, Y^{0*}} + \phi_x^0 \cdot g = 0$$

with the initial data $\theta_1(T, x) = 0$.

Proof. We can show $\theta_1(s, x)$ has the following form:

$$\theta_1(s, x) = E_{s,x} \int_x^T \phi_x^0(t, \xi^{0*}(t)) \cdot g(\xi^{0*}(t)) dt.$$

For $\delta > 0$, let

$$\theta_1^\delta = \delta^{-1}(\phi^\delta - \phi^0).$$

By (1.7 $^\delta$) and (1.7 0), θ_1^δ satisfies

$$(4.3^\delta) \quad (\theta_1^\delta)_s + \frac{1}{2} \text{tr} \{ \sigma \sigma' (\theta_1^\delta)_{xx} \} + (\theta_1^\delta)_x \cdot f^{\delta, (Y^{\delta*} + Y^{0*})/2} + \phi_x^0 \cdot g = 0.$$

Let ζ^δ be the solution of

$$d\zeta^\delta(t) = f^{\delta, (Y^{\delta*} + Y^{0*})/2}(t, \zeta^\delta) dt + \sigma(t) dw$$

with initial data $\zeta^\delta(s) = x$. Since $\frac{1}{2}(Y^{\delta*} + Y^{0*}) \in \mathcal{Y}$, the solution ζ^δ exists which is unique in the usual sense and all of its moments are bounded. In a way similar to the proof of Lemma 4.3, we can show that $\zeta^\delta \rightarrow \xi^{0*}$ in probability as $\delta \rightarrow 0$. Also by Lemma 2.3,

$$\theta_1^\delta(s, x) = E_{s,x} \int_s^T \phi_x^0(t, \zeta^\delta(t)) \cdot g(\zeta^\delta(t)) dt$$

where $\phi_x \cdot g$ satisfies polynomial growth condition, i.e., $E_{s,x} |\phi_x^0(t, \zeta^\delta(t)) \cdot g(\zeta^\delta(t))|^2$ is bounded independent of δ . Hence

$$(4.4) \quad \lim_{\delta \rightarrow 0} E_{s,x} \int_s^T \phi_x^0(t, \zeta^\delta(t)) \cdot g(\zeta^\delta(t)) dt = E_{s,x} \int_s^T \phi_x^0(t, \xi^{0*}(t)) \cdot g(\xi^{0*}(t)) dt.$$

Thus $\theta_1^\delta \rightarrow \theta_1$ as $\delta \rightarrow 0$. The convergence is uniform on any compact set. Hence the lemma is proved. Q.E.D.

Now we consider formula (1.12) with $k = 1$.

LEMMA 4.5. $\delta_x^\delta = \phi_x^0 + \delta(\theta_1)_x + o(\delta)$ where $\delta^{-1}o(\delta) \rightarrow 0$ as $\delta \rightarrow 0$ uniformly on any compact set.

Proof. It is equivalent to show $(\theta_1^\delta)_x \rightarrow (\theta_1)_x$ as $\delta \rightarrow 0$. Using (4.3 $^\delta$), we find that $(\theta_1^\delta)_{x_i}$ satisfies

$$(4.5^\delta) \quad (\theta_1^\delta)_{x_i} + \frac{1}{2} \text{tr} \{ \sigma \sigma' (\theta_1^\delta)_{xx} \} + (\theta_1^\delta)_x \cdot f^{\delta, Y^{\delta*}} + (\theta_1^\delta)_x \cdot f_{x_i}^{\delta, Y^{0*}} + (\phi_x^0 \cdot g)_{x_i} = 0$$

with $\theta_{1x_i}^\delta(T, x) = 0$. Let W^δ be the principal matrix solution at initial time s of

$$dW^\delta = f^{\delta, Y^{0*}}(t, \xi^{\delta*}(t)) W^\delta dt.$$

In fact, $f_x^{\delta, Y^{0*}} = A + \delta g_x - BN^{-1}B'K$. Using a proof similar in technique to that for the boundedness of W , we can show that W^δ is bounded and the bound does not depend on x . In a way similar to the proof of Lemma 3.2 we can show

$$(4.6^\delta) \quad (\theta_1^\delta)_x(s, x) = E_{s,x} \int_s^T (\phi_x^0 \cdot g)_x W^\delta|_{(t, \xi^{\delta*}(t))} dt.$$

Since moments of $\xi^{\delta*}(t)$ are bounded and $(\phi_x^0 \cdot g)_x W^\delta \in C_p$, then

$$(4.7) \quad \lim_{\delta \rightarrow 0} (\theta_1^\delta)_x(s, x) = E_{s,x} \int_s^T (\phi_x^0 \cdot g)_x W^0|_{(t, \xi^{0*}(t))} dt$$

where W^0 is the principal matrix solution at initial time s of

$$dW^0 = (A - BN^{-1}B'K)W^0 dt.$$

It is easy to see the right-hand side of (4.7) is just $(\theta_1)_x(s, x)$. Indeed, from (4.3⁰) we have

$$(\theta_{1x_i})_s + \frac{1}{2} \text{tr} \{ \sigma \sigma' (\theta_{1x_i})_{xx} \} + (\theta_{1x_i})_x \cdot f^{0, Y^{0*}} + (\theta_1)_x \cdot f_{x_i}^{0, Y^{0*}} + (\phi_x^0 \cdot g)_{x_i} = 0.$$

By a similar procedure

$$(\theta_1)_x(s, x) = E_{s,x} \int_s^T (\phi_x^0 \cdot g)_x W^0|_{(t, \xi^{0*}(t))} dt.$$

Hence $(\theta_1^\delta)_x \rightarrow (\theta_1)_x$ as $\delta \rightarrow 0$ uniformly on any compact set. Q.E.D.

LEMMA 4.6. $\phi^\delta = \phi^0 + \delta \theta_1 + \delta^2 \theta_2 + o(\delta^2)$ where $\delta^{-2} o(\delta^2) \rightarrow 0$ as $\delta \rightarrow 0$ uniformly on any compact set and θ_2 is defined by

$$(4.8^0) \quad (\theta_2)_s + \frac{1}{2} \text{tr} \{ \sigma \sigma' (\theta_2)_{xx} \} + (\theta_2)_x \cdot f^{0, Y^{0*}} + \frac{1}{2} \theta_{1x} H_{pp} \theta_{1x} + (\theta_1)_x \cdot g = 0$$

with initial data $\theta_2(T, x) = 0$.

Proof. Let

$$\theta_2^\delta = \delta^{-1} (\theta_1^\delta - \theta_1) = \delta^{-2} (\phi^\delta - \phi^0 - \delta \theta_1).$$

Then the problem is equivalent to $\theta_2^\delta \rightarrow \theta_2$ as $\delta \rightarrow 0$. By (4.3⁰), (4.3⁰), θ_2^δ satisfies

$$(4.8^\delta) \quad (\theta_2^\delta)_s + \frac{1}{2} \text{tr} \{ \sigma \sigma' (\theta_2^\delta)_{xx} \} + (\theta_2^\delta)_x \cdot f^{\delta, (Y^{\delta*} + Y^{0*})/2} + \frac{1}{2} (\theta_1)_x H_{pp}^0 (\theta_1^\delta)_x + (\theta_1)_x \cdot g = 0.$$

Using Lemma 2.3, we obtain

$$(4.9^\delta) \quad \theta_2^\delta(s, x) = E_{s,x} \int_s^T \left(\frac{1}{2} \theta_{1x}(t, \xi^\delta(t)) H_{pp}^0 \theta_{1x}(t, \xi^\delta(t)) \right. \\ \left. + \theta_{1x}(t, \xi^\delta(t)) \cdot g(\xi^\delta(t)) \right) dt.$$

Since $\xi^\delta \rightarrow \xi^{0*}$ in probability as $\delta \rightarrow 0$ and by (4.6⁰), (4.6⁰) we can see that both $(\theta_1)_x$ and $(\theta_1^\delta)_x$ belong to C_p , i.e., the integrand of (4.9⁰) also belongs to C_p ; hence

$$\lim_{\delta \rightarrow 0} \theta_2^\delta(s, x) = E_{s,x} \int_s^T \left[\frac{1}{2} \theta_{1x}(t, \xi^{0*}(t)) H_{pp}^0 \theta_{1x}(t, \xi^{0*}(t)) \right. \\ \left. + \theta_{1x}(t, \xi^{0*}(t)) \cdot g(\xi^{0*}(t)) \right] dt$$

and the right-hand side is just $\theta_2(s, x)$. In fact from (4.8⁰),

$$(4.9^0) \quad \theta_2(s, x) = E_{s,x} \int_s^T \left(\frac{1}{2} \theta_{1x}(t, \xi^{0*}(t)) H_{pp}^0 \theta_{1x}(t, \xi^{0*}(t)) \right. \\ \left. + \theta_{1x}(t, \xi^{0*}(t)) \cdot g(\xi^{0*}(t)) \right) dt.$$

Therefore, $\theta_2^\delta \rightarrow \theta_2$ as $\delta \rightarrow 0$ uniformly on any compact set. Q.E.D.

We can continue the procedure and finally we have

THEOREM 4.1. *The expansions of (1.11), (1.12) are valid for any k to l and hold uniformly on any compact set.*

COROLLARY 4.1.

$$Y^{\delta*}(s, x) = Y^{0*}(s, x) - (\delta/2) N^{-1}(s) B'(s) (\theta_1)'_x(s, x) \\ - \dots - (\delta^k/2) N^{-1}(s) B'(s) (\theta_k)'_x(s, x) + o(\delta^k)$$

where $k \leq l$.

Proof. Use (2.2) and Theorem 4.1. Q.E.D.

Now we consider the goodness of $Y^{0*}(s, x)$ in the perturbed problem. By Corollary 4.1 we know Y^{0*} gives approximately the optimal control policy in the perturbed problem for small δ . It is also plausible that Y^{0*} should give approximately the optimum in the perturbed problem. The above lemmas and their method of proof put this rough statement on a quantitative basis. Let

$$\Phi^\delta(s, x) = J^\delta(s, x; Y^{0*}).$$

In particular, $\Phi^0(s, x) = \phi^0(s, x)$. For $\delta > 0$, $\Phi^\delta(s, x) - \phi^\delta(s, x)$ represents how much Y^{0*} fails to be optimal in the perturbed problem. It is known that $\Phi^\delta \in \mathcal{F}_0$ and satisfies the linear parabolic equation

$$(4.10) \quad (\Phi^\delta)_s + \frac{1}{2} \text{tr}(\sigma\sigma' \Phi^\delta_{xx}) + \Phi^\delta_x \cdot f^{\delta, Y^{0*}} + x' Mx + Y^{0*'} N Y^{0*} = 0$$

with initial data $\Phi^\delta(T, x) = 0$. Let us write

$$(4.11) \quad \Phi^\delta = \phi^0 + \delta\chi_1 + \delta^2\chi_2 + o(\delta^2).$$

By the same procedure as before, we have for $k = 1, 2$

$$(4.12) \quad (\chi_k)_s + \frac{1}{2} \text{tr}(\sigma\sigma' (\chi_k)_{xx}) + (\chi_k)_x \cdot f^{0, Y^{0*}} + (\chi_{k-1})_x \cdot g = 0$$

where $\chi_0 = \phi^0$. Let $\chi_1^\delta = \delta^{-1}(\Phi^\delta - \phi^0)$, $\chi_2^\delta = \delta^{-2}(\Phi^\delta - \phi^0 - \delta\chi_1)$. Hence for $k = 1, 2$

$$(\chi_k^\delta)_s + \frac{1}{2} \text{tr}(\sigma\sigma' (\chi_k^\delta)_{xx}) + (\chi_k^\delta)_x \cdot f^{\delta, Y^{0*}} + (\chi_{k-1})_x \cdot g = 0.$$

Then

$$\begin{aligned} \chi_1 &= \theta_1, \\ \chi_2 &= E_{s,x} \int_s^T \theta_{1x}(t, \xi^0(t)) \cdot g(\xi^0(t)) dt. \end{aligned}$$

By the same procedure, we can prove $\chi_1^\delta \rightarrow \chi_1$, $\chi_2^\delta \rightarrow \chi_2$ as $\delta \rightarrow 0$ uniformly on any compact set. By comparing with Lemma 4.6, we find that

$$(4.13) \quad \Phi^\delta(s, x) - \phi^\delta(s, x) = -\frac{1}{2} \delta^2 E_{s,x} \int_s^T \theta_{1x}(t, \xi^0(t)) H_{pp}^0 \theta_{1x}(t, \xi^0(t)) dt + o(\delta^2).$$

Formula (4.13) shows that Y^{0*} gives within order the square of the intensity of perturbation δ of the optimum.

Example. Let ξ^δ be the solution of the scalar Itô equation

$$d\xi^\delta = (-\delta(\xi^\delta)^3 + u(t)) dt - \sigma dw$$

with $\xi^\delta(x) = x$. The control set is R . The criterion of performance is

$$J(s, x; u) = E_{s,x} \int_s^1 (\xi^{\delta^2} + u^2) dt.$$

Here $g_x(x) = -3x^2$. It is bounded above so that Theorem 4.1 and Corollary 4.1 can be applied. By [5, § 6.5], we have

$$\phi^0(s, x) = K(s)x^2 + q(s)$$

where $K(s) = \tanh(T-s)$ and $q(s) = \sigma^2 \ln \cosh(T-s)$. By Theorem 4.1

$$\begin{aligned}\theta_1(s, x) &= -2 \int_s^T k(t) E_{s,x} \xi^{04}(t) dt \\ &= -2(\beta_1(s)x^4 + \beta_2(s)x^2 + \beta_3(s))\end{aligned}$$

where

$$\begin{aligned}\beta_1(s) &= \frac{1}{4} \left(1 - \frac{1}{\cosh^4(T-s)} \right), \\ \beta_2(s) &= \frac{6\sigma^2}{4\cosh^2(T-s)} \left(\tanh(T-s)(\cosh^4(T-s) - 1) \right. \\ &\quad \left. - \frac{1}{8} \sin 4(T-s) + \frac{T-s}{2} \right), \\ \beta_3(s) &= 6\sigma^4(\tanh^2(T-s)(\cosh^4(T-s) - 1) \\ &\quad - \tanh(T-s) \left(\frac{1}{8} \sin^4(T-s) - \frac{T-s}{2} \right) \\ &\quad + \frac{1}{8} \left(1 - \cosh^2(T-s) + \frac{1}{2\cosh^2(T-s)} (\cosh^4(T-s) - 1) \right)).\end{aligned}$$

Then

$$\begin{aligned}\phi^\delta(x, x) &= \phi^0(s, x) + \delta\theta_1(s, x) + o(\delta), \\ Y^{\delta*}(s, x) &= Y^{0*}(s, x) + \delta(4\beta_1(s)x^3 + 2\beta_2(s)x) + o(\delta).\end{aligned}$$

5. Acknowledgment. The author would like to thank W. H. Fleming for his guidance and encouragement through the course of this study.

REFERENCES

- [1] J. M. BISMUT, *Theorie probabilistic du contrôle des diffusions*, Mem. Amer. Math. Soc., (1976), no. 167.
- [2] W. H. FLEMING, *Stochastic control for small noise intensities*, this Journal, 9 (1971), pp. 483–517.
- [3] ———, *Duality and a priori estimates in Markovian optimization problems*, J. Math. Anal. Appl., 16 (1966) pp. 254–279; Erratum, Ibid., 19 (1966), p. 204.
- [4] ———, *Stochastically perturbed dynamic system*, Rocky Mountain J. Math., 4 (1974), no. 3, pp. 407–433.
- [5] W. H. FLEMING AND R. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York 1975.
- [6] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.
- [7] H. J. KUSHNER, *Stochastic Stability and Control*, Academic Press, New York, 1967.
- [8] ———, *Stability and existence of diffusions with discontinuous or rapidly growing drift terms*, J. Differential Equations, 11 (1972), no. 1, pp. 156–168.
- [9] V. B. KOLMANOVSKII, *Optimal control of certain quasilinear stochastic systems*, Prikl. Math. Meh., 39 (1975), no. 4, pp. 724–727; J. Appl. Math. Mech., 39 (1975), no. 4, pp. 692–696.
- [10] Y. NISHIKAWA, N. SANNOMIYA AND H. ITAKURA, *Power-series approximations for the optimal feedback control noisy nonlinear systems*, J. Optimization Theory Appl., 19 (1976), no. 4, pp. 633–644.

A STOCHASTIC GAME MODEL OF A WEAPONS DEVELOPMENT COMPETITION*

WAYNE WINSTON†

Abstract. We model a weapons development competition between two countries as a two person zero-sum stochastic game. Conditions are given which ensure that each player's optimal strategy is monotone in the state, horizon length and discount factor. A non-zero-sum stochastic game model of a weapons development competition is also discussed.

1. Introduction. In this paper we investigate the structure of optimal policies for discrete and continuous time two person, zero-sum stochastic games which model a weapons development competition between two nations. Utilization of mathematical models to explain the dynamics of arms races began with Richardson [4]. To date, however, very few arms race models have employed game theory, and those that do have been deterministic models (see [7] and [8]). While our model has some defects, we feel it does an adequate job of incorporating the effects of uncertainty on a nation's weapons expenditures.

In § 2 we begin by developing a discrete time two person zero-sum stochastic game model of an arms race. Optimal strategies for both finite and infinite horizon models are characterized in § 3. Finally, the arms competition between M nations is modeled in § 4 as a non-zero-sum M -person continuous time stochastic game. For a special case of this model, a Nash equilibrium point is explicitly computed.

2. The model. We consider a weapons development competition between countries, country 1 and country 2. During any period t ($t = 1, 2, \dots$) we let i denote the 'state' of the weapons development competition. If the state during a period is i , this means that the difference between country 1's and country 2's weapons level (measured with reference to country 1) is given by i . It will be assumed that i is an integer, and the set of all nonnegative integers will be denoted by I . Also, for any positive integer k we let $I_k^+ = \{1, \dots, k\}$ and $I_k^- = \{0, -1, \dots, -k\}$.

During each period both countries must decide how much to invest on weapons development. Country 1's weapons development expenditures must be a member of $A^1 = \{b(1), b(2), \dots, b(l)\}$ ($b(1) < b(2) < \dots < b(l)$) while country 2's expenditures must be a member of $A^2 = \{c(1), c(2), \dots, c(m)\}$ ($c(1) < c(2) < \dots < c(m)$). If the state during a period is N^* we assume country 1 has 'won' the competition and country 2 must pay country 1 an amount $R(N^*)$ during each successive period. Similarly, if the state during a period is N_* , we assume country 2 has 'won' the arms race and country 1 must pay country 2 an amount $R(N_*)$ (probably negative) during each successive period. We allow for the cases where N^* and N_* do not exist by writing $N^* = \infty$ and $N_* = -\infty$.

To give a possible interpretation of N^* and N_* suppose country 1 is the U.S.S.R. and country 2 is the U.S.A. and we are concerned with nuclear weapons. Then N^* might correspond to a level of nuclear superiority which would enable the U.S.S.R. to launch a successful 'first strike' which would result in a 'Pax Soviet', while N_* might correspond to a level of American nuclear superiority which would enable us to 'make the world safe for democracy.'

Consider a period t during which the state is i ($N_* < i < N^*$) and country 1 spends $b(j)$ on weapons development and country 2 spends $c(k)$ on weapons development.

* Received by the editors November 30, 1976 and in revised form August 1, 1977.

† School of Business, Indiana University, Bloomington, Indiana 47401.

Then with probability $\lambda(j)$ the state during period $t+1$ will be $i+1$ (corresponding to the development of a new weapon by country 1), with probability $\mu(k)$ the state during period $t+1$ (corresponding to the development of a new weapon by country 2) will be $i-1$, and with probability $1-\lambda(j)-\mu(k)$ the state during period $t+1$ will be i [of course, we require $\lambda(j)+\mu(k)\leq 1$]. Thus we are modeling the progress of an arms race as a controlled random walk in which an upward transition corresponds to the development of a new weapon by country 1 and a downward transition corresponds to the development of a new weapon by country 2. In our model at most one transition per period is permissible. This will be a reasonable assumption if the length of a period is chosen to be sufficiently small. If our state variable referred to absolute armament levels rather than weapons development a random walk model would be less appropriate, since transitions into nonadjacent states would clearly be possible. Unfortunately, no analytical results have been obtained for models in which transitions to nonadjacent states are permissible. We also note that other random walk models might be more appropriate. For example, the assumption that the probability of a transition from i to $i+1$ (resp. $i-1$) is given by $\lambda(j)/[\lambda(j)+\mu(k)+z]$ (resp. $\mu(k)/[\lambda(j)+\mu(k)+z]$), $z>0$, incorporates the fact that the effects of the expenditures of one country are influenced by the action of the other country. Unfortunately, this interaction has made it impossible for us to obtain any results for this model.

Finally, during any period in which the state is i country 1 receives a payoff of $R(i)$ from country 2. This $R(i)$ is a measure of how a nation's well-being depends on their relative position in the weapons development competition.

This model will now be formulated as a two person, zero-sum stochastic game. As defined by Shapley [6], a two person stochastic game is characterized by the following:

1. A state space S .
2. The finite set of actions A_s^v available to player v ($v=1, 2$) in state $s\in S$.
3. The reward or payoff $R_i(j, k)$ which is paid from player 2 to player 1 during a period in which the state is i and player 1 chooses action j and player 2 chooses action k . Rewards are discounted by a factor β .
4. A set of transition probabilities $\{p_{is}(j, k), i, s\in S, j\in A^1, k\in A^2\}$, where $p_{is}(j, k)$ is the probability that the state during period $t+1$ will be s given that the state during period t was i , player 1 chose action j , and player 2 chose action k .

To formulate our problem as a zero-sum stochastic game we assume that each country (player) wishes to maximize how they are doing vis à vis their opponent. Then our model is a zero sum stochastic game with state space $S = I_{N_*}^- \cup I_{N_*}^+$, action spaces $A_i^1 = \{b(1), b(2), \dots, b(l)\}$ and $A_i^2 = \{c(1), c(2), \dots, c(m)\}$, transition probabilities given by

$$\begin{aligned} p_{i,i+1}(b(j), c(k)) &= \lambda(j), & i \neq N_* \text{ or } N^*, \\ p_{i,i-1}(b(j), c(k)) &= \mu(k), & i \neq N_* \text{ or } N^*, \\ p_{ii}(b(j), c(k)) &= 1 - \lambda(j) - \mu(k), & i \neq N_* \text{ or } N^*, \\ p_{N_*N_*}(b(j), c(k)) &= p_{N^*N^*}(b(j), c(k)) = 1, \\ p_{ij}(b(j), c(k)) &= 0, & \text{otherwise,} \end{aligned}$$

and rewards

$$R_i(b(j), c(k)) = R(i) - b(j) + c(k).$$

Actually it would probably be more realistic to formulate our problem as a non-zero-sum game in which the one period payoff to country 1 is $R(i) - b(j)$ and the one period

payoff to country 2 is $-R(i) - c(k)$. Unfortunately, we have been unable to obtain any analytical results for such a formulation except for the case in which $R(i)$ is linear.

We define $\Delta(\Pi)$ to be the set of *stationary strategies* for country 1 (country 2). Then $\delta \in \Delta$ is a set of probability vectors of the form $\delta(i) = (\delta_1(i), \dots, \delta_l(i))$ where $\delta_j(i)$ is the probability that country 1 will spend $b(j)$ on arms during any period in which the state is i . Similarly $\pi \in \Pi$ is a set of probability vectors of the form $\pi(i) = (\pi_1(i), \pi_2(i), \dots, \pi_m(i))$, where $\pi_k(i)$ is the probability that country 2 will spend $c(k)$ on arms during any period in which the state is i .

A stationary strategy $\delta[\pi]$ is a pure strategy if for each $i \in S$ there exist $F(i) [\bar{F}(i)]$ such that $\delta_{F(i)}(i) = 1 [\pi_{\bar{F}(i)}(i) = 1]$. Let $V(i, \delta, \pi)$ be the expected discounted payoff received by country 1 during an infinite number of periods when country 1 follows stationary strategy δ , country 2 follows stationary strategy π , and the initial state is i . If the game is played for an infinite number of periods we seek stationary strategies $\hat{\delta} \in \Delta$ and $\hat{\pi} \in \Pi$ (termed infinite horizon optimal) which satisfy

$$(1) \quad V(i, \hat{\delta}, \pi) \geq V(i, \hat{\delta}, \hat{\pi}) \geq V(i, \delta, \hat{\pi}), \quad \pi \in \Pi, \quad \delta \in \Delta, \quad i \in S.$$

We will write $V(i, \hat{\delta}, \hat{\pi}) = V(i)$ and call $V(i)$ the *value* of the game to country 1 if the initial state is i . For S finite, the existence of infinite horizon optimal strategies was demonstrated by Shapley [6]. Shapley also proved that $V(i)$ does not depend on the choice of $\hat{\delta}$ and $\hat{\pi}$. Since the action spaces in our problem are finite, the results of Denardo [2] show that if S is finite and the game is played for an infinite number of periods, our restriction to stationary strategies is without loss of generality. For infinite horizon problems, we therefore restrict our attention to stationary strategies.

Suppose the game is to be played for $T < \infty$ periods. For any $\delta = (\delta^{T_1}, \dots, \delta^{T_T}) \in \times_{t=1}^T \Delta = \Delta^T$ and $\pi = (\pi^{T_1}, \dots, \pi^{T_T}) \in \times_{t=1}^T \Pi = \Pi^T$ let $V_t^T(i, \delta, \pi)$ be the discounted expected payoff accruing to country 1 during periods $t, t+1, \dots, T$ when country 1 follows δ^{T_k} during period k , country 2 follows π^{T_k} during period k and the state is i at the beginning of period t . We seek $\hat{\delta}^T \in \Delta^T$ and $\hat{\pi}^T \in \Pi^T$ (termed T -period optimal) satisfying

$$(2) \quad V_t^T(i, \hat{\delta}^T, \pi^T) \geq V_t^T(i, \hat{\delta}^T, \hat{\pi}^T) \geq V_t^T(i, \delta^T, \hat{\pi}^T), \\ \delta^T \in \Delta^T, \quad \pi^T \in \Pi^T, \quad i \in S, \quad t = 1, 2, \dots, T.$$

It is well known that T -period optimal policies exist and may be computed by backwards induction. Since our game is zero-sum, $V_t^T(i, \hat{\delta}^T, \hat{\pi}^T)$ is independent of the choice of $\hat{\delta}^T$ and $\hat{\pi}^T$. We therefore write $V_t^T(i, \hat{\delta}^T, \hat{\pi}^T) = V_t^T(i)$. We define $V_{T+1}^T(i) = 0$. For T -period optimal strategies $\hat{\delta}^T$ and $\hat{\pi}^T$ define $\hat{\delta}^{T_t}(i) = (\hat{\delta}_1^{T_t}(i), \dots, \hat{\delta}_l^{T_t}(i))$ and $\hat{\pi}^{T_t}(i) = (\hat{\pi}_1^{T_t}(i), \dots, \hat{\pi}_m^{T_t}(i))$, where $\hat{\delta}_j^{T_t}(i) [\hat{\pi}_k^{T_t}(i)]$ is the probability that country 1 (country 2) will spend $b(j)$ ($c(k)$) during period t given that country 1 (country 2) is following $\hat{\delta}^T$ ($\hat{\pi}^T$), and the state during period t is i . Finally, let $\Delta V_t^T(i) = V_t^T(i+1) - V_t^T(i)$ and define for $\hat{\delta} = (\delta_1, \dots, \delta_l)$ and $\hat{\pi} = (\pi_1, \dots, \pi_m)$

$$(3) \quad W_t^T(i, \hat{\delta}, \hat{\pi}) = \sum_{j=1}^l \sum_{k=1}^m \delta(j) \pi(k) W_t^T(i, j, k),$$

where

$$W_t^T(i, j, k) = R(i) - b(j) + c(k) \\ + \beta \{ \lambda(j) \Delta V_{t+1}^T(i) - \mu(k) \Delta V_{t+1}^T(i-1) + V_{t+1}^T(i) \}, \quad N_* < i < N^*, \\ W_t^T(N_*, j, k) = R(N_*) - b(j) + c(k) + \beta V_{t+1}^T(N_*), \\ W_t^T(N^*, j, k) = R(N^*) - b(j) + c(k) + \beta V_{t+1}^T(N^*).$$

3. Structure of optimal policies: Discrete time. In this section we derive the structure of optimal strategies when the game of § 2 is played for $T < \infty$ periods. These results are then utilized to characterize the structure of infinite horizon optimal stationary strategies. Our first result follows easily from (2) and (3).

THEOREM 1. *For country 1 (country 2) there exists a T -period optimal strategy $\delta^T(\pi^T)$ such that for $t = 1, 2, \dots, T$, $\delta^{Tt}(\tilde{\pi}^{Tt})$ is a pure strategy.*

Proof. Define $j_t^T(i)$ and $k_t^T(i)$ by

$$(4) \quad \beta\lambda[j_t^T(i)] \Delta V_{t+1}^T(i) - b[j_t^T(i)] = \max_{j \in I_1^+} \{\beta\lambda(j) \Delta V_{t+1}^T(i) - b(j)\},$$

$$(5) \quad c[k_t^T(i)] - \beta\mu[k_t^T(i)] \Delta V_{t+1}^T(i-1) = \min_{k \in I_m^+} \{c(k) - \beta\mu(k) \Delta V_{t+1}^T(i-1)\}.$$

It follows from (2)–(5) that it is optimal for player 1 [2] to always choose $j_t^T(i)$ [$k_t^T(i)$] when the state during period t is i .

For the sake of definiteness we now let $j_t^T(i)$ and $k_t^T(i)$ be the smallest values of j and k attaining the extrema in (4) and (5). For all t and T it is clear that $j_t^T(N^*) = k_t^T(N^*) = j_t^T(N_*) = k_t^T(N_*) = 1$. We now characterize the dependence of these quantities on t , β , and i . Our results require that $R(\cdot)$, $\lambda(\cdot)$, $\mu(\cdot)$ all be nondecreasing functions. Assuming this, we can now prove

THEOREM 2. *For $i \in S$ and $t = 1, 2, \dots, T-1$, $j_t^T(i) \geq j_{t+1}^T(i)$ and $k_t^T(i) \geq k_{t+1}^T(i)$.*

Proof. Together (2)–(5) imply that it suffices to show that for $j_2 > j_1$ and $N_* < i < N^*$,

$$W_{t+1}^T(i, j_2, k_{t+1}^T(i)) + W_t^T(i, j_1, k_t^T(i)) \leq W_{t+1}^T(i, j_1, k_{t+1}^T(i)) + W_t^T(i, j_2, k_t^T(i)),$$

and for $k_2 > k_1$,

$$W_{t+1}^T(i, j_{t+1}^T(i), k_2) + W_t^T(i, j_t^T(i), k_1) \geq W_{t+1}^T(i, j_{t+1}^T(i), k_1) + W_t^T(i, j_t^T(i), k_2).$$

Both these inequalities will follow if $\Delta V_t^T(i) \geq \Delta V_{t+1}^T(i)$. Since $R(i)$ is nondecreasing, this inequality is clearly true for $t = T-1$. We therefore assume that it is valid for $t+1$ and verify it for t . From (2) it follows that

$$\Delta V_t^T(i) \geq W_t^T(i+1, j_{t+1}^T(i+1), k_t^T(i+1)) - W_t^T(i, j_t^T(i), k_{t+1}^T(i)),$$

and

$$\Delta V_{t+1}^T(i) \leq W_{t+1}^T(i+1, j_{t+1}^T(i+1), k_t^T(i+1)) - W_{t+1}^T(i, j_t^T(i), k_{t+1}^T(i)).$$

The desired result now follows from these inequalities, the induction hypothesis, and $\lambda(\cdot) + \mu(\cdot) \leq 1$.

Our subsequent results require the following lemma:

LEMMA 1. *For $t = 1, 2, \dots, T$ and $i \in S - \{N^*\}$, $\Delta V_t^T(i) \geq 0$.*

Proof. The result is trivially true for $t = T$ and $t = T-1$, so we assume it to be valid for $t+1$ and verify it for t . From (2), the induction hypothesis, and the fact that $R(i)$ is nondecreasing it follows that

$$\Delta V_t^T(i) \geq W_t^T(i+1, j_t^T(i), k_t^T(i+1)) - W_t^T(i, j_t^T(i), k_t^T(i+1)) \geq 0.$$

Let $j_t^T(i)_\beta$, $k_t^T(i)_\beta$, and $V_t^T(i)_\beta$ indicate the dependence of these quantities on β . We can now prove

THEOREM 3. *For $\beta_2 > \beta_1$, $j_t^T(i)_{\beta_2} \geq j_t^T(i)_{\beta_1}$ and $k_t^T(i)_{\beta_2} \geq k_t^T(i)_{\beta_1}$.*

Proof. Reasoning analogous to that used to prove Theorem 2 shows that it suffices to prove $\beta_2 \Delta V_t^T(i)_{\beta_2} \geq \beta_1 \Delta V_t^T(i)_{\beta_1}$. Since $\Delta V_t^T(i)_\beta \geq 0$ the first inequality will

follow if $\Delta V_i^T(i)_{\beta_2} \geq \Delta V_i^T(i)_{\beta_1}$. The last inequality follows from a line of argument identical to that used to show that $\Delta V_i^T(i) \geq \Delta V_{i+1}^T(i)$.

Thus we see that the optimal arms expenditures of both nations are nondecreasing functions of the length of their planning horizon and the importance they attach to future developments (as reflected by the discount factor).

Our next result requires $N^* = \infty$ and $N_* = -\infty$.

THEOREM 4. *If $R(i)$ is convex (concave), then both $j_i^T(i)$ and $k_i^T(i)$ are nondecreasing (nonincreasing) functions of i .*

Proof. We consider the case where $R(i)$ is convex; that is,

$$\Delta R(i) = R(i+1) - R(i) \geq R(i) - R(i-1) = \Delta R(i-1).$$

Then the desired result follows if

$$\begin{aligned} W_i^T(i+1, j_2, k_i^T(i+1)) + W_i^T(i, j_1, k_i^T(i)) \\ \geq W_i^T(i+1, j_1, k_i^T(i+1)) + W_i^T(i, j_2, k_i^T(i)), \quad j_2 > j_1, \end{aligned}$$

and

$$\begin{aligned} W_i^T(i+1, j_i^T(i+1), k_2) + W_i^T(i, j_i^T(i), k_1) \\ \leq W_i^T(i+1, j_i^T(i+1), k_1) + W_i^T(i, j_i^T(i), k_2), \quad k_2 > k_1. \end{aligned}$$

These inequalities reduce to $\Delta V_{i+1}^T(i+1) \geq \Delta V_{i+1}^T(i)$ and $\Delta V_{i+1}^T(i) \geq \Delta V_{i+1}^T(i-1)$. It therefore suffices to prove that $V_i^T(i)$ is convex. For $t = T$ the convexity of $V_T^T(i)$ follows immediately from the convexity of $R(i)$. We therefore assume that $V_{i+1}^T(i)$ is convex and verify that $V_i^T(i)$ is convex.

If we define $j_1^* = j_i^T(i)$, $k_1^* = k_i^T(i+1)$, and $k_2^* = k_i^T(i-1)$, it follows from (2) that

$$\Delta V_i^T(i) \geq W_i^T(i+1, j_1^*, k_1^*) - W_i^T(i, j_1^*, k_1^*)$$

and

$$\Delta V_i^T(i-1) \leq W_i^T(i, j_1^*, k_2^*) - W_i^T(i-1, j_1^*, k_2^*).$$

After simplification the last two inequalities yield

$$\begin{aligned} & \Delta V_i^T(i) - \Delta V_i^T(i-1) \\ & \geq \Delta R(i) - \Delta R(i-1) \\ (6) \quad & + \beta \{ \lambda(j_1^*) \Delta V_i^T(i+1) + (1 - \mu(k_1^*) - 2\lambda(j_1^*)) \Delta V_i^T(i) \\ & + (\mu(k_1^*) + \lambda(j_1^*) + \mu(k_2^*) - 1) \Delta V_i^T(i-1) - \mu(k_2^*) \Delta V_i^T(i-2) \} \geq 0, \end{aligned}$$

where the last inequality follows from the convexity of $R(i)$, $\lambda(\cdot) + \mu(\cdot) \leq 1$, and the induction hypothesis.

The proof for the case where $R(i)$ is concave is similar and is therefore omitted.

By a method of proof analogous to that used to prove Theorem 4 the following results can be proven.

THEOREM 5. If $N_* = -\infty$ and $R(i)$ is convex, then $j_i^T(i)$ and $k_i^T(i)$ are nondecreasing functions of i .

THEOREM 6. If $N^* = +\infty$ and $R(i)$ is concave, then $j_i^T(i)$ and $k_i^T(i)$ are nonincreasing functions of i .

The hypotheses of Theorem 5 might be applicable to the U.S.S.R.–U.S.A. arms race rivalry if we assume that the Russians would conquer the world if the state is N^* and that the U.S.A. does not derive any benefit from arms superiority. If this were the case, then $R(i) = 0$, $i < N^*$, and $R(N^*) = M$, for some large M , would reasonably reflect reality.

We note that if $R(i)$ is linear, Theorem 4 implies that $j_i^T(i)$ and $k_i^T(i)$ are constant. Thus, if $R(i)$ is linear each player has a state-independent T -period optimal strategy.

Unfortunately, it seems unreasonable for $R(i)$ to be convex or concave (unless $R(i)$ is linear). For example, if we consider both countries to be identical, then $R(i) = -R(-i)$ should be valid. If $R(i)$ is not linear, however, such a payoff structure precludes $R(i)$ from being convex or concave. We can, however, obtain interesting results for the model described by the following parameters:

$$\begin{aligned}
 (7) \quad & R(i) = -R(-i), \quad i \geq 0, \\
 & 0 < N^* = -N_*, \\
 & A^1 = A^2 = \{b(1), b(2), \dots, b(l)\}, \\
 & \lambda(i) = \mu(i), \quad i = 1, 2, \dots, l.
 \end{aligned}$$

The model described by (7) represents weapons development competition between two identical countries. As another example for which (7) is an appropriate model consider two identical firms that produce the same product. Let i denote the measure of superiority of firm 1's product over firm 2's product. We assume the two firms are competing for a market of fixed size. If $R(i)$ reflects the dependence of firm 1's sales revenue less firm 2's sales revenue on the relative merits of the two firms' products, then the assumption of $R(i) = -R(-i)$ seems reasonable. In order to improve their product each firm may spend money on research, with improvement in the quality of their product being more likely if more money is spent on research. If each firm wishes to maximize their expected discounted profits vis à vis their opponent, then the zero-sum stochastic game associated with (7) is a reasonable model of the problem facing both firms.

For the model specified by (7) we now prove

THEOREM 7. If (7) is satisfied, then

$$(8) \quad j_i^T(i) = k_i^T(-i), \quad N^* \geq i \geq 0,$$

$$(9) \quad k_i^T(i) = j_i^T(-i), \quad N^* \geq i \geq 0$$

and

$$(10) \quad \Delta V_i^T(i) = \Delta V_i^T(-i-1), \quad N^*-1 \geq i \geq 0.$$

Proof. For $t = T$ the result follows directly from (7). We therefore assume that (8)–(10) are valid for $t+1$ and verify them for t . Combining (10) of the induction hypothesis with (4) and (5) immediately yields (8) and (9). To prove (10) define $j_1 = j_i^T(i+1)$, $j_2 = j_i^T(i)$, $k_1 = k_i^T(i+1)$, and $k_2 = k_i^T(i)$. Then for $1 \leq i \leq N^*-1$ (7)–(9)

imply

$$\begin{aligned}
 \Delta V_i^T(i) &= \Delta R(i) + b(j_1) - b(k_1) - b(j_2) + b(k_2) \\
 &\quad + \beta \{ \lambda(j_1) \Delta V_{i+1}^T(i+1) + (1 - \lambda(k_1) - \lambda(j_2)) \Delta V_{i+1}^T(i) \\
 &\quad \quad \quad + \lambda(k_2) \Delta V_{i+1}^T(i-1) \} \\
 &= W_{i+1}^T(-i, k_2, j_2) - W_{i+1}^T(-i-1, k_1, j_1) \\
 &\quad \quad \quad [\text{by (10) of the induction hypothesis}] \\
 &= \Delta V_i^T(-i-1) \quad [\text{by (7)–(9)}].
 \end{aligned}$$

For $i = N^* - 1$ and $i = 0$ the result follows in analogous fashion.

We now prove the analogue of Theorem 4 for a model satisfying (7).

THEOREM 8. *If (7) and $\Delta R(i) \geq \Delta R(i-1)$ ($i \geq 0$) are satisfied, then for $N^* - 1 \geq i \geq 0$*

$$(11) \quad j_i^T(i+1) \geq j_i^T(i),$$

$$(12) \quad k_i^T(i+1) \geq k_i^T(i),$$

and

$$(13) \quad \Delta V_i^T(i) \geq \Delta V_i^T(i-1).$$

Proof. It is trivial to see that (11)–(13) are valid for $t = T$. We therefore assume that (11)–(13) are valid for $t+1$ and demonstrate their validity for t . For $i > 0$, (11) and (12) follow from (4) and (5) plus (13) of the induction hypothesis; for $i = 0$, (12) follows from (4) and (5), plus (10). Inequality (13) follows from (6), (10), $\Delta R(i) \geq \Delta R(i-1)$, $\lambda(\cdot) + \mu(\cdot) \leq 1$, and the induction hypothesis.

If $\Delta R(i) \leq \Delta R(i-1)$ ($i \geq 0$), then a proof that is virtually identical to the proof of Theorem 7 shows that (11)–(13) are valid if ‘greater than or equal’ is replaced by ‘less than or equal’. If $R(i)$ is linear for $i \geq 0$, Theorems 4 and 7 imply the existence of \bar{j}_i^T and \bar{k}_i^T such that for $i \geq 0$,

$$j_i^T(i) = k_i^T(-1) = \bar{j}_i^T \quad \text{and} \quad k_i^T(i) = k_i^T(-i) = \bar{k}_i^T.$$

We now extend some of our results to the infinite horizon case. For technical reasons we now assume $N^* < \infty$ and $N_* > -\infty$.¹ Then the results of Denardo [2] imply

$$(14) \quad \lim_{T \rightarrow \infty} V_1^T(i) = V(i).$$

Theorems 1 and 2 plus the finiteness of each player’s state and action space ensure the existence of T_0 , $j(i)$, and $k(i)$ such that for $T > T_0$, $j(i) = j_1^T(i)$ and $k(i) = k_1^T(i)$. Let $\delta(\pi)$ be the pure stationary strategy for player 1 (player 2) for which player 1 (player 2) always takes action $j(i)$ ($k(i)$) whenever the state is i . It follows from (2) that for $T > T_0$, $\pi \in \Pi$, and $\delta \in \bar{\Delta}$, $W_1^T(i, \delta, \pi) \geq W_1^T(i, \delta, \pi) \geq W_1^T(i, \delta, \pi)$. Using (14) and letting $T \rightarrow \infty$ yields $V(i, \delta, \pi) \geq V(i, \delta, \pi) \geq V(i, \delta, \pi)$. This proves that δ and π are infinite horizon optimal strategies. The following results now follow directly from Theorems 1, 3, 7, and 8.

¹ The reason for the restriction to the finite state space is that (to the author’s knowledge) it is not known if an optimal stationary strategy exists for a stochastic game with a countable state space and unbounded payoffs.

THEOREM 9. *For each country there exists an infinite horizon optimal strategy that is a pure strategy.*

THEOREM 10. *For $\beta_2 > \beta_1$, $j(i)_{\beta_2} \geq j(i)_{\beta_1}$ and $k(i)_{\beta_2} \geq k(i)_{\beta_1}$.*

THEOREM 11. *If (7) is satisfied, then for $N^* \geq i \geq 0$, $j(i) = k(-i)$ and $k(i) = j(-i)$.*

THEOREM 12. *If (7) and $\Delta R(i) \geq \Delta R(i-1)$ ($i \geq 0$) are satisfied, then for $N^* - 1 \geq i \geq 0$, $j(i+1) \geq j(i)$ and $k(i+1) \geq k(i)$.*

4. A non-zero-sum game arms race model. We now develop a non-zero-sum continuous time stochastic game model of an arms race. For a simple payoff structure we give an explicit characterization of a discounted equilibrium point. The model considered here is similar to a model discussed in [1].

We assume that M countries are competing in an arms race. At an instant during which the arms level of country k is i_k the state of the world will be specified by (i_1, i_2, \dots, i_M) . At any instant country k must spend money on arms at a rate $d^k \in [d^k, \bar{d}^k] = A^k$. If country k spends money at a rate d^k , then the state changes to $(i_1, i_2, \dots, i_k + 1, \dots, i_M)$ according to an exponential distribution with rate $\lambda_k(d^k)$. At an instant during which the state is (i_1, i_2, \dots, i_M) country k earns a payoff at a rate $c_1^k i_1 + c_2^k i_2 + \dots + c_M^k i_M + b^k$. Payoffs and expenditures are continuously discounted at a rate α . Let F^k be the set of all probability measures on $B(A^k)$, the Borel subsets of A^k . Then a stationary strategy δ^k for player k is a mapping from $S = \{(i_1, i_2, \dots, i_M) : i_k \in I\}$ onto F^k where the measure δ_s^k is a randomized strategy for player k ; that is, for $B \in B(A^k)$, $\delta_s^k(B)$ is the probability that country k will spend at a rate $d^k \in B$ when the state is $s \in S$. Thus, $\Delta^k = \times_{s \in S} F^k$ is the set of all δ^k and $\Pi = \Delta^1 \times \Delta^2 \times \dots \times \Delta^M$ is the product set of all country's strategies.

For $\delta \in \Pi$, let $V^k(s, \delta)$ be the expected discounted reward earned over an infinite horizon by player k when all players follow δ and the initial state is s . We seek a $\delta \in \Pi$ that is stable in the sense that no player can benefit by a unilateral change in strategy. More precisely, a stationary strategy $\hat{\delta} = (\hat{\delta}^1, \hat{\delta}^2, \dots, \hat{\delta}^M)$ is said to be a *discounted equilibrium point* (DEP) if

$$\begin{aligned} V^k(s, (\hat{\delta}^1, \hat{\delta}^2, \dots, \hat{\delta}^k, \dots, \hat{\delta}^M)) &\geq V^k(s, (\hat{\delta}^1, \hat{\delta}^2, \dots, \delta^k, \dots, \hat{\delta}^M)), \\ k &\in \{1, 2, \dots, M\}, \\ \delta^k &\in \Delta^k, \text{ and } s \in S. \end{aligned}$$

The concept of a DEP for a non-zero-sum stochastic game was introduced by Rogers [5] and Sobel [9].

Define q^k to be any value attaining the maximum in

$$(15) \quad \max_{d^k \in A^k} (\lambda(d^k)c_k^k/\alpha - d^k).$$

Consider the strategy $\hat{\delta}$ for which $\hat{\delta}_s^k(q^k) = 1$ for all $s \in S$ and $k \in \{1, 2, \dots, M\}$. We will show that $\hat{\delta}$ is a DEP. Thus if each country always spends money on arms at a rate q^k no country can benefit by a unilateral shift to a different stationary strategy. Before proving this result, we require the following lemma.

LEMMA 2. *For $k = 1, 2, \dots, M$ and $s \in S$*

$$(16) \quad V^k(s, \hat{\delta}) = (b^k - q^k)/\alpha + \sum_{r=1}^{r=M} \lambda(q^r)c_r^k/\alpha^2 + \sum_{r=1}^{r=M} i_r c_r^k/\alpha.$$

Proof. From Theorem 1 of Lippman [3] and the results of Denardo [2] it follows that $V^k(s, \delta)$ is the (unique) solution to the following denumerably infinite system of equations.

$$(17) \quad \alpha V^k(s, \delta) = \sum_{r=1}^{r=M} i_r c_r^k - d^k + b^k + \sum_{r=1}^{r=M} \lambda(d^r) \Delta^r V^k(s, \delta),$$

$$s = (i_1, \dots, i_M) \in S,$$

where

$$\begin{aligned} \Delta^r V^k(s, \delta) &= V^k((i_1, \dots, i_r + 1, \dots, i_M), \delta) \\ &\quad - V^k((i_1, \dots, i_r, \dots, i_M), \delta). \end{aligned}$$

A direct substitution shows that $V^k(s, \delta)$ as specified by (15) satisfies (16). We now prove that δ is a DEP.

THEOREM 13. δ is a DEP.

Proof. By Lemma 1 of Lippman [3], the results of Denardo [2], the definition of a DEP, and Lemma 2, δ will be a DEP if for $k = 1, 2, \dots, M$ and $s \in S$, d^k attains the maximum in

$$\max_{d^k \in A^k} \left\{ \sum_{r=1}^{r=M} i_r c_r^k - d^k + b^k + \sum_{r \neq k} \lambda(d^r) \Delta^r V^k(s, \delta) + \lambda(d^k) \Delta^k V^k(s, \delta) \right\}.$$

By Lemma 2 and (16), d^k attains this maximum. Therefore δ is a DEP and the proof of the Theorem is complete.

REFERENCES

- [1] S. C. ALBRIGHT AND W. WINSTON, *A birth death model of advertising*, submitted to J. Economic Theory, July 1976.
- [2] E. V. DENARDO, *Contraction mappings in the theory underlying dynamic programming*, SIAM Review, 9 (1967), pp. 165–177.
- [3] S. LIPPMAN, *On dynamic programming with unbounded rewards*, Management Sci., 21 (1975), pp. 1225–1233.
- [4] L. F. RICHARDSON, *Arms and Insecurity*, The Boxwood Press, Chicago, 1960.
- [5] P. D. ROGERS, *Non-zero-sum stochastic games*, Rep. ORC 698, Operations Research Center, University of California, Berkeley, 1969.
- [6] L. SHAPLEY, *Stochastic games*, Proc. Nat. Acad. Sci. U.S.A., 39 (1953), pp. 1095–1100.
- [7] M. SIMAAN AND J. B. CRUZ, *A Differential game example of armament race*, Proceedings of the Seventh Annual Princeton Conference on Information Sciences and Systems, March 1973.
- [8] ———, *Nash equilibrium strategies in armament race and control*, Management Sci., 22 (1975), pp. 96–105.
- [9] M. SOBEL, *Noncooperative stochastic games*, Ann. Math. Statist., 42 (1971), pp. 1930–1935.

SEQUENTIAL DECISION PROBLEMS WITH EXPECTED UTILITY CRITERIA. III: UPPER AND LOWER TRANSIENCE*

DAVID M. KREPS†

Abstract. This paper concerns sequential decision problems with expected utility criteria. A class of such problems, which combines and extends the standard cases of transient, positive and negative dynamic programming is proposed and analyzed. The principal results are characterizations of when conserving and unimprovable strategies are optimal.

Introduction. This paper concerns sequential decision problems with expected utility criteria. These are the problems of choosing actions at a discrete sequence of times, thereby controlling a stochastic decision chain, so as to maximize the expectation of a utility function defined on the complete history (or path) of the chain. Such problems are analyzed in [9], [10], the former dealing with basic definitions and analogues to positive and negative dynamic programming and the latter with stationarity. In this paper, a class of problems is identified and analyzed which extends the standard case of transient dynamic programming (cf. [2], [6], [15]) and combines this case with the cases of positive and negative dynamic programming. General conditions are provided under which conserving and unimprovable strategies are optimal in such problems.

Pertinent definitions and results from [9] are given in § 1. (The reader is strongly advised to read that paper before proceeding.) In § 2, definitions of *upper* and *lower transient* strategies and problems are given. These generalize negative and positive dynamic programming respectively, but with conditions imposed jointly on the chain being controlled and the utility function. Then mathematical analysis of such strategies and problems is given. The principal results are: Conserving upper transient strategies are optimal. If one strategy is a “one step improvement” on a second and the first is upper transient, then the first is “better than” the second. (See § 1 for definitions of these terms.) Unimprovable strategies are “as good as” all lower transient strategies—if the problem is lower transient, then unimprovable strategies are optimal. Section 2 concludes with interpretations of the definitions. In § 3, some examples of upper and lower transient problems are given. Problems with upper and lower convergent utility are upper and lower transient, respectively. The standard case of transient dynamic programming, where utility is additive or multiplicative and there is a single absorbing state (which absorbs “quickly enough”) is shown to be both upper and lower transient. This is generalized to problems with nonseparable utility and to problems where there are absorbing classes of states. Examples of optimal stopping problems (with and without separable utility) which have upper transient strategies are also given. In § 4, the regularity conditions upon which the analysis is based are discussed.

In this paper, no mention is made of stationarity and no results are given concerning the optimality or “almost optimality” of stationary strategies. Results of this sort for upper and lower transient problems are given in [10]. To avoid measure theoretic difficulties, only problems with countable state and history spaces are considered. (Kertz and Nachman [7], [8] analyze problems with uncountable state spaces.)

* Received by the editors March 30, 1977, and in final revised form September 21, 1977.

† Graduate School of Business, Stanford University, Stanford California 94305. This work was supported in part by a grant from the Atlantic Richfield Foundation to the Stanford Graduate School of Business.

Throughout, expected utility criteria are used. Analyses of sequential decision problems with other types of ordinal and cardinal criteria are given in [11], [12], [14].

1. Sequential decision problems with expected utility criteria.

DEFINITION. A (countable state) *sequential decision problem with expected utility criterion* is a collection (X, H_0, A, p, U) of

- (a) countable *state spaces* X_t , with generic element x_t , for $t = 1, 2, \dots$,
- (b) a countable *initial history space* H_0 , with generic h_0 ,
- (c) *partial history spaces* H_t , for $t = 1, 2, \dots$, recursively constructed from H_0 and the X_t by $H_t := H_{t-1} \times X_t$, with generic elements of H_t denoted $h_t = (h_0, x_1, \dots, x_t)$,
- (d) sets of *feasible actions* $A_t(h_t)$, with generic a_t , for each $t = 0, 1, \dots$ and $h_t \in H_t$,
- (e) *transition probabilities* $p_t(h_t, a_t, x_{t+1})$ for each t, h_t, a_t and x_{t+1} , such that $p_t(h_t, a_t, x_{t+1}) \geq 0$ and $\sum_{x_{t+1}} p_t(h_t, a_t, x_{t+1}) = 1$, and
- (f) an extended real valued *utility function* U defined on the space of *complete histories* $H := H_0 \times X_1 \times X_2 \times \dots$ (with generic $h = (h_0, x_1, x_2, \dots)$).

All random variables will be over H , and the notation for the sample point h will usually be omitted. Projections from H onto H_t and X_t and from H_t to X_t will be denoted $h_t(h)$, $x_t(h)$ and $x_t(h_t)$ respectively. Functions on H_t will be written as functions on H as well, where projection onto H_t is understood. For example, for a function f_t on H_t , $f_t(h)$ will be written for $f_t(h_t(h))$.

For the above problem, construct the *strategy space* $\Pi := \prod_{t=0}^{\infty} \prod_{h_t} A_t(h_t)$, with generic π . For each π and h_t , *conditional probabilities* $P^\pi[\cdot | h_t]$ and *expectations* $E^\pi[\cdot | h_t]$ using π (at time t and thereafter) given h_t are constructed on H from the transition probabilities in the usual fashion.

Throughout, one of the following two regularity conditions is assumed:

$$(1) \quad \lim_{M \rightarrow \infty} \sup_{\pi} E^\pi[U \cdot I(\{U > M\}) | h_t] = 0 \quad \text{for all } t \text{ and } h_t,$$

or

$$(2) \quad \lim_{M \rightarrow -\infty} \inf_{\pi} E^\pi[U \cdot I(\{U < M\}) | h_t] = 0 \quad \text{for all } t \text{ and } h_t.$$

(Throughout, $I(G)$ for G a subset of H will denote the indicator function of G .) If (1) or (2) hold, we can define for each π and h_t

$$(3) \quad v_t(\pi, h_t) := E^\pi[U | h_t],$$

the *expected utility using π from time t onward given h_t* . Note that if (1) holds, $v_t(\pi, h_t) = -\infty$ is possible, and if (2) holds, $v_t(\pi, h_t) = \infty$ is not ruled out. But (1) or (2) ensure that U is at least quasi-integrable; thus the integral in (3) is well defined. Also define

$$f_t(h_t) := \sup_{\pi} v_t(\pi, h_t) \quad \text{and} \quad e_t(h_t) := \inf_{\pi} v_t(\pi, h_t);$$

$f_t(h_t)$ is the *optimal expected utility given h_t* and $e_t(h_t)$ is, for lack of a better term, the *worst expected utility given h_t* . Strategy π is optimal if $v_t(\pi, h_t) = f_t(h_t)$ for all t and h_t .

The fundamental functional equations of dynamic programming are

$$(4) \quad v_t(\pi, h_t) = E^\pi[v_{t+1}(\pi, h_{t+1}) | h_t]$$

for all π , t and h_t , and

$$(5) \quad f_t(h_t) = \sup_{\pi} E^{\pi}[f_{t+1}(h_{t+1})|h_t],$$

for all t and h_t . Equations (5), the *optimality equations*, must be proved: See the proof in [9, proposition 1] for the general method. Note that the *sup* here replaces the *max* in [9], as $A_t(h_t)$ is not necessarily finite. Also, the monotone convergence argument needed in cases where (2) is assumed is valid because of (2)—cf. the lemma following.

Strategy π is *conserving* if $f_t(h_t) = E^{\pi}[f_{t+1}(h_{t+1})|h_t]$ for all t and h_t . Strategy π is *unimprovable* (in a single step) if $v_t(\pi, h_t) = \sup_{\pi' \in \Pi} E^{\pi'}[v_{t+1}(\pi, h_{t+1})|h_t]$ for all t and h_t . Strategy π is an (one-step) *improvement* on π' if $E^{\pi}[v_{t+1}(\pi', h_{t+1})|h_t] \geq v_t(\pi', h_t)$ for all t and h_t with strict inequality for some t and h_t . Strategy π is *as good as* π' if $v_t(\pi, h_t) \geq v_t(\pi', h_t)$ for all t and h_t , and π is *better than* π' if, in addition, $v_t(\pi, h_t) > v_t(\pi', h_t)$ for some h_t . By virtue of (4), if π is as good as π' and π is an improvement on π' , then π is better than π' .

From (4) and (5), optimal strategies are both conserving and unimprovable. Well known examples (see e.g., [13]) show that the converse fails in general. So we seek conditions under which conserving and/or unimprovable strategies are optimal. Also, for purposes of *strategy iteration*, we seek conditions under which: π improves on π' implies that π is as good as (hence, better than) π' .

Notational conventions which will be employed include: U^+ , f_t^+ , etc., will denote the positive part of U , f_t , etc., and U^- , f_t^- , etc., will denote the negative part. (That is, $U^-(h) = |U(h) \wedge 0|$.) Throughout, expressions such as $\lim_{T \rightarrow \infty}$ will be abbreviated \lim_T .

The results that will be given require a technical lemma concerning the “uniform integrability” conditions (1) and (2) (in quotes because more than one probability measure is involved).

LEMMA. If (1) holds, then for every t , h_t and $\varepsilon > 0$, there exists $\delta > 0$ such that for each π and $G \subseteq H$, $P^{\pi}[G|h_t] < \delta$ implies that $\sup_{T \geq t} E^{\pi}[f_T^+ \cdot I(G)|h_t] \leq \varepsilon$ and $E^{\pi}[U^+ \cdot I(G)|h_t] \leq \varepsilon$. If (2) holds, then for every t , h_t and $\varepsilon > 0$ there exists $\delta > 0$ such that for each π and $G \subseteq H$, $P^{\pi}[G|h_t] < \delta$ implies $\sup_{T \geq t} E^{\pi}[e_T^- \cdot I(G)|h_t] \leq \varepsilon$ and $E^{\pi}[U^- \cdot I(G)|h_t] \leq \varepsilon$.

Proof. Suppose (1) holds. Fix t , h_t and ε . Then $\sup_{\pi} E^{\pi}[U^+|h_t] = B$ for some $B < \infty$. So by Chebyshev's inequality, for every π , N and $T \geq t$

$$P^{\pi}[f_T \geq N|h_t] \leq E^{\pi}[f_T^+|h_t]/N \leq \sup_{\pi} E^{\pi}[U^+|h_t]/N = B/N.$$

Choose M sufficiently large so that $\sup_{\pi} E^{\pi}[U \cdot I(\{U > M\})|h_t] < \varepsilon/6$. For any π , if $P^{\pi}[G|h_t] < \varepsilon/(6M)$, then

$$\begin{aligned} E^{\pi}[U^+ \cdot I(G)|h_t] &\leq E^{\pi}[U^+ \cdot I(G \cap \{U \leq M\})|h_t] + E^{\pi}[U^+ \cdot I(\{U > M\})|h_t] \\ &\leq M\varepsilon/(6M) + \varepsilon/6 = \varepsilon/3. \end{aligned}$$

Thus, if $N > 6BM/\varepsilon$, then for all π and T

$$\begin{aligned} E^{\pi}[f_T \cdot I(\{f_T > N\})|h_t] &\leq E^{\pi}[\sup_{\pi'} E^{\pi'}[U^+|h_T] \cdot I(\{f_T > N\})|h_t] \\ &\leq E^{\pi''}[E^{\pi''}[U^+ \cdot I(\{f_T > N\})|h_T]|h_t] + \varepsilon/6 \end{aligned}$$

(where π'' is π up to T and any $\varepsilon/6$ -optimal strategy thereafter)

$$= E^{\pi''}[U^+ \cdot I(\{f_T > N\})|h_t] + \varepsilon/6 \leq \varepsilon/3 + \varepsilon/6 = \varepsilon/2.$$

So if $\delta = \min \{\varepsilon/(6M), \varepsilon/(2N)\}$ and $P^\pi[G|h_t] < \delta$, then $E^\pi[U^+ \cdot I(G)|h_t] \leq \varepsilon$ (by previous inequalities) and, for all $T \geq t$

$$\begin{aligned} E^\pi[f_T^+ \cdot I(G)|h_t] &\leq E^\pi[f_T^+ \cdot I(G \cap \{f_T \leq N\})|h_t] + E^\pi[f_T \cdot I(\{f_T > N\})|h_t] \\ &\leq N\varepsilon/(2N) + \varepsilon/2 = \varepsilon. \end{aligned}$$

The proof for (2) follows similarly. \square

2. Upper and lower transience.

DEFINITIONS. For a given problem, define

$$\bar{H} := \{h \in H : \limsup_t f_t(h) > U(h)\}$$

and

$$\underline{H} := \{h \in H : \liminf_t e_t(h) < U(h)\}.$$

For problems for which (1) holds, a strategy π is *upper transient* if $P^\pi[\bar{H}|h_t] = 0$ for all h_t . The problem is *upper transient* if all strategies are upper transient. For problems for which (2) holds, a strategy π is *lower transient* if $P^\pi[\underline{H}|h_t] = 0$ for all h_t . The problem is *lower transient* if all strategies are lower transient. A problem is *transient* if it is both upper and lower transient.

Interpretation will follow the results. Note that the term upper transient is used only when (1) holds, and lower transient is used only when (2) holds.

THEOREM. *If π is an upper transient strategy, then for all h_t ,*

$$(6) \quad v_t(\pi, h_t) = \lim_T E^\pi[f_T|h_t].$$

If π is lower transient, then for all h_t ,

$$(7) \quad v_t(\pi, h_t) = \lim_T E^\pi[e_T|h_t].$$

Proof. Suppose π is upper transient. Let $H^T = \{h : \sup_{T \geq t} f_T(h) > (U(h) + 1)^+\}$ and let G^T be the complement of H^T . Then

$$\limsup_T E^\pi[f_T|h_t] \leq \limsup_T E^\pi[f_T \cdot I(H^T)|h_t] + \limsup_T E^\pi[f_T \cdot I(G^T)|h_t].$$

Since π is upper transient, $\lim_T P^\pi[H^T|h_t] = 0$; thus by the lemma, $\limsup_T E^\pi[f_T \cdot I(H^T)|h_t] \leq 0$. Since $f_T \cdot I(G^T) \leq (U + 1)^+$ and $E^\pi[(U + 1)^+|h_t] < \infty$, Fatou's lemma yields

$$\limsup_T E^\pi[f_T \cdot I(G^T)|h_t] \leq E^\pi[\limsup_T f_T \cdot I(G^T)|h_t] \leq E^\pi[U|h_t] = v_t(\pi, h_t),$$

where the second inequality is due to π being upper transient, so that $P^\pi[\limsup_T f_T \cdot I(G^T) \leq U|h_t] = 1$. Of course, for all T , $E^\pi[f_T|h_t] \geq E^\pi[v_T(\pi, \cdot)|h_t] = v_t(\pi, h_t)$; thus $\lim_T E^\pi[f_T|h_t] = v_t(\pi, h_t)$. The proof of (7) is similar and left to the reader. \square

COROLLARY. (a) *If π is upper transient and conserving, then π is optimal.*

(b) *If π is upper transient and π improves π' , then π is as good as π' .*

(c) *If π is unimprovable, then π is as good as any lower transient strategy.*

(d) *If either the problem is lower transient, or if (2) holds and $f_t(h_t) = \sup v_t(\pi, h_t)$ for every h_t , where the supremum is over lower transient strategies, then any unimprovable strategy is optimal.*

Proof. If π is conserving, then for $T \geq t$, $E^\pi[f_T|h_t] = f_t(h_t)$. So if π is also upper transient, then $v_t(\pi, h_t) = \lim_T E^\pi[f_T|h_t] = f_t(h_t)$. If π improves on π' , then for $T \geq t$, $E^\pi[v_T(\pi', h_T)|h_t] \geq v_T(\pi', h_t)$. So if π is also upper transient, then $v_t(\pi, h_t) = \lim_T E^\pi[f_T|h_t] \geq \liminf_T E^\pi[v_T(\pi', h_T)|h_t] \geq v_t(\pi', h_t)$. If π is unimprovable, then for every $T \geq t$ and π' , $v_t(\pi, h_t) \geq E^\pi[v_T(\pi, h_T)|h_t]$. So if π' is lower transient, then $v_t(\pi', h_t) = \lim_T E^{\pi'}[e_T(h_T)|h_t] \leq \limsup_T E^{\pi'}[v_T(\pi, h_T)|h_t] \leq v_t(\pi, h_t)$. Part (d) is a trivial consequence of (c). \square

Some paraphrase of the mathematics may be enlightening. The condition for a strategy to be upper transient, $P^\pi[\bar{H}|h_t] = 0$, roughly means that π can safely be used by an “optimist”. An optimist would get into trouble if he didn’t avoid \bar{H} , that being the set of h where the optimistic prognostication of what may happen is better than what eventually does happen (that is, $\limsup_t f_t(h)$ exceeds $U(h)$). Similarly, $P^\pi[H|h_t] = 0$ means that π can be used safely by a “pessimist”. Using a conserving strategy is an exercise in optimism, as it conserves the optimal value, “hoping” for an optimal strategy to eventually take over. If a conserving strategy is also upper transient, then this optimism is warranted. Similar stories can be told about parts (b) and (c) of the corollary.

In the corollary (which is the major result), only (6) and (7) are used. One might therefore wish to redefine an upper transient strategy as one which satisfies (6) and a lower transient strategy as one which satisfies (7). Then (1) and $P^\pi[\bar{H}|h_t] = 0$ could be given as a sufficient condition for upper transience, and (2) and $P^\pi[H|h_t] = 0$ could be given as a sufficient condition for lower transience. Of course, if neither (1) nor (2) is assumed at the outset, then some other regularity condition will be needed to ensure that all integrals are well defined and that (4) and (5) hold.

With regard to this possible redefinition, note the following. The regularity conditions (1) and (2) are not advanced as being necessary for the results; see the discussion in § 4. But under any regularity conditions where the integrals are well defined and (4) and (5) hold, $P^\pi[\bar{H}|h_t] = 0$ is *necessary* for (6) and $P^\pi[H|h_t] = 0$ is *necessary* for (7). This can be shown as follows. To rule out trivial pathologies, assume that $v_t(\pi, h_t)$ and $f_t(h_t)$ are finite. Then $\{v_{t+k}(\pi, \cdot); U\}$ is a closed martingale and $\{f_{t+k}; U\}$ is a closed supermartingale with respect to $P^\pi[\cdot|h_t]$. By the convergence theorem $P^\pi[\lim_T v_T(\pi, \cdot) = U|h_t] = 1$ and $P^\pi[\lim_T f_T \text{ exists and is } \geq U|h_t] = 1$ (see [1, Theorem 9.4.6 and the corollary following]). Thus if $P^\pi[\bar{H}|h_t] > 0$, then there exists $\varepsilon > 0$ such that $\lim_T P^\pi[f_T - v_T(\pi, \cdot) > \varepsilon|h_t] > \varepsilon$, which would contradict (6). A similar argument works for (7) and $P^\pi[H|h_t] = 0$. Since $P^\pi[\bar{H}|h_t] = 0$ is necessary for (6) and $P^\pi[H|h_t]$ is necessary for (7), and since these conditions have more intuitive appeal than (6) and (7), they are used to define upper and lower transience.

Condition (6) is clearly analogous to the well known “equalizing” property for problems with additive utility (see, for example, [5, Definition 4.3]). Since part (a) of the corollary relies only on (6) and since (6) is clearly necessary for an optimal strategy, the result can be paraphrased: A strategy is optimal if and only if it is conserving and equalizing.

3. Examples.

A. Problems with upper and lower convergent utility. For every h , let $\bar{U}_T(h) = \sup \{U(h') : h_T(h') = h_T(h)\}$ and let $\underline{U}_T(h) = \inf \{U(h') : h_T(h') = h_T(h)\}$. In [9], U was said to be an *upper convergent* utility function if $\bar{U}_0(h) < \infty$ and $\lim_T \bar{U}_T(h) = U(h)$ for

all h , and U was said to be *lower convergent* if $\bar{U}_0(h) > -\infty$ and $\lim_T \bar{U}_T(h) = U(h)$ for all h . Since $\bar{U}_T(h) \leq e_T(h)$ and $\bar{U}_T(h) \geq f_T(h)$, it is apparent that a problem with upper convergent utility is upper transient and one with lower convergent utility is lower transient. (For (1), note that $E^\pi[U \cdot I(\{U > \bar{U}_0(h)\})|h_t] = 0$ for all π .) Very special cases are $U(h) = \sum_t r_t(x_t)$: If $r_t \leq 0$ (the utility function of negative dynamic programming), then U is upper convergent, while if $r_t \geq 0$ (the utility function of positive dynamic programming), then U is lower convergent. Hinderer's [4] definitions of essentially negative and essentially positive decision problems are likewise subsumed by these categories. See [9] for further details and for several other special cases, with emphasis on generalizations to nonseparable utility.

B. Problems with absorbing states. The standard problem in transient dynamic programming concerns a stationary Markov decision chain with an absorbing state and undiscounted additive utility. That is $H_0 = X_t \equiv X$ for some set X , and $A_t(h_t) = A(x_t(h_t))$ and $p_t(h_t, \cdot, \cdot) = p(x_t(h_t), \cdot, \cdot)$ for functions A and p . There exists $x^* \in X$ such that $p(x^*, \cdot, x^*) \equiv 1$; and $U(h_0, x_1, \dots) = \sum_{t=1}^\infty r(x_t)$ for some function r with $r(x^*) = 0$. If r is bounded and for each h_0 there exists $\alpha < 1$ and K such that $P^\pi[x_t \neq x^* | h_0] < K \cdot \alpha^t$, then the problem is "transient". To get the bound on $P^\pi[x_t \neq x^* | h_0]$, it is helpful to specialize to the case where X and each $A(x)$ are finite. Then K and α as above exist if and only if $P^\pi[X_n \neq x^* | h_0] < 1$ for every h_0 and stationary π , where n is the cardinality of X . (See [12]. Note that in much of the literature, the absorbing state x^* is not explicitly modeled. Instead, the transition probabilities are substochastic. It is necessary here to have transition probabilities which sum to 1, however. This is because with nonseparable utility, the value functions must be the expected utility of the entire history—it is meaningless to speak of the (incremental) value of continuing. Thus, the absorbing state is explicitly mentioned.)

It is straightforward to show that this problem is transient in the sense of § 2. Moreover, generalizations are easily obtained. Suppose that instead of a single state x^* there is an "absorbing set of states" $X^* \subseteq X$ such that $P^\pi[x_{t+1} \in X^* | x_t \in X^*, h_0] = 1$ for all π . If (i) r is bounded above on X , (ii) $r \leq 0$ on X^* and (iii) for every h_0 there exists K and $\alpha < 1$ such that $\sup_\pi P^\pi[x_t \notin X^* | h_0] < K \cdot \alpha^t$, then the problem is upper transient. If r is bounded below on X , $r \geq 0$ on X^* and (iii) holds, then the problem is lower transient.

It is more interesting to generalize to nonseparable utility functions. For example, suppose that the chain is as in the original example and that $U(h) = V(h_0 + \sum_{t=1}^\infty r(x_t))$ for functions V and r with $r(x^*) = 0$. If r is bounded and V is of less than exponential order (that is, for all $\gamma > 1$, $|V(y)| = o(\gamma^{|y|})$), then the problem is transient. To accommodate multiplicative utility, V can be of exponential order but must "grow" more slowly than the chain "dies". If V is nondecreasing and concave (which is natural in economic applications), then the problem is upper transient, although lower transience may fail.

Problems as above but with unbounded r can also be accommodated. For example, suppose that in the immediately preceding example, r is unbounded but satisfies $\lim_M \sup_{h_0 \neq x^*} \sup_\pi E^\pi[(r(x_1) - r(h_0)) \cdot I(\{r(x_1) - r(h_0) > M\}) | h_0, x_1 \neq x^*] = 0$. Then if either V is of less than exponential order or if V is concave and nondecreasing, the problem is upper transient. (The messy details are left to the reader.) Analogous results for lower transience follow from similar conditions. Compare with [3], where the analogous condition is that $\sup_{h_0 \neq x^*} \sup_\pi E^\pi[|r(x_1) - r(h_0)| | h_0, x_1 \neq x^*] < \infty$. (To make the comparison, several superficial differences in setup must be disregarded.)

C. Optimal stopping problems. Consider the standard "house hunting" problem.

The state spaces are (say) $H_0 = X_t \equiv \{0, 1, \dots, \Delta\}$. From each $x_t \in \{0, 1, \dots\}$, there are two possible actions—"sample again", in which case x_{t+1} has some distribution on $\{0, 1, \dots\}$ independent of h_t , and "stop", in which case $x_{t+1} = \Delta$ with certainty. State Δ is absorbing: $p_t(h_t, \cdot, \Delta) \equiv 1$ if $x_t(h_t) = \Delta$. The utility function has the form $U(h) = \sum_{t=1}^{\infty} r(x_{t-1}, x_t)$ where if $x_t \in \{0, 1, \dots\}$, then $r(x_{t-1}, x_t) = c$ for $c < 0$ (there is a fixed cost of sampling), if $x_{t-1} \in \{0, 1, \dots\}$, then $r(x_{t-1}, \Delta) = s(x_{t-1})$ for some bounded nonnegative function s (if you stop, you get some positive reward based on the last observed state), and $r(\Delta, \Delta) = 0$. It is trivial to show that (1) holds (as s is bounded) and that if π stops with probability 1, then π is upper transient. Thus, any conserving strategy which stops with probability 1 is optimal.

Numerous generalizations are possible. Among them are:

(i) Let $U(h) = V(\sum_{t=1}^{\infty} r(x_{t-1}, x_t))$ for r as before and V any nondecreasing function.

(ii) Let U be as in (i), but with s unbounded above and with state dependent sampling costs. It will, of course, be necessary to make some assumptions about the function s so that (1) will hold.

(iii) Transition probabilities may be state or history dependent, and more than one "sample again" action may be available. The former is natural if the "distribution of house prices" is unknown and Bayesian inference is employed. For the latter, compare with the general optimal control problems with additive utility which are analyzed by Hordijk in [5, Chaps. 3 and 4].

4. Regularity conditions. Note that in the optimal stopping problems just discussed, condition (2) does not hold. Thus, we cannot conclude (on the basis of the results in § 2) that an unimprovable strategy is as good as any strategy which stops with probability 1. Indeed, if the transition probabilities for the "stop" action are modified so that, say, $p(h, \text{"stop"}, \Delta) = .99$, then the strategy which always "samples again" is unimprovable. This strategy is certainly worse than the strategy which always "stops" (and which will stop with probability 1).

The roles played by (1) and (2) are technical in nature. Condition (1), for example, allows the "pointwise optimism is safe" condition, $P^\pi[\bar{H}|h_t] = 0$, to be converted into a statement that "optimism is safe in the average", $\lim_T E^\pi[f_T|h_t] = v_t(\pi, h_t)$. Certainly, *some* regularity conditions like (1) or (2) are needed, as the following simple example indicates. Consider a three state Markov decision problem: $H_0 = X_t \equiv \{0, 1, 2\}$ for all t . States 0 and 1 are absorbing—from each there is only one feasible action, and the next state is 0 or 1, respectively, with certainty. From state 2 there are two feasible actions. The first leads to the next state being 1 with certainty. If the second action is used, the next state is either 2 or 0, each with probability 1/2. If $U(h_0, x_1, \dots) = h_0 \cdot x_1 \cdot x_2 \cdot \dots$, it is easy to see that $f_0(h_0) = h_0$. The strategy which always takes the second action in state 2 is conserving and leads with certainty to state 0, i.e., to $h \notin \bar{H}$. But this strategy is not optimal. (Of course, (1) does not hold.) And if $U(h_0, x_1, \dots) = -h_0 \cdot x_1 \cdot x_2 \cdot \dots$, then $f_0(2) = f_0(0) = 0$ while $f_0(1) = -1$. All strategies lead with certainty to $h \notin \bar{H}$, and the strategy which always takes the first action in state 2 is unimprovable, but this strategy is not optimal. (Of course, (2) does not hold.)

But it is not contended that (1) or (2) are necessary conditions for the results given. For example, in the optimal stopping problem, suppose that π is unimprovable and that $f_t(h_t) - v_t(\pi, h_t)$ is uniformly bounded in t and h_t . Since (1) holds, all integrals are well defined and (6) holds. Also, π is clearly as good as any strategy π' such that $v_t(\pi', h_t) = -\infty$, so to show that π is optimal, we can assume that $v_t(\pi', h_t) > -\infty$, and

therefore that π' leads to stopping with probability 1. Then

$$\begin{aligned}
 v_i(\pi', h_i) &= \lim_T E^{\pi'}[f_T | h_i] \\
 &= \lim_T E^{\pi'}[f_T \cdot I(\{x_T = \Delta\}) | h_i] + \lim_T E^{\pi'}[(f_T - v_T(\pi, \cdot)) \cdot I(\{x_T \neq \Delta\}) | h_i] \\
 &\quad + \lim_T E^{\pi'}[v_T(\pi, \cdot) \cdot I(\{x_T \neq \Delta\}) | h_i] \\
 &= \lim_T E^{\pi'}[v_T(\pi, \cdot) \cdot I(\{x_T = \Delta\}) | h_i] + 0 + \lim_T E^{\pi'}[v_T(\pi, \cdot) \cdot I(\{x_T \neq \Delta\}) | h_i] \\
 &\quad (\text{since (i) if } x_T = \Delta, \text{ then } f_T(h_T) = v_T(\pi, h_T), \text{ (ii) } f_T - v_T(\pi, \cdot) \\
 &\quad \text{is uniformly bounded and (iii) } P^{\pi}[x_T \neq \Delta | h_i] \rightarrow 0) \\
 &= \lim_T E^{\pi'}[v_T(\pi, h_T) | h_i] \leq V_i(\pi, h_i),
 \end{aligned}$$

since π is unimprovable. Thus π is optimal (although (2) does not hold).

The point to be made is simply that at the “heart” of upper and lower transience are $P^{\pi}[\bar{H} | h_i] = 0$ and $P^{\pi}[H | h_i] = 0$, respectively. Regularity conditions such as (1) or (2) are required to get the desired results, and (1) and (2) are formally incorporated into the definitions given here. But weaker regularity conditions can be used in particular contexts.

Acknowledgment. The author gratefully acknowledges the helpful comments of Professor Evan L. Porteus and the referees.

REFERENCES

- [1] KAI LAI CHUNG, *A Course in Probability Theory*, 2nd ed., Academic Press, New York, 1974.
- [2] ERIC V. DENARDO AND URIEL G. ROTHBLUM, *A Markov decision model having multiplicative utility*, draft, 1975.
- [3] J. MICHAEL HARRISON, *Discrete dynamic programming with unbounded rewards*, Ann. Math. Statist., 43 (1972), pp. 636–644.
- [4] K. HINDERER, *Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter*, Springer-Verlag, New York, 1970.
- [5] A. HORDIJK, *Dynamic Programming and Markov Potential Theory*, Mathematisch Centrum, Amsterdam 1974.
- [6] STRATTON C. JAQUETTE, *Utility optimal policies in an undiscounted Markov decision process*, draft 1975.
- [7] ROBERT P. KERTZ AND DAVID C. NACHMAN, *Optimal plans for discrete-time non-stationary dynamic programming with general total reward function I: The topology of weak convergence case*, Working Paper MS-77-1, College of Industrial Management, Georgia Inst. of Tech., Atlanta, 1977.
- [8] ———, *Optimal plans for discrete-time non-stationary dynamic programming with general total reward function II: the WS^{∞} -topology case*, Working Paper MS-77-6, College of Industrial Management, Georgia Inst. of Tech., Atlanta, 1977.
- [9] DAVID M. KREPS, *Decision problem with expected utility criteria. I: Upper and lower convergent utility*, Math. Operations Res., 2 (1977), pp. 45–53.
- [10] ———, *Decision problems with expected utility criteria. II: Stationarity*, Ibid., 2 (1977), pp. 266–274.
- [11] DAVID M. KREPS AND EVAN L. PORTEUS, *On the optimality of structured policies in countable stage decision problems. II: Positive and negative problems*, SIAM J. Appl. Math., 32 (1977), pp. 456–466.

- [12] EVAN L. PORTEUS, *On the optimality of structured policies in countable stage decision problems*, Management Sci., 22 (1975), pp. 148–157.
- [13] SHELDON M. ROSS, *Dynamic programming and gambling models*, Advances in Appl. Probability, 6 (1974), pp. 593–606.
- [14] MATTHEW J. SOBEL, *Ordinal dynamic programming*, Management Sci., 21 (1975), pp. 967–975.
- [15] ARTHUR F. VEINOTT, JR., *Markov decision chains*, Studies in Optimization, G. B. Dantzig and C. Eaves, eds., Mathematical Association of America, Washington, D.C., 1975.

A SUFFICIENT CONDITION FOR FUNCTION SPACE CONTROLLABILITY OF A LINEAR NEUTRAL SYSTEM*

HERNAN RIVERA RODAS† AND C. E. LANGENHOP†

Abstract. A proof is given for the following conjecture. When $\text{rank}[b, A_{-1}b, \dots, A_{-1}^{n-1}b] = n$, a sufficient condition for function space controllability of $\dot{x}(t) = A_{-1}\dot{x}(t-h) + A_0x(t) + A_1x(t-h) + bu(t)$ is that $K(\lambda)\zeta(e^{-\lambda h}) \neq 0$ for all complex λ , where $K(\lambda)$ is a $n \times n$ polynomial matrix in λ constructed from A_{-1} , A_0 , A_1 , b and $\zeta(S)$ is the transpose of $[1, S, \dots, S^{n-1}]$.

1. Introduction. Let $h > 0$, let A_{-1} , A_0 , A_1 be $n \times n$ constant real matrices and let b be an $n \times 1$ constant real matrix. Function space controllability of

$$(1.1) \quad \dot{x}(t) = A_{-1}\dot{x}(t-h) + A_0x(t) + A_1x(t-h) + bu(t)$$

was studied in [1] and [3]. Here x is real $n \times 1$ and the control function is real. When $\tau > nh$ the question of controllability of (1.1) on $[0, \tau]$ was reduced to the question of whether or not there is a nontrivial solution ω to a related boundary value problem

$$(1.2) \quad K(D)\omega(t) = 0, \quad t \in [\tau-h, \tau],$$

$$(1.3) \quad D^i\omega_j(\tau-h) = D^i\omega_{j+1}(\tau), \quad i = 0, 1, \dots, n-1 \quad j = 1, \dots, n-1.$$

(See [3, Remark 4.1].) Here $K(D)$ is a certain $n \times n$ polynomial matrix in the derivative operator D constructed from A_{-1} , A_0 , A_1 and b ; ω_j denotes the j th component of the $n \times 1$ function ω . It was also shown in [3] (Corollary 4.6) when the matrix $C_n[A_{-1}, b] = [b, A_{-1}b, \dots, A_{-1}^{n-1}b]$ has rank n that a necessary condition for controllability of (1.1) is that $K(\lambda)\zeta(e^{-\lambda h}) \neq 0$ for all complex numbers λ where $\zeta(e^{-\lambda h})$ is the transpose of $[1, e^{-\lambda h}, \dots, e^{-(n-1)\lambda h}]$. In [2] it was shown that the condition is also sufficient in case $n = 2$ and in [3] it was conjectured that this was true in case $n \geq 3$ (see [3, Remark 4.4]). We give a proof here of this conjecture.

2. Notation and definitions. In most instances we use the notation in [3] but for the convenience of the reader we repeat some of the principal concepts here. If ν is a positive integer, $W_2^{(\nu)}([\alpha, \beta], R^n)$ denotes the space of functions $X: [\alpha, \beta] \rightarrow R^n$ such that $D^{\nu-1}x$ is absolutely continuous on $[\alpha, \beta]$ and $D^\nu x$ is square integrable, i.e., $D^\nu x \in L_2([\alpha, \beta], R^n)$. We use the convention $W_2^{(0)}([\alpha, \beta], R^n) = L_2([\alpha, \beta], R^n)$. For any function whose domain of definition contains the interval $[t-h, t]$ we use x_t to denote the function $\theta \mapsto x_t(\theta)$ defined on $[-h, 0]$ by $x_t(\theta) = x(t+\theta)$.

Given $u \in W_2^{(0)}([0, \tau], R^n)$ where $\tau > 0$, the value at t of the function $x \in W_2^{(1)}([-h, \tau], R^n)$ satisfying (1.1) a.e. on $[0, \tau]$ and the initial condition $x_0 = \phi \in W_2^{(1)}([-h, 0], R^n)$ is denoted by $x(t; \phi, u)$. Controllability of (1.1) on $[0, \tau]$ as defined in [1] was shown to reduce to the condition that for every $\psi \in W_2^{(1)}([-h, 0], R^n)$ there is a $u \in W_2^{(0)}([0, \tau], R)$ such that $x_\tau(\cdot; 0, u) = \psi$ [3, p. 102]. Thus for studying controllability of (1.1), one may consider x and u to be zero on $(-\infty, 0]$ and replace (1.1) by

$$(2.1) \quad Q(D, S)x(t) = bu(t), \quad t \leq \tau,$$

where

$$(2.2) \quad Q(D, S) = ID - A_{-1}DS - A_0 - A_1S,$$

* Received by the editors May 13, 1977, and in revised form September 7, 1977.

† Department of Mathematics, Southern Illinois University, Carbondale, Illinois 62901.

in which the shift operator S satisfies $(Sx)(t) = x(t-h)$. That is, let $W_{2,0}^{(\nu)}(\tau, R^p)$ denote the subspace of functions in $W_2^{(\nu)}((-\infty, \tau], R^p)$ which are zero on $(-\infty, 0]$ and let $t \mapsto x(t, u)$ denote the solution (unique) in $W_{2,0}^{(1)}(\tau, R^n)$ of (2.1) with $u \in W_{2,0}^{(0)}(\tau, R)$. Then controllability of (1.1) on $[0, \tau]$ is equivalent to the condition that for every $\psi \in W_2^{(1)}([-h, 0], R^n)$ there is such a u so that $x_\tau(\cdot, u) = \psi$.

Interpreting D and S as scalar indeterminates and $Q(D, S)$ as a matrix of polynomials in D, S , one introduces the $n \times n$ polynomial matrix $P(D, S) = \text{adj } Q(D, S)$, the transposed matrix of cofactors of $Q(D, S)$. With

$$(2.3) \quad \zeta(S) = \begin{bmatrix} 1 \\ S \\ \cdot \\ \cdot \\ S^{n-1} \end{bmatrix}$$

there is then a unique $n \times n$ polynomial matrix $K(D)$ defined by the relation

$$(2.4) \quad K(D)\zeta(S) = P(D, S)b.$$

Since $K(D)$ is at most of degree $n-1$, one may write

$$(2.5) \quad K(D) = \sum_{i=0}^{n-1} K_i D^{n-1-i}$$

for some $n \times n$ constant matrices K_i , $i = 0, 1, \dots, n-1$.

The following lemma is of fundamental importance in our consideration of controllability conditions for (1.1).

LEMMA 2.1. Let $\omega \in W_2^{(n)}([\alpha, \beta], R^n)$ satisfy

$$(2.6) \quad K(D)\omega(t) = 0, \quad t \in [\alpha, \beta].$$

If $\text{rank } C_n[A_{-1}, b] = n$, then

$$(2.7) \quad \omega(t) = \sum_{i=1}^p \sum_{k=0}^{m_i} c_{ik} t^k e^{\lambda_i t}, \quad t \in [\alpha, \beta]$$

for some vectors $c_{ik} \in \mathbb{C}^n$ where $\lambda_1, \dots, \lambda_p$ are the distinct zeros of $\det K(\lambda) = 0$ and $0 \leq m_i \leq n_i - 1$ with n_i the multiplicity of λ_i , $i = 1, 2, \dots, p$.

Proof. It follows from Lemma 4.1 in [3] that K_0 is nonsingular since $\text{rank } C_n[A_{-1}, b] = n$. That ω has the form (2.7) is then implicit in the proof of Chrystal's theorem given in [4 pp. 326–327]. To give a more complete derivation we note that (2.6) may be written in the form

$$D^{n-1}\omega = \sum_{i=1}^{n-1} \hat{K}_i D^{n-1-i}\omega$$

where $\hat{K}_i = -K_0^{-1}K_i$, $i = 1, \dots, n-1$. This in turn is equivalent to

$$(2.8) \quad D\Omega = G\Omega$$

where Ω is $\delta \times 1$ and G is $\delta \times \delta$ with $\delta = n(n-1)$, and these are given explicitly by

$$\Omega = \begin{bmatrix} \omega \\ D\omega \\ \cdot \\ \cdot \\ D^{n-2}\omega \end{bmatrix}, \quad G = \begin{bmatrix} 0 & I & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & I \\ \hat{K}_{n-1} & \hat{K}_{n-2} & \cdots & \hat{K}_1 \end{bmatrix}$$

in which I is the $n \times n$ identity. If we introduce the $n \times \delta$ matrix $E = [I, 0, \cdots, 0]$, then by virtue of (2.8) we may write

$$\omega(t) = Ee^{Gt}\Omega_0$$

for some $\Omega_0 \in \mathbb{C}^\delta$. Now let $\lambda_1, \cdots, \lambda_p$ be the distinct eigenvalues of G . It was shown in [5, p. 270] that $\det P(\lambda) = \det(\lambda I - G)$ where

$$P(\lambda) = \lambda^{n-1}I_n - \sum_{i=1}^{n-1} \hat{K}_i \lambda^{n-1-i} = K_0^{-1}K(\lambda).$$

Hence $\lambda_1, \cdots, \lambda_p$ are also the distinct zeros of $\det K(\lambda)$ with the same multiplicities as in the characteristic polynomial of G . The expression (2.7) now clearly follows from the known form of the matrix e^{Gt} . \square

3. Controllability conditions. If $\tilde{\omega} \in W_{2,0}^{(n)}(\tau, R)$, then

$$(3.1) \quad x = K(D)\zeta(S)\tilde{\omega}$$

is in $W_{2,0}^{(1)}(\tau, R)$. It follows from (2.4) that $Q(D, S)x = Q(D, S)P(D, S)b\tilde{\omega} = b \det Q(D, S)\tilde{\omega}$ so x defined by (3.1) satisfies (2.1) with

$$(3.2) \quad u = \det Q(D, S)\tilde{\omega}.$$

The resulting u is in $W_{2,0}^{(0)}(\tau, R)$. These relations were exploited in [1, Thm. 5.1] to establish the following.

THEOREM 3.1. *Let $\tau > nh$. If $\psi \in W_2^{(1)}([-h, 0], R^n)$, there is a $u \in W_{2,0}^{(0)}(\tau, R)$ such that the corresponding solution $x \in W_{2,0}^{(1)}(\tau, R^n)$ of (2.1) satisfies $x_\tau = \psi$ if and only if there is $\tilde{\omega} \in W_2^{(n)}([\tau - nh, \tau], R)$ such that*

$$(3.3) \quad K(D)\zeta(S)\tilde{\omega}(t) = \psi(t - \tau), \quad t \in [\tau - h, \tau].$$

Remark 3.1. It should be noted that the condition here can be replaced by the condition that there is $\omega \in W_2^{(n)}([\tau - h, \tau], R^n)$ such that $K(D)\omega(t) = \psi(t - \tau)$ for $t \in [\tau - h, \tau]$ and that the boundary conditions (1.3) are satisfied. One then defines $\tilde{\omega}$ by $\tilde{\omega}(t - (j-1)h) = \omega_j(t)$, $t \in [\tau - h, \tau]$, $j = 1, 2, \cdots, n$ so that $\zeta(S)\tilde{\omega}(t) = \omega(t)$ for $t \in [\tau - h, \tau]$.

If $\tau > nh$, then a function $\tilde{\omega} \in W_2^{(n)}([\tau - nh, \tau], R)$ satisfying (3.3) can be extended on $(-\infty, \tau - nh)$ so as to produce an $\tilde{\omega} \in W_{2,0}^{(n)}(\tau, R)$. The x and u in Theorem 3.1 are then given by (3.1) and (3.2).

Under the condition that K_0 in (2.5) is nonsingular it was shown in [3, Remark 4.1] that Theorem 3.1 leads to the following result.

THEOREM 3.2. *Let $\tau > nh$. In order that (1.1) be controllable on $[0, \tau]$ it is necessary and sufficient that $\text{rank } C_n[A_{-1}, b] = n$ and that the only solution $\tilde{\omega} \in$*

$W_2^{(n)}([\tau - nh, \tau], R)$ of

$$(3.4) \quad K(D)\zeta(S)\tilde{\omega}(t) = 0, \quad t \in [\tau - h, \tau],$$

be the identically zero function.

The same condition is valid if instead we allow $\tilde{\omega} \in W_2^{(n)}([\tau - nh, \tau], \mathbb{C})$ for the matrix $K(D)\zeta(S)$ has real coefficients as a polynomial matrix in D and S , so if $\tilde{\omega}$ is a complex valued solution of (3.4), then both $\operatorname{Re} \tilde{\omega}$ and $\operatorname{Im} \tilde{\omega}$ would be real valued solutions. Since $K(D)\zeta(S)e^{\lambda t} = K(\lambda)\zeta(e^{-\lambda h})e^{\lambda t}$ it is clear then that under the hypotheses of Theorem 3.2 controllability of (1.1) implies

$$(3.5) \quad K(\lambda)\zeta(e^{-\lambda h}) \neq 0$$

for all complex numbers λ . It was conjectured in [3, Remark 4.4] that the converse is also valid and we prove this conjecture below. To this end we prove the following.

THEOREM 3.3. *Let $\tau > nh$ and $\operatorname{rank} C_n[A_{-1}, b] = n$. If $\tilde{\omega} \in W_2^{(n)}([\tau - nh, \tau], R)$ satisfies (3.4) then*

$$(3.6) \quad \tilde{\omega}(t) = \sum_{i=1}^p \sum_{k=0}^{m_i} b_{ik} t^k e^{\lambda_i t}, \quad t \in [\tau - nh, \tau],$$

for some complex numbers b_{ik} where $\lambda_1, \dots, \lambda_p$ are the distinct zeros of $\det K(\lambda) = 0$ and $0 \leq m_i \leq n_i - 1$, n_i being the multiplicity of λ_i , $i = 1, \dots, p$.

Proof. First extend $\tilde{\omega}$ to $(-\infty, \tau - nh)$ so the resulting $\tilde{\omega} \in W_{2,0}^{(n)}(\tau, R)$ and define $x \in W_{2,0}^{(1)}(\tau, R^n)$ and $u \in W_{2,0}^{(0)}(\tau, R)$ through (3.1) and (3.2), respectively, for this extended $\tilde{\omega}$. Taking $T = \tau + (n-1)h$, we now extend u to $(\tau, T]$ by defining $u(t) = 0$ for $t > \tau$. Let $\hat{\omega}$ be the unique solution in $W_{2,0}^{(n)}(T, R)$ of $\det Q(D, S)\hat{\omega}(t) = u(t)$, $t \leq T$. Clearly $\hat{\omega}(t) = \tilde{\omega}(t)$ for $t \leq \tau$. Then $\hat{x} = K(D)\zeta(S)\hat{\omega}$ defines a function in $W_{2,0}^{(1)}(T, R^n)$ which agrees with x on $(-\infty, \tau]$. Moreover, for $t \in (\tau, T]$ we have $Q(D, S)\hat{x}(t) = bu(t) = 0$. Since $\hat{x}(t) = 0$ for $t \in [\tau - h, \tau]$ by (3.4), we conclude that $\hat{x}(t) = 0$ for $t \in [\tau - h, T]$. Hence $K(D)\zeta(S)\hat{\omega}(t) = 0$ for $t \in [\tau - h, \tau + (n-1)h]$. For $t \in [\tau - h, \tau + (n-1)h]$ we have $S^{n-1}\hat{\omega}(t) = \tilde{\omega}(t - (n-1)h)$ so it follows from Lemma 2.1 that $\tilde{\omega}$ has the form given in (3.6) when $t \in [\tau - nh, \tau]$. \square

Now let a, b and k be nonnegative integers and let D_λ denote differentiation with respect to λ . Then

$$\begin{aligned} D^a S^b (t^k e^{\lambda t}) &= D^a S^b D_\lambda^k e^{\lambda t} = D_\lambda^k (D^a S^b e^{\lambda t}) \\ &= D_\lambda^k [\lambda^a (e^{-\lambda h})^b e^{\lambda t}] \\ &= \sum_{j=0}^k \binom{k}{j} t^j e^{\lambda t} D_\lambda^{k-j} [\lambda^a (e^{-\lambda h})^b] \end{aligned}$$

Since

$$K(D)\zeta(S) = \sum_{a=0}^{n-1} \sum_{b=0}^{n-1} w_{ab} D^a S^b$$

for some vectors $w_{ab} \in R^n$ it follows that if $b_k \in \mathbb{C}$, $k = 0, 1, \dots, m$, then

$$(3.7) \quad K(D)\zeta(S) \sum_{k=0}^m b_k t^k e^{\lambda t} = \sum_{j=0}^m t^j e^{\lambda t} \sum_{k=j}^m b_k \binom{k}{j} D_\lambda^{k-j} [K(\lambda)\zeta(e^{-\lambda h})].$$

LEMMA 3.1. *Let $\tilde{\omega}^i$ be given by*

$$(3.8) \quad \tilde{\omega}^i(t) = \sum_{k=0}^{m_i} b_{ik} t^k e^{\lambda_i t}, \quad t \in [\tau - nh, \tau],$$

for some complex numbers b_{ik} and let $\tilde{\omega}(t) = \sum_{i=1}^p \tilde{\omega}^i(t)$, $t \in [\tau - nh, \tau]$. If $\lambda_1, \dots, \lambda_p$ are distinct and if $\tilde{\omega}$ satisfies (3.4), then

$$(3.9) \quad \sum_{k=j}^{m_i} b_{ik} \binom{k}{j} [D_\lambda^{k-j} K(\lambda) \zeta(e^{-\lambda h})]_{\lambda=\lambda_i} = 0$$

for $i = 1, 2, \dots, p$ and $j = 0, 1, \dots, m_i$.

Proof. From (3.4) and (3.7) it follows that

$$\sum_{i=1}^p \sum_{j=0}^{m_i} t^j e^{\lambda_i t} \sum_{k=j}^{m_i} b_{ik} \binom{k}{j} [D_\lambda^{k-j} K(\lambda) \zeta(e^{-\lambda h})]_{\lambda=\lambda_i} = 0$$

for $t \in [\tau - h, \tau]$. Since the functions $t^j e^{\lambda_i t}$ are linearly independent the conditions (3.9) must hold. \square

We may now establish the following.

THEOREM 3.4. Let $\tau > nh$. Then for (1.1) to be controllable on $[0, \tau]$ it is necessary and sufficient that $\text{rank } C_n[A_{-1}, b] = n$ and $K(\lambda) \zeta(e^{-\lambda h}) \neq 0$ for every $\lambda \in \mathbb{C}$.

Proof. The necessity was proved in [3, Cor. 4.6] and partly sketched above. To establish the sufficiency suppose (1.1) is not controllable on $[0, \tau]$ but $\text{rank } C_n[A_{-1}, b] = n$. By Theorem 3.2 there is then a nontrivial $\tilde{\omega} \in W_2^{(n)}([\tau - nh, \tau], R)$ satisfying (3.4). By Theorem (3.3) $\tilde{\omega}$ must have the form given in (3.6) with some $b_{ik} \neq 0$ since $\tilde{\omega}$ is not identically zero. We may take m_i so that $b_{im_i} \neq 0$ for some i . It follows from Lemma 3.1 that the conditions (3.9) hold. Taking $j = m_i$ in (3.9) for the i such that $b_{im_i} \neq 0$ we conclude $K(\lambda_i) \zeta(e^{-\lambda_i h}) = 0$. \square

4. Example. Let $n = 3$ and

$$[A_{-1}, A_0, A_1] = \begin{bmatrix} 0 & 1 & 0 & 1 & 2 & 0 & 0 & -1 & -2 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Then

$$P(D, S) = \begin{bmatrix} D(D-1) & D^2S + 2D - DS & (D-1)^2S^2 \\ 0 & D(D-1) & (D-1)^2S \\ 0 & 0 & (D-1)^2 \end{bmatrix}.$$

Let e_i be the i th column of the identity matrix and consider the system

$$(4.1) \quad Q(D, S)x = e_3u, \quad u \in W_{2,0}^{(0)}(\tau, R).$$

Note that $\text{rank } C_3[A_{-1}, e_3] = 3$. Denoting by $K^1(D)$ the operator $K(D)$ corresponding to this system, we obtain

$$K^1(D) = \begin{bmatrix} 0 & 0 & (D-1)^2 \\ 0 & (D-1)^2 & 0 \\ (D-1)^2 & 0 & 0 \end{bmatrix}$$

and $\det K^1(\lambda) = -(\lambda - 1)^6$. Since $K^1(1) = 0$ then (4.1) is not controllable. We now consider whether there is a $b_1 \in R^3$ such that for $B = [e_3, b_1]$ the system

$$(4.2) \quad Q(D, S)x = Bu, \quad u \in W_{2,0}^{(0)}(\tau, R^2),$$

is controllable. Let $b_1 = e_2$ so $B = [e_3, e_2]$. Clearly $C_3[A_{-1}, B]$ has full rank. From Theorem 6, p. 86 of [6] we find that there is a $c \in R^2$ such that $b = Bc$ and $\text{rank } C_3[A_{-1}, b] = 3$. If $c = [c_1, c_2]^T$, where T denotes transpose, then this is true if and only

if $c_1 \neq 0$. Moreover,

$$K(D) = c_1 K^1(D) + c_2 K^2(D),$$

where

$$K^2(D) = \begin{bmatrix} 2D & D(D-1) & 0 \\ D(D-1) & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

It follows that $\det K(\lambda) = -c_1^3(\lambda-1)^6$ and

$$K(1)\zeta(e^{-h}) = \begin{bmatrix} 2c_2 \\ 0 \\ 0 \end{bmatrix} \neq 0$$

if and only if $c_2 \neq 0$. We conclude that the system (1.1) with $b = [e_3, e_2]c$ is controllable for any $c = [c_1, c_2]^T$ such that $c_1 \neq 0$ and $c_2 \neq 0$ and only for such c .

Although the analysis of this example can also be effected by the use of Theorem 3.2 with just elementary computations, the required calculations are much more tedious than those used here. If one were to analyze the same questions for the system above with A_{-1} replaced by

$$(4.3) \quad A_{-1} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix},$$

then the use of Theorem 3.2 would be nearly impossible. Indeed, in that case with $b = [0, c_2, c_1]^T$ we find

$$K(D) = \begin{bmatrix} 2c_2D & c_2D(D-1) & c_1(D-1)^2 \\ c_2D(D-1) & c_1(D-1)^2 & 2c_2D \\ c_1(D-1)^2 & 2c_2D & c_2D(D-1) \end{bmatrix}.$$

The circulant form of this matrix enables one to factor $\det K(D)$ and we find

$$\det K(D) = -(\alpha + \beta + \gamma)(\alpha + \omega\beta + \omega^2\gamma)(\alpha + \omega^2\beta + \omega\gamma)$$

where ω is a complex cube root of unity and

$$\alpha = c_1(D-1)^2, \quad \beta = c_2D(D-1), \quad \gamma = 2c_2D.$$

If $c_1 \neq 0$ and $c_2 \neq 0$, then it is clear that the zeros of $\det K(\lambda)$ are generally quite complicated so use of Theorem 3.2 would be extremely cumbersome. On the other hand, Theorem 3.4 could be much more readily employed with the aid of some computer calculations.

For this modified example we might note some special cases. In particular, with A_{-1} as in (4.3) and $b = [0, -c_1, c_1]^T$ we find that $\text{rank } C_3[A_{-1}, b] < 3$ whereas $\text{rank } C_3[A_{-1}, b] = 3$ when $b = [0, 0, c_1]^T$ and $c_1 \neq 0$. However, for this latter b one still finds the system (1.1) is uncontrollable since in this case $K(1) = 0$ so $K(1)\zeta(e^{-h}) = 0$. With $b = [0, c_2, 0]^T$, $c_2 \neq 0$, one finds $\text{rank } C_3[A_{-1}, b] = 3$ but now $K(0) = 0$ so again the system (1.1) is uncontrollable by virtue of Theorem 3.4.

REFERENCES

- [1] H. T. BANKS, M. Q. JACOBS AND C. E. LANGENHOP, *Characterization of the controlled states in $W_2^{(1)}$ of linear hereditary systems*, this Journal, 13 (1975), pp. 611–649.
- [2] M. Q. JACOBS AND C. E. LANGENHOP, *Controllable two dimensional neutral systems*, Mathematical Control Theory (Proceedings of a Conference, Zakopane, 1974), S. Dolecki, C. Olech, J. Zabczyk, eds., Polish Scientific Publishers, Warsaw, 1976, pp. 107–113.
- [3] ———, *Criteria for function space controllability of linear neutral systems*, this Journal, 14 (1976), pp. 1009–1048.
- [4] P. LANCASTER AND H. K. WIMMER, *Zur Theorie der λ -Matrizen*, Math. Nachr., 68 (1975), pp. 325–330.
- [5] C. E. LANGENHOP, *The inverse of a matrix polynomial*, Linear Algebra Appl., 16 (1977), pp. 267–284.
- [6] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

DIFFERENCE APPROXIMATIONS TO CONTROL PROBLEMS WITH FUNCTIONAL ARGUMENTS*

F. H. MATHIS† AND G. W. REDDIEN‡

Abstract. The midpoint difference method is analyzed as a method of discretizing control problems with delays in the state equation. Attention is given to construction of mesh spacings so that $O(h^2)$ convergence can be achieved. Several numerical examples are given and compared to recent numerical results that have appeared for these problems.

1. Introduction. The purpose of this note is to prove the convergence of a special discretization for a family of optimal control problems with functional arguments, e.g. a time lag, and to compare the method numerically with two new methods recently proposed for these problems. The problems studied are nonlinear and in the linear case include those recently studied by Banks and Burns [1, pp. 10–25], [2], Banks, Burns, Cliff and Thrift [3] and also some of the linear problems of Banks and Manitius [4]. The first three of these papers use a projection method with averaging projectors while the fourth uses a projection series technique. We will analyze a finite difference method and thus will generalize and improve some of the results of Budak, Berkovich and Solov'eva [5]. However, the results here are more than a generalization of those of [5] to the retarded case. We will use a different numerical method here, namely the midpoint difference method, on the state equation as opposed to the Euler method studied in [5], and we will use the midpoint integration rule on the cost functional. With this specific method, our goal will be to establish a convergence rate, not just convergence as in [2] or [5]. In order to obtain high order convergence, the mesh points for our scheme will have to be selected carefully. In particular, $O(h^2)$ convergence (where h is the mesh norm) will be shown to be possible with only piecewise smooth solutions. The solutions to control problems with delays typically have jump discontinuities in low order derivatives which makes the development of high order methods difficult. The results given here will apply to problems without functional arguments, and so our method is a general one that can be used to achieve high order convergence for control problems in the presence of jump discontinuities. Rates of convergence for several finite difference approximations to control problems without functional arguments are contained in Hager [8], but the scheme presented here is not treated and the problem of piecewise continuous solutions is not examined. Euler's method as a discretization for problems with simple delays only is considered in Lasiecka [12]. Our more accurate method is analyzed here for general functional arguments. Also, the problem of discontinuities is not discussed in [12].

In § 2, the problems treated and the numerical method are defined, and needed assumptions are made. Section 3 contains the convergence results, and in § 4 several numerical examples are presented and compared with results reported in [4] and also with some of the very extensive numerical results reported in [3].

2. Problem formulation. We shall adopt the following notational conventions throughout. Let r and T be fixed positive numbers with $T > r$. Denote by C the space of R^k -valued continuous functions over $[-r, 0]$ with the usual supremum norm, and denote by $L^\infty[0, T]$ the space of R^l -valued essentially bounded functions. Let $\|\cdot\|$ denote the maximum norm in R^m for any m , let $\|\cdot\|_C$ denote the norm on C , let $\|\cdot\|_2$

* Received by the editors September 2, 1976, and in revised form August 1, 1977.

† Department of Mathematics, Vanderbilt University, Nashville, Tennessee. Now at Department of Mathematics, Auburn University, Auburn, Alabama 36830.

‡ Department of Mathematics, Vanderbilt University, Nashville, Tennessee 37235.

denote the usual norm on $L^2[0, T]$, and let $\|\cdot\|_\infty$ denote the essential supremum norm. If $x: [-r, T] \rightarrow R^k$, denote by x_t the element of C given by $x_t(s) = x(t+s)$, s in $[-r, 0]$.

Let U be a nonempty subset of $L^\infty[0, T]$ of the form $\{u \in L^\infty[0, T]: \|u\|_\infty \leq K\}$. We will consider the problem to minimize the functional

$$(2.1) \quad J(u) = \int_0^T g(x(s), u(s), s) ds + P(x_T)$$

subject to the constraints

$$(2.2) \quad \begin{aligned} \dot{x}(t) &= f(x_t, u(t), t) \quad \text{a.e. for } 0 \leq t \leq T, \\ x_0 &= \phi \quad \text{with } \phi \text{ in } C, \end{aligned}$$

and

$$u \in U.$$

Here $\dot{x}(t)$ denotes the time derivative, $g: R^k \times R^l \times R \rightarrow R$, $f: C \times R^l \times R \rightarrow R^k$, and $P: C \rightarrow R$.

In order to define our finite difference method, we need to introduce a partition of the interval $[-r, T]$. We want to do this to take into account the discontinuities of the solution u^* and x^* of (2.1)–(2.2) so that high order convergence can be obtained. Later, we will assume that x^* is piecewise three times continuously differentiable and that these points of discontinuity can be identified. This identification problem is discussed later in this section. It may be the case that the assumed smoothness is absent or that the points of discontinuity can not be found a priori. In such cases, convergence for our method can be shown using the proofs given here, but without any convergence rates. The case of guaranteed $O(h^2)$ convergence only will be explicitly studied here. We proceed as follows. Let $\{t_i\}$ be the set (assumed finite) of discontinuities of the first three derivatives of x^* and the first two derivatives of u^* on $[0, T]$ plus the points 0 and T . Add the set (also finite) of discontinuities of the first two derivatives of ϕ and denote this basic partition as $\Sigma = \{\hat{t}_i\}$. All other partitions chosen will be refinements of Σ . Other restrictions on the meshes may need to be added. This is discussed in § 3. We will denote these partitions by $\Sigma_n: -r = t_{-N}^n < t_{-N+1}^n < \dots < t_0^n = 0 < t_1^n < \dots < t_n^n = T$ and define $h_i^n = t_i^n - t_{i-1}^n$, $h_n = \max_i h_i^n$ and $\Sigma'_n = \Sigma_n \cap [0, T]$.

Our approximation method will be to minimize the discrete functional

$$(2.3) \quad J_n(u_n) = \sum_{i=1}^n h_i^n g((x_{i-1}^n + x_i^n)/2, u_i^n, t_{i-1/2}^n) + P_n(x_n^n),$$

subject to the constraints

$$(2.4) \quad \begin{aligned} x_{i+1}^n &= x_i^n + h_{i+1}^n f_n(x_{i+1/2}^n, u_{i+1}^n, t_{i+1/2}^n), \quad i = 0, 1, \dots, n-1, \\ x_i^n &= \phi(t_i^n), \quad i = 0, -1, \dots, -N, \end{aligned}$$

and

$$\{u_i^n: i = 1, 2, \dots, n\} \text{ is in } U_n,$$

where we employ the following notations: The mesh function x_i^n is to approximate x^* at t_i^n , u_i^n is to approximate u^* at $t_{i-1/2}^n$ where $t_{i-1/2}^n \equiv t_{i-1}^n + h_i^n/2$, $\beta(i, n)$ is chosen so that $t_{\beta(i, n)}^n$ is the first mesh point less than or equal to $t_{i+1/2}^n - r$, $x_{\beta(i, n)}^n$ denotes the set of mesh function values $\{x_{\beta(i, n)}^n, x_{\beta(i, n)+1}^n, \dots, x_{i+1}^n\}$, $x_{\gamma(n)}^n$ denotes the set of mesh function values $\{x_{\gamma(n)}^n, x_{\gamma(n)+1}^n, \dots, x_n^n\}$ where $\gamma(n)$ is the index of the first mesh point less than or equal to $t_n^n - r$, and where f_n and P_n are discrete approximations to f and P

discussed further below. The difference method in (2.4) is implicit, and in specific examples is chosen so that the truncation error of (2.4) as an approximation to (2.2) is order $O((h_i^n)^2)$. Examples are given later in this section and in § 4. The set $u_n = \{u_i^n: i = 1, 2, \dots, n\}$ is an element of U_n if $\max_i |u_i^n| \leq K$. The set U_n is then simply a discretization of U . The basic idea of the method is relatively simple. It is essentially based on the midpoint integration rule which is a consequence of Taylor's formula: if x is in C^3 , then $\int_0^h \dot{x}(t) dt = h\dot{x}(h/2) + O(h^3)$.

In many cases it is possible through the use of necessary conditions to locate the discontinuities of u^* and x^* a priori. For example, suppose we wished to minimize the unconstrained problem

$$(2.5) \quad J(u) = \int_0^T (x^T(s)Wx(s) + u^T(s)Vu(s)) ds$$

with

$$(2.6) \quad \dot{x}(t) = A_1(t)x(t-r) + A_2(t)x(t) + B(t)u(t)$$

for $0 \leq t \leq T$ and $x(t) = \phi(t)$ for $-r \leq t \leq 0$. Here A_1 , A_2 , B , W and V are smooth matrix functions of the proper dimension with W symmetric positive semi-definite and V symmetric positive definite. By a result of [6], if u^* and x^* are an optimal solution to (2.5)–(2.6), then there is a constant $\eta_0 < 0$ and an R^l -valued function $\eta(t)$ so that

$$(2.7) \quad \dot{\eta}(t) = -\eta(t+r)^T A_1(t+r) - \eta(t)^T A_2(t) - 2\eta_0 x^*(t)^T W \quad \text{for } 0 \leq t < T-r,$$

$$(2.8) \quad \dot{\eta}(t) = -\eta(t)^T A_2(t) - 2\eta_0 x^*(t)^T W \quad \text{for } T-r \leq t \leq T, \quad \eta(t) = 0, \quad t \geq T,$$

and

$$(2.9) \quad u^*(t) = -\frac{1}{2\eta_0} V^{-1} B(t) \eta(t).$$

From (2.6) we have that x^* is continuous and from (2.7)–(2.8) we have η is continuous. Thus by (2.9), u^* is continuous. Now returning to (2.6) we have that \dot{x}^* is continuous. We may continue in this “boot-strapping” fashion and locate possible jump discontinuities for the third derivative of x^* and the second of u^* among r , $2r$, $T-r$ and $T-2r$ only, and then include them in our mesh. Further necessary conditions are contained in the paper of Jacobs and Kao [10].

Restricting u to a bounded set U is necessary for our convergence proof. The identification of points of discontinuity for the constrained problem is a difficult one. For an unconstrained problem, one can more easily locate these points as indicated above. Also for the unconstrained problem, one can in many cases bound the control a priori. Then one may artificially add the constraint u in U with U taken sufficiently large to apply our theoretical results. This is the approach of [5].

We now make the following assumptions on the problem (2.1)–(2.2).

(a) There are constants M_1 and M_2 so that

$$|f(\zeta, u, t) - f(\psi, v, t)| \leq M_1 \|\zeta - \psi\|_C + M_2 |u - v|,$$

for all t in $[0, T]$, ζ, ψ in C , and u, v in U . Also we assume f has two continuous derivatives with respect to each of its arguments.

(b) For any bounded set K in R^k there are constants M_3 and M_4 so that

$$|g(x, u, t) - g(y, v, t)| \leq M_3 |x - y| + M_4 |u - v|$$

for all t in $[0, T]$, x, y in K and u, v in U .

(c) For any bounded set K in C there is a constant M_5 so that

$$|P(\zeta) - P(\psi)| \leq M_5 \|\zeta - \psi\|_C$$

for all ζ and ψ in K .

(d) An optimal control u^* exists with corresponding response x^* defined by (2.2). The first three derivatives of x^* on $[0, T]$ are assumed piecewise continuous with the set of these discontinuities being finite and the discontinuities all being finite jumps. The functions u^* and ϕ are assumed to have two piecewise continuous derivatives with the set of these discontinuities finite and the discontinuities all being finite jumps. Designate this set of discontinuities plus 0, T and $-r$ by Σ .

Hypothesis (a) is satisfied by linear problems. It is also stronger than is actually required for applications. Since controls are only considered in the bounded set U , if it is assumed that solutions $x(\cdot, u)$ to (2.2) are uniformly bounded, then hypothesis (a) can be relaxed to a local Lipschitz assumption on f . The uniform boundedness of these solutions can often be verified by a variety of devices, not just a global Lipschitz condition. For certain delay problems, say $f(x_t, u, t) = f(x(t-r), u(t), t)$, an argument analogous to the method of steps [7] can be used for continuous functions f to bound $x(t)$ in terms of f , u and ϕ . Since u is restricted to a bounded set, it then follows that x remains in a bounded set and so hypothesis (a) would be satisfied over the relevant set of solutions $x = x(\cdot, u)$ with f continuously differentiable. See the last example in § 4.

The differentiability condition on f need only be satisfied piecewise in t in order to later prove convergence. One simply puts mesh points at the discontinuities. This is a well-known advantage of the midpoint method (see for example Keller [11]) as compared to other difference methods including the trapezoid rule. We will not discuss this further here but will simply assume f has two continuous derivatives. However, see the third example in § 4.

If $g(x, u, t)$ has the scalar form $a(t)x^2(t) + b(t)u^2(t)$, then

$$g(x, u, t) - g(y, v, t) = a(t)(x(t) + y(t))(x(t) - y(t)) + b(t)(u(t) + v(t))(u(t) - v(t)).$$

Thus this quadratic g will satisfy hypothesis (b). Similar remarks hold true for P in hypothesis (c) when P has the typical form $P(\phi) = (\phi(T))^2$.

The mappings f_n are taken to be defined for the discrete mesh functions $x_{i+1/2}^n$ defined earlier, u_n in U_n and $t_{i+1/2}^n$, $i = 0, 1, \dots, n-1$, so that $f_n(x_{i+1/2}^n, u_n, t_{i+1/2}^n)$ is in R^k . P_n is taken to be a mapping from the discrete mesh functions $x_{i+1/2}^n$ defined earlier to the reals.

We assume that the approximations f_n and P_n satisfy the following conditions:

(e) Let f_n be defined so that for any two mesh functions x_n, y_n mapping Σ_n into R^k and for any u_n in U_n one has

$$\begin{aligned} & |f_n(x_{i+1/2}^n, u_{i+1}^n, t_{i+1/2}^n) - f_n(y_{i+1/2}^n, u_{i+1}^n, t_{i+1/2}^n)| \\ & \leq M_6 \max \{ |x^n(t_{\beta(i,n)+j}^n) - y^n(t_{\beta(i,n)+j}^n)| : 0 \leq j \leq i+1 - \beta(i, n) \}, \end{aligned}$$

where M_6 is independent of i, n, x_n, y_n and u_n in U_n .

(f) If $x(t)$ defined over $[-r, T]$ has three piecewise continuous derivatives over $[0, T]$ and two piecewise continuous derivatives over $[-r, 0]$ with only jump discontinuities all included in the points of Σ , then there is a constant M_7 that depends only on the maximum of the second or third derivative of $x(t)$ so that

$$|f(x_{i+1/2}^n, u_{i+1}^n, t_{i+1/2}^n) - f_n(x_{i+1/2}^n, u_{i+1}^n, t_{i+1/2}^n)| \leq M_7 (h_{i+1}^n)^2$$

holds uniformly in i, n and u in U_n where $x_{i+1/2}^n = \{x(t_{\beta(i,n)}^n), \dots, x(t_{i+1}^n)\}$.

(g) Define K to be the set of mesh functions mapping Σ_n into R^k so that $\max |x_n(t_{\gamma(n)+j}^n)| \leq K : 0 \leq j \leq n - \gamma(n)$. We assume there is a constant M_8 depending on K so that for any two mesh functions x_n and y_n in K ,

$$|P_n(x_n^n) - P_n(y_n^n)| \leq M_8 \max \{|x_n(t_{\gamma(n)+j}^n) - y_n(t_{\gamma(n)+j}^n)| : 0 \leq j \leq n - \gamma(n)\}$$

where $M_8(K)$ is assumed independent of n . Finally,

(h) If x and x_n are as in (f), then there is a constant M_9 so that

$$|P(x_{t_n}) - P_n(x_{t_n}^n)| \leq M_9 h_n^2,$$

for all n , where M_9 depends continuously on the maximum of value of the second or third derivative of x on $[0, T]$.

Typically, P has the simple form $P(x) = (x(T))^2$. In such cases P_n would be taken to be P itself and so (h) would be trivially satisfied. Hypothesis (e) can be relaxed to a local condition in an analogous manner to the relaxation of hypothesis (a) as discussed earlier. To simplify the exposition somewhat, we will simply use hypotheses (a) and (e) as stated.

We next consider pairs f, f_n meeting the hypotheses (e) and (f).

Example 1. Suppose $f(x, u, t) = a(t)x(t) + b(t)x(t-r) + c(t)u(t)$ and one chooses the mesh spacings to be uniform and so that if t_i^n is a mesh point, then so is $t_i^n - r$. Then $f_n(x_{t_{i+1/2}^n}^n, u_{i+1}^n, t_{i+1/2}^n)$ would be taken to be

$$a(t_{i+1/2}^n)(x_{t_{i+1/2}^n}^n + x_{t_{i+1/2}^n - r}^n)/2 + b(t_{i+1/2}^n)(x_i^n + x_{i+1}^n)/2 + c(t_{i+1/2}^n)u_{i+1}^n.$$

The Lipschitz assumption of hypothesis (e) will be satisfied and the truncation error assumption of (f) can be seen to be met by an application of Taylor's theorem since $f(x_{t_{i+1/2}^n}^n, u_{i+1}^n, t_{i+1/2}^n) = a(t_{i+1/2}^n)x(t_{i+1/2}^n - r) + b(t_{i+1/2}^n)x(t_{i+1/2}^n) + c(t_{i+1/2}^n)u_{i+1}^n$.

Example 2. Consider the same example as Example 1 but suppose the mesh spacings are not uniform. Then $t_{i+1/2}^n - r$ will lie between two mesh points, namely $t_{\beta(i,n)}^n$ and $t_{\beta(i,n)+1}^n$, but will not necessarily be their midpoint. In this case, a weighted average need be taken i.e.

$$f_n(x_{t_{i+1/2}^n}^n, u_{i+1}^n, t_{i+1/2}^n) = a(t_{i+1/2}^n)(\theta_i^n x_{\beta(i,n)}^n + (1 - \theta_i^n) x_{\beta(i,n)+1}^n) + b(t_{i+1/2}^n) \cdot (x_i^n + x_{i+1}^n)/2 + c(t_{i+1/2}^n)u_{i+1}^n$$

where $\theta_i^n = (t_{\beta(i,n)+1}^n - (t_{i+1/2}^n - r)) / (t_{\beta(i,n)+1}^n - t_{\beta(i,n)}^n)$ and $0 \leq \theta_i^n \leq 1$. Hypotheses (e) and (f) will then follow as in the preceding example.

Example 3. Suppose that $f(x, u, t)$ has the form $a(t) \int_{t-r}^t x(s) ds + u(t)$. Suppose the mesh spacings are as in Example 1. Then take

$$f_n(x_{t_{i+1/2}^n}^n, u_{i+1}^n, t_{i+1/2}^n) = a(t_{i+1/2}^n) \sum_{j=\beta(i,n)+1}^i h_n x_j^n + u_{i+1}^n,$$

i.e., the midpoint integration rule. Again, hypotheses (e) and (f) will be satisfied.

3. Convergence. The convergence of the optimal values of (2.3)–(2.4) to the optimal value of (2.1)–(2.2) is established by the following theorem, which is the main result of this section. An additional technical hypothesis on the mesh, (i), is needed and is discussed before Lemma 3.3 in this section.

THEOREM 3.1. *Let assumptions (a) through (i) hold and let the sequence of meshes $\{\Sigma_n\}$ chosen be refinements of Σ and satisfy $h_n \leq \lambda/n$ for some constant $\lambda > 0$ and all n sufficiently large. Let u_n^* be an optimal control for (2.3)–(2.4) and let u^* be an optimal control for (2.1)–(2.2). Then*

$$|J_n(u_n^*) - J(u^*)| = O(h_n^2).$$

The proof of this theorem will be established in this section through a sequence of lemmas. We first give some continuity results about the problem (2.1)–(2.2). The first lemma follows from results of Hale [9].

LEMMA 3.1. *Let (a) and (d) hold. Then for any control u in $L^2[0, T]$, an absolutely continuous solution to (2.2) exists and is unique.*

LEMMA 3.2. *Let (a) hold. Suppose that u and v are elements of $L^2[0, T]$ and that x and y are the corresponding solutions of (2.2) with the initial condition $x_0 = y_0 = \phi$. Then there is a constant c independent of u and v such that*

$$\max_{0 \leq t \leq T} |x(t) - y(t)| \leq c \|u - v\|_{L^2[0, T]}.$$

Proof. Let $\sigma(t) = \max_{-r \leq s \leq t} |x(s) - y(s)|$ and note that σ is nonnegative and non-decreasing. Choose t in $[0, T]$. Then using (a) and the Schwarz inequality, we have

$$\begin{aligned} |x(t) - y(t)| &\leq \int_0^t |f(x_s, u(s), s) - f(y_s, v(s), s)| ds \\ &\leq M_1 \int_0^t \|x_s - y_s\|_C ds + M_2 \int_0^t |u(s) - v(s)| ds \\ &\leq M_1 \int_0^t \sigma(s) ds + M_2 \sqrt{T} \|u - v\|_{L^2[0, T]}. \end{aligned}$$

Thus

$$\begin{aligned} \sigma(t) &\leq \max_{0 \leq z \leq t} \left\{ M_1 \int_0^z \sigma(s) ds \right\} + M_2 \sqrt{T} \|u - v\|_{L^2} \\ &\leq M_1 \int_0^t \sigma(s) ds + M_2 \sqrt{T} \|u - v\|_{L^2}, \end{aligned}$$

and so by Gronwall's inequality

$$\sigma(t) \leq M e^{MT} \sqrt{T} \|u - v\|_{L^2}$$

where $M = \max \{M_1, M_2\}$ and the lemma follows.

The next lemma essentially establishes the stability of (2.4). However, we will need one additional hypothesis on the mesh spacings and f . Denote the solution of (2.2) corresponding to a piecewise constant function u_n over Σ'_n as $x(t, u_n)$. We additionally assume: (i) that $x(t, u_n)$ has three piecewise continuous derivatives with discontinuities only at points of Σ'_n , and that the first three derivatives of $x(t, u_n)$ are bounded by a constant independent of n and u_n in U_n . Since $\dot{x} = f(x_t, u_n, t)$, assumption (i) says that discontinuities of x are generated only at points of discontinuity of u_n and also points that can be predicted based on the nature of the dependence of f on x_t . As an example, suppose $f(x_t, u, t) = ax(t-r) + bx(t) + u(t)$. Then differentiating the equation $\dot{x} = f(x_t, u_n, t)$ shows that discontinuities in \ddot{x} and \ddot{x} will occur at points of the form t_i^n , $t_i^n + r$ and $t_i^n + 2r$. Thus, in this case, our meshes $\Sigma_n = \{t_i^n\}$ should be such that for $0 \leq t_i^n$, the points $t_i^n + r \leq T$ and $t_i^n + 2r \leq T$ should also be in Σ_n . The boundedness of the derivatives of x follows from Lemma 3.2 and the form of the equation. As a second example, suppose $f(x_t, u, t) = \int_{t-r}^t x(s) dx + bu(t)$. Then differentiating shows that discontinuities in \ddot{x} and \ddot{x} are introduced at points of the forms t_i^n and $t_i^n + r$.

We assume throughout the remainder of this section that the hypotheses of Theorem 3.1 hold.

LEMMA 3.3. Let $x_n = x_n(t, u_n)$ denote the solution of (2.4) corresponding to u_n in U_n and let $x(t, u_n)$ denote the solution to (2.2) corresponding to u_n in U_n where $u_n(t) = u_i^n$ on $(t_{i-1}^n, t_i^n]$. Then there is a constant c independent of n and u_n in U_n so that

$$\max_{0 \leq i \leq n} |x(t_i^n, u_n) - x_i^n| \leq ch_n^2.$$

Proof. We first note that with $h_n M_6 < 1$ and assumption (e), a simple contraction mapping argument shows that $x_n = x_n(t, u_n)$ is well defined. Define $x(t) = x(t, u_n)$. By assumption (i), it follows using the midpoint rule that

$$(3.1) \quad x(t_{i+1}^n) = x(t_i^n) + h_{i+1}^n f(x_{t_{i+1}^n/2}^n, u_{i+1}^n, t_{i+1}^n/2) + B_{i,1}(h_{i+1}^n)^3$$

where $B_{i,1}$ depends on the third derivative of x . By use of (3.1), it follows that

$$(3.2) \quad \begin{aligned} x(t_{i+1}^n) - x_{i+1}^n &= x(t_i^n) - x_i^n + h_{i+1}^n f(x_{t_{i+1}^n/2}^n, u_{i+1}^n, t_{i+1}^n/2) \\ &\quad - h_{i+1}^n f_n(x_{t_{i+1}^n/2}^n, u_{i+1}^n, t_{i+1}^n/2) + B_{i,1}(h_{i+1}^n)^3. \end{aligned}$$

Define the function $e(i)$ for $i = -N, -N+1, \dots, 0, 1, \dots, n$ by $e(i) = \max_{j \leq i} |x(t_j^n) - x_j^n|$. Note that $x_i^n = x(t_i^n) = \phi(t_i^n)$ for $i = 0, -1, \dots, -N$ so that $e(i) = 0$ for $i = 0, -1, \dots, -N$. Now from (3.2) and using assumptions (e) and (f) we have

$$\begin{aligned} |x(t_{i+1}^n) - x_{i+1}^n| &\leq |x(t_i^n) - x_i^n| + M_7(h_{i+1}^n)^3 \\ &\quad + M_6 h_{i+1}^n e(i+1) + B_1(h_{i+1}^n)^3, \end{aligned}$$

and so

$$e(i+1) \leq e(i) + h_{i+1}^n M_6 e(i+1) + B_2(h_{i+1}^n)^3$$

where $|B_{i,1}| \leq B_1$ for all i and n and $B_2 = M_7 + B_1$. Then an induction argument shows that for $n \geq n_0$, n_0 sufficiently large, and some constant $\alpha > 0$

$$\begin{aligned} e(i+1) &\leq \left(\sum_{j=1}^{i+1} (1 - h_j^n M_6)^{-j} h_j^n \right) h_n^2 B_2 \leq \left(\sum_{j=1}^{i+1} e^{\alpha j h_n} h_j^n \right) h_n^2 B_2 \\ &\leq T e^{\alpha \lambda} h_n^2 B_2 = O(h_n^2), \end{aligned}$$

which completes the proof.

LEMMA 3.4. If $u_n = \{u_i^n\} = \{u^*(t_{i+1/2}^n)\}$ where u^* is a solution to (2.1)–(2.2), then there is a constant c independent of n so that

$$\max_{0 \leq i \leq n} |x^*(t_i^n) - x_i^n| \leq ch_n^2$$

where $x_i^n = x_n(t_i^n, u_n)$.

Proof. By assumption (d), it follows that

$$x^*(t_{i+1}^n) = x^*(t_i^n) + h_{i+1}^n f(x_{t_{i+1}^n/2}^*, u_{i+1}^n, t_{i+1}^n/2) + B_i^*(h_{i+1}^n)^3$$

where B_i^* depends on the third derivative of x^* . This is the same equation as (3.1), and so the proof follows as in Lemma 3.3 with no change.

LEMMA 3.5. There is a constant $c > 0$ so that

$$|J(w_n) - J_n(w_n)| \leq ch_n^2$$

for all w_n in U_n , considered as step functions over Σ'_n .

Proof. Let $x(t) = x(t, w_n)$ denote the solution of (2.2) with w_n treated as a step function and let $x_n(t) = x_n(t, w_n)$ be the solution of (2.4). Then

$$|P(x_T) - P_n(x_{t_n}^h)| \leq |P(x_T) - P_n(z_{t_n}^h)| + |P_n(z_{t_n}^h) - P_n(x_{t_n}^h)|$$

where $z_i^n = x(t_i^n)$. The first term on the right in this inequality is less than or equal to $M_9 h_n^2$ by assumption (h). By assumption (g) we have

$$|P_n(z_{t_n}^h) - P_n(x_{t_n}^h)| \leq M_8 \max_j |z_j^n - x_j^n|,$$

and applying Lemma 3.3, we thus have a similar estimate for the second term.

Now for $i = 0, 1, \dots, n-1$,

$$\begin{aligned} (3.3) \quad & \left| \int_{t_i^n}^{t_{i+1}^n} \left[g(x(t), u_{i+1}^n, t) - g\left(\frac{x_{i+1}^n + x_i^n}{2}, u_{i+1}^n, t_{i+1/2}^n\right) \right] dt \right| \\ & \leq \left| \int_{t_i^n}^{t_{i+1}^n} \left[g(x(t), u_{i+1}^n, t) - g\left(\frac{x(t_i^n) + x(t_{i+1}^n)}{2}, u_{i+1}^n, t_{i+1/2}^n\right) \right] dt \right| \\ & \quad + \left| \int_{t_i^n}^{t_{i+1}^n} \left[g\left(\frac{x(t_i^n) + x(t_{i+1}^n)}{2}, u_{i+1}^n, t_{i+1/2}^n\right) \right. \right. \\ & \quad \left. \left. - g\left(\frac{x_{i+1}^n + x_i^n}{2}, u_{i+1}^n, t_{i+1/2}^n\right) \right] dt \right|. \end{aligned}$$

The first term on the right in (3.3) is $O((h_{i+1}^n)^3)$ since it may be estimated by using the error term for the midpoint integration rule. The second term is $O((h_{i+1}^n)^3)$ from hypothesis (b) and Lemma 3.3. Since $\sum_{i=1}^n O((h_i^n)^3) = O(h_n^2)$, the result follows.

We now may prove Theorem 3.1. First note that

$$J_n(u_n^*) - J(u^*) \leq J_n(w_n) - J(u^*)$$

where $w_{i+1}^n = u^*(t_{i+1/2}^n)$. Then

$$\begin{aligned} J_n(w_n) - J(u^*) &= J_n(w_n) - J(w_n) + J(w_n) - J(u^*) \\ &\leq ch_n^2 + J(w_n) - J(u^*). \end{aligned}$$

Now

$$J(w_n) - J(u^*) = \sum_{i=0}^{n-1} \int_{t_i^n}^{t_{i+1}^n} (g(x(t, w_n), w_n, t) - g(x(t, u^*), u^*, t)) dt.$$

By two applications of the midpoint integration rule,

$$\begin{aligned} & \int_{t_i^n}^{t_{i+1}^n} (g(x(t, w_n), w_n, t) - g(x(t, u^*), u^*, t)) dt \\ &= \int_{t_i^n}^{t_{i+1}^n} [g(x(t, w_n), u^*(t_{i+1/2}^n), t) - g(x(t_{i+1/2}^n, u^*), u^*(t_{i+1/2}^n), t_{i+1/2}^n)] dt \\ & \quad + \int_{t_i^n}^{t_{i+1}^n} [g(x(t_{i+1/2}^n, u^*), u^*(t_{i+1/2}^n), t_{i+1/2}^n) - g(x(t, u^*), u^*, t)] dt \\ &= \int_{t_i^n}^{t_{i+1}^n} [g(x(t, w_n), u^*(t_{i+1/2}^n), t) - g(x(t_{i+1/2}^n, u^*), u^*(t_{i+1/2}^n), t_{i+1/2}^n)] dt + O((h_{i+1}^n)^3) \\ &= \int_{t_i^n}^{t_{i+1}^n} [g(x(t, w_n), u^*(t_{i+1/2}^n), t) - g(x(t_{i+1/2}^n, w_n), u^*(t_{i+1/2}^n), t_{i+1/2}^n)] dt \end{aligned}$$

$$\begin{aligned}
& + \int_{t_i^n}^{t_{i+1}^n} [g(x(t_{i+1/2}^n, w_n), u^*(t_{i+1/2}^n), t_{i+1/2}^n) - g(x(t_{i+1/2}^n, u^*), u^*(t_{i+1/2}^n), t_{i+1/2}^n)] dt \\
& \quad + O((h_{i+1}^n)^3) \\
& = \int_{t_i^n}^{t_{i+1}^n} [g(x(t_{i+1/2}^n, w_n), u^*(t_{i+1/2}^n), t_{i+1/2}^n) - g(x(t_{i+1/2}^n, u^*), u^*(t_{i+1/2}^n), t_{i+1/2}^n)] dt \\
& \quad + O((h_{i+1}^n)^3).
\end{aligned}$$

Then using the Lipschitz continuity of g , we have that the integral on the right in this last equality is less than or equal to

$$\begin{aligned}
M_3 h_{i+1}^n |x(t_{i+1/2}^n, w_n) - x(t_{i+1/2}^n, u^*)| & \leq M_3 h_{i+1}^n |x(t_{i+1/2}^n, w_n) - (x_i^n(w_n) + x_{i+1}^n(w_n))/2| \\
& \quad + M_3 h_{i+1}^n |(x_i^n(w_n) + x_{i+1}^n(w_n))/2 - x(t_{i+1/2}^n, u^*)|.
\end{aligned}$$

Using the triangle inequality, we have

$$\begin{aligned}
\left| x(t_{i+1/2}^n, w_n) - \frac{x_i^n(w_n) + x_{i+1}^n(w_n)}{2} \right| & \leq \left| x(t_{i+1/2}^n, w_n) - \frac{x(t_i^n, w_n) + x(t_{i+1}^n, w_n)}{2} \right| \\
& \quad + \left| \frac{x(t_i^n, w_n) + x(t_{i+1}^n, w_n)}{2} - \frac{x_i^n(w_n) + x_{i+1}^n(w_n)}{2} \right|.
\end{aligned}$$

The first term on the right hand side of this inequality can be bounded by $O(h_n^2)$ using Taylor's theorem and assumption (i). The second term can be similarly bounded by use of Lemma 3.3. The expression $|(x_i^n(w_n) + x_{i+1}^n(w_n))/2 - x(t_{i+1/2}^n, u^*)|$ is bounded by $O(h_n^2)$ in a similar manner by the use of Taylor's theorem and Lemma 3.4. We thus have

$$(3.4) \quad J_n(u_n^*) - J(u^*) \leq O(h_n^2).$$

Now by Lemma 3.5,

$$J_n(u_n^*) - J(u^*) = J_n(u_n^*) - J(u_n^*) + J(u_n^*) - J(u^*) \geq -ch_n^2 + J(u_n^*) - J(u^*).$$

But $J(u^*) \leq J(u_n^*)$ by definition of $J(u^*)$, so

$$(3.5) \quad J_n(u_n^*) - J(u^*) \geq -ch_n^2.$$

Combining (3.4) and (3.5), we have proved the theorem.

Theorem 3.1 says that we have a minimizing sequence and so with the usual convexity assumptions on g , one can in certain cases deduce L^2 -convergence of the discrete controls u_n^* to u^* . We omit statements of these results and refer the reader to [5] for details.

4. Numerical examples. Consider first the scalar problem

$$\min J(u) = \frac{1}{2} \gamma [x(3)]^2 + \frac{1}{2} \int_0^3 (u(t))^2 dt$$

subject to the constraints

$$\begin{aligned}
\dot{x}(t) &= x(t-1) + u(t), & 0 < t < 3, \\
x(t) &= 1, & -1 \leq t \leq 0.
\end{aligned}$$

The optimal control, u^* , which solves this problem is [3]

$$u^*(t) = \begin{cases} \delta\{-(t-2)^2/2-1.5\}, & 0 \leq t \leq 1, \\ \delta(t-3), & 1 < t \leq 2, \\ -\delta, & 2 < t \leq 3, \end{cases}$$

where $\delta = 37\gamma[6(1+319\gamma/30)]$. Note that u^* is continuous on $[0, 3]$ but its first derivative is discontinuous at $t = 2$ and its second derivative is discontinuous at $t = 1$. The value $J(u^*)$ is $.5(37/6)^2/(1/\gamma+319/30)$. We solved this problem numerically using a sequence of uniform mesh spacings with $h = 1/n$. The discretization (2.3)–(2.4) then leads to a quadratic programming problem with $6n$ variables. For our discretization, we chose $P_n = P$ and $f_n = f$. Thus $f_n(x_{i+1/2}^n, u_{i+1}^n, t_{i+1/2}^n)$ was in this case taken to be $(x_{i-n}^n + x_{i-n+1}^n)/2 + u_{i+1}^n$. Without the superscripts, (2.4) became in this case

$$\frac{x_{i+1} - x_i}{h} = \frac{x_{i-n} + x_{i-n+1}}{2} + u_{i+1}, \quad i = 1, \dots, 3n-1,$$

and (2.3) was taken to be

$$\frac{1}{2}\gamma(x_{3n})^2 + \sum_{i=1}^{3n} (u_i)^2 h.$$

The following table, Table 1, contains numerical results that we obtained for several values of n . The parameter γ was set to 3 and $J(u^*) \approx 1.73379$. The calculations were done in single precision. Under the assumption that the error has the form $C(1/n)^\beta$, the observed value of β has been computed although the numbers are too accurate for β to be correct to more than one digit. See also the second example.

TABLE 1

n	$J_n(u_n^*)$	$ J(u^*) - J_n(u_n^*) $	β
5	1.73460	$8.1 \cdot 10^{-4}$	—
9	1.73402	$2.3 \cdot 10^{-4}$	2.1
13	1.73388	$0.9 \cdot 10^{-4}$	2.5

Table 2 reports values of u_n^* at several points.

TABLE 2

t	$u_5^*(t)$	$u_9^*(t)$	$u_{13}^*(t)$	$u^*(t)$
.5	-1.47877	-1.47688	-1.47645	-1.47606
1.5	-0.84341	-0.84345	-0.84346	-0.84346
2.5	-0.56227	-0.56230	-0.56231	-0.56231

The following table, Table 3, contains numerical results for this problem reported in [3] using the projection method of Banks and Burns [2]. Their discretization involves a parameter n that leads to an approximating control system with an $(n+1)$ -dimensional system of ordinary differential equations replacing the state equation and a performance criterion in this case involving $(n+1) \times (n+1)$ matrices. They produced solutions by integrating the resulting necessary condition using two schemes, both involving fourth order discretizations, and their computations were done in single precision. Assuming their method behaves as $C(1/n)^\beta$, we also computed an observed

β for their results. Their method uses orthogonal projections into spaces of piecewise constant functions over a mesh with spacing $1/n$ to define their approximating equations. If they had discretized both the state equation and cost functional directly with a spacing of $1/n$, then our two methods would lead to nonlinear programming problems of exactly the same size. Thus the n in Table 3 is the same as the n in Tables 1 and 2. We also integrated the necessary conditions using the midpoint rule and saw no significant change in the values of our approximations.

TABLE 3

n	$ J_n - J(u^*) $	β
9	$1.89 \cdot 10^{-2}$	—
13	$1.32 \cdot 10^{-2}$.98
17	$1.01 \cdot 10^{-2}$	1.00
20	$8.60 \cdot 10^{-3}$.99

They also obtained $u_9^*(1.5) = -.8240$ and $u_{20}^*(1.5) = -.8309$ to compare with our Table 2. Their method appears to be linearly convergent whereas ours is second order. In addition, our absolute error estimates are smaller.

As a second example, consider the problem

$$\min J(u) \equiv \frac{1}{2} \gamma [y(2)]^2 + \frac{1}{2} \int_0^2 (u(t))^2 dt$$

with

$$\ddot{y}(t) + \dot{y}(t-1) + y(t) = u(t), \quad 0 < t < 2,$$

$$y(t) \equiv 10, \quad -1 \leq t \leq 0.$$

In order to treat this problem by our method, we need to rewrite the state equation as a first order vector system by defining $x_1(t) = y(t)$, $x_2(t) = \dot{y}(t)$ and then letting $x(t) = [x_1(t) x_2(t)]^T$. As in the first example, we let $h = 1/n$. The dimension of the discrete problem is $6n$, $4n$ for the state variables and $2n$ for the control. The solution to this problem is [3]

$$u^*(t) = \begin{cases} \delta \sin(2-t) + \frac{\delta}{2}(1-t) \sin(t-1), & 0 \leq t \leq 1, \\ \delta \sin(2-t), & 1 < t \leq 2, \end{cases}$$

where $\delta = \gamma 2.655625 / (1 + \gamma .937378)$. Also $J(u^*) = 7.052344 / 2(1/\gamma + .937378)$. For $\gamma = 10$, $J(u^*) = 3.39912$ and we obtained the numerical results in single precision shown in Table 4.

TABLE 4

n	$J_n(u_n^*)$	$ J_n(u_n^*) - J(u^*) $	β
5	3.27916	.12	—
9	3.36199	.04	1.99
13	3.38131	.02	2.00

In this problem, the second derivative of u^* has a discontinuity at $t = 1$. We also obtained the following values for the approximating controls.

TABLE 5

t	$u_5^*(t)$	$u_9^*(t)$	$u_{13}^*(t)$	$u^*(t)$
.5	2.20222	2.23300	2.24012	2.24670
1.5	1.20369	1.22003	1.22383	1.22730

The following table, Table 6, contains the results of [3] for this problem. They also rewrite the second order state equation as a first order vector system, and so the dimensions of the fully discretized approximating systems are the same.

TABLE 6

n	J_n	$ J_n - J(u^*) $	β
5	2.3391	1.06	—
9	2.7390	.66	.81
13	2.9205	.48	.87
17	3.0238	.38	.91
48	3.2587	.14	.95

A comparison of Table 4 with Table 6 shows that our value for $n = 13$ is about a factor seven better than their value for $n = 48$. We also appear to have computationally confirmed Theorem 3.1.

We applied our method to a problem with a target constraint on the response to compare with an example of Banks and Manitius [4], where their new projection series method was used. We have not as yet been able to extend Theorem 3.1 to this class of problems. The problem is

$$\min J(u) = \int_0^2 u^2(t) dt$$

subject to

$$\dot{x}(t) = \frac{1}{\sqrt{2}}x(t) + \frac{1}{\sqrt{2}}x(t-1) + u(t), \quad 0 < t < 2,$$

and $x(t) = 1$ on $[-1, 0]$, $x(t) = 0$ on $[1, 2]$. The exact solution [4] $J(u^*) \approx 4.3975$, and u^* has a jump discontinuity at $t = 1$. We obtained the numerical results in the next table, Table 7.

TABLE 7

n	$ J_n(u_n) - J(u^*) $	β
5	$4.41 \cdot 10^{-4}$	—
10	$1.11 \cdot 10^{-4}$	1.98
20	$2.79 \cdot 10^{-5}$	2.00

We treated the target objective as additional constraints and so had $4n$ variables. The method of [2] involves considering controls in an n dimensional space. If they used a discretization with spacing $1/n$ on the state equation, they would have a nonlinear programming problem with $3n$ variables. They obtained the results in Table 8.

TABLE 8

n	$ J_n^* - J(u^*) $	β
5	$6.45 \cdot 10^{-2}$	—
11	$3.45 \cdot 10^{-2}$.79
21	$2.05 \cdot 10^{-2}$.80
41	$1.25 \cdot 10^{-2}$.74

Their method also appears to be linearly convergent at best. The midpoint rule is necessary here in our numerical method since u^* has a jump discontinuity. Using the trapezoidal rule on both (2.3) and (2.4) produced the results in Table 9.

TABLE 9

n	$ J_n(u_n) - J(u^*) $	β
5	$4.71 \cdot 10^{-2}$	—
10	$2.11 \cdot 10^{-2}$	1.16
20	$1.00 \cdot 10^{-2}$	1.07

As a final example, we considered the simple nonlinear problem

$$\min J(u) = \int_0^2 u^2(t) dt + 5x^2(2)$$

subject to

$$\dot{x}(t) = x^2(t-1) + u(t), \quad 0 < t < 2,$$

and $x(t) = 1$ on $[-1, 0]$. The solution to this problem gives $J(u^*) \approx 3.094428$ where

$$u^*(t) = \begin{cases} cke^{kt} + (c-1)ke^{-kt} - 1, & 0 \leq t \leq 1, \\ -k^2/2, & 1 < t \leq 2, \end{cases}$$

and $c = (2 + 2ke^{-k} - k^2)/(2k(e^k + e^{-k}))$ and $k^2 \approx 1.76640557$. We obtained the numerical results in the following table, Table 10. The discrete problem was solved using the method of steepest descent.

TABLE 10

n	$ J_n(u_n) - J(u^*) $	β
5	$3.33 \cdot 10^{-3}$	—
9	$1.02 \cdot 10^{-3}$	2.0
13	$49 \cdot 10^{-3}$	2.0

REFERENCES

- [1] H. T. BANKS AND J. A. BURNS, *An abstract framework for approximate solutions to optimal control systems governed by hereditary systems*, International Conference on Differential Equations, H. A. Antosiewicz, ed., Academic Press, New York, 1975.
- [2] ———, *Hereditary control problems: Numerical methods based on averaging approximations*, this Journal, 16 (1978), pp. 169–208.
- [3] H. T. BANKS, J. A. BURNS, E. M. CLIFF AND P. R. THRIFT, *Numerical solutions of hereditary control problems via an approximation technique*, CDS Tech. Rep. 75–6, Division of Applied Mathematics, Brown University, Providence, RI, 1975.
- [4] H. T. BANKS AND A. MANITIUS, *Projection series for retarded functional differential equations with applications to optimal control problems*, J. Differential Equations, 18 (1975), pp. 296–332.
- [5] B. M. BUDAK, E. M. BERKOVICH AND E. N. SOLOV'YEV, *Difference approximations in optimal control problems*, this Journal, 7 (1969), pp. 18–31.
- [6] D. H. CHUNG AND E. B. LEE, *Linear optimal control systems with time delays*, this Journal, 4 (1966), pp. 568–575.
- [7] L. E. EL'SGOL'TS, *Introduction to the Theory of Differential Equations with Deviating Arguments*, Holden-Day, San Francisco, 1966.
- [8] W. W. HAGER, *Rates of convergence for discrete approximations to unconstrained control problems*, SIAM J. Numer. Anal., 13 (1976), pp. 449–472.
- [9] J. HALE, *Functional Differential Equations*, Springer-Verlag, New York, 1971.
- [10] M. A. JACOBS AND T. J. KAO, *An optimum settling problem for time lag systems*, J. Math. Anal. Appl., 40 (1972), pp. 687–707.
- [11] H. B. KELLER, *Accurate difference methods for linear ordinary differential systems subject to linear constraints*, SIAM J. Numer. Anal., 6 (1969), pp. 8–30.
- [12] I. LASIECKA, *Finite difference approximations of optimal control for systems described by nonlinear differential equations with delay*, Control and Cybernetics, 5 (1976), pp. 35–67.

MODAL CONTROL OF CERTAIN FLEXIBLE DYNAMIC SYSTEMS*

MARK J. BALAS†

This is dedicated to T. N. Edelbaum, my good friend and a great source of encouragement in this and other work; his untimely death is an immeasurable personal and professional loss.

Abstract. Interest has increased in the active control of vibrations in mechanically flexible systems, e.g. attitude control of flexible spacecraft, ride quality improvement of air and surface transportation, and active optics. To insure satisfactory performance of such systems, their distributed parameter nature must be taken into account in control system design. In this paper, we obtain feedback control of N nodes of a flexible system and treat the problem of control "spillover" into the uncontrolled modes.

We consider the class of flexible systems that can be described by a generalized wave equation, $u_{tt} + Au = F$, which relates the displacement $u(x, t)$ of a body Ω in n -dimensional space to the applied control forces $F(x, t)$. The operator A is a time-invariant, symmetric differential operator with a discrete semibounded spectrum and the control forces $F(x, t) = \sum_{i=1}^M b_i(x)f_i(t)$ are provided by M actuators with influence functions $b_i(x)$. The displacements are measured by P sensors $y_j(t)$, $j = 1, 2, \dots, P$, located at various points along the body. This class of distributed parameter systems includes interior and boundary control of vibrating strings, membranes, thin beams, and thin plates.

A feedback control is found by applying known state variable methods to a lumped parameter approximation of the above system. Such a discretization can be obtained by expanding the system state in the eigenmodes of the operator A and truncating the expansion. The N controlled modes are selected and the resulting finite dimensional approximate system must meet certain controllability and observability conditions which we display.

The feedback control so obtained will stabilize the N selected modes of the flexible system, and the energy in the uncontrolled modes is usually ignored in the design. However, in all but some very special situations, the actuators will excite these uncontrolled modes. In vibrating systems, such control "spillover" can severely degrade the system performance. We prove a result that guarantees suitable system performance with the above modal feedback control, provided that the controlled modes are controllable and observable, the remaining modes are unobservable, and the control spillover meets an a priori bound determined by the desired performance. This result justifies the use of a simple and direct modal feedback control for flexible systems which satisfy the spillover condition, and it issues a warning for modal control of systems which have excessive control spillover. To illustrate these results, a modal feedback control is obtained for a simple beam with pinned ends and the performance is evaluated.

1. Introduction. Interest has increased in the active control of vibrations in mechanically flexible systems, e.g. attitude control of flexible spacecraft [4], [11], ride quality improvement of air and surface transportation [5], [7], [18], and active optics [2]. To insure satisfactory performance of such systems, their distributed parameter nature must be taken into account in control system design. In this paper we obtain feedback control of N modes of a flexible system and treat the problem of control "spillover" into the uncontrolled modes.

We consider the class of flexible systems that can be described by a generalized wave equation

$$(1.1) \quad u_{tt} + Au = F$$

which relates the displacement $u(x, t)$ of a body Ω , a bounded open set with smooth boundary $\partial\Omega$ in n -dimensional Euclidean space R^n , to the applied control forces $F(x, t)$. The operator A is a time-invariant, symmetric differential operator with compact resolvent and lower semibounded spectrum. The domain $\mathcal{D}(A)$ of A is dense

* Received by the editors May 4, 1977, and in revised form August 18, 1977.

† Charles Stark Draper Laboratory, Cambridge, Massachusetts 02139. This work was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Deputy of Advanced Space Programs, Space and Missile Systems Organization under Contract FO-4701-76-C-0178.

in the Hilbert space $L^2(\Omega)$ with (\cdot, \cdot) denoting the usual inner product and $\|\cdot\|$ denoting the associated norm. The control forces

$$(1.2) \quad F(x, t) = Bf(t) = \sum_{i=1}^M b_i(x)f_i(t)$$

are provided by M actuators with influence functions $b_i(x)$. The displacements are measured by P averaging sensors

$$(1.3) \quad \underline{y}(t) = Cu(x, t)$$

where $y_j(t) = \int_{\Omega} c_j(x)u(x, t) dx$ with $j = 1, 2, \dots, P$. The actuator and sensor functions $b_i(x)$, $c_j(x)$ are in $L^2(\Omega)$ and normalized to have unit integral. When the support of $b_i(x)$ is in a small neighborhood of a point x_i , we say it is a *point actuator* and, similarly, we define a *point sensor*. The point actuator and point sensor situation is of special interest here. This class of distributed parameter systems includes interior and boundary control of vibrating strings, membranes, thin beams, and thin plates.

It is well known [6, p. 277] that the spectrum of A contains only isolated eigenvalues λ_k with corresponding eigenvectors ϕ_k such that

$$\lambda_1 \leq \lambda_2 \leq \dots$$

and $A\phi_k = \lambda_k\phi_k$. Without loss of generality, we will assume that λ_1 is positive. Thus A satisfies

$$(1.4) \quad (Au, u) \geq \varepsilon \|u\|^2, \quad \varepsilon > 0,$$

and has a square root $A^{1/2}$. Every vector $u \in L^2(\Omega)$ has a unique representation

$$(1.5) \quad u(x) = \sum_{k=1}^{\infty} u_k \phi_k(x)$$

where $u_k = \int_{\Omega} u \phi_k dx$ and we define the orthogonal projections P_0, Q_0 by

$$(1.6) \quad P_0 u = \sum_{k=1}^n u_k \phi_k, \quad Q_0 u = \sum_{k=n+1}^{\infty} u_k \phi_k.$$

Let $H \equiv L^2(\Omega) \times L^2(\Omega)$ with the usual inner product and let V be the domain of A and W be the domain of $A^{1/2}$. A new operator \bar{A} is defined in H by

$$(1.7a) \quad \mathcal{D}(\bar{A}) = V \times W \equiv H_1 \text{ where } \mathcal{D}(\bar{A}) \text{ denotes the domain of } \bar{A},$$

$$(1.7b) \quad \bar{A} \begin{bmatrix} v \\ w \end{bmatrix} = \begin{bmatrix} w \\ -Av \end{bmatrix} \text{ for every } v \in V, w \in W.$$

The *energy inner product* $(\cdot, \cdot)_E$ is defined on H_1 by

$$(1.8) \quad \left(\begin{bmatrix} v_1 \\ w_1 \end{bmatrix}, \begin{bmatrix} v_2 \\ w_2 \end{bmatrix} \right)_E \equiv (v_1, Av_2) + (w_1, w_2)$$

for $v_1, v_2 \in V$ and $w_1, w_2 \in W$. The associated *energy norm* is denoted by $\|\cdot\|_E$ and is a measure of the total potential and kinetic energy in (u, u_t) where u is a solution of (1.1). Let $v = [u, u_t]^T$ and write (1.1), (1.2), and (1.3) as

$$(1.9) \quad \begin{aligned} v_t &= \bar{A}v + \bar{B}f, & v_0 &\in H_1, \\ \underline{y} &= \bar{C}v, \end{aligned}$$

where $\bar{B} = \begin{bmatrix} 0 \\ B \end{bmatrix}$ and $\bar{C} = [C \ 0]$. It is easy to see that

$$(1.10a) \quad v = \sum_{k=1}^{\infty} \begin{bmatrix} u_k \\ \dot{u}_k \end{bmatrix} \phi_k$$

and

$$(1.10b) \quad \|v\|_{\mathbb{E}}^2 = \sum_{k=1}^{\infty} [\lambda_k u_k^2 + \dot{u}_k^2]$$

and, in the energy norm, \bar{A} generates a unitary group $U(t)$ [23, p. 446] such that (for $\omega_k \equiv \lambda_k^{1/2}$)

$$(1.11) \quad U(t)v_0 = \sum_{k=1}^{\infty} \begin{bmatrix} \cos \omega_k t & \sin \omega_k t / \omega_k \\ -\omega_k \sin \omega_k t & \cos \omega_k t \end{bmatrix} \begin{bmatrix} u_k(0) \\ \dot{u}_k(0) \end{bmatrix} \phi_k$$

for any $v_0 = \sum_{k=1}^{\infty} \begin{bmatrix} u_k(0) \\ \dot{u}_k(0) \end{bmatrix} \phi_k$ in H_1 . This means that, when no external forces are present (i.e., $f=0$), the energy in solutions of (1.9) is preserved:

$$(1.12) \quad \|U(t)v_0\|_{\mathbb{E}} = \|v_0\|_{\mathbb{E}}$$

for any $v_0 \in H_1$. When $f(t)$ is continuously differentiable for $t \geq 0$, the solution to (1.9) is given by [6, p. 486]

$$(1.13) \quad v(t) = U(t)v_0 + \int_0^t U(t-\tau)B\dot{f}(\tau) d\tau$$

for any $v_0 \in H_1$.

Recalling (1.6), we define the projections P, Q on H_1 by

$$(1.14) \quad P \begin{bmatrix} v \\ w \end{bmatrix} \equiv \begin{bmatrix} P_0 v \\ P_0 w \end{bmatrix}, \quad Q \begin{bmatrix} v \\ w \end{bmatrix} \equiv \begin{bmatrix} Q_0 v \\ Q_0 w \end{bmatrix}$$

and note that they are orthogonal in the energy inner product and that PH_1, QH_1 form \bar{A} invariant subspaces of H_1 . Consequently, from (1.9),

$$(1.15) \quad Pv_t = \bar{A}Pv + P\bar{B}\dot{f},$$

$$(1.16) \quad Qv_t = \bar{A}Qv + Q\bar{B}\dot{f},$$

$$(1.17) \quad y = \bar{C}Pv + \bar{C}Qv$$

and

$$(1.18) \quad \|v\|_{\mathbb{E}}^2 = \|Pv\|_{\mathbb{E}}^2 + \|Qv\|_{\mathbb{E}}^2.$$

We refer to solutions Pv of the finite dimensional subsystem (1.15) as the *controlled modes* of (1.9) and solutions Qv of the infinite dimensional subsystem (1.16) as the *residual (or uncontrolled) modes* of (1.9). From (1.18), the total energy of a solution v of (1.9) is the sum of the energies in the controlled and residual modes.

Control and controllability of the distributed parameter system (1.9) have been studied in [3], [15], [16], [17], [21], [22], and [25]. In these papers, the intention has been to control or stabilize the full state v assuming exact knowledge of it. It is our intention to develop a feedback control $\dot{f}(t)$ based only on the sensor information $y(t)$. Our philosophy is similar to R. Triggiani's [24] with generalized heat equations, where \bar{A} generates a holomorphic semigroup, in that we will develop our control for the

finite dimensional subsystem (1.15) instead of the full system (1.9). We must pay a greater price for doing this with a flexible system since the residual subsystem (1.16) does not have the strongly damped characteristic of the heat equation; however, we intend to account for the effect of the residual system to determine what performance can be expected from such a modal control. Modal control schemes similar to ours have been developed in the past for particular flexible systems [8], [10], [11], [12]. Although the degradation of performance due to the residual system has been recognized, it has not been analyzed or compensated. Two exceptions are [20] where corrections are made to the state estimator to approximate the effect of the residual system and [14] where analysis of the residual system effect is done on a class of flexible satellites.

We call the effect of control $f(t)$ entering the residual sub-system (1.16) through the term $Q\bar{B}f$ —*control spillover*—and the contamination of the sensor measurements (1.17) with residual subsystem information through the term $\bar{C}Qv$ —*observation spillover*. Both effects can be reduced by careful location and shaping of actuator and sensor influence functions but this freedom is rarely available to the control system designer; the locations are often already determined by structural considerations. Also dependence on exact location and shape can lead to highly sensitive designs. However, the effect of observation spillover can be reduced substantially by prefiltering the sensor data to remove the residual system information [1]. Consequently, in this paper, we restrict our attention to the effect of control spillover and *assume no observation spillover*, i.e.,

$$(1.17') \quad \underline{y} = \bar{C}Pv.$$

The closed loop instabilities that arise when both control and observation spillover are present are analyzed and illustrated in [1].

The modal control scheme is developed in § 2 and the controllability and observability requirements are presented there. The main results are estimates of the system behavior with control spillover present; these are presented in §§ 3 and 4. An example of modal control of a simply supported thin beam is presented in § 5 and the results are summarized with conclusions in § 6.

2. Modal control. In this section, we develop the control $f(t)$ for the finite dimensional subsystem (1.15) using well known state variable methods [9]. This *modal control scheme* consists of

1) a Luenberger observer which accepts the sensor measurements $\underline{y}(t)$ and produces an exponentially convergent estimate of Pv ,

2) a linear state variable feedback control law which multiplies the estimate of Pv by constant gains to produce the actuator commands $f(t)$.

The feedback gains can be obtained by pole allocation [19] or by optimal regulator design.

From (1.6) and (1.14), we have

$$(2.1) \quad Pv = \sum_{k=1}^N \begin{bmatrix} u_k \\ \dot{u}_k \end{bmatrix} \phi_k$$

and we define

$$(2.2) \quad \underline{v}_N \equiv \begin{bmatrix} \underline{u}_N \\ \underline{\dot{u}}_N \end{bmatrix}$$

where $\underline{u}_N \equiv [u_1, \dots, u_N]^T$ and $\dot{\underline{u}}_N \equiv [\dot{u}_1, \dots, \dot{u}_N]^T$. From (1.15) and (1.17')

$$(2.3) \quad \begin{aligned} \dot{\underline{v}}_N &= \bar{A}_N \underline{v}_N + \bar{B}_N \underline{f}, \\ \underline{y} &= \bar{C}_N \underline{v}_N \end{aligned}$$

where

$$\bar{A}_N \equiv \begin{bmatrix} 0 & I_N \\ -A_N & 0 \end{bmatrix}, \quad \bar{B}_N \equiv \begin{bmatrix} 0 \\ B_N \end{bmatrix}, \quad \bar{C}_N \equiv [C_N \quad 0]$$

and

$$A_N \equiv \begin{bmatrix} \omega_1^2 & & 0 \\ & \ddots & \\ 0 & & \omega_N^2 \end{bmatrix}, \quad B_N \equiv \begin{bmatrix} (b_1, \phi_1) & \cdots & (b_M, \phi_1) \\ \vdots & & \\ (b_1, \phi_N) & \cdots & (b_M, \phi_N) \end{bmatrix}$$

and

$$C_N \equiv \begin{bmatrix} (c_1, \phi_1) & \cdots & (c_1, \phi_N) \\ \vdots & & \\ (c_P, \phi_1) & \cdots & (c_P, \phi_N) \end{bmatrix}.$$

THEOREM 2.1. *The system $(\bar{A}_N, \bar{B}_N, \bar{C}_N)$ is controllable and observable if and only if (A_N, B_N, C_N) is controllable and observable.*

Proof. The proof is by direct examination of the controllability and observability matrices; the ones for $(\bar{A}_N, \bar{B}_N, \bar{C}_N)$ have rank $2N$ if and only if the ones for (A_N, B_N, C_N) have rank N . \square

In the case of one space dimension where the eigenvalues λ_k have unit multiplicity, the conditions of Theorem 2.1 are satisfied if and only if each row of B_N and each column of C_N have a nonzero entry. Consequently, controllability and observability can be accomplished with one point actuator and one point sensor if they are located away from the zeros of the N modes we wish to control. In the case where an eigenvalue has greater than unit multiplicity, the conditions of Theorem 2.1 are satisfied if the diagonal Jordan block of A_N associated with such an eigenvalue and the corresponding blocks of the B_N and C_N matrices form a controllable and observable subsystem; this will be true when the ranks of these corresponding blocks equal the eigenvalue multiplicity. Thus a single point actuator and sensor cannot produce a controllable and observable subsystem (2.3) in more than one space dimension. Hereafter, we will *assume* that the conditions of Theorem 2.1 have been satisfied.

The modal control for the subsystem (2.3) is given by

$$(2.4) \quad \begin{aligned} \underline{f} &= \bar{G}_N \hat{\underline{v}}_N, \\ \dot{\hat{\underline{v}}}_N &= (\bar{A}_N + \bar{B}_N \bar{G}_N - \bar{K}_N \bar{C}_N) \hat{\underline{v}}_N + \bar{K}_N \underline{y}, \quad \hat{\underline{v}}_N(0) = 0 \end{aligned}$$

with control gain matrix \bar{G}_N and estimator gain matrix \bar{K}_N . Let $\underline{e}_N \equiv \hat{\underline{v}}_N - \underline{v}_N$, the estimator error, and (2.3) becomes

$$(2.5) \quad \begin{aligned} \dot{\underline{v}}_N &= (\bar{A}_N + \bar{B}_N \bar{G}_N) \underline{v}_N + \bar{B}_N \bar{G}_N \underline{e}_N, \\ \dot{\underline{e}}_N &= (\bar{A}_N - \bar{K}_N \bar{C}_N) \underline{e}_N, \quad \underline{e}_N(0) = -\underline{v}_N(0). \end{aligned}$$

Since $(\bar{A}_N, \bar{B}_N, \bar{C}_N)$ is controllable and observable, pole allocation or optimal regulator methods can be used to achieve any desired stability of the composite system (2.5). Assume that \bar{G}_N, \bar{K}_N have been chosen so that the spectrum of the composite system

(2.5) is to the left of a vertical line in the complex plane drawn through the point $(-\sigma, 0)$ where $\sigma > 0$. But

$$(2.6) \quad \begin{bmatrix} \underline{v}_N \\ \underline{\hat{v}}_N \end{bmatrix} = \begin{bmatrix} I_N & 0 \\ I_N & I_N \end{bmatrix} \begin{bmatrix} \underline{v}_N \\ \underline{e}_N \end{bmatrix}$$

is a nonsingular transformation and therefore the spectrum of the equivalent composite system

$$(2.7) \quad \begin{bmatrix} \dot{\underline{v}} \\ \dot{\underline{\hat{v}}} \end{bmatrix} = \begin{bmatrix} \bar{A}_N & \bar{B}_N \bar{G}_N \\ \bar{K}_N \bar{C}_N & \bar{A}_N + \bar{B}_N \bar{G}_N - \bar{K}_N \bar{C}_N \end{bmatrix} \begin{bmatrix} \underline{v}_N \\ \underline{\hat{v}}_N \end{bmatrix}$$

also lies to the left of the vertical line through $(-\sigma, 0)$. Therefore, using $\underline{\hat{v}}_N(0) = 0$, we obtain

$$(2.8) \quad \underline{v}_N^T \underline{v}_N + \underline{\hat{v}}_N^T \underline{\hat{v}}_N \leq K^2 e^{-2\sigma t} \underline{v}_N^T(0) \underline{v}_N(0).$$

with $K^2 \equiv \rho(D_N^* D_N) \rho([D_N D_N^*]^{-1}) \geq 1$ where $\rho(\cdot)$ denotes the spectral radius of a matrix and D_N is the nonsingular matrix that transforms the composite system (2.7) into Jordan canonical form.

THEOREM 2.2. *The following inequalities hold for $v_0 \in H_1$:*

- 1) $\|Pv\|_E \leq K e^{-\sigma t} \sqrt{\delta/\mu} \|Pv_0\|_E$,
- 2) $|\underline{\hat{v}}_N| \leq K e^{-\sigma t} (1/\sqrt{\mu}) \|Pv_0\|_E$

where $\delta \equiv \max_{1 \leq k \leq N} (1, \lambda_k)$ and $\mu \equiv \min_{1 \leq k \leq N} (1, \lambda_k)$ and K is defined in (2.8).

Proof. From (2.1) and (1.10), we have $\|Pv\|_E^2 = \sum_{k=1}^N (\lambda_k u_k^2 + \dot{u}_k^2)$. By definition of \underline{v}_N in (2.2), $\mu \underline{v}_N^T \underline{v}_N \leq \|Pv\|_E^2 \leq \delta \underline{v}_N^T \underline{v}_N$. Using (2.8), we obtain parts 1) and 2). \square

Note that, if no estimator is present (i.e., $\underline{e}_N = 0$), then K can be defined by

$$K^2 \equiv \rho(D_N^* D_N) \rho([D_N D_N^*]^{-1})$$

where D_N transforms $\bar{A}_N + \bar{B}_N \bar{G}_N$ to Jordan canonical form.

3. Energy estimates with control spillover. In this section, we estimate the total energy in the solution of v of the system (1.9) with the modal control $\underline{f}(t)$ in (2.4).

THEOREM 3.1. *Let $v_0 \in H_1$; then*

- 1) $Qv = U(t)Qv_0 + \int_0^t U(t-\tau)Q\bar{B}\underline{f}(\tau) d\tau$,
- 2) $\|Qv\|_E \leq \|Qv_0\|_E + (K\beta/\sigma)(1 - e^{-\sigma t})\|Pv_0\|_E$

where K, σ are as in Theorem 2.2 and the control spillover coefficient β is defined by

$$\beta^2 \equiv \frac{\rho(\bar{G}_N^T \bar{G}_N)}{\mu} \sum_{i=1}^M \|Q_0 b_i\|^2, \quad \mu \equiv \min_{1 \leq k \leq N} (1, \lambda_k).$$

Proof. By (2.4), $\underline{f} = \bar{G}_N \underline{\hat{v}}_N$ and $\underline{\hat{v}}_N$ is a linear function of the sensor data \underline{y} which is continuous in t from (1.3). Therefore, \underline{f} is continuously differentiable for $t \geq 0$ and we can apply (1.13) to (1.16) and obtain 1). Using the fact that $U(t)$ is a unitary group, we find that, from 1),

$$(3.1) \quad \|Qv\|_E \leq \|Qv_0\|_E + \int_0^t \|Q\bar{B}\underline{f}\|_E d\tau.$$

But $Q\bar{B} = \begin{bmatrix} 0 \\ Q_0 B \end{bmatrix}$, by definition of \bar{B} in (1.9), and thus

$$\begin{aligned} \|Q\bar{B}f\|_{\mathbb{E}}^2 &= \|Q_0 B f\|^2 = \sum_{k=N+1}^{\infty} \left(\sum_{i=1}^M (b_i, \phi_k) f_i \right)^2 \\ &\leq \sum_{k=N+1}^{\infty} \left(\sum_{i=1}^M (b_i, \phi_k)^2 \sum_{i=1}^M f_i^2 \right) \\ &= \sum_{i=1}^M \left(\sum_{k=N+1}^{\infty} (b_i, \phi_k)^2 \right) \sum_{i=1}^M f_i^2 \\ &= \left(\sum_{i=1}^M \|Q_0 b_i\|^2 \right) f^T f \end{aligned}$$

by the definition of Bf in (1.2) and the Cauchy inequality

$$\sum_{i=1}^M g_i f_i \leq \left(\sum_{i=1}^M g_i^2 \right)^{1/2} \left(\sum_{i=1}^M f_i^2 \right)^{1/2}.$$

However, $f^T f \leq \rho(\bar{G}_N^T \bar{G}_N) \hat{v}_N^T \hat{v}_N \leq \rho(\bar{G}_N^T \bar{G}_N) (K^2 e^{-2\sigma t} / \mu) \|Pv_0\|_{\mathbb{E}}^2$ from Theorem 2.2 part 2). Substitution of these inequalities in (3.1) gives 2). \square

THEOREM 3.2. Let K, β, σ be as in Theorem 3.1.

1) Let $v_0 \in H_1$. Then, for $K_1 \equiv K(\delta/\mu)^{1/2}$ and $\beta \neq 0$,

$$\|v(t)\|_{\mathbb{E}} \leq \left[1 + \left(\frac{K_1 \beta}{\sigma} \right)^2 \right]^{1/2} \|v_0\|_{\mathbb{E}}$$

for t sufficiently large, i.e.,

$$t \geq T_0 \equiv \max \left(0, \frac{1}{\sigma} \ln \frac{1 + (\beta/\sigma)^2}{2(\beta/\sigma)^2} \right).$$

When $\beta = 0$, the modal control f produces a contraction for $t \geq (1/\sigma) \ln K_1$.

2) Let $v_0 \in PH_1$ then, for $\beta \neq 0$,

$$\|v(t)\|_{\mathbb{E}} \leq \left(\frac{K_1 \beta}{\sigma} \right) \|v_0\|_{\mathbb{E}}$$

for $t \geq T_0$ and, in particular, for $K_1 \beta / \sigma \leq 1$ the modal control f produces a contraction on the subspace PH_1 for $t \geq T_0$.

Proof. For part 1) use Theorem 2.2 part 1), $\|Pv\|_{\mathbb{E}} \leq K_1 e^{-\sigma t} \|Pv_0\|_{\mathbb{E}}$, $K_1 \geq 1$, and use (1.18), $\|v\|_{\mathbb{E}}^2 = \|Pv\|_{\mathbb{E}}^2 + \|Qv\|_{\mathbb{E}}^2$. If $\beta = 0$ then from Theorem 3.1 part 2) $\|Qv\|_{\mathbb{E}} \leq \|Qv_0\|_{\mathbb{E}}$ and therefore $\|v\|_{\mathbb{E}}^2 \leq K_1^2 e^{-2\sigma t} \|Pv_0\|_{\mathbb{E}}^2 + \|Qv_0\|_{\mathbb{E}}^2 \leq \|Pv_0\|_{\mathbb{E}}^2 + \|Qv_0\|_{\mathbb{E}}^2 = \|v_0\|_{\mathbb{E}}^2$ for $t \geq (1/\sigma) \ln K_1 > 0$.

If $\beta \neq 0$, take $\varepsilon \equiv (\beta K / \sigma)(1 - e^{-\sigma t})$ for $t > 0$. Then, from Theorem 3.1 part 2),

$$(3.2) \quad \|v\|_{\mathbb{E}}^2 \leq K_1^2 e^{-2\sigma t} \|Pv_0\|_{\mathbb{E}}^2 + (\|Qv_0\|_{\mathbb{E}} + \varepsilon \|Pv_0\|_{\mathbb{E}})^2.$$

However,

$$\begin{aligned} (a+b)^2 &\leq a^2 + b^2 + 2ab \\ &= a^2 + b^2 + 2(\varepsilon a) \left(\frac{b}{\varepsilon} \right) \quad (\text{cont'd.}) \end{aligned}$$

$$\begin{aligned}
&\leq a^2 + b^2 + (\varepsilon a)^2 + \left(\frac{b}{\varepsilon}\right)^2 \\
&= (1 + \varepsilon^2)a^2 + \left(1 + \frac{1}{\varepsilon^2}\right)b^2 = (1 + \varepsilon^2)\left(a^2 + \frac{b^2}{\varepsilon^2}\right)
\end{aligned}$$

since $\varepsilon \neq 0$. Take $a \equiv \|Qv_0\|_{\mathbb{E}}$ and $b \equiv \varepsilon \|Pv_0\|_{\mathbb{E}}$ and (3.2) becomes

$$\begin{aligned}
\|v\|_{\mathbb{E}}^2 &\leq K_1^2 e^{-2\sigma t} \|Pv_0\|_{\mathbb{E}}^2 + (1 + \varepsilon^2)(\|Qv_0\|_{\mathbb{E}}^2 + \|Pv_0\|_{\mathbb{E}}^2) \\
&\leq [K_1^2 M(t) + 1] \|v_0\|_{\mathbb{E}}^2, \quad \text{where } M(t) \equiv e^{-2\sigma t} + \varepsilon^2,
\end{aligned}$$

because $\|Pv_0\|_{\mathbb{E}} \leq \|v_0\|_{\mathbb{E}}$ and $\mu/\delta \leq 1$. Now, let $z \equiv e^{-\sigma t}$ so that $0 \leq z \leq 1$ and consider

$$\begin{aligned}
M(t) - (\beta/\sigma)^2 &= z^2 + (\beta/\sigma)^2(1 - z)^2 - (\beta/\sigma)^2 \\
&= z([1 + (\beta/\sigma)^2]z - 2(\beta/\sigma)^2) \leq 0
\end{aligned}$$

if and only if $0 \leq z \leq \min(1, 2(\beta/\sigma)^2/(1 + (\beta/\sigma)^2))$. Consequently, $M(t) \leq (\beta/\sigma)^2$ if and only if $t \geq T_0$; hence $\|v\|_{\mathbb{E}}^2 \leq [1 + (K_1\beta/\sigma)^2] \|v_0\|_{\mathbb{E}}^2$ as required in 1). For part 2), if $v_0 \in PH_1$ then $Qv_0 = 0$ and, repeating the above procedure for $\beta \neq 0$, we obtain

$$\|v\|_{\mathbb{E}}^2 \leq K_1^2 M(t) \|v_0\|_{\mathbb{E}}^2 \leq \left(\frac{K_1\beta}{\sigma}\right)^2 \|v_0\|_{\mathbb{E}}^2$$

when $t \geq T_0$. \square

In general only “dissipative” perturbations of \bar{A} , the generator of the original contraction group $U(t)$, can lead to new contraction semigroup generators [13, p. 84]. Even with perfect feedback of the infinite dimensional state, it is not always possible to achieve this with (1.2); e.g., see [16, p. 346]. Therefore, it is not surprising that, in general, modal control does not produce a contraction because of control spillover ($\beta \neq 0$) as seen in Theorem 3.2 part 1).

THEOREM 3.3. *Let $v_0 \in H_1$; then*

$$\|v\|_{\mathbb{E}} \leq \|Qv_0\|_{\mathbb{E}} + K_1 h(t) \|Pv_0\|_{\mathbb{E}}$$

where $h(t) \equiv e^{-\sigma t} + (\beta/\sigma)(1 - e^{-\sigma t})$ is decreasing for $\beta/\sigma < 1$ and $\lim_{t \rightarrow \infty} h(t) = \beta/\sigma$ where K_1, β, σ are as in Theorem 3.2.

Proof. Use the triangle inequality, Theorem 2.2 part 1), and Theorem 3.1 part 2) to obtain

$$\begin{aligned}
\|v\|_{\mathbb{E}} &\leq \|Pv\|_{\mathbb{E}} + \|Qv\|_{\mathbb{E}} \leq K_1 e^{-\sigma t} \|Pv_0\|_{\mathbb{E}} + \|Qv_0\|_{\mathbb{E}} \\
&\quad + \frac{K\beta}{\sigma} (1 - e^{-\sigma t}) \|Pv_0\|_{\mathbb{E}} \\
&\leq \|Qv_0\|_{\mathbb{E}} + K_1 h(t) \|Pv_0\|_{\mathbb{E}}
\end{aligned}$$

because $\mu/\delta \leq 1$. Since $h(t) = (1 - \beta/\sigma)e^{-\sigma t} + \beta/\sigma$, it is decreasing for $1 - \beta/\sigma > 0$ and clearly $\lim_{t \rightarrow \infty} h(t) = \beta/\sigma$. \square

Note that using $\|Pv_0\|_{\mathbb{E}} \leq \|v_0\|_{\mathbb{E}}$ and $\|Qv_0\|_{\mathbb{E}} \leq \|v_0\|_{\mathbb{E}}$ in Theorem 3.3 does not give a better estimate than Theorem 3.2 part 1) because, for $\beta/\sigma < 1$,

$$1 + \left(\frac{K_1\beta}{\sigma}\right)^2 < \left(1 + \frac{K_1\beta}{\sigma}\right)^2 < (1 + K_1 h(t))^2.$$

4. Extension of system response time due to control spillover. In this section, we will consider v_0 in the initial set

$$\theta(\alpha_P, \alpha_Q) \equiv \{v_0 \in H_1 \mid \|v_0\|_E^2 \leq E_0, \|Pv_0\|_E^2 \leq \alpha_P E_0, \|Qv_0\|_E^2 \leq \alpha_Q E_0\}$$

where α_P, α_Q are in $[0, 1]$, and we want to establish when (if ever) the modal control f of § 2 will bring the system to a target set $\theta_T \equiv \{v \mid \|v\|_E^2 \leq E_T\}$. We will assume

$$(4.1) \quad \alpha_Q < \frac{E_T}{E_0} < \min(1, \alpha_P + \alpha_Q).$$

Then, without control spillover ($\beta = 0$), from Theorem 3.3,

$$(4.2) \quad \|v\|_E \leq H_0(t) \equiv \|Qv_0\|_E + K_1 e^{-\sigma t} \|Pv_0\|_E$$

and hence v is in the target set θ_T when

$$(4.3) \quad T \equiv \frac{1}{\sigma} \ln \frac{K_1 \alpha_P^{1/2}}{(E_T/E_0)^{1/2} - \alpha_Q^{1/2}}$$

since $H_0(t) \leq H_0(T) = E_T^{1/2}$ for $t \geq T$, where $T > 0$ because, due to (4.1) and $K_1 \geq 1$,

$$0 < \frac{(E_T/E_0)^{1/2} - \alpha_Q^{1/2}}{K_1 \alpha_P^{1/2}} < 1.$$

The time T is the *response time* of the closed-loop system when no spillover is present. Control spillover increases this response time by ΔT so that $v \in \theta_T$ for $t \geq T + \Delta T$ and this extension of response time is given by

THEOREM 4.1. *Let $v_0 \in \theta(\alpha_P, \alpha_Q)$ and assume (4.1). When control spillover is present ($\beta \neq 0$), if $\beta/\sigma < e^{-\sigma T}$, then $v \in \theta_T$ for $t \geq T + \Delta T$ where*

$$\Delta T = \frac{1}{\sigma} \ln \frac{1 - \beta/\sigma}{1 - (\beta/\sigma) e^{\sigma T}}$$

and T is given by (4.3).

Proof. From Theorem 3.3,

$$\|v\|_E \leq H_\beta(t) \equiv H_0(t) + \frac{K_1 \beta}{\sigma} (1 - e^{-\sigma t}) \|Pv_0\|_E.$$

Thus

$$(4.4) \quad H_\beta(T + \Delta T) = H_0(T) + K_1 \|Pv_0\|_E (h(T + \Delta T) - e^{-\sigma T}).$$

If $\beta/\sigma < e^{-\sigma T}$, then $\beta/\sigma < 1$ and $h(t)$ decreasing; consequently,

$$H_\beta(t) = \|Qv_0\|_E + K_1 h(t) \|Pv_0\|_E$$

is also decreasing and

$$\|v\|_E \leq H_\beta(t) \leq H_\beta(T + \Delta T) \quad \text{for } t \geq T + \Delta T.$$

But, if

$$\Delta T \equiv \frac{1}{\sigma} \ln \frac{1 - \beta/\sigma}{1 - (\beta/\sigma) e^{\sigma T}}$$

where

$$0 < \frac{1 - (\beta/\sigma) e^{\sigma T}}{1 - \beta/\sigma} \leq 1$$

then $h(T + \Delta T) = (1 - \beta/\sigma) e^{-\sigma(T + \Delta T)} + \beta/\sigma = e^{-\sigma T}$. Also ΔT is unique because $h(t)$ is decreasing. Therefore $H_\beta(T + \Delta T) = H_0(T)$ and for $t \geq T + \Delta T$

$$\|v\|_E < H_\beta(T + \Delta T) = H_0(T) \leq \sqrt{E_T}$$

by (4.3); hence $v \in \theta_T$ for $t \geq T + \Delta T$. \square

THEOREM 4.2. Let $v_0 \in \theta(1, 0)$ and assume $0 < E_T/E_0 < 1$ and $(K_1\beta/\sigma)^2 \leq E_T/E_0$. When no spillover is present ($\beta = 0$), $v \in \theta_T$ for $t \geq T \equiv 1/\sigma \ln(K_1/(E_T/E_0)^{1/2})$. When spillover is present ($\beta \neq 0$), $v \in \theta_T$ for $t \geq \min(T + \Delta T, T_0)$ where

$$\Delta T = \frac{1}{\sigma} \ln \frac{1 - \beta/\sigma}{1 - (K_1\beta/\sigma)(E_0/E_T)^{1/2}}$$

and

$$T_0 = \ln \frac{1 + (\beta/\sigma)^2}{2(\beta/\sigma)^2}.$$

Proof. Since $v_0 \in \theta(1, 0)$, we have $v_0 \in PH_1$. Applying Theorem 3.2 part 2) when $\beta \neq 0$, gives

$$\|v\|_E^2 \leq \left(\frac{K_1\beta}{\sigma}\right)^2 \|v_0\|_E^2 \leq \frac{E_T}{E_0}(E_0) = E_T \quad \text{for } t \geq T_0 = \ln \frac{1 + (\beta/\sigma)^2}{2(\beta/\sigma)^2}$$

because $(\beta/\sigma)^2 \leq (E_T/E_0)(1/K_1^2) < 1$. Applying Theorem 4.1, gives the remaining results. \square

5. Application: A simply supported beam. In this section, we consider control of the transverse vibrations of a homogeneous Euler–Bernoulli beam:

$$(5.1) \quad u_{tt} + u_{xxxx} = Bf$$

where the beam constants have all been normalized to unity. The operator

$$(5.2) \quad Au = u_{xxxx}$$

is defined on $\mathcal{D}(A)$ containing all sufficiently differentiable functions on $[0, 1]$ which satisfy the boundary conditions for simple (pinned) support:

$$(5.3) \quad \begin{aligned} u(0, t) &= u(1, t) = 0, \\ u_{xx}(0, t) &= u_{xx}(1, t) = 0. \end{aligned}$$

The operator A has compact resolvent and the eigenvalues are

$$(5.4) \quad \lambda_k = (k\pi)^4$$

with corresponding eigenvectors

$$(5.5) \quad \phi_k(x) = \sqrt{2} \sin k\pi x.$$

We want to develop a modal control that will stabilize the first N modes of this system. We will assume

- 1) no observation spillover,
- 2) no estimator error (this is for convenience and is not essential),

- 3) $v_0 \in PH_1$ (i.e., none of the residual modes is initially excited)
 4) $b(x)$ is a single point actuator, e.g.,

$$b(x) = \begin{cases} \frac{1}{2\varepsilon}, & x_1 - \varepsilon \leq x \leq x_1 + \varepsilon, \\ 0 & \text{elsewhere,} \end{cases}$$

- 5) the N mode system is controllable (i.e., $(b, \phi_k) \neq 0$ for $1 \leq k \leq N$ by Theorem 2.1)

If we use the definitions of § 2, the N mode system is

$$(5.6) \quad \begin{aligned} \dot{v}_N &= \bar{A}_N v_N + \bar{B}_N f, \\ \underline{f} &= \bar{G}_N v_N \end{aligned}$$

where \underline{f} is a scalar control. Define

$$J_N \equiv \int_0^\infty [\underline{v}_N^T Q_N \underline{v}_N + R_N \underline{f}^2] dt$$

where

$$Q_N \equiv \begin{bmatrix} A_N & 0 \\ 0 & I_N \end{bmatrix} \quad \text{and} \quad R_N > 0.$$

Find the optimal steady state regulator for the system $(\bar{A}_N + \sigma \bar{I}_N, \bar{B}_N)$ where $\sigma > 0$. This entails solving a steady state Riccati equation to obtain the appropriate \bar{G}_N . Then the spectrum of $\bar{A}_N + \bar{B}_N \bar{G}_N$ will be to the left of the vertical line passing through $(-\sigma, 0)$ and thus guarantees K_1 such that

$$(5.7) \quad \|Pv\|_E \leq K_1 e^{-\sigma t} \|Pv_0\|_E$$

where $\|Pv\|_E^2 = \underline{v}_N^T Q_N \underline{v}_N$.

From Theorem 3.1, the spillover coefficient is given by

$$(5.8) \quad \beta^2 = \rho(\bar{G}_N^T \bar{G}_N) \|Q_0 b\|^2.$$

Therefore, from Theorem 4.2, given $\|v_0\|_E^2 \leq E_0$ and $0 < E_T/E_0 < 1$, if

$$(5.9) \quad \left(\frac{K_1 \beta}{\sigma} \right)^2 = \frac{K_1^2 \rho(\bar{G}_N^T \bar{G}_N)}{\sigma^2} \|Q_0 b\|^2 < \frac{E_T}{E_0}$$

then $v \in \theta_T$ for t sufficiently large. Since $\|Q_0 b\|^2 = \sum_{k=N+1}^\infty (b, \phi_k)^2 \rightarrow 0$ as $N \rightarrow \infty$ because $b \in L^2(\Omega)$, we will assume that N is large enough that (5.9) is satisfied. This assumption may not always be satisfied since K_1 and $\rho(\bar{G}_N^T \bar{G}_N)$ depend on N as well. However, it can be satisfied in the special case

$$(5.10) \quad K_1^2 \rho(\bar{G}_N^T \bar{G}_N) \leq \sigma^2.$$

Therefore, N modes of the Euler-Bernoulli beam can be controlled with a single point actuator if it is properly located. The effect of control spillover into the residual modes will be to extend the response time for an initial disturbance $v_0 \in \theta(1, 0)$ to be brought to the desired target set θ_T . But it will eventually arrive there, if the a priori bound (5.10) is satisfied and if the number of controlled modes is sufficiently large. If these conditions are not satisfied, no conclusions can be drawn from results about the effects of control spillover.

The situation is better when a distributed actuator is used instead of a point actuator. In this case, the distributed actuator influence function $b(x)$ may be well-approximated by a few modes $\phi_k(x)$. Consequently, $\|Q_0 b\|^2$ may be extremely small when only a few modes are controlled. Then it would not be difficult for the system to satisfy (5.9) and the spillover results would apply. In particular, if

$$(5.11) \quad b(x) = \sum_{k=1}^N (b, \phi_k) \phi_k$$

then the spillover coefficient β is zero if N modes are controlled. Hence, the closed-loop system is a contraction by Theorem 3.2 and by Theorem 4.1 it takes T seconds for an initial disturbance $v_0 \in (1, 0)$ to be brought to the target set θ_T where

$$T = \frac{1}{\sigma} \ln K_1 \left(\frac{E_0}{E_T} \right)^{1/2}.$$

6. Conclusions. For the class of flexible systems described by (1.1), (1.2), (1.3) we have obtained stable feedback control of N modes using standard state variable methods. We have examined the effect of control spillover from this controller into the residual modes of the system when no observation spillover is present. The instabilities present when both control and observation spillover occur are shown in [1].

In Theorem 3.1, we have shown that control spillover can increase the energy norm of the residual modes by as much as $(K\beta/\sigma)(E_0^c)^{1/2}$ where E_0^c is the initial energy in the controlled modes and β is the spillover coefficient which depends on the L^2 projection of the actuator influence functions onto the residual modes. Under special circumstances, the closed-loop system eventually becomes dissipative but, in general, from Theorem 3.2, the energy of the total system state is eventually less than $1 + (K_1\beta/\sigma)^2$ of the initial state energy.

The closed-loop system response time T to bring an initial disturbance to a prescribed energy target set is obtained in (4.3) when no control spillover is present. When the system meets the a priori bound of Theorem 4.1 or Theorem 4.2, the effect of control spillover is only to increase the response time by ΔT which can be calculated in advance. This result makes it possible to judge the performance of a modal controller on a flexible system if the system satisfies the a priori bound on control spillover. It also issues a warning to users of modal control in systems with excessive control spillover.

These results are illustrated with a simply supported Euler-Bernoulli beam. Distributed actuators can be made to perform very well under the hypotheses of the theory presented here. On the other hand, the bounds obtained for control spillover (in particular, Theorem 2.2) may be too conservative for point actuators and may place an unnecessarily stringent spillover requirement on systems using them; for example, the bound in (5.10) for the beam may be quite difficult to satisfy. In particular applications these bounds can no doubt be substantially reduced.

Acknowledgment. I would like to thank Patricia Balas for many helpful conversations on this work.

REFERENCES

- [1] M. BALAS, *Active control of flexible systems*, AIAA Symp. Dynamics and Control of Large Flexible Spacecraft, Blacksburg, VA, June 13-15, 1977.
- [2] J. F. CREEDON, *Control of the optical surface of a thin deformable primary mirror with application to an orbiting astronomical observatory*, IFAC 3rd Symp. on Auto. Control in Space, Toulouse, France, March 2-6, 1970.

- [3] H. FATTORINI, *On complete controllability of linear systems*, J. Differential Equations, 3 (1967), pp. 391–402.
- [4] C. HARRIS AND J. MILES, *Adaptive control of nonlinear flexible spacecraft*, Proc. IFAC 6th World Conf., Boston, MA, August 24–30, 1975.
- [5] T. JOHNSON, *The aerodynamic surface location problem in optimal control of flexible aircraft*, MIT-ESL Rep. ESL-R-387, Mass. Inst. of Tech., Cambridge, MA, June 1969.
- [6] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
- [7] R. KLEIN AND R. HUGHES, *The distributed parameter control of torsional bending in seagoing ships*, Joint Automatic Control Conf., 1971.
- [8] M. KOEHNE, *Optimal feedback control of flexible mechanical systems*, IFAC Symp. Distributed Parameter Control Systems, Banff, Canada, 1971.
- [9] H. KWAKERNAAK AND R. SIVAN, *Linear Optimal Control Systems*, Wiley-Interscience, New York, 1972.
- [10] V. LARSON AND P. LIKINS, *An application of modern control theory to an elastic spacecraft*, ESA Symp. on Dynamics and Control of Non-Rigid Space Vehicles, Frascati, Italy, May 1976.
- [11] V. LARSON, P. LIKINS AND E. MARSH, *Optimal estimation and attitude control of a solar electric propulsion spacecraft*, IEEE Trans. Aerospace and Electron. Systems, AES-13 (1977), pp. 35–37.
- [12] L. MEIROVITCH, H. VANLANDINGHAM AND H. ÖZ, *Control of Spinning Flexible Spacecraft by Modal Synthesis*, Int. Aero. Fed. 27th Congress, Anaheim, CA, October 10–16, 1976.
- [13] A. PAZY, *Semigroups of linear operators and applications to partial differential equations*, Lecture Notes #10, Univ. of Maryland, College Park, Jan. 1974.
- [14] R. QUARTARARO, *Mode control—a simple concept for controlling flexible spacecraft*, Adv. Tech. Lab. Prog. for Large Space Structures, I, App. B, Rockwell International, Anaheim, CA, SD-76-SA-0210, Nov. 1976.
- [15] D. RUSSELL, *Control theory of hyperbolic equations related to certain questions in harmonic analysis and spectral theory*, J. Math. Anal. Appl., 40 (1972), pp. 336–368.
- [16] ———, *Decay rates for weakly damped systems in Hilbert space obtained with control-theoretic methods*, J. Differential Equations, 19 (1975), pp. 344–370.
- [17] ———, *Linear stabilization of the linear oscillator in Hilbert space*, J. Math. Anal. Appl., 25 (1969), pp. 663–675.
- [18] M. SANDFORD, I. ABEL AND D. GRAY, *Development and demonstration of a flutter-suppression system using active controls*, NASA TR-R-450, National Technical Information Service, Washington, D.C., Dec. 1975.
- [19] J. SIMON AND S. MITTER, *A theory of modal control*, Information and Control, 13 (1968), pp. 216–353.
- [20] R. SKELTON AND P. LIKINS, *Techniques of modelling and model error compensation in linear regulator problems*, Advances in Control and Dynamic Systems, 14, C. Leondes, ed., Academic Press, New York, 1977.
- [21] M. SLEMROD, *The linear stabilization problem in Hilbert space*, J. Functional Analysis, 11 (1972), pp. 334–345.
- [22] ———, *A note on complete controllability and stabilizability for linear control systems in Hilbert space*, this Journal, 12 (1974), pp. 500–508.
- [23] F. TREVES, *Basic Linear P.D.E.*, Academic Press, New York, 1975.
- [24] R. TRIGGIANI, *On the stabilizability problem in Banach space*, J. Math. Anal. Appl., 52 (1975), pp. 383–403.
- [25] K. TSUJIOKA, *Remarks on controllability of second order evolution equations in Hilbert spaces*, this Journal, 8 (1970), pp. 90–99.

ON THE STABILITY OF THE EXISTENCE OF SINGULAR CONTROLS UNDER PERTURBATION OF THE CONTROL SYSTEM*

DEBORAH REBHUN†

Abstract. Here we discuss generic stability of certain properties under the perturbation of control systems. The systems under consideration are nonlinear systems with scalar control appearing linearly. We find that our results vary with dimension of the Euclidean space or manifold on which the control system is defined.

I. Introduction. If we consider the set of linear time optimal control systems on \mathbb{R}^n , an open dense subset of such systems which are controllable is known. For the controllable systems, it is generically true that optimal controls are nonsingular and piecewise constant with at most finitely many discontinuities. See [2], [3], [8].

Here we consider nonlinear systems with scalar control in the compact interval $[0, 1]$. For these, and even for more general nonlinear systems, controllability is a generic property. See [9], [11]. The question of the existence of piecewise constant and bang-bang optimal controls or of the existence of optimal singular controls is a harder one.

When a control system of the type we are interested in is defined on a two dimensional manifold, we investigate whether it is generically true that optimal controls are piecewise constant, bang-bang and nonsingular. For manifolds of dimension greater than two, we investigate the existence of strictly singular controls stable under perturbation of the control system. We define strictly singular in § II. The next two theorems and the conjecture that follows summarize our conclusions.

I.1. THEOREM. *For any $k \geq 2$, let \mathcal{A} be the set of C^∞ control systems in the Whitney C^k topology on a two dimensional manifold M . (See [4] for a definition of the topology.)*

(1) For $a \in \mathcal{A}$, subject to the hypothesis that a certain rank condition is satisfied, we show the existence of a function $q_a: M \rightarrow \mathbb{R}$ such that if $q_a(x) \neq 0$ at some $x \in M$, then in a sufficiently small neighborhood of x optimal controls are bang-bang, piecewise constant and involve at most one switch.

(2) There is an open dense subset \mathcal{O} of \mathcal{A} such that if $a \in \mathcal{A}$ then the rank condition mentioned in (1) is satisfied everywhere with the possible exception of a closed embedded curve depending on a . The set $q_a^{-1}(0)$ is either empty or contained in a closed, embedded curve in M . In particular, if $a \in \mathcal{O}$, a strictly singular control cannot be optimal unless its trajectory lies in one of those two curves.

(3) If $a \in \mathcal{O}$, no control other than a bang-bang control can even satisfy the maximum principle, a necessary condition for optimality except along one of the curves mentioned in (2).

(4) We show that it is possible for strictly singular optimal controls for a to exist along $q_a^{-1}(0)$ and that this property is stable under perturbation. That is, if $a' \in \mathcal{A}$ is sufficiently close to a in a C^3 sense, then singular optimal controls for a' exist along $q_{a'}^{-1}(0)$ which is nonempty.

* Received by the editors May 13, 1976, and in final revised form May 24, 1977.

† Department of Mathematics, Vassar College, Poughneepsie, New York 12601. This research was completed while the author was visiting Massachusetts Institute of Technology under the support of a grant from the Radcliffe Institute, Cambridge, Massachusetts.

I.2. THEOREM. *Let M be a three dimensional manifold and for any $k \geq 3$, \mathcal{B} be the set of C^∞ optimal control systems on M in the Whitney C^k topology.*

(1) *Given $b \in \mathcal{B}$, subject to the hypothesis that a certain rank condition is satisfied, strictly singular controls of first order can exist on an open subset of M . Under C^3 small perturbations of b , such strictly singular controls will also exist for the perturbed control system. The singular control in question is unique when it exists. Its existence and value as a feedback control is completely determined by a certain Lie algebra determined by the control system.*

(2) *There is an open dense subset of \mathcal{U} of \mathcal{B} such that if $b \in \mathcal{U}$, singular controls of second order or higher cannot exist except at the points of a (possibly empty) pair of hypersurfaces that depend on b . Off the hypersurfaces, singular controls are completely determined by the results of (1).*

We remark that the main result of Theorem I.2 is that when $\dim M = 3$, we cannot guarantee as we did in dimension two that non bang-bang controls satisfying the maximum principle cannot exist except on a hypersurface. We also remark that the assumption that \mathcal{A} and \mathcal{B} are C^∞ control systems is not really necessary. We could have assumed they were both C^3 and come to all the same conclusions.

I.3. Conjecture. When $\dim M \geq 4$, we will use the symbol \mathcal{C} to denote the C^∞ control systems on M in the Whitney C^k topology. Here k will depend on $\dim M$. There is an open dense subset \mathcal{V} of \mathcal{C} such that if $c \in \mathcal{V}$, then there is an open dense subset U_c of M such that for every $x \in U_c$ there are infinitely many controls strictly singular for c steering us through x .

We will give evidence to support the conjecture later.

There are statements made which are not as precise but which seem to be related to I.3, Theorems I.2(2) and I.1(3) in Johnson [5, Appendix]. We believe we have given the first actual proof of these results.

II. Definitions and basic facts. Given a C^∞ paracompact connected manifold M , a control system on M is determined by a pair (X, Y) of C^∞ vector fields on M . For the moment we will take (X, Y) to be autonomous but as we go along we will comment on modifications in the arguments to be made when (X, Y) are nonautonomous.

II.1. DEFINITION. Given a control system on M , a *trajectory* of the control system is a curve γ_u defined by the differential equation

$$(1) \quad \frac{d}{dt} \gamma_u(t) = u(t)X(\gamma_u(t)) + (1 - u(t))Y(\gamma_u(t)) \quad \text{a.e.}$$

here $u: [t, t''] \rightarrow [0, 1]$ is a piecewise differentiable regulated curve. That is right and left hand limits of u and its derivatives exist at each point. We call u a *control*, γ_u its *trajectory* and say that u *steers* $x \in M$ to $x'' \in M$ if $\gamma_u(t) = x$ and $\gamma_u(t'') = x''$. When it is clear what control u we are referring to, we will write γ instead of γ_u .

II.2. DEFINITION. Let $f: M \rightarrow \mathbb{R}$ be a fixed positive C^∞ function. We define a functional on trajectories by

$$(2) \quad J(\gamma_u) = \int_t^{t''} f(\gamma_u(s)) ds.$$

We say that a control u' steering x to x'' is *optimal* if $J(\gamma_{u'}) \leq J(\gamma_u)$ for any other u steering x to x'' . When $f \equiv 1$, we say $\gamma_{u'}$ is *time optimal*.

II.3. *The maximum principle.* It is known that if the control u' is optimal and γ is the associated trajectory then there is a curve

$$t \rightarrow (\lambda(t), \hat{\lambda}(t)) \in (T_{\gamma(t)}^*M) \times \mathbb{R}$$

such that $(\lambda(t), \hat{\lambda}(t))$ satisfies a differential equation adjoint to (1), $(\lambda(t), \hat{\lambda}(t))$ is never zero at any point, $\lambda(t) \leq 0$, such that

$$(3) \quad \begin{aligned} & \lambda(t)[u'(t)X(\gamma(t)) + (1 - u'(t))Y(\gamma(t))] + \hat{\lambda}(t)f(\gamma(t)) \\ &= \max_{u \in [0,1]} \lambda(t)[uX(\gamma(t)) + (1 - u)Y(\gamma(t))] + \hat{\lambda}(t)f(\gamma(t)) = 0. \end{aligned}$$

If $f \equiv 1$, we can take $\hat{\lambda}(t) \equiv -1$. For details see [6], [7].

Note that if $\lambda(t)[X(\gamma(t)) - Y(\gamma(t))] \neq 0$, then $u'(t)$ is determined by (3) and must take the value zero or one.

II.4. DEFINITION. If a control u' satisfies an equation of type (3) with $[\lambda(t)[X(\gamma(t)) - Y(\gamma(t))] = 0$ a.e. on an open interval, then we say u' is *singular* on the given interval. If u' is singular on the given interval and takes values in $(0, 1)$ a.e., then we say u' is *strictly singular*.

III. PROOFS. Let us fix a positive smooth function $f: M \rightarrow \mathbb{R}$. When we say a control steers x to x'' optimally, we mean that this f is used to define $J(\gamma)$. See Definition II.2.

III.1. *Proof of Theorem I.1(1).* Given $(X, Y) \in \mathcal{A}$, the set of C^∞ control systems on M , let $U = \{x \in M: X(x), Y(x) \text{ are linearly independent}\}$. We can define a one form on U by

$$\langle \omega, X(x) \rangle = \langle \omega, Y(x) \rangle = f(x).$$

If γ is a trajectory of the system (X, Y) lying completely in U , then

$$J(\gamma) = \int_{\gamma} \omega.$$

If γ_1, γ_2 are trajectories of (X, Y) such that

- (i) both γ_1 and γ_2 lie in U ,
- (ii) both γ_1 and γ_2 start at x' and end at x'' ,
- (iii) the closed curve δ obtained by following first γ_1 then the reverse of γ_2 is simple and oriented counterclockwise,

then if Δ is the interior of δ by Stokes theorem

$$J(\gamma_1) - J(\gamma_2) = \int_{\delta} \omega = \int_{\Delta} d\omega.$$

Now, since $X(x), Y(x)$ are independent for each $x \in U$, we can express the Lie bracket $[X, Y]$ in a unique way as $[X, Y] = aX + bY$ where a and b are smooth functions on U . Clearly $d\omega = h \, dx \, dy$ where

$$h = d\omega\left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)$$

is a smooth function.

If U_{Δ} is the component of U containing Δ , the orientation of any pair $X(x), Y(x)$ is the same on U_{Δ} . Let us consider the case when this orientation is counterclockwise.

The case when the orientation is clockwise is handled similarly. On U_Δ ,

$$d\omega(X, Y) = hk$$

where $k > 0$ is a smooth function on U_Δ .

Now, $d\omega(X, Y) = X\langle\omega, Y\rangle - Y\langle\omega, X\rangle - \omega([X, Y]) = Xf - Yf - (a + b)f$. Thus

$$\text{A. if } a(x) + b(x) \leq \frac{(Xf)(x) - (Yf)(x)}{f(x)} = g(x) \text{ on } \Delta, \text{ then } J(\gamma_1) \geq J(\gamma_2);$$

$$\text{B. if } a(x) + b(x) \geq g(x) \text{ on } \Delta, \text{ then } J(\gamma_1) \leq J(\gamma_2).$$

Now let $x' \in U$. Choose $\varepsilon > 0$ so small that the map F defined by $(t, \tau) \rightarrow X_\tau Y_t(x')$ is a diffeomorphism of the closed cube $C_\varepsilon = \{(t, \tau) : |t| \leq \varepsilon, |\tau| \leq \varepsilon\}$ onto a compact neighborhood $V \subset U$ of x' . Here X_τ, Y_t are the time τ and t flows of X and Y respectively. The compact set $C_\varepsilon^+ = \{(t, \tau) \in C_\varepsilon : t \geq 0, \tau \geq 0\}$ is mapped by F to a compact subset V^+ of V . It is easy to see that V^+ is precisely the set of points that can be reached from x' by a trajectory that stays within V .

This shows that if $x' \in U$, then bang-bang controls of the form $u(t) = 1$ for $t \in [t', t''']$, $u(t) = 0$ for $t \in [t''', t'']$, henceforth called YX controls, suffice near x' to reach anything that can be reached from x' . Similarly, XY controls suffice. Also the reachable set for small time is locally closed at x .

Now, assume $g(x') < a(x') + b(x')$. We can choose a neighborhood $V \subset U$ of x' so small that everything reachable from x' within V is reachable by an XY control and $g(x) < a(x) + b(x)$ for all $x \in V$. Then if $x'' \in V$ is reachable from x' within V , let γ_2 be the curve from x' to x'' that corresponds to a YX control, and let γ_1 be the curve from another trajectory that goes from x' to x'' within V . The curves γ_1 and γ_2 satisfy (1)(i), (1)(ii) and (1)(iii). We then conclude from A that γ_1 is the optimal trajectory from x' to x'' .

If instead we have $g(x') > a(x') + b(x')$, we show in a similar way that XY controls are optimal near x' .

Let $q_{(X,Y)}(x) = [Xf - Yf - (a + b)f](x)$. Clearly $q_{(X,Y)}$ is defined on U the set where X and Y are independent and if $q_{(X,Y)}(x) \neq 0$, then $g(x) \neq a(x) + b(x)$. Thus if $q_{(X,Y)}(x) \neq 0$, in a neighborhood of x optimal controls are bang-bang and involve at most one switch. This concludes the proof of Theorem I.1(1). We remark that everything could have been modified to account for the situation in which X and Y were time dependent. Before we go on to prove Theorem I.1(2) we will need two lemmas.

III.2. LEMMA. *There is an open dense subset \mathcal{O}' of \mathcal{A} such that if $(X, Y) \in \mathcal{O}'$, then $X(x)$ and $Y(x)$ are linearly independent at all $x \in M$ with the possible exception of a closed embedded hypersurface C that depends on (X, Y) .*

III.3. LEMMA. *Let (\tilde{X}, \tilde{Y}) be a control system on $M \times \mathbb{R}$ defined by*

$$\tilde{X} = \begin{pmatrix} X \\ f \end{pmatrix}, \quad \tilde{Y} = \begin{pmatrix} Y \\ f \end{pmatrix}.$$

There is an open dense subset \mathcal{O}_f of \mathcal{A} such that if $(X, Y) \in \mathcal{O}_f$, then $\tilde{X}(x, t), \tilde{Y}(x, t), [\tilde{X}, \tilde{Y}](x, t)$ are linearly dependent at every $(x, t) \in M \times \mathbb{R}$ unless x is some element of a closed embedded hypersurface C_f in M .

We will need transversality theory to prove these. We will sketch the proof of Lemma III.2. The proof of Lemma III.3 is similar. We also remark that if we take $f \equiv 1$ and the vector fields (X, Y) are nonautonomous, then the results are still true except that the hypersurfaces of M must be replaced by hypersurfaces in $M \times \mathbb{R}$.

Proof of Lemma III.2. Consider the bundle T^2M over M whose fiber T_x^2M over x consists of $(T_xM) \times (T_xM)$. It is clear that a control system can be considered a section of T^2M . We can partition T^2M into three disjoint manifolds R_0, R_1, R_2 as follows: An element of T^2M has the form (X_1, X_2) where $X_1 \in T_xM, X_2 \in T_xM$. We say $(X_1, X_2) \in R_i$ if the space spanned by X_1, X_2 has dimension i in T_xM . Since a pair $(X, Y) \in \mathcal{A}$ determines a section in T^2M , which we will again denote by (X, Y) , we can use standard arguments to show that there is an open dense subset \mathcal{O}' of \mathcal{A} such that if $(X, Y) \in \mathcal{O}'$ the section it determines is transversal to R_0 and R_1 . Note that codimension $R_0 = 4$ and codimension $R_1 = 1$. In particular, if $(X, Y) \in \mathcal{O}'$, then $(X, Y)^{-1}(R_0)$ is empty and $(X, Y)^{-1}(R_1) = (X, Y)^{-1}(R_0 \cup R_1)$ the set of points x at which $X(x)$ and $Y(x)$ are linearly dependent is a closed embedded manifold C of codimension one. See [4], [10] for standard transversality arguments.

III.4. *Proof of Theorem I.1(2).* Let $\mathcal{O} = \mathcal{O}_f \cap \mathcal{O}'$ where \mathcal{O}_f and \mathcal{O}' are the sets defined in the preceding lemmas. If $(X, Y) \in \mathcal{O}$, then with the exception of the points of a (possibly empty) curve C , $X(x)$ and $Y(x)$ are independent everywhere. In particular, $q_{(X,Y)}$ is defined except on C . Let us show that if $q_{(X,Y)}(x) = 0$, then x must be a point of the curve C_f along which \hat{X}, \hat{Y} and $[\hat{X}, \hat{Y}]$ are linearly dependent. If this is so, then Theorem I.1(2) is proven.

Note that

$$\hat{X} = \begin{pmatrix} X \\ f \end{pmatrix}, \quad \hat{Y} = \begin{pmatrix} Y \\ f \end{pmatrix}, \quad \text{and} \quad [\hat{X}, \hat{Y}] = \begin{pmatrix} [X, Y] \\ Xf - Yf \end{pmatrix}$$

are linearly independent at $x \in M$ if and only if

$$\hat{Y} = \begin{pmatrix} Y \\ f \end{pmatrix}, \quad (\hat{X} - \hat{Y}) = \begin{pmatrix} X - Y \\ 0 \end{pmatrix}, \quad \text{and} \quad [\hat{X}, \hat{Y}] - c\hat{Y} = \begin{pmatrix} [X, Y] - cY \\ 0 \end{pmatrix}$$

are linearly independent with $c(x) = (1/f(x))(Xf - Yf)(x)$.

In particular, note that since $f(x) \neq 0$, either of these is equivalent to the linear independence of $(X - Y)(x)$ and $[X, Y](x) - c(x)Y(x)$. If $q_{(X,Y)}(x) = 0$, then $c(x) = a(x) + b(x)$. As a result,

$$\begin{aligned} [X, Y](x) - c(x)Y(x) &= a(x)X(x) + b(x)Y(x) - a(x)Y(x) - b(x)Y(x) \\ &= a(x)(X - Y)(x), \end{aligned}$$

contradicting linear independence. If $q_{(X,Y)}(x) = 0$, then $x \in C_f$. As a result, if $x \notin (C \cup C_f)$, then optimal controls steering x to nearby points are bang-bang and involve at most one switch.

III.5. *Proof of Theorem I.1(3).* If u is a singular control with associated trajectory γ , then along γ we have

$$(2) \quad \lambda Y + \hat{\lambda} f \equiv 0$$

$$(3) \quad \lambda(X - Y) \equiv 0. \quad (\text{See II.3.})$$

If we differentiate this last equation we find

$$(4) \quad \frac{d}{dt} \lambda(X - Y) = -\lambda[Y, X - Y] - \hat{\lambda}(Yf - Xf) = 0.$$

See [6]. If $\gamma(t) \notin C \cup C_f$, then $[Y, X - Y] = [Y, X] = -aX - bY$ at $\gamma(t)$. Substituting back into (4) and using (2) and (3), we find

$$-a\hat{\lambda}f - b\hat{\lambda}f - \hat{\lambda}(Yf - Xf) = 0.$$

In particular, if $\hat{\lambda}(t) \neq 0$, $[Xf - Yf - (a+b)f](\gamma(t)) = 0$ which contradicts the hypothesis that $\gamma(t) \notin C_f$. If $\hat{\lambda}(t) = 0$, then $\lambda(t) \neq 0$. Equations (2) and (3) then imply that $\lambda(t)X(\gamma(t)) = \lambda(t)Y(\gamma(t)) = 0$. This is impossible unless $X(\gamma(t))$ and $Y(\gamma(t))$ are linearly dependent which contradicts the hypothesis that $\gamma(t) \notin C$. We conclude that any singular control must steer us in $C \cup C_f$ for all time.

Again, we can prove this same result for nonautonomous (X, Y) if $f \equiv 1$.

III.6. *Proof of Theorem I.1(4)*. Let us fix (X, Y) and abbreviate $q_{(X,Y)}$ to q . Assume $x' \in M$ is a fixed point such that the hypotheses

$$(5) \quad q(x') = 0,$$

$$(6) \quad (Xq)(x') > 0,$$

$$(7) \quad (Yq)(x') < 0$$

hold. By (6), $dq \neq 0$ since $\langle dq, X \rangle(x') \neq 0$. As a result the equation $q(x) = 0$ defines a hypersurface H through x' in a neighborhood of x' . Thus there is a unique $\alpha(x)$, $0 < \alpha(x) < 1$, such that $\alpha(x)(Yq)(x) + [1 - \alpha(x)](Xq)(x) = 0$. Let $Z(x) = \alpha(x)X(x) + (1 - \alpha(x))Y(x)$. Then $Zq = 0$ so Z is tangent to H . Let $t \rightarrow \gamma(t)$ be the integral curve of Z through x' . Then

$$\dot{\gamma} = u(t)X(\gamma(t)) + (1 - u(t))Y(\gamma(t))$$

where $u(t) = \alpha(\gamma(t))$. That is γ is a trajectory of the system (X, Y) lying in H .

If we can show that u is optimal, then u must be a strictly singular control. It is also clear that if (X', Y') is C^3 close enough to (X, Y) , then X', Y' and $q_{X', Y'}$ will satisfy analogues of hypotheses (5), (6), (7) at some point near x and will also have a strictly singular optimal control.

Let us show that u is optimal. Near x' , the hypersurface H divides the plane into two parts, the side H_x towards which X points and the side H_Y towards which Y points. Let a trajectory $\tilde{\gamma}$ of (X, Y) go from x' to x'' with $x'' \in H$ and assume $\tilde{\gamma}$ is contained in H_x . Then (1)(i), (1)(ii), and (1)(iii) hold with $\gamma = \gamma_1$ and $\tilde{\gamma} = \gamma_2$. Since $q = 0$ on H and $Xq > 0$ near x' , it follows that $q > 0$ on H_x . Hence by A, $J(\tilde{\gamma}) \geq J(\gamma)$. A similar argument shows $J(\tilde{\gamma}) \geq J(\gamma)$ if $\tilde{\gamma}$ is contained in H_Y .

Finally, if $\tilde{\gamma}$ is an arbitrary trajectory from x' to $x'' \in H$, the preceding argument shows that every piece of $\tilde{\gamma}$ between two consecutive intersections with H can be replaced by a piece of γ in such a way that J will be decreased. This completes the proof of Theorem I.1.

Before going on to prove Theorem I.2, let us recall that we now assume $\dim M = 3$. We will assume $f \equiv 1$ and prove everything under that assumption. It is not fundamentally different from the general case but the computational details are much simpler and this special case still demonstrates the basic idea of the theorem: in dimensions above two we cannot guarantee that in the generic case singular controls will occur only along a hypersurface.

III.7. *Proof of Theorem I.2(1)*. Again, if u is a singular control, we can use the maximum principle to show that along $\gamma(t)$ the associated trajectory (see II.3)

$$(8) \quad \lambda Y \equiv 1,$$

$$(9) \quad \lambda(X - Y) \equiv 0,$$

$$(10) \quad \frac{d}{dt} \lambda(X - Y) = -\lambda[Y, X - Y] = -\lambda[Y, X] = 0,$$

$$(11) \quad \frac{d^2}{dt^2} \lambda (X - Y) = \lambda [Y[Y, X]] + u \lambda [X - Y[Y, X]] = 0.$$

If we know that

$$(12) \quad (X - Y), [Y, X], \text{ and } [Y[Y, X]] \text{ are linearly independent}$$

at $x \in M$ and if $x = \gamma(t)$, then in some neighborhood of $\gamma(t)$ we can find smooth functions c_1, c_2, c_3 on M such that

$$[X - Y[Y, X]] = c_1(X - Y) + c_2[Y, X] + c_3[Y[Y, X]].$$

Substituting this last expression into the equation (11) and using (9) and (10) we find

$$(13) \quad \lambda [Y[Y, X]] + u c_3 \lambda [Y[Y, X]] = 0.$$

By hypothesis (12) since $\lambda(t) \neq 0$, we see that $\lambda(t)$ cannot take all three basis vectors of $T_{\gamma(t)}M$ to zero. By (9) and (10), it takes $(X - Y)(\gamma(t))$ and $[Y, X](\gamma(t))$ to zero. So, $\lambda(t)[Y[Y, X]](\gamma(t))$ is nonzero and by (13) $1 + u c_3 = 0$. In particular, the only possible choice for $u(t)$ is $-1/c_3(\gamma(t))$. If $-1/c_3(\gamma(t)) \notin [0, 1]$, then we see that $u(t)$ cannot be singular. Also note that $-1/c_3(x)$ is a singular feedback control determined by the Lie algebra generated by the vector fields X and Y . Let us show

III.8. LEMMA. *If $-1/c_3(x') \in (0, 1)$, then the feedback control $u(t) = -1/c_3(\gamma(t))$ is a strictly singular control steering us through x' provided that*

$$(14) \quad Y(x'), (X - Y)(x') \text{ and } [Y, X](x') \text{ are linearly independent.}$$

Proof. Note that if hypothesis (21) holds at x' , then it holds for all x in a neighborhood of x' . Let $\text{ad}^0 Y(x) = X$ and let $\text{ad}^i Y(x) = [Y, \text{ad}^{i-1} Y(x)]$ for $i \geq 1$. With this notation, along any trajectory $\gamma_u(t)$ of the system (X, Y) for any solution λ of the adjoint equation,

$$\frac{d}{dt} \lambda \text{ad}^2 Y(X) = -\lambda \text{ad}^3 Y(X) - u \lambda \text{ad}^1(X - Y) \text{ad}^2 Y(X).$$

If we express $\text{ad}^3 Y(X)$ and $\text{ad}^1(X - Y) \text{ad}^2 Y(X)$ as linear combinations of the basis vector fields in (12), then we can conclude that

$$\frac{d}{dt} \lambda \text{ad}^2 Y(X) = a_1(t) \lambda (X - Y) + a_2(t) \lambda [Y, X] + a_3(t) \lambda [Y[Y, X]].$$

In particular, if we define $y_1(t) = \lambda(t)[X(\gamma(t)) - Y(\gamma(t))]$, $y_2(t) = \lambda(t)[Y, X](\gamma(t))$ and $y_3(t) = \lambda(t)[Y[Y, X]](\gamma(t))$, then

$$\frac{d}{dt} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 0 & -1 & 0 \\ -u c_1 & -u c_2 & -1 - u c_3 \\ a_1 & a_2 & a_3 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}.$$

If $u(t) = -1/c_3(\gamma(t))$, then the fundamental matrix for this system of linear equations would have the form

$$\Phi(t) = \begin{pmatrix} 1 & -t & 0 \\ ? & ? & 0 \\ ? & ? & ? \end{pmatrix}.$$

In particular, if we could choose $y_1(0) = y_2(0) = 0$ and $y_3(0) \neq 0$, our feedback control would be singular. Note that $t \rightarrow \lambda(t)$ is any solution of an adjoint differential equation

so we may choose $\lambda(0)$ as we wish. By (13), we may choose $\lambda(0)$ so that $\lambda(0)Y(x') = 1$, and $\lambda(0)(X - Y)(x') = \lambda(0)[Y, X](x') = 0$. This concludes the proof.

Note that this result implies that the existence of singular controls of first order on open subsets of M is stable under C^3 small perturbations of the control system. We have not shown that if the singular control is optimal at all points of the open set then the singular control for nearby perturbed systems is optimal. We conjecture that this is true.

Theorem I.2(2) follows immediately from the following two lemmas whose proofs we omit because they are so similar to the proofs of Lemmas III.2 and III.3.

III.9. LEMMA. *There is an open dense subset \mathcal{U}' of \mathcal{B} such that if $(X, Y) \in \mathcal{U}'$, then $Y, X - Y$ and $[Y, X]$ are linearly independent at all points of M with the possible exception of points on a closed embedded hypersurface of M that depends on (X, Y) .*

III.10. LEMMA. *There is an open dense subset \mathcal{U}'' of \mathcal{B} such that if $(X, Y) \in \mathcal{U}''$, then $X - Y, [Y, X]$ and $[Y[Y, X]]$ are linearly independent at all points of M with the possible exception of points on a closed embedded hypersurface of M that depends on (X, Y) .*

If we set $\mathcal{U} = \mathcal{U}' \cap \mathcal{U}''$, Theorem I.2(2) follows from the proof of Theorem I.2(1).

We remark that if we had taken \mathcal{B} to be the set of nonautonomous vector fields, we would have had to replace X and Y and their brackets by

$$\hat{X} = \begin{pmatrix} X \\ 1 \end{pmatrix} \quad \text{and} \quad \hat{Y} = \begin{pmatrix} Y \\ 1 \end{pmatrix}$$

and their respective brackets. The hypersurfaces in M would have become varieties of codimension one in $M \times \mathbb{R}$.

Now let us discuss the case when $\dim M = m \geq 4$. Let $(X, Y) \in \mathcal{C}$. If $\text{ad}^i Y(X)$ are linearly independent at x for $i = 0, \dots, m-1$, and if the vector fields $Y, \text{ad}^i Y(X)$ are linearly independent at x for $i = 0, 1, 3, \dots, m-1$, then if we use a similar proof, an analogue of Lemma III.8 holds. That is, if $[X - Y, [Y, X]] = \sum_{i=0}^{m-1} a_i \text{ad}^i Y(X)$ for smooth functions a_0, \dots, a_{m-1} defined in a neighborhood of x , then $u(t) = -1/a_2(\gamma(t))$ is a strictly singular control in a neighborhood of x whenever $-1/(a_2(x)) \in (0, 1)$. As a result singular controls of first order do exist on open sets and their existence is stable under C^k small perturbations of the system (X, Y) for $k \geq m$.

If $m \geq 4$, there is no guarantee that these are the only singular controls through x . We conjecture the existence of a C^k open dense subset \mathcal{V} of \mathcal{C} such that if $(X, Y) \in \mathcal{V}$, then with the possible exception of a closed set of positive codimension, there will be infinitely many singular controls through each $x \in M$.

To support the conjecture, let us assume $m = 4$. Let

$$\tilde{Y} = \begin{pmatrix} Y \\ 1 \end{pmatrix}, \quad \tilde{X} = \begin{pmatrix} X \\ 1 \end{pmatrix}$$

be vector fields on $M \times \mathbb{R}$, let $\tilde{X}^0 = \tilde{X}$ and let $\tilde{X}^i = \text{ad}^i \tilde{Y}(\tilde{X})$. If we are at a point where the \tilde{X}^i are linearly independent for $i = 0, \dots, 4$, we can make the following change of coordinates in a neighborhood of $y \in M \times \mathbb{R}$:

$$(t_1, t_2, \dots, t_5) \rightarrow \tilde{Y}_{t_1} \tilde{X}_{t_2}^0 \tilde{X}_{t_3}^1 \tilde{X}_{t_4}^2 \tilde{X}_{t_5}^3(y).$$

Recall that the subscript t indicates time t flow of the given vector field.

Identifying \tilde{X} and \tilde{Y} with their new coordinate representations and using Taylor's theorem we find $y = (0, 0, 0, 0, 0)$ and

$$\tilde{Y} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \tilde{X} = \begin{pmatrix} \varepsilon_1 \\ 1 + \varepsilon_2 \\ t_1 + \varepsilon_3 \\ \frac{1}{2}t_1^2 + \varepsilon_4 \\ \frac{1}{3}t_1^3 + \varepsilon_5 \end{pmatrix}$$

where $\varepsilon_i(t_1, \dots, t_5) = b_{i1}(t_1, \dots, t_5)t_1t_2 + b_{i2}(t_1, \dots, t_5)t_1^2t_3 + b_{i3}(t_1, \dots, t_5)t_1^3t_4$.

If A_0 is the set of all points in the coordinate neighborhood that can be reached from the origin by moving forward in time along trajectories of (\tilde{X}, \tilde{Y}) , and if $\hat{\gamma}_u$ lies in the boundary of A_0 for all time, then u satisfies the maximum principle. See [1].

In particular, a control that minimized t_4 would satisfy the maximum principle. We would expect such a control to steer us in the hypersurface H defined by

$$\frac{\partial}{\partial t_1} \left(\frac{1}{2}t_1^2 + \varepsilon_4 \right) = 0$$

where the derivative of t_4 (the fourth coordinate of \tilde{X}) is a minimum as a function of t_1 for t_2, \dots, t_5 fixed. It is not hard to check that the condition

$$-1/a_2(0, \dots, 0) = -1/b_{41}(0, \dots, 0) \in (0, 1)$$

is the same as requiring \tilde{X}, \tilde{Y} to point to opposite sides of H and if $u(0) = -1/a_2(0, \dots, 0)$ then $u(0)\tilde{X}(0, \dots, 0) + (1 - u(0))\tilde{Y}(0, \dots, 0)$ is tangent to H at the origin.

Now if $b_{51}(0, \dots, 0) \neq 0$, we should be able to find a constant $c \neq 0$ such that \tilde{X}, \tilde{Y} are transversal and point to opposite sides of the hypersurface H' defined by

$$\frac{\partial}{\partial t_1} \left[\frac{1}{2}t_1^2 + \varepsilon_4 + c \left(\frac{1}{3}t_1^3 + \varepsilon_5 \right) \right] = 0.$$

Intuitively, the unique control steering us into H' would minimize $t_4 + ct_5$, thus steer us into the boundary of A_0 and be strictly singular. Since there are infinitely many distinct c 's with these properties, we would expect the existence of infinitely many singular controls through x whenever $b_{51}(0, \dots, 0) \neq 0$. Generically we would expect this last inequality to occur on the complement of a set of positive codimension in M .

Acknowledgment. I would like to thank the referee for suggestions that simplified the proof of Theorem I phenomenally, the Radcliffe Institute and its fellows for financial and moral support and the mathematics department of M.I.T. for kind assistance and hospitality.

REFERENCES

- [1] F. ALBRECHT, *Topics in Control Theory*, Springer-Verlag, Berlin, 1968.
- [2] F. COSTAL, *Classes d'équivalence de systèmes linéaires nonautonomes*, Analyse appliquée et informatique no. 7504, U.E.R. de Mathématiques et d'informatique, Université de Bordeaux 1, France.
- [3] ———, *El problema del tiempo minimo en ciertos sistemas lineales*, Publ. Depto. Anal. Mat., Univ. Santiago de Compostela, Spain, 1974.
- [4] M. GOLUBITSKY AND V. GUILLEMIN, *Stable Mappings and their Singularities*, Springer-Verlag, New York, 1974.

- [5] C. D. JOHNSON, *Singular solutions in problems of optimal control*, Advances in Control Systems Theory and Applications, Vol. 2, C. T. Leondes, ed., Academic Press, New York, 1965, pp. 209–267.
- [6] A. KRENER, *The high order maximal principle and its applications*, preprint.
- [7] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
- [8] C. LOBRY, *Quelques aspects qualitatifs de la théorie de la commande*, Ph.D. Thesis, l'Université de Grenoble, France, 1972.
- [9] ———, *Un propriété générique des couples de champs de vecteurs*, Czechoslovak Math. J., 22 (1972), pp. 230–237.
- [10] J. MARTINET, *Sur les singularités des formes différentielles*, Ann. Inst. Fourier (Grenoble), 20 (1970), pp. 95–178.
- [11] D. REBHUHN, *On the set of attainability of nonlinear nonautonomous control systems*, this Journal, 15 (1977), pp. 803–812.

ON FINDING THE MAXIMAL RANGE OF VALIDITY OF A CONSTRAINED SYSTEM*

SHMUEL GAL,[†] BORIS BACHELIS[†] AND AHARON BEN-TAL[‡]

Abstract. For a given set of k constraints which depend on a real parameter x , $x \in [0, 1]$ (or any other interval) it is desired to find the maximal subinterval $[0, \lambda]$ where all these constraints are satisfied. In order to find λ one can make n sequential observations where each of them can be made on any one of the constraints at a chosen point.

If $k = 1$ then the minimax procedure (bisection) guarantees an accuracy of $(\frac{1}{2})^n$ in locating λ . The methods considered in this paper are applicable for $k > 1$. The surprising property of some of these methods is the fact that for any k , they guarantee an accuracy of $(\frac{1}{2} + \varepsilon_k)^n$ where ε_k rapidly decreases to zero as n tends to infinity. Thus, for large n the asymptotic behavior is "almost" like the bisection procedure for one constraint. Numerical experiments show that these methods are efficient even for relatively small n and that they can handle effectively any number of constraints.

CONTENTS

1. Introduction	473
2. General description of the results obtained in this work	475
3. Optimization among simplified procedures	477
4. Overall optimization for $k=2$	481
5. Optimization for $k > 2$	489
6. Numerical evaluation of the asymptotically optimal procedures for $k > 2$	496
7. Randomized procedures	498
8. Points for further research	500
Appendix	500
References	503

1. Introduction. The problem considered in this work arises in the design of experiments in several fields and in mathematical programming as will be illustrated by the following two examples:

A. Capability tests. A manufacturer produces a system which is composed of several subsystems (or components). The system is so built that a failure of one of the subsystems causes the whole system to fail. The manufacturer wishes to find the maximal value of a certain crucial environmental condition (such as temperature) which the system can tolerate. The usual procedure is to test the whole system as one block, at several levels of the environmental condition. In the case of expensive or of destructive testing it may be advantageous to test each subsystem separately and to perform the tests sequentially. In this case there exists a problem of optimal sequential choice of the environmental condition level for testing each subsystem, in order to determine, as accurately as possible, the maximal level which can be tolerated by the whole system.

A similar problem of this type arises in testing a set of assets, each of them being designed to perform the same mission, in order to choose the asset which has the maximal capability. Experimental problems of this type also occur in medical and other fields. A list of such problems is presented in [6].

B. Mathematical programming. A typical problem of this field is to find: $\min g(\mathbf{x})$ subject to $\mathbf{x} \in F$ where the "feasible set" F is defined by:

$$F = \{\mathbf{x} : g_i(\mathbf{x}) \leq 0, i = 1, \dots, k, \mathbf{x} \in R^m\}.$$

* Received by the editors August 12, 1976, and in revised form June 20, 1977.

[†] IBM Israel Scientific Center, Haifa 32000, Israel.

[‡] Computer Science Department, Technion-Israel Institute of Technology, Haifa, Israel. The work of this author was supported in part by the National Science Foundation under Grant NSF ENG76-10260.

Several methods of minimization which are called “methods of feasible directions” [5], [7], [8] and [9] use the following procedure: Starting from a nonoptimal point $\mathbf{x}_0 \in F$ they find (by methods which will not be discussed here) a direction $\mathbf{b} \in R^m$ such that for some $\lambda > 0$: $\mathbf{x}_0 + \lambda \mathbf{b} \in F$ and $g(\mathbf{x}_0 + \lambda \mathbf{b}) < g(\mathbf{x}_0)$; such a \mathbf{b} is called “feasible direction of descent.” Next these methods attempt to find the minimal point along the ray which starts from \mathbf{x}_0 in the direction \mathbf{b} ; i.e. to find the point $\mathbf{x}_0 + \lambda_0 \mathbf{b}$ where λ_0 is the maximal number such that for all λ , $0 \leq \lambda < \lambda_0$,

$$(1) \quad \frac{dg(\mathbf{x}_0 + \lambda \mathbf{b})}{d\lambda} \leq 0 \quad \text{and} \quad g_i(\mathbf{x}_0 + \lambda \mathbf{b}) \leq 0, \quad i = 1, \dots, k.$$

λ_0 will be called “the optimal step size.”

In the case when the constraints presented in (1) can not be easily solved or if they are given as “black boxes,” one has to find λ_0 by observing these constraints for several values of λ . In problems with many constraints one has to choose the observation points efficiently because otherwise the optimization will use too much computer time. This brings us back to problems of the type discussed in example A.

The same problem arises in several modifications of the feasible direction method which are based on the following procedure: At each stage, one starts from an interior point of the feasible set F and has to find the boundary point of F along the segment which joins this feasible point to a certain infeasible point. Such a problem is discussed in [1].

Problems of the type described above will be discussed in the following framework: Let $H = \{f_1, \dots, f_k\}$ be a set of functions defined on an interval (for convenience this interval will be taken as $[0, 1]$). It is assumed that the functions f_i , $i = 1, \dots, k$, satisfy the following condition:

Either there exists a unique point z_i , $0 < z_i < 1$ such that

$$f_i(x) < 0 \quad \text{for } 0 \leq x < z_i \quad \text{and} \quad f_i(x) > 0 \quad \text{for } z_i < x \leq 1$$

(2) or

$$f_i(x) < 0 \quad \text{for all } 0 \leq x \leq 1.$$

The point z_i will be referred to as “root.” Our aim is to locate the minimal root z

$$(3) \quad z = \min_{1 \leq i \leq k} z_i$$

(if $f_i(x) < 0$ for all $0 \leq x < 1$ and all $1 \leq i \leq k$ then z is defined to be 1).

The interval $[0, z]$ is the maximal subinterval of $[0, 1]$ where all $f_i(x)$ are negative (or in other words, the maximal range of validity of all the constraints).

In order to locate z one may gain information by making observations on each of the functions $f_i \in H$ (one at a time) at points in the interval $[0, 1]$ which are determined sequentially, using the results of the previous observations. No regularity conditions except (2) are assumed about the functions $f_i \in H$ so that the only relevant information for locating the minimal root z is the sign of the function f_i at points where it has been observed. Specifically, let $S_k(n)$ denote the class of n -observation (nonrandomized) procedures $s_k(n)$; such an $s_k(n)$ consists of a collection of n pairs,

$$(4) \quad s_k(n) = \{(j_i, x_i), i = 1, \dots, n\},$$

so that in the i th stage of $s_k(n)$ one observes the sign of $f_{j_i}(x_i)$ where both j_i and x_i are functions of the previous observations.

Before starting the observations, the only knowledge about z is that $z \in (0, 1]$ but the additional information obtained after applying $s_k(n)$ can be used to assert that for each i , $1 \leq i \leq k$, $z_i \in [L_i, U_i]$, where L_i is equal to the maximal point at which f_i has been observed and found to be negative (or to 0 if no such point exists). Similarly, U_i is equal to the minimal point at which f_i has been observed and found to be positive (or to 1 if no such point exists). It would then be possible to assert that $z \in [L, U]$ where

$$(5) \quad L = \min_{1 \leq i \leq k} L_i, \quad U = \min_{1 \leq i \leq k} U_i.$$

The interval $[L, U]$ will be called the “interval of uncertainty.” For each $s_k(n)$ let $V(s_k(n))$ be the length of this interval for the worst possible choice of f_1, \dots, f_k (i.e. the longest interval that can be obtained if one uses $s_k(n)$). Thus, the procedure $s_k(n)$ can assure us of locating z in an interval whose length does not exceed $V(s_k(n))$ which will be called the “value of $s_k(n)$.”

The maximal achievable accuracy in locating z is given by

$$(6) \quad V_k(n) = \inf_{s_k(n) \in S_k(n)} V(s_k(n)).$$

$\bar{s}_k(n)$ will be called a “minimax procedure” if

$$(7) \quad V(\bar{s}_k(n)) = V_k(n).$$

An important property of search problems is the asymptotic behavior (as a function of n) of $V_k(n)$. Of special importance is the number r_k defined by:

$$(8) \quad r_k = \lim_{n \rightarrow \infty} [V_k(n)]^{1/n}$$

(if it exists). The number $1 - r_k$ measures the asymptotic proportion of reduction per observation of the interval of uncertainty. It has been proved in [2] that for a wide class of problems the right side of (8) does converge.

We shall also refer to the maximal achievable accuracy via a certain subclass $S'_k(n) \subset S_k(n)$ of search procedures. In this case the expression

$$V_k(n) = \inf_{s_k(n) \in S'_k(n)} V(s_k(n))$$

will be referred to as the “minimax value of procedures of subclass $S'_k(n)$.” Similarly, we shall use the expression

$$r_k = \lim_{n \rightarrow \infty} [V_k(n)]^{1/n}$$

for the same subclass $S'_k(n)$ (the notation for $V_k(n)$ and r_k is identical to (6) and (8) since no confusion is bound to occur).

2. General description of the results obtained in this work. Let us consider the problem of locating the minimal root z , as a function of the number of observable functions, k . If $k = 1$ then the minimax procedure is bisection so that $V_1(n) = (\frac{1}{2})^n$ and $r_1 = \frac{1}{2}$.

Let us now consider $k > 1$: A naive procedure is to separately bisect the interval of uncertainty for each z_i (or alternatively to observe all the functions f_i at $\frac{1}{2}$ and then at $\frac{3}{4}$ if

$f_i(\frac{1}{2}) < 0$ for all i or $\frac{1}{4}$ if $f_i(\frac{1}{2}) > 0$ for at least one i , etc.). Procedures of this type have the following value:

$$(9) \quad V(s_k(n)) \sim \left(\frac{1}{2}\right)^{n/k} \quad \text{so that for the class of naive procedures} \quad r_k = \left(\frac{1}{2}\right)^{1/k}.$$

Thus, if k is large then the amount of reduction per measurement, $1 - (\frac{1}{2})^{1/k}$, is very small.

To be sure, this is a naive approach and one attempt to improve it is presented in § 3 where we shall define a simple class of procedures, which will be referred to as “simplified procedures.” The best one among them has an asymptotic behavior $V(s_k(n)) \sim (r_k)^n$ where r_k satisfies

$$(10) \quad (r_k)^k + r_k = 1.$$

A similar result has been obtained in [1]. This is an improvement over (9) but still, as k gets large r_k approaches 1, which is an undesirable behavior for large k .

It is therefore very surprising that we were able to find procedures $s_k(n)$ which have the following property:

$$(11) \quad [V(s_k(n))]^{1/n} \rightarrow \frac{1}{2}.$$

Moreover, the procedures that we found satisfy a stronger condition:

$$(12) \quad V(s_k(n)) = \left(\frac{1}{2}\right)^{n - l_k(n)}$$

where

$$(13) \quad l_k(n) \leq (k-1) \log_2 n + c$$

(c being a constant).

Since for one constraint $V_1(n) = (\frac{1}{2})^n$, it is natural to call $l_k(n)$ the “number of lost observations” (relatively to the case: $k = 1$). It follows from (13) that for large n , $l_k(n)/n$ rapidly approaches zero. Thus, by comparing the numbers of observations needed to locate z very accurately for 1 versus k constraints it is seen that they differ only by a small fraction, which means that in this case we do not expend much for checking additional constraints.

Properties (12) and (13) are equivalent to:

$$(14) \quad V(s_k(n)) = \left[\frac{1}{2}(1 + \varepsilon_k)\right]^n \quad \text{where} \quad \varepsilon_k \leq \frac{(k-1) \log_e n + c'}{n}$$

and c' is a constant.

Since any search procedure satisfies $V(s_k(n)) > (\frac{1}{2})^n$ it is natural to refer to all procedures which satisfy (11) as asymptotically optimal. However we shall use the term “asymptotically optimal procedures” only for those procedures which satisfy the stronger condition (13). As can be expected, these procedures are quite sophisticated and most of the paper is dedicated to their derivation.

In § 4 we shall analyze the case of $k = 2$. In this case it is possible to calculate numerically the optimal (minimax) procedure using dynamic programming, as will be done in § 4.1. In § 4.2 we shall find two asymptotically optimal procedures and prove that their performance is close to the optimal even for small n . Numerical comparisons will be made among the optimal, asymptotic optimal, and simplified procedures.

In § 5 we shall consider the case of $k > 2$. In this case it is practically impossible to use dynamic programming in order to calculate numerically the optimal procedure; however we were able to find asymptotically optimal procedures, which can be very simply implemented, and to test them numerically. It turned out that these procedures are effective even for a comparatively small number of observations n .

Actually, the numerical results which are presented in §§ 5 and 6 show that if $k \leq 20$ then by using no more than $n = 2k + 20$ observations one can be sure that the interval of uncertainty will be about 10^{-5} or less, while if $k = 100$ then $1.7k$ observations (i.e. only 1.7 observations per constraint!) are sufficient for this purpose.

In § 7 we discuss the effect of randomization for a possible further reduction in the interval of uncertainty. Finally, some problems for further research are presented in § 8.

3. Optimization among simplified procedures.

3.1. Description of the simplified procedures. Assume that each of the observable functions f_1, \dots, f_k has a unique root in the sense of (2) and that one wishes to locate the minimal root z .

If at a point x , $0 < x < 1$, $f_i(x) > 0$ for some $i \in \{1, \dots, k\}$ then clearly $z < x$ so that it is reasonable to make all the subsequent observations only at points in the interval $(0, x)$. On the other hand if $f_i(x) < 0$ for all $i = 1, \dots, k$ then $z > x$ and the subsequent observation should be taken only to the right of x . If $f_i(x) \leq 0$ for $i = j+1, \dots, k$ and $f_j(x) > 0$ then $z < x$ and none of the functions f_{j+1}, \dots, f_k possesses the minimal root so that they can be ignored during the continuation of the search procedure. Thus a class of reasonable procedures which will be called "simplified procedures" can be described as follows:

Suppose that during the search procedure we have k' ($k' \leq k$) current relevant functions $f_1, \dots, f_{k'}$, n' ($n' \leq n$) remaining observations and no observation has yet been made in the current interval of uncertainty (a, b) . This situation will be referred to as "starting point." The next observation is made at $f_{k'}(x)$ where $a < x < b$; if $f_{k'}(x) > 0$ then the interval of uncertainty will become (a, x) , the number of relevant functions k' will remain the same, the number of remaining observations will be reduced by one, and we will have returned to a starting point.

If, on the other hand $f_{k'}(x) \leq 0$ then we next observe $f_{k'-1}(x)$ and continue to observe the other functions at the same point x until one of the two following possibilities occur: If $f_i(x) \leq 0$ for $i = j+1, \dots, k'$ and $f_j(x) > 0$, then the interval of uncertainty will be (a, x) , the relevant functions will be f_1, \dots, f_j (that is, the new k' is j), the number of remaining observations will be $n' - (k' - j + 1)$ and we again shall be at a starting point.

The other possibility is that $f_i(x) \leq 0$ for all $i = 1, \dots, k'$. If for some j , $f_j(x) = 0$ then we are in the trivial situation where we know that $z = x$. If $f_i(x) < 0$ for all $i = 1, \dots, k'$ then the current interval of uncertainty will be (x, b) , the number of relevant functions will remain k' , the number of remaining observations will be $n' - k'$ and we shall again be at a starting point.

The search procedure is finished when we have used all the n observations.

Thus, in this class of simplified procedures one has to choose x as a function of k' , n' and the current interval of uncertainty (a, b) . Actually, it can be easily seen that the only problem is to choose $(x - a)/(b - a)$ as a function of k' and n' so that it can be assumed that $a = 0$ and $b = 1$.

3.2. The optimal simplified procedure. It is easy to see that for any current number of constraints k' and current number of remaining observations n' , the value of the optimal simplified procedure is proportional to the interval of uncertainty. This fact

together with the description given in § 3.1 implies that the following recurrence relation holds for the minimax value of simplified procedures $V_k(n)$:

$$(15) \quad \text{For } 0 \leq n < k, \quad V_k(n) = 1$$

and for $n \geq k$:

$$(16) \quad V_k(n) = \min_{0 \leq x \leq 1} \max [x V_k(n-1), x V_{k-1}(n-2), \dots, x V_1(n-k), (1-x) V_k(n-k)].$$

In order to explicitly solve equation (16) we will show:

PROPOSITION 1. For all $1 \leq i < k$:

$$(17) \quad V_k(n-1) \geq V_{k-i}(n-i-1).$$

Proof. If $n-1 < k$ then $n-i-1 < k-i$ and both sides of (17) are equal to one. If $n-1 \geq k$ then by (16):

$$\begin{aligned} V_k(n-1) &= \min_x \max [x V_k(n-2), \dots, x V_{k-i}(n-i-2), \dots, \\ &\quad x V_1(n-k-1), (1-x) V_k(n-k-1)] \\ &\geq \min_x \max [x V_{k-i}(n-i-2), \dots, x V_1(n-k-1), (1-x) V_k(n-k-1)] \\ &= V_{k-i}(n-i-1). \quad \text{Q.E.D.} \end{aligned}$$

It follows from Proposition 1 that (16) reduces to

$$(18) \quad V_k(n) = \min_{0 \leq x \leq 1} \max [x V_k(n-1), (1-x) V_k(n-k)]$$

(i.e. the worst case occurs when the number of relevant functions does not decrease).

Equation (18) with the initial condition (15) can be explicitly solved as follows: Let x_n^* be the point where the minimum of the right side of (18) is obtained; then it readily follows that for all $n \geq k$

$$(19) \quad V_k(n) = x_n^* V_k(n-1) = (1-x_n^*) V_k(n-k);$$

thus

$$(20) \quad x_n^* = \frac{V_k(n)}{V_k(n-1)} = \frac{V_k(n-k)}{V_k(n-1) + V_k(n-k)}.$$

It follows from (15), (19) and (20) that $V_k(n)$ satisfies the following recursive relations:

$$(21) \quad V_k(n) = \begin{cases} 1 & \text{for } 0 \leq n < k, \\ \frac{V_k(n-1) V_k(n-k)}{V_k(n-1) + V_k(n-k)} & \text{for } k \leq n. \end{cases}$$

Let $L_k(n) = 1/V_k(n)$; then it can be obtained from (21) that $L_k(n)$ satisfies

$$(22) \quad L_k(n) = \begin{cases} 1 & \text{for } 0 \leq n < k \\ L_k(n-1) + L_k(n-k) & \text{for } n \geq k. \end{cases}$$

$L_k(n)$ can be interpreted as the largest initial interval of uncertainty which can be reduced for sure to the unit interval, using n observations. It follows from (22) that $L_k(n)$ is always an integer.

The optimal searching point x_n^* is given by

$$(23) \quad x_n^* = \frac{L_k(n-1)}{L_k(n)}.$$

The value of the optimal among simplified procedures will be denoted by SP.

3.3. Asymptotic behavior of simplified procedures. Equation (22) can be solved explicitly, $L_k(n)$ being expressed as a combination of geometric terms. We shall use only the fact that for large n :

$$(24) \quad \frac{1}{V_k(n)} = L_k(n) \sim \left(\frac{1}{a_k}\right)^n \quad \text{where } a_k \text{ satisfies:}$$

$$0 < a_k < 1 \quad \text{and} \quad (a_k)^k + a_k = 1.$$

Thus, for the class of simplified procedures $r_k = \lim_{n \rightarrow \infty} [V_k(n)]^{1/n} = a_k$.

For $k = 2$, $L_2(n)$ are Fibonacci numbers and $a_2 = 2/(1 + \sqrt{5}) \approx .62$ is the golden section. For large k a_k is close to 1 so that the asymptotic reduction per observation $1 - a_k$ is close to zero which shows that simplified procedures are not efficient for large k . In the following sections we shall present methods for which the asymptotic reduction per observation tends to $\frac{1}{2}$ for all k .

A simple example which illustrates the difference between the optimal among simplified procedures and the overall optimal procedure is presented in § 3.4.

3.4. The case: $k = 2$, $n = 4$.

For simplified procedures. Using (22) we obtain: $L_2(2) = 2$, $L_2(3) = 3$ and $L_2(4) = 5$; thus for simplified procedures $V_2(4) = \frac{1}{5}$. It follows from (23) that one should start by observing f_2 at $x_4^* = L_2(3)/L_2(4) = \frac{3}{5}$.

$$\text{If } f_2\left(\frac{3}{5}\right) > 0 \text{ then next observe } f_2 \text{ at } x_3^* = \frac{L_2(2)}{L_2(3)} x_4^* = \frac{2}{5} \text{ etc.}$$

$$\text{If } f_2\left(\frac{3}{5}\right) \leq 0 \text{ then next observe } f_1 \text{ at } \frac{3}{5} \text{ etc.}$$

The final interval of uncertainty will not exceed $\frac{1}{5}$.

For the overall optimal procedure. In the next section it will be shown how to obtain the overall optimal procedure for $k = 2$. If $n = 4$ one can be assured of obtaining an interval of uncertainty whose length does not exceed $\frac{1}{6}$. This procedure, which is quite complicated, is presented in Table 1.

It should be noted that this table illustrates the difference between simplified and overall optimal procedures only for very small n . The huge difference between the performance of these procedures becomes significant when n or k is large. It should also be noted that although the overall optimal procedure is very complicated and can be practically calculated only for $k = 2$ (and maybe also for $k = 3$), the asymptotically optimal procedures are relatively simple and can be used very easily by the searcher.

TABLE 1
Optimal procedure for $k = 2$ and $n = 4$.

First observation	Second observation	Third observation	Fourth observation	Interval uncertainty
$f_2(\frac{3}{6})$	$f_2(\frac{2}{6})$	> 0	immaterial	$\subset (0, \frac{1}{6})$
		$f_2(\frac{1}{6})$		
		< 0	> 0	$(0, \frac{1}{6})$
			$f_1(\frac{1}{6})$	
			< 0	$(\frac{1}{6}, \frac{2}{6})$
	< 0		> 0	$(0, \frac{1}{6})$
			$f_1(\frac{1}{6})$	
		> 0	< 0	$(\frac{1}{6}, \frac{2}{6})$
		$f_1(\frac{2}{6})$		
		< 0	immaterial	$\subset (\frac{2}{6}, \frac{3}{6})$
< 0	$f_1(\frac{4}{6})$		> 0	$(0, \frac{1}{6})$
			$f_1(\frac{1}{6})$	
		> 0	< 0	$(\frac{1}{6}, \frac{2}{6})$
		$f_1(\frac{2}{6})$		
		< 0	> 0	$(\frac{2}{6}, \frac{3}{6})$
			$f_1(\frac{3}{6})$	
	< 0		< 0	$(\frac{3}{6}, \frac{4}{6})$
		> 0	> 0	$(\frac{3}{6}, \frac{4}{6})$
			$f_2(\frac{4}{6})$	
		< 0	< 0	$(\frac{4}{6}, \frac{5}{6})$
		$f_2(\frac{5}{6})$		
	< 0		> 0	$(\frac{4}{6}, \frac{5}{6})$
		> 0	> 0	$(\frac{4}{6}, \frac{5}{6})$
			$f_1(\frac{5}{6})$	
		< 0	< 0	$(\frac{5}{6}, 1)$

4. Overall optimization for $k = 2$.

4.1. General description. Consider the problem of finding the minimal root z of the two functions $f_1(x)$ and $f_2(x)$, each of which has a unique root in the sense of (2), using n sequential observations. Each observation can be made on either f_1 or f_2 at a point chosen according to the results of the prior observations. Since no regularity assumptions except (2) are imposed on f_i , the only relevant information is the sign of f_i at the observed points.

In considering possible search procedures we shall have to describe the situations encountered during each search. To this end we shall use the notion of "relevant intervals" of the functions. Furthermore, we shall be interested in certain transformations which these intervals undergo during the process. To illustrate what is meant by a relevant interval of a function f_i , let us look upon the following situation: Assume that the first observation was made on f_1 at a point x , $0 < x < 1$. If $f_1(x) > 0$ then its root satisfies $0 < z_1 < x$ so that the relevant interval of f_1 at this stage is $(0, x)$. Moreover, we claim that, using a minimax procedure, it is not worthwhile to observe f_2 at any point $y \in [x, 1]$ because if $f_2(y) > 0$ then no further information for locating z will have been obtained. (It is true that if $f_2(y)$ were negative then much would be gained by this observation because this would imply that $z = z_1$, but we are interested in the worst case.)

Thus, before starting the observations, the relevant intervals are $(0, 1)$ and $(0, 1)$. After observing f_1 at x then if $f_1(x) > 0$ the relevant intervals become $(0, x)$ and $(0, x)$ while if $f_1(x) < 0$ then these intervals will be $(x, 1)$ and $(0, 1)$. It should be noted that in any event, the right end of the relevant intervals for f_1 and f_2 coincide. In general, if at any stage the relevant intervals are (y_1, b) and (y_2, b) and f_1 is observed at a point x , $y_1 < x < b$, then if $f_1(x) < 0$, the relevant intervals become (x, b) and (y_2, b) . If $f_1(x) > 0$ then two cases are possible:

Case 1. If $x > y_2$ then the relevant intervals will be (y_1, x) and (y_2, x) .

Case 2. If $y_1 < x < y_2$ then it is known that $f_1(x) > 0$ while $f_2(x) < 0$ (because $f_2(y_2) < 0$ and $x < y_2$); thus $z = z_1$ and f_2 can be eliminated from the search. In this case the subsequent observations should be made only on f_1 at points of the relevant interval (y_1, x) .

In § 5.1 it will be proved by Proposition 2 that the optimal procedure always observes the function f_i which corresponds to the maximal current relevant interval; thus if $k = 2$ and the current relevant intervals are (y_1, b) and (y_2, b) where $y_1 \leq y_2$, it may be assumed that at this stage one has to observe f_1 . The transformation of the relevant interval after an additional observation is presented in Table 2.

TABLE 2
Transformation of relevant intervals after observing $f_1(x)$.

Results	New relevant intervals
$f_1(x) < 0$	(x, b) and (y_2, b)
$f_1(x) > 0$	$x > y_2$ (y_1, x) and (y_2, x)
	$x \leq y_2$ (y_1, x)

Table 2 will be used in the sequel when we establish recursive relations for the relevant intervals obtained by using certain search procedures. It should also be noted

that the maximal relevant interval at the final stage of the search is equal to the final interval of uncertainty.

4.2. Optimal procedure using dynamic programming. Assume that at a certain stage of the search procedure, the relevant intervals for f_1 and f_2 are (y_1, b) and (y_2, b) where $y_1 \leq y_2$, and the number of remaining observations is m . Let $a_1 = b - y_1$ and $a_2 = b - y_2$; denote:

$$(25) \quad G_m(a_1, a_2) \stackrel{\text{def}}{=} \text{The length of the final (assured) interval of uncertainty when the next } m \text{ observations are executed optimally, given that the lengths of the current relevant intervals are } a_1, a_2 \text{ with } a_1 \geq a_2.$$

Obviously, $G_n(1, 1) = V_2(n)$ and $G_o(a_1, a_2) = a_1$. It is also easy to see that

$$(26) \quad G_m(a_1, a_2) = a_1 G_m\left(1, \frac{a_2}{a_1}\right) = a_1 g_m\left(\frac{a_2}{a_1}\right)$$

where

$$(27) \quad g_m(x) \stackrel{\text{def}}{=} G_m(1, x).$$

The function g_m can be calculated recursively as follows:

$g_0(x) = 1$ and for $m \geq 1$:

If one has $m + 1$ remaining observations and wishes to make the next observation at $1 - d$, then

$$(28) \quad g_{m+1}(x) = G_{m+1}(1, x) = \min \left[\min_{0 < d < x} \max(G_m(1 - d, x - d), G_m(x, d)); \min_{x \leq d < 1} \max\left(\frac{1 - d}{2^m}, G_m(d, x)\right) \right];$$

thus by (26), equation (28) is equivalent to:

$$(29) \quad g_{m+1}(x) = \min \left[\min_{0 < d < x} \max\left((1 - d)g_m\left(\frac{x - d}{1 - d}\right), xg_m\left(\frac{d}{x}\right)\right); \min_{x \leq d < 1} \max\left(\frac{1 - d}{2^m}, dg_m\left(\frac{x}{d}\right)\right) \right].$$

The evaluation of the righthand side of (28), and similar expressions, which will appear in the proofs of Theorem 1 and 2, is based on the following simple lemma:

LEMMA 1. Let f^1, f^2 be continuous decreasing functions and g^1, g^2 continuous increasing functions on the interval $[0, b]$ and such that for $a \in [0, b]$

$$(30) \quad f^1(a) = f^2(a); \quad g^1(a) = g^2(a).$$

Suppose further that

$$(31) \quad g^1(0) \leq f^1(0); \quad g^2(b) \geq f^2(b).$$

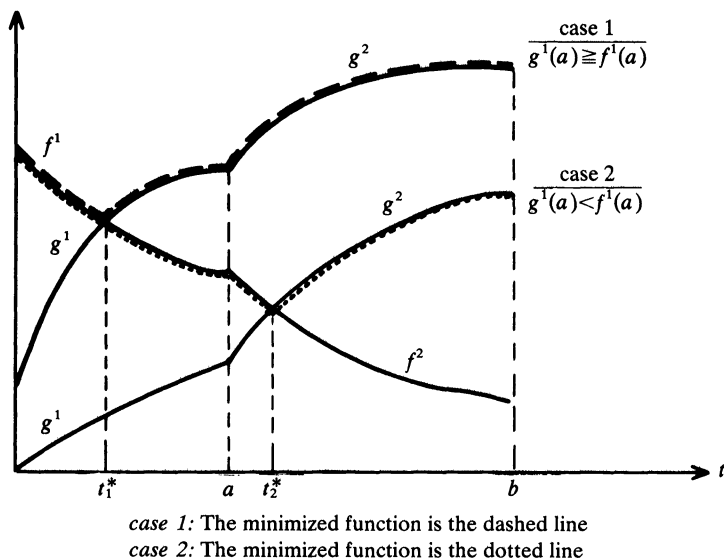


FIG. 1

Then

$$(32) \quad \min \left\{ \min_{0 \leq t \leq a} \max(f^1(t), g^1(t)); \min_{a \leq t \leq b} \max(f^2(t), g^2(t)) \right\} = \begin{cases} f^1(t_1^*) & \text{if } g^1(a) \geq f^1(a), \\ f^2(t_2^*) & \text{if } g^1(a) < f^1(a), \end{cases}$$

where t_i^* ($i = 1, 2$) satisfies

$$(33) \quad f^i(t_i^*) = g^i(t_i^*).$$

The proof can be readily deduced from Fig. 1.

Turning back to equation (28) we see that the lemma can be used with $b = 1$, $a = x$, $f^1(d) = G_m(1-d, x-d)$, $f^2(d) = (1-d)/2^m$, $g^1(d) = G_m(x, d)$ and $g^2(d) = G_m(d, x)$. The condition $g^1(a) \geq f^1(a)$ becomes here $x \geq x^*$ where

$$x^* = \frac{1}{1 + 2^m g_m(1)}$$

and so

$$(34) \quad g_{m+1}(x) = \begin{cases} (1-d_1)g_m\left(\frac{x-d_1}{1-d_1}\right) & \text{if } x^* \leq x \leq 1, \\ \frac{1-d_2}{2^m} & \text{if } 0 \leq x < x^*, \end{cases}$$

where d_1 satisfies

$$(1-d_1)g_m\left(\frac{x-d_1}{1-d_1}\right) = xg_m\left(\frac{d_1}{x}\right)$$

and d_2 satisfies

$$\frac{1-d_2}{2^m} = d_2g_m\left(\frac{x}{d_2}\right).$$

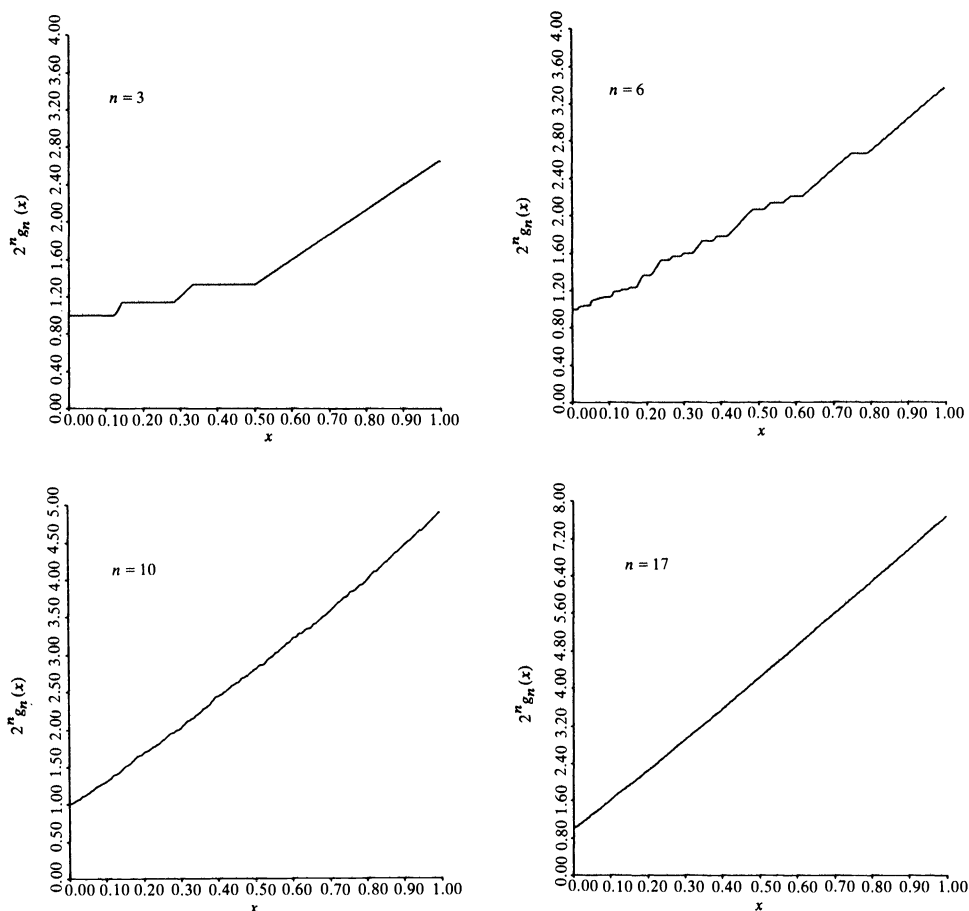


FIG. 2

Using equation (34) we wrote a computer program which calculates the optimal procedure and its value $V_2(n) = g_n(1)$. Table 3 (see § 4.3) presents the results and also compares the performance of the optimal, the optimal among simplified procedures, and the asymptotically optimal procedures which are presented in the next subsection. It should be noted that $1/V_2(n)$ is an integer, as can be expected.

We also obtained graphs of $g_n(x)$ for $n = 1, \dots, 17$. Figure 2 presents the function $2^n g_n(x)$, $0 \leq x \leq 1$, for $n = 3, 6, 10, 17$. By inspecting Fig. 2 it appears as if $2^n g_n(x)$ approaches a linear function of the form $1 + u_n x$ for $n \geq 10$. The fact that $g_n(x)$ behaves "almost" as such a function will be proved rigorously in § 4.3 where it will be used as a tool for finding asymptotically optimal procedures.

4.3. Asymptotically optimal procedures. In this, and the following subsections, we shall use the function $G_m(a_1, a_2)$ defined by (25). To repeat: $G_m(a_1, a_2)$ will denote the length of the final interval of uncertainty where one can make m additional observations having the following information: $f_1(1 - a_1) < 0$, $f_2(1 - a_2) < 0$ with $a_1 \geq a_2$ and either $f_1(1) > 0$ or $f_2(1) > 0$. Actually, it can be assumed that both f_1 and f_2 are positive at 1, for if $f_i(x) < 0$, $x \in (1 - a_i, 1)$, then instead of $f_i(x)$ one could consider $\bar{f}_i(x)$ where $\bar{f}_i(x) = f_i(x)$ for $x \in (1 - a_i, 1)$ and $\bar{f}_i(1) > 0$, and the results obtained would never be

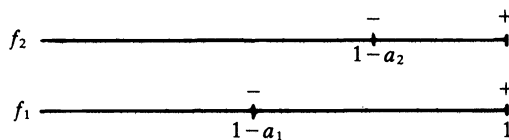


FIG. 3

improved (recall that we are interested in the worst case). This situation is depicted in Fig. 3.

It can be assumed, as before, that $a_1 \geq a_2$ (remember that in this case we shall prove in § 5.1 that it is optimal to make the next observation on f_1). The following theorem will be used as a tool for finding an asymptotically optimal procedure.

THEOREM 1.

$$(35) \quad G_m(a_1, a_2) \leq \frac{a_1 + u_m a_2}{2^m}$$

where u_m satisfies

$$(36) \quad u_0 = 0 \quad \text{and} \quad u_{m+1} = \frac{u_m^2 + u_m + 1}{u_m + 1/2}.$$

Proof. We shall present an inductive construction of a procedure to be denoted by $SU_2(m)$, which satisfies the following: Starting from the situation depicted in Fig. 3 and using m additional observations this procedure assures us of obtaining the value

$$(37) \quad V(SU_2(m)) < \frac{a_1 + u_m a_2}{2^m}.$$

The procedure $SU_2(m)$ is based on the following principle: At any stage of the search, if $n+1$ observations are left, then the current observation point is so chosen as to minimize the maximal value of the upper bound given in (35) where a_1 and a_2 are the relevant intervals at the next stage.

The induction proof is constructed as follows: For $m=0$, $u_0=0$ and since $G_0(a_1, a_2) = a_1$, it follows that (35) holds. Next assume that (37) holds for all $m \leq n$.

Similar to the derivation of the recursion (28) we obtain

$$(38) \quad \begin{aligned} G_{n+1}(a_1, a_2) &= \min \left[\min_{0 \leq d \leq a_2} \max(G_n(a_1 - d, a_2 - d), G_n(a_2, d)); \right. \\ &\quad \left. \min_{a_2 < d \leq a_1} \max\left(\frac{a_1 - d}{2^n}, G_n(d, a_2)\right) \right] \\ &\leq \min \left[\min_{0 \leq d \leq a_2} \max\left(\frac{a_1 - d + u_n(a_2 - d)}{2^n}, \frac{a_2 + u_n d}{2^n}\right); \right. \\ &\quad \left. \min_{a_2 < d \leq a_1} \max\left(\frac{a_1 - d}{2^n}, \frac{d + u_n a_2}{2^n}\right) \right] \quad \text{by the induction hypothesis} \end{aligned}$$

$$(39) \quad = \begin{cases} \frac{a_2 + u_n d_1}{2^n} & \text{where } d_1 = \frac{a_1 + (u_n - 1)a_2}{2u_n + 1} \text{ if } a_1 \leq (2 + u_n)a_2, \\ \frac{a_1 - d_2}{2^n} & \text{where } d_2 = \frac{a_1 - u_n a_2}{2} \text{ if } a_1 > (2 + u_n)a_2 \end{cases} \quad \text{by Lemma 1}$$

$$(40) \quad$$

$$\begin{aligned}
&= \begin{cases} \frac{1}{2^{n+1}} \left(\frac{u_n}{u_n + 1/2} a_1 + \frac{u_n^2 + u_n + 1}{u_n + 1/2} a_2 \right) & \text{if } a_1 \leq (2 + u_n)a_2, \\ \frac{1}{2^{n+1}} (a_1 + u_n a_2) & \text{if } a_1 \geq (2 + u_n)a_2 \end{cases} \\
&\equiv \frac{a_1 + u_{n+1}a_2}{2^{n+1}} \quad \text{by (36).} \quad \text{Q.E.D.}
\end{aligned}$$

It follows from (36) that

$$(41) \quad u_{m+1} = u_m + \frac{1}{2} + \frac{3/4}{u_m + 1/2}$$

this implies that

$$(42) \quad u_n \sim \frac{n}{2} + \frac{3}{2} \log_e n$$

(because $u_n \geq n/2$ so that $u_n + 1/2 \leq u_{n+1} \leq u_n + 1/2 + (3/2)/(n+1)$ etc.)

In any case, numerical calculation shows that

$$(43) \quad u_n < \frac{n}{2} + \frac{3}{2} \log_e n$$

for all $n \geq 5$.

The search procedure $SU_2(n)$ constructed in the proof of Theorem 1, which calculates the current observation point $1-d$ (where $d = d_1$ or d_2 according to whether a_1 is smaller or greater than $(2+u_n)a_2$) satisfies inequality (37). Thus, it follows from (37) and (42) that for large n

$$(44) \quad V(SU_2(n)) \leq \frac{1+u_n}{2^n} \leq \frac{n/(2-\varepsilon)}{2^n}$$

where ε approaches zero as n becomes large; thus

$$(45) \quad V_2(SU_2(n)) = \left(\frac{1}{2}\right)^{n-l_2(n)}$$

where the number of “lost observations” $l_2(n)$ satisfies

$$(46) \quad l_2(n) \leq \log_2 n - \log_2 (2 - \varepsilon).$$

Actually, the procedure $SU_2(n)$ has a value $V(SU_2(n))$ which is strictly less than $(1+u_n)/2^n$. This value has been calculated numerically by using a recursion formula and is presented in Table 3.

Although the procedure $SU_2(n)$ is simple to use, it is worthwhile to introduce another procedure to be denoted by $SN_2(n)$. In this procedure too, the observation point d is given by $d = d_1$ or $d = d_2$ of (39) or (40), but here the weight u_n is equal to $n/2$ instead of being defined by the recursion formula (36). The reason for introducing $SN_2(n)$ is that its extension for $k > 2$, which will be considered in § 5, will be found useful. It will also be shown there that this procedure is asymptotically optimal. In fact, it is seen from Table 3 that $SN_2(n)$ is only slightly inferior to $SU_2(n)$.

TABLE 3
The value $V(S_2(n))$ of several search procedures.

n	Optimal procedure	$SU_2(n)$	$SN_2(n)$	Simplified procedure
1	1	1	1	1
2	.50	.71	.75	.50
3	.33	.53	.50	.33
4	.17	.31	.31	.20
5	.10	.17	.19	.12
6	5.3E-2	9.8E-2	10.9E-2	7.7E-2
7	2.9E-2	5.4E-2	6.2E-2	4.8E-2
8	1.6E-2	2.9E-2	3.5E-2	2.9E-2
9	8.8E-3	1.6E-2	1.9E-2	1.8E-2
10	4.8E-3	8.5E-3	1.1E-2	1.1E-2
11	2.6E-3	4.6E-3	5.9E-3	6.9E-3
12	1.4E-3	2.4E-3	3.2E-3	4.3E-3
13	7.4E-4	1.3E-3	1.7E-3	2.6E-3
14	3.9E-4	6.8E-4	9.2E-4	1.6E-3
15	2.1E-4	3.6E-4	4.9E-4	1.0E-3
16	1.1E-4	1.9E-4	2.6E-4	6.3E-4
17	5.9E-5	9.8E-5	1.4E-4	3.9E-4
18	3.1E-5	5.1E-5	7.3E-5	2.4E-4
30		1.9E-8	2.9E-8	7.4E-7
50		2.7E-14	4.5E-14	4.9E-11

Both $SU_2(n)$ and $SN_2(n)$ are asymptotically optimal procedures; they are better than the simplified procedure for $n > 8$ and $n > 10$ respectively. Of course, the difference is more significant for large n because the value of $SU_2(n)$ behaves like $(n/2)/2^n$ while the value of the simplified procedure behaves like 0.62^n . The difference between the simplified and the asymptotically optimal procedures becomes more significant for large k , as will be demonstrated in § 5.

Since the procedure $SU_2(n)$ can be used simply while the optimal procedure is quite complicated, it is interesting to compare the performance of these two procedures. Table 3 shows that for all $n \geq 4$, the value obtained by using $SU_2(n)$ is better than the one obtained by using the optimal procedure with $n - 1$ observations; i.e. by using $SU_2(n)$, at most one observation is wasted. This is of course an experimental fact; however, in the next subsection we shall prove a theorem which gives a lower bound for the value of the optimal procedure. From this theorem it will follow that for any n , the number of wasted observations in $SU_2(n)$ as compared to the optimal procedure cannot exceed 2.

4.4. Lower bound for the value of the optimal procedure. In the preceding subsection we proved a theorem which presented a search method $SU_2(n)$ whose value behaves like $(n/2)/2^n$; we shall now show that any search procedure $S_2(n)$ satisfies

$$(47) \quad V(S_2(n)) > \frac{n/8}{2^n}.$$

Since the asymptotic amount of reduction (of the value) per observation tends to $\frac{1}{2}$ and since the upper and the lower bounds differ by a factor of $(\frac{1}{2})^2$, it will follow that the

number of wasted observations of $SU_2(n)$, in comparison with the optimal procedure, is at most 2. Practically, no more than one observation is wasted as is shown in Table 3.

THEOREM 2.

$$(48) \quad G_m(a_1, a_2) \geq \frac{a_1/2 + (m/8)a_2}{2^m}.$$

(see definition (25)).

Proof. Inequality (48) clearly holds for $m = 0$; assume that it holds for $m \leq n$ and that one is faced with the situation described by Fig. 3 (i.e. the relevant intervals have lengths $a_1 \geq a_2$) having $n + 1$ additional observations. It will be proved in Proposition 2 of § 5.1 that $a_1 \geq a_2$ implies that for a minimax procedure the next observation should be made on f_1 .

Now

$$(49) \quad G_{n+1}(a_1, a_2) = \min \left[\min_{0 < d < a_2} \max(G_n(a_1 - d, a_2 - d), G_n(a_2, d)); \right. \\ \left. \min_{a_2 \leq d < a_1} \max\left(\frac{a_1 - d}{2^n}, G_n(d, a_2)\right) \right].$$

Since, for $0 < d < a_2$

$$G_n(a_1 - d, a_2 - d) \geq \frac{a_1 - d}{2^n} > \frac{a_1 - a_2}{2^n}$$

it follows that

$$(50) \quad G_{n+1}(a_1, a_2) \geq \min \left[\frac{a_1 - a_2}{2^n}, \min_{a_2 \leq d < a_1} \max\left(\frac{a_1 - d}{2^n}, G_n(d, a_2)\right) \right] \quad (\text{by the induction hypothesis})$$

$$\geq \min \left[\frac{a_1 - a_2}{2^n}, \min_{a_2 \leq d < a_1} \max\left(\frac{a_1 - d}{2^n}, \frac{d/2 + (n/8)a_2}{2^n}\right) \right].$$

The second minimand in the square brackets is equal to $(a_1 - d^*)/2^n$ where d^* is defined by

$$(51) \quad a_1 - d^* = \frac{d^*}{2} + \frac{n}{8}a_2 \quad \text{i.e.} \quad d^* = \frac{2}{3}a_1 - \frac{n}{12}a_2$$

provided $d^* \geq a_2$. The latter holds if and only if

$$(52) \quad a_1 \geq \left(\frac{n}{8} + \frac{3}{2}\right)a_2$$

which is also a necessary and sufficient condition for the second minimand in (50), $(a_1 - d^*)/2^n$, to be smaller than the first minimand $(a_1 - a_2)/2^n$. Therefore, under condition (52)

$$G_{n+1}(a_1, a_2) \geq \frac{a_1 - d^*}{2^n} = \frac{1}{2^n} \left[\frac{a_1}{3} + \frac{n}{12}a_2 \right] > \frac{1}{2^{n+1}} \left[\frac{a_1}{2} + \frac{n+1}{8}a_2 \right] \quad (\text{by (51)}).$$

If (52) does not hold we proceed as follows:

$$\begin{aligned}
 G_{n+1}(a_1, a_2) &\geq \min \left[\min_{0 < d \leq a_2} \max (G_n(a_1 - d, a_2 - d), G_n(a_2, d)); G_n(a_2, a_2) \right] \\
 (53) \qquad \qquad \qquad &\qquad \qquad \qquad \text{(by the induction hypothesis)} \\
 &\geq \frac{1}{2^n} \min \left[\min_{0 < d \leq a_2} \max \left(\frac{a_1 - d}{2} + \frac{n(a_2 - d)}{8}, \frac{a_2}{2} + \frac{nd}{8} \right); \frac{a_2}{2} + \frac{na_2}{8} \right]
 \end{aligned}$$

The first minimand in the square brackets is equal to $a_2/2 + nd^*/8$ where

$$\frac{a_1 - d^*}{2} + \frac{n(a_2 - d^*)}{8} = \frac{a_2}{2} + \frac{nd^*}{8}$$

i.e.

$$(54) \qquad \qquad \qquad d^* = \frac{(n-4)a_2 + 4a_1}{2n+4}$$

provided $d^* \leq a_2$. The latter is equivalent to

$$(55) \qquad \qquad \qquad a_1 \leq \left(\frac{n}{4} + 2 \right) a_2$$

and this is implied by the negation of (52). Moreover, (55) implies also that the first minimand in (53), $a/2 + nd^*/8$, is smaller than the second minimand $a_2/2 + na_2/8$. Therefore

$$\begin{aligned}
 G_{n+1}(a_1, a_2) &\geq \frac{1}{2^n} \left[\frac{a_2}{2} + \frac{nd^*}{8} \right] \\
 &= \frac{1}{2^{n+1}} \left[a_2 + \frac{n(n-4)a_2 + 4a_1}{2n+4} \right] \quad \text{(by (54))} \\
 &= \frac{1}{2^{n+1}} \left[\left(\frac{1}{2} - \frac{1}{n+2} \right) a_1 + \frac{n^2 + 4n + 16}{8n + 16} a_2 \right] \\
 &\geq \frac{1}{2^{n+1}} \left[\frac{a_1}{2} - \frac{(n/8 + 3/2)}{n+2} a_2 + \frac{n^2 + 4n + 16}{8n + 16} a_2 \right] \quad \text{(by (55))} \\
 &= \frac{1}{2^{n+1}} \left[\frac{a_1}{2} + \frac{n^2 + 3n + 4}{8(n+2)} a_2 \right] \\
 &\geq \frac{1}{2^{n+1}} \left[\frac{a_1}{2} + \frac{n+1}{8} a_2 \right].
 \end{aligned}$$

We conclude that (48) is valid for $m = n + 1$ whether (52) or the opposite inequality holds. Q.E.D.

Thus for any search procedure $s_2(n)$

$$V(s_2(n)) \geq \frac{n/8 + 1/2}{2^n}.$$

5. Optimization for $k > 2$.

5.1. General discussion. Consider k observable functions $f_1(x), \dots, f_k(x)$ which have unique roots in the sense of (2). The notion of "relevant intervals" of the functions during any stage of the search could be defined just as in § 4.1.

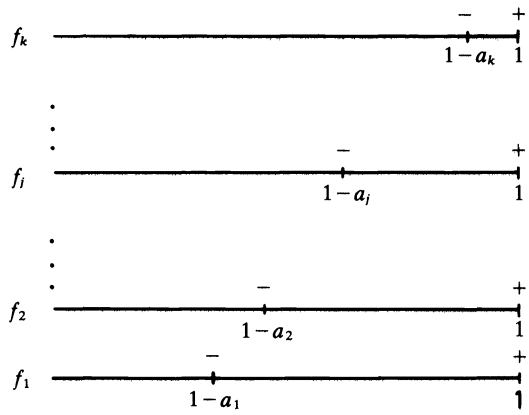


FIG. 4

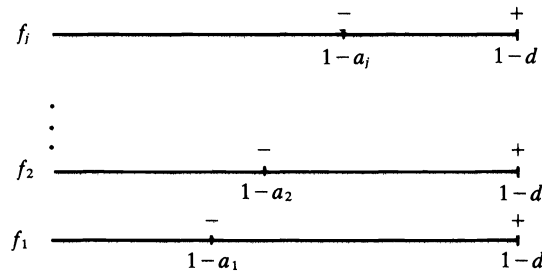


FIG. 5

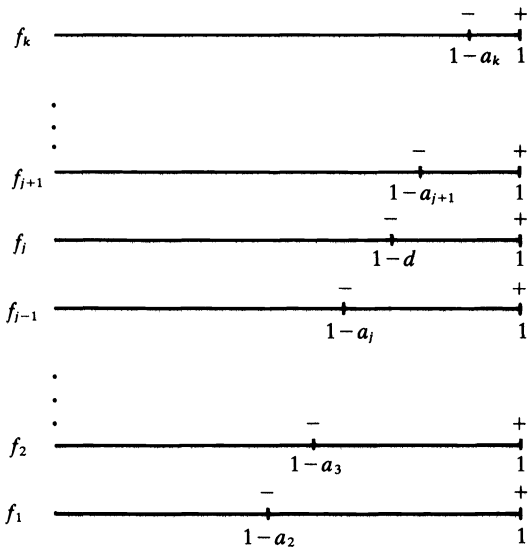


FIG. 6

Thus, it can be assumed that at any stage of the search, the situation is as depicted in Fig. 4, so that $a_1 \geq a_2 \geq \dots \geq a_k$ and for all i , $1 \leq i \leq k$, $f_i(1 - a_i) < 0$ and $f_i(1) > 0$.

(56) The length of the final interval of uncertainty obtained by a minimax strategy starting with the situation depicted in Fig. 4 and having m additional observations will be denoted by $G_m(a_1, \dots, a_k)$.

We shall use the following fact: If at the situation depicted by Fig. 4, one observes f_1 at a point $1 - d$ where d satisfies $a_{j+1} \leq d < a_j$ for some $j \in \{1, \dots, k\}$ (for convenience we define a_{k+1} to be zero) then: If $f_1(1 - d) > 0$ the situation is transformed to the one depicted in Fig. 5, while if $f_1(1 - d) < 0$ then (after re-indexing f_i according to the lengths of the relevant intervals) we obtain Fig. 6.

We shall now prove the following result which has been already used in § 4:

PROPOSITION 2. *For the situation depicted by Fig. 4 there exists a minimax procedure which first observes f_1 .*

Proof. Let $s_k(m)$ be any search procedure. Suppose that $s_k(m)$ first observes f_i , $i > 1$, at $1 - d$ and then continues in an optimal manner. If d satisfies: $a_{j+1} \leq d < a_j$; then the value of $s_k(m)$ (starting with the situation of Fig. 4) $V(s_k(m))$ is given by

$$V(s_k(m)) = \max [G_{m-1}(a_1 - d, a_2 - d, \dots, a_j - d), \\ G_{m-1}(a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_j, d, a_{j+1}, \dots, a_k)].$$

Now let $\bar{s}_k(m)$ be defined almost like $s_k(m)$, the only difference being that $\bar{s}_k(m)$ first observes $f_1(1 - d)$; then

$$(57) \quad V(\bar{s}_k(m)) = \max [G_{m-1}(a_1 - d, a_2 - d, \dots, a_j - d), \\ G_{m-1}(a_2, \dots, a_i, a_{i+1}, \dots, a_j, d, a_{j+1}, \dots, a_k)].$$

Since $G_{m-1}(x_1, \dots, x_k)$ is monotonic nondecreasing in each of its arguments and since $a_1 \geq a_2 \geq \dots \geq a_k$ it follows that

$$V(\bar{s}_k(m)) \leq V(s_k(m)). \quad \text{Q.E.D.}$$

Let us denote the right side of (57) by $R_j(d)$; then the overall optimal search procedure could be defined similarly to the case $k = 2$ by

$$G_m(a_1, \dots, a_k) = \min_{0 \leq j \leq k} \min_{a_{j+1} \leq d \leq a_j} R_j(d).$$

Note that

$$R_j(d) = \max (f^j(d), g^j(d))$$

where

$$f^j(d) \stackrel{\text{def}}{=} G_{m-1}(a_1 - d, a_2 - d, \dots, a_j - d)$$

and

$$g^j(d) \stackrel{\text{def}}{=} G_{m-1}(a_2, \dots, a_i, a_{i+1}, \dots, d, a_{j+1}, \dots, a_k)$$

and observe that f^j is a decreasing function while g^j is an increasing function. These observations enable one to evaluate $G_m(a_1, \dots, a_k)$ by a straightforward extension of Lemma 1. This extension is used in the proofs of Theorems 3 and 4.

However, for $k > 2$ (or 3) it is not practical to use the recursion formula for obtaining the optimal search procedure. Thus, we shall concentrate on finding asymptotically optimal methods which are easily computable and have good performance.

5.2. Asymptotically optimal solution. For the case $k = 2$, asymptotically optimal solutions were constructed by using a linear upper bound for $G_n(a_1, a_2)$ of the form $(1/2^n)(w_{n1}a_1 + w_{n2}a_2)$ where $w_{n1} = 1$ and w_{n2} was either defined recursively by (36) (procedure $SU_2(n)$) or was chosen to be $n/2$ (procedure $SN_2(n)$).

At first we shall present an extension of $SN_2(n)$ for $k \geq 2$. This procedure which will be denoted by $SN_k(n)$ is asymptotically optimal and can be easily implemented. It is based on the following theorem:

THEOREM 3. *For every natural number n_0 there exists $c = c(n_0)$ and $\varepsilon = \varepsilon(n_0)$ ($\varepsilon \rightarrow 0$ as $n_0 \rightarrow \infty$) such that for all $m \geq n_0$*

$$(58) \quad G_m(a_1, \dots, a_k) \leq \frac{c}{2^m} \sum_{i=1}^k w_{mi} a_i \stackrel{\text{def}}{=} Q_m(a_1, \dots, a_k)$$

where

$$(59) \quad w_{mi} = \frac{m^{i-1}}{(i-1)!(2-\varepsilon)^{i-1}}.$$

Proof. By induction: Choose c such that (58) will hold for $m = n_0$. Assume that it holds for all m , $n_0 \leq m \leq n$ and that we are in the situation depicted by Fig. 4 with $n+1$ additional observations to make. For convenience define a_{k+1} as zero and denote:

$$(60) \quad R_j^+ = \sum_{i=1}^j w_{ni}(a_i - a_j)$$

and

$$(61) \quad R_j^- = \sum_{i=2}^j w_{ni-1}a_i + w_{nj}a_j + \sum_{i=j+1}^k w_{ni}a_i.$$

Since $a_i > 0$, $w_{ni} > 0$ it follows that:

$$0 = R_1^+ < R_1^-.$$

Also, since $a_1 \geq a_2, \dots \geq a_k$, $a_{k+1} = 0$ and $w_{ni} > 0$ it follows that:

$$R_{k+1}^+ = \sum_{i=1}^k w_{ni}a_i > \sum_{i=2}^k w_{ni-1}a_i = R_{k+1}^-.$$

Thus there exists $j \in \{1, 2, \dots, k\}$ such that

$$(62) \quad R_j^+ < R_j^- \quad \text{and} \quad R_{j+1}^+ \geq R_{j+1}^-.$$

It follows from (60), (61) and (62) that there exists $d > 0$ which satisfies:

$$(63) \quad a_{j+1} \leq d < a_j$$

and

$$(64) \quad \sum_{i=1}^j w_{ni}(a_i - d) = \sum_{i=2}^j w_{ni-1}a_i + w_{nj}d + \sum_{i=j+1}^k w_{ni}a_i$$

i.e.

$$(65) \quad Q_n(a_1 - d, \dots, a_j - d) = Q_n(a_2, \dots, a_j, d, a_{j+1}, \dots, a_k)$$

where Q_n is defined by (58).

Now, if one observes f_1 at $1 - d$ where d satisfies (63) and (64), then

$$\begin{aligned} G_{n+1}(a_1, \dots, a_k) &\leq \max [Q_n(a_1 - d, \dots, a_j - d), Q_n(a_2, \dots, a_j, d, a_{j+1}, \dots, a_k)] \\ &= Q_n(a_2, \dots, a_j, d, a_{j+1}, \dots, a_k) \end{aligned}$$

Thus, we have to prove that

$$(66) \quad \frac{c}{2^n} \left(\sum_{i=2}^j w_{ni-1} a_i + w_{nj} d + \sum_{i=j+1}^k w_{ni} a_i \right) \leq \frac{c}{2^{n+1}} \sum_{i=1}^k w_{n+1,i} a_i.$$

This can be proved by substituting d by the linear function of a_1, \dots, a_k obtained from solving (64), and w_{ni} by the value presented in (59). By this substitution, both sides of (66) take the form of a linear combination of a_1, \dots, a_k where the coefficients are rational functions of n . Now, it can be shown (by reducing the coefficients of a_i in both sides of (66) to a common factor) that the nominators of the left and the right side of (66) turn out to be polynomials in n such that the highest power of n has the same coefficient while the next power of n has coefficients

$$\frac{2}{(2-\varepsilon)^{i-1}(i-1)!} \quad \text{and} \quad \frac{1+2/(2-\varepsilon)}{(2-\varepsilon)^{i-1}(i-1)!} \quad \text{respectively.}$$

Thus it is possible to choose an ε so that the difference in the coefficient of the next highest power of n will compensate for all the terms which correspond to lower powers of n . It follows that a suitable choice of ε can assure that the coefficients of a_1, \dots, a_k of the left side of (66) do not exceed those of the right side. It also follows that as n_0 grows, ε can be chosen to be close to zero. This completes the proof.

The upper bound (58) can be used as a tool for finding a search procedure $SN_k(n)$ which can guarantee that the final interval of uncertainty will not exceed this upper bound. At each step of this procedure, the observed point $1 - d$ is chosen by (64) where the weights w_{ni} are given by (59).

It follows from (58) that for $n > n_0$ the value of $SN_k(n)$ satisfies

$$\begin{aligned} V(SN_k(n)) &\leq \frac{c}{2^n} \sum_{i=1}^k \frac{n^{i-1}}{(i-1)!(2-\varepsilon)^{i-1}} < \frac{c'n^{k-1}}{2^n} \\ &= \left(\frac{1}{2}\right)^{n-(k-1)\log_2 n + c''} \end{aligned}$$

where c' and c'' are constants. Thus the procedure $SN_k(n)$ is asymptotically optimal in the sense of (13).

Let us now consider a procedure, to be denoted by $SU_k(n)$, which is the next extension for $k > 2$ of the procedure $SU_2(n)$. In this procedure which produces better results than $SN_k(n)$, the weights are not given explicitly but rather are defined recursively. It is based on the following theorem:

THEOREM 4. Let $G_m(a_1, \dots, a_k)$ be defined by (56); then for all m

$$(67) \quad G_m(a_1, \dots, a_k) \leq \frac{1}{2^m} \sum_{i=1}^k u_{mi} a_i$$

where u_{ni} are defined recursively as follows:

$$(68) \quad u_{n1} = 1 \quad \text{for all } n \geq 0; \quad u_{02} = u_{03} = \cdots = u_{0k} = 0$$

and for $n \geq 1, 2 \leq i \leq k$

$$(69) \quad u_{n+1i} = \max \left[\max_{1 \leq j < i} \left(2 - \frac{u_{nj}}{T_{nj}} \right) u_{ni}; \max_{i \leq j \leq k} \left(\left(2 - \frac{u_{nj}}{T_{nj}} \right) u_{ni-1} + \frac{u_{nj}}{T_{nj}} u_{ni} \right) \right]$$

where

$$(70) \quad T_{nj} = u_{nj} + \frac{1}{2} \sum_{l=1}^{j-1} u_{nl}.$$

Proof. The same principle used in proving Theorem 3 will also be used here. The proof will be given by induction: Since $G_0(a_1, \dots, a_k) = a_1$ it follows that (67) holds for $m = 0$. Now assume that (67) holds for $m \leq n$. If we are in the situation depicted by Fig. 4 with $n + 1$ additional observations to make, then similarly to (63) and (64) it can be shown that there exist $d > 0$ and $j \in \{1, 2, \dots, k\}$ such that:

$$(71) \quad a_{j+1} \leq d < a_j$$

and

$$(72) \quad \sum_{i=1}^j u_{ni}(a_i - d) = \sum_{i=2}^j u_{ni-1}a_i + u_{nj}d + \sum_{i=j+1}^k u_{ni}a_i.$$

By solving equation (72) for d we obtain:

$$(73) \quad d = \frac{1}{2T_{nj}} \left[a_1 + \sum_{i=2}^j (u_{ni} - u_{ni-1})a_i - \sum_{i=j+1}^k u_{ni}a_i \right]$$

where T_{nj} is defined by (70).

Now, it easily follows from (72) and the induction hypothesis that if one observes f_1 at $1 - d$ then:

$$(74) \quad \begin{aligned} G_{n+1}(a_1, \dots, a_k) &\leq \frac{1}{2^n} \left[\sum_{i=2}^j u_{ni-1}a_i + u_{nj}d + \sum_{i=j+1}^k u_{ni}a_i \right] \quad (\text{by (73)}) \\ &= \frac{1}{2^{n+1}} \left[\frac{u_{nj}}{T_{nj}} a_1 + \sum_{i=2}^j \left[\left(2 - \frac{u_{nj}}{T_{nj}} \right) u_{ni-1} + \frac{u_{nj}}{T_{nj}} u_{ni} \right] a_i \right. \\ &\quad \left. + \sum_{i=j+1}^k \left(2 - \frac{u_{nj}}{T_{nj}} \right) u_{ni}a_i \right]. \end{aligned}$$

It follows from (70) that

$$(75) \quad \frac{u_{nj}}{T_{nj}} \leq 1 = u_{n+11}.$$

Also, for $i \in \{2, \dots, j\}$, (69) implies that

$$u_{n+1i} \geq \left(2 - \frac{u_{nj}}{T_{nj}} \right) u_{ni-1} + \frac{u_{nj}}{T_{nj}} u_{ni}$$

while for $i \in \{j+1, \dots, k\}$ (69) implies that

$$u_{n+1i} \geq \left(2 - \frac{u_{nj}}{T_{nj}} \right) u_{ni}.$$

Thus it follows from (74) and (75) that

$$G_{n+1}(a_1, \dots, a_k) \leq \frac{1}{2^{n+1}} \sum_{i=1}^k u_{n+1,i} a_i. \quad \text{Q.E.D.}$$

Theorem 4 can be used to define the procedure $SU_k(n)$ which is based on the following principle: When $n+1$ observations are left, one has to order the relevant intervals of the functions in a descending order $a_1 \geq a_2 \geq \dots \geq a_k$ and to observe the function which corresponds to a_1 , at $1-d$ where d is given by (73), and the weights u_{ni} are defined recursively by (68), (69) and (70). In order to find the suitable j one can use the process described at the beginning of the proof of Theorem 3.

Using the relations (68), (69) and (70) it will be proved in the Appendix that $u_{ni} \leq c_i n^{i-1}$ where c_i are constants, so that

$$(76) \quad V(SU_k(n)) \leq \frac{1}{2^n} \sum_{i=1}^k c_i n^{i-1} \leq \frac{cn^{k-1}}{2^n}$$

which implies that $SU_k(n)$ is asymptotically optimal in the sense of (13).

It was found by numerical calculation (for $k \leq 100$, $n \leq 300$) that starting with the initial conditions (68), the maximum in expression (69) has been always attained at $j = i$; thus $u_{n+1,i}$ can be defined more simply by

$$(77) \quad u_{n+1,i} = \left[\left(2 - \frac{u_{ni}}{T_{ni}} \right) u_{ni-1} + \frac{u_{ni}}{T_{ni}} u_{ni} \right]$$

where T_{ni} is given by (70).

If the recursive definition (77) is used for constructing the weights u_{ni} then these weights are independent of the total number of constraints k . Thus, for all k_0 and n_0 one can build a file of weights which has $k_0 n_0$ numbers and this file will be used by the procedures $SU_k(n)$ for all $k \leq k_0$ and $n \leq n_0$.

Theorem 4 presents an upper bound for $V_k(n)$ to be denoted by SU .

$$SU = \frac{1}{2^n} \sum_{i=1}^n u_{ni}.$$

We now present two tables which compare SU to the value obtained by using simplified procedures which will be denoted by SP . In Table 4 we compare SU to SP for $k = 3, 5, 10, 20$; and $n = 3k, 4k$ and $5k$. It should be kept in mind that the values which can be obtained by the naive procedure (separate bisection) for these case are $\frac{1}{8}$, $\frac{1}{16}$ and $\frac{1}{32}$ respectively. These tables demonstrate the fact that the advantage of the procedure $SU_k(n)$ becomes much more significant as n or k grows. It should also be noted that SP

TABLE 4
A comparison between the upper bound given by Theorem 4— SU —and the simplified procedures— SP .

$k \backslash n$	3		5		10		20	
	SU	SP	SU	SP	SU	SP	SU	SP
3k	7.1E-2	5.2E-2	3.0E-2	2.9E-2	2.7E-3	1.1E-2	1.7E-5	3.6E-3
4k	1.3E-2	1.7E-2	2.1E-3	7.1E-3	1.9E-5	1.8E-3	1.4E-9	3.7E-4
5k	2.2E-3	5.3E-3	1.3E-4	1.8E-3	9.5E-8	3.0E-4	4.7E-14	4.0E-5

gives the exact value for the simplified procedures while SU gives an upper bound for the performance of $SU_k(n)$. The numerical experiments which were made using the procedure $SU_k(n)$ show that its performance is better than the upper bound SU; these experiments are described in the next section. In Table 5, SU and SP are compared for $k = 100$.

TABLE 5
Comparison between SU and SP for $k = 100$.

n	SU	SP
150	3.8E-2	1.9E-2
160	4.5E-3	1.6E-2
170	4.2E-4	1.4E-2
180	3.2E-5	1.3E-2
190	2.1E-6	1.1E-2
200	1.1E-7	9.7E-3
210	4.9E-9	5.6E-3
220	1.9E-10	2.8E-3
230	6.6E-12	1.6E-3
240	2.0E-13	1.0E-3
250	5.3E-15	6.8E-4

6. Numerical evaluation of the asymptotically optimal procedures for $k > 2$.

6.1. General discussion. For the two-constraints case, $k = 2$, it was possible to calculate numerically the values of the asymptotically optimal procedures, $V(SU_k(n))$ and $V(SN_k(n))$ by recursion formulas. This method is practically infeasible as k gets bigger whereas the value SU based on Theorem 4 and presented in Tables 4 and 5 is only an upper bound for $V(SU_k(n))$.

In order to get a practical estimation for the performance of the procedures $SU_k(n)$ and $SN_k(n)$, we used Monte Carlo methods where the roots z_i , $i = 1, \dots, k$ of the observed functions were chosen as random variables. We used two models for producing z_1, \dots, z_k :

Model I. z_i , $i = 1, \dots, k$ were chosen to be k independent random variables each of which is uniformly distributed in $(0, 1)$.

Model II. $z_1 = z_2 = \dots = z_k = z$ where z is uniformly distributed in $(0, 1)$.

The first model represents a "natural" case while Model II is an extreme case when the number of relevant constraints does not decrease during the search process. In fact, Model II contains configurations which correspond to the worst case for the simplified procedures as can easily be deduced from the analysis presented in § 3.2.

Given k , n and the search procedure, (which was either $SU_k(n)$ or $SN_k(n)$ or certain variations of them) fifty lotteries were executed for each of the models and both the maximal and the average value of the interval of uncertainty were recorded. The numerical results which appear in the tables in the next subsections correspond to the worst case among those two models; for $SU_k(n)$ this was always Model I, while for $SN_k(n)$ the model which gave the more pessimistic results was normally Model II (except for the maximal length of the interval for $K = 5$, $n = 15$ and $n = 20$).

In testing the search procedures $SU_k(n)$ and $SN_k(n)$ we used a simple modification which is sometimes very useful: Before making the observations and while having in mind an estimate of the desired accuracy needed for the location of the minimal root, we find that it is advantageous to define a tolerance interval, TOL, which is one or two

orders of magnitude smaller than the desired accuracy. Then, if during the search process, the relevant interval of one or several functions is smaller than TOL, we omit these functions in the subsequent calculation. This modification increases the efficiency of the search procedures when k is large. It is easy to see that if one uses this modification and obtains a final interval of uncertainty (R_1, R_2) then the minimal root z must lie in $(\min(R_1, R_2 - \text{TOL}), R_2)$. Thus if $R_2 - R_1 \geq \text{TOL}$ then the final interval of uncertainty remains the same while if $R_2 - R_1 < \text{TOL}$ then the final interval of uncertainty is $(R_2 - \text{TOL}, R_2)$. The numerical results are presented in the next subsection.

6.2. Numerical results. Table 6 presents the maximal (MAX) and average (AV) value of the final interval of uncertainty for $SU_k(n)$ which were obtained by the lotteries described in § 6.1, for the number of constraints which correspond to those considered in Tables 4 and 5; namely $k = 3, 5, 10, 20$ and 100 .

TABLE 6
The interval of uncertainty of $SU_k(n)$.

$k \backslash n$	3		5		10		20		100	
	MAX	AV	MAX	AV	MAX	AV	MAX	AV	MAX	AV
1.5k									3.0E-4	1.5E-4
2k							7.8E-3	2.5E-3	1.2E-8	6.7E-9
3k	5.4E-2	3.4E-2	1.9E-2	1.0E-2	1.4E-3	7.0E-4	7.2E-6	5.0E-6		
4k	1.1E-2	7.7E-3	1.5E-3	1.1E-3	1.2E-5	8.9E-6	8.1E-10	7.0E-10		
5k	1.9E-3	1.5E-3	1.0E-4	8.0E-5	6.8E-8	6.0E-8				

By comparing Table 6 to Tables 4 and 5 it is seen that $SU_k(n)$ usually produces results which are better (and for large k even much better) than the upper bound of Theorem 4.

As we have already noted before, the results are very satisfactory. For example, if there are 100 constraints then by only 2 observations per constraint, it is possible to locate z in accuracy of order 10^{-8} . The simplified procedure guarantees only 10^{-2} and a naive approach only 0.25.

We shall now describe the results obtained by using procedures which are based on Theorem 3. Recall that this theorem gives an asymptotic upper bound which need not hold for small n . It does define a procedure $SN_k(n)$ which uses weights w_{ni} which are approximately

$$(78) \quad w_{ni} = \frac{n^{i-1}}{(i-1)! 2^{i-1}}$$

and calculates the observation point $1-d$ using formula (64).

We made experiments with $SN_k(n)$ and several modifications of it. A modification which was usually quite effective is described as follows: Instead of considering all the weights w_{ni} , we tried to concentrate on the four most significant weights. For $n \geq 2k - 2$ it follows from (78) that $w_{nk} > w_{nk-1} > \dots > w_{n1}$ and thus these four weights were chosen to be w_{nk} , w_{nk-1} , w_{nk-2} and w_{nk-3} ; for $n \leq 2k - 2$ the maximal weight is w_{nj} where $J = [(n+2)/2]$; thus in this case w_{nJ} , w_{nJ-1} , w_{nJ-2} and w_{nJ-3} were used. All the

other weights were neglected (i.e. taken to be zero). The calculation of d in formula (64) depends only upon relative weights which are in this case:

$$(79) \quad 1, \quad \frac{2(i-1)}{n}, \quad \frac{4(i-1)(i-2)}{n^2}, \quad \text{and} \quad \frac{8(i-1)(i-2)(i-3)}{n^3}.$$

where i is equal to k or to J .

The advantage of using only a small number of weights (in this case four weights) is that the calculations of the observation points are quicker. Also, since the weights are given simply by (79), one does not have to keep the file of weights needed by the method $SU_k(n)$.

The reason for choosing only the weights corresponding to the intervals a_k, a_{k-1}, a_{k-2} and a_{k-3} was the experimental fact that for $k \leq 20$ the distance between the observation point and 1, d satisfies in most cases $d \leq a_{k-3}$. It should be noted that the intuitive reason for considering d only in the interval $0 < d \leq a_i$, say $i = k - 3$, instead of $0 < d \leq a_1$ is that in the case where the optimal d is bigger than a_{k-3} , if we choose $d = a_{k-3}$ instead of the optimal d , then (as can be easily seen) the worst case would be $f_1(a_{k-3}) > 0$, which means that even for this case this observation will eliminate the four functions f_k, f_{k-1}, f_{k-2} and f_{k-3} in the subsequent stages of the search. The results for this modification of $SN_k(n)$ are presented in Table 7.

TABLE 7
The interval of uncertainty of $SN_k(n)$ (modified).

$n \backslash k$	3		5		10		20		100	
	MAX	AV	MAX	AV	MAX	AV	MAX	AV	MAX	AV
1.5k									2.0E-1	6.7E-2
2k							3.4E-2	1.5E-2	1.1E-3	2.7E-4
3k	6.5E-2	3.6E-2	2.0E-2	8.7E-3	1.7E-3	1.1E-3	2.6E-5	1.0E-5		
4k	1.4E-2	6.3E-3	1.9E-3	6.0E-4	1.5E-5	4.2E-6	6.4E-10	1.4E-10		
5k	1.9E-3	1.2E-3	8.3E-5	4.8E-5	5.1E-8	2.1E-8				

By comparing Table 7 to Table 6 it is seen that $SU_k(n)$ almost always produces better results in MAX than $SN_k(n)$. If $k \leq 20$ and $n \geq 3k$ then there is not a big difference between the performance of the two methods. However, for $k = 100$ the performance of $SU_k(n)$ is much better than this version of $SN_k(n)$. Some other versions of $SN_k(n)$ gave better results for $k = 100$ but they too were inferior to $SU_k(n)$.

It should be noted that when evaluating the performance of search procedure $s_k(n)$ one should also take into account the time required by $s_k(n)$ to calculate the observation point. For example, a search procedure that takes into account only a small portion of the (most significant) weights may be much quicker than the procedure $SU_k(n)$. Thus, there is still more to be done in finding some variations of $SU_k(n)$ or $SN_k(n)$ which will work quickly and will also perform satisfactorily for large k .

7. Randomized procedures. The search procedures considered so far were pure procedures, that is: For each stage, the next observation point is a deterministic function of the results obtained in the prior observations. The question which is now being

considered is: Can randomization substantially improve the performance of the search methods?

A randomized procedure is defined as the probabilistic mixture of several pure procedures (of the type considered before). The performance of such a randomized procedure is measured by the expected length of the final interval of uncertainty for the worst possible choice of z_1, \dots, z_k .

In the simple case of one constraint when one can make n observations then the length of the interval of uncertainty guaranteed by bisection (which is the minimax pure strategy) is $(\frac{1}{2})^n$, and it can be easily seen that this value cannot be improved by any randomization method.

Now let us look at the case $k > 1$. If one uses the naive approach of separately bisecting each constraint, then in § 2 it was noted that if a certain accuracy V is needed and

$$(80) \quad n_0 = -\log_2 V$$

then to get the same accuracy for $k > 1$ one has to make $k \cdot n_0$ observations. It is possible to reduce this number by using the following randomization: Choose the first function f_j to be bisected randomly with probability of $1/k$ for each function and make n_0 observations on it. When this state is finished it is known that the root of f_j lies in an interval $(x, x + V)$. Next choose randomly the next function f_i to be bisected and observe f_i at x . If $f_i(x) \leq 0$ then discard it and choose randomly one among the remaining functions, while if $f_i(x) > 0$ make n_0 observations on it. This process is continued until all the functions have been observed at least once. It is easy to see that the procedure just described uses an expected number of observation which is about $((k + 1)/2)n_0$ instead of $k \cdot n_0$.

It turns out that the optimal search method for $k > 1$ can also be improved by randomization as will be illustrated by the following simple example: Assume that $k = 2$ and $n = 3$. It is easily seen that the value guaranteed by a pure procedure is greater or equal to $\frac{1}{3}$. On the other hand it is possible to obtain an expected value of $\frac{3}{10}$ by the following randomized procedure:

Choose the index of the first function to be observed j_1 (see (4)) such that $\Pr(j_1 = 1) = \Pr(j_1 = 2) = \frac{1}{2}$. Denote the chosen function by f_{j_1} and the other function by \bar{f}_{j_1} . At first, observe f_{j_1} at $x = \frac{4}{10}$; if $f_{j_1}(\frac{4}{10}) > 0$ then observe both f_{j_1} and \bar{f}_{j_1} at $x = \frac{2}{10}$. (In the notation of (4), $j_2 = j_1$, $x_2 = \frac{2}{10}$, $j_3 = 3 - j_1$, $x_3 = \frac{2}{10}$.)

If $f_{j_1}(\frac{4}{10}) \leq 0$ then observe \bar{f}_{j_1} at $\frac{7}{10}$; if $\bar{f}_{j_1}(\frac{7}{10}) > 0$ then next observe $\bar{f}_{j_1}(\frac{4}{10})$ while if $\bar{f}_{j_1}(\frac{7}{10}) \leq 0$ then next observe $f_{j_1}(\frac{7}{10})$. (Again, one could use the notation of (4) but it seems unnecessary to write it explicitly.)

Call this procedure S (S is a mixture of two pure procedures each with probability $\frac{1}{2}$) and denote its expected value of the interval of uncertainty by $EV(S)$. Now consider the three possible configurations for the roots z_1, z_2 of f_1 and f_2 :

1. $z_1, z_2 \leq \frac{4}{10} \rightarrow V(S) = \frac{2}{10}$,
2. $z_1, z_2 \geq \frac{4}{10} \rightarrow V(S) = \frac{3}{10}$,
3. $z_1 < \frac{4}{10} < z_2$ or $z_2 < \frac{4}{10} < z_1 \rightarrow EV(S) = \frac{1}{2} \cdot \frac{2}{10} + \frac{1}{2} \cdot \frac{4}{10} = \frac{3}{10}$.

It follows that randomization can improve the performance of the optimal pure procedure. However, we do not know the amount which can be gained and whether it substantially improves the asymptotic behavior of the interval of uncertainty. This is a point for further research.

8. Points for further research. The methods presented in this paper for finding the maximal range of validity of many constraints perform very satisfactorily. However, some problems of theoretical or practical interest are still open. The problems that we consider as important are listed below:

A. *Lower bound for $k > 2$.* For $k = 2$, Theorems 1 and 2 establish the fact that if the interval of uncertainty is given by $(\frac{1}{2})^{n-l_2(n)}$ then the number of “lost observations” $l_2(n)$ behaves like $\log_2 n$. For $k > 2$, Theorem 3 shows that $l_k(n) \leq (k-1) \log_2 n + \text{const.}$; it follows from Theorem 2 that $l_k(n) > l_2(n) \geq \log_2 n + \text{const.}$, but we could not prove that $l_k(n) \geq (k-1) \log_2 n + \text{const.}$

B. *Further research on the procedure $SU_k(n)$.* Theorem 4 establishes an upper bound for the interval of uncertainty and presents the search procedure $SU_k(n)$ which was later found to be effective. It is of interest to investigate the behavior of the weights u_{ni} defined by (69) and to show that starting with the initial conditions (68), it is sufficient to use (77) for defining the weights. In addition, it may be important to investigate more deeply the performance of the method $SU_k(n)$ and try to further improve it (if possible).

C. *Randomization.* It remains an open problem to investigate the amount of improvement which can be achieved by using randomized procedures.

In addition there are some further extensions which were not considered in the paper such as:

D. The case of different prices of observations for each of the functions.

E. The case when one wishes to find the maximum of a function $f_0(x)$ $0 \leq x \leq 1$ subject to: $f_i(x) \leq 0$, $i \in \{1, \dots, k\}$ without calculating the derivative of $f_0(x)$.

F. The case when the functions f_i have additional regularity conditions besides (2).

G. The case when the statistical behavior of the roots z_i is known.

To conclude: It seems to us that a wide field of problems of both theoretical and practical interest has yet to be investigated.

Appendix. The objective of this Appendix is to prove that the weights u_{ni} , $1 \leq i \leq k$, defined in Theorem 4 satisfy

$$(A.1) \quad u_{ni} \leq C_i n^{i-1}$$

which would imply that the procedure $SU_k(n)$ is asymptotically optimal.

For convenience, we shall re-write the definition of these weights: The initial conditions are given by

$$(A.2) \quad u_{n1} = 1 \quad \text{for all } n \geq 0 \quad \text{and} \quad u_{02} = u_{03} = \dots = u_{0k} = 0$$

and the recursion equation for $n \geq 1$, $2 \leq i \leq k$, is

$$(A.3) \quad u_{n+1i} = \max \left[\max_{1 \leq j < i} \left(2 - \frac{u_{nj}}{T_{nj}} \right) u_{ni}; \max_{i \leq j \leq k} \left(\left(2 - \frac{u_{nj}}{T_{nj}} \right) u_{ni-1} + \frac{u_{nj}}{T_{nj}} u_{ni} \right) \right]$$

where

$$(A.4) \quad T_{nj} = u_{nj} + \frac{1}{2} \sum_{l=1}^{j-1} u_{nl}.$$

In order to show that (A.1) holds we shall prove the following proposition:

PROPOSITION A.1. *There exist two positive sequences b_i , B_i , $2 \leq i \leq k$, and a natural number n_0 ($n_0 = n_0(k)$) so that for all $2 \leq i \leq k$ and $n \geq n_0$*

$$(A.5) \quad \frac{u_{ni}}{u_{ni-1}} \geq B_i n$$

and

$$(A.6) \quad \frac{u_{ni}}{u_{ni-1}} \leq b_i n.$$

Proposition A.1 and (A.2) imply that for all $n \geq n_0$:

$$u_{ni} = \frac{u_{ni}}{u_{ni-1}} \frac{u_{ni-1}}{u_{ni-2}}, \dots, \frac{u_{n2}}{u_{n1}} u_{n1} \leq b_2, \dots, b_i n^{i-1}$$

which is equivalent to (A.1).

Proof of Proposition A.1. We shall use induction on i : If $i = 2$ then by (A.2), (A.3) and (A.4)

$$(A.7) \quad \begin{aligned} u_{n+12} &= \max \left[u_{n2}; \max_{2 \leq j \leq k} \left(\left(2 - \frac{u_{nj}}{T_{nj}} \right) + \frac{u_{nj}}{T_{nj}} u_{n2} \right) \right] \\ &\geq \left(2 - \frac{u_{n2}}{T_{n2}} \right) + \frac{u_{n2}^2}{T_{n2}} = \frac{u_{n2}^2 + u_{n2} + 1}{u_{n2} + 1/2} \\ &\geq u_{n2} + \frac{1}{2}; \end{aligned}$$

thus $u_{n2} > n/2$ which proves (A.5) for $i = 2$.

On the other hand it follows from (A.4) that $0 < u_{nj}/T_{nj} \leq 1$; thus (A.7) implies that

$$u_{n+12} \leq \max(u_{n2}, 1) + 1$$

so that $u_{n2} \leq n$ which proves (A.6) for $i = 2$.

Assume that (A.5) and (A.6) hold for $i \leq l$. This implies that there exist two positive sequences c_i and C_i such that for all $i \leq l$ and $n \geq n_0(l)$

$$(A.8) \quad C_i n^{i-1} \leq u_{ni} \leq c_i n^{i-1}.$$

We shall now prove (A.5) for $i = l+1$: Choose any ε , $0 < \varepsilon < 1$ and an n_1 , $n_1 \geq n_0(\varepsilon)$ so that for all $n \geq n_1$

$$(A.9) \quad \sum_{j=1}^{l-1} u_{nj} < \varepsilon u_{nl};$$

such an n_1 exists by (A.8). It follows from (A.3) that

$$\begin{aligned} u_{n+1l+1} &\geq \left(2 - \frac{u_{nl+1}}{T_{nl+1}} \right) u_{nl} + \frac{u_{nl+1}^2}{T_{nl+1}} \quad (\text{by (A.4)}) \\ &= \frac{u_{nl} \sum_{j=1}^{l+1} u_{nj} + u_{nl+1}^2}{u_{nl+1} + \frac{1}{2} \sum_{j=1}^l u_{nj}} \\ &= \frac{(u_{nl+1} + \frac{1}{2} u_{nl})(u_{nl+1} + \frac{1}{2} \sum_{j=1}^l u_{nj}) + \frac{3}{4} u_{nl} \sum_{j=1}^l u_{nj} - \frac{1}{2} u_{nl+1} \sum_{j=1}^{l-1} u_{nj}}{u_{nl+1} + \frac{1}{2} \sum_{j=1}^l u_{nj}} \quad (\text{by (A.9)}) \\ &\geq u_{nl+1} + \frac{1-\varepsilon}{2} u_{nl}. \end{aligned}$$

Thus, for all $n \geq n_1$

$$(A.10) \quad u_{n+1l+1} \geq u_{n_1l+1} + \frac{1-\varepsilon}{2} \sum_{m=n_1}^n u_{ml} \geq \frac{1-\varepsilon}{2} C_l \sum_{m=n_1}^n m^{l-1} \quad (\text{by (A.8)}).$$

Hence, there exists a positive number C_{l+1} so that for all $n \geq n_1$

$$(A.11) \quad u_{n+l+1} \geq C_{l+1} n^l.$$

It follows from (A.8) and (A.11) that

$$\frac{u_{n+l+1}}{u_{nl}} \geq \frac{C_{l+1} n^l}{c_l n^{l-1}} = B_{l+1} n$$

where $B_{l+1} = C_{l+1}/c_l$ so that (A.5) holds for $n \geq n_1$.

At last we prove (A.6) for $i = l + 1$. This will be done by induction on n . We choose a large enough integer $n_2 \geq n_0(l)$ so that for all $n \geq n_2$

$$(A.12) \quad \frac{\sum_{j=1}^{l-2} u_{nj}}{u_{n+l-1}} < \frac{B_l}{2},$$

such an n_2 exists by (A.8).

Next we choose a positive number b_{l+1} which satisfies $b_{l+1} \geq b_l$ and also

$$(A.13) \quad \frac{u_{m+l+1}}{u_{ml}} \leq b_{l+1} m$$

for $m = n_2$.

Assume that (A.13) is satisfied for $n_2 \leq m \leq n$; we shall show that it holds for $m = n + 1$ by considering the following three possible cases:

Case 1.

$$u_{n+1+l+1} = \left(2 - \frac{u_{nj}}{T_{nj}}\right) u_{nl} + \frac{u_{nj}}{T_{nj}} u_{n+l+1} \quad \text{for some } j, l+1 \leq j \leq k.$$

Since $j \geq l$ it follows from (A.3) that

$$u_{n+1+l} \geq \left(2 - \frac{u_{nj}}{T_{nj}}\right) u_{n+l-1} + \frac{u_{nj}}{T_{nj}} u_{nl}.$$

Since $0 \leq u_{nj}/T_{nj} \leq 1$ it follows that

$$\begin{aligned} \frac{u_{n+1+l+1}}{u_{n+1+l}} &\leq \max\left(\frac{u_{nl}}{u_{n+l-1}}, \frac{u_{n+l+1}}{u_{nl}}\right) && \text{(by (A.6) and (A.13))} \\ &\leq \max(b_l n, b_{l+1} n) = b_{l+1} n < b_{l+1}(n+1). \end{aligned}$$

Case 2.

$$u_{n+1+l+1} = \left(2 - \frac{u_{nj}}{T_{nj}}\right) u_{n+l+1} \quad \text{for some } j, 1 \leq j \leq l-1.$$

Since $u_{n+1+l} \geq (2 - u_{nj}/T_{nj}) u_{nl}$ it follows that

$$\frac{u_{n+1+l+1}}{u_{n+1+l}} \leq \frac{u_{n+l+1}}{u_{nl}} \leq b_{l+1} n < b_{l+1}(n+1).$$

Case 3.

$$u_{n+1+l+1} = \left(2 - \frac{u_{nl}}{T_{nl}}\right) u_{n+l+1}.$$

Using (A.3) again we obtain $u_{n+1l} \geq (2 - u_{nl}/T_{nl})u_{nl-1} + u_{nl}^2/T_{nl}$ so that:

$$\begin{aligned}
 \frac{u_{n+1l+1}}{u_{n+1l}} &\leq \frac{(2T_{nl} - u_{nl})u_{nl+1}}{(2T_{nl} - u_{nl})u_{nl-1} + u_{nl}^2} \\
 &= \frac{u_{nl+1}}{u_{nl-1} + u_{nl}^2 / \sum_{j=1}^l u_{nj}} && \text{(by (A.4))} \\
 &\leq \frac{u_{nl+1}}{u_{nl-1} + u_{nl}^2 / (u_{nl} + (1 + B_l/2)u_{nl-1})} && \text{(by (A.12))} \\
 &= \frac{u_{nl+1}}{u_{nl}} \frac{1}{(u_{nl-1}/u_{nl}) + 1 / (1 + (1 + B_l/2)(u_{nl-1}/u_{nl}))} \\
 &\leq \frac{u_{nl+1}}{u_{nl}} \frac{1}{(u_{nl-1}/u_{nl}) + 1 - (1 + B_l/2)(u_{nl-1}/u_{nl})} \\
 &= \frac{u_{nl+1}}{u_{nl}} \frac{1}{1 - (B_l/2)(u_{nl-1}/u_{nl})} \\
 &\leq \frac{u_{nl+1}}{u_{nl}} \frac{1}{1 - 1/2n} && \text{(by (A.5) for } i = l) \\
 &\leq \frac{n+1}{n} \frac{u_{nl+1}}{u_{nl}} \leq \frac{n+1}{n} b_{l+1}n && \text{(by (A.13))} \\
 &= b_{l+1}(n+1).
 \end{aligned}$$

Thus (A.6) holds for $i = l+1$ and $n \geq n_2$. It follows that both (A.5) and (A.6) hold for $i = l+1$ and $n \geq n_0(l+1) = \max(n_1, n_2)$. Q.E.D.

REFERENCES

- [1] J. H. BEAMER AND D. J. WILDE, *A minimax search plan for constrained optimization problems*, J. Optimization Theory Appl., 12 (1973), pp. 439–446.
- [2] S. GAL, *Multidimensional minimax search for a maximum*, SIAM J. Appl. Math., 23 (1972), pp. 513–526.
- [3] S. GAL AND W. L. MIRANKER, *Sequential and parallel search for finding a root*, Tech. Rep. 30, IBM Israel Scientific Center, Haifa, Israel, 1975.
- [4] J. KIEFER, *Optimum sequential search and approximation methods under minimum regularity assumptions*, SIAM J. Appl. Math., 5 (1957), pp. 105–136.
- [5] O. L. MANGASARIAN, *Dual feasible direction algorithms*, Techniques of Optimization, A. Balakrishnan, ed., Academic Press, New York, 1972, pp. 67–88.
- [6] Y. MILMAN, *Search problems in optimization theory*, Ph.D. thesis, Technion–Israel Institute of Technology, Haifa, 1972.
- [7] D. M. TOPKINS AND A. F. VEINOTT, *On the convergence of some feasible direction algorithms for nonlinear programming*, this Journal, 5 (1967), pp. 268–279.
- [8] G. ZOUTENDIJK, *Methods of Feasible Directions*, Elsevier, Amsterdam, 1960.
- [9] ———, *Nonlinear programming: A numerical survey*, this Journal, 4 (1966), pp. 194–210.

ON A DIRECT ALGORITHM FOR NONLINEAR COMPLEMENTARITY PROBLEMS*

G. J. HABETLER† AND M. M. KOSTREVA‡

Abstract. We extend Murty's scheme for solving linear complementarity problems to the generalized nonlinear complementarity problem on an orthant of R^n . Reformulating to accommodate a type of nonlinear function generalizing the P -matrices of Fiedler and Pták, we present an algorithm which is simple, efficient and does not require monotonicity for convergence.

1. Introduction. In 1974, Murty [17] introduced a scheme for solving the linear complementarity problem. He showed that the scheme would work when the matrix involved in the problem was a P -matrix. In this paper we extend the scheme to nonlinear complementarity problems, showing that the scheme will work when the function involved is a nondegenerate P -function.

Given a function $f: R^n \rightarrow R^n$, the *complementarity problem associated with f* is: Find $x \in R^n$ such that for each index $i \in N = \{1, 2, \dots, n\}$

$$(1.1) \quad x_i \geq 0,$$

$$(1.2) \quad f_i(x) \geq 0,$$

$$(1.3) \quad x_i \cdot f_i(x) = 0.$$

We refer to (1.1)–(1.3) by using the symbol CPf.

In this paper we will be interested in a direct algorithm for solving the complementarity problem (1.1)–(1.3) generalized to any orthant \mathcal{O} in R^n [6], [7] i.e.: Find $x \in R^n$ such that

$$(1.4) \quad x \in \mathcal{O}, \quad f(x) \in \mathcal{O}^* = \mathcal{O}, \quad x^t f(x) = 0.$$

Condition (1.3) is referred to as the complementarity condition. Note that if x and $f(x) \in \mathcal{O}$, then the complementarity condition is equivalent to $x^t f(x) = 0$.

DEFINITION 1.1. If $y = f(x)$, CPf is *nondegenerate* if for each $x \in R^n$, at most n of the $2n$ variables (y, x) simultaneously vanish.

DEFINITION 1.2. We say that $s \in R^n$ is a *complementary point of f* if there exists an $x \in R^n$ satisfying the complementarity condition and $s = x + f(x)$. For an x satisfying the complementarity condition we say $s = x + f(x)$ is the *complementary point associated with x* . A subset $I \subseteq N$ will be called an *index set*. We say that an index set I *leads to a complementary point s* if there exists an x satisfying the complementarity condition such that $x_i = 0$ for $i \in N - I$, $f_i(x) = 0$ for $i \in I$, and s is the complementary point associated with x .

By (1.3) there exists a maximal index set $I_x \subseteq N$ so that $f_i(x) = 0$ for $i \in I_x$ and $x_i = 0$ for $i \in N - I_x$, and then $s_i = x_i$ for $i \in I_x$ and $s_i = f_i(x)$ for $i \in N - I_x$. If CPf is nondegenerate, so is (1.4), and I_x is the only index set associated with x .

In the algorithm discussed in § 3 we move from index set to index set (and so from complementary point to complementary point), until we find an index set I^* which leads to a complementary point s^* in \mathcal{O} .

* Received by the editors February 4, 1977, and in revised form August 4, 1977. This research was supported in part by the National Science Foundation under Grant MCS 76-06958.

† Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, New York 12181.

‡ Department of Mathematics, University of Maine at Orono, Orono, Maine 04473.

2. P -functions. Following Fiedler and Pták [5] we have

DEFINITION 2.1. An $n \times n$ matrix is said to be a P -matrix if all its principal minors are positive.

An extension to nonlinear functions was given by Moré and Rheinboldt [16] when they defined P -functions.

DEFINITION 2.2. If $f: R^n \rightarrow R^n$, then f is a P -function on a set S if for all $x, y \in S$ with $x \neq y$, there exists an index $i = i(x, y)$ such that $(x_i - y_i)(f_i(x) - f_i(y)) > 0$.

Note that P -functions must be one-to-one, but need not be continuous. It should be observed that the addition of any constant vector to such a function will not alter its being a P -function on S . The following theorem contains an important geometric insight for P -functions, which appeared for the special case $\mathcal{O} = R_+^n$ in Moré [14].

THEOREM 2.3. Let $f: R^n \rightarrow R^n$ be a P -function on R^n . Then (i) each orthant $\mathcal{O} \subseteq R^n$ contains at most one complementary point s , (ii) for each complementary point s there is a unique x such that s is the complementary point associated with x , (iii) if the index sets $I^{(k_1)}$ and $I^{(k_2)}$ lead to the same complementary point s , then $s_i = 0$ for

$$i \in [I^{(k_1)} \cup I^{(k_2)}] - [I^{(k_1)} \cap I^{(k_2)}].$$

Proof. For (i) assume two complementary points s and t in \mathcal{O} : $s = x + f(x)$ and $t = y + f(y)$. Note that from the complementarity condition $x, y, f(x)$, and $f(y)$ all lie in \mathcal{O} . Then, for each $i \in N$, $(x_i - y_i)(f_i(x) - f_i(y)) = -x_i f_i(y) - y_i f_i(x) \leq 0$. Since f is a P -function we must have $x = y$ and so $s = t$. To prove (ii) let $s = x + f(x) = y + f(y)$. Following the same proof as in (i) we find $x = y$. To prove (iii) we note from (ii) that the same x must be used with the two index sets. So $x_i = 0$ for $i \in (N - I^{(k_1)}) \cup (N - I^{(k_2)})$ and $f_i(x) = 0$ for $i \in I^{(k_1)} \cup I^{(k_2)}$. Thus $s_i = x_i + f_i(x) = 0$ for $i \in I^{(k_1)} \cap (N - I^{(k_2)})$ and $i \in I^{(k_2)} \cap (N - I^{(k_1)})$.

DEFINITION 2.4. Let $f: R^n \rightarrow R^n$. Then for each index set $I^{(k)}$, $k = 1, 2, \dots, 2^n$, we define $f^{(k)}$ by

$$(2.1) \quad f_i^{(k)}(x) = \begin{cases} f_i(x), & i \in I^{(k)}, \\ x_i, & i \in N - I^{(k)}. \end{cases}$$

It is obvious that x satisfies the complementarity condition if and only if there exists a k such that

$$(2.2) \quad 0 = f^{(k)}(x).$$

It is also fairly easy to see that a given index set does not necessarily lead to a complementary point. For example, if we take f to be the P -function

$$(2.3) \quad f(x_1, x_2, x_3) = \begin{pmatrix} x_1 + x_2 - 1 \\ -\exp(-x_2) + x_3 - 3 \\ -x_2 + 2x_3 - 2 \end{pmatrix}$$

we find that CPf is nondegenerate, but that several index sets do not lead to complementary points. Thus there are fewer than 8 complementary points and so for some orthants of R^3 there is no solution to (1.4). Thus we would like to restrict our attention to a proper class of P -functions which we will call nondegenerate P -functions. The terminology is in keeping with the use of the adjective nondegenerate [18] when applied to a matrix (as opposed to a complementarity problem as in Definition 1.1) meaning a matrix having all its principal minors nonzero.

DEFINITION 2.5. Let $f: R^n \rightarrow R^n$ be a P -function on R^n . Then if for each $k = 1, \dots, 2^n$ the function $f^{(k)}$, given by (2.1), is a function from R^n onto R^n we say f is a *nondegenerate P -function*.

For such a function, (2.2) will always have a unique solution and so $I^{(k)}$ leads to a unique complementary point

$$(2.4) \quad s^{(k)} = [f^{(k)}]^{-1}(0) + f([f^{(k)}]^{-1}(0)).$$

While discontinuous nondegenerate P -functions are impossible when $n = 1$, the following is an example for $n = 2$. Let $f(x_1, x_2) = (-1, -2)^t + h(x_1, x_2)$ where

$$h(x_1, x_2) = \begin{cases} \sqrt{2}/2 \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} & \text{if } x_1^2 + x_2^2 < 1, \\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} & \text{if } x_1^2 + x_2^2 \geq 1. \end{cases}$$

Because continuity is not required, the results we present generalize [4], [8], [9], [13], [15].

THEOREM 2.6. Let $f: R^n \rightarrow R^n$ be a nondegenerate P -function. Then for each $\mathcal{O} \subseteq R^n$, (1.4) has a unique solution.

Proof. Theorem 2.3 says that there is at most one such solution. We must show that there is at least one solution. Consider two cases:

Case 1 (Nondegenerate problem). We assume each complementary point of the problem has no zero components, and so lies in only one orthant. Theorem 2.3(iii) and our discussion above says that there are 2^n complementary points. Theorem 2.3(i) says that there is at most one per orthant. So a unique s lies in \mathcal{O} and the unique x that leads to it (the unique solution x of (2.2) for the corresponding index set) is the solution we seek.

Case 2 (Degenerate problem). We assume that there exists at least one complementary point s such that $s_i = 0$ for some i . We will analyze the case of one such degenerate complementary point. All of them can be handled in the same manner.

Suppose s is a degenerate complementary point with exactly r zeros, say $s_i = 0$ for $i \in \{i_1, \dots, i_r\}$. Then s lies in 2^r orthants. Since f is a P -function on R^n , there are no other complementary points in these orthants.

The complementary point s corresponds to *at most* 2^r index sets. For if we take any two of them they must agree on the set $N - \{i_1, \dots, i_r\}$ by Theorem 2.3(iii). That is, any i in this set must be either in both of the sets or in neither. Thus only the presence or absence of elements from $\{i_1, \dots, i_r\}$ is possible in the index sets and this leads to at most 2^r index sets.

On the other hand, there are at least 2^r index sets leading to s . There is *at least* one since each complementary point corresponds to a solution of (2.2) for some index set $I^{(k)}$. But then it is seen from the form of the equations (2.2) that once we find one such index set we are free to add or subtract from it any $i \in \{i_1, \dots, i_r\}$ without changing the solution since $s_i = 0$ for $i \in \{i_1, \dots, i_r\}$, i.e. $s_i = x_i = f_i(x) = 0$ for $i \in \{i_1, \dots, i_r\}$.

Thus 2^r index sets lead to this single s which lies in exactly 2^r orthants. Now consider the remaining $2^n - 2^r$ index sets. If there is no other degenerate complementary point, then by Theorem 2.3 they will lead to $2^n - 2^r$ distinct complementary points which will be distributed one to each of the remaining $2^n - 2^r$ orthants. Thus each orthant has a unique solution to (1.4). (Note that for the degenerate s each of the 2^r index sets gives the same x).

If there were another degenerate complementary point, it would be handled in the same manner as the first. Thus the result for Case 2 is established.

In what follows we will show how Murty's scheme can be extended to find the solution shown to exist by Theorem 2.6. To do so we will need some results concerning such solutions. We establish these in the remainder of this section.

DEFINITION 2.7. Partition the f and x in (1.4) by

$$f(x) = \begin{pmatrix} g \\ f_n \end{pmatrix}, \quad x = \begin{pmatrix} y \\ x_n \end{pmatrix}.$$

Denote by \mathcal{O}_1 the orthant of R^{n-1} corresponding to the first $n-1$ components of \mathcal{O} and let $h: R^{n-1} \rightarrow R^{n-1}$ be defined by

$$h(y) = g\left(\begin{pmatrix} y \\ 0 \end{pmatrix}\right).$$

Then the following is a *principal subproblem* of (1.4): find $y \in R^{n-1}$ such that

$$(2.5) \quad y \in \mathcal{O}_1, \quad h(y) \in \mathcal{O}_1^* = \mathcal{O}_1, \quad y^t h(y) = 0.$$

One important property that nondegenerate P -functions possess in common with P -matrices is the inheritance property. That is, principal subfunctions are also nondegenerate P -functions.

THEOREM 2.8. *Given a nondegenerate P -function f on R^n , then h , as defined in Definition 2.7, is a nondegenerate P -function on R^{n-1} .*

Proof. The function h is a P -function on R^{n-1} . To see this, suppose there are two points $y^1 \neq y^2$ in R^{n-1} such that

$$(y_k^1 - y_k^2)(h_k(y^1) - h_k(y^2)) \leq 0, \quad k = 1, 2, \dots, n-1.$$

Let

$$x^1 = \begin{pmatrix} y^1 \\ 0 \end{pmatrix}, \quad x^2 = \begin{pmatrix} y^2 \\ 0 \end{pmatrix};$$

then $x^1 \neq x^2$ and

$$\begin{aligned} & (x_k^1 - x_k^2)(f_k(x^1) - f_k(x^2)) \\ &= \begin{cases} (y_k^1 - y_k^2)(h_k(y^1) - h_k(y^2)) \leq 0, & k = 1, 2, \dots, n-1, \\ 0, & k = n, \end{cases} \end{aligned}$$

contradicting f being a P -function on R^n .

Next we show that h is a nondegenerate P -function on R^{n-1} . Consider $h^{(k_1)}$ where $I^{(k_1)}$ is some index set from the power set of $\{1, 2, \dots, n-1\}$. Then consider $f^{(k)}$ where $I^{(k)} = I^{(k_1)}$. Note that

$$h_i^{(k_1)}(y) = f_i^{(k)}(y_1, y_2, \dots, y_{n-1}, 0), \quad i = 1, 2, \dots, n-1.$$

Since $f^{(k)}$ is a bijection of R^n ,

$$[h_i^{(k_1)}]^{-1}(y) = [f_i^{(k)}]^{-1}(y_1, y_2, \dots, y_{n-1}, 0), \quad i = 1, 2, \dots, n-1.$$

Thus we see that since $f^{(k)}$ is a bijection of R^n , $h^{(k_1)}$ is a bijection of R^{n-1} .

THEOREM 2.9. *Let f be a nondegenerate P -function. Then:*

(a) *Suppose $x_n = 0$ where x is the unique solution to (1.4). Then $y = (x_1, x_2, \dots, x_{n-1})$ is the unique solution to the principal subproblem (2.5).*

(b) Let $x = (y_1, \dots, y_{n-1}, 0)$ where $y = (y_1, \dots, y_{n-1})$ is the unique solution to (2.5). Suppose $f(x) \in \mathcal{O}$. Then x is the unique solution to (1.4).

Before we examine Murty's scheme we will also need some information concerning principal pivots of nondegenerate P -functions.

DEFINITION 2.10. We say g is a *principal pivot* of $f: R^n \rightarrow R^n$ if for some $r \in N$, $g_r: R^n \rightarrow R$ is a unique map satisfying $z_r = g_r(x)$ if and only if $x_r = f_r(x_1, \dots, x_{r-1}, z_r, x_{r+1}, \dots, x_n)$ and for $i \neq r$ we have $g_i(x) = f_i(x_1, \dots, x_{r-1}, g_r(x), x_{r+1}, \dots, x_n)$.

Note that principal pivots with respect to all $r \in N$ exist when f is a nondegenerate P -function since f_r maps R^n onto R and is an increasing function of its r th component. Thus as a function of its r th component it has an inverse.

To conclude our discussion of P -functions, we establish the following nonlinear version of a result of Tucker [19]:

THEOREM 2.11. Suppose f is a nondegenerate P -function. Then any principal pivot, g , of f is also a nondegenerate P -function. Moreover, f and g have the same complementary points.

Proof. Let g be a principal pivot of f , and $r \in N$ be the subscript for which the pivot is made as described in Definition 2.10. To show that it is a P -function, let us define the bijection $J: R^n \rightarrow R^n$ by $Jx = (x_1, \dots, x_{r-1}, g_r(x), x_{r+1}, \dots, x_n)$. Then we see that for x and y in R^n

$$(x_i - y_i)(g_i(x) - g_i(y)) = \begin{cases} ((Jx)_i - (Jy)_i)(f_i(Jx) - f_i(Jy)), & i \neq r, \\ (f_r(Jx) - f_r(Jy))((Jx)_r - (Jy)_r), & i = r. \end{cases}$$

Since f is a P -function we see that if all of the quantities on the right are ≤ 0 , we must have $Jx = Jy$. But then this says that $x = y$. Thus g is a P -function.

To show that g is nondegenerate, consider $g^{(k)}$ for some index set $I^{(k)}$. Now if $r \in I^{(k)}$, it is easy to verify from the definition that

$$g_i^{(k)}(x) = \begin{cases} f_i(Jx), & i \in I^{(k)} - \{r\}, \\ (Jx)_i, & i \in (N - I^{(k)}) \cup \{r\}, \end{cases}$$

and so $g^{(k)}(x) = f^{(\hat{k})}(Jx)$ where \hat{k} is such that $I^{(\hat{k})} = I^{(k)} - \{r\}$. Since $f^{(\hat{k})}$ and J are bijective so is $g^{(k)}$. A similar argument holds if $r \in N - I^{(k)}$.

Finally, we wish to establish that the complementary points of f and g are the same. Let s be a complementary point of f , and x and $k \in N$ the unique vector and index such that $0 = f^{(k)}(x)$, $s = f^{(m)}(x)$ where m is such that $I^{(k)} \cup I^{(m)} = N$. From what we've seen above this means $g^{(\hat{k})}(J^{-1}x) = 0$ for some index set $I^{(\hat{k})}$ obtained from $I^{(k)}$ by adding or dropping a member. But then it is easily verified that for \hat{m} such that $I^{(\hat{k})} \cup I^{(\hat{m})} = N$ we have $g^{(\hat{m})}(J^{-1}x) = f^{(m)}(x) = s$ and so $J^{-1}x$ leads to s which is thus shown to be a complementary point of g . Thus all complementary points of f are also complementary points of g . Since f is also a principal pivot of g , it follows that the set of complementary points for f and g agree.

3. The direct algorithm. By a *direct algorithm* for the generalized CPf (1.4) we mean a method for recursively determining a sequence $(I^{(k)})_{k=1}^m$ of index sets along with some way of determining at each step whether or not $I^{(k)}$ leads to the desired complementary point. We say that the *algorithm is successful* if the desired index set is found (in which case some scheme for numerically solving (2.2) for that index set will have to be employed) and no index set is considered more than once. Clearly one successful direct algorithm in enumeration; i.e. list all the index sets and try them in order. It is not a very efficient algorithm and so other schemes have been introduced. We are interested in the following one:

Murty's scheme. Let $I^{(1)} = \emptyset$. Then given $I^{(k)}$, determine $I^{(k+1)}$ as follows. Suppose $I^{(k)}$ leads to the complementary point $s^{(k)}$. Let $r = \min \{j | s_j^{(k)} \cdot t_j < 0\}$ where t is some element in the interior of \mathcal{O} . (If no such r exists, the scheme terminates and $s^{(k)}$ is the desired complementary point.) If $s_r^{(k)} = x_r^{(k)}$, then $I^{(k+1)} = I^{(k)} - \{r\}$. If $s_r^{(k)} = f_r(x^{(k)})$, then $I^{(k+1)} = I^{(k)} \cup \{r\}$.

Note that as we proceed we need only determine the sign pattern of $s^{(k)}$ (except for the final step, of course). Note further the connection with principal pivoting. For example, for $k=2$, we are interested in the complementary point $s^{(2)}$ given by the x satisfying $0 = f^{(2)}(x)$. But if we pivot on r , the r of the scheme, we find $f^{(2)}(x) = g^{(1)}(J^{-1}x)$. Thus we are interested in the complementary point of g that $I^{(1)}$ leads to. In this manner we see that Murty's method can be interpreted as examining the complementary points $I^{(1)}$ leads to for a sequence of functions, where each element of the sequence is a principal pivot of the preceeding one. We can then establish the following result:

THEOREM 3.1. *If $f: R^n \rightarrow R^n$ is a nondegenerate P -function, then Murty's scheme will be successful when applied to the generalized complementarity problem (1.4).*

Proof. For such an f , $x^{(k)}$ and $s^{(k)}$ are defined for any $I^{(k)} \subseteq N$ and so the scheme is well defined.

We proceed by induction on the size of the problem, as in Murty [17]. If $n=1$, the algorithm terminates in at most one step. For if $I^{(1)}$ is not correct, the scheme tells us to try $I^{(2)}$, which will have to be correct.

Now let n be greater than 1 and assume that the theorem holds for all problems of order $n-1$ or less. Let \hat{x} be the unique solution of (1.4).

Case 1. $\hat{x}_n = 0$. By Theorem 2.9, $\hat{y} = (\hat{x}_1, \dots, \hat{x}_{n-1})$ is the unique solution to the principal subproblem. When we apply the scheme to the original problem we see that since the scheme never changes x_n from 0 until the first $n-1$ components of x agree with the sign pattern in \mathcal{O} , the scheme actually works on the first $n-1$ components as if Murty's scheme were being applied to the principal subproblem. But by our induction hypothesis, Murty's scheme on the principal subproblem will produce \hat{y} in a finite number of steps. Thus the scheme applied to the original problem will produce $\hat{x} = (\hat{y}, 0)$ in the same number of steps.

Case 2. $\hat{x}_n \neq 0$. When we apply Murty's scheme to our problem, our induction assumption assures us that we will generate a sequence of index sets $I^{(1)}, I^{(2)}, \dots, I^{(p)}$, none of which contains n , and only the last of which has the signs of the first $n-1$ components in agreement with \mathcal{O} . The last sign of this complementary point will not agree with \mathcal{O} since $f^{(p)}(x) = 0$ will lead to $x_n = 0$ and so the x here will be of the form $(x_1, x_2, \dots, x_{n-1}, 0)$. This cannot lead to the same s that \hat{x} leads to. Thus Murty's scheme will say to go on to $I^{(p+1)}$ where $n \in I^{(p+1)}$, making $x_n \neq 0$ and $f_n = 0$. But interpreting Murty's scheme by means of principal pivots, we see that we are now investigating a function g , obtained from f by a sequence of principal pivots, where the x leading to the complementary point we are after now has $x_n = 0$. This is because the last bijection J accompanying that last principal pivot has $(Jx)_n = f_n(x)$. Thus starting with $I^{(p+2)}$ we are back in Case 1 with g replacing f . Then our induction hypothesis says we will succeed in finding the desired complementary point of g in \mathcal{O} . By Theorem 2.11, this is also the desired complementary point of f .

4. Discussion. At first glance the above scheme bears some resemblance to an earlier constructive method given in Cottle [4]. Both methods use principal pivot operations and on certain examples both will execute the same sequence of principal pivots. However, Cottle's method requires $f(x)$ to be continuously differentiable and

CPF to be nondegenerate. We do not require the nondegeneracy assumption because we do not need the concept of “critical value”, which was used by Cottle to show that the number of nonnegative basic variables monotonically increases during successive iterations of the method. We do not get such monotonicity—index set $I^{(k)}$ may lead to $s^{(k)}$ which disagrees with \mathcal{O} in only one component, while $I^{(k+1)}$ leads to $s^{(k+1)}$ which disagrees with \mathcal{O} in all but one component (the worst possible case on a P -function). Kojima [10] has noted that solutions are difficult to calculate with Cottle’s method unless the function f is affine. This difficulty arises because of the use of “critical value” (a kind of nonlinear min-ratio rule). The scheme above avoids this complication.

Some experimentation [11] seems to indicate that Murty’s scheme is comparable with other schemes for linear problems involving diagonally dominant P -matrices. Thus it should be considered as a possible scheme for solving similar nonlinear problems.

It should be noted that at each step of the scheme one only has to be sure that we have the sign pattern of the computed complementary point correct before proceeding to the next step. Using this fact, the less “degenerate” a problem is, the easier the direct scheme can solve it. Theoretically, Murty’s scheme will solve even degenerate problems, although some device should be incorporated to try to avoid index set changes based only on roundoff effects on complementary points.

At the k th step of the process we have the nonlinear problem $f^{(k)}(x) = 0$ to solve “approximately”. Present means for the computer solution of systems of simultaneous nonlinear equations should be adequate for producing appropriate approximations in all but the most general cases. Our computer implementation for C^1 functions has used Brown’s method [2], starting with an initial guess of $x = 0$ for each system $f^{(k)}(x) = 0$. Another likely method is that of Boggs [1], which is robust even when only a poor initial approximation to the solution is known. Finally, the techniques of simplicial approximations have recently provided homotopy methods [3], which converge globally under mild restrictions generally satisfied by continuous nondegenerate P -functions. Since numerical methods do not presently exist for discontinuous systems, examples of such problems must be dealt with on an individual basis.

A recent paper [12] showed the equivalence of the complementarity problem CPF to a system of n nonlinear equations. The scheme we outline in this paper seems, at first glance, to involve more work than that in the cited paper, in that perhaps many systems have to be solved. However, only one system, the final one, has to be solved to high accuracy. Each system encountered can be reduced in dimensionality by substituting $x_i = 0$ for $i \in N - I^{(k)}$. Hopefully, the form of the equations, preserving any sparsity which is inherent in $f(x)$, lends itself more readily to available nonlinear methods.

Acknowledgment. The authors thank the referees for their valuable comments.

REFERENCES

- [1] P. T. BOGGS, *The solution of nonlinear systems of equations by A-stable integration techniques*, SIAM J. Numer. Anal., 8 (1971), pp. 767–785.
- [2] K. M. BROWN, *A quadratically convergent Newton-like method based upon Gaussian elimination*, Ibid., 6 (1969), pp. 560–569.
- [3] A. CHARNES, C. B. GARCIA AND C. E. LEMKE, *Constructive proofs of theorems relating to $F(x) = y$ with applications*, Math. Programming, 12 (1977), pp. 328–343.
- [4] R. W. COTTLE, *Nonlinear programs with positively bounded Jacobians*, SIAM J. Appl. Math., 14 (1966), pp. 147–158.

- [5] M. FIEDLER AND V. PTÁK, *On matrices with non-positive off-diagonal elements and positive principal minors*, Czechoslovak Math. J., 12 (1962), pp. 382–400.
- [6] G. J. HABETLER AND A. L. PRICE, *Existence theory for generalized nonlinear complementarity problems*, J. Optimization Theory Appl., 7 (1971), pp. 223–239.
- [7] S. KARAMARDIAN, *The generalized complementarity problem*, Ibid., 8 (1971), pp. 161–168.
- [8] ———, *The nonlinear complementarity problem with applications, Parts I and II*, J. Optimization Theory Appl., 4 (1969), pp. 87–98 and pp. 167–181.
- [9] M. KOJIMA, *A Unification of the existence theorems of the nonlinear complementarity problem*, Math. Programming, 9 (1975), pp. 257–277.
- [10] ———, *Computational methods for solving the nonlinear complementarity problem*, Keio Engineering Reports, vol. 27, no. 1, Keio University, Yokohama, Japan, 1974.
- [11] M. M. KOSTREVA, *Direct algorithms for complementarity problems*, Doctoral thesis, Rensselaer Polytechnic Institute, Troy, New York, 1976.
- [12] O. L. MANGASARIAN, *Equivalence of the complementarity problem to a system of nonlinear equations*, SIAM J. Appl. Math., 31 (1976), pp. 89–92.
- [13] N. MEGIDDO AND M. KOJIMA, *On the existence and uniqueness of solutions in nonlinear complementarity theory*, Math. Programming, 12 (1977), pp. 110–130.
- [14] J. MORÉ, *Classes of functions and feasibility conditions in nonlinear complementarity problems*, Ibid., 6 (1974), pp. 327–338.
- [15] ———, *Coercivity conditions in nonlinear complementarity problems*, SIAM Rev., 16 (1974), pp. 1–16.
- [16] J. MORÉ AND W. RHEINOLDT, *On P - and S -functions and related classes of n -dimensional nonlinear mappings*, Linear Algebra Appl., 6 (1973), pp. 45–68.
- [17] K. G. MURTY, *Note on a Bard-type scheme for solving the complementarity problem*, Opsearch, 11 (1974), pp. 123–130.
- [18] ———, *On the number of solutions to the complementarity problem and the spanning properties of complementary cones*, Linear Algebra Appl., 5 (1972), pp. 65–108.
- [19] A. W. TUCKER, *Principal pivot transforms of square matrices* (Abstract), SIAM Rev., 5 (1963), p. 305.

WEAK TANGENT CONES AND OPTIMIZATION IN A BANACH SPACE*

J. BORWEIN†

Abstract. A general notion of a τ -tangent cone is introduced and developed for optimization purposes. This includes as special cases both the weak and strong tangent cones that appear in the literature.

First order conditions with and without constraint qualification are examined and particular examples are provided to demonstrate that these conditions properly subsume those previously in the literature. Emphasis is placed on weak Kuhn–Tucker sufficiency conditions.

1. Introduction. Suppose that X and Y are real Banach spaces and that $f: X \rightarrow \mathbb{R}$, $g: X \rightarrow Y$ are Fréchet differentiable at \bar{x} . Let (P) denote the general program

$$(P) \quad \text{minimize } f(x) \quad \text{subject to } g(x) \in B, \quad x \in C,$$

where $B \subset X$, $C \subset Y$ are arbitrary sets. (P) is presumed when necessary conditions are discussed for a program to assume a minimum at \bar{x} .

This general programming problem, which includes the standard Kuhn–Tucker program [17], has been studied by Varaiya [25], Guignard [12], Zlobec and Massam [28] and others. These authors use the notion of a tangent cone, introduced by Abadie [1], which extends the Kuhn–Tucker notion of a feasible direction.

In this paper the concept of a τ -tangent cone is applied to produce first order necessary and sufficient optimality conditions for (P) analogous to those in [4], [12], [27]. Examples are given to show that in both directions these are stronger than previous results. Fritz John conditions are also considered.

2. Preliminaries. For any two topological spaces X and Y , $L[X, Y]$ will denote the continuous, linear mappings between X and Y . For any locally convex real topological vector space X , X' denotes $L[X, \mathbb{R}]$ which will be given the weak* topology $\sigma(X', X)$. When M is a subset of X or X' , M^0 , \bar{M} and $[M]$ denote respectively the interior, closure and closed convex hull of M (in the appropriate topology). Convergence in norm is denoted \rightarrow while convergence in another linear topology τ is denoted \rightarrow_τ or $\rightarrow(\tau)$. $N(T)$ and $R(T)$ denote the null space and range of a linear operator T . T^* denotes the adjoint of T .

DEFINITION 1. $C \subset X$ ($C' \subset X'$) is a *cone* if for $x \in C$ ($x' \in C'$), $\alpha x \in C$ ($\alpha x' \in C'$) for all $\alpha \geq 0$.

DEFINITION 2. (i) For a cone $C \subset X$, one defines

$$C^+ = \{x' \in X' : x'(x) \geq 0 \text{ for all } x \in C\}.$$

(C^+ is called the *dual cone* of C and is always closed and convex.)

(ii) For a cone $C' \in X'$, one defines

$$C'^+ = \{x \in X : x'(x) \geq 0 \text{ for all } x' \in C'\}.$$

$C^- = -(C^+) = (-C)^+$. Similarly $(C')^- = (-C')^+$.

DEFINITION 3 [15]. Let $f: X \rightarrow \mathbb{R} \cup \{\infty\}$ be a convex function. The *subgradient set* of f at x , denoted $\partial f(x)$, is the set of vectors in X' satisfying

$$x^*(y - x) \leq f(y) - f(x) \quad \forall y \in X.$$

* Received by the editors December 11, 1975, and in revised form June 30, 1977.

† Department of Mathematics, Dalhousie University, Halifax, Nova Scotia, Canada B3H 3T5. This work was supported in part by the National Research Council of Canada under Account A4493.

If we let $i(x|A)$ denote the indicator function of a set A ($i(x|A)=0$ if $x \in A$, $i(x|A)=\infty$ if $x \notin A$) we then have the relationship

$$\partial i(0|C) = C^-.$$

PROPOSITION 1. Suppose C, D , are convex cones in a locally convex space X and that C' is a convex cone in X' .

- (i) $(C^+)^+ = \bar{C}$; $(C')^{++} = \bar{C}'$.
- (ii) $(\bar{C} \cap \bar{D})^+ = \bar{C}^+ + \bar{D}^+$.

Proof. These are proved in normal spaces in [21]. The extensions offer no problems. \square

Remark. Guignard [12] claims (i) and (ii) with the norm topology on X' . This is only true when X is reflexive. As an example consider $X = l_1$ and the sequence space c_0 as a (closed) cone in l^∞ . Then c_0 is norm closed in l^∞ but c_0 is $\sigma(l^\infty, l_1)$ dense in l^∞ . Thus $(c_0)^{++} = l^\infty$.

PROPOSITION 2. Suppose that C and D are closed cones in a locally convex space X and that $C^0 \cap D \neq \emptyset$. Then

$$(C \cap D)^+ = C^+ + D^+.$$

Proof. This is a consequence of the subgradient relationship $\partial f(x) + \partial g(x) = \partial(f+g)(x)$ applied to $f(x) = i(X|C)$ and $g(x) = i(X|D)$ which holds with the hypothesis $C^0 \cap D \neq \emptyset$. The general relationship is proved by Rockafellar in [24].

Ritter has proved a similar result in [21] for normed spaces. \square

DEFINITION 4 [25]. h is tangent to A at \bar{x} if there is a sequence $\{x_n\}$ in A with $x_n \rightarrow \bar{x}$ and a sequence $\{\lambda_n\}$ of nonnegative real numbers with $\lambda_n(x_n - \bar{x}) \rightarrow h$. The set $T(A, \bar{x})$ of all tangents to A at \bar{x} is called the *tangent cone* to A at \bar{x} .

DEFINITION 5 [12]. $P(A, \bar{x})$, the closed convex hull of $T(A, \bar{x})$, is called the *pseudotangent cone* to A at \bar{x} .

Let τ denote another locally convex (Hausdorff) topology on X which is (1) coarser than the norm topology s and satisfies the property (2) that τ -convergent sequences are s -bounded. Property (2) is always satisfied by any topology of the dual pair (X, X') [22, Thm 1, p. 67] and so in particular by the weak topology σ . If X is the dual of a Banach space it is also true of the weak* topology σ^* [22, Cor. 1, p. 66]. We can now introduce the following notions.

DEFINITION 6 [4]. We say the vector h is a τ -tangent to A at \bar{x} if there is a sequence $\{x_n\}$ in A with $x_n \rightarrow \bar{x}$ and a sequence $\{\lambda_n\}$ of nonnegative real numbers such that $\lambda_n(x_n - \bar{x}) \rightarrow (\tau)h$. $T_\tau(A, \bar{x}) (\subset T_\tau([A], \bar{x}))$ which will be called the τ -tangent cone to A at \bar{x} , consists of all such vectors h .

DEFINITION 7. $P_\tau(A, \bar{x})$ will denote the closed convex hull of $T_\tau(A, \bar{x})$ and will be called the τ -pseudotangent cone to A at \bar{x} .

Remarks. (i) Thus $T(A, x) = T_s(A, x)$ and $P(A, x) = P_s(A, x)$. In this case we will generally drop the subscript.

(ii) It is clear that if $\bar{x} \in \bar{A}$ then $0 \in T(A, \bar{x})$, and that if $\bar{x} \in A^0$ then $T(A, \bar{x}) = X$.

(iii) Since $P_\tau(A, \bar{x})$ is norm closed and convex it is also weakly closed.

(iv) The tangent cone is always a closed cone, but need not be convex. $T_\tau(A, \bar{x})$ need not, in general, be closed.

(v) It is clear that as τ gets weaker the corresponding cones increase.

DEFINITION 8. A_1 is said to be τ -pseudoconvex with respect to A_2 at \bar{x} when $A_1 - \bar{x} \subset P_\tau(A_2, \bar{x})$. When $A_1 = A_2$, A_1 is simply said to be τ -pseudoconvex at \bar{x} .

This final case with $\tau = s$ corresponds to Guignard's definition of pseudoconvexity. Again we will often drop the s . It is easy to verify that if there is some set S , with $A_1 \subset S \subset A_2$, such that S is starshaped at \bar{x} then A_1 is pseudoconvex with respect to A_2 at \bar{x} . In particular, any closed convex set is pseudoconvex at *all* its members. The set $A = \{1/n\}_{n=1} \cup \{0\}$ is a simple example of a disconnected set which is pseudoconvex at 0. From now on it is supposed that $\bar{x} \in \bar{A}$.

PROPOSITION 3. *Suppose τ is a topology of the dual pair. Then*

(i) $T(A, \bar{x}) \subset T_\tau(A, \bar{x}) \subset T_\tau([A], \bar{x}) = T([A], \bar{x}) = P([A], \bar{x})$.

(ii) *If A is τ -pseudoconvex at \bar{x} then $P_\tau(A, \bar{x}) = P([A], \bar{x})$.*

Proof. (i) The first two containments are immediate. The final equality follows from $T([A], \bar{x}) = \bigcup_{t \geq 0} t([A] - \bar{x})$ (see [25]). Moreover, since $\bigcup_{t \geq 0} t([A] - \bar{x})$ is convex, it has the same weak and norm closures, and thus $T_\tau([A], \bar{x}) = T([A], \bar{x})$. (ii) When A is τ -pseudoconvex at \bar{x} one has $A - \bar{x} \subset P_\tau(A, \bar{x})$. Since this last set is convex and closed $\bigcup_{t \geq 0} t([A] - \bar{x}) \subset P_\tau(A, \bar{x})$. In conjunction with $P_\tau(A, \bar{x}) \subset P([A], \bar{x})$, this proves (ii). \square

The next proposition generalizes a result in Rockafellar [24] from convex to τ -pseudoconvex sets.

PROPOSITION 4. *Suppose that τ is a topology of the dual pair and that the following conditions hold:*

(i) $[A] \cap [B] = [A \cap B]$,

(ii) $[A]^0 \cap [B]^0 \neq \emptyset$,

(iii) $A \cap B$ is τ -pseudoconvex at \bar{x} .

Then $P_\tau(A, \bar{x}) \cap P_\tau(B, \bar{x}) = P_\tau(A \cap B, \bar{x})$.

Proof. It clearly suffices to show that $P_\tau(A; \bar{x}) \cap P_\tau(B, \bar{x}) \subset P_\tau(A \cap B, \bar{x})$. By (i) and Proposition 3 (ii) we have

$$P_\tau(A \cap B, \bar{x}) = P_\tau([A \cap B], \bar{x}) = P([A] \cap [B], \bar{x}).$$

However, $P([A] \cap [B], \bar{x}) = P([A], \bar{x}) \cap P([B], \bar{x})$ since (ii) holds. This is proved in R^n in [24] and the proof is unchanged. Since $P([A], \bar{x})$ contains $P_\tau(A, \bar{x})$ and $P([B], \bar{x})$ contains $P_\tau(B, \bar{x})$ we have the required containment. \square

All the conditions are necessary, in that easy examples exist in R^2 which satisfy any two of the hypotheses and not the conclusion.

Borwein and O'Brien [6] have examined a variety of tangent cone properties. In particular, they have shown that a weakly compact set in a normed space which is pseudoconvex at all its points is convex. This can be extended using a recent result of Lau's [18] to show that whenever A is a closed set in a reflexive space

$$(3) \quad \cap \{P(A, a) + a : a \in A\} = \text{star } A,$$

where $\text{star } A = \{\bar{a} \in A : t\bar{a} + (1-t)a \in A \text{ for } 0 \leq t \leq 1, a \in A\}$. This, in conjunction with requiring that A is pseudoconvex at all its members, now implies that A is convex.

In general the notion of τ -pseudoconvexity is strictly weaker than that of pseudoconvexity as the next example shows.

Example 1. Let $X = C[0, 1]$, the continuous functions with sup norm. Then $A = \{f : f \in C[0, 1], \|f\| = 1, f(x) = 1 \text{ for some } x \in [0, 1]\}$ is weakly pseudoconvex at all $a \in \bar{A}$ and is not pseudoconvex at some $a \in \bar{A}$.

Proof. Obviously A is norm closed. To show that it is weakly pseudoconvex, we fix f in A and show that for any other g in A , $g - f$ is in $T_\sigma(A, f)$.

Since for some x_0 in $[0, 1]$ $f(x_0) = 1$, there exist points x_m^1, x_m^2 in $[0, 1]$ with $x_m^1 < x_m^2$, $x_0 \notin [x_m^1, x_m^2]$, $x_m^1 \rightarrow x_0$ and $x_m^2 \rightarrow x_0$ and such that $f(y) > 1 - (1/m)$ for all

$y \in [x_m^1, x_m^2]$. Set $k_m(x) = ((m-1)/m)f(x) + (1/m)g(x)$ and

$$h_m(x) = \begin{cases} k_m(x) & \text{if } x \notin [x_m^1, x_m^2], \\ \lambda + (1-\lambda)k_m(x_m^1) & \text{if } x = \lambda\left(\frac{x_m^1 + x_m^2}{2}\right) + (1-\lambda)x_m^1, \quad 0 \leq \lambda \leq 1, \\ \lambda + (1-\lambda)k_m(x_m^2) & \text{if } x = \lambda\left(\frac{x_m^1 + x_m^2}{2}\right) + (1-\lambda)x_m^2, \quad 0 \leq \lambda \leq 1. \end{cases}$$

By construction $h_m \in A$ and

$$\|f - h_m\| < \frac{2}{m}, \quad \text{so } h_m \rightarrow f \text{ in norm.}$$

Also for any $x \notin [x_m^1, x_m^2]$, $m(k_m(x) - f(x)) = g(x) - f(x)$ so $m(h_m - f) \rightarrow g - f$ pointwise. Since $\|m(h_m - f)\| < 2$ for all m , the Lebesgue bounded convergence theorem implies $\mu(m(h_m - f)) \rightarrow \mu(g - f)$ for any continuous linear functional μ on $C[0, 1]$. Thus $m(h_m - f)$ converges weakly to $g - f$ and $g - f \in T_\sigma(A, f)$.

It remains to show that A is not pseudoconvex. It is easy to do this by showing that if $i(x) = x$ is the identity function on $[0, 1]$, then $P(A, i) \subset \{\lambda(g - i) : g \in A, g(1) = 1\}$. Thus A is not pseudoconvex at i . \square

So at least in a nonreflexive Banach space, a bounded weakly pseudoconvex set need not be convex.

DEFINITION 9. $f: X \rightarrow R$ is *quasiconvex on a convex set C* if

$$\{x \in X : x \in C, f(x) \leq r\}$$

is convex for all $r \in R$.

DEFINITION 10. $f: X \rightarrow R$ is *pseudoconvex on a set A* at \bar{x} if f is Fréchet differentiable and if $x \in A$, $f'(\bar{x})(x - \bar{x}) \geq 0$ implies $f(x) \geq f(\bar{x})$.

The next proposition gives a simple relation between these two concepts. It generalizes an observation of Guignard [12].

PROPOSITION 5. Suppose $f: X \rightarrow R$ is quasiconvex on a convex set C and differentiable. Suppose that, for some $y \in C$, $f'(\bar{x})(y - \bar{x}) > 0$. Then f is pseudoconvex at \bar{x} on C .

Proof. Suppose $f'(\bar{x})(x - \bar{x}) \geq 0 \quad \forall x \in C$. Then, for $0 < \alpha \leq 1$, $(1-\alpha)f'(\bar{x})(x - \bar{x}) + \alpha f'(\bar{x})(y - \bar{x}) > 0 \quad \forall x \in C$. Since $x, y \in C$ and C is convex $x_\alpha = (1-\alpha)x + \alpha y \in C$ and $f'(\bar{x})(x_\alpha - \bar{x}) > 0$. Since f is supposed quasiconvex on C it follows easily that $f(x_\alpha) > f(\bar{x})$ for $0 < \alpha \leq 1$. Since f is continuous, $f(x) \geq f(\bar{x})$.

Guignard uses this lemma in the case that $C = X$ and $f'(\bar{x}) \neq 0$.

DEFINITION 11. $g: X \rightarrow Y$ is τ_1 - τ_2 continuous if it is continuous from the τ_1 topology on the X to the τ_2 topology on Y . If $\tau_1 = \sigma$ and $\tau_2 = s$ we say g is *completely continuous*.

The complete continuity of g implies that of $g'(\bar{x})$ (on bounded sets) [26]. Since $g'(\bar{x})$ is linear and continuous it is completely continuous when it is compact (maps bounded sets into compact sets). For simplicity in many of the following results one may assume compactness of $g'(\bar{x})$ in cases in which complete continuity suffices. This in particular insures that $g'(\bar{x})$ is τ - s continuous whenever τ is a dual topology.

3. First order optimality conditions. Consider (P). Following Guignard [12], let Δ denote $\{x : g(x) \in B\}$, and let A denote the feasible set $\Delta \cap C$. The first proposition of this section gives a τ -tangent cone containment analogous to the one proven in [12].

PROPOSITION 6. Suppose that $g'(\bar{x})$ is τ_1 - τ_2 continuous. Then

$$g'(\bar{x})(P_{\tau_1}(\Delta, \bar{x})) \subset P_{\tau_2}(B, g(\bar{x})).$$

Proof. Suppose h is nonzero and belongs to $T_{\tau_1}(\Delta, \bar{x})$. Then there is a sequence $\{h_n\}$, $h_n = \lambda_n(x_n - \bar{x}) \rightarrow (\tau_1)h$, with $\{x_n\} \subset A$, $x_n \rightarrow \bar{x}$ and $\lambda_n \geq 0$ for all $n \in N$. Since $h \neq 0$, one may assume that $\lambda_n \rightarrow \infty$ and $\lambda_n \geq 1$. g is supposed Fréchet differentiable at \bar{x} . It follows that

$$(4) \quad \frac{g(\bar{x} + tk) - g(\bar{x})}{t} - g'(\bar{x})(x) \rightarrow 0,$$

uniformly on bounded sets in X , as $t \rightarrow 0$. In particular, $\{h_n\}$, being a τ_1 -convergent sequence, is bounded by assumption (1) on τ_1 . Thus

$$(5) \quad \frac{g(\bar{x} + \lambda_n^{-1}h_n) - g(\bar{x})}{\lambda_n^{-1}} - g'(\bar{x})(h_n) \rightarrow 0,$$

as $n \rightarrow \infty$. Moreover, $g'(\bar{x})$ is, by hypothesis, τ_1 - τ_2 continuous. It follows that $\lambda_n(g(\bar{x} + \lambda_n^{-1}h_n) - g(\bar{x}))$ converges τ_2 to $g'(\bar{x})(h)$. (This uses assumption (2).) Since $\bar{x} + \lambda_n^{-1}h_n = x_n \in \Delta$, and $x_n \rightarrow \bar{x}$ one has $g(x_n) \in B$ and $g(x_n) \rightarrow g(\bar{x})$. (g is continuous as it is Fréchet differentiable.) It follows that $g'(\bar{x})(h) \in T_{\tau_2}(B, g(\bar{x}))$.

Since $g'(\bar{x})$ is continuous and linear, it preserves the closed convex hulls of sets and $g'(\bar{x})(P_{\tau_1}(\Delta, \bar{x})) \subset P_{\tau_2}(B, g(\bar{x}))$. \square

COROLLARY 1. If $g'(\bar{x})$ is compact then $g'(\bar{x})(P_{\sigma}(\Delta, \bar{x})) \subset P(B, g(\bar{x}))$.

Remarks. (i) Note that $g'(\bar{x})$, being a Fréchet derivative, is always s - s continuous and so, being linear, is always σ - σ continuous. Similarly if X and Y are dual spaces and $g'(\bar{x})$ is an *adjoint* map it is σ^* - σ^* continuous.

(ii) These results hold true in locally convex spaces if one uses bounded derivatives ([1]). This remains true for the rest of the development. If one is interested only in strong cones one can replace Fréchet derivatives by Gateaux derivatives of locally Lipschitz functions.

The next result generalizes Theorem 1 of Guignard [12].

THEOREM 1. Suppose that $f'(\bar{x})$ is τ -continuous. A necessary condition for \bar{x} to minimize f over A is that $f'(\bar{x}) \in P_{\tau}(A, \bar{x})^+$. If f is pseudoconvex over A at \bar{x} and A is τ -pseudoconvex at x_0 , then the condition is also sufficient.

Proof. (i) *Necessity:* Suppose that $h \in T_{\tau}(A, \bar{x})$. As in Proposition 6 there is a sequence $\{x_n\}$, $x_n \in A$, $x_n \rightarrow \bar{x}$ and a sequence of positive real numbers $\{\lambda_n\}$ with $h_n = \lambda_n(x_n - \bar{x})$ and $h_n \rightarrow h$. Again, as in Proposition 6,

$$(6) \quad \lambda_n(f(x_n) - f(\bar{x})) = \lambda_n(f(\bar{x} + \lambda_n^{-1}h_n) - f(\bar{x})) \rightarrow f'(\bar{x})(h),$$

because $f'(\bar{x})$ is assumed τ -continuous. By assumption $x_n \in A$ so that $f(x_n) \geq f(\bar{x})$. Since $\lambda_n \geq 0$, one has $f'(\bar{x})(h) \geq 0$. This holds for all $h \in T_{\tau}(A, \bar{x})$. By continuity and linearity of $f'(\bar{x})$, one has in fact $f'(\bar{x}) \in P_{\tau}(A, \bar{x})^+$.

(ii) *Sufficiently:* Suppose $f'(\bar{x}) \in P_{\tau}(A, \bar{x})^+$. Then $f'(\bar{x})(h) \geq 0$ for all h in $P_{\tau}(A, \bar{x})$. Since A is τ -pseudoconvex at \bar{x} , $f'(\bar{x})(x - \bar{x}) \geq 0$ for all $x \in A$. This in turn implies that $f(x) \geq f(\bar{x})$ for $x \in A$, because f is supposed pseudoconvex. Thus \bar{x} minimizes f over A . \square

When τ is a topology of the dual pair $f'(\bar{x})$ is always τ -continuous. If X is a dual space E the condition that $f'(\bar{x})$ is σ^* -continuous requires that $f'(\bar{x}) \in E$ not only that $E'' = X'$. By Corollary 5 of [2] this is true if f is a σ^* -lower semicontinuous convex function which is Fréchet differentiable at \bar{x} .

Guignard proved Theorem 1 with $P_{\tau}(A, \bar{x})$ replaced by $P(A, \bar{x})$ and, in the sufficiency proof with A assumed pseudoconvex. Since $P(A, \bar{x}) \subset P_{\tau}(A, \bar{x})$, it is immediate that Theorem 1 is more general. Moreover, as the next example shows, in some cases Theorem 1 gives more information.

Example 2. (a) Let $X = l_2(N)$ and $A = \{(e_n + e_1)/n\}_{n=2}^\infty \cup \{0\}$, where e_n is the sequence with a one in the n th position and zeros elsewhere. Suppose $f: X \rightarrow \mathbb{R}$ has a minimum at 0 over A . Now $T(A, 0) = P(A, 0) = 0$ while $e_1 \in T_\sigma(A, 0) \subset P_\sigma(A, 0)$. Thus Theorem 1 implies that $f'(0)(e_1) \geq 0$ while Theorem 1 in [12] gives no information.

(b) Now consider the same example with $X = l_1(N)$. Again $T(A, 0) = P(A, 0) = 0$ while now $T_\sigma(A, 0) = 0$ since weak and strong sequential convergence agree in $l_1(N)$. By considering $X = c_0'$ we can verify that in $\sigma^* = \sigma(l_1, c_0)$ we have $e_1 \in T_{\sigma^*}(A, 0)$. Thus if $f'(\bar{x}) \in c_0$ we can get information from Theorem 1 again even though weak tangent cones provide no information.

Let $K(\tau) = \{h \in X : g'(\bar{x})(h) \in P_\tau(B, g(\bar{x}))\}$ and let $H(\tau) = \{h' \in X' : h' = u^+ g'(\bar{x}), u^+ \in P_\tau(B, g(\bar{x}))^+\}$.

PROPOSITION 7. $K(\tau)^+ = \overline{H(\tau)}$. In addition when one has $R(g'(\bar{x})) \cap P_\tau(B, g(\bar{x}))^0 \neq \emptyset$, then $K(\tau)^+ = H(\tau)$.

Proof. Suppose $h \in K(\tau)$ and $h' \in H(\tau)$. Then $h' = u^+ g'(\bar{x})$, with $u^+ \in P_\tau(B, g(\bar{x}))$, and $h'(h) = (u^+ g'(\bar{x}))(h) = u^+(g'(\bar{x})(h)) \geq 0$. Thus $K(\tau) \subset H^+(\tau)$. If $h_1 \notin K(\tau)$, there is, by the strong separation theorem, some $u^+ \in P_\tau(B, g(\bar{x}))^+$ with $u^+ g'(\bar{x})(h_1) < 0$. Thus $h_1 \notin H^+(\tau)$ and $H^+(\tau) = K(\tau)$. Proposition 1(ii) shows that $K^+(\tau) = H^{++}(\tau) = \overline{H(\tau)}$. Suppose that the auxiliary condition is met.

Theorem 3 of [15, p. 5] guarantees that in a locally convex setting

$$(7) \quad \partial f(x) = T^* \partial d(Tx),$$

where d is a convex function, T is a continuous linear operator and $f(x) = d(Tx)$. For (7) to hold one needs d continuous at some point of $R(T)$. One quickly verifies that this is true for $d(y) = i(y|P_\tau(B, g(\bar{x})))$ and $T = g'(\bar{x})$. Apply (7) with $x = 0$ and one has (see the discussion below Definition 3)

$$K(\tau)^- = \partial f(0) = g'(\bar{x})^* \partial d(0) = g'(\bar{x})^* P_\tau(B, g(\bar{x}))^-,$$

which clearly yields the desired equivalence. \square

Proposition 7 gives conditions which exclude the example given by Zlobec in [27] which exhibits a case in which H is not closed but $R(g'(\bar{x}))$ is. In his example the interior condition is violated.

THEOREM 2 (The generalized Kuhn–Tucker conditions). *Suppose that $H(\tau_2)$ is closed and that $G \subset X$ is a closed convex cone with $K(\tau_2) \cap G \subset P_{\tau_1}(A, \bar{x})$. Suppose that $K^+(\tau_2) + G^+$ is closed. A necessary condition for \bar{x} to minimize (P) when $f'(\bar{x})$ is τ_1 -continuous is that there exist $u^+ \in P_{\tau_2}(B, g(\bar{x}))^+$ with*

$$f'(\bar{x}) - u^+ g'(\bar{x}) \in G^+.$$

This condition is also sufficient if (i) G is a closed convex cone containing $A - \bar{x}$, (ii) $g'(\bar{x})$ is τ_1 - τ_2 continuous, (iii) A is τ_1 -pseudoconvex with respect to Δ at \bar{x} , and if (iv) f is pseudoconvex over A at \bar{x} .

Proof. (i) *Necessity:* By Theorem 1 $f'(\bar{x}) \in P_{\tau_1}(A, \bar{x})^+$. Since $K(\tau_2)^+ + G^+$ is assumed closed and $K(\tau_2) \cap G \subset P_{\tau_2}(A, \bar{x})$ one has

$$f'(\bar{x}) \in \overline{K(\tau_2)^+ + G^+} = K(\tau_2)^+ + G^+.$$

This uses Proposition 1(ii). Proposition 7 and the assumption that $H(\tau_2)$ is closed mean that $f'(\bar{x}) \in P_{\tau_1}(A, \bar{x})^+ \subset H(\tau_2) + G^+$. This is the required result.

(ii) *Sufficiency:* Since $A - \bar{x} \subset G$ one has

$$f'(\bar{x})(x - \bar{x}) \geq u^+ g'(\bar{x})(x - \bar{x}) \quad \forall x \in A.$$

Since $A - \bar{x} \subset P_{\tau_1}(\Delta, \bar{x})$ by assumption, and since $g'(\bar{x})(P_{\tau_1}(\Delta, \bar{x}))$ is contained in $P_{\tau_2}(B, g(\bar{x}))$ by Proposition 6, $g'(\bar{x})(x - \bar{x}) \in P_{\tau_2}(B, g(\bar{x}))$ for all x in A . Since $u^+ \in P_{\tau_2}(B, g(\bar{x}))^+$, this implies that $f'(\bar{x})(x - \bar{x}) \geq 0$ for all x in A . Because f is assumed pseudoconvex over A , $f(x) \geq f(\bar{x})$ for x in A . \square

Let $H(G, \tau_1, \tau_2)$ denote the hypotheses that $K^+(\tau_2) + G^+$ and $H(\tau_2)$ are $\sigma(X', X)$ closed with $K(\tau_2) \cap G \subset P_{\tau_1}(A, \bar{x})$.

COROLLARY 2. $H(G, \tau_1, \tau_2)$ is satisfied when $G \cap K(\tau_2) \subset P_{\tau_1}(A, \bar{x})$ and

- (i) $G^0 \cap K(\tau_2) \neq \emptyset$ or $G \cap K(\tau_2)^0 \neq \emptyset$,
- (ii) $(Rg'(\bar{x})) \cap P_{\tau_2}(B, g(\bar{x}))^0 \neq \emptyset$.

Proof. This follows from Proposition 2 and Proposition 7. \square

THEOREM 2'. The following conditions are also sufficient in Theorem 2:

- (i) $A - \bar{x} \subset G$,
- (ii) $g(A)$ is τ_2 -pseudoconvex with respect to B at $g(\bar{x})$,
- (iii) the function $f - u^+g$ is pseudoconvex over A at \bar{x} .

Proof. As before we have

$$(f'(\bar{x}) - u^+g'(\bar{x}))(x - \bar{x}) \geq 0 \quad \forall x \in A.$$

Since $f - u^+g$ is assumed pseudoconvex at \bar{x} we have

$$f(x) - u^+g(x) \geq f(\bar{x}) - u^+g(\bar{x}) \geq 0 \quad \forall x \in A.$$

Or

$$(8) \quad f(x) - f(\bar{x}) \geq u^+(g(x) - g(\bar{x})) \quad \forall x \in A.$$

Hypothesis (ii) implies that $g(x) - g(\bar{x}) \in P_{\tau_2}(B, g(\bar{x}))$ for each $x \in A$. As $u^+ \in P_{\tau_2}(B, g(\bar{x}))^+$, (8) implies that \bar{x} minimizes f over A . \square

Remarks. (i) Hypothesis (ii) is guaranteed when B is convex while hypothesis (iii) is certainly weaker than asking for f to be convex and g to be concave. Indeed simple examples can be created to show that (ii) and (iii) are not subsumed in Theorem 2 partially because these conditions make use of u^+ which the previous ones do not.

(ii) Example 1 and the discussion preceeding it shows that weak pseudoconvexity is a good deal less restrictive a property than pseudoconvexity.

In finite dimensions polyhedralness of the various cones suffices to guarantee closure in the corollary. Guignard has indicated in [12] how $H(G, \tau_1, \tau_2)$ ($H(G)$ in her case) includes the standard Kuhn-Tucker conditions in finite dimensions where all linear topologies agree. It is also easy to verify that the standard infinite dimensional "constraint qualifications" are subsumed.

The asymptotic results of Zlobec [27] and Zlobec and Massam [28] can also be proved correspondingly for τ -tangent cones, as can the Pareto optimum results in [5]. Proposition 4 can be used to give conditions under which $P_{\tau}(A, \bar{x}) = P_{\tau}(\Delta, \bar{x}) \cap P_{\tau}(C, \bar{x})$. This, in conjunction with the requirement that $P_{\tau}(\Delta, \bar{x}) = K(\tau)$, gives conditions for Theorem 2 to hold with $G = P_{\tau}(C, \bar{x})$. In particular, $P(\Delta, \bar{x}) = K$ when B is a closed convex cone with interior and $g'(\bar{x})(\bar{h}) + g(\bar{x}) \in B^0$ for some $\bar{h} \in X$. This last condition also guarantees that Proposition 7 applies and $H(\tau)$ is closed.

4. Fritz John conditions. In the case that $H(G, \tau_1, \tau_2)$ does not hold one can still often give a necessary condition for optimality. Such a condition is generally called a Fritz John condition [16]. The theorem stated below is the τ -tangent cone analogue of one in Nagahisa and Sakawa [20].

THEOREM 3. Suppose that \bar{x} is minimal for (P) with B a closed convex cone with interior. Suppose $f'(x)$ is τ -continuous and $g'(\bar{x})$ is τ -s continuous. For any closed convex

cone M in $T_\tau(C, \bar{x})$ there exist $r^+ \geq 0$ and $u^+ \in B^+$ with $u^+g(\bar{x}) = 0$, not both zero, such that

$$r^+f'(\bar{x}) - u^+g'(\bar{x}) \in M^+.$$

Proof. As in [20], consider $S_1 = \{(r, z) | f'(\bar{x})(h) \leq r, g'(\bar{x})(h) + g(\bar{x}) - z \in B, \text{ for some } h \in M\}$ and $S_2 = \{(r, z) | r \leq 0, z \in B\}$. It suffices to show that S_1 and S_2 are convex sets with $S_1 \cap S_2^0 = \emptyset$ since then they can be separated and the separating hyperplane $z^+ = (r^+, u^+)$ has the requisite properties exactly as in [20]. Suppose, therefore, that $S_1 \cap S_2^0 \neq \emptyset$. There is then some $\bar{h} \in M$ with $f'(\bar{x})(\bar{h}) < 0$ and $g'(\bar{x})(\bar{h}) + g(\bar{x}) \in B^0$. Since $\bar{h} \in M$, there is a sequence $\{x_n\}$ in C with $x_n \rightarrow \bar{x}$, and $\lambda_n \in R^+$ with $h_n = \lambda_n(x_n - \bar{x}) \rightarrow (\tau)\bar{h}$. Now $g'(\bar{x})$ is τ -s continuous, so that there exists some $n_0 \in N$ such that $g'(\bar{x})(h_n) + g(\bar{x}) \in B^0$ for $n \geq n_0$. Similarly, for $n \geq n_1$, $f'(\bar{x})(h_n) < 0$. It follows from (5) of Proposition 6 that for n large enough one has

$$\lambda_n(g(\bar{x} + \lambda_n^{-1}h_n) - g(\bar{x})) + g(\bar{x}) \in B^0.$$

λ_n is positive so that

$$g(x_n) \in B^0 + (1 - \lambda_n^{-1})g(\bar{x}).$$

Since $\lambda_n^{-1} \rightarrow 0$ and $g(\bar{x}) \in B$ one must have $g(x_n) \in B^0 + B \subset B^0$ for n sufficiently large. A similar argument shows that $f(x_n) < f(\bar{x})$ for large n . Since $x_n \in C$ this contradicts the optimality of \bar{x} . Thus $S_1 \cap S_2^0 = \emptyset$. Since both S_1 and S_2 are convex they can be separated and the proof proceeds as in [20]. \square

Remark. (i) The requirement that $g'(\bar{x})$ be τ -s continuous can be replaced by the condition that $g'(\bar{x})$ be τ - τ_2 continuous and B have interior in the τ_2 topology as one can still separate S_1 and S_2 in that case.

(ii) If τ is a dual topology one can require that $g'(\bar{x})$ is compact, or, of course, finite dimensional.

As a consequence of this result one can extend the theorems in Craven [8] and Craven and Mond [9] on Fritz John conditions with equality constraints. Two definitions are necessary.

DEFINITION 12 [8]. A continuous linear map $B: X \rightarrow Z$ is said to be *adequate* if (i) $R(B)$ is closed in Z , and if (ii) $R(B) = Z$ then there is a continuous projection of X onto $N(B)$.

DEFINITION 13. Let $h: X \rightarrow Z$ be a Fréchet differentiable map between two Banach spaces. h will be called (G, τ) -regular at \bar{x} when $R(h'(x))$ is closed and when $R(h'(x)) = Z$ implies $N(h'(x)) \cap G \subset T_\tau(N(h), \bar{x})$ for some closed convex cone G with $G^0 \cap N(h'(\bar{x})) \neq \emptyset$.

In particular, it follows, from Flett's results in [10], that a *continuously* Fréchet differentiable map with a surjective derivative at \bar{x} is (G, τ) -regular for $G = X$, $\tau = s$. In general (G, τ) -regularity caters to the possibility that $g'(x)$ is not continuous in x .

THEOREM 4. Suppose that in Theorem 3 $C = N(h)$ where h is (G, τ) -regular at \bar{x} . Then there exist $r^+ \geq 0$, $u^+ \in B^+$, $z^+ \in Z'$, not all zero, such that $r^+f'(\bar{x}) - u^+g'(\bar{x}) - z^+h'(\bar{x}) \in G^+$; $u^+g(\bar{x}) = 0$. Moreover, when $h'(\bar{x})$ is surjective, one of u^+ , r^+ can be

Proof. In the case that $\overline{R(h'(\bar{x}))} \neq Z$ one can find some nonzero $z^+ \in Z'$ with $z^+h'(\bar{x}) = 0$. Setting $u^+ = 0$ and $r^+ = 0$, one is done. Suppose now that $R(h'(\bar{x})) = Z$. Then $N(h'(\bar{x})) \cap G \subset T_\tau(N(h), \bar{x})$. Since $N(h'(\bar{x})) \cap G$ is a closed convex cone in $T_\tau(C, \bar{x})$, (with $C = N(h)$) one can apply Theorem 3 to derive that $r^+ \geq 0$, $u^+ \in B^+$ with $u^+g(\bar{x}) = 0$ exist such that $r^+f'(\bar{x}) - u^+g'(\bar{x}) \in (N(h'(\bar{x})) \cap G)^+$. Since $G^0 \cap N(h'(\bar{x})) \neq$

by hypothesis this can be rewritten as

$$r^+ f'(\bar{x}) - u^+ g'(\bar{x}) \in N(h'(\bar{x})) + G^+$$

A simple application of the Farkas lemma [15] shows that any member y^+ of $N(h'(\bar{x}))^+$ can be written as $z^+ h'(\bar{x})$ with $z^+ \in Z'$. This concludes the proof. \square

Craven and Mond's Theorem 4 in [9] is essentially a corollary of the present Theorem 4 and $G = X$. There it is proved for an adequate continuously differentiable h . As has been seen such a mapping is G -regular with $G = X$. This theorem is in itself a generalization of the Fritz John results in [16]. Note that adequacy is entirely too strong a notion.

Zlobec and Massam state Theorem 3 for (strong) tangent cones in locally convex spaces with an erroneous formulation. Theorem 3 of this paper can also be extended to locally convex spaces with bounded differentiation.

We conclude this section by observing that the condition in Theorem 3 and 4 that B be a cone can be removed as follows:

THEOREM 5. *Assume the hypotheses of Theorem 3 save that B need now only be a closed convex set with interior. Then the conclusions hold except that now*

$$u^+ \in (B - g(\bar{x}))^+ = P(B, g(\bar{x}))^+.$$

Proof. Consider $g_1: X \rightarrow Y \times R$ given by $g_1(x) = (g(x), 1)$. Let $S = \{(x, r): x \in rB, r \geq 0\}$. We can check that

$$g(x) \in B \Leftrightarrow g_1(x) \in \bar{S}.$$

Moreover \bar{S} is a closed convex cone with interior and so we may apply Theorem 3 to deduce that

$$(9) \quad r^+ f'(\bar{x}) - s^+ g'_1(\bar{x}) \in M^+, \quad s^+ g_1(\bar{x}) = 0,$$

with $r^+ \geq 0$, $s^+ \in S^+$ not both zero. By construction of S , $s^+ = (b^+, r_1^+)$ with

$$(10) \quad b^+(b) + r_1^+ \geq 0 \quad \forall b \in B.$$

Also

$$(11) \quad b^+ g(\bar{x}) + r_1^+ = s^+ g_1(\bar{x}) = 0.$$

Together (10) and (11) show that $b^+ \in (B - g(\bar{x}))^+$. From (11) it also follows that $s^+ \neq 0$ implies $b^+ \neq 0$. Finally, since

$$(12) \quad s^+ g'_1(\bar{x}) = b^+ g'(\bar{x}) + r_1^+ 0 = b^+ g'(\bar{x}),$$

we obtain the desired conclusion by substituting (12) into (9). \square

Applications. We do not give an application of these results here. However, Theorem 4 can be used to give an analogous application to optimal control to that given by Craven and Mond [9]. One merely relaxes their condition of adequacy to G -regularity. Similar remarks apply to the Kuhn-Tucker theorems for which examples (mainly finite dimensional) are given in [12].

5. On the "necessity" of the constraint qualification. Various technical results are possible when τ_1 is a topology which makes the unit ball $B(X)$ in X a compact set. It is known in this case that $X = E'$ where E is a Banach space [14]. For simplicity we

confine ourselves to the case in which X is reflexive and $\tau_1 = \sigma(X, X')$. As in [4] we let

$$F(A, \bar{x}) = \{f: f \text{ is Fréchet differentiable at } \bar{x} \text{ and achieves a (local) minimum over } A \text{ at } \bar{x}\},$$

$$M(A, \bar{x}) = \{f'(\bar{x}): f \in F(A, \bar{x})\}.$$

It is proved in [4], that when X is reflexive

$$(13) \quad P_\sigma(A, \bar{x})^+ \subset M(A, \bar{x}).$$

(An inspection of the proof of (13) in [4] shows that if $X = E'$ and τ_1 makes $B(X)$ sequentially compact one in fact has $P_{\tau_1}(A, \bar{x})^+ \cap E \subset M(A, \bar{x}) \cap E$.)

Consider the special case of (P) in which $C = X$.

$$(P') \quad \min \{f(x) : g(x) \in B\}.$$

THEOREM 6. Suppose X is reflexive and $g'(x)$ is σ - τ continuous and that

$$(14) \quad f'(\bar{x}) - u^+ g'(\bar{x}) = 0; \quad u^+ \in P_\tau(B, g(\bar{x}))^+$$

holds for all $f \in F(A, \bar{x})$. Then $H(X, \sigma, \tau)$ holds. In other words the constraint qualification is necessary in this case.

Proof. From (13) it follows that $M(A, \bar{x}) \subset H(\tau)$. Now

$$H(\tau) \subset K(\tau)^+ \subset P_\sigma(A, \bar{x})^+,$$

where the first inequality follows from Proposition 7 and the second from Proposition 6. Since (14) holds we have

$$P_\sigma(A, \bar{x})^+ = H(\tau) = K(\tau)^+,$$

which is $H(X, \sigma, \tau)$. \square

This extends the corresponding results in [4], from the case in which B is the orthant in R^n . A moment's reflection on Example 2(b) and Theorem 1 shows that (13) fails in a more general dual space since we may have $P_\sigma(A, \bar{x})^+ = X'$ while $P_{\sigma^*}(A, \bar{x})^+ \neq X'$. The latter will serve as a constraint on at least some members of $M(A, \bar{x})$. One might hypothesize that (13) in fact characterizes reflexive spaces. In any case the general question of necessity of $H(G, \tau_1, \tau_2)$ still remains open even for $G = X$ and $\tau_1 = \sigma^*$. It would also be interesting to characterize the existence of a bounded or compact set of multipliers along the lines of the work in [11].

Theorem 6 thus illustrates another reason for examining τ -tangent cones. Clarke [7] has defined and exploited a different tangent cone $T_E(x)$. It agrees with this notion on convex sets but it is always convex and satisfies $T_E(x) \subset T(E, x)$. Thus in general his normal cone $N_E(x) = T_E(x)^+$ is strictly larger than $T(E, x)^+$. It therefore constrains a derivative less to belong to $N_E(x)$ than to $T_E(x)^+$. Clarke's optimal condition applies to all generalized derivatives [7] while this tangent cone approach requires something close to Fréchet derivatives to exist. The author intends to systematically explore the comparisons between the two cone notions in an upcoming paper.

Acknowledgments. I would like to thank Dr. M. A. H. Dempster, Dr. R. O'Brien and the referee for their generous suggestions.

REFERENCES

- [1] J. M. ABADIE, *Problèmes d'optimisation*, Institut Blaise Pascal, Paris, 1965.
- [2] E. ASPLUND AND R. T. ROCKAFELLAR, *Gradients of convex functions*, Trans. Amer. Math. Soc., 139 (1969), pp. 443-467.

- [3] V. T. AVERBUKH AND O. G. SMOLYANOV, *The various definitions of the derivative in linear topological spaces*, Russian Math. Surveys, 23 (1968), no. 4, pp. 67–114.
- [4] M. S. BAZARAA, J. J. GOODE, M. F. NASHED AND C. M. SHETTY, *Nonlinear programming without differentiability in Banach Spaces. Necessary and sufficient constraint qualifications*, J. Appl. Analysis, 5 (1976), pp. 165–173.
- [5] J. BORWEIN, *Proper efficient points for maximizations with respect to cones*, this Journal, 15 (1977), pp. 57–63.
- [6] J. BORWEIN AND R. O'BRIEN, *Tangent cones and convexity*, Canad. Math. Bull., 19 (1976), pp. 257–261.
- [7] F. H. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.
- [8] B. D. CRAVEN, *Nonlinear programming in locally convex spaces*, J. Optimization Theory Appl., 10 (1970), pp. 197–210.
- [9] B. D. CRAVEN AND B. MOND, *Transposition theorems for cone-convex functions*, SIAM J. Appl. Math., 24 (1973), pp. 603–612.
- [10] T. M. FLETT, *On differentiation in normed vector spaces*, J. London Math. Soc., 42 (1967), pp. 523–533.
- [11] J. GAUVIN, *A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming*, Math. Programming, 12 (1977), pp. 136–138.
- [12] M. GUIGNARD, *Generalized Kuhn–Tucker conditions for mathematical programming in a Banach space*, this Journal, (1969), pp. 232–241.
- [13] H. HALKIN, *Implicit functions and optimization problems without continuous differentiability of the data*, C.O.R.E. Discussion Paper no. 7206, Centre for Operations Research and Econometrics, (1972).
- [14] R. B. HOLMES, *Geometric Functional Analysis and its Applications*, Springer-Verlag, New York, 1975).
- [15] A. D. IOFFE AND V. L. LEVIN, *Differentials of convex functions*, Trans. Moscow Math. Soc., 26 (1972), pp. 1–71.
- [16] F. JOHN, *Extremum problems with inequalities as subsidiary conditions*, Studies and Essays, Courant Anniversary Volume, Interscience, New York, 1948, pp. 187–204.
- [17] H. W. KUHN AND A. W. TUCKER, *Nonlinear programming*, Proc. Second Berkeley Symposium on Math. Statistics and Probability, J. Neyman, ed., University of California Press, Berkeley, CA, 1951.
- [18] KA-SING LAU, *Almost Chebyshev subsets in reflexive Banach spaces*, to appear.
- [19] O. L. MANGASARIAN AND S. FROMOVITZ, *The Fritz John necessary optimality conditions in the presence of equality and inequality constraints*, J. Math. Anal. Appl., 17 (1967), pp. 37–47.
- [20] Y. NAGAHISA AND Y. SAKAWA, *Nonlinear programming in Banach spaces*, J. Optimization Theory Appl., 4 (1969), pp. 182–190.
- [21] K. RITTER, *Optimization in linear spaces I, II, III*, Math. Ann., 182 (1969), pp. 189–206, 183 (1969), pp. 169–180.
- [22] A. P. AND W. J. ROBERTSON, *Topological Vector Spaces*, Cambridge University Press, London, 1964.
- [23] R. T. ROCKAFELLAR, *An extension of Fenchel's duality theorem for convex functions*, Duke Math. J., 33 (1966), pp. 81–90.
- [24] ———, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [25] P. P. VARAIYA, *Nonlinear programming in Banach spaces*, SIAM J. Appl. Math., 19 (1967), pp. 239–244.
- [26] S. YAMAMURO, *Differential Calculus in Topological Linear Spaces*, Springer-Verlag, New York, 1974.
- [27] S. ZLOBEC, *Asymptotic Kuhn–Tucker conditions for mathematical programming problems in Banach space*, this Journal, 8 (1970), pp. 505–512.
- [28] S. ZLOBEC AND H. MASSAM, *Various definitions of the derivative in mathematical programming*, Math. Programming, 7 (1974), pp. 144–161.

ON DECOMPOSITION OF GENERATORS*

JERZY ZABCZYK†

Abstract. Let \mathcal{A} be the infinitesimal generator of a strongly continuous semigroup on a Banach space \mathcal{E} . Two classes of bounded operators \mathcal{P} on \mathcal{E} are introduced for which the operators $\mathcal{A}\mathcal{P}$ and $\mathcal{P}\mathcal{A}$ also generate semigroups on \mathcal{E} . It is shown that many important operators can be decomposed into the form $\mathcal{A}\mathcal{P} + \mathcal{R}$ or $\mathcal{P}\mathcal{A} + \mathcal{R}$ where \mathcal{A} is a generator of a simpler structure and \mathcal{R} is a bounded operator. Applications of decompositions of this type to infinite dimensional system theory are discussed.

1. Introduction. If \mathcal{A} is the infinitesimal generator of a strongly continuous semigroup on a Banach space \mathcal{E} and \mathcal{P} is a bounded operator on \mathcal{E} , then the operators $\mathcal{A}\mathcal{P}$ and $\mathcal{P}\mathcal{A}$ defined on $\mathcal{P}^{-1}\mathcal{D}(\mathcal{A})$ and $\mathcal{D}(\mathcal{A})$ respectively are not, in general, generators on \mathcal{E} even if \mathcal{P} is an isomorphism. The simplest counterexample is probably $\mathcal{P} = -\mathcal{I}$ (identity): the operators \mathcal{A} and $-\mathcal{A}$ generate semigroups iff the operator \mathcal{A} generates a group. On the other hand if the operator $\mathcal{I} - \mathcal{P}$ has sufficiently small norm and \mathcal{A} generates a holomorphic semigroup then $\mathcal{P}\mathcal{A}$ and $\mathcal{A}\mathcal{P}$ are generators of holomorphic semigroups, as easily follows from [7, Thm 2.4, p. 497]. In § 2 of this paper we give different sufficient conditions which imply that $\mathcal{P}\mathcal{A}$ and $\mathcal{A}\mathcal{P}$ are generators.

From Theorem 1 of § 2 we obtain a corollary of some significance in system theory: if \mathcal{A} generates a holomorphic semigroup and \mathcal{F} is an \mathcal{A} -compact operator then also the operator $\mathcal{A} + \mathcal{F}$ generates a holomorphic semigroup. The results of § 2 are applied in the next sections, the applications being of two types. In § 3 we show that some unbounded operators generate semigroups, by decomposing them into the form $\mathcal{A}\mathcal{P} + \mathcal{R}$ or $\mathcal{P}\mathcal{A} + \mathcal{R}$. This way we obtain simple proofs that Cauchy problems for the heat equation with nonlocal or nonhomogenous boundary conditions are well posed. We obtain also a "representation" theorem for delay equations in the spirit of the papers [1] and [4]. In § 4 we show how it is possible to use the decomposition of generators to solve or to simplify some problems in infinite dimensional system theory.

The main contributions of the paper are the perturbation results contained in Theorem 1, Proposition 1 and 2 and examples of decomposition procedures contained in § 4. The author hopes that decomposition of generators provides a method of discussing in a unified way many concrete problems and examples scattered throughout the control literature.

The present paper is a rewritten version of the report [12].

2. General results. Let \mathcal{E} be a Banach space and \mathcal{A} the infinitesimal generator of a strongly continuous semigroup on \mathcal{E} . We prove in this section that, under certain conditions

- (1) *the operator $\mathcal{A}\mathcal{P} + \mathcal{R}$ with the domain $\mathcal{P}^{-1}\mathcal{D}(\mathcal{A})$, and the operator $\mathcal{P}\mathcal{A} + \mathcal{R}$ with the domain $\mathcal{D}(\mathcal{A})$ generate semigroups on \mathcal{E} .*

The conditions imposed are of two types. We require that either a) the operator $\mathcal{F} = \mathcal{I} - \mathcal{P}$ is compact or b) it is of a special structure. We also derive two corollaries concerning finite dimensional perturbations and semigroups on product spaces. Throughout the paper \mathcal{P} and \mathcal{R} are bounded operators.

* Received by the editors March 18, 1977.

† Control Theory Centre, University of Warwick, Coventry CV4 7AL, England, on leave from the Institute of Mathematics, Polish Academy of Sciences, Warsaw, Poland.

THEOREM 1. a) Define $\mathcal{F} = \mathcal{I} - \mathcal{P}$ and assume that \mathcal{E} is a reflexive Banach space, then (1) holds provided

- 1) \mathcal{A} generates a holomorphic semigroup and \mathcal{F} is a compact operator or,
 - 2) \mathcal{P} is an invertible operator and $\mathcal{F}(\mathcal{E}) \subset \mathcal{D}(\mathcal{A})$.
- b) The property (1) holds also if

$$\mathcal{E} = \begin{matrix} X \\ \times \\ Y \end{matrix}, \quad \mathcal{A} = \begin{pmatrix} A, & 0 \\ 0, & B \end{pmatrix}, \quad \mathcal{P} = \begin{pmatrix} I, & -F \\ 0, & I \end{pmatrix}$$

where X, Y are Banach spaces, A and B generators on X and Y respectively, and F a bounded operator from X into Y . It is assumed that

$$\mathcal{D}(\mathcal{A}) = \begin{matrix} \mathcal{D}(A) \\ \times \\ \mathcal{D}(B) \end{matrix}.$$

Before proving the theorem we recall that an operator \mathcal{F} is a compact operator (in narrow sense) if and only if there exists a sequence (\mathcal{G}_n) of finite dimensional operators such that $\|\mathcal{F} - \mathcal{G}_n\| \rightarrow 0$ as $n \rightarrow \infty$. A semigroup $(\mathcal{T}_t)_{t \geq 0}$ is holomorphic, see [7, p. 488], if and only if the resolvent set $\rho(\mathcal{A})$ contains the set

$$\left\{ \lambda : |\arg(\lambda - \lambda_0)| < \frac{\pi}{2} + \omega_0, \lambda \neq \lambda_0 \right\}$$

for some $\omega_0 > 0$ and $\lambda_0 \in (-\infty, +\infty)$ and for any $\varepsilon > 0$ there exists a constant M_ε such that

$$|\mathcal{R}_\lambda| \leq \frac{M_\varepsilon}{|\lambda - \lambda_0|} \quad \text{if} \quad |\arg(\lambda - \lambda_0)| \leq \frac{\pi}{2} + \omega_0 - \varepsilon.$$

Proof of the theorem. a) We can obviously assume that $\mathcal{R} = 0$. Reflexivity of the space \mathcal{E} implies that, see [11, p. 233], the operator \mathcal{A}^* generates a semigroup on \mathcal{E}^* and thus that the set $\mathcal{D}(\mathcal{A}^*)$ is dense in \mathcal{E}^* . Let us assume that 1) holds and that \mathcal{F} is a finite dimensional operator, representable in a form:

$$\mathcal{F} = \sum_{j=1}^m e_j f_j, \quad e_j \in \mathcal{E}, \quad f_j \in \mathcal{E}^*, \quad j = 1, 2, \dots, m.$$

If now $\delta > 0$ is any positive number then there exist functionals $\bar{f}_j \in \mathcal{D}(\mathcal{A}^*)$ such that $\sum_{j=1}^m |f_j - \bar{f}_j| |e_j| < \delta$. Therefore for $x \in \mathcal{D}(\mathcal{A})$

$$\begin{aligned} |\mathcal{F}(\mathcal{A}x)| &\leq \sum_{j=1}^m |e_j(f_j - \bar{f}_j)(\mathcal{A}x)| + \sum_{j=1}^m |\mathcal{A}^* \bar{f}_j(x)| \\ &\leq \delta |\mathcal{A}x| + \left(\sum_{j=1}^m |\mathcal{A}^* \bar{f}_j| \right) |x|. \end{aligned}$$

This proves that the operator $\mathcal{F}\mathcal{A}$ has \mathcal{A} -bound zero and consequently by Corollary 2.5 in [7, p. 498] the operator $\mathcal{P}\mathcal{A} = \mathcal{A} - \mathcal{F}\mathcal{A}$ generates a holomorphic semigroup. On the other hand if the operator \mathcal{P} is invertible then $\mathcal{A}\mathcal{P} = \mathcal{P}^{-1}(\mathcal{P}\mathcal{A})\mathcal{P}$ and we see also that $\mathcal{A}\mathcal{P}$ generates a holomorphic semigroup on \mathcal{E} . Since the domain $\mathcal{D}(\mathcal{A})$ is dense in \mathcal{E} we can find a finite dimensional operator \mathcal{F}_0 into $\mathcal{D}(\mathcal{A})$ such that $\|\mathcal{F} - \mathcal{F}_0\| < 1$. Consequently $\mathcal{P} = \mathcal{I} - (\mathcal{F} - \mathcal{F}_0) - \mathcal{F}_0 = \mathcal{P}_0 - \mathcal{F}_0$, \mathcal{P}_0 is invertible and $\mathcal{A}\mathcal{P} = \mathcal{A}\mathcal{P}_0 - \mathcal{A}\mathcal{F}_0$. Now since $\mathcal{A}\mathcal{F}_0$ is a bounded operator and $\mathcal{A}\mathcal{P}_0$ generates, by the above considerations, a holomorphic semigroup, the operator $\mathcal{A}\mathcal{P}$ is also a holomorphic generator. Assume now that \mathcal{F} is a compact operator approximated by a sequence of

finite dimensional operators \mathcal{G}_n , $n = 1, 2, 3, \dots$. Since

$$\begin{aligned} |\mathcal{F}\mathcal{A}(x)| &\leq |(\mathcal{F} - \mathcal{G}_n)(\mathcal{A}x)| + |\mathcal{G}_n(\mathcal{A}x)| \\ &\leq |\mathcal{F} - \mathcal{G}_n| |\mathcal{A}x| + |\mathcal{G}_n(\mathcal{A}x)|, \end{aligned}$$

the operator $\mathcal{F}\mathcal{A}$ has \mathcal{A} -bound equal zero and consequently the operator $\mathcal{P}\mathcal{A} = \mathcal{A} - \mathcal{F}\mathcal{A}$ is an infinitesimal generator on \mathcal{E} . Density of $\mathcal{D}(\mathcal{A})$ implies that there exists a finite dimensional operator \mathcal{G} into $\mathcal{D}(\mathcal{A})$ such that $|\mathcal{F} - \mathcal{G}| < 1$. Thus the operator $\mathcal{I} - (\mathcal{F} - \mathcal{G})$ is invertible and $\mathcal{A}\mathcal{G}$ is a bounded operator and we see that $\mathcal{A}\mathcal{P}$ is equal to $\mathcal{A}(\mathcal{I} - (\mathcal{F} - \mathcal{G})) - \mathcal{A}\mathcal{G}$ and generates a semigroup as well.

Suppose now that 2) holds, then by the closed graph theorem the operator $\mathcal{A}\mathcal{F}$ is bounded, therefore $\mathcal{A}\mathcal{P} = \mathcal{A} - \mathcal{A}\mathcal{F}$ is generator. But then also the operator $\mathcal{P}\mathcal{A}$ is generator because \mathcal{P} is an invertible operator.

b) As before put $\mathcal{R} = 0$ and denote by $(T_t)_{t \geq 0}$ and $(S_t)_{t \geq 0}$ semigroups generated respectively by A and B . Let us define for every $t \geq 0$ operators \mathcal{T}_t on \mathcal{E} by the formula:

$$\mathcal{T}_t \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} FS_t y - T_t Fy + T_t x \\ S_t y \end{pmatrix}.$$

Then it is easy to check that $(\mathcal{T}_t)_{t \geq 0}$ is a strongly continuous semigroup on \mathcal{E} . From the identity:

$$\frac{1}{t} \left(\mathcal{T}_t \begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} x \\ y \end{pmatrix} \right) = \begin{pmatrix} F \left(\frac{1}{t} (S_t y - y) \right) + \frac{1}{t} [T_t(x - Fy) - (x - Fy)] \\ \frac{1}{t} (S_t y - y) \end{pmatrix}$$

it follows that the generator of the semigroup $(\mathcal{T}_t)_{t \geq 0}$ is defined on the set $\left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in \mathcal{E}, y \in \mathcal{D}(B), x - Fy \in \mathcal{D}(A) \right\}$, which is equal to $\mathcal{P}^{-1}\mathcal{D}(\mathcal{A})$, and that it is equal to $\mathcal{A}\mathcal{P}$. The invertibility of the operator \mathcal{P} implies that the operator $\mathcal{P}\mathcal{A}$ also generates a semigroup on \mathcal{E} .

Remark 1. The reflexivity of the space \mathcal{E} was used only to ensure density of the set $\mathcal{D}(\mathcal{A}^*)$ in \mathcal{E}^* . Thus the part (a) of the theorem remains true if the reflexivity assumption is replaced by the continuity of the conjugate semigroup $(\mathcal{J}_t^*)_{t \geq 0}$.

The following consequence of the Theorem 1 is of some significance in applications, see Remark 2.

PROPOSITION 1. *Let \mathcal{A} generate a holomorphic semigroup on a reflexive space \mathcal{E} and let an operator \mathcal{F} defined on $\mathcal{D}(\mathcal{A})$ be \mathcal{A} -compact then the operator $\mathcal{A} + \mathcal{F}$ also generates a holomorphic semigroup on \mathcal{E} .*

Proof. Assume first that \mathcal{F} is a finite dimensional \mathcal{A} -bounded operator and define on the graph $\Gamma(\mathcal{A})$ the operator $\tilde{\mathcal{F}}(x, y) = \mathcal{F}(x)$ for all $x \in \mathcal{D}(\mathcal{A})$ and $y = \mathcal{A}x$. Since $\tilde{\mathcal{F}}$ is a bounded finite dimensional operator on closed subspace $\Gamma(\mathcal{A})$ it has a continuous finite dimensional extension to the whole $\mathcal{E} \times \mathcal{E}$, (by the Hahn-Banach theorem). This extension is of the form $\tilde{\mathcal{F}}(x, y) = \mathcal{F}_0 x + \mathcal{F}_1 y$ where \mathcal{F}_0 and \mathcal{F}_1 are finite dimensional operators. Consequently for every $x \in \mathcal{D}(\mathcal{A})$, $\mathcal{F}(x) = \mathcal{F}_0 x + \mathcal{F}_1 \mathcal{A}x$. Thus in the case of \mathcal{F} finite dimensional, Proposition 1 follows from the part (a) of the theorem. If now \mathcal{F} is an \mathcal{A} -compact (in the narrow sense) operator then by choosing an appropriate \mathcal{A} -bounded finite dimensional operator \mathcal{G} the \mathcal{A} -bound of the operator $\tilde{\mathcal{F}} = \mathcal{F} - \mathcal{G}$ can be made arbitrarily small and therefore, $\mathcal{A} + \tilde{\mathcal{F}}$ generates holomorphic semigroup.

Moreover if the \mathcal{A} -bound of \mathcal{F} is less than 1 then \mathcal{G} is also continuous with respect to $\mathcal{A} + \mathcal{F}$ and thus $\mathcal{A} + \mathcal{F} = \mathcal{A} + \mathcal{F} + \mathcal{G}$ generates also a holomorphic semigroup.

Remark 2. Many "systems" can be modeled by the equation $\dot{x} = \mathcal{A}x + \mathcal{B}u$ where \mathcal{A} is a generator of a holomorphic semigroup and \mathcal{B} is a compact operator from U into \mathcal{E} . Proposition 1 allows us to consider all feedback laws $u = Cx$ where C is any \mathcal{A} -bounded operator.

Remark 3. In the course of the proof of the above Proposition 1 and Theorem 1 we have shown that if the set $\mathcal{D}(\mathcal{A}^*)$ is dense in \mathcal{E}^* then every finite dimensional \mathcal{A} -bounded operator has \mathcal{A} -bound zero. The special case of \mathcal{A} -bounded functional $\varphi(x) = \int_0^1 x(s)\mu(ds)$ where μ is a delta measure on $[0, 1]$, $\mathcal{A} = d/(ds)$ and $\mathcal{E} = L^p[0, 1]$, $p > 1$, was proved in [7, p. 193] by means of specific properties of the operator d/ds .

Proposition 2 below will be used in the study of delay equations in the next section. Recently A. Chojnowska-Michalik [2] applied a generalized version of the Proposition 2 to investigate properties of a general class of stochastic delay equations.

In the formulation of Proposition 2 the operator \mathcal{A} is defined on the set

$$(2) \quad \mathcal{D}(\mathcal{A}) = \left\{ \begin{pmatrix} x \\ y \end{pmatrix}; \begin{pmatrix} x - Fy \\ y \end{pmatrix} \in \bigtimes_{\mathcal{D}(A)}^{\mathcal{D}(B)} \right\}$$

by the formula

$$(3) \quad \mathcal{A} = \begin{pmatrix} A, 0 \\ 0, B \end{pmatrix} \begin{pmatrix} I, -F \\ 0, I \end{pmatrix} + \begin{pmatrix} 0, 0 \\ C, 0 \end{pmatrix}$$

where A, B generate semigroups $(T_t)_{t \geq 0}, (S_t)_{t \geq 0}$ respectively on X and Y , F is a linear bounded operator and C is, in general, an unbounded and nonlinear operator from $\mathcal{D}(C) = \{x \in X, x - Fy \in \mathcal{D}(A) \text{ for some } y \in \mathcal{D}(B)\}$ into Y . It follows from Theorem 1, b) that \mathcal{A} generates a semigroup on

$$\mathcal{E} = \bigtimes_{Y}^X$$

provided that C is a bounded linear operator.

PROPOSITION 2. Let $z(t) = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}$, $t \geq 0$ be a continuously differentiable solution of the equation

$$\dot{z}(t) = \mathcal{A}z(t),$$

$$z(0) = \begin{pmatrix} x \\ y \end{pmatrix}, \quad t \geq 0.$$

Then

$$(4) \quad x(t) = T_t x - A \left(\int_0^t T_{t-s} F y(s) ds \right),$$

$$(5) \quad \dot{y}(t) = B y(t) + C \left(T_t x - A \int_0^t T_{t-s} F y(s) ds \right).$$

In the proof we shall need the following lemma, which easily follows from [7, Thm 1.19, p. 486].

LEMMA 1. If $g(t)$, $t \geq 0$ is a continuously differentiable function, $f(t)$, $t \geq 0$ a continuous function from $[0, +\infty)$ into X and

$$\dot{g}(t) = A g(t) + f(t), \quad t \geq 0,$$

then

$$g(t) = T_t g(0) + \int_0^t T_{t-s} f(s) ds.$$

Proof of Proposition 2. From the assumptions of the proposition it follows that $\dot{x}(t) = A(x(t) - Fy(t))$ and consequently

$$\frac{d}{dt}(x(t) - Fy(t)) = A(x(t) - Fy(t)) - F\dot{y}(t).$$

From Lemma 1

$$x(t) - Fy(t) = T_t x - T_t Fy - \int_0^t T_{t-s} F\dot{y}(s) ds.$$

But again, [7, Thm 1.19, p. 486] implies

$$\begin{aligned} \frac{d}{dt} \int_0^t T_{t-s} Fy(s) ds &= A \int_0^t T_{t-s} Fy(s) ds + Fy(t), \\ \frac{d}{dt} \int_0^t T_{t-s} Fy(s) ds &= T_t Fy + \int_0^t T_{t-s} F\dot{y}(s) ds. \end{aligned}$$

Therefore finally we obtain

$$\begin{aligned} x(t) &= T_t x - T_t Fy + Fy(t) - \left[A \int_0^t T_{t-s} Fy(s) ds + Fy(t) - T_t Fy \right] \\ &= T_t x - A \int_0^t T_{t-s} Fy(s) ds. \end{aligned}$$

Example 1. Assume that Y is any B -space and X a B -space of functions from $[-h, 0]$ into Y satisfying the following two conditions:

- a) all continuous functions from $[-h, 0]$ into Y are contained in X ,
- b) the left shift semigroup $T_t: X \rightarrow X$, $T_t y(s) = y(t+s)$ if $t+s \leq 0$ and 0 otherwise, is a strongly continuous semigroup on X .

Then for any continuous function $y: [0, +\infty) \rightarrow Y$, $x \in X$ and $t \geq 0$ the function $x(t)(\cdot)$ given by (4) is exactly the “segment” of y :

$$x(t)(s) = \begin{cases} x(t+s) & \text{if } t+s \leq 0, \\ y(t+s) & \text{if } t+s \geq 0, \end{cases}$$

provided $(F\bar{y})(s) \equiv \bar{y}$ on $[-h, 0]$, $\bar{y} \in Y$.

3. Examples of decomposition procedures. In this section we decompose some unbounded operators into the form $\mathcal{A}_0 \mathcal{P} + \mathcal{R}$, where \mathcal{A}_0 is a generator on a Banach space, proving this way that a given operator generates a semigroup. We start the discussion with parabolic equations on $[0, 1]$. For more general examples we refer to [13]. We show how Theorem 1 allows us to “perturb” the boundary conditions.

3.1. Heat equation on $[0, 1]$. In Propositions 3 and 4 below $W^{2,2} = W^{2,2}[0, 1; R^n]$ and the operator $A_0 = d^2/ds^2$ is defined either on

$$(a) \quad \mathcal{D}(A_0) = \{x \in W^{2,2}; x(0) = x(1) = 0\}$$

or on

$$(b) \quad \mathcal{D}(A_0) = \left\{ x \in W^{2,2}; \frac{dx}{ds}(0) = x(1) = 0 \right\}.$$

It is well known that A_0 is a self-adjoint operator on $L^2[0, 1; R^n]$ and generates a holomorphic semigroup.

By $\mathcal{L}(R^n)$ we denote the space of all $n \times n$ matrices. If $f(\cdot) \in L^2[0, 1; \mathcal{L}(R^n)]$ then $\int f$ denotes the transformation $\int fx = \int_0^1 f(s)x(s) ds$ from $L^2[0, 1; R^n]$ into R^n .

PROPOSITION 3. For every function $f \in L^2[0, 1; \mathcal{L}(R^n)]$ the operator $\mathcal{A} = d^2/ds^2$ defined on

$$(a.1) \quad \mathcal{D}(\mathcal{A}) = \left\{ x \in W^{2,2}; x(0) = \int fx, x(1) = 0 \right\},$$

or on

$$(b.1) \quad \mathcal{D}(\mathcal{A}) = \left\{ x \in W^{2,2}; \frac{dx}{ds}(0) = \int fx, x(1) = 0 \right\}$$

can be decomposed into the form $\mathcal{A} = A_0\mathcal{P}$ where $\mathcal{D}(A_0)$ is defined respectively by (a) or (b) and $\mathcal{P} - \mathcal{P}$ is a finite dimensional operator. Thus \mathcal{A} generates a holomorphic semigroup.

Proof. We consider, for instance, the case (b.1). Let us define $e(s) = -s\mathcal{I} + \mathcal{I}$ (\mathcal{I} identity matrix) and $\mathcal{P}x = x + e \int fx$. Then $x \in \mathcal{D}(A_0\mathcal{P})$ iff $x + e \int fx \in W^{2,2}$ and $(d/ds)(x + e \int fx)(0) = (x + e \int x)(1) = 0$. Or equivalently if and only if $x \in W^{2,2}$, $(dx/ds)(0) = (-de/ds) \int fx$ and $x(0) = -e(1) \int fx$. It is also obvious that

$$A_0(x + e \int fx) = \frac{d^2x}{ds^2}.$$

This finishes the proof.

Let us now put

$$\mathcal{E} = \begin{matrix} X \\ \times \\ Y \end{matrix}$$

where $X = L^2[0, 1; R^n]$ and

$$Y = \begin{matrix} R^n \\ \times \\ R^n \end{matrix}.$$

Assume also that B_{ij} , $i, j = 1, 2$ are given $n \times n$ matrices and C_i , $i = 1, 2$ functions belonging to $L^2[0, 1; \mathcal{L}(R^n)]$.

PROPOSITION 4. The operator \mathcal{A} defined by

$$(a.2) \quad \mathcal{A} \begin{pmatrix} x \\ x(0) \\ x(1) \end{pmatrix} = \begin{pmatrix} \frac{d^2x}{ds^2} \\ B_{11}x(0) + B_{12}x(1) + \int C_1x \\ B_{21}x(0) + B_{22}x(1) + \int C_2x \end{pmatrix}, \quad \mathcal{D}(\mathcal{A}) = \left\{ \begin{pmatrix} x \\ x(0) \\ x(1) \end{pmatrix}, x \in W^{2,2} \right\}$$

$$(b.2) \quad \mathcal{A} \begin{pmatrix} x \\ \frac{dx}{ds}(0) \\ x(1) \end{pmatrix} = \begin{pmatrix} \frac{d^2x}{ds^2} \\ B_{11} \frac{dx}{ds}(0) + B_{12}x(1) + \int C_1x \\ B_{21} \frac{dx}{ds}(0) + B_{22}x(1) + \int C_2x \end{pmatrix}, \quad \mathcal{D}(\mathcal{A}) = \left\{ \begin{pmatrix} x \\ \frac{dx}{ds}(0) \\ x(1) \end{pmatrix}, x \in W^{2,2} \right\}$$

can be decomposed into the form

$$\mathcal{A} = \begin{pmatrix} A_0, & 0 \\ 0 & \begin{pmatrix} B_{11}, B_{12} \\ B_{21}, B_{22} \end{pmatrix} \end{pmatrix} \mathcal{P} + \mathcal{R}, \quad \text{where } \mathcal{P} = \begin{pmatrix} 1, -F \\ 0, 1 \end{pmatrix}, \quad \mathcal{R} = \begin{pmatrix} 0, & 0 \\ \begin{pmatrix} \int C_1 \\ \int C_2 \end{pmatrix}, & 0 \end{pmatrix},$$

and F is a bounded operator. Thus \mathcal{A} generates a holomorphic semigroup on \mathcal{E} .

Proof. We consider only the operator \mathcal{A} defined by (b.2). The case (a.2) can be treated in an analogous way. Let us define the transformation $F: Y \rightarrow X$ by the formula

$$F \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}(s) = y_1 s + y_2 - y_1, \quad s \in [0, 1].$$

Then

$$x - F \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \in \mathcal{D}(A_0) \quad \text{iff} \quad x \in W^{2,2} \text{ and } y_1 = \frac{dx}{ds}(0), y_2 = x(1).$$

It is also clear that

$$A_0 \left(x - F \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right) = \frac{d^2 x}{ds^2}.$$

Remark 4. The nonhomogenous equation

$$(6) \quad \dot{z} = \mathcal{A}z + \mathcal{B}u, \quad z_0 \in \mathcal{E}, \quad u \in L^2[0, t_0; u],$$

where \mathcal{A} is given by (a.2) with $B_{1,2} = B_{2,1} = 0$, and $C_1 \equiv C_2 \equiv 0$ appeared in [5], in an implicit way. The semigroup formulation presented here gives a simple proof that the equation (6) has a unique weak solution, (this problem was treated in [5] in a classical way). It enables us also to treat the regulator problem for (6) from the more general point of view and write almost immediately the corresponding Riccati equations, see § 4 of this paper. On the other hand, the study of the Riccati equation, which was carried out in [5] needs a special effort and does not follow from the general theory.

3.2. Equations of nuclear reactor dynamics. Even decompositions of the form $\mathcal{A} = \mathcal{A}_0 + \mathcal{R}$ are sometimes of some practical use. This is, for instance, in the case of the linearized version of the nuclear reactor dynamics equation. This equation is of the form, see [6], [8]:

$$(7) \quad \frac{d}{dt} \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} A_0, R_0 \\ R_1, R_2 \end{pmatrix} \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \mathcal{A} \begin{pmatrix} x(t) \\ y(t) \end{pmatrix},$$

where

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} \in \mathcal{E} = \begin{matrix} L^2(\Omega, R^n) \\ \times \\ R^n \end{matrix} = \begin{matrix} X \\ \times \\ Y \end{matrix},$$

R_0, R_1, R_2 are bounded operators and A_0 is an elliptic operator which generates a semigroup on X . Thus

$$\mathcal{A} = \begin{pmatrix} A_0, 0 \\ 0, 0 \end{pmatrix} + \begin{pmatrix} 0, R_0 \\ R_1, R_2 \end{pmatrix},$$

and consequently \mathcal{A} generates a holomorphic semigroup. In [6] this fact was proved by checking coercitivity of \mathcal{A} and in [8] by defining \mathcal{A} through appropriate bilinear form and checking that it satisfies Lion's conditions.

3.3. Delay equations. Let us fix the notations $X = L^2[-h, 0; R^n]$, $Y = R^n$,

$$\mathcal{E} = \begin{matrix} X \\ \times \\ Y \end{matrix}.$$

The semigroup treatment of the delay equation

$$(8) \quad \dot{y}(t) = \int_{-h}^0 N(ds)y(t+s), \quad t \geq 0,$$

in the space \mathcal{E} was initiated in papers [1] and [4]. It was shown in [1] and [4] that, under some condition on the matrix valued measure N , the operator

$$(9) \quad \mathcal{A} \begin{pmatrix} x \\ x(0) \end{pmatrix} = \begin{pmatrix} \frac{dx}{ds} \\ \int_{-h}^0 N(ds)x(s) \end{pmatrix}, \quad \mathcal{D}(\mathcal{A}) = \left\{ \begin{pmatrix} x \\ x(0) \end{pmatrix}; x \in W^{1,2} \right\}$$

defines a semigroup \mathcal{T}_t , $t \geq 0$ on \mathcal{E} , with the second coordinate satisfying (8). Under the condition N has no atoms on $[-h, 0)$, this result follows from the general results of § 2 and is contained in Proposition 5 and Corollary 1. The decomposition obtained for the general case in Proposition 6 will be used in the next section. The contents of Proposition 5 and Proposition 6 can be summarized as follows:

Solutions of delay equations are linear images of trajectories of the left shift semigroup. We recall that the left shift semigroup T_t , $t \geq 0$ on X is defined by the formula: $T_t x(s) = x(t+s)$, if $t+s \leq 0$, and 0 otherwise. The generator of $(T_t)_{t \geq 0}$ is $A_0 = d/ds$ with the domain $\mathcal{D}(A_0) = \{x \in W^{1,2}, x(0) = 0\}$. In the formulation below N_0 is an $n \times n$ matrix and $N_1 \in L^2[-h, 0; \mathcal{L}(R^n)]$. Other notations are as in § 3.1.

PROPOSITION 5. *The operator \mathcal{A} defined by*

$$(10) \quad \mathcal{A} \begin{pmatrix} x \\ x(0) \end{pmatrix} = \begin{pmatrix} \frac{dx}{ds} \\ N_0 x(0) + \int N_1 x \end{pmatrix}, \quad \text{on } \mathcal{D}(\mathcal{A}) = \left\{ \begin{pmatrix} x \\ x(0) \end{pmatrix}, x \in W^{1,2} \right\}$$

can be decomposed into the form:

$$\mathcal{A} = \begin{pmatrix} A_0 & 0 \\ 0 & N_0 \end{pmatrix} \begin{pmatrix} I & -F \\ 0 & I \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ \int N_1 & 0 \end{pmatrix},$$

and thus it generates a strongly continuous semigroup on \mathcal{E} .

Proof. The proposition follows in an analogous way as Proposition 4 if we define the “boundary” operator F by the formula:

$$(Fy)(s) \equiv y, \quad \text{for all } s \in [-h, 0] \text{ and } \mathcal{R} = \begin{pmatrix} 0 & 0 \\ \int N_1 & 0 \end{pmatrix}.$$

COROLLARY 1. *Let $y(t)$, $t \geq 0$ be the second coordinate of the trajectory*

$$\left(\mathcal{T}_t \begin{pmatrix} x \\ y \end{pmatrix} \right)_{t \geq 0}, \quad \begin{pmatrix} x \\ y \end{pmatrix} \in \mathcal{D}(\mathcal{A})$$

of the semigroup generated by (10), then

$$\dot{y}(t) = N_0 y(t) + \int_{-h}^0 N_1(s)y(t+s) ds.$$

Proof. The statement of the corollary follows from Example 1.

Remark 5. As was pointed out to us by R. Vinter, the same approach “from semigroups to delay equations” was carried out by G. F. Webb [9, Prop. 5.2, Prop. 5.3] for nonlinear delay equations. However our Proposition 2 applies also to nonlinear delay equations, as follows from Example 1. Moreover our formulation, which involves the boundary operator F , allows us to consider noncoercive operators N_0 , Banach spaces Y , and general operators C not treated in [9].

PROPOSITION 6. *Let N be any finite measure with values in $\mathcal{L}(R^n)$. Then the operator \mathcal{A} defined by*

$$\mathcal{A} \begin{pmatrix} x \\ x(0) \end{pmatrix} = \begin{pmatrix} \frac{dx}{ds} \\ \int_{-h}^0 N(ds)x(s) \end{pmatrix}, \quad \mathcal{D}(\mathcal{A}) = \left\{ \begin{pmatrix} x \\ x(0) \end{pmatrix}; x \in W^{1,2} \right\}$$

can be decomposed into the form

$$(11) \quad \mathcal{A} = \begin{pmatrix} I, & 0 \\ -G, & I \end{pmatrix} \begin{pmatrix} A_0, & 0 \\ 0, & N\{0\} \end{pmatrix} \begin{pmatrix} I, & -F \\ 0, & I \end{pmatrix}$$

where F and G are some bounded operators.

Proof. It is well known that if N is a finite measure on $[-h, 0]$, g its distribution function: $g(s) = N[-h, s]$, $s \in [-h, 0]$ and $x \in W^{1,2}[-h, 0]$ then:

$$\int_{-h}^0 N(ds)x(s) = N\{0\}x(0) - \int_{-h}^0 g(s)\dot{x}(s) ds.$$

Consequently

$$\int_{-h}^0 N(ds)x(s) = N\{0\}x(0) - \int_{-h}^0 g(s)A_0(x - Fy)(s) ds$$

where

$$Fy(s) = y, \quad s \in [-h, 0].$$

Therefore decomposition (11) holds with $F: Y \rightarrow X$, $Fy(s) = y$ for $s \in [-h, 0]$ and $G: X \rightarrow Y$, $Gx = \int_{-h}^0 gx = \int_{-h}^0 g(s)x(s) ds$.

4. Applications to infinite dimensional system theory. In this section we illustrate how the obtained results can be used to solve, or to simplify, some problems in system theory.

4.1. Regulator problem. It is well known that the solution of the linear regulator problem for the infinite dimensional system

$$(12) \quad \dot{z} = \mathcal{A}z + \mathcal{B}u$$

can be expressed through the solution $P = P(t)$, $t \geq 0$ of the inner product Riccati equation, see [10]. If the operator \mathcal{A} can be decomposed into the form $\mathcal{A} = \mathcal{A}_0\mathcal{P}$ or $\mathcal{A} = \mathcal{Q}\mathcal{A}_0\mathcal{P}$ then the Riccati equation can be written in the terms of the operator \mathcal{A}_0 and its domain $\mathcal{D}(\mathcal{A}_0)$ only. Let us take, for instance, into account the situation where the “regulation” is implemented through a new dynamical system. This is the case of the controlled delay systems and systems considered by M. Giurgiu in [5]. The state space is then of the form

$$\mathcal{E} = \begin{matrix} X \\ \times \\ Y \end{matrix}$$

and the generator \mathcal{A} is given by

$$\mathcal{A} = \begin{pmatrix} A, 0 \\ 0, B \end{pmatrix} \begin{pmatrix} I, -F \\ 0, I \end{pmatrix} + \mathcal{R};$$

compare Proposition 4 and Proposition 5.

After standard transformations we obtain the Riccati equation

$$\begin{aligned} \frac{d}{dt} \left\langle \begin{pmatrix} I, 0 \\ F^*, I \end{pmatrix} P \begin{pmatrix} I, F \\ 0, I \end{pmatrix} \omega, z \right\rangle &= \left\langle \begin{pmatrix} I, 0 \\ F^*, I \end{pmatrix} P \begin{pmatrix} A, 0 \\ 0, B \end{pmatrix} \omega, z \right\rangle \\ &+ \left\langle P \begin{pmatrix} I, F \\ 0, I \end{pmatrix} \omega, \begin{pmatrix} A, 0 \\ 0, B \end{pmatrix} z \right\rangle + \left\langle \begin{pmatrix} I, 0 \\ F^*, I \end{pmatrix} (P\mathcal{R} + \mathcal{R}^*P + S) \begin{pmatrix} I, F \\ 0, I \end{pmatrix} \omega, z \right\rangle \\ &- \left\langle \begin{pmatrix} I, 0 \\ F^*, I \end{pmatrix} P\mathcal{B}^*R^{-1}\mathcal{B}P \begin{pmatrix} I, F \\ 0, I \end{pmatrix} \omega, z \right\rangle, \quad \omega, z \in \begin{matrix} \mathcal{D}(A) \\ \times \\ Y \end{matrix}, \quad t \geq 0. \end{aligned}$$

Taking as (ω, z) pairs

$$\left(\begin{pmatrix} x \\ 0 \end{pmatrix}, \begin{pmatrix} \bar{x} \\ 0 \end{pmatrix} \right), \quad \left(\begin{pmatrix} 0 \\ y \end{pmatrix}, \begin{pmatrix} 0 \\ \bar{y} \end{pmatrix} \right)$$

and

$$\left(\begin{pmatrix} x \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ y \end{pmatrix} \right)$$

we derive three equations (a system of Riccati equations) which were obtained and studied for delay case in [4] and for parabolic case in [5].

4.2. Conjugate systems. Assume that \mathcal{P}, \mathcal{Q} are bounded operators on \mathcal{E} , \mathcal{P} is an invertible operator and \mathcal{A} a closed, densely defined operator on \mathcal{E} . Then the following proposition, see [7, p. 195], holds.

PROPOSITION 7. *The domain of the operator $(\mathcal{Q}\mathcal{A}\mathcal{P})^*$ is exactly $(\mathcal{Q}^*)^{-1}\mathcal{D}(\mathcal{A}^*)$ and*

$$(\mathcal{Q}\mathcal{A}\mathcal{P})^* = \mathcal{P}^* \mathcal{A}^* \mathcal{Q}^*.$$

From this proposition it follows that if

$$\mathcal{A} = \begin{pmatrix} I, 0 \\ G, I \end{pmatrix} \begin{pmatrix} A, 0 \\ 0, B \end{pmatrix} \begin{pmatrix} I, F \\ 0, I \end{pmatrix}, \quad \mathcal{D}(\mathcal{A}) = \left\{ \begin{pmatrix} x \\ y \end{pmatrix}; \begin{matrix} x + Fy \in \mathcal{D}(A) \\ y \in \mathcal{D}(B) \end{matrix} \right\},$$

then

$$\mathcal{A}^* = \begin{pmatrix} I, 0 \\ F^*, I \end{pmatrix} \begin{pmatrix} A^*, 0 \\ 0, B^* \end{pmatrix} \begin{pmatrix} I, G^* \\ 0, I \end{pmatrix}, \quad \mathcal{D}(\mathcal{A}^*) = \left\{ \begin{pmatrix} x^* \\ y^* \end{pmatrix}; \begin{matrix} x^* + G^*y^* \in \mathcal{D}(A^*) \\ y^* \in \mathcal{D}(B^*) \end{matrix} \right\}.$$

Since the generator corresponding to the general delay equation (6) is of the above form, see Proposition 6, we get

PROPOSITION 8. *The conjugate operator to \mathcal{A}*

$$\mathcal{A} \begin{pmatrix} x \\ x(0) \end{pmatrix} = \begin{pmatrix} \frac{dx}{ds} \\ \int_{-h}^0 N(ds)x(s) \end{pmatrix}, \quad \mathcal{D}(\mathcal{A}) = \left\{ \begin{pmatrix} x \\ x(0) \end{pmatrix}, x \in W^{1,2} \right\},$$

is given by the formula

$$\mathcal{A}^* \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -\frac{d}{dx}(x - g^*y) \\ x(0) \end{pmatrix}, \quad \mathcal{D}(\mathcal{A}^*) = \left\{ \begin{pmatrix} x \\ y \end{pmatrix}; \begin{matrix} x - g^*y \in W^{1,2} \\ x(-h) = N_0^*y \end{matrix} \right\}.$$

Remark 6. The above proposition generalizes Theorem 5.2 of [10]. In [10] the form of \mathcal{A}^* was found under the condition that the measure N is a sum of a finite number of atoms and of a measure with bounded density.

Remark 7. It is obvious that Proposition 8 holds if the space $X = L^2[-h, 0; R^n]$ is replaced by $X = L^p[-h, 0; R^n]$ and $W^{1,2}$ by $W^{1,p}$, $p \geq 1$. The same Proposition 1 gives form of the conjugate operators for all unbounded operators which were decomposed in § 3.

4.3. Determination of spectrum. If the decomposition of an operator \mathcal{A} in the form $\mathcal{A} = \mathcal{A}_0\mathcal{P} + \mathcal{R}$ is known then it is possible to obtain some information about the spectrum of \mathcal{A} provided that rather complete knowledge about \mathcal{A}_0 is available. For instance the following proposition easily follows.

PROPOSITION 9. *If $\lambda \in \rho(\mathcal{A}_0)$ and \mathcal{P} is an invertible operator, then $\lambda \in \rho(\mathcal{A}_0\mathcal{P} + \mathcal{R})$ if and only if the operator*

$$\mathcal{P} + (\lambda I - \mathcal{P}\mathcal{R}\mathcal{P}^{-1})R_\lambda$$

is invertible, where R_λ denotes the resolvent of \mathcal{A}_0 .

In the case of operators on the product space

$$\mathcal{E} = \begin{matrix} X \\ \times \\ Y \end{matrix},$$

more explicit conditions can be given. The proposition below covers generator introduced in Proposition 4 and Proposition 5. In its formulation

$$(13) \quad \mathcal{A}_0 = \begin{pmatrix} A, & 0 \\ 0, & B \end{pmatrix}, \quad \mathcal{P} = \begin{pmatrix} I, & -F \\ 0, & I \end{pmatrix}, \quad \mathcal{R} = \begin{pmatrix} 0, & 0 \\ C, & 0 \end{pmatrix}.$$

Details of the proof can be found in [12].

PROPOSITION 10. *If operators \mathcal{A}_0 , \mathcal{P} , \mathcal{R} are given by (13) and $\lambda \in \rho(A)$ then $\lambda \in \rho(\mathcal{A}_0\mathcal{P} + \mathcal{R})$ if and only if the transformation*

$$\lambda I - (B + CF - \lambda CR_\lambda)$$

is invertible, where R_λ denotes the resolvent of A .

Proposition 10 when applied to delay systems (see Proposition 5) gives a well known result: $\lambda \in \rho(\mathcal{A})$ iff the matrix $\lambda I - N_0 - \int_{-h}^0 e^{\lambda s} N_1(s) ds$ is invertible. In [12] applications to parabolic systems are discussed.

4.4. Stochastic systems. Generators introduced in the second part of Theorem 1 provide a good description (see § 4.1) of systems whose “boundary conditions” evolve according to evolution equations. The same is true if “boundary conditions” have stochastic nature and are described by Ito’s stochastic equation. The simplest example of such system is as follows:

$$(14) \quad \begin{aligned} \frac{\partial x}{\partial t} &= \frac{\partial^2 x}{\partial s^2}, & t > 0, s \in (0, 1), \\ dx(t, 0) &= B_{11}x(t, 0) dt + B_{12}x(t, 1) dt + db^1(t), \\ dx(t, 1) &= B_{21}x(t, 0) dt + B_{22}x(t, 1) dt + db^2(t), \\ x(0, \cdot) &= x_0(\cdot) \in L^2[0, 1], \end{aligned}$$

where b^1, b^2 are Wiener processes. The solution of (14) is exactly the first coordinate of the mild solution of the equation

$$dz = \mathcal{A}z \, dt + dB_t,$$

where \mathcal{A} is the generator introduced in Proposition 4.

Now it is clear how the results of Theorem 1 can be applied to filtering and stochastic and control problems. As we mentioned before, a generalization of Proposition 2 enabled A. Chojnowska-Michalik [2] to prove that stochastic delay equations of the form

$$dy(t) = \left[\int_{-h}^0 N(ds)y(t+s) \right] dt + Du(t) \, dt + dM(y),$$

with the state dependent martingale noise $dM(y)$, can be also treated as the second coordinate of the equation

$$dz = \mathcal{A}z \, dt + \mathcal{D}u(t) \, dt + dM(z),$$

where \mathcal{A} is the generator corresponding to general delay equations.

Acknowledgments. I would like to thank Dr. A. J. Pritchard and Dr. A. Ichikawa for some helpful discussions.

REFERENCES

- [1] J. G. BORISOVIC AND A. S. TURBABIN, *On the Cauchy problem for linear non-homogeneous differential equations with retarded argument*, Soviet Math. Dokl., 10 (1969), pp. 401–405.
- [2] A. CHOJNOWSKA-MICHALIK, *Representation theorem for general stochastic delay equations*, submitted for publication, Bull. Acad. Polon. Sci. Sér. Math. Astronom. Phys.
- [3] R. F. CURTAIN AND A. J. PRITCHARD, *The infinite dimensional Riccati equation*, J. Math. Anal. Appl., 47 (1974), pp. 43–57.
- [4] M. C. DELFOUR AND S. K. MITTER, *Controllability, observability and optimal feedback control of hereditary differential systems*, this Journal, 10 (1972), pp. 298–328.
- [5] M. GIURGIU, *A feedback solution of a linear quadratic problem for boundary control of heat equation*, Rev. Roumaine Math. Pures Appl., XX (1975), pp. 927–954.
- [6] E. E. INFANTE AND J. A. WALKER, *On the stability properties of an equation arising in reactor dynamics*, J. Math. Anal. Appl., 55 (1976), No. 1, pp. 112–125.
- [7] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Heidelberg, New York, 1966.
- [8] Y. KURODA AND A. MAKINO, *Some Problems Arising in Distributed Parameter Reactor Systems*, Colloques IRIA, Analyse de Systèmes et ses Orientations Nouvelles, Versailles, 13–17 December, 1976, to appear.
- [9] G. F. WEBB, *Functional differential equations and nonlinear semigroups in B-spaces*, J. Differential Equations, 20 (1976), pp. 71–89.
- [10] R. B. VINTER, *On the evolution of the state of linear differential delay equation in M^2 : properties of the generator*, technical report, Imperial College, London.
- [11] K. YOSHIDA, *Functional Analysis*, Springer-Verlag, Heidelberg, New York, 1974.
- [12] J. ZABCZYK, *On semigroups corresponding to non-local boundary conditions with applications to system theory*, report 49, Control Theory Centre, University of Warwick, Coventry, England, 1976.
- [13] ———, *A Semigroup Approach to Boundary Value Control*, Proceedings of the IFAC 2nd Symposium on Control of Distributed Parameter Systems, 28 June–1 July 1977, University of Warwick, Coventry, England, to appear.

UNBOUNDED CONTROL AND OBSERVATION SYSTEMS AND THEIR DUALITY*

A. J. PRITCHARD† AND A. WIRTH†

Abstract. The paper considers some questions arising from a general theory of observation and control by Russell and Dolecki.

For a distributed parameter system the observation may be restricted to the boundary or some other subset, so the observation operator may be unbounded. If this operator satisfies certain conditions the dual control system operator can be defined. If also there is an appropriate Green's formula, the dual system is interpreted as one of control on a subset. For systems associated with biorthogonal sequences it is shown that the attainable sets cannot be fully characterized in terms of the sequence unless it is basic.

1. Introduction. For distributed parameter systems governed by partial differential equations there are severe practical restrictions on the positioning of sensors and actuators. Usually the observation and control processes are restricted to subsets, boundaries or even points of the region over which the distributed system is defined. This often means that the observation and control operators are unbounded on the state and control spaces. Dolecki and Russell [3] have developed a general theory for such problems, and in this report we examine a number of questions which arise from their work. In particular we assume that the system is determined by a semigroup and derive a set of conditions on the observation operator in terms of an intermediate space so that the dual control system operator can be defined. Then by means of a Green's formula we are able to associate the control problem with a controlled partial differential equation. In §§ 2, 3 and 4 we describe these duality results and also prove a generalization of a theorem in [3] which gives conditions under which a semigroup may be extended to a group. Dolecki and Russell also introduce the concept of a reconstruction operator and derive sufficient conditions for the existence and uniqueness of an optimal operator. Their conditions are not sufficient, but by adding an additional one we find a sufficient set. Finally in § 5, we examine biorthogonal systems and boundary control for a special class of problems considered in [4] and [7]. We show that not every control which steers the system from a given initial state to a desired final state need be expressed in terms of the biorthogonal sequence; then we determine conditions involving basic and minimal sequences for which this is the case.

2. Duality for unbounded systems. We consider the control problem which we write formally as

$$(1) \quad \dot{x} = Ax + Bu, \quad x(0) = x_0$$

where A is the infinitesimal generator of a strongly continuous semigroup T_t on a Banach space X and B is a, possibly unbounded, operator from a control Banach space U to X . The interpretation given to (1) will in fact be the integral equation

$$(2) \quad x(t) = T_t x_0 + \int_0^t T_{t-s} B u(s) ds.$$

If $B \in \mathcal{L}(U, X)$ and $u \in L^q(0, T; U)$, $q > 1$, such a solution is known as a mild solution and we know $x \in C(0, T; X)$ although in general we are not able to differentiate (2) to obtain (1). We will generalize the concept of a mild solution to the case where B is an

* Received by the editors July 21, 1976, and in revised form May 24, 1977.

† Control Theory Centre, University of Warwick, Coventry CV4 7AL, Warwickshire, England.

unbounded operator. The problem of observation we write formally as

$$(3) \quad \dot{z} = \mathcal{A}z, \quad z(0) = z_0,$$

$$(4) \quad y = Cz$$

where \mathcal{A} is the infinitesimal generator of a strongly continuous semigroup S_t on a Banach space Z and C is a (possibly unbounded) operator with dense domain in Z and range a Banach space Y . Then

$$y(t) = CS_t z_0$$

where we need to interpret the operator CS_t where C is unbounded. In fact in order to exploit the duality between the controllability and observability problems we shall assume

$$X = Z^*, \quad Z \text{ reflexive}, \quad B = C^*, \quad T_t = S_t^* \quad \text{and} \quad U = Y^*.$$

For the observability problem we make the assumptions:

(a) CS_t can be extended for each $t > 0$ to a bounded linear operator $\overline{CS_t}$ in $\mathcal{L}(Z, Y)$

and

(b) $\|\overline{CS_t}\|_{\mathcal{L}(Z, Y)} \leq g(t)$ for some $g \in L^p(0, T)$, $p > 1$.

Now define $\mathcal{C}: Z \rightarrow L^p(0, T; Y)$ by $(\mathcal{C}z)(t) = \overline{CS_t}(z)$; then

$$\mathcal{C} \in \mathcal{L}(Z, L^p(0, T; Y)).$$

With the above assumptions it is possible to derive duality theorems as in [3]; see also the next section. However, to interpret the solution of the controllability problem not merely in terms of \mathcal{C}^* but in terms of A and B it is necessary to make further assumptions. We assume there exists a Banach space W such that $\overline{W} = Z$:

(α) $Z \supset D(C) \supset W$;

(β) $C \in \mathcal{L}(W, Y)$;

(γ) $S_t \in \mathcal{L}(Z, W)$ for all $t > 0$;

(δ) $\|S_t z\|_W \leq g(t)\|z\|_Z$ for all $z \in Z$ where $g \in L^p(0, T)$.

Clearly $\|CS_t z\|_Y \leq \|C\|_{\mathcal{L}(W, Y)}\|S_t z\|_W \leq g(t)\|C\|\|z\|$, so conditions (α) and (β) are satisfied.

Remark 1. Property (γ) is similar to the assumption that S_t is an analytic semigroup for which $S_t Z \rightarrow D(\mathcal{A})$, $t > 0$. However we are not able to set $W = D(\mathcal{A})$ with the graph norm since

$$\|\mathcal{A}S_t\| \leq M/t \quad \text{for some } M > 0 \quad \text{if } S_t \text{ is analytic}$$

and (δ) may not hold.

Remark 2. If $f \in L^q(0, T; Z)$ with $1/p + 1/q = 1$, then $S_{t-s}f(s) \in W$ for almost all $s < t$, and is Bochner integrable with respect to W , so since W is a Banach space, we have

$$C \int_0^t S_{t-s}f(s) ds = \int_0^t CS_{t-s}f(s) ds.$$

The conditions (α)–(γ) may seem strange but they become clearer when we unravel the dual problem. In general unless we impose smoothness constraints on Z trace operators onto boundaries or manifolds interior to the region are densely defined on Z but are not closable. Hence if we do not assume (γ), i.e. that S_t smooths the space, then we are only able to conclude that

$$(CS_t)^* \supset S_t^* C^* = T_t B.$$

Moreover in this case B may have trivial domain so that it is not clear how to define the control system $x(t) = T_t x_0 + \int_0^t T_{t-s} B u(s) ds$, with $u \in L^q(0, T; U)$.

If (α) – (δ) hold we have

$$\begin{aligned} B &\in \mathcal{L}(U, W^*), \\ T_t &\in \mathcal{L}(W^*, X) \quad \text{for all } t > 0, \\ \|T_t B u\|_X &\leq g(t) \|u\|_U. \end{aligned}$$

So $(CS_t)^* = T_t^* B^*$ and (2) is well defined with $x \in C(0, T; X)$.

Let us illustrate assumptions (α) – (δ) and the duality with the following:

Example 1. Consider the heat conduction Dirichlet problem

$$\dot{\theta} = \mathcal{A}\theta, \quad \theta(0) = \theta_0$$

where

$$\mathcal{A}\theta = \Delta\theta, \quad \theta \in D(\mathcal{A}),$$

$$D(\mathcal{A}) = H^2(\Omega) \cap H_0^1(\Omega) \quad \text{and} \quad \Omega = (0, 1) \times (0, 1).$$

Then \mathcal{A} generates an analytic semigroup on $L^2(\Omega) = Z$. So putting $W = H^2(\Omega)$ we have $S_t \in \mathcal{L}(L^2(\Omega), H^2(\Omega))$ for all $t > 0$.

Suppose that we can observe $\text{grad } \theta$; then we are able to associate an unbounded closed linear C with this observation $C: D(C) \rightarrow L^2(\Omega) \times L^2(\Omega) = Y$ with $H^2(\Omega) \subset D(C) \subset Z$. Also $C \in \mathcal{L}(H^2(\Omega), L^2(\Omega) \times L^2(\Omega))$. Moreover

$$\|CS_t\|_{\mathcal{L}(Z, Y)} \leq \frac{M}{t^{1/2}} \quad \text{for some } M > 0,$$

since $\|CS_t z\|^2 = \sum \sum \alpha_{mn}^2 (m^2 + n^2) \pi^2 e^{-2(m^2 + n^2) \pi^2 t} \leq (M^2/t) \|z\|^2$, where $\|z\|^2 = \sum \sum \alpha_{mn}^2$.

Thus $g \in L^p(0, T)$ for $p < 2$ and the dual problem will be well-defined if the control $u \in L^q(0, T; L^2(\Omega) \times L^2(\Omega))$ with $q > 2$.

Now suppose instead that we can observe θ on $x = x_1$, $0 < x_1 < 1$, so that $C\theta(\cdot, \cdot, t) = \theta(x_1, \cdot, t)$. Let $Y = L^2(0, 1)$; then $C \in \mathcal{L}(H^2(\Omega), Y)$ and $\|CS_t\| \leq M/t^{1/4}$ for some $M > 0$, since

$$\|CS_t z\|^2 \leq \sum_n e^{-2n^2 \pi^2 t} \left(\sum_m \alpha_{mn}^2 \right) \left(\sum_m e^{-2m^2 \pi^2 t} \right) \leq \frac{M^2}{t^{1/2}} \|z\|^2$$

by Cauchy–Schwarz. Thus $g \in L^p(0, T)$ for $p < 4$. If we consider C as a map from $\text{dom } C \subset L^2(\Omega)$ into $L^2(0, 1)$ then C^* is $\delta(x - x_1)$ the Dirac delta function. So we need to interpret the control problem

$$x(t) = T_t x_0 + \int_0^t T_{t-s} C^* u(s) ds$$

for $u \in L^q(0, T; L^2(0, 1))$ with $q > \frac{4}{3}$. It will be shown in Example 3 that we may interpret (2) as the weak solution of the control problem

$$\dot{\theta} = \mathcal{A}\theta \quad \text{with} \quad - \left[\frac{\partial \theta}{\partial x} \right]_{x_1-}^{x_1+} = u(y, t).$$

3. Controllability and observability. We assume as in the previous section that

$$Z \text{ is reflexive, } X = Z^*, \quad U = Y^* \quad \text{and} \quad T_t = S_t^*$$

and (α) – (δ) hold.

We follow [3] in making the following definitions.

DEFINITION 1. (\mathcal{A}, C) is *initially observable on* $(0, T)$ if $\ker \mathcal{C} = \{0\}$.

DEFINITION 2. (\mathcal{A}, C) is *continuously initially observable on* $(0, T)$ if there exists $\kappa > 0$ such that

$$\|z\|_Z \leq \kappa \|\mathcal{C}z\|_{L^p(0,T;Y)} \text{ for all } z \in Z.$$

DEFINITION 3. (\mathcal{A}, C) is *finally observable on* $(0, T)$ if

$$\ker \mathcal{C} \subset \ker S_T.$$

DEFINITION 4. (\mathcal{A}, C) is *continuously finally observable on* $(0, T)$ if there exists $\kappa > 0$ such that

$$\|S_T z\|_Z \leq \kappa \|\mathcal{C}z\|_{L^p(0,T;Y)}$$

DEFINITION 5. (A, B) is *approximately controllable on* $(0, T)$ if

$$\text{cl} \left\{ \int_0^T T_{T-s} B u(s) ds : u \in L^q(0, T; U) \right\} = X.$$

DEFINITION 6. (A, B) is *exactly controllable on* $(0, T)$ if

$$\left\{ \int_0^T T_{T-s} B u(s) ds : u \in L^q(0, T; U) \right\} = X.$$

DEFINITION 7. (A, B) is *approximately null controllable on* $(0, T)$ if

$$\text{cl} \left\{ \int_0^T T_{T-s} B u(s) ds : u \in L^q(0, T; U) \right\} \supset \text{range } T_T.$$

DEFINITION 8. (A, B) is *exactly null controllable on* $(0, T)$ if

$$\left\{ \int_0^T T_{T-s} B u(s) ds : u \in L^q(0, T; U) \right\} \supset \text{range } T_T.$$

THEOREM 1 [3]. *With the assumptions stated at the beginning of this section:*

- (i) (\mathcal{A}, C) is *initially observable if and only if* (A, B) is *approximately controllable*,
- (ii) (\mathcal{A}, C) is *continuously initially observable if and only if* (A, B) is *exactly controllable*,
- (iii) (\mathcal{A}, C) is *finally observable if and only if* (A, B) is *approximately null controllable*,
- (iv) (\mathcal{A}, C) is *continuously finally observable if and only if* (A, B) is *exactly null controllable*,

all statements referring to the interval $(0, T)$.

In general, initial continuous observability is not possible if S_t is only a semigroup, as is indicated by the following generalization of Theorem 1.3 of [3] to unbounded observation operators.

THEOREM 2. *If* $\text{cl range } S_t = Z$ *for some* $t > 0$ *and* (\mathcal{A}, C) *is continuously initially observable on* $(0, T)$ *then* $\{S_t : t \geq 0\}$ *has an extension to a strongly continuous group of bounded operators* $\{S_t : -\infty < t < \infty\}$.

Proof. Since (\mathcal{A}, C) is continuously initially observable there exists $\kappa > 0$ such that $\|z\|^p \leq \kappa \int_0^T \|CS_t z\|^p dt$ for all $z \in Z$. Suppose no S_t , $t > 0$, is bounded below; that is

for each $t > 0$ and each $\varepsilon > 0$, there exists z_t such that $\|z_t\| = 1$ and $\|S_t z_t\| < \varepsilon$. Now if $\|z\| = 1$

$$\int_0^T \|CS_t z\|^p dt \leq \int_0^s \|CS_t\|^p dt + \int_s^T \|CS_{t-s}\|^p \|S_s z\|^p dt.$$

Let $s \rightarrow 0$ and choose $\|z_s\| = 1$ so that $\|S_s z_s\| \rightarrow 0$. Then $\int_0^T \|CS_t z_s\|^p dt \rightarrow 0$ as $s \rightarrow 0$, but $\int_0^T \|CS_t z_s\|^p dt \geq 1/\kappa$ for all $s > 0$. Hence there exists $t_0 > 0$, $M > 0$ such that $\|S_{t_0} z\| \geq M\|z\|$ for all $z \in Z$. Also there exists N such that $\|S_t\| \leq N$ for all $t \in [0, T]$. If $t = nt_0 + t_1$ where $0 \leq t_1 < t_0$ then $\|S_t z\| \geq M^n \|S_{t_1} z\|$ and $M\|z\| \leq \|S_{t_0} z\| \leq \|S_{(t_0-t_1)}\| \|S_{t_1} z\| \leq N \|S_{t_1} z\|$. Hence $\|S_t z\| \geq M^{n+1} N^{-1} \|z\|$, so every S_t is bounded below.

By hypothesis S_α has dense range for some $\alpha > 0$, and hence S_α is invertible. If $t = n\alpha + t_1$, $0 \leq t_1 < \alpha$ then $S_t S_{(\alpha-t_1)} (S_\alpha)^{(-n-1)} = I$. So range $S_t = Z$ for each $t > 0$ and hence each S_t , $t > 0$ is invertible. The rest is as for [3].

Example 2. Suppose that Z is a Hilbert space with orthonormal basis

$$\{\phi_{ni} : n = 1, 2, \dots; i = 1, \dots, m_n\},$$

such that $S_t z = \sum_{n=1}^{\infty} e^{-\lambda_n t} \sum_{i=1}^{m_n} \alpha_{ni} \phi_{ni}$ where $z = \sum \alpha_{ni} \phi_{ni}$, and $\lambda_n \uparrow \infty$. Suppose also that $\limsup ((\log n)/\lambda_n)$ is finite and that $m_n \leq \exp(\varepsilon \lambda_n)$ for some $\varepsilon > 0$. For example, the solution of the diffusion equation in a region $\Omega \subset R^r$ satisfies the above conditions (see [6, p. 38] and [7]); in fact, then $(\log n)/\lambda_n \rightarrow 0$.

Clearly $\text{cl range } S_t = Z$ for all $t > 0$. But the system is not continuously initially observable for any C for which $\mathcal{C} \in \mathcal{L}(Z, L^p(0, T; Y))$. If it were, then by Theorem 2 we would have range $S_t = Z$ for each $t > 0$. Now choose $t_0 > 0$ so that $t_0 - \varepsilon > \limsup ((\log n)/\lambda_n)$; then the series $\sum e^{-\lambda_n(t_0-\varepsilon)}$ is convergent [2]. Let $z = \sum \sum e^{-\lambda_n t_0/2} \phi_{ni} \in Z$. So there exists $z_1 = \sum \alpha_{ni} \phi_{ni}$ such that $S_{t_0} z_1 = z$. So $e^{-\lambda_n t_0} \alpha_{ni} = e^{-\lambda_n t_0/2}$ for each n and i . But $\{\alpha_{ni}\}$ is bounded, a contradiction.

We now assume, more generally, that we have a continuous linear map $\mathcal{C} : Z \rightarrow \tilde{Y}$ where \tilde{Y} is a Banach space, for example $L^p(0, T; Y)$.

The notions of initial and final observability can be both included in that of F -observability [3], where $F : Z \rightarrow V$ is a bounded linear map into a Banach space V . For initial observability $F = I$ and for final observability $F = S_T$, with $V = Z$. We say (\mathcal{A}, C) is F -observable if $\|Fz\| \leq M\|\mathcal{C}z\|$ for some $M > 0$. So clearly if we have F -observability there exists $\mathcal{G}_0 : \text{range } \mathcal{C} \rightarrow V$ such that \mathcal{G}_0 is linear, continuous and $\mathcal{G}_0 \mathcal{C} = F$.

Dolecki and Russell [3] ask whether there exists a ‘reconstruction’ operator $\mathcal{G} : \tilde{Y} \rightarrow V$ such that \mathcal{G} extends \mathcal{G}_0 , i.e. $\mathcal{G}\mathcal{C} = F$. If $\|\mathcal{G}\|$ is minimal we call \mathcal{G} an *optimal reconstruction operator*. Such an operator minimizes the reconstruction error $\|Fz - \mathcal{G}(\mathcal{C}z + w)\| \leq \|\mathcal{G}\| \|w\|$, where w is the error in the observation. Russell and Dolecki [3] state a result about sufficient conditions for the existence and uniqueness of an optimal reconstruction operator. Their conditions are, however, insufficient. The following conditions ensure *uniqueness*: $\|\mathcal{G}_0 w\| = \gamma \|w\|$ for some $\gamma > 0$, range $\mathcal{G}_0 = V$, \tilde{Y} is smooth and reflexive. \tilde{Y} is said to be *smooth* if at each point of the unit sphere there is only one supporting hyperplane of the unit ball. Equivalently (Day [1, p. 144]) the norm is Gateaux differentiable at each point of the unit sphere, i.e. $\lim_{\alpha \rightarrow 0} (1/\alpha) \{\|x + \alpha y\| - \|x\|\}$ exists for each $\|x\| = 1$, y . The result about uniqueness of the operator follows from Lemma 5.2 of [3] and the following lemma.

LEMMA 1. *Let \tilde{Y} be a smooth reflexive normed vector space, M a closed subspace. If $f \in M^*$ then f has a unique extension to an element of \tilde{Y}^* with same norm.*

Proof. Without loss of generality $\|f\| = 1$. Since \tilde{Y} is reflexive, so is M [6]. So there exists $x_0 \in M$ such that $f(x_0) = 1$ and $\|x_0\| = 1$. If $y \notin M$ and if F is an extension of f with $\|F\| = \|f\|$ and $F(y) = \alpha$ then by the standard Hahn–Banach argument (Day [1, p. 10])

$$-\|tx_0 + y\| - \|tx_0\| \leq \alpha \leq \|tx_0 + y\| - \|tx_0\| \quad \text{for all } t > 0$$

and

$$-\|tx_0 - y\| + \|tx_0\| \leq \alpha \leq \|tx_0 - y\| + \|tx_0\| \quad \text{also for all } t > 0.$$

But by the existence of the Gateaux derivative,

$$\lim_{t \rightarrow \infty} (\|tx_0 + y\| - \|tx_0\|) = \lim_{t \rightarrow \infty} (-\|tx_0 - y\| + \|tx_0\|)$$

so α is uniquely defined. Hence the extension is unique.

In particular if $\tilde{Y} = L^p(0, T; Y)$ with $1 < p < \infty$ and Y a reflexive Banach space then \tilde{Y} is smooth.

4. Green's formula and weak solutions. In the previous sections we have interpreted (2) in the sense that $B \in \mathcal{L}(U, W^*)$, and we now show how (2) can be related to a specific partial differential equation. First we prove the following lemma.

LEMMA 2. *If $\psi \in C(0, T; X)$ then $x \in C(0, T; X)$ satisfies*

$$\int_0^T \langle \psi(t), x(t) \rangle_X dt + \int_0^T \langle C\phi(t), u(t) \rangle_{Y^*} + \langle \phi(0), x_0 \rangle = 0$$

where $\phi(t) = -\int_t^T S_{s-t}\psi(s) ds$ if and only if x satisfies (2).

Proof. Substitute (2) into the left side of the above equation to obtain

$$\begin{aligned} \int_0^T \left\langle \psi(t), T_t x_0 + \int_0^t T_{t-s} B u(s) ds \right\rangle dt + \int_0^T \left\langle -C \int_t^T S_{s-t} \psi(s) ds, u(t) \right\rangle dt \\ - \left\langle \int_0^T S_s \psi(s) ds, x_0 \right\rangle = 0, \end{aligned}$$

since by assumption $C \int_t^T S_{s-t} \psi(s) ds = \int_t^T C S_{s-t} \psi(s) ds$, (see Remark 2) and $(CS_t)^* = T_t B$ by $(\alpha) - (\delta)$. Conversely substitution for ϕ in the above equation yields

$$\int_0^T \left\langle \psi(t), x(t) - T_t x_0 - \int_0^t T_{t-s} B u(s) ds \right\rangle_X dt = 0$$

for all $\psi \in C(0, T; X)$.

Remark 3. If $\psi \in C^1(0, T; X)$ then $\dot{\phi} + \mathcal{A}\phi = \psi$, $\phi(T) = 0$, so x defined by (2) satisfies

$$\int_0^T \langle \dot{\phi}(t) + \mathcal{A}\phi(t), x(t) \rangle_X dt + \int_0^T \langle C\phi(t), u(t) \rangle_{Y^*} dt + \langle \phi(0), x_0 \rangle = 0.$$

This suggests that x is a weak solution of a controlled partial differential equation. To see how we can determine the equation we assume Y is a Banach space of functions on a subset Γ_1 of $\bar{\Omega}$ where the operator \mathcal{A} is defined on an open bounded set Ω . Z and X are Hilbert spaces with X and Z identified, and the differential operator \mathcal{A} is defined on a space of functions on Ω with appropriate conditions on Γ , the boundary of Ω . Let $\tilde{\mathcal{A}}$ be formally the same operator but now defined on the same space of functions restricted to $\Omega \setminus \Gamma$, with the same boundary conditions on $\Gamma \setminus \Gamma_1$. Let \tilde{A} be the

adjoint of $\tilde{\mathcal{A}}$ and assume we have the following Green's formula:

$$\langle x_1, \tilde{A}x_2 \rangle_X = \langle \mathcal{A}x_1, x_2 \rangle_X + \langle Cx_1, Dx_2 \rangle_{Y^*} + \langle Ex_1, Fx_2 \rangle_{Y^*}$$

for some linear operators D, E, F .

We now show that (2) is a weak solution of

$$(5) \quad \dot{x} = \tilde{A}x, \quad Dx = u, \quad Fx = 0, \quad x(0) = x_0.$$

DEFINITION 9. A *weak solution* of (5) is a solution of

$$(6) \quad \int_0^T \langle \psi(t), x(t) \rangle_X dt + \int_0^T \langle C\phi(t), u(t) \rangle_{Y^*} dt + \langle \phi(0), x_0 \rangle = 0$$

where $\dot{\phi} + \mathcal{A}\phi = \psi$, $\phi(T) = 0$, $\psi \in C^1(0, T; X)$.

To see that (6) is a reasonable definition let $x_1 = \phi$ and $x_2 = x$ in the Green's formula:

$$\begin{aligned} 0 &= \int_0^T \langle \phi(t), \dot{x} - \tilde{A}x \rangle_X dt \\ &= \langle \phi(T), x(T) \rangle_X - \langle \phi(0), x_0 \rangle_X \\ &\quad - \int_0^T \langle \dot{\phi} + \mathcal{A}\phi, x \rangle_X dt - \int_0^T \langle C\phi(t), u(t) \rangle_{Y^*} dt \\ &\quad - \int_0^T \langle E\phi(t), Fx \rangle_{Y^*} dt. \end{aligned}$$

But $\phi(T) = 0$ and $\dot{\phi} + \mathcal{A}\phi = \psi$ so we obtain (6). Lemma 2 and Remark 2 yield the following

THEOREM 3. With the above assumptions $x(t) = T_t x_0 + \int_0^t T_{t-s} B u(s) ds$ is a weak solution of $\dot{x} = \tilde{A}x$, $Dx = u$, $Fx = 0$, $x(0) = x_0$.

Example 3. Let $\Omega = (0, 1)$, $X = L^2(\Omega)$, $D(\mathcal{A}) = H^2(\Omega) \cap H_0^1(\Omega)$, $\mathcal{A}\theta_1 = \Delta\theta_1$, $\dot{\theta}_1 = \mathcal{A}\theta_1$ and $C\theta_1 = \theta_1(\alpha)$ where $0 < \alpha < 1$. Since S_t is analytic and $H^2(\Omega) \subset C^1(\bar{\Omega})$, C is well-defined and in fact $C \in \mathcal{L}(H^2(\Omega), R)$. Now $D(\tilde{A}) = H^2(\Omega \setminus \{\alpha\}) \cap H_0^1(\Omega \setminus \{\alpha\})$, $\tilde{A}\theta_2 = \Delta\theta_2$, $\dot{\theta}_2 = \tilde{A}\theta_2$ and the Green's formula is

$$\int_0^1 \theta_1 \frac{\partial^2 \theta_2}{\partial x^2} dx = \int_0^1 \frac{\partial^2 \theta_1}{\partial x^2} \theta_2 dx + C\theta_1 \left[\frac{\partial \theta_2}{\partial x} \right]_{\alpha^-}^{\alpha^-} + \frac{\partial \theta_1}{\partial x}(\alpha) [\theta_2]_{\alpha^-}^{\alpha+}.$$

So

$$D\theta_2 = - \left[\frac{\partial \theta_2}{\partial x} \right]_{\alpha^-}^{\alpha+} \quad \text{and} \quad F\theta_2 = [\theta_2]_{\alpha^-}^{\alpha+}.$$

Hence the dual system is the diffusion equation on $(0, 1)$ with $\theta_2(0) = \theta_2(1) = 0$, θ_2 continuous at α but a discontinuity of $\partial\theta_2/\partial x$ at α such that $-[\partial\theta_2/\partial x]_{\alpha^-}^{\alpha+} = u$.

Note that $H^2(\Omega \setminus \{\alpha\}) \not\subset C(\bar{\Omega})$ since the open set $\Omega \setminus \{\alpha\}$ is not locally on one side of its boundary.

Now let

$$D(\mathcal{A}) = H^2(\Omega) \cap \left\{ \theta : \frac{\partial \theta}{\partial x} \Big|_{\Gamma} = 0 \right\}$$

and let $C\theta = \theta(1)$. Then

$$D(\tilde{A}) = H^2(\Omega) \cap \left\{ \theta : \frac{\partial \theta}{\partial x}(0) = 0 \right\}.$$

The condition

$$\frac{\partial \theta}{\partial x}(1) = 0$$

is not necessarily satisfied by θ_2 since now $\Gamma_1 = \{1\}$. Then

$$D\theta_2 = \frac{\partial \theta_2}{\partial x}(1) \quad \text{and} \quad F = 0.$$

5. Biorthogonal systems and boundary control. The following problem of boundary control has been studied extensively by Russell and others [4], [7]. For the sake of completeness we briefly summarize it.

Let Ω be a bounded open connected subset of R^r with piecewise smooth boundary Γ . Let $\Gamma = \Gamma_0 \cup \Gamma_1$ such that $\Gamma_0 \cap \Gamma_1 = \emptyset$, Γ_0 is relatively open in Γ and the relative interior of Γ_1 is nonempty, also let (Ω, Γ_1) be star-complemented.

Consider

$$\begin{aligned} \frac{\partial w}{\partial t} &= \sum_{i=1}^r \frac{\partial^2 w}{\partial x_i^2} \quad \text{for all } x \in \Omega, \quad t \geq 0, \\ w(x, t) &= 0 \quad \text{for all } x \in \Gamma_0, \quad t \geq 0, \\ \frac{\partial w}{\partial \nu} &= g \quad \text{for all } x \in \Gamma_1, \quad t \geq 0, \end{aligned}$$

where ν is the unit outward normal, $g \in L^2((0, T) \oplus \Gamma_1)$, $w \in H^2(\Omega)$. Let $\{-\lambda_n\}$, $\{\phi_{ni}\}$ be the eigenvalues with multiplicities m_n and eigenfunctions of Δ with domain

$$\left\{ z \in H^2(\Omega): z|_{\Gamma_0} = 0, \frac{\partial z}{\partial \nu} \Big|_{\Gamma_1} = 0 \right\}.$$

If

$$w(\cdot, 0) = \sum_{n=1}^{\infty} \sum_{i=1}^{m_n} \alpha_{ni} \phi_{ni} \in L^2(\Omega) \quad \text{and} \quad w(\cdot, T) = \sum \sum \beta_{ni} \phi_{ni} \in L^2(\Omega),$$

then by an application of the divergence theorem it follows that

$$(7) \quad \beta_{ni} - e^{-\lambda_n T} \alpha_{ni} = \langle p_{ni}, \tilde{g} \rangle_{L^2((0, T) \oplus \Gamma_1)}$$

where $p_{ni}(t) = e^{-\lambda_n t} \phi_{ni}|_{\Gamma_1}$ and $p_{ni} \in L^2((0, T) \oplus \Gamma_1)$ and $\tilde{g}(x, t) = g(x, T - t)$. It can be shown that there exist $q_{mj} \in L^2((0, T) \oplus \Gamma_1)$ such that $\langle p_{ni}, q_{mj} \rangle = \delta_{nm} \delta_{ij}$, i.e. $\{(p_{ni}, q_{ni})\}$ is a *biorthogonal system* in $L^2((0, T) \oplus \Gamma_1)$.

So the control steering the state $\{\alpha_{ni}\}$ to $\{\beta_{ni}\}$ is

$$(8) \quad g(x, T - t) = \sum \sum (\beta_{ni} - e^{-\lambda_n T} \alpha_{ni}) q_{ni}(x, t)$$

provided $g \in L^2((0, T) \oplus \Gamma_1)$. Russell shows that $\|q_{ni}\| \leq M_0 e^{M_1 \sqrt{\lambda_n}}$ for some constants M_0, M_1 thus providing sufficient conditions for the convergence of (8). He also shows by a simple example for $r = 1$ that the conditions he obtains are not necessary.

We show that either Russell's bounds can be improved or there exist controls not expressible in the form (8).

The problem of solving (7) may be stated abstractly thus:

Let X be a Hilbert space, let $\{p_n\} \subset X$ and let $\{a_n\} \subset R$. Does there exist $y \in X$ such that

$$(9) \quad \langle y, p_n \rangle = a_n \quad \text{for all } n?$$

We say that the sequence $\{p_n\}$ is *minimal* if $d_n > 0$ for each n , where $d_n = \inf \|p_n - X_n\|$ and X_n is the closed span in X of the set $\{p_i: i \neq n\}$. We say that (p_n, q_n) is a *biorthogonal system* if $\langle p_m, q_n \rangle = \delta_{mn}$. It is well-known, (Singer [10, p. 54], that $\{p_n\}$ is minimal if and only if there exists a sequence $\{q_n\}$ such that (p_n, q_n) is biorthogonal. The sequence $\{p_n\}$ is said to be *basic* if $\{p_n\}$ is a Schauder basis in $\text{cl span } \{p_n\}$, i.e. if for every $x \in \text{cl span } \{p_n\}$ there exists a unique sequence $\{\alpha_n\}$ such that $x = \sum_{n=1}^{\infty} \alpha_n p_n$. Every basic sequence is minimal (Singer [10, p. 51]).

LEMMA 3. *If $\{p_n\}$ is minimal then there exists $\{q_n\}$ such that (p_n, q_n) is a biorthogonal system with minimum norm, $\|q_n\| = 1/d_n$ and $\text{cl span } \{p_n\} \supset \text{cl span } \{q_n\}$. The sequence $\{p_n\}$ is basic if and only if $\{q_n\}$ is basic and $\text{cl span } \{p_n\} = \text{cl span } \{q_n\}$.*

Proof. Since $p_n \notin \text{cl span } \{p_i: i \neq n\} = X_n$, there exists a unique $r_n \in X_n$ such that $\|p_n - r_n\| = d_n$. Let $q_n = (p_n - r_n)d_n^{-2}$. Clearly $\langle p_m, q_n \rangle = \delta_{mn}$ and if (p_n, t_n) is also biorthogonal then $(t_n - q_n) \perp q_n$. So $\|t_n\|^2 = \|t_n - q_n\|^2 + \|q_n\|^2$. Clearly by construction $\text{cl span } \{p_n\} \supset \text{cl span } \{q_n\}$. By Singer [10, Thm. 12.1, p. 112] if $\{p_n\}$ is basic then so is $\{q_n\}$. Also then $x \in \text{cl span } \{p_n\}$ implies that $x = \sum \alpha_n p_n$ for some $\{\alpha_n\}$, so $\alpha_n = \langle x, q_n \rangle = 0$ for all n , if $x \perp q_n$. Hence $\text{cl span } \{p_n\} = \text{cl span } \{q_n\}$.

Conversely if $\text{cl span } \{p_n\} = \text{cl span } \{q_n\}$ and $\{q_n\}$ is basic then $\{p_n\}$ is basic by Singer [10, Cor. 12.1, p. 113].

THEOREM 4. *Suppose $\{p_n\}$ is minimal.*

If $\{p_n\}$ is basic then every solution of $\langle y, p_n \rangle = a_n$ is of the form

$$(10) \quad y = \sum \alpha_n q_n + z \quad \text{where } z \in \{\text{cl span } \{p_n\}\}^\perp,$$

and so a necessary condition for (9) to have a solution is that $\|\alpha_n q_n\| \rightarrow 0$.

If $\{p_n\}$ is not basic then there exists $y \in \text{cl span } \{p_n\}$ such that $\sum \langle y, p_n \rangle q_n$ is divergent, and hence not every solution of (9) is of the form (10).

Proof. If $\{p_n\}$ is basic, then by Lemma 3, if y is a solution of (9) $y = y_1 + z$ where $y_1 \in \text{cl span } \{q_n\}$, $z \in \{\text{cl span } \{p_n\}\}^\perp$. Since now $\{q_n\}$ is basic, $y_1 = \sum \alpha_n q_n$ for some $\{\alpha_n\}$. Clearly in fact $\alpha_n = a_n$.

If $\{p_n\}$ is not basic, then either $\{q_n\}$ is not basic or $\text{cl span } \{p_n\} \not\supset \text{cl span } \{q_n\}$. If $\{q_n\}$ is not basic there exists $y_0 \in \text{cl span } \{q_n\}$ such that for no sequence $\{\alpha_n\}$ is $y_0 = \sum \alpha_n q_n$. So clearly $\sum \langle y_0, p_n \rangle q_n$ is divergent, yet the system $\langle y, p_n \rangle = \langle y_0, p_n \rangle$ is soluble.

If there exists $y_1 \in \text{cl span } \{p_n\} \setminus \text{cl span } \{q_n\}$ then $\sum \langle y_1, p_n \rangle q_n$ is divergent for otherwise $y_2 = \sum \langle y_1, p_n \rangle q_n \in \text{cl span } \{q_n\}$ and $\langle y_1 - y_2, p_n \rangle = 0$ for all n . Hence $y_1 = y_2$, a contradiction.

THEOREM 5. *If $\{p_n\}$ is basic then there exists $M > 0$ such that*

$$\frac{1}{\|p_n\|} \leq \|q_n\| \leq \frac{M}{\|p_n\|} \quad \text{for all } n.$$

Proof. Apply Singer [10, Thm. 3.1, p. 20] to the space $\text{cl span } \{p_n\}$.

Russell's sufficient condition for (7) to be soluble, assuming for simplicity $\alpha_{ni} = 0$, is that $|\beta_{ni}| \leq \kappa e^{-(M_1 + \epsilon)\sqrt{\lambda_n}}$ where M_1 is the constant of the upper bound to $\|q_{ni}\|$ ($\|q_{ni}\| \leq M_0 e^{M_1 \sqrt{\lambda_n}}$) and κ and $\epsilon > 0$ are arbitrary. This condition is not necessary as seen from Russell's example: $r = 1$, $\alpha_{ni} = 0$, $g \equiv 1$, $\Omega = (0, 1)$, $\Gamma_0 = \{0\}$, $\Gamma_1 = \{1\}$. Then $m_n = 1$ for all n , $\lambda_n = (n - \frac{1}{2})^2 \pi^2$, $\phi_n(x) = \sqrt{2} \sin[(n - \frac{1}{2})\pi x]$ and $\beta_n n^2 \rightarrow \text{const.}$, so clearly for all κ, M there exists n such that $|\beta_n| \not\leq \kappa e^{-M(n-1/2)}$. The point here is that $\{p_n\}$ is not basic in $L^2((0, T) \oplus \Gamma_1)$, i.e. $p_n(t) = \sqrt{2}(-1)^{n-1} e^{-(n-1/2)^2 \pi^2 t}$ is not basic in $L^2(0, T)$.

THEOREM 6. *The sequence $\{p_n\}$ defined above is not basic in $L^2(0, T)$.*

Proof. By Schwartz [9, Théorème II, p. 58], $f \in \text{cl span } \{p_n\}$ if and only if $f \in$

$L^2(0, T)$ and $f(t) = \sum \alpha_n p_n(t)$ almost everywhere in $(0, T)$. Now

$$\sum_{n=1}^{\infty} (-1)^n (2n+1) x^{n(n+1)} = \prod_{n=1}^{\infty} (1-x^{2n})^3 \quad \text{for } |x| < 1$$

by Hardy and Wright [5, Thm. 357, p. 285]. So

$$\sum_{n=1}^{\infty} (-1)^n (2n+1) x^{n(n+1)} x^{1/4} = x^{1/4} \prod_{n=1}^{\infty} (1-x^{2n})^3 \quad \text{for } |x| < 1.$$

Now let $x \rightarrow 1^-$. Then the right-hand side $\rightarrow 0$. Put $x = e^{-\pi^2 t}$. So

$$f(t) = \sum_{m=2}^{\infty} (-1)^{m-1} (2m-1) e^{-(m-1/2)^2 \pi^2 t} \rightarrow 0 \quad \text{as } t \rightarrow 0+.$$

Clearly $f \in L^2(0, T)$ and so $f \in \text{cl span } \{p_n\}$. But

$$f \neq \sum_{m=2}^{\infty} \frac{1}{\sqrt{2}} (2m-1) p_m,$$

in the $L^2(0, T)$ sense, since

$$\|(2m-1)p_m\| \sim 2/\pi \not\rightarrow 0 \quad \text{as } m \rightarrow \infty.$$

So the sequence $\{p_n\}$ is not basic.

Hence we cannot expect a necessary and sufficient condition for the solution of the moment problem in terms of the convergence of $\sum \alpha_n q_n$.

The results above imply that we have exact null controllability for the diffusion equation with control on Γ_1 . This is so because for exact null controllability it is necessary and sufficient that the system

$$-e^{\lambda_n t} \alpha_{ni} = \langle p_{ni}, \tilde{g} \rangle$$

have a solution for each $\sum \alpha_{ni} \phi_{ni}$ in $L^2(\Omega)$. By Russell's estimate for $\|q_{ni}\|$, it is sufficient to have

$$|\alpha_{ni}| \leq \kappa e^{-(M_1 + \varepsilon)\sqrt{\lambda_n} + \lambda_n T}$$

for given M , and arbitrary $\kappa, \varepsilon > 0$. This is always the case since $\{\alpha_{ni}\}$ is bounded. The dual result is that the diffusion equation with $w(x, t) = 0$ for all $x \in \Gamma_0, t \geq 0$, $\partial w / \partial \nu(x, t) = 0$ for all $x \in \Gamma_1, t \geq 0$, and $Cw(t) = w(t)|_{\Gamma_1}$ is continuously finally observable. Dolecki [2] discusses observation at a general point for the one-dimensional case. Sakawa [8] also considers point observations in higher dimensions, but his analysis via biorthogonal systems is void. In fact as we show one point sensor, even if it suffices for observability in higher dimensions, can never yield continuous final observability.

Suppose that X is a Hilbert space with $\{\phi_{ni}\}$ an orthonormal basis and

$$u(t) = \sum_{n=1}^{\infty} e^{-\lambda_n t} \sum_{i=1}^{m_n} \alpha_{ni} \phi_{ni}, \quad \text{where } \sum \alpha_{ni}^2 < \infty,$$

and $\lambda_n \uparrow \infty$.

So u can be considered as the solution of $\dot{u} = Au$ where A is self-adjoint with compact resolvent, densely defined on some Hilbert space of real-valued functions on $\Omega \subset \mathbb{R}^r$. Let $x_n(t) = e^{-\lambda_n t}$, $x_n \in L^2(0, T)$. It is well-known (Schwartz [9, p. 54]) that x_n is minimal if and only if $\sum_{\lambda_n \neq 0} 1/\lambda_n < \infty$. But for the operators that Sakawa considers, $\lim (n^{2/r}/\lambda_n) = \text{const.}$, so x_n is not minimal if $r \geq 2$.

If $C: \text{dom } C \subset X \rightarrow R$ with $\text{dom } C$ dense in X and

$$C(\sum \alpha_{ni} \phi_{ni}) = \sum \sum \alpha_{ni} C \phi_{ni}$$

provided the right-hand side exists, we do not have continuous final observability if $r \geq 2$. Because for $r \geq 2$, $\{x_n\}$ is not minimal. Assume the system is observable, that is, $C\phi_n \neq 0$ for all n ; then there exist for each $\varepsilon > 0$, $\alpha_2, \dots, \alpha_{N(\varepsilon)}$ such that

$$\left\| C\phi_1 x_1 - \sum_{i=2}^{N(\varepsilon)} \alpha_i C\phi_i x_i \right\|_{L^2(0, T)} < \varepsilon.$$

Let

$$u(0) = \phi_1 - \sum_{i=2}^{N(\varepsilon)} \alpha_i \phi_i;$$

then $\|u(T)\| \geq e^{-\lambda_1 T}$ but $\int_0^T \|Cu(t)\|^2 dt < \varepsilon$, so we do not have continuous final observability. The positive results mentioned above for $r \geq 2$ refer to observation on a subset Γ_1 with nonempty interior relative to Γ .

REFERENCES

- [1] M. M. DAY, *Normed Linear Spaces*, third ed., Springer-Verlag, Berlin, 1973.
- [2] S. DOLECKI, *Observability for the one-dimensional heat equation*, *Studia Math.*, 48 (1973), pp. 291–305.
- [3] S. DOLECKI AND D. L. RUSSELL, *A general theory of observation and control*, Math. Research Center, Univ. of Wisconsin Rep. 1519, 1975.
- [4] H. O. FATTORINI AND D. L. RUSSELL, *Uniform bounds on biorthogonal functions for real exponentials with an application to the control theory of parabolic equations*, *Quart. Appl. Math.*, 32 (1974), pp. 45–69.
- [5] G. H. HARDY AND E. M. WRIGHT, *An Introduction to the Theory of numbers*, Oxford University Press, London, 1959.
- [6] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, Colloquium Publications, vol. 31, American Mathematical Society, Providence, RI, 1957.
- [7] D. L. RUSSELL, *Problems of control and stabilization for partial differential equations*, Preprint, IFAC 6th Congress, (Boston, 1975).
- [8] Y. SAKAWA, *Observability and related problems for partial differential equations of parabolic type*, this *Journal*, 13 (1975), pp. 14–27.
- [9] L. SCHWARTZ, *Etude des Sommes d'Exponentielles*, second ed., Hermann, Paris, 1959.
- [10] I. SINGER, *Bases in Banach Spaces I*, Springer-Verlag, Berlin, 1970.

THE EQUIVALENCE OF STRONG AND WEAK FORMULATIONS FOR CERTAIN PROBLEMS IN OPTIMAL CONTROL*

RICHARD B. VINTER† AND RICHARD M. LEWIS‡

Abstract. It is shown that a wide class of control problems may be cast equivalently as convex mathematical programming problems over a space of Radon measures. The particularly simple structure of the equivalent problems will enable us to obtain new conditions for optimality, of dynamic programming type, by application of convex analysis. These implications are pursued in a companion paper.

1. Introduction. This is the first of two companion papers which, together, develop conditions for optimality for a wide class of terminally constrained control problems. The conditions are of a flavor reminiscent of the earlier “sufficiency conditions” literature [1], [3] which characterizes optimality through a partial differential equation, the Bellman equation—the novelty here is that, while dispensing with all a priori assumptions concerning the optimal controls (such as existence of an optimal feedback synthesis with some “admissible set of discontinuities”), our conditions are necessary as well as sufficient for optimality. This is accomplished by replacing the partial differential equation by a partial differential inequality, and requiring this to hold only in a limiting sense.

Derivation of the results proceeds in two steps. We first cast the problem as a constrained minimization problem over a certain set of Radon measures. The recast problem is a *convex* mathematical programming problem. The final results emerge as a statement of “strong” duality between this recast problem and its “Fenchel” dual.

In this, the first paper, we establish equivalence of the original (or “strong”) problem and the recast (or “weak”) problem, thereby validating the development just outlined.

Proof of the equivalence is carried out as follows: we embed elements feasible for the strong problem in the class of elements feasible for the weak problem which in turn are embedded in feasible elements for a “parametric” problem. The object here is to make possible application of a refinement of a result due to L. C. Young concerning density of “polygonal flows” in elements feasible for the parametric problem; the result enables us to establish that a solution to the parametric problem may be selected which under the inverse embeddings is feasible for the original strong problem. Equivalence follows immediately.

Because of the introduction of the parametric problem, which plays no part in the statement of results, the proof of equivalence may appear unnecessarily roundabout. While acknowledging the desirability of a direct proof, we feel that such a proof would necessarily be long and involved. The major task would be in obtaining an analogue for nonparametric problems of Young’s density theorem. A density theorem for the parametric problem presents considerable difficulties, even when we exploit Young’s device of defining a linear space of “boundary conditions”; for the nonparametric problem where the device is no longer available to us, the difficulties would be even more severe. Our development may therefore be viewed as economizing on having to prove a difficult density theorem at the cost of introducing indirect arguments to graft Young’s results into our proofs.

* Received by the editors February 28, 1977, and in revised form August 12, 1977.

† Department of Computing and Control, Imperial College of Science and Technology, London SW7 2BZ, England.

‡ Department of Computing and Control, Imperial College of Science and Technology, London, England. Now at School of Mathematics, University of Bath, Bath BA2 7AY, England.

Finally we mention that Rubio too has studied the weak problem [8], [9] for fixed endpoints. The emphasis in these papers is on existence of solutions to the weak problem (under rather more general conditions) and equivalence with the strong problem, in the sense that it is required to set up optimality conditions as in the companion paper [12], is not established.

Some generalizations of the results reported here and in the companion paper are given in [6].

2. Notations. Let S be a compact space. Then $C(S)$ denotes the Banach space of continuous real-valued functions on S with the sup norm. In the case that S is a subset of \mathbb{R}^k , $C^1(S)$ denotes the subset of $C(S)$ comprising restrictions to S of continuously differentiable functions $\mathbb{R}^k \rightarrow \mathbb{R}$.

$C^*(S)$ is written for the (normed) dual of $C(S)$. It is well known that each $f' \in C^*(S)$ has unique representation through a (signed) Radon measure μ on (the Borel sets of) S , thus

$$f'(g) = \int_S g \, d\mu, \quad \text{all } g \in C(S)$$

[11, p. 397]. In the sequel, we shall not distinguish between bounded, linear functionals on $C(S)$ and the measures which represent them, writing $\int_S g \, d\mu$ for the action of $\mu \in C^*(S)$ on $g \in C(S)$.

$P(S)$ denotes the subset of $C(S)$ comprising functions taking only nonnegative values, while $P^\oplus(S)$ is the positive polar cone of this subset, i.e.

$$P^\oplus(S) = \left\{ \mu \in C^*(S) \mid \int g \, d\mu \geq 0 \right\}, \quad \text{all } g \in P(S).$$

$|\cdot|$ denotes the norm over the linear space determined by context (we shall need only in Appendix A to distinguish between different norms on the same linear space in the notation). If this is \mathbb{R}^k , then $|\cdot|$ is the Euclidean norm. $C^*(S)$ carries the dual norm

$$|\mu| = \sup \left\{ \int g \, d\mu \mid |g| = 1 \right\}, \quad \text{for } \mu \in C^*(S).$$

We write

$$P^n(S) = \{ \mu \in P^\oplus(S) \mid |\mu| = 1 \}.$$

Finally, if $f: S \rightarrow R$ is a measurable map between measurability spaces (S, \mathcal{S}) , (R, \mathcal{R}) and ν is a measure on (S, \mathcal{S}) , then $\nu \circ f^{-1}$ denotes the measure which ν induces on (R, \mathcal{R}) under f .

3. Relaxed controls and admissible pairs. Let be given

$$\begin{aligned} Q &\subset \mathbb{R}^n, & \Omega &\subset \mathbb{R}^m; \\ (x, t, u) &\mapsto l(x, t, u): \mathbb{R}^{n+m+1} \rightarrow \mathbb{R}; \\ (x, t, u) &\mapsto f(x, t, u): \mathbb{R}^{n+m+1} \rightarrow \mathbb{R}^n; \\ x_0 &\in \mathbb{R}^n; t_0, T \in \mathbb{R} \quad \text{with } T > t_0; \\ \Gamma &\subset Q \times [t_0, T]. \end{aligned}$$

The following assumptions are made:

(i) Q, Ω, Γ are compact.

(ii) f, l are continuous.

A $C^*(\Omega)$ -valued function $\{\mu_t; t_0 \leq t \leq t_1\}$, $[t_0, t_1] \subset [t_0, T]$, defined modulo (Lebesgue) null functions, will be termed a *relaxed control* in case

(i) $\mu_t \in P^n(\Omega)$, a.e. $t \in [t_0, t_1]$

(ii) $t \mapsto \int_{\Omega} \phi(t, u) \mu_t(u)$ is (Lebesgue) measurable for every $\phi \in C^1([t_0, t_1] \times \Omega)$ (cf. [13, Chap. 4] where relaxed controls are defined in a more general context).

Relaxed controls are of course an enlargement of the class of "ordinary" controls $\{u(t); t_0 \leq t \leq t_1\}$ (Lebesgue measurable Ω -valued functions), ordinary controls being embedded in the larger class thus

$$u(\cdot) \mapsto \{\delta(u(t)); t_0 \leq t \leq t_1\}$$

($\delta(v)$, unit measure concentrated at v).

A couple $\{\mu_t, x(t); t_0 \leq t \leq t_1\}$, with $\{\mu_t; t_0 \leq t \leq t_1\}$ a relaxed control and $\{x(t); t_0 \leq t \leq t_1\}$ absolutely continuous, taking values in Q will be termed an *admissible (control trajectory) pair* in case

$$(i) \quad x(t_0) = x_0, \quad (x(t_1), t_1) \in \Gamma, \quad \text{and}$$

$$(3.1) \quad (ii) \quad \dot{x}(t) = \int f(x(t), t, u) d\mu_t(u) \quad \text{a.e. } t \in [t_0, t_1].$$

Existence of an admissible pair is assumed in the sequel.

4. The control problem and its weak formulation. We shall be concerned with the control problem (the *Strong Problem*)

$$(S) \quad \begin{aligned} &\text{Minimize } \int_{t_0}^{t_1} \int l(x(t), t, u) d\mu_t(u) dt \\ &\text{over admissible pairs } \{\mu_t, x(t); t_0 \leq t \leq t_1\}. \end{aligned}$$

Let A be a cube in \mathbb{R}^n containing Q , and write $\mathbf{A} = A \times [t_0, T]$. In the spirit of [14, p. 282ff], we view admissible pairs $\{\mu_t, x(t); t_0 \leq t \leq t_1\}$ as defining elements in $C^*(\mathbf{A} \times \Omega)$ thus

$$(4.1) \quad g \mapsto \int_{t_0}^{t_1} \int g(x(t), t, u) d\mu_t(u) dt \quad \text{for } g \in C(\mathbf{A} \times \Omega)$$

(simple verification that the mapping is well-defined, linear and bounded is omitted).

The Strong Problem (S) may then be cast as that of minimizing the value of a functional

$$\mu \mapsto \int l(x, t, u) d\mu(x, t, u)$$

over elements $\mu \in C^*(\mathbf{A} \times \Omega)$, subject to the constraint that the μ 's have representation (4.1) for some admissible control trajectory pair $\{\mu_t, x(t); t_0 \leq t \leq t_1\}$.

Let us explore in more detail the nature of elements $\mu \in C^*(\mathbf{A} \times \Omega)$ arising in this way. Evidently

$$(i) \quad \mu \in P^{\oplus}(\mathbf{A} \times \Omega).$$

We notice also that, for $\phi \in C^1(\mathbf{A})$ and Lipschitz continuous $\{x(t), t_0 \leq t \leq t_1\}$, $t \mapsto \phi(x(t), t)$ is Lipschitz continuous and

$$\frac{d}{dt} \phi(x(t), t) = \phi_t(x(t), t) + \int \phi_x(x(t), t) f(x(t), u) d\mu_t(u) \quad \text{a.e. } t \in [t_0, t_1].$$

In consequence,

$$\begin{aligned}\phi(x(t_1), t_1) - \phi(x(t_0), t_0) &= \int_{t_0}^{t_1} \frac{d}{dt} \phi(x(t), t) dt \\ &= \int_{t_0}^{t_1} \int (\phi_t(x(t), t) + \phi_x(x(t), t)f(x(t), t, u)) d\mu_t(u) dt.\end{aligned}$$

This imposes the requirement that there exists $(x_1, t_1) \in \Gamma$ with

$$(4.2) \quad \int (\phi_t + \phi_x f) d\mu = \phi(x_1, t_1) - \phi(x_0, t_0), \quad \text{all } \phi \in C^1(\mathbf{A}),$$

recalling that for admissible pairs, $(x(t_1), t_1) \in \Gamma$. Equation (4.2) implies, in particular, existence of some $\beta \in P^n(\Gamma)$ such that

$$(ii) \quad \int (\phi_t + \phi_x f) d\mu = \int \phi d\beta - \phi(x_0, t_0), \quad \text{all } \phi \in C^1(\mathbf{A}).$$

Finally we remark that

$$(iii) \quad \int \text{dist}\{x, Q\} d\mu = 0$$

(dist denotes Euclidean distance), for the left-hand side is none other than

$$\int_{t_0}^{t_1} \text{dist}\{x(t), Q\} dt,$$

which is equal to zero since admissible trajectories take values in Q .

The weak formulation of the problem may now be given. This amounts to a weakening of the constraints to require only (i), (ii) and (iii) to hold:

$$\begin{aligned}(W) \quad & \text{Minimize } \int l d\mu \text{ over } \mu \in P^\oplus(\mathbf{A} \times \Omega) \\ & \text{subject to } \int \text{dist}\{x, Q\} d\mu = 0 \\ & \text{and there exists some } \beta \in P^n(\Gamma) \\ & \text{with } \int (\phi_t + \phi_x f) d\mu = \int \phi d\beta - \phi(x_0, t_0), \quad \text{all } \phi \in C^1(\mathbf{A}).\end{aligned}$$

The significance of this weakening of the constraints and of casting the problem into this framework is that we thereby obtain a convex mathematical programming problem. We remark that feasible elements for the *Weak Problem* (W) are constrained to satisfy (ii), rather than (4.2) precisely to convexify the constraint set.

The machinery of convex analysis is now at hand to characterize solutions and to supply lower bounds for the minimum value of the objective functional. Such a development is pursued in a companion paper [12]. In the present paper we take the crucial step of proving equivalence (in an appropriate sense) of the Strong and the Weak Problems.

5. The main results on equivalence of the Strong and Weak Problems. We introduce some terminology and notation. In the two problems (S) and (W) of § 4, elements (admissible pairs or measures) satisfying the constraints will be termed *feasible*. The infimum of the values of the functional to be minimized over feasible

elements will be termed the *value* of the minimization problem. The notation $\eta(S)$ (or $\eta(W)$) attaches to the value of (S) (or (W)).

We establish equivalence only under the following technical condition:

Condition 5.1. The function

$$(x, t, u) \mapsto \min_{w \in \Omega} \{l(x, t, w) \mid f(x, t, u) = f(x, t, w)\}$$

mapping $\mathbf{A} \times \Omega \rightarrow \mathbb{R}$ is continuous.

THEOREM 5.1. *Suppose that Condition 5.1 is met. Then the problems (S), (W) both have solutions and*

$$\eta(S) = \eta(W).$$

Further, if the admissible pair $\{\mu_t, x(t) \mid t_0 \leq t \leq t_1\}$ solves the Strong Problem (S), then the element $\mu \in C^(\mathbf{A} \times \Omega)$ defined by*

$$(5.1) \quad g \mapsto \int_{t_0}^{t_1} \int g(x(t), t, u) d\mu_t(u) dt \quad \text{for } g \in C^*(\mathbf{A} \times \Omega)$$

solves the Weak Problem.

The substance of the present paper is proof of this result. Of course existence of a solution to the Strong Problem (S) is well-known under considerably more general conditions than those applying here [10].

Some comments about Condition 5.1 are in place. While it is easy to construct counterexamples in which the condition is violated, the condition clearly holds if either

- (i) $u \mapsto f(x, t, u)$ is injective, all $(x, t) \in \mathbf{A}$, or
- (ii) there exists some $g \in C(\mathbf{A} \times \mathbb{R}^n)$ such that

$$l(x, t, u) = g(x, t, f(x, t, u)), \quad \text{all } (x, t, u) \in \mathbf{A} \times \Omega$$

((ii) may be interpreted as penalization of the velocity, rather than the control, in the cost). Thus the condition is seen to hold in many cases of interest.

Actually we may always arrange that Condition 5.1 is satisfied by augmenting the state variables. Indeed consider the minimization problem over pairs $\{(x(t), \tilde{x}(t)), \mu_t(\cdot); t_0 \leq t \leq t_1\}$ with $\{(x(t), \tilde{x}(t)); t_0 \leq t \leq t_1\}$ an absolutely continuous \mathbb{R}^{n+m} -valued function, $\{\mu_t(\cdot); t_0 \leq t \leq t_1\}$ a relaxed control as above:

$$\begin{aligned} & \text{minimize } \int_{t_0}^{t_1} \int l_a(x(t), \tilde{x}(t), t, u) d\mu_t(u) dt \\ & \text{subject to } (\dot{x}(t), \dot{\tilde{x}}(t)) = f_a(x(t), \tilde{x}(t), t, u) d\mu_t(u) \quad \text{a.e.}, \\ (S_a) \quad & (x(t_0), \tilde{x}(t_0)) = (x_0, 0), \quad (t_1, x(t_1), \tilde{x}(t_1)) \in \Gamma_a, \quad (x(t), \tilde{x}(t)) \in A, \\ & t_0 \leq t \leq t_1, \quad t_0 \text{ fixed.} \end{aligned}$$

Here,

$$l_a(x, \tilde{x}, t, u) = l(x, t, u), \quad f_a(x, \tilde{x}, t, u) = (f(x, t, u), u), \quad \Gamma_a = \Gamma \times \Sigma$$

(Σ a ball in \mathbb{R}^m containing $\{T \cdot u \mid u \in \Omega\}$), $A_a = A \times \Sigma$. (S) and (S_a) are equivalent to the extent that $\{x(t), \mu_t(\cdot); t_0 \leq t \leq t_1\}$ solves (S) if $\{(x(t), \tilde{x}(t)), \mu_t(\cdot); t_0 \leq t \leq t_1\}$ solves (S_a) and $\{(x(t), \int_{t_0}^t u d\mu_s(u)), \mu_t(\cdot); t_0 \leq t \leq t_1\}$ solves (S_a) if $\{x(t), \mu_t(\cdot); t_0 \leq t \leq t_1\}$ solves (S). The new problem satisfies Condition 5.1 (in addition to the assumptions introduced in § 3), because $u \mapsto f_a(x, t, u)$ is injective for all $(x, t) \in \mathbb{R}^{n+m}$.

Notice however that satisfaction of Condition 5.1 is achieved only at the cost of increasing the dimension of the state space. By Theorem 5.1 $\eta(S_a) = \eta(W_a)$ ((W_a) the weak problem corresponding to (S_a)); however we cannot conclude from this fact that $\eta(S) = \eta(W)$. As to whether Condition 5.1 can be dropped from Theorem 5.1 remains an interesting open question.

Finally we remark that (in the absence of convexity assumptions) it is essential that the Strong Problem be posed over *relaxed* controls. Indeed examples of control problems with terminal constraints are well known [13], where the infimum of the cost over ordinary controls is achieved and is strictly greater than the infimum over relaxed controls. In such examples, equivalence of the Strong Problem (if posed over ordinary controls) and the Weak Problem would fail.

6. Some results on the structure of generalized flows. We bring together here two theorems which form the basis of establishing the equivalence of the Strong and the Weak Problems. Again let \mathbf{A} be a cube in \mathbb{R}^{n+1} and let $\Gamma \subset \mathbf{A}$ be a closed subset. We write

$$\mathbf{B} = \{\dot{y} \in \mathbb{R}^{n+1} \mid |\dot{y}| = 1\}.$$

Let $\{y_1(\sigma), \dots, y_k(\sigma); 0 \leq \sigma \leq 1\}$ be a collection of Lipschitz continuous curves from a fixed point $y_0 \in \mathbb{R}^n$ to Γ taking values in \mathbf{A} , and let $\alpha_1, \dots, \alpha_k$ be positive coefficients. The collection of curves and coefficients is termed a *positive mixture of curves* and defines a bounded linear functional μ on $C(\mathbf{A} \times \mathbf{B})$ through¹

$$g \mapsto \sum_i \alpha_i \int_0^1 g(y_i(\sigma), \dot{y}_i(\sigma)/|\dot{y}_i(\sigma)|) |\dot{y}_i(\sigma)| d\sigma$$

(the integrand takes value zero when $\dot{y}(\sigma) = 0$).

We observe that

$$(6.1) \quad \int \phi_y \dot{y} d\mu = \int \phi d\beta - \phi(y_0), \quad \text{all } \phi \in C^1(\mathbf{A}),$$

when β is the measure with finite support:

$$\beta = \sum \alpha_i \delta(y_i(1)).$$

The first result asserts that any element $\mu \in P^\oplus(\mathbf{A} \times \mathbf{B})$ satisfying the “boundary condition” (6.1) for some $\beta \in P^n(\Gamma)$ may be approximated by positive mixtures of “polygonal arcs” from y_0 to Γ and “closed polygonal arcs”.

THEOREM 6.1 (Approximation of generalized flows with a prescribed boundary). *Take $\nu \in P^\oplus(\mathbf{A} \times \mathbf{B})$, and suppose that there exist $y_0 \in \mathbf{A}$, $\beta \in P^n(\Gamma)$ such that*

$$\int \phi_y \dot{y} d\nu = \int \phi d\beta - \phi(y_0), \quad \text{all } \phi \in C^1(\mathbf{A}).$$

Then ν is the weak* limit of a sequence $\{\nu_i\}$ in $C^*(\mathbf{A} \times \mathbf{B})$ where

$$(6.2) \quad \nu_i = \sum_j \alpha_j^i \nu_j^i + \sum_j \beta_j^i \tau_j^i.$$

¹ To motivate introduction of functionals defined in this way, observe that in parametric problems in the calculus of variations we deal with integrands $\tilde{g}(y, \dot{y})$ defined on $\mathbf{A} \times \mathbb{R}^n$, positively homogeneous in their argument \dot{y} . Positive homogeneity assures that \tilde{g} is actually specified by its restriction g to $\mathbf{A} \times \mathbf{B}$ and $\int_0^1 g(y(\sigma), \dot{y}(\sigma)/|\dot{y}(\sigma)|) |\dot{y}(\sigma)| d\sigma$ is nothing but the line integral $\int_0^1 \tilde{g}(y(\sigma), \dot{y}(\sigma)) d\sigma$ along the curve y .

In (6.2), the number of terms in the summations is finite, but may depend on i . The α_j^i 's, β_j^i 's are positive with $\sum_j \alpha_j^i = 1$, each i , and ν_j^i has the representation

$$(6.3) \quad g \mapsto \int_0^1 g(y(\sigma), \dot{y}(\sigma)/|y(\sigma)|) |\dot{y}(\sigma)| d\sigma, \quad \text{all } g \in C(\mathbf{A} \times \mathbf{B}),$$

for some continuous, piecewise linear function $\{y(\sigma); 0 \leq \sigma \leq 1\}$ with values in \mathbf{A} and such that $y(0) = y_0$, $y(1) \in \Gamma$. Finally, each τ_j^i has representation (6.3) for some continuous piecewise linear function $\{y(\sigma); 0 \leq \sigma \leq 1\}$ with values in \mathbf{A} and such that $y(0) = y(1)$.

Theorem 6.1 is proved in Appendix A.

The second theorem concerns the structure of "generalized curves":

DEFINITION 6.1. $\nu \in C^*(\mathbf{A} \times \mathbf{B})$ is called a *generalized curve* in case it is the weak* limit of a sequence $\{\nu^i\}$ in $C^*(\mathbf{A} \times \mathbf{B})$ where each ν^i has representation (6.3) for some Lipschitz continuous function $\{y(\sigma); 0 \leq \sigma \leq 1\}$ taking values in \mathbf{A} .

THEOREM 6.2 (Representation of generalized curves). *Let $\nu \in C^*(\mathbf{A} \times \mathbf{B})$ be a generalized curve. Then there exists a Lipschitz continuous function $\{y(\sigma); 0 \leq \sigma \leq 1\}$, taking values in \mathbf{A} and a family $\{\mu_\sigma \in P^\oplus(\mathbf{B}); 0 \leq \sigma \leq 1\}$ with $|\mu_\sigma|$ uniformly bounded such that:*

- (i) $\sigma \mapsto \int g(y(\sigma), \dot{y}) d\mu_\sigma(\dot{y})$ is (Lebesgue) measurable for each $g \in C([0, 1] \times \mathbf{B})$,
- (ii) $\int g d\mu = \int_0^1 \int g(y(\sigma), \dot{y}) d\mu_\sigma(\dot{y}) d\sigma$, for each $g \in C(\mathbf{A} \times \mathbf{B})$, and
- (iii) $\dot{y}(\sigma) = \int \dot{y} d\mu_\sigma(\dot{y})$ a.e. $\sigma \in [0, 1]$,

where $\{\dot{y}(\sigma); 0 \leq \sigma \leq 1\}$ is the derivative of $y(\cdot)$.

This result is proved in [14, p. 171].

7. The parametric problem. We shall cast the Weak Problem as a parametric problem in order to apply the results of § 6. We begin with a few informal comments to motivate our particular choice of problem. Because Theorems 6.1, 6.2 are stated in terms of trajectories, we treat the original Strong Problem as a minimization problem over trajectories and introduce satisfaction of the differential equation (for some control) as a side constraint.

If we take

$$F = \{\dot{x} \in \mathbb{R}^n \mid |\dot{x}| \leq K\}$$

as a ball containing $f(\mathbf{A} \times \Omega)$, this translates into posing the Weak Problem over elements $\nu \in C^*(\mathbf{A} \times F)$ (rather than $C^*(\mathbf{A} \times \Omega)$) subject to the constraints

- (a) there exists some $\beta \in P^n(\Gamma)$ such that

$$\int \{\phi_t + \phi_x \dot{x}\} d\nu = \int \phi d\beta - \phi(x_0, t_0), \quad \phi \in C^1(\mathbf{A}),$$

- (b) $\int \text{dist}\{\dot{x}, f(x, t, \Omega)\} d\nu = 0$,

- (c) $\int \text{dist}\{x, Q\} d\nu = 0$.

Of course (b) corresponds to the side constraint in the Strong Problem:

$$\dot{x}(t) \in f(x(t), t, \Omega).$$

What is the appropriate choice of objective functional when the minimization is performed over trajectories rather than admissible pairs? It is natural to introduce

$$q(x, t, x) = \min_{u \in \Omega} \{l(x, t, u) \mid x = f(x, t, u)\}$$

and, loosely speaking, to seek to minimize

$$(7.1) \quad \int q(x(t), t, \dot{x}(t)) dt$$

for one would hope that the value of this functional is the minimum of the values of the cost of the original Strong Problem over all controls which correspond to a given fixed trajectory. Unfortunately the function q is not continuous on its domain in general. However we have:

LEMMA 7.1. *Under Condition 5.1, q is continuous on its domain and may be extended continuously to all of $\mathbf{A} \times \mathbb{R}^n$ such that the extension is uniformly bounded.*

This result is proved in Appendix B.

Henceforth we assume Condition 5.1 to be in force and q denotes the extension (which is of course not unique). (7.1) now translates into

$$\int q(x, t, \dot{x}) d\nu$$

in the corresponding Weak Problem. Finally we must adopt parametric form: a simple change of variables would suggest replacing \dot{x} by \dot{x}/i and multiplying the integrands by $|i|$. We need also to introduce an additional constraint to ensure that time is monotone nondecreasing in the new parameter. Accordingly we study the *Parametric Problem* (P):

$$\text{Minimize } \int q(x, t, \dot{x}/i)|i| d\nu \quad \text{over } \nu \in P^\oplus(\mathbf{A} \times \mathbf{B})$$

subject to: there exists $\beta \in P^n(\Gamma)$ with

$$(7.2) \quad \int \{\phi, i + \phi_x \dot{x}\} d\nu = \int \phi d\beta - \phi(x_0, t_0), \quad \text{all } \phi \in C^1(\mathbf{A}),$$

$$(P) \quad \int \text{dist}\{\dot{x}/i, f(x, t, \Omega)\}|i| d\nu = 0,$$

$$\int \text{dist}\{x, Q\}|i| d\nu = 0,$$

$$\int \max\{-i, 0\} d\nu = 0.$$

Here \mathbf{A} , Γ are the subsets of \mathbb{R}^{n+1} introduced in § 3. Consistent with § 6, \mathbf{B} is $\{(\dot{x}, i) \in \mathbb{R}^{n+1} | \dot{x}^2 + i^2 = 1\}$. In the sequel we shall find it convenient often to write couples (x, t) , or (\dot{x}, i) (in \mathbb{R}^{n+1}) as y , or \dot{y} .

$$(y, \dot{y}) \mapsto q(x, t, \dot{x}/i)|i|, \quad (y, \dot{y}) \mapsto \text{dist}\{\dot{x}/i, f(x, t, \Omega)\}|i|$$

are assigned the values 0 and 1 respectively when $i = 0$. With this convention the two functions become continuous on $\mathbf{A} \times \mathbf{B}$ (see Lemma B.1 in Appendix B), and consequently well-define elements in the predual of $C^*(\mathbf{A} \times \mathbf{B})$. The other integrands in (P) obviously share this property. We employ the shorthand $\mathbf{q}(y, \dot{y})$, $\mathbf{d}(y, \dot{y})$, $\mathbf{s}(y, \dot{y})$, $\mathbf{m}(y, \dot{y})$ for $q(x, t, \dot{x}/i)|i|$, $\text{dist}\{\dot{x}/i, f(x, t, \Omega)\}|i|$, $\text{dist}\{x, Q\}|i|$ and $\max\{-i, 0\}$ respectively.

8. Existence of solutions to the parametric problem. Our first result concerns existence of a solution to (P). This proceeds by a standard compactness argument, following development of two technical lemmas which supply a bound on elements feasible for the Parametric Problem.

The first lemma gives some useful implications of the constraints:

LEMMA 8.1. *Suppose that $\nu \in P^\oplus(\mathbf{A} \times \mathbf{B})$ takes value zero on \mathbf{d} , $\mathbf{m} \in C(\mathbf{A} \times \mathbf{B})$; then*

$$\text{supp}\{\nu\} \subset \{(y, \dot{y}) \in \mathbf{A} \times \mathbf{B} | i \geq 0, \dot{x} \in if(x, t, \Omega)\}$$

(supp denotes support of the measure).

The lemma is proved in Appendix B.

In simple consequence of the above we have:

LEMMA 8.2. *The subset of $P^\oplus(\mathbf{A} \times \mathbf{B})$ comprising all elements ν feasible for (P) is (norm) bounded.*

Proof. Let ν be feasible for (P). By Lemma 8.1,

$$\int |i| d\nu = \int i d\nu = \int (\phi_t i + \phi_x \dot{x}) d\nu$$

(for $\phi(x, t) = t$).

By (7.2) then

$$\int |i| d\nu = \int_{\Gamma} t d\beta - t_0$$

(for some $\beta \in P^n(\Gamma)$) whence

$$\int |i| d\nu \leq T - t_0.$$

Now since points (\dot{x}, i) in \mathbf{B} are constrained by $\dot{x}^2 + i^2 = 1$,

$$\begin{aligned} |\nu| &= \int 1 d\nu = \int (\dot{x}^2 + i^2)^{1/2} d\nu \leq \int (|\dot{x}| + |i|) d\nu \\ &\leq (1 + K) \int |i| d\nu \end{aligned}$$

(K , a bound on $f(\mathbf{A}, \Omega)$, as in § 7). The last inequality here follows from Lemma 8.1.

We have shown that feasible ν are bounded by $(1 + K)(T - t_0)$.

PROPOSITION 8.1. *(Under Condition 5.1), there exists a solution to the Parametric Problem (P).*

Proof. The constraint set for (P) is nonempty. Indeed by assumption there exists some admissible pair $\{\mu_t, x(t); t_0 \leq t \leq t_1\}$. Define $\nu \in C^*(\mathbf{A} \times \mathbf{B})$ as follows: for $g \in C(\mathbf{A} \times \mathbf{B})$,

$$g \mapsto \int_{t_0}^{t_1} \left\{ \int \left[g\left(x(t), t, \frac{\dot{x}}{(1 + |\dot{x}|^2)^{1/2}}, \frac{1}{(1 + |\dot{x}|^2)^{1/2}}\right) (1 + |\dot{x}|^2)^{1/2} \right]_{\dot{x}=f(x(t), t, u)} d\mu_t(u) \right\} dt$$

with $(\mu_t, x(t))$ as above. We omit the simple verification that this is feasible for (P).

Let Σ be a closed ball in $C^*(\mathbf{A} \times \mathbf{B})$ containing all feasible elements for (P) (see Lemma 8.2). Equip Σ with the induced weak* topology. Let us show that the set of feasible elements \mathcal{F} is compact in Σ . Since Σ is compact, metrizable [2, p. 424], it suffices to show that \mathcal{F} is sequentially closed. Accordingly let $\{\nu_i\}$ be a sequence in \mathcal{F} . Write β_i for the unit measures associated with the “endpoints” of the ν_i ’s, and suppose that

$$\nu_i \rightarrow \nu \quad (\text{weakly}^*).$$

By weak* convergence and closedness of $P^\oplus(\mathbf{A} \times \mathbf{B})$ in $C^*(\mathbf{A} \times \mathbf{B})$, we have

$$\int \mathbf{d} d\nu = \int \mathbf{s} d\nu = \int \mathbf{m} d\nu = 0, \quad \nu \in P^\oplus(\mathbf{A} \times \mathbf{B}).$$

It remains to check (7.2). By weak* compactness of $P^n(\Gamma)$ we may extract a subsequence such that

$$\beta_i \rightarrow \beta_0 \quad (\text{weakly})^* \quad \text{for some } \beta_0 \in P^n(\Gamma).$$

Given arbitrary $\phi \in C^1(\mathbf{A})$ we have for this subsequence

$$\begin{aligned} \int \{\phi, \dot{y}\} d\nu &= \lim_i \int \{\phi, \dot{y}\} d\nu_i = \lim_i \left\{ \int \phi d\beta_i - \phi(x_0, t_0) \right\} \\ &= \int \phi d\beta_0 - \phi(x_0, t_0). \end{aligned}$$

Thus (7.2) holds and $\mu \in \mathcal{F}$. But this establishes that \mathcal{F} is compact in S . Now $\nu \mapsto \int q d\nu$ is weakly* continuous on Σ . The problem therefore reduces to minimization of a continuous function on a (nonempty) compact set; it has a solution by an elementary result.

9. A generalized curve solution to the parametric problem. In this section the approximation theorem of § 6 is applied to extract a generalized curve solution to (P). We shall require

LEMMA 9.1. *Suppose that $\nu \in C^*(\mathbf{A} \times \mathbf{B})$ has representation*

$$\int g d\nu = \sum_i \alpha_i \int_0^1 g(y_i(\sigma), \dot{y}_i(\sigma)) d\sigma \quad \text{for } g \in C(\mathbf{A} \times \mathbf{B})$$

where the y_i 's are a finite collection of Lipschitz continuous functions on $[0, 1]$ satisfying $y(1) = y(0)$, and the real numbers α_i are positive. Then

$$\int \text{dist} \{ \dot{x}/t, f(x, t, \Omega) \} |t| d\nu + 2(K+1) \int \max \{-t, 0\} d\nu \geq |\nu|$$

(K , bound on $f(\mathbf{A}, \Omega)$).

The result is proved in Appendix B.

PROPOSITION 9.1. *The parametric problem (P) has a solution which is a generalized curve with endpoints in $\{y_0\} \times \Gamma$.*

Proof. Let ν_0 solve (P). Existence of such a ν_0 , is assured by Proposition 8.1. Let $\{\bar{\nu}^i\}$ be a sequence of elements in $C^*(\mathbf{A} \times \mathbf{B})$ of the special structure described in Theorem 6.1, converging (weakly*) to ν_0 . Thus

$$\bar{\nu}^i = \nu^i + \tau^i, \quad i = 1, 2, \dots,$$

with ν^i a "unit mixture of polygonal arcs" from $y_0 = (x_0, t_0)$ to Γ ,

$$\nu^i = \sum_j \bar{\alpha}_j^i \bar{\nu}_j^i$$

and τ^i a "mixture of closed polygonal arcs",

$$\tau^i = \sum_j \beta_j^i \tau_j^i.$$

By weak* convergence

$$\int \mathbf{d} d\bar{\nu}^i, \quad \int \mathbf{s} d\bar{\nu}^i, \quad \int \mathbf{m} d\bar{\nu}^i \rightarrow 0.$$

Now

$$\nu^i, \tau^i \in P^\oplus(\mathbf{A} \times \mathbf{B}) \quad \text{and} \quad \mathbf{d}, \mathbf{m} \in P(\mathbf{A} \times \mathbf{B}).$$

It follows that

$$\int \mathbf{d} d\tau^i, \quad \int \mathbf{m} d\tau^i \rightarrow 0.$$

By Lemma 9.1 then, $|\tau^i| \rightarrow 0$.

We have therefore

$$\nu^i \rightarrow \nu_0 \quad (\text{weakly}^*)$$

and in consequence

$$(9.1) \quad \begin{aligned} \int \mathbf{q} \, d\nu^i &\rightarrow \eta(P), \\ |\nu^i| &= \int 1 \, d\nu^i \rightarrow |\nu_0|, \\ \int \mathbf{d} \, d\nu^i, \int \mathbf{s} \, d\nu^i, \int \mathbf{m} \, d\nu^i &\rightarrow 0. \end{aligned}$$

Now for each i , the collection of real numbers

$$\left\{ \sum_j \bar{\alpha}_j^i \int \mathbf{q} \, d\bar{\nu}_j^i, \sum_j \bar{\alpha}_j^i \int \mathbf{d} \, d\bar{\nu}_j^i, \sum_j \bar{\alpha}_j^i \int \mathbf{s} \, d\bar{\nu}_j^i, \sum_j \bar{\alpha}_j^i \int \mathbf{m} \, d\bar{\nu}_j^i, \sum_j \bar{\alpha}_j^i \int 1 \, d\bar{\nu}_j^i \right\}$$

defines a point in the convex hull of the set

$$\left\{ \xi \in \mathbb{R}^5 \mid \xi_1 = \int \mathbf{q} \, d\bar{\nu}_j^i, \xi_2 = \int \mathbf{d} \, d\bar{\nu}_j^i, \xi_3 = \int \mathbf{s} \, d\bar{\nu}_j^i, \xi_4 = \int \mathbf{m} \, d\bar{\nu}_j^i, \xi_5 = \int 1 \, d\bar{\nu}_j^i, \text{ for some } j \right\}.$$

By Caratheodory's theorem [7, p. 153] there exist collections $\{\alpha_j^i\}_{j=1}^6$, $i = 1, 2, \dots$, of nonnegative coefficients summing to unity with the property: $\sum_{j=1}^6 \alpha_j^i \nu_j^i$ takes the same value as ν^i on $\mathbf{q}, \mathbf{d}, \mathbf{s}, \mathbf{m}$, $1 \in C(\mathbf{A} \times \mathbf{B})$. Here, $\{\nu_j^i\}$ is a reordering of (at most six of the components of) the $\{\bar{\nu}_j^i\}$, for fixed i . Since $\{\alpha_j^i \nu_j^i\} \subset P^{\oplus}(\mathbf{A} \times \mathbf{B})$, $\mathbf{d}, \mathbf{s}, \mathbf{m} \in P(\mathbf{A} \times \mathbf{B})$

$$(9.2) \quad \alpha_j^i \int \mathbf{d} \, d\nu_j^i \rightarrow 0, \quad \alpha_j^i \int \mathbf{s} \, d\nu_j^i \rightarrow 0, \quad \alpha_j^i \int \mathbf{m} \, d\nu_j^i \rightarrow 0$$

as $i \rightarrow \infty$, for $j = 1, 2, \dots, 6$.

By extracting subsequences, we may arrange that

$$(9.3) \quad \begin{aligned} \lim_i \alpha_j^i \int \mathbf{q} \, d\nu_j^i, \quad \lim_i \alpha_j^i |\nu_j^i| \\ \alpha_j = \lim_i \alpha_j^i \text{ exist,} \end{aligned}$$

$j = 1, \dots, 6$. Evidently $\sum_j \alpha_j = 1$, $\alpha_j \geq 0$, each j .

We assume without loss of generality that the nonzero entries occur first in $\{\alpha_1, \dots, \alpha_6\}$. We have

$$\begin{aligned} \eta(P) &= \lim_i \int \mathbf{q} \, d\nu^i = \sum_j \lim_i \alpha_j^i \int \mathbf{q} \, d\nu_j^i \\ &\cong \sum_j \alpha_j \lim_i \int \mathbf{q} \, d\nu_j^i \end{aligned}$$

where the final expression is justified by (9.3) and the summation is taken over j 's corresponding to nonzero α_j 's. Suppose, again without loss of generality,

$$\lim_i \int \mathbf{q} \, d\nu_1^i \leq \lim_i \int \mathbf{q} \, d\nu_j^i$$

for j 's corresponding to nonzero α_j 's. Then

$$(9.4) \quad \begin{aligned} \eta(P) &\cong \sum_j \alpha_j \lim_i \int \mathbf{q} \, d\nu_j^i \cong \left(\sum_j \alpha_j \right) \lim_i \int \mathbf{q} \, d\nu_1^i \\ &= \lim_i \int \mathbf{q} \, d\nu_1^i. \end{aligned}$$

By (9.3), $\{\nu_1^i\}$ is norm bounded and, in view of (9.2),

$$(9.5) \quad \lim_i \int \mathbf{d} \, d\nu_1^i = 0, \quad \lim_i \int \mathbf{s} \, d\nu_1^i = 0, \quad \lim_i \int \mathbf{m} \, d\nu_1^i = 0.$$

By weak* compactness of closed balls in $C^*(\mathbf{A} \times \mathbf{B})$, we may extract a weak* convergent subsequence of $\{\nu_j^i\}$ having some limit $\bar{\nu} \in C^*(\mathbf{A} \times \mathbf{B})$. In view of (9.5), $\bar{\nu}$ takes value zero on $\mathbf{d}, \mathbf{s}, \mathbf{m} \in C(\mathbf{A} \times \mathbf{B})$.

Now recall that each ν_1^i is a polygonal arc from y_0 to the closed set Γ . $\bar{\nu}$ then is a generalized curve as the weak* limit of polygonal arcs. A standard argument, involving examination of particular "exact integrands" establishes that the curve has left endpoint y_0 and right endpoint in Γ . It follows that

$$\int \phi_y \dot{y} \, d\bar{\nu} = \int \phi \, d\beta - \phi(x_0, t_0), \quad \text{all } \phi \in C^1(\mathbf{A}),$$

where β is a concentrated unit measure with support in Γ . We have shown that $\bar{\nu}$ is feasible for the Parametric Problem. It solves the problem in view of (9.4). The proposition is proved.

By Theorem 6.2, we conclude:

COROLLARY 9.1. *There exists a Lipschitz continuous function $y(\cdot) = \{\tilde{x}(\sigma), t(\sigma); 0 \leq \sigma \leq 1\}$ taking values in \mathbf{A} , and a family $\{\nu_\sigma \in P^\oplus(\mathbf{B}); 0 \leq \sigma \leq 1\}$ satisfying properties (i)–(iii) of Theorem 6.2 such that $\nu \in C^*(\mathbf{A} \times \mathbf{B})$ defined by*

$$g \rightarrow \int_0^1 \int g(\tilde{x}(\sigma), t(\sigma), \dot{x}, i) \, d\nu_\sigma(\dot{x}, i) \, d\sigma$$

solves problem (P).

10. Proof of the equivalence theorem. We present first of all a "change of variables" lemma, a slight extension of [14, Lemma 69.1, p. 180].

LEMMA 10.1. *Let be given real-valued functions $t(\sigma), \xi(\sigma), 0 \leq \sigma \leq 1$, with $t(0) = t_0, t(1) = t_1$. Suppose that $t(\cdot)$ is monotone nondecreasing, absolutely continuous and that $\xi(\cdot)$ is Lebesgue measurable, essentially bounded. Write*

$$\Sigma = \{\sigma \in [0, 1] | i(\sigma) \neq 0\} \quad \text{and} \quad T = t(\Sigma).$$

Then the map $\sigma \mapsto t(\sigma)$ carries Lebesgue sets of full measure into Lebesgue sets of full measure. $t(\cdot)$ is one-to-one on Σ , T is Lebesgue measurable with $|T| = t_1 - t_0$ ($|\cdot|$ denotes Lebesgue measure), and defining $\hat{\xi}(t), t_0 \leq t \leq t_1$, as

$$\hat{\xi}(t) = \begin{cases} \xi(\sigma) & \text{when } t \in T, \quad t(\sigma) = t \\ 0 & \text{otherwise} \end{cases}$$

we have that $\hat{\xi}(\cdot)$ is essentially bounded, Lebesgue measurable on $[t_0, t_1]$ and

$$(10.1) \quad \int_{t_0}^{t(\sigma)} \hat{\xi}(t) \, dt = \int_0^{\bar{\sigma}} \xi(\sigma) i(\sigma) \, d\sigma \quad \text{for any } \bar{\sigma} \in [0, 1].$$

The lemma is proved in Appendix B.

The crucial steps are taken in the next lemma of translating the “generalized curve” solution to the Parametric Problem, obtained in § 9, into a solution to the Strong Problem.

LEMMA 10.2. *Let $\{\tilde{x}(\sigma), t(\sigma), \nu_\sigma; 0 < \sigma \leq 1\}$ be a generalized curve solution to the Parametric Problem (Corollary 9.1). Set $t_0 = t(0)$, $t_1 = t(1)$.*

We have that $\{x(t); t_0 \leq t \leq t_1\}$, given by $x(t) = \tilde{x}(t^{-1}(t))$, well-defines an \mathbb{R}^n -valued function and $x(\cdot)$ is Lipschitz continuous. Let the subset T of $[t_0, t_1]$ be defined as

$$T = t(\Sigma), \quad \Sigma = \{\sigma \in [0, 1] | i(\sigma) \neq 0\}.$$

$t(\cdot)$ is one-to-one on Σ , T is Lebesgue measurable and $|T| = t_1 - t_0$.

Define $t \mapsto \tilde{\nu}_t: [t_0, t_1] \rightarrow C^(F)$ by*

$$(10.2) \quad \int g(\dot{x}) d\tilde{\nu}_t = \begin{cases} \frac{1}{i(\sigma)} \int g(\dot{x}/i) i d\nu_\sigma(\dot{x}, i), & t \in T, \quad t = t(\sigma) \\ 0 & \text{otherwise} \end{cases}$$

for $g \in C^(F)$. Then, given $\bar{g} \in C([t_0, t_1] \times F)$*

$$t \mapsto \int \bar{g} d\tilde{\nu}_t$$

is Lebesgue measurable, essentially bounded, and

$$(10.3) \quad \tilde{\nu}_t \in P^n(F) \quad \text{a.e. } t \in [t_0, t_1],$$

$$(10.4) \quad \dot{x}(t) = \int \dot{x} d\nu_t \quad \text{a.e. } t \in [t_0, t_1],$$

$$(10.5) \quad \text{supp } \{\tilde{\nu}_t\} \subset \{\dot{x} | \dot{x} \in f(x(t), t, \Omega)\} \quad \text{a.e. } t \in [t_0, t_1],$$

$$(10.6) \quad \int_{t_0}^{t_1} \int q(x(t), t, \dot{x}) d\tilde{\nu}_t(\dot{x}) dt = \int_0^1 \int q(\tilde{x}(\sigma), t(\sigma), \dot{x}/i) i d\nu_\sigma(\dot{x}, i) d\sigma.$$

Before supplying proof of this result we remark that continuous functions g on the closed ball

$$F = \{\dot{x} \in \mathbb{R}^n | |\dot{x}| \leq K\}$$

(K , bound on $f(\mathbf{A}, \Omega)$) are understood as extended to all of \mathbb{R}^n as

$$g(x) = g\left(\frac{K}{|\dot{x}|} \cdot \dot{x}\right), \quad |\dot{x}| > K.$$

Notice that the extension so defined is continuous on \mathbb{R}^n and bounded by $\max \{g(\dot{x}) | \dot{x} \in F\}$. Continuous functions \bar{g} on $\mathbf{A} \times F$ are understood as extended to $\mathbf{A} \times \mathbb{R}^n$ in similar fashion.

Proof. Observe that

$$(10.7) \quad \begin{aligned} i(\sigma) &\geq 0 & \text{a.e. } \sigma \in [0, 1], \\ |\dot{\tilde{x}}(\sigma)| &\leq K |i(\sigma)| & \text{a.e. } \sigma \in [0, 1] \end{aligned}$$

where as usual K is a bound on $f(\mathbf{A}, \Omega)$. Indeed, arguing as in the proof of Lemma 8.1, we have that

$$\text{supp } \{\nu_\sigma\} \subset \{(\dot{x}, i) \in \mathbf{B} | i \geq 0, \dot{x} = if(\tilde{x}(\sigma), t(\sigma), \Omega)\}, \quad \text{a.e. } \sigma \in [0, 1],$$

whence

$$i(\sigma) = \int i \, d\nu_\sigma \geq 0, \quad \text{a.e. } \sigma \in [0, 1]$$

and

$$\begin{aligned} |\dot{x}(\sigma)| &= \left| \int \dot{x} \, d\nu_\sigma \right| \leq \int |\dot{x}| \, d\nu_\sigma \leq K \int |i| \, d\nu_\sigma \\ &= K \int i \, d\nu_\sigma = K |i(\sigma)|, \quad \text{a.e. } \sigma \in [0, 1]. \end{aligned}$$

$\sigma \mapsto t(\sigma)$ is continuous on $[0, 1]$ and therefore assumes all values in $[t_0, t_1]$. It follows that $t^{-1}(\bar{t})$ is nonempty, each $\bar{t} \in [t_0, t_1]$, so that $x(\cdot)$ is well-defined (as a set-valued function on $[t_0, t_1]$). For $\sigma_2 > \sigma_1$, write $t'' = t(\sigma_2)$, $t' = t(\sigma_1)$. Then by (10.7)

$$\begin{aligned} \tilde{x}(\sigma_2) - \tilde{x}(\sigma_1) &\leq \int_{\sigma_1}^{\sigma_2} |\dot{x}(\sigma)| \, d\sigma \leq K \int_{\sigma_1}^{\sigma_2} |i(\sigma)| \, d\sigma \\ &= K \int_{\sigma_1}^{\sigma_2} i(\sigma) \, d\sigma = K(t'' - t'), \end{aligned}$$

which establishes that $t \mapsto x(t)$ is single-valued and Lipschitz continuous with constant K .

Since the Lipschitz continuous function $t(\sigma)$ is monotone nondecreasing, the asserted properties of the subsets T, Σ follow directly from Lemma 10.1.

Choose $g \in C(F)$. With our extension convention $(\dot{x}, i) \mapsto g(\dot{x}/i)$ is continuous on **B** and bounded by $\max \{ |g(\dot{x})| \mid \dot{x} \in F \}$. It follows that (10.2) well-defines a $C^*(F)$ -valued function ν_t on $[t_0, t_1]$.

Now suppose $\bar{g} \in C([t_0, t_1] \times F)$. $(\sigma, \dot{x}, i) \mapsto \bar{g}(t(\sigma), \dot{x}/i)i$ is continuous on $[t_0, t_1] \times \mathbf{B}$ (again recall the extension convention), so that

$$\sigma \mapsto \int \bar{g}(t(\sigma), \dot{x}/i)i \, d\nu_\sigma(\dot{x}, i)$$

is Lebesgue measurable and essentially bounded.

The function $\xi(\cdot)$ on $[0, 1]$ defined as

$$\xi(\sigma) = \begin{cases} \frac{1}{i(\sigma)} \int \bar{g}(t(\sigma), \dot{x}/i)i \, d\nu_\sigma(\dot{x}, i) & \text{when } i(\sigma) \neq 0, \\ 0 & \text{otherwise} \end{cases}$$

is Lebesgue measurable since $\int \bar{g}i \, d\nu_\sigma$, $i(\sigma)$ are measurable. We see also that on the subset of full Lebesgue measure $I = \{\sigma \in [0, 1] \mid \int i \, d\nu_\sigma = i(\sigma)\}$, $\xi(\sigma)$ is uniformly bounded. We may therefore apply Lemma 10.1 to $\{t(\sigma), \xi(\sigma); 0 \leq \sigma \leq 1\}$. Notice that

$$\xi(\sigma)i(\sigma) = \int \bar{g}(t(\sigma), \dot{x}/i)i \, d\nu_\sigma(\dot{x}, i), \quad \sigma \in I,$$

while

$$\hat{\xi}(t) = \int \bar{g}(t, \dot{x}) \, d\nu_b, \quad t \in T$$

($\hat{\xi}$ as in Lemma 10.1). We have

- (i) $t \mapsto \int \bar{g}(t, \dot{x}) d\tilde{\nu}_t$ is measurable, essentially bounded on $[t_0, t_1]$ and, by (10.1),
 (ii) $\int_0^{\sigma} \int \bar{g}(t(\sigma), \dot{x}/i) i d\nu_{\sigma}(\dot{x}, i) d\sigma = \int_{t_0}^{t(\sigma)} \bar{g}(t, \dot{x}) d\tilde{\nu}_t dt$ for $\bar{\sigma} \in [0, 1]$.

ν_t evidently takes values in $P^{\oplus}(F)$. Also, since $\int i d\nu_{\sigma}(\dot{x}, i) = i(\sigma)$, a.e. $\sigma \in [0, 1]$, and since the map $\sigma \mapsto t(\sigma)$ carries sets of full (Lebesgue) measure into sets of full measure (Lemma 10.1), we have from the definition of ν_t ,

$$\int d\tilde{\nu}_t = \frac{1}{i(\sigma)} \int i d\nu_{\sigma}(\dot{x}, i)|_{\sigma=t^{-1}(t)} = 1 \quad \text{a.e. } t \in [t_0, t_1].$$

This establishes (10.3). Set $\bar{g}(t, \dot{x}) = \dot{x}$; then (ii) gives

$$\begin{aligned} x(\bar{t}) - x(t_0) &= \int_0^{\bar{\sigma}} \dot{x} d\sigma = \int_0^{\bar{\sigma}} \int \dot{x} d\nu_{\sigma}(\dot{x}, i) d\sigma \\ &= \int_{t_0}^{\bar{t}} \int \dot{x} d\tilde{\nu}_t dt \end{aligned}$$

for $\bar{t} = t(\bar{\sigma})$, $\bar{\sigma} \in [0, 1]$. It follows that

$$\dot{x}(t) = \int \dot{x} d\tilde{\nu}_t \quad \text{a.e. } t \in [t_0, t_1].$$

This proves (10.4). Next we set $\bar{g} = \text{dist} \{ \dot{x}, f(\tilde{x}(t), t, \Omega) \}$ in (ii). This gives

$$\begin{aligned} 0 &= \int_0^1 \int \text{dist} \{ \dot{x}/i, f(\tilde{x}(\sigma), t(\sigma), \Omega) \} i d\nu_{\sigma} d\sigma \\ &= \int_{t_0}^{t_1} \int \text{dist} \{ \dot{x}, f(x(t), t, \Omega) \} d\tilde{\nu}_t dt. \end{aligned}$$

Since $t \mapsto \tilde{\nu}_t$ takes values in $P^{\oplus}(F)$,

$$\int \text{dist} \{ \dot{x}, f(x(t), t, \Omega) \} d\tilde{\nu}_t = 0, \quad \text{a.e. } t \in [t_0, t_1]$$

and arguing as in the proof of Lemma 8.1, we conclude

$$\text{supp} \{ \tilde{\nu}_t \} \subset \{ \dot{x} | \dot{x} \in f(x(t), t, \Omega) \}, \quad \text{a.e. } t \in [t_0, t_1].$$

Thus (10.5) is proved. Finally set $\bar{g}(t, \dot{x}) = q(x(t), t, \dot{x})$ in (ii) to obtain

$$\int_0^1 \int q(\tilde{x}(\sigma), t(\sigma), \dot{x}/i) i d\nu_{\sigma}(\dot{x}, i) d\sigma = \int_{t_0}^{t_1} q(x(t), t, \dot{x}) d\tilde{\nu}_t(\dot{x}) dt.$$

This establishes (10.6) and completes the proof.

Proof of Theorem 5.1. We first show that

$$(10.8) \quad \eta(S) \cong \eta(W) \cong \eta(P).$$

Suppose that $\{ \mu_t, x_t; t_0 \leq t \leq t_1 \}$ is an admissible pair. Define $\mu \in C^*(\mathbf{A} \times \Omega)$ by

$$g \xrightarrow{\mu} \int_{t_0}^{t_1} \int g(x(t), t, u) d\mu_t(u) dt, \quad g \in C(\mathbf{A} \times \Omega).$$

We omit the simple steps in checking that μ is feasible for (W) and

$$(10.9) \quad \int l d\mu = \int_{t_0}^{t_1} \int l(x(t), t, u) d\mu_t(u) dt.$$

Thus an element, feasible for (S), defines an element, feasible for (W) and the value of the functional to be minimized is unaltered; the first inequality is proved.

Now let μ be feasible for (W), that is $\mu \in P^\oplus(\mathbf{A} \times \Omega)$; there exists $\beta \in P^n(\Gamma)$ such that

$$\int (\phi_t + \phi_x f) d\mu = \int \phi d\beta - \phi(x_0, t_0), \quad \text{all } \phi \in C^1(\mathbf{A})$$

and

$$\int \text{dist}\{x, Q\} d\mu = 0.$$

Take $f^*(x, t, u) = (x, t, f(x, t, u))$ on $\mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^m$ and set $\tilde{\nu} = \mu \circ f^{*-1}$. Now define $\nu \in C^*(\mathbf{A} \times \mathbf{B})$ as

$$g \mapsto \int g\left(x, t, \frac{\dot{x}}{(\dot{x}^2 + 1)^{1/2}}, \frac{1}{(\dot{x}^2 + 1)^2}\right) (\dot{x}^2 + 1)^{1/2} d\tilde{\nu}$$

for $g \in C(\mathbf{A} \times \mathbf{B})$. Let us verify that ν is feasible for the Parametric Problem. For $\phi \in C^1(\mathbf{A})$

$$\begin{aligned} \int (\phi_t i + \phi_x \dot{x}) d\nu &= \int (\phi_t + \phi_x \dot{x}) d\tilde{\nu} = \int (\phi_t + \phi_x f) d\mu \\ &= \int \phi d\beta - \phi(x_0, t_0). \end{aligned}$$

Furthermore

$$\begin{aligned} \int \text{dist}\{\dot{x}/i, f(x, t, \Omega)\} i d\nu &= \int \text{dist}\{\dot{x}, f(x, t, \Omega)\} d\tilde{\nu} \\ &= \int \text{dist}\{f(x, t, u), f(x, t, \Omega)\} d\mu = 0, \\ \int \text{dist}\{x, Q\} i d\nu &= \int \text{dist}\{x, Q\} d\tilde{\nu} = \int \text{dist}\{x, Q\} d\mu = 0, \\ \int \max\{-i, 0\} d\nu &= \int \max\{-1, 0\} d\tilde{\nu} = 0. \end{aligned}$$

This proves feasibility. We note also, in view of the manner in which q is defined

$$\begin{aligned} \int q(x, t, \dot{x}/i) i d\nu &= \int q(x, t, f(x, t, u)) d\mu \\ &\equiv \int l(x, t, u) d\mu. \end{aligned}$$

Thus corresponding to an element feasible for (W) is an element feasible for (P) having not larger cost. We conclude the second inequality in (10.8).

Now consider the pair $\{x(t), \tilde{\nu}_t(t); t_0 \leq t \leq t_1\}$ introduced in Lemma 10.2. Define $p(t), t_0 \leq t \leq t_1$, as

$$p(t) = \int_{t_0}^t \int q(x(\tau), \tau, \dot{x}) d\tilde{\nu}_\tau d\tau.$$

Then $\begin{bmatrix} x \\ p \end{bmatrix}(\cdot)$ is an absolutely continuous \mathbb{R}^{n+1} -valued function with

$$(10.10) \quad \frac{d}{dt} \begin{bmatrix} x \\ p \end{bmatrix}(t) = \int \begin{bmatrix} \dot{x} \\ q(x(t), t, \dot{x}) \end{bmatrix} d\tilde{\nu}_t(\dot{x}) \quad \text{a.e. } t \in [t_0, t_1].$$

Select t such that (10.10) holds and furthermore

$$\text{supp } \{\tilde{\nu}_t\} \subset \{x | \dot{x} \in f(x(t), t, \Omega)\}, \quad \tilde{\nu}_t \in P^n(\Gamma)$$

(in view of Lemma 10.2 such t 's comprise a set of full measure). Suppose for the moment that $\tilde{\nu}_t$ has finite support (this assumption will be removed). Then

$$\frac{d}{dt} \begin{bmatrix} x \\ p \end{bmatrix}(t) = \sum_i \alpha_i \begin{bmatrix} f(x(t), t, u_i) \\ q(x(t), t, f(x(t), t, u_i)) \end{bmatrix}$$

for some simplex $\{u_1, \dots, u_k; \alpha_1, \dots, \alpha_k\}$ in Ω

$$= \sum_i \alpha_i \begin{bmatrix} f(x(t), t, \bar{u}_i) \\ q(x(t), t, f(x(t), t, \bar{u}_i)) \end{bmatrix}$$

(with $\bar{u}_i \in \{u \in \Omega | l(x(t), t, u) = \min \{l(x(t), t, v) | f(x(t), t, v) = f(x(t), t, u_i)\}\}$)

$$= \sum_i \alpha_i \begin{bmatrix} f(x(t), t, \bar{u}_i) \\ l(x(t), t, \bar{u}_i) \end{bmatrix}, \quad \text{by definition of } q,$$

$$\in \text{co} \left\{ \begin{bmatrix} f \\ l \end{bmatrix}(x(t), t, \Omega) \right\}$$

(co denotes convex hull). Now, in view of the weak* density of unit measures with finite support in $P^n(F)$ and the closedness of $\text{co} \left\{ \begin{bmatrix} f \\ l \end{bmatrix}(x(t), t, \Omega) \right\}$,

$$\frac{d}{dt} \begin{bmatrix} x \\ p \end{bmatrix}(t) \in \text{co} \left\{ \begin{bmatrix} f \\ l \end{bmatrix}(x(t), t, \Omega) \right\}$$

without the restriction that $\tilde{\nu}_t$ have finite support. By a well-known selection theorem [14, p. 297], there exists a relaxed control $\{\mu_t; t_0 \leq t \leq t_1\}$ with

$$\left. \begin{aligned} \dot{x}(t) &= \int f(x(t), t, u) d\mu_t(u) \\ \dot{p}(t) &= \int l(x(t), t, u) d\mu_t(u) \end{aligned} \right\} \quad \text{a.e. } t \in [t_0, t_1]$$

and, taking note of (10.6), we have

$$(10.11) \quad \begin{aligned} \int_{t_0}^{t_1} \int q(\tilde{x}(\sigma), t(\sigma), \dot{x}/t) d\nu_\sigma &= \int_{t_0}^{t_1} \int q(x(t), t, x) d\tilde{\nu}_t(\dot{x}) dt \\ &= \int_{t_0}^{t_1} \dot{p}(t) dt = \int_{t_0}^{t_1} \int l(x(t), t, u) d\mu_t(u) dt. \end{aligned}$$

$x(\cdot)$ takes values in Q and has endpoints in $\{(x_0, t_0)\} \times \Gamma$. It follows that $\{\mu_t, x(t); t_0 \leq t \leq t_1\}$ is a feasible pair for the Strong Problem.

But the first integral in (10.11) will be recognized as the value of the Parametric Problem; since the last integral is the cost of the Strong Problem for this feasible pair,

$$\eta(S) \leq \eta(P).$$

But then (10.8) implies

$$(10.12) \quad \eta(S) = \eta(W) = \eta(P).$$

Taking $\{\mu_t, x(t); t_0 \leq t \leq t_1\}$ a solution to (S), in view of (10.9) and (10.12), we have that μ defined by (5.1) solves (W). The theorem is proved.

Comment. Suppose

$$\left\{ \begin{pmatrix} l(x, t, u) \\ f(x, t, u) \end{pmatrix} \in \mathbb{R}^{n+1} \mid u \in \Omega \right\}$$

is convex for each $(x, t) \in Q \times [t_0, T]$. Then $\{\mu_t; t_0 \leq t \leq t_1\}$ in the final sections of the proof of Theorem 5.1 may be chosen an ordinary control [14, p. 297]. We conclude that (under this extra assumption) the Weak Problem has a solution having representation through an ordinary control, trajectory pair, and that Theorem 5.1 still applies with the Strong Problem posed over ordinary controls.

Appendix A. Proof of Theorem 6.1. Theorem 6.1 is a refinement of [14, Thm. 86.1, p. 201] and its proof requires introducing terminology and definitions from [14, p. 194ff.] used in rather informal fashion at the beginning of § 6.

As in § 6 we take \mathbf{A} to be a cube in \mathbb{R}^{n+1} , \mathbf{B} the surface of the unit ball in \mathbb{R}^{n+1} and Γ a compact subset of \mathbf{A} . In the present section it will be convenient to diverge from previous notation and write the operation of $\nu \in C^*(\mathbf{A} \times \mathbf{B})$ on $g \in C(\mathbf{A} \times \mathbf{B})$ as (ν, g) (rather than $\int g d\nu$). We will need to introduce a number of norms on $C^*(\mathbf{A} \times \mathbf{B})$ and distinguish them in the notation. $|\cdot|$ is the usual “dual” norm. $|\cdot|'$ and $\|\cdot\|$ will denote respectively the “dashed” and “consistent” norms (introduced below).

Given $p \in C^*(\mathbf{A} \times \mathbf{B})$, the restriction of p to elements of the form $\phi_y \dot{y}$ ($\phi \in C^1(\mathbf{A})$), so-called exact integrands, is the *boundary* of p , and is written ∂p .

Elements in $P^\oplus(\mathbf{A} \times \mathbf{B})$ are termed *flows*. A *polygonal flow* is a flow with representation

$$(A.1) \quad g \mapsto \sum_i \alpha_i \int_0^1 g(y_i(\sigma), \dot{y}_i(\sigma)/|\dot{y}_i(\sigma)|) |\dot{y}_i(\sigma)| d\sigma,$$

with the α_i 's positive, and the $\{y_i(\sigma); 0 \leq \sigma \leq 1\}$ a finite collection of directed line segments. If the α_i 's take value one and $y_i(1) = y_{i+1}(0)$, $i = 1, 2, \dots$, then the polygonal flow is a *polygonal arc*. If additionally the y_i 's close, then the polygonal flow is a *closed polygonal arc*. Boundaries of polygonal flows are called *simplicial*.

For b the boundary of an element in $C^*(\mathbf{A} \times \mathbf{B})$ we write

$$(A.2) \quad |b| = \inf \{|p|(\text{dual norm}) \mid p \in P^\oplus(\mathbf{A} \times \mathbf{B}), b = \partial p\}$$

(any boundary of an element in $C^*(\mathbf{A} \times \mathbf{B})$ is also the boundary of a flow, permitting us to take the infimum over P^\oplus [14, p. 171]). This defines a norm on the linear space of boundaries [14, p. 200]. The norm of a simplicial boundary may be taken as given by (A.2) where now the infimum is taken over polygonal flows [14, p. 200].

For $p \in C^*(\mathbf{A} \times \mathbf{B})$, the *consistent norm* $\|\cdot\|$ is now introduced as

$$\|p\| = \max \{|p|', |\partial p|\}.$$

Here $|\cdot|'$, the *dashed* norm, is a norm on $C^*(\mathbf{A} \times \mathbf{B})$ consistent with weak* convergence [14, p. 114ff.].

The completion of the class of polygonal flows with respect to the consistent norm is termed the class of *consistent flows*. *Consistent boundaries* are boundaries of consistent flows.

LEMMA A.1. Suppose that $\nu \in P^\oplus(\mathbf{A} \times \mathbf{B})$ has boundary

$$\partial\nu = \left\{ \phi_y \dot{y} \mapsto \int_{\Gamma} \phi \, d\beta - \phi(y_0) \right\}$$

for some $\beta \in P^n(\Gamma)$, $y_0 \in A$. Then ν is a consistent flow. Furthermore, there exists a sequence of polygonal flows $\{p_i\}$ with

$$\partial p_i \rightarrow \partial\nu \quad (\text{with respect to boundary norm})$$

where each p_i is a unit mixture (i.e., a positive linear combination with coefficients summing to one) of line segments from y_0 to Γ .

Proof. By a standard construction we obtain a sequence of partitions into Borel sets of the compact set Γ ,

$$\Gamma = \Gamma_{i1} \cup \cdots \cup \Gamma_{iN_i}$$

having the property

$$\max_j |\Gamma_{ij}| \rightarrow 0 \quad \text{as } i \rightarrow \infty$$

($|\Gamma_{ij}|$ denotes diameter of Γ_{ij}).

Pick $y_{ij} \in \Gamma_{ij}$ and set

$$\alpha_{ij} = \int_{\Gamma_{ij}} d\beta.$$

By a well-known result

$$\sum_j \alpha_{ij} \delta(y_{ij}) \rightarrow \beta$$

weakly* in $C^*(\Gamma)$.

We introduce the notation $s(y, \bar{y})$ for the flow defined through the line segment from y to \bar{y} in \mathbf{A} . It is readily verified that, for fixed $g \in C(\mathbf{A} \times \mathbf{B})$,

$$y \mapsto \chi_g(y)$$

is a continuous functional. Here $\chi_g(y) = (s(y_0, y), g)$.

Now define the flow p_0 by

$$p_0 = \int s(y_0, y) \, d\beta(y),$$

which for our purposes is taken to be the linear functional

$$g \mapsto \int_{\Gamma} \chi_g(y) \, d\beta(y).$$

The functional is evidently positive and bounded, so defines a flow.

We see immediately that p_0 has boundary

$$\partial p_0 = \left\{ \phi_y \dot{y} \mapsto \int \phi \, d\beta - \phi(y_0) \right\}.$$

In other words, with ν as in the lemma statement, p_0 and ν have the same boundary.

Now consider the sequence $\{p_i\}$ in $P^\oplus(\mathbf{A} \times \mathbf{B})$ defined by

$$p_i = \sum_j \alpha_{ij} s(y_0, y_{ij}).$$

Clearly each p_i is a unit mixture of segments.

For $g \in C(\mathbf{A} \times \mathbf{B})$,

$$\begin{aligned}(p_i, g) &= \sum_j \alpha_{ij} \chi_g(y_{ij}) = \int_{\Gamma} \chi_g(y) d\left(\sum_j \alpha_{ij} \delta(y_{ij})\right) \\ &\rightarrow \int_{\Gamma} \chi_g(y) d\beta = (p_0, g)\end{aligned}$$

as $i \rightarrow \infty$. We have shown that

$$(A.3) \quad p_i \rightarrow p_0 \text{ (weakly*)}.$$

We now show that $\partial p_0 - \partial p_i \rightarrow 0$ (with respect to the boundary norm). For the exact integrand $\phi_y \dot{y}$,

$$\begin{aligned}(p_0 - p_i, \phi_y \dot{y}) &= \int_{\Gamma} (\phi(y) - \phi(y_0)) d\beta(y) - \left(\sum_j \alpha_{ij} \phi(y_{ij}) - \phi(y_0)\right) \\ &= \sum_j \int_{\Gamma_{ij}} (\phi(y) - \phi(y_{ij})) d\beta(y) = (\hat{p}_i, \phi_y \dot{y})\end{aligned}$$

where \hat{p}_i is the flow defined by

$$\sum_j \int_{\Gamma_{ij}} s(y_{ij}, y) d\beta(y), \quad i = 1, 2, \dots$$

(we justify this is a flow as in the definition of p_0). Let us examine the dual norm of \hat{p}_i :

$$\begin{aligned}|\hat{p}_i| &= (\hat{p}_i, 1) = \sum_j \int_{\Gamma_{ij}} |y_{ij} - y| d\beta(y) \\ &\leq \max_j |\Gamma_{ij}| \sum_j \alpha_{ij} = \max_j |\Gamma_{ij}| \rightarrow 0 \quad \text{as } i \rightarrow \infty.\end{aligned}$$

A fortiori

$$(A.4) \quad \partial p_0 - \partial p_i \rightarrow 0 \quad \text{(with respect to the boundary norm)}.$$

We conclude from (A.3), (A.4) that p_0 is the consistent limit of a sequence of polygonal flows and, as such, is a consistent flow. But then ∂p_0 , and hence $\partial \nu$, is a consistent boundary. We conclude that ν is a consistent flow (flows with consistent boundaries are consistent flows [14, Thm. 86.1, p. 201]). Along the way, we have constructed a sequence of polygonal flows $\{p_i\}$ with the properties required to complete the proof of the lemma.

Proof of Theorem 6.1. By definition, polygonal flows are dense in the class of consistent flows (with respect to the consistent norm). We have shown (Lemma A.1) that ν is consistent, and accordingly there exists a sequence $\{q_i\}$ of polygonal flows such that

$$q_i \rightarrow \nu \quad \text{(with respect to the consistent norm)}.$$

The theorem will have been proved if we can exhibit a sequence of polygonal flows $\{\nu_i\}$ such that

$$(A.5) \quad \nu_i \rightarrow \nu \quad \text{(with respect to the consistent norm)}$$

$$(A.6) \quad \partial \nu_i = \partial p_i, \quad i = 1, 2, \dots$$

For then in particular ν_i will converge to ν (weakly*), while (A.6) ensures that each ν_i has the properties asserted in the theorem. (Here the p_i are as in Lemma A.1.)

Set

$$\rho_i = \partial p_i - \partial q_i, \quad i = 1, 2, \dots$$

$\{\rho_i\}$ is a sequence of simplicial boundaries. In view of the definition of norm on the class of boundaries and foregoing remarks, we may choose polygonal flows r_i , $i = 1, 2, \dots$, such that

$$\partial r_i = \rho_i, \quad |r_i| \leq 2|\rho_i|,$$

whence, as the dashed norm is weaker than the dual norm,

$$\|r_i\| \leq 2|\rho_i|.$$

But $|\rho_i| = |\partial p_i - \partial q_i| \rightarrow 0$, since $\{\partial p_i\}$, $\{\partial q_i\}$ have a common limit. So, taking

$$\nu_i = q_i + r_i$$

we have

$$\partial \nu_i = \partial q_i + (\partial p_i - \partial q_i) = \partial p_i.$$

We have verified (A.6). Finally

$$\lim_i \nu_i = \lim_i q_i = \nu \quad (\text{limits with respect to the consistent norm})$$

since $\|r_i\| \rightarrow 0$. This establishes (A.5) and concludes the proof.

Appendix B. Here we prove a number of technical results needed in the main development. (Recall y (\dot{y}) denotes (x, t) ((\dot{x}, \dot{t})).

Proof of Lemma 7.1. Recalling that f is continuous and Ω is compact, we see that $D = \{(x, t, \dot{x}) \in \mathbb{R}^{2n+1} | \dot{x} \in f(x, t, \Omega)\}$ is compact. It suffices to show therefore that the function q is continuous on D . For then the continuous function q is bounded on the closed set D , and may be continuously extended to all of \mathbb{R}^{2n+1} in such a way that it remains uniformly bounded [4, p. 242].

Let $\{(x_i, t_i, \dot{x}_i)\}$ be an arbitrary sequence in D converging to $(\bar{x}, \bar{t}, \bar{\dot{x}})$. Let $\{u_i\}$ be a corresponding sequence in Ω such that

$$u_i \in \arg \min \{l(x_i, t_i, u) | \dot{x}_i = f(x_i, t_i, u), u \in \Omega\}$$

and extract a subsequence such that $u_i \rightarrow \bar{u}$. By the continuity of f then

$$\bar{\dot{x}} = (f(\bar{x}, \bar{t}, \bar{u})).$$

In view of the definition of q and Condition 5.1, it follows that for the subsequence

$$\begin{aligned} q(x_i, t_i, \dot{x}_i) &= \min_v \{l(x_i, t_i, v) | f(x_i, t_i, u_i) = f(x_i, t_i, v)\} \\ &\rightarrow \min_v \{l(\bar{x}, \bar{t}, v) | f(\bar{x}, \bar{t}, \bar{u}) = f(\bar{x}, \bar{t}, v)\} \\ &= \min_v \{l(\bar{x}, \bar{t}, v) | \bar{\dot{x}} = f(\bar{x}, \bar{t}, v)\} = q(\bar{x}, \bar{t}, \bar{\dot{x}}). \end{aligned}$$

But the original sequence was arbitrary; it follows that

$$q(x_i, t_i, \dot{x}) \rightarrow q(\bar{x}, \bar{t}, \bar{\dot{x}})$$

for the *original* sequence. The lemma is proved.

LEMMA B.1. *The functions $(y, \dot{y}) \mapsto q(y, \dot{x}/i)|i|$, $(y, \dot{y}) \mapsto \text{dist}\{\dot{x}/i, f(y, \Omega)\}|i|$ mapping $\mathbf{A} \times \mathbf{B} \rightarrow \mathbb{R}$ introduced in § 7 are continuous.*

Proof. Continuity of the first function is almost immediate; it is continuous off $A \times \{(\dot{x}, i) | i = 0\}$ by continuity of q , yet if $\{y_i, \dot{y}_i\}$ is a sequence in $\mathbf{A} \times \mathbf{B}$ with $t_i \rightarrow 0$, $q(y_i, \dot{x}_i/i)|i| \rightarrow 0$ (by uniform boundedness of q). Thus the function is continuous on $A \times \{(\dot{x}, i) | i = 0\}$ also, since it takes the value zero on this set.

Consider now the second function:

(a) We show that $(p, y) \mapsto \text{dist}\{p, f(y, \Omega)\}$ is continuous on $\mathbb{R}^n \times \mathbb{R}^{n+1}$. Indeed let $\{(p_i, y_i)\} \rightarrow (p, y)$. By compactness of $f(\bar{y}, \Omega)$, each $\bar{y} \in \mathbb{R}^{n+1}$, we may choose u_i 's in Ω such that

$$\text{dist}\{p_i, f(y_i, \Omega)\} = |p_i - f(y_i, u_i)|.$$

Extract a subsequence $\{(p_i, y_i, u_i)\}$ converging to (p, y, \bar{u}) for some $\bar{u} \in \Omega$. Let $u_0 \in \Omega$ be such that

$$\text{dist}\{p, f(y, \Omega)\} = |p - f(y, u_0)|.$$

By definition of the distance function

$$(B.1) \quad |p - f(y, u_0)| \leq |p - f(y, \bar{u})|$$

and

$$\text{dist}\{p_i, f(y_i, \Omega)\} = |p_i - f(y_i, u_i)| \leq |p_i - f(y_i, u_0)|, \quad i = 1, 2, \dots$$

Taking the limit for the subsequence (which is permissible since f is continuous), we obtain

$$\lim \text{dist}\{p_i, f(y_i, \Omega)\} = |p - f(y, \bar{u})| \leq |p - f(y, u_0)| = \text{dist}\{p, f(y, \Omega)\}.$$

By (B.1) then

$$\text{dist}\{p_i, f(y_i, \Omega)\} \rightarrow \text{dist}\{p, f(y, \Omega)\}.$$

But the limit is independent of the particular subsequence; it follows that the original sequence was convergent, and the function is continuous as stated.

(b) We conclude the proof. By (a), $(y, \dot{y}) \mapsto \text{dist}\{\dot{x}/t, f(y, \Omega)\}|i|$ is continuous off $A \times \{(\dot{x}, i) | i = 0\}$. To complete the proof, suppose that $\{(y_i, \dot{y}_i)\} \rightarrow (y, (\dot{x}, 0))$. Choose the u_i 's in Ω , arbitrarily if $i = 0$, and otherwise such that

$$|(\dot{x}_i/i) - f(y_i, u_i)| \cdot |i| = \text{dist}\{(\dot{x}_i/i), f(y_i, \Omega)\}|i|.$$

We have for each i (with $\text{sgn}\{t\} = 1$ ($= -1$) when $i \geq 0$ ($i < 0$))

$$|(\dot{x}_i/i) - f(y_i, u_i)| \cdot |i| = |\text{sgn}\{i\} \dot{x}_i - f(y_i, u_i)| |i|.$$

It follows that

$$|\dot{x}| - K|i| \leq \text{dist}\{\dot{x}_i/i, f(y_i, \Omega)\}|i| \leq |\dot{x}_i| + K|i|$$

where K bounds $f(\mathbf{A}, \Omega)$. Clearly

$$\text{dist}\{(\dot{x}_i/i), f(y_i, \Omega)\}|i| \rightarrow |\dot{x}|.$$

Since $|\dot{x}| = 1$ (recall defining relation for elements in \mathbf{B}), the lemma is proved.

Proof of Lemma 8.1. We show that ν has support in $\{(y, \dot{y}) | \mathbf{m}(y, \dot{y}) = 0\}$ or equivalently in $\{(y, \dot{y}) | \mathbf{m}(y, \dot{y}) \leq 0\}$ since $\mathbf{m} \in P(\mathbf{A} \times \mathbf{B})$. Write R for the complement of this second set in $\mathbf{A} \times \mathbf{B}$ and define $R_\varepsilon = \{(y, \dot{y}) | \mathbf{m}(y, \dot{y}) > \varepsilon\}$. We have $R = \bigcup_{\varepsilon > 0} R_\varepsilon$, ε rational. It suffices, in view of the sigma-additivity of ν , to show that for any $\xi \in C(\mathbf{A} \times \mathbf{B})$, any $\varepsilon > 0$, we have $\int_{R_\varepsilon} \xi d\nu = 0$. Suppose this were not the case. Then, with a possible sign change, there exists some $\xi \in C(\mathbf{A} \times \mathbf{B})$ with $\int_{R_\varepsilon} \xi d\nu > 0$. However, choosing k sufficiently large

$$\xi(y, \dot{y}) < k|\varepsilon| \quad \text{for all } (y, \dot{y}) \in \mathbf{A} \times \mathbf{B}.$$

Recalling that $\nu \in P^\oplus(\mathbf{A} \times \mathbf{B})$, $\mathbf{m} \in P(\mathbf{A} \times \mathbf{B})$, we have

$$0 < \int_{R_\varepsilon} \xi d\nu \leq k \int_{R_\varepsilon} \varepsilon d\nu \leq k \int_{R_\varepsilon} \mathbf{m} d\nu \leq k \int_{\mathbf{A} \times \mathbf{B}} \mathbf{m} d\nu = 0.$$

This contradiction proves the assertion above.

Substituting \mathbf{d} for \mathbf{m} throughout the foregoing, we have also proved that ν has support in $\{(y, \dot{y}) | \mathbf{d}(y, \dot{y}) = 0\}$. Combining the two results, we conclude ν has support in $\{(y, \dot{y}) | \text{dist}\{(\dot{x}/i), f(y, \Omega)\} |i| = 0, i \geq 0\}$. But this set may be written $\{(y, \dot{y}) | \dot{x} \in if(y, \Omega), i \geq 0\}$, and the lemma is proved.

Proof of Lemma 9.1. Since ν has the stated representation, $\int i d\nu = 0$. This means

$$\int_{i \geq 0} i d\nu - \int_{i < 0} (-i) d\nu = 0.$$

However

$$\int_{i \geq 0} i d\nu + \int_{i < 0} (-i) d\nu = \int |i| d\nu$$

whence

$$(B.2) \quad \int \max\{-i, 0\} d\nu = \frac{1}{2} \int |i| d\nu.$$

Also

$$\begin{aligned} \int \text{dist}\{(\dot{x}/i), f(x, t, \Omega)\} |i| d\nu &\geq \int \text{dist}\{(\dot{x}/i), \{\dot{x} | |\dot{x}| \leq K\}\} |i| d\nu \\ &\geq \int |\dot{x}| d\nu - K \int |i| d\nu. \end{aligned}$$

From (B.2), then,

$$\begin{aligned} \int \text{dist}\{(\dot{x}/i), f(x, t, \Omega)\} |i| d\nu + 2(K+1) \int \max\{-i, 0\} d\nu \\ \geq \int |\dot{x}| d\nu + \int |i| d\nu \geq \int (|\dot{x}|^2 + |i|^2)^{1/2} d\nu = |\nu|. \end{aligned}$$

Proof of Lemma 10.1. Let us first prove that, for any Lebesgue measurable set $\mathcal{S} \subset [0, 1]$, $t(\mathcal{S})$ is Lebesgue measurable and

$$(B.3) \quad |t(\mathcal{S})| = \int_{\mathcal{S}} i(\sigma) d\sigma.$$

By the monotonicity and absolute continuity of $t(\sigma)$, (B.3) holds for all finite unions of intervals and therefore, by a limiting argument, for any Borel set. Now take \mathcal{S} an

arbitrary Lebesgue subset. Since Lebesgue measure is regular, there exist Borel sets $\mathcal{S}_i, \mathcal{S}_0$ such that $\mathcal{S}_i \subset \mathcal{S} \subset \mathcal{S}_0$ and $|\mathcal{S}_0 \setminus \mathcal{S}_i| = 0$.

But by (B.3)

$$|t(\mathcal{S}_0)| = \int_{\mathcal{S}_0} i(\sigma) d\sigma = \int_{\mathcal{S}_i} i(\sigma) d\sigma = |t(\mathcal{S}_i)|,$$

which implies that $t(\mathcal{S}_0) \setminus t(\mathcal{S}_i)$ is a null set. It follows that $t(\mathcal{S}) \Delta t(\mathcal{S}_i)$ is a null set as a subset of $t(\mathcal{S}_0) \setminus t(\mathcal{S}_i)$. Thus $t(\mathcal{S})$ differs from a Borel set by a null set and is therefore Lebesgue measurable. (B.3) then holds for \mathcal{S} any Lebesgue measurable set by additivity.

It is clear from (B.3) that $t(\cdot)$ carries sets of full measure into sets of full measure. Also

$$t_1 - t_0 = \int_{[0,1]} i(\sigma) d\sigma = \int_{\Sigma} i(\sigma) d\sigma = |t(\Sigma)| = |T|$$

which establishes that T is of full measure. $t(\sigma)$ is obviously one-to-one on Σ .

To test measurability of $\hat{\xi}(\cdot)$, since T is of full measure, it suffices to show that, for K any interval, $T \cap \hat{\xi}^{-1}(K)$ is Lebesgue measurable. But

$$T \cap \hat{\xi}^{-1}(K) = t(\Sigma \cap \xi^{-1}(K))$$

and this is Lebesgue measurable since $\xi^{-1}(K)$ is Lebesgue measurable under the assumptions and the property that $t(\cdot)$ carries Lebesgue sets into Lebesgue sets. The subset of $[t_0, t_1]$ on which $|\hat{\xi}(t)| > \text{ess sup } \{|\xi|\}$ is a subset of $t(\sigma) \mid |\xi(\sigma)| > \text{ess sup } \{|\xi|\}$, which is a null set by (B.3). It follows that $\hat{\xi}(\cdot)$ is essentially bounded.

By a well-known change of variables lemma [5, p. 155] applied to the integrable function $\hat{\xi}(\cdot)$ we have

$$\int_{t_0}^{t_1} \hat{\xi}(t) dt = \int_0^1 \hat{\xi}(t(\sigma)) i(\sigma) d\sigma = \int_{\{\sigma \mid i(\sigma) \neq 0\}} \xi(\sigma) i(\sigma) d\sigma = \int_0^1 \xi(\sigma) i(\sigma) d\sigma.$$

This proves the final assertion of the lemma.

REFERENCES

- [1] V. G. BOLTYANSKII, *Sufficient conditions for optimality and the justification of the dynamic programming method*, this Journal, 4 (1966), pp. 326–361.
- [2] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators Part I*, Interscience, New York, 1957.
- [3] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.
- [4] J. L. KELLEY, *General Topology*, Van Nostrand, Princeton, NJ, 1955.
- [5] J. F. C. KINGMAN AND S. J. TAYLOR, *Introduction to Measure and Probability*, Cambridge University Press, Cambridge, England, 1966.
- [6] R. M. LEWIS, *Problem equivalence and necessary conditions of dynamic programming type in optimal control*, Ph.D. thesis, Imperial College, University of London.
- [7] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Mathematical Series no. 28, Princeton University Press, Princeton, NJ, 1970.
- [8] J. E. RUBIO, *Extremal points and optimal control theory*, Annali di Matema, to appear.
- [9] ———, *An existence theorem for control problems in Hilbert spaces*, Bull. London Math. Soc., to appear.
- [10] E. J. MCSHANE, *Relaxed controls and variational problems*, this Journal, 5 (1967), pp. 438–485.
- [11] A. E. TAYLOR, *Introduction to Functional Analysis*, McGraw-Hill, New York, 1958.

- [12] R. B. VINTER AND R. M. LEWIS, *Necessary and sufficient conditions for optimality of dynamic programming type, making no a priori assumptions on the controls*, this Journal, to appear.
- [13] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
- [14] L. C. YOUNG, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.

A NECESSARY AND SUFFICIENT CONDITION FOR OPTIMALITY OF DYNAMIC PROGRAMMING TYPE, MAKING NO A PRIORI ASSUMPTIONS ON THE CONTROLS*

RICHARD B. VINTER[†] AND RICHARD M. LEWIS[‡]

Abstract. A well-known sufficient condition for optimality in control theory, given in terms of a solution to the Bellman partial differential equation, is considered. It is shown that if the equation is relaxed to inequality, and if the resulting inequality is required to be satisfied only in a limiting sense, then the condition becomes also *necessary* for optimality. In contrast to previous results of a similar nature, this is accomplished without making regularity assumptions about optimal feedback controls. The results here are obtained through application of convex analysis to the weak version of the control problem studied in a companion paper.

1. Introduction. In this, the second of two companion papers, we build on our earlier results [10] to obtain a new necessary and sufficient condition for optimality of relaxed control trajectory pairs for a wide class of control problems.

Our condition modifies a well-known sufficient condition for optimality, given in terms of a continuously differentiable solution, ϕ , to the Bellman partial differential equation (see e.g. [9, p. 192]). Other authors have supplied refinements of the sufficient condition, with a view to enlarging the class of control problems whose solutions may be characterized and to obtain necessary conditions of optimality ([1], and in a slightly different setting [3]). These refinements require continuous differentiability of ϕ only in certain regions. These regions are required to satisfy certain complicated conditions. In the case that the problem admits a feedback solution, and that the optimal feedback function has certain regularity properties, the refined conditions become necessary for optimality. Thus the condition is refined to the point of becoming necessary for optimality at the cost of greater complexity and a severe a priori assumption on optimal controls.

Our modification differs from those just described by characterizing optimality through a *sequence* of continuously differentiable functions. It is perhaps remarkable that our condition is both simply stated and, while making no a priori assumptions about optimal controls (of a synthesizability nature or otherwise), is necessary as well as sufficient for optimality. We permit general relaxed controls, and consequently waive convexity assumptions. The cost function and dynamics are required merely to be continuous.

The main result of the companion paper [10] was to establish equivalence between a control problem and a related “weak” problem. The weak problem is a *convex* mathematical programming problem. The new optimality condition is obtained by setting up the Fenchel dual of this weak problem, and establishing that the values of the two problems coincide. In the use of convex analysis, the methodology here is strongly reminiscent of L. C. Young’s generalization of “Huygen’s variational algorithm” in the specialized setting of parametric problems in the calculus of variations; indeed our findings provide something of a control counterpart of [11, Thm. 86.3, p. 202].

As is customary when duality ideas are brought to bear, we recover as subsidiary results lower bounds on the value of the optimization problem as well as fresh

* Received by the editors February 28, 1977, and in revised form August 12, 1977.

[†] Department of Computing and Control, Imperial College of Science and Technology, London SW7 2BZ, England.

[‡] Department of Computing and Control, Imperial College of Science and Technology, London, England. Now at School of Mathematics, University of Bath, Bath BA2 7AY, England.

interpretations of earlier results. The earlier sufficient condition will be seen to correspond to existence of a solution to a dual problem in a certain subset of the feasible region.

We point out that characterization of optimal controls resembling that given here to the extent that it is given through a sequence of continuously differentiable ϕ_i 's underlies results in [2], [4], though with reference to a different class of control problems. These results differ in spirit from our own in that the ϕ_i 's are smooth solutions of a family of perturbed Bellman equations, rather than a single partial differential inequality as here.

A familiarity with the technical details of [10] is not presupposed. The main equivalence theorem of [10] is quoted; otherwise the present paper is largely self-contained.

2. Preliminaries. We adhere to the notation and terminology of [10]. $C(S)$ is the Banach space of continuous (real-valued) functions on S with the sup norm (for S a compact space). In the case that $S \subset \mathbb{R}^k$, $C^1(S)$ is the subset of $C(S)$ comprising restrictions of continuously differentiable functions on \mathbb{R}^k to S . $C^*(S)$ denotes the (normed) dual of $C(S)$. We do not distinguish between elements μ in $C^*(S)$ and the (signed) Radon measures that represent them, writing

$$g \mapsto \int g d\mu$$

for the action of μ on $g \in C(S)$. $P(S)$ is the usual closed positive cone in $C(S)$, and $P^\oplus(S)$ is its positive polar cone in $C^*(S)$. $P^n(S)$ denotes points in $P^\oplus(S)$ of unit norm. $|\cdot|$ denotes norm: the Euclidean norm in the case of \mathbb{R}^k , and the dual norm in the case of $C^*(S)$. Let be given

$$\begin{aligned} Q &\subset \mathbb{R}^n, & \Omega &\subset \mathbb{R}^m; \\ [t_0, T] &\subset \mathbb{R} & \text{with } T > t_0; \\ \Gamma &\subset Q \times [t_0, T], & x_0 &\in \mathbb{R}^n; \\ (x, t, u) &\mapsto l(x, t, u): \mathbb{R}^{n+1+m} \rightarrow \mathbb{R}; \\ (x, t, u) &\mapsto f(x, t, u): \mathbb{R}^{n+1+m} \rightarrow \mathbb{R}^n. \end{aligned}$$

We assume:

- (i) Q, Ω are compact; l, f are continuous.
- (ii) \mathbf{A} denotes a cube in \mathbb{R}^{n+1} containing $Q \times [t_0, T]$.

We need to impose additionally the technical condition introduced in § 5 of [10]:

Condition 2.1. The function

$$(x, t, u) \mapsto \min_{v \in \Omega} \{l(x, t, v) | f(x, t, u) = f(x, t, v)\}: \mathbf{A} \times \Omega \rightarrow \mathbb{R}$$

is continuous.

Relaxed controls $\{\mu_i; t_0 \leq t \leq t_1\}$, $[t_0, t_1] \subset [t_0, T]$ are $C^*(\Omega)$ -valued functions satisfying

- (i) $\mu_i \in P^n(\Omega)$ a.e. $t \in [t_0, t_1]$,
- (ii) $t \mapsto \int g(t, u) d\mu_i(u)$ is (Lebesgue) measurable for each $g \in C(\Omega \times [t_0, t_1])$.

Ordinary controls (measurable Ω -valued functions) are embedded in the class of relaxed controls in an obvious way.

Admissible (control trajectory) pairs are couples $\{\mu_i, x(t); t_0 \leq t \leq t_1\}$ of relaxed controls $\{\mu_i; t_0 \leq t \leq t_1\}$ and absolutely continuous \mathbb{R}^n -valued functions $\{x(t); t_0 \leq t \leq$

$t_1\}$ satisfying: $x(t) \in Q$ for all $t \in [t_0, t_1]$; $dx(t)/dt = \int f(x(t), t, u) d\mu_t(u)$, a.e. $t \in [t_0, t_1]$; and the endpoint conditions $x(t_0) = x_0$, $(x(t_1), t_1) \in \Gamma$. Existence of an admissible pair is assumed.

The control problem (the Strong Problem) may now be stated:

$$(S) \quad \begin{aligned} & \text{Minimize } \int_{t_0}^{t_1} \int l(x(t), t, u) d\mu_t(u) dt \\ & \text{over admissible pairs } \{\mu_t, x(t); t_0 \leq t \leq t_1\}. \end{aligned}$$

This is the Strong Problem introduced in [10].

Condition 2.1 is introduced to assure equivalence of the Strong Problem and its weak formulation as given in [10] and restated below. Recall that the condition is met if either

- (i) $u \mapsto f(x, t, u)$ is injective for all $(x, t) \in \mathbf{A}$, or
- (ii) there exists $g \in C(\mathbf{A} \times \mathbb{R}^n)$ such that

$$l(x, t, u) = g(x, t, f(x, t, u)) \quad \text{for all } (x, t, u) \in \mathbf{A} \times \Omega.$$

Finally we define the *reachable set*

$$\mathcal{R} = \{(x(t), t) \in \mathbb{R}^{n+1} | t \in [t_0, T]; \{\mu_s, x(s); t_0 \leq s \leq t_1\}, t \leq t_1\},$$

is an admissible pair (except we do not require $(x(t_1), t_1) \in \Gamma$). \mathcal{R} is a closed subset of \mathbb{R}^{n+1} , as may be deduced from results in [6].

Evidently the class of admissible pairs $\{\mu_t, x(t); t_0 \leq t \leq t_1\}$ is unaltered by strengthening the endpoint constraint to

$$x(t_0) = t_0, \quad (x(t_1), t_1) \in \Gamma \cap \mathcal{R}.$$

3. Sufficient conditions for optimality and refinements. Let $\{\mu_t, x(t); t_0 \leq t \leq t_1\}$ be an admissible pair. Suppose that there exists some $\phi \in C^1(\mathbf{A})$ with the following properties:

$$(3.1) \quad (i) \quad \phi_t(x, t) + \max_{u \in \Omega} \{\phi_x(x, t)f(x, t, u) - l(x, t, u)\} = 0, \quad \text{all } (x, t) \in Q \times [t_0, T],$$

$$\phi(x, t) = 0, \quad \text{all } (x, t) \in \Gamma;$$

$$(3.2) \quad (ii) \quad \phi_t(x(t), t) + \int \{\phi_x(x(t), t)f(x(t), t, u) - l(x(t), t, u)\} d\mu_t(u) = 0, \\ \text{a.e. } t \in [t_0, T];$$

then $\{\mu_t, x(t), t_0 \leq t \leq t_1\}$ solves the Strong Problem. Indeed let $\{\bar{\mu}_t, \bar{x}(t); t_0 \leq t \leq \bar{t}_1\}$ be any admissible pair. $t \mapsto \phi(\bar{x}(t), t)$ is Lipschitz continuous and may be expressed through the integral of its derivative to give

$$\begin{aligned} -\phi(x_0, t_0) &= \phi(\bar{x}(\bar{t}_1), \bar{t}_1) - \phi(x_0, t_0) \\ &= \int_{t_0}^{\bar{t}_1} \left\{ \phi_t(\bar{x}(t), t) + \int \phi_x(\bar{x}(t), t)f(\bar{x}(t), t, u) d\bar{\mu}_t(u) \right\} dt \\ &\leq \int_{t_0}^{\bar{t}_1} \int l(\bar{x}(t), t, u) d\mu_t(\bar{u}) dt \end{aligned}$$

(we have used (3.1)).

On the other hand, in view of (3.2),

$$-\phi(x_0, t_0) = \int_{t_0}^{t_1} \int l(x(t), t, u) d\mu_t(u) dt.$$

These two assertions amount to a statement that $\{\mu_t, x(t); t_0 \leq t \leq t_1\}$ is a solution to the Strong Problem.

We have in fact enunciated a sufficient condition for optimality very similar to, for example, [1, p. 339] or [9, p. 192].

This sufficient condition admits certain refinements. Firstly in (3.1) we may relax the partial differential equation to inequality; the boundary condition too may be relaxed to inequality and be required to apply only on $\Gamma \cap \mathbb{R}$; thus

$$(3.1)' \quad \begin{aligned} \phi_t + \max_{u \in \Omega} \{\phi_x t - l\} &\leq 0, \quad \text{all } (x, t) \in Q \times [t_0, T], \\ \phi(x, t) &\geq 0, \quad \text{all } (x, t) \in \Gamma \cap \mathbb{R}. \end{aligned}$$

Secondly, we may state the condition, not in terms of some $\phi \in C^1(\mathbf{A})$ satisfying (3.1)', but rather through a sequence of ϕ_i 's in $C^1(\mathbf{A})$ satisfying (3.1)'. Namely we require of the admissible pair $\{\mu_t, x(t); t_0 \leq t \leq t_1\}$

$$(3.2)' \quad \begin{aligned} \lim_i \left\{ t \mapsto \left(\phi_i^i(x(t), t) + \int \{ \phi_x^i(x(t), t) f(x(t), t, u) - l(x(t), t, u) \} d\mu_t(u) \right) \right\} &= 0 \\ &\text{(strong } L^1(t_0, t_1) \text{ limit),} \\ \phi^i(x(t_1), t_1) &\rightarrow 0, \quad \text{as } i \rightarrow \infty. \end{aligned}$$

Simple verification that any admissible pair satisfying (3.2)' solves the problem is omitted.

A desire to enlarge the class of control problems having solutions characterized through conditions similar to (3.1), (3.2) motivates these refinements. The main results of this paper is that (3.1)', (3.2)' in a sense provide the ultimate refinement, for these conditions are also *necessary* for optimality.

4. The main results.

THEOREM 4.1. *The Strong Problem has a solution. There exists a sequence $\{\phi^i\}$ in $C^1(\mathbf{A})$ such that*

$$\begin{aligned} \phi^i(x, t) + \sup_{u \in \Omega} \{ \phi_x^i(x, t) f(x, t, u) - l(x, t, u) \} &\leq 0, \quad (x, t) \in Q \times [t_0, T], \\ \phi^i(x, t) &\geq 0, \quad \text{all } (x, t) \in \Gamma \cap \mathbb{R}, \end{aligned}$$

for $i = 1, 2, \dots$, and having the property: the admissible pair $\{\mu_t, x(t); t_0 \leq t \leq t_1\}$ solves the control problem if and only if

$$\begin{aligned} \lim_i \left\{ t \mapsto \left(\phi^i(x(t), t) + \int \{ \phi_x^i(x(t), t) f(x(t), t, u) - l(x(t), t, u) \} d\mu_t(u) \right) \right\} &= 0 \\ &\text{(strong limit in } L^1(t_0, t_1)), \\ \lim_i \phi^i(x(t_1), t_1) &= 0. \end{aligned}$$

In addition we provide the following new, *tight* lower bound on the value $\eta(S)$ of the Strong Problem.

THEOREM 4.2.

$$\eta(S) = \sup \{ -\phi(x_0, t_0) \}$$

where the supremum is taken over $\phi \in C^1(\mathbf{A})$ such that

$$\begin{aligned} \phi_t(x, t) + \max_{u \in \Omega} \{ \phi_x(x, t) f(x, t, u) - l(x, t, u) \} &\leq 0, & (x, t) \in Q \times [t_0, T], \\ \phi(x, t) &\geq 0, & \text{all } (x, t) \in \Gamma \cap \mathbb{R}. \end{aligned}$$

It is crucial for the above results that we admit relaxed controls. Indeed no sequence $\{\phi^i\}$ exists characterizing optimal controls (and trajectories) as in Theorem 4.1 over the class of *ordinary* controls when the minimum cost over ordinary controls is achieved and is strictly greater than the minimum cost over relaxed controls; with the endpoint constraints considered here, this may easily occur.

Characterization of optimal controls through a smooth solution to the Bellman equation (§ 3) assumed to exist, is actually supplied for a whole family of control problems parametrized by the initial condition (x_0, t_0) . In contrast, the sequence $\{\phi^i\}$ in Theorem 4.1 characterizes solutions for a particular initial condition. Notice however that characterization is supplied here even in the case that the domain of the value function (that is, the minimum cost as a function of the initial condition) has empty interior. We observe also that, for the particular initial condition, the characterization is of *all* solutions.

5. The Weak Problem and its formulation as a Fenchel problem. Corresponding to the Strong Problem (S) of § 2 we introduce the Weak Problem as in [10, § 4]:

$$\begin{aligned} \text{(W)} \quad & \text{Minimize } \int l \, d\mu \quad \text{over } \mu \in C^*(\mathbf{A} \times \Omega) \\ & \text{subject to } \mu \in \mathcal{M} \cap P^\oplus(\mathbf{A} \times \Omega). \end{aligned}$$

Here

$$\begin{aligned} \mathcal{M} = \left\{ \mu \in C^*(\mathbf{A} \times \Omega) \mid \text{there exists } \beta \in P^n(\Gamma \cap \mathbb{R}) \text{ s.t. } \int (\phi_t + \phi_x f) \, d\mu = \int \phi \, d\beta \right. \\ \left. - \phi(x_0, t_0), \text{ all } \phi \in C^1(\mathbf{A}) \right\} \\ \cap \left\{ \mu \mid \int \text{dist}\{x, Q\} \, d\mu = 0 \right\} \end{aligned}$$

($\text{dist}\{x, Q\}$ denotes Euclidean distance from $x \in \mathbb{R}^n$ to the set $Q \in \mathbb{R}^n$). Notice that the “target set” in the Strong Problem has been taken as $\Gamma \cap \mathbb{R}$ rather than Γ (recall the concluding remarks of § 2).

Problem (W) represents a weakening of the constraints on the Strong Problem. Indeed feasible elements for (S), that is admissible pairs $\{\mu_t, x(t); t_0 \leq t \leq t_1\}$, define feasible elements μ for (W); thus

$$(5.1) \quad g \mapsto \int_{t_0}^{t_1} \int g(x(t), t, u) \, d\mu_t(u) \, dt$$

and

$$\int_{t_0}^{t_1} \int l(x(t), t, u) \, d\mu_t(u) \, dt = \int l \, d\mu.$$

Define the $\mathbb{R} \cup \{+\infty\}$ -valued function p on $C^*(\mathbf{A} \times \Omega)$ as

$$p(\mu) = \begin{cases} \int l \, d\mu & \text{if } |\mu| \leq (T - t_0), \mu \in P^\oplus(\mathbf{A} \times \Omega), \\ +\infty & \text{otherwise} \end{cases}$$

and the $\mathbb{R} \cup \{-\infty\}$ -valued function q on $C^*(\mathbf{A} \times \Omega)$ as

$$q(\mu) = \begin{cases} 0 & \mu \in \mathcal{M}, \\ -\infty & \text{otherwise.} \end{cases}$$

We have

LEMMA 5.1. *Equip $C^*(\mathbf{A} \times \Omega)$ with the weak* topology. Then p is lower semicontinuous, convex, and q is upper semicontinuous, concave with*

$$(5.2) \quad \{\mu \in C^*(\mathbf{A} \times \Omega) | p(\mu) \neq +\infty, q(\mu) \neq -\infty\} \neq \emptyset.$$

Proof. Let μ be feasible for (W). Taking $\phi(x, t) = t$, we see that

$$|\mu| = \int 1 \, d\mu = \int t \, d\beta - t_0 \quad (\text{for some } \beta \in P^n(\Gamma \cap \mathbb{R})), \\ \leq (T - t_0).$$

Thus we do not augment the constraints on the Weak Problem by imposing the additional constraints $|\mu| \leq (T - t_0)$. It follows that $\mu \in C^*(\mathbf{A} \times \Omega)$ is feasible for the Weak Problem if and only if $p(\mu) - q(\mu) < \infty$ and that, given μ feasible for the Weak Problem,

$$\int l \, d\mu = p(\mu) - q(\mu).$$

However there is an element feasible for the Weak Problem, given by the admissible control trajectory pair, whose existence is assumed, through the embedding (5.1). Condition (5.2) is proved.

Next we show that \mathcal{M} is weak* closed in $C^*(\mathbf{A} \times \Omega)$. Let $\{\mu_p | p \in D\}$ be a generalized sequence in \mathcal{M} converging in the weak* topology to $\mu_0 \in C^*(\mathbf{A} \times \Omega)$. Evoking the axiom of choice, we have a corresponding generalized sequence $\{\beta_p | p \in D\}$ in $P^n(\Gamma \cap \mathbb{R})$ (D , the partially ordered set above). But $P^n(\Gamma \cap \mathbb{R})$ is weak* compact, so $\{\beta_p\}$ has a weak* cluster point $\beta_0 \in P^n[\Gamma \cap \mathbb{R}]$.

Fix $\phi \in C^1(\mathbf{A})$. Then, for given $\varepsilon > 0$ there exists $p \in D$ with $\mu_p \in \mathcal{M}$, $\beta_p \in P^n(\Gamma \cap \mathbb{R})$ lying in the neighborhoods $\{\mu | |\int (\phi_t + \phi_x f) \, d(\mu - \mu_0)| < \varepsilon/2\}$, $\{\beta | |\int \phi \, d(\beta - \beta_0)| < \varepsilon/2\}$ respectively. It follows that $|\int (\phi_t + \phi_x f) \, d\mu_0 - \int \phi \, d\beta_0 - \phi(x_0, t_0)| = |\int (\phi_t + \phi_x f) \, d(\mu_0 - \mu_p) - \int \phi \, d(\beta_0 - \beta_p)| < \varepsilon$. Evidently then

$$\int (\phi_t + \phi_x f) \, d\mu_0 = \int \phi \, d\beta_0 - \phi(x_0, t_0),$$

and this relation holds for all $\phi \in C^1(\mathbf{A})$. Since $\int \text{dist}\{x, Q\} \, d\mu_0 = 0$ ($\text{dist}\{\beta, Q\}$ is continuous on A), it follows that \mathcal{M} is weak* closed in $C^*(\mathbf{A} \times \Omega)$.

Thus q on its effective domain (points at which it is finite valued) is the restriction of a linear function, continuous with respect to the weak* topology, to a weak* closed, convex subset of $C^*(\mathbf{A} \times \Omega)$; likewise for p , since as is well known $\{\mu | |\mu| \leq T - t_0\}$ and therefore $\{\mu | |\mu| \leq T - t_0\} \cap (\cap_{g \in P(\mathbf{A} \times \Omega)} \{\mu | \int g \, d\mu \geq 0\}) = \{\mu | |\mu| \leq T - t_0\} \cap P^\oplus(\mathbf{A} \times \Omega)$ is weak* closed. It is now a simple matter to show that q is convex, lower semicontinuous, and that p is concave, upper semicontinuous.

Actually we have additionally shown:

PROPOSITION 5.1. *If μ is feasible for the Weak Problem (and such a feasible element exists), then*

$$\int l \, d\mu = p(\mu) - q(\mu).$$

Otherwise $p(\mu) - q(\mu) = +\infty$.

Thus we may treat the Weak Problem as that of minimizing $(p-q)(\mu)$ over $\mu \in C^*(\mathbf{A} \times \Omega)$.

The main result of [10] may now be written as

THEOREM 5.1.

$$\eta(S) = \inf \{ (p-q)(\mu) \mid \mu \in C^*(\mathbf{A} \times \Omega) \}$$

($\eta(S)$, the value of the Strong Problem).

6. Computation of the dual functionals. $C(\mathbf{A} \times \Omega)$ with the topology of uniform convergence, and $C^*(\mathbf{A} \times \Omega)$ with the weak* topology are topological linear spaces in duality. It is natural therefore to introduce the dual functions p^*, q^* on $C(\mathbf{A} \times \Omega)$,

$$p^*(\xi) = \sup \left\{ \int \xi d\mu - p(\mu) \mid \mu \in C^*(\mathbf{A} \times \Omega) \right\}$$

$$q^*(\xi) = \inf \left\{ \int \xi d\mu - q(\mu) \mid \mu \in C^*(\mathbf{A} \times \Omega) \right\}$$

for $\xi \in C(\mathbf{A} \times \Omega)$.

We now give more explicit representation of these dual functions. Here and subsequently we write $g^+ \in C(\mathbf{A} \times \Omega)$, given $g \in C(\mathbf{A} \times \Omega)$, for the function

$$g^+(x, t, u) = \begin{cases} g(x, t, u) & \text{when } g(x, t, u) \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

We also abbreviate the element $\text{dist } \{x, Q\}$ in $C^*(\mathbf{A} \times \Omega)$ as d . Γ_r denotes $\Gamma \cap \mathbb{R}$. It will sometimes be convenient to write pairs (x, t) in \mathbb{R}^{n+1} as y . In particular y_0 denotes (x_0, t_0) .

PROPOSITION 6.1. For $\xi \in C(\mathbf{A} \times \Omega)$

$$p^*(\xi) = \max \{ (\xi(x, t, u) - l(x, t, u))^+ \mid (x, t, u) \in \mathbf{A} \times \Omega \} \cdot (T - t_0).$$

If we define

$$W = \{ \xi \in C(\mathbf{A} \times \Omega) \mid \xi(x, t, u) = \phi_t(x, t) + \phi_x(x, t)f(x, t, u) + \alpha \text{ dist } \{x, Q\}, \\ \text{for some } \phi \in C^1(\mathbf{A}), \alpha \in \mathbb{R} \},$$

then

$$q^*(\xi) \begin{cases} < -\infty & \text{for } \xi \in \text{closure } \{W\}, \\ = -\infty & \text{otherwise,} \end{cases}$$

and given sequences $\{\phi^i\}$ and $\{\alpha^i\}$ in $C^1(\mathbf{A}), \mathbb{R}$, respectively, such that

$$\xi(x, t, u) = \lim_i \{ \phi_t^i(x, t) + \phi_x^i(x, t)f(x, t, u) + \alpha^i \text{ dist } \{x, Q\} \}$$

(strong limit in $C(\mathbf{A} \times \Omega)$) exists, then $q^*(\xi)$ is well-defined as

$$q^*(\xi) = \lim_i \left\{ \min_{(x, t) \in \Gamma \cap \mathbb{R}} \phi^i(x, t) - \phi^i(x_0, t_0) \right\}.$$

Proof.

$$\begin{aligned}
 p^*(\xi) &= \sup \left\{ \int (\xi - l) d\mu \mid \mu \in P^\oplus(\mathbf{A} \times \Omega), |\mu| \leq T - t_0 \right\} \\
 &= \sup \left\{ \int (\xi - l)^+ d\mu - \int (l - \xi)^+ d\mu \mid \mu \in P^\oplus(\mathbf{A} \times \Omega), |\mu| \leq T - t_0 \right\} \\
 &\leq \sup \left\{ \int (\xi - l)^+ d\mu \mid \mu \in P^\oplus(\mathbf{A} \times \Omega), |\mu| \leq T - t_0 \right\} \\
 &\leq \max \{ (\xi - l)^+(x, t, u) \mid (x, t, u) \in \mathbf{A} \times \Omega \} \cdot (T - t_0) = \int (\xi - l) d\mu_0 \leq p^*(\xi).
 \end{aligned}$$

Here, $\mu_0 = 0$ if $\xi - l \leq 0$; otherwise $\mu_0 \in P^\oplus(\mathbf{A} \times \Omega)$ has support in the nonempty set $\arg \max (\xi - l)$, and is such that $\int 1 d\mu_0 = T - t_0$. This deals with p^* .

We now turn to q^* . For $\xi \in C(\mathbf{A} \times \Omega)$, $q^*(\xi)$ is the infimum of $\int \xi d\mu$ over the (nonempty) set

$$\mathcal{M} = \left\{ \mu \in C^*(\mathbf{A} \times \Omega) \mid \text{there exists } \beta \in P^n(\Gamma_r) \text{ subject to } \int \{ \phi_t + \phi_x f \} d\mu = \int \phi d\mu - \phi(y_0), \text{ for all } \phi \in C^1(\mathbf{A}) \right\} \cap \left\{ \mu \mid \int d\mu = 0 \right\}.$$

Take $\xi \in W$. Then $\xi = \phi_t + \phi_x f + \alpha d$, for some $\phi \in C^1(\mathbf{A})$ some $\alpha \in \mathbb{R}$. We have

$$q^*(\xi) = \inf \left\{ \int \phi d\beta \right\} - \phi(y_0)$$

where the infimum is over \mathcal{B} , the subset of $P^n(\Gamma_r)$ comprising elements corresponding to which there exist $\mu \in C^*(\mathbf{A} \times \Omega)$ such that

$$\begin{aligned}
 (6.1) \quad & \int (\phi_t + \phi_x f) d\mu = \int \phi d\beta - \phi(y_0) \quad \text{for all } \phi \in C^1(\mathbf{A}), \\
 & \int d\mu = 0.
 \end{aligned}$$

Evidently,

$$\int \phi d\beta \geq \int \phi d\beta_0, \quad \beta \in \mathcal{B},$$

where β_0 is a unit measure concentrated at a point y_1 minimizing ϕ over $\Gamma \cap \mathbb{R}$. But $\beta_0 \in \mathcal{B}$ since $y_1 \in \mathcal{R}$; indeed a μ satisfying (6.1), for $\beta = \beta_0$, is given by an admissible control trajectory pair with endpoints (y_0, y_1) under the embedding (5.1).

We have then that

$$q^*(\xi) \left(= \int \phi d\beta_0 - \xi(y_0) \right) = \min_{y \in \Gamma_r} \phi(y) - \phi(y_0)$$

for $q = \phi_t + \phi_x f + \alpha d$, $\phi \in C^1(\mathbf{A})$, $\alpha \in \mathbb{R}$.

q^* has been characterized on W . Let us now examine q^* on the complement of \bar{W} , the closure of W . Choose $\xi \notin \bar{W}$. Then there is a closed hyperplane strictly separating ξ and the subspace \bar{W} , that is, there exists $\bar{\mu} \in C^*(\mathbf{A} \times \Omega)$ with

$$\int \bar{\xi} d\bar{\mu} = 0, \quad \bar{\xi} \in \bar{W},$$

$$\int \xi d\bar{\mu} \neq 0$$

[8, p. 186]. Choose $\mu \in \mathcal{M}$ (such an element exists by assumption). We readily check that, for any real α , $\mu + \alpha \bar{\mu} \in \mathcal{M}$. But $\int \xi d(\mu + \alpha \bar{\mu})$ may then be taken arbitrarily

negative by appropriate choice of α . We have shown that

$$q^*(\xi) = -\infty, \quad \xi \notin \bar{W}.$$

It remains to examine the case when $\xi \in \bar{W}$. Let

$$\xi^i \rightarrow \xi \quad (\text{strongly in } C(\mathbf{A} \times \Omega))$$

with

$$\xi^i = \phi^i + \phi_x^i f + \alpha^i d \quad (\phi^i \in C^1(\mathbf{A}), \quad \alpha^i \in \mathbb{R})$$

$i = 1, 2, \dots$. We have shown that

$$q^*(\xi^i) = \min_{y \in \Gamma_r} \phi^i(y) - \phi^i(y_0).$$

Let y_i achieve the minimum of ϕ^i over Γ_r and let $\{\mu_i\}$ be a sequence in $C^*(\mathbf{A} \times \Omega)$ defined through admissible pairs with right endpoints y_i via the embedding. We have then

$$(6.2) \quad q^*(\xi^i) = \phi^i(y_i) - \phi^i(y_0) = \int \xi^i d\mu_i.$$

Now admissible pairs define elements in $C^*(\mathbf{A} \times \Omega)$ which are uniformly bounded in the dual norm. By extracting a subsequence therefore, we can arrange that

$$\mu_i \rightarrow \mu_0 \quad (\text{weakly}^*)$$

for some $\mu \in C^*(\mathbf{A} \times \Omega)$. Since \mathcal{M} is weak* closed (§ 5), $\mu_0 \in \mathcal{M}$.

Recalling that ξ^i converges strongly to ξ , we have, for the subsequence,

$$\int \xi^i d\mu_i \rightarrow \int \xi d\mu_0,$$

and from (6.2)

$$\lim_i q^*(\xi^i) = \int \xi d\mu_0.$$

But q^* , as a concave conjugate functional, is upper semicontinuous so that $q^*(\xi) \geq \int \xi d\mu_0$. However, as we have observed, $\mu_0 \in \mathcal{M}$. Since $q(\xi)$ is the infimum of $\int \xi d\mu$ over $\mu \in \mathcal{M}$, $q^*(\xi) \leq \int \xi d\mu_0$. It follows that, for the subsequence,

$$(6.3) \quad q^*(\xi) = \int \xi d\mu_0 = \lim_i \left\{ \min_{y \in \Gamma_r} \phi^i(y) - \phi^i(y_0) \right\}.$$

The limit is independent of the particular subsequence whence (6.3) applies for the original sequence. The proposition is proved.

7. The Fenchel dual problem. The Fenchel dual of the Weak Problem (making identifications of p and q as in § 5) is given by

$$(D) \quad \begin{aligned} &\text{Maximize } q^*(\xi) - \Gamma^*(\xi) \\ &\text{over } \xi \in C(\mathbf{A} \times \Omega). \end{aligned}$$

The following proposition gives a useful characterization of the value $\eta(D)$ of the Dual Problem.

PROPOSITION 7.1.

$$\eta(D) = \sup \{ -\phi(x_0, t_0) \}$$

where the supremum is taken over $\phi \in C^1(\mathbf{A})$ such that

$$\begin{aligned}\phi_t(x, t) + \phi_x(x, t)f(x, t, u) - l(x, t, u) &\leq 0, & (x, t, u) \in Q \times [t_0, T] \times \Omega, \\ \phi(x_1, t_1) &\geq 0, & (x_1, t_1) \in \Gamma \cap \mathbb{R}.\end{aligned}$$

Proof. In view of Proposition 6.1,

$$\begin{aligned}\eta(D) &= \sup \{ (q^* - p^*)(\xi) \mid \xi \in C(\mathbf{A} \times \Omega) \} \\ &= \sup \{ \bar{z}(\phi, \alpha) \mid (\phi, \alpha) \in C^1(\mathbf{A}) \times \mathbb{R} \}\end{aligned}$$

where

$$(7.1) \quad \bar{z}(\phi, \alpha) = \min_{y \in \Gamma_r} \phi(y) - \phi(y_0) - (T - t_0) \cdot \max \{ (\phi_t + \phi_x f + \alpha d - l)^+ \mid \mathbf{A} \times \Omega \}.$$

(Notice that we use the continuity of the map $g \mapsto \max \{ g^+ \mid \mathbf{A} \times \Omega \} : C(\mathbf{A} \times \Omega) \rightarrow \mathbb{R}$ to justify taking the supremum here over W (W as in Proposition 6.1), rather than over its closure.)

Let us write

$$m_{\phi, \alpha}(x, t, u) = \min_{y \in \Gamma_r} \phi(y) - \phi(y_0) - (T - t_0) \cdot (\phi_t + \phi_x f + \alpha d - l)^+(x, t, u)$$

for $(x, t, u) \in \mathbf{A} \times \Omega$, so that

$$\bar{z}(\phi, \alpha) = \min \{ m_{\phi, \alpha} \mid \mathbf{A} \times \Omega \}.$$

Evidently, $\alpha \mapsto \bar{z}(\phi, \alpha)$ is monotone nonincreasing, and we conclude that

$$(7.2) \quad \eta(D) = \sup \left\{ \lim_{\alpha \rightarrow -\infty} \bar{z}(\phi, \alpha) \mid \phi \in C^1(\mathbf{A}) \right\}.$$

For $\varepsilon \geq 0$, define

$$G_\varepsilon = \{ (x, t, u) \in \mathbf{A} \times \Omega \mid \text{dist} \{ x, Q \} \leq \varepsilon \}$$

and \tilde{G}_ε its complement in $\mathbf{A} \times \Omega$. For arbitrary α

$$(7.3) \quad \bar{z}(\phi, \alpha) \leq \min \{ m_{\phi, \alpha} \mid G_0 \}$$

because $G_0 \subset \mathbf{A} \times \Omega$ ($d = 0$ on $G_0 = Q \times [t_0, T] \times \Omega$). For $\varepsilon > 0$ and $\alpha \leq -(1/\varepsilon) \cdot \max \{ (\phi_t + \phi_x f - l)^+ \mid \mathbf{A} \times \Omega \}$ however

$$\begin{aligned}\bar{z}(\phi, \alpha) &= \min \{ \min \{ m_{\phi, \alpha} \mid G_\varepsilon \}, \min \{ m_{\phi, \alpha} \mid \tilde{G}_\varepsilon \} \} \\ &= \min \{ m_{\phi, \alpha} \mid G_\varepsilon \} \\ &\geq \min \{ m_{\phi, 0} \mid G_\varepsilon \}.\end{aligned}$$

By the uniform continuity of the continuous function $(x, t, u) \mapsto m_{\phi, 0}(x, t, u)$ on the compact set $\mathbf{A} \times \Omega$

$$\min \{ m_{\phi, 0} \mid G_\varepsilon \} \rightarrow \min \{ m_{\phi, 0} \mid G_0 \}$$

as $\varepsilon \rightarrow 0^+$. It follows that

$$\lim_{\alpha \rightarrow -\infty} \bar{z}(\phi, \alpha) \geq \min \{ m_{\phi, 0} \mid G_0 \}$$

whence, by (7.3)

$$\lim_{\alpha \rightarrow -\infty} \bar{z}(\phi, \alpha) = \min \{ m_{\phi, 0} \mid G_0 \}.$$

By (7.2) then

$$\eta(D) = \sup \{z(\phi) | \phi \in C^1(\mathbf{A})\}$$

where

$$z(\phi) = \min_{y \in \Gamma_r} \phi(y) - \phi(y_0) - (T - t_0) \cdot \max \{(\phi_t + \phi_x f - l)^+ | Q \times [t_0, T] \times \Omega\}.$$

For given $\tilde{\phi} \in C^1(\mathbf{A})$, define $\tilde{\tilde{\phi}}$ as

$$\tilde{\tilde{\phi}}(x, t) = \psi(x, t) - \min_{(x, t) \in \Gamma_r} \psi(x, t) \quad \text{for } (x, t) \in \mathbf{A},$$

where

$$\psi(x, t) = \tilde{\phi}(x, t) - \max \{(\tilde{\phi}_t + \tilde{\phi}_x f - l)^+ | Q \times [t_0, T] \times \Omega\} \cdot (t - t_0), \quad (x, t) \in \mathbf{A}.$$

We readily check that

$$\tilde{\tilde{\phi}} \in \mathcal{S} \cap \left\{ \phi \in C^1(\mathbf{A}) \mid \min_{y \in \Gamma_r} \phi(y) = 0 \right\},$$

$$\mathcal{S} = \{ \phi \in C^1(\mathbf{A}) | \phi_t + \phi_x f - l \leq 0 \text{ on } Q \times [t_0, T] \times \Omega; \phi(x, t) \geq 0, (x, t) \in \Gamma_r \}$$

and

$$z(\tilde{\tilde{\phi}}) \geq z(\tilde{\phi}).$$

It follows that

$$\eta(D) = \sup \{z(\phi) | \phi \in C^1(\mathbf{A})\} = \sup \left(-\phi(x_0, t_0) | \phi \in \mathcal{S} \cap \left\{ \phi \mid \min_{y \in \Gamma_r} \phi(y) = 0 \right\} \right).$$

However, given $\phi \in \mathcal{S}$, $\tilde{\phi}$ defined by

$$\tilde{\phi}(x, t) = \phi(x, t) - \min_{y \in \Gamma} \phi(y)$$

satisfies

$$-\tilde{\phi}(x_0, t_0) \geq -\phi(x_0, t_0), \quad \tilde{\phi} \in \mathcal{S} \cap \left\{ \phi \mid \min_{y \in \Gamma_r} \phi(y) = 0 \right\}.$$

We conclude

$$\eta(D) = \sup \{ -\phi(x_0, t_0) | \phi \in \mathcal{S} \}.$$

This proves the proposition.

8. Proof of the main results.

PROPOSITION 8.1.

$$\eta(W) = \eta(D).$$

Proof. We have shown that the Weak Problem may be formulated as

$$\text{minimize } p(\mu) - q(\mu) \text{ over } \mu \in C^*(\mathbf{A} \times \Omega)$$

where p, q are respectively convex lower semicontinuous, concave upper semicontinuous. Further the dual function p^* has been computed as $p^*(\xi) = \max \{(\xi - l)^+ | \mathbf{A} \times \Omega\} \cdot (T - t_0)$, for $\xi \in C(\mathbf{A} \times \Omega)$. We notice that this function is finite

valued on $C(\mathbf{A} \times \Omega)$ and everywhere continuous (with respect to the topology of uniform convergence). The dual function q^* does not take value $-\infty$ everywhere on its domain. It follows that q^* is finite at some point in its domain where p^* is continuous. This implies that the values of the Weak and the Dual Problems coincide [7].

Proof of Theorems 4.1 and 4.2. We have from Propositions 7.1, 8.1

$$\eta(W) = \sup \{ -\phi(x_0, t_0) \mid \phi \in C^1(\mathbf{A}) \text{ subject to } \phi_t + \phi_x f - l \leq 0 \text{ on } \\ Q \times [t_0, T] \times \Omega \text{ and } \phi(x_1, t_1) \geq 0, (x_1, t_1) \in \Gamma \cap \mathbb{R} \}.$$

However the values of the Strong Problem and the Weak Problem coincide (Theorem 5.1). Theorem 4.2 is proved.

Turning to Theorem 4.1, the Strong Problem has a solution [10]. To prove the rest of the theorem, suppose that $\{\phi^i\}$ is a sequence in $C^1(\mathbf{A})$ such that

$$\begin{aligned} \phi_t^i(x, t) + \phi_x^i(x, t)f(x, t, u) &\leq 0, & (x, t, u) \in Q \times [t_0, T] \times \Omega, \\ \phi^i(x_1, t_1) &\geq 0, & (x_1, t_1) \in \Gamma \cap \mathbb{R}, \end{aligned}$$

and

$$\eta(S) = \lim_i \{ -\phi^i(x_0, t_0) \}.$$

Existence of such a sequence follows from Theorem 4.2.

First let $\{\mu_i, x(t) : t_0 \leq t \leq t_1\}$ be an admissible pair solving the Control Problem. Then

$$\eta(S) = \int_{t_0}^{t_1} l(x(t), t, u) d\mu_t(u) dt = \lim_i \{ -\phi^i(x_0, t_0) \}.$$

However because of the properties of admissible pairs,

$$\begin{aligned} &\phi^i(x(t_1), t_1) - \phi^i(x_0, t_0) \\ &= \int_{t_0}^{t_1} \left\{ \phi_t^i(x(t), t) + \int \{ \phi_x^i(x(t), t)f(x(t), t, u) \} d\mu_t(u) \right\} dt, \quad i = 1, 2, \dots \end{aligned}$$

It follows that

$$\begin{aligned} \lim_i \left\{ -\phi^i(x(t_1), t_1) + \int_{t_0}^{t_1} \left\{ \phi_t^i(x(t), t) \right. \right. \\ \left. \left. + \int \{ \phi_x^i(x(t), t)f(x(t), t, u) - l(x(t), t, u) \} d\mu_t(u) \right\} dt \right\} = 0. \end{aligned}$$

But $\phi^i(x(t_1), t_1) \geq 0$ and the integrands are everywhere nonpositive. In consequence

$$\begin{aligned} &\phi^i(x(t_1), t_1) \rightarrow 0, \\ &\left\{ t \mapsto \int (\phi_t^i + \phi_x^i f - l)(x(t), t, u) d\mu_t(u) \right\} \rightarrow 0 \quad (\text{strongly in } L^1(t_0, t_1)) \end{aligned}$$

as $i \rightarrow \infty$.

Conversely, if an admissible pair $\{\bar{\mu}, \bar{x}(t); t_0 \leq t \leq t_1\}$ satisfies the conditions of Theorem 4.1, then

$$\begin{aligned} & \int_{t_0}^{t_1} \int l(\bar{x}(t), t, u) d\bar{\mu}_t(u) dt \\ &= \lim_i \int_{t_0}^{t_1} \left\{ \phi_i^t(\bar{x}(t), t) + \phi_x^i(\bar{x}(t), t) \int f(\bar{x}(t), t, u) d\bar{\mu}_t(u) \right\} dt \\ &= \lim_i \{ \phi^i(x(t_1), t_1) - \phi^i(x_0, t_0) \} = \lim_i \{ -\phi^i(x_0, t_0) \} = \eta(S). \end{aligned}$$

Thus $\{\mu_t, x(t); t_0 \leq t \leq t_1\}$ solves the Strong Problem and the theorem is proved.

9. Some concluding remarks. The simple sufficient condition of § 3 (relaxing the partial differential equation (3.1) to the partial differential inequality (3.1')) may be seen to apply to the special situation where the dual problem has a solution in the subspace \mathcal{W} comprising ξ 's expressible as $\xi = \phi_t + \phi_x f$ for some continuously differentiable ϕ . The "effective domain" of the dual problem is the closure of this subspace. This situation is special indeed, for we cannot guarantee that the dual problem even has a solution, let alone a solution in the subset \mathcal{W} .

Under the further assumption that

$$\left\{ \begin{pmatrix} l(x, t, u) \\ f(x, t, u) \end{pmatrix} \in \mathbb{R}^{n+1} \mid u \in \Omega \right\}$$

is convex, for each $(x, t) \in Q \times [t_0, T]$, Theorem 4.1 applies for the Strong Problem posed over ordinary control, trajectory pairs. Theorem 4.1 is a consequence of the equivalence of the Strong and the Weak Problems and such equivalence also applies with reference to ordinary controls under the extra assumption [10, § 10].

For discussion of extensions of the results reported here to allow terminal costs, explicit state constraints, etc., we refer to [5].

REFERENCES

- [1] V. G. BOLTYANSKI, *Sufficient conditions for optimality and the justification of the dynamic programming method*, this Journal (1966), pp. 326–361.
- [2] T. F. BRIDGLAND, *On the problem of approximate synthesis of optimal controls*, this Journal (1967), pp. 326–344.
- [3] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.
- [4] H. HERMES, *The equivalence and approximation of optimal control problems*, J. Differential Equations, 1 (1965), pp. 409–426.
- [5] R. M. LEWIS, *Problem Equivalence and Necessary Conditions of Dynamic Programming Type in Optimal Control*, Ph.D. thesis, Imperial College, London, 1977.
- [6] E. J. MCSHANE, *Relaxed controls and variational problems*, this Journal (1967), pp. 438–485.
- [7] R. T. ROCKAFELLAR, *Extensions of Fenchel's duality theory for convex functions*, Duke Math J., 33 (1966), pp. 81–89.
- [8] A. E. TAYLOR, *Introduction to Functional Analysis*, John Wiley, New York, 1958.
- [9] P. P. VARAIYA, *Notes on Optimization*, Van Nostrand Reinhold, New York, 1972.
- [10] R. B. VINTER AND R. M. LEWIS, *The equivalence of strong and weak formulations for certain problems in optimal control*, this Journal, 16 (1978), pp. 546–570.
- [11] L. C. YOUNG, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.

A DISTRIBUTED FILTER DERIVATION WITHOUT RICCATI EQUATIONS*

JON H. DAVIS†

Abstract. A Wiener–Hopf based solution of the stationary distributed Kalman–Bucy filtering problem is presented. We derive an explicit representation for the optimal filter gains, from which both optimality and stability of the resulting filter follow. The derivation is concise, and avoids entirely consideration of the distributed Riccati equation.

1. Introduction. This paper considers “frequency domain” methods for the derivation of Kalman–Bucy filters for a certain class of distributed systems.

These methods lead (as expected) to the usual form for the optimal filter. The derivation, however, is quite short, and avoids entirely any consideration of a “distributed Riccati equation”, or related operator valued differential equations. We proceed instead from an explicit formula for the feedback gains of the optimal filter. Under the usual [1], [2] assumptions of finite dimensionality of the observation space, this requires calculation only of the “vectors” corresponding to the optimal gains.

Our interest in this approach arose from experience with a “dual problem” involving a distributed least-squares controller [3]. The frequency domain calculation used in [3] was found computationally efficient, while a Riccati equation truncation method proved numerically difficult to handle. The optimal gain representation used below is the result “dual” to that presented in [3]. Both results are motivated by the work presented in [4, § 26], and [5].

We consider the stationary estimation problem below. This is motivated largely by the possibility of explicit frequency domain calculations for various distributed systems, and the availability of algorithms for numerical spectral factorization [6].

2. Basic problem formulation. We consider a stationary estimation problem for the Itô evolution equations [1], [2]

$$(1) \quad \begin{aligned} dx_t &= Ax_t dt + B dw_i(t), \\ dz_t &= Cx_t dt + dw_0(t), \quad t > -\infty. \end{aligned}$$

Here A represents the infinitesimal generator of a strictly stable strongly continuous semi-group $\{S_t\}$ of class C_0 in a separable Hilbert space H , with resolvent operator $R(\cdot; A)$;

$$\|S_t\| \leq Me^{-wt}, \quad \text{for some } w > 0.$$

We assume that B is a bounded linear operator from the separable Hilbert space H_i to H ; $\{w_i(t)\}$ is an H_i -valued Wiener process [1] with (trace class) covariance operator Q . C is a bounded linear operator from the Hilbert space H to the finite dimensional Hilbert space H_0 .

Our standing assumptions are that the noise increments are independent, of mean zero, and independent of past (and current) values of the state.

We consider the stationary estimation problem. Alternative formulations are to consider (1) for $t > 0$, and specify the “steady state” covariance operator as the initial

* Received by the editors July 6, 1977, and in revised form November 10, 1977.

† Department of Mathematics, Queen's University, Kingston, Canada K7L 3N6.

state covariance, or to specify the state directly as (the weak solution)

$$(2) \quad \begin{aligned} x^* x_t &= \int_{-\infty}^t x^* S_{t-\tau} B dw_i(\tau) \\ &= \lim_{t_0 \rightarrow \infty} \int_{-t_0}^t x^* S_{t-\tau} B dw_i(\tau), \end{aligned}$$

for each bounded linear functional x^* .

We assume that the observation noise is full rank. Without loss of generality, we take the incremental covariance of the observation noise as the identity operator on H_0 .

We seek the best linear estimate of the current state in the form of a mean-square convergent integral

$$(3) \quad \hat{x}_t = \int_{-\infty}^t T(t-\tau) dz(\tau).$$

More specifically, we seek a weighting pattern (operator valued, from H_0 to H ; hence effectively vector valued) such that

$$(4) \quad \int_0^\infty \|T(t)\|^2 dt < \infty,$$

while for each bounded linear functional x^*

$$E\{|x^*(x_t - \hat{x}_t)|^2\}$$

is minimized.

This, of course, is essentially the standard Wiener filtering problem, with the assumption of an underlying Hilbert state space model generating the observed process. By the standard arguments, then, we conclude that the optimal weighting pattern must satisfy the classical Wiener-Hopf equation

$$(5) \quad x^* T(\tau) + x^* [T^* R_{yy}(\tau)] = x^* R_{xy}(\tau), \quad \tau \geq 0, \quad \text{for all } x^* \in H^*.$$

In the above, $R_{yy}(\cdot)$ is the output correlation, with Fourier transform

$$(6) \quad S_{yy}(\omega) = CR(i\omega; A)BQB^*(R(i\omega; A))^*C^*;$$

$x^* R_{xy}(\cdot)$ is the estimate-output cross correlation, with Fourier transform

$$(7) \quad x^* S_{xy}(\omega) = x^* R(i\omega; A)BQB^*(R(i\omega; A))^*C^*.$$

Our assumptions suffice to guarantee that a matrix representation of the observation power spectral density operator possesses a spectral factorization

$$(8) \quad I + S_{yy}(\omega) = F^+(\omega)F^-(\omega).$$

This follows from [7], once one establishes that the matrix elements in the representation are Fourier transforms of integrable functions. From the assumption that A generates a stable semi-group, it follows that for $y \in H_0$, $v \in H$, the scalar valued function

$$y^* S_t v, \quad t \geq 0,$$

belongs to $L_1(0, \infty)$, with norm estimate

$$(9) \quad \|y^* S_t v\|_{L_1} \leq \frac{M}{w} \|y\| \cdot \|v\|.$$

Hence its Fourier transform

$$y^* R(i\omega; A)v$$

belongs to \hat{L}_1 .

Since Q is a trace class operator, BQB^* has a spectral representation

$$(10) \quad BQB^* = \sum_{i=1}^{\infty} \lambda_i v_i v_i^*,$$

with $\sum_{i=1}^{\infty} \lambda_i < \infty$. From these estimates it follows easily that (for each $y \in H_0$)

$$y^* CR(i\omega; A)BQB^*[R(i\omega; A)]^* C^* y$$

is the transform of an $L_1(-\infty, \infty)$ function.

From the usual method of solution of the Wiener–Hopf equation (5), it follows at once that the (Fourier transform of the) optimal weighting pattern is given by

$$(11) \quad x^* \hat{T}(\omega) = P_+ \{x^* S_{xy}(\omega) [F^-(\omega)]^{-1}\} [F^+(\omega)]^{-1}.$$

In the above, the operator P_+ represents projection onto functions of positive support.

Of course, the advantage of the Kalman–Bucy approach over the classical Wiener approach to this problem is that an automatic synthesis of the optimal filter is provided by the former. That is, we expect that the optimal filter has a realization of the form

$$(12) \quad d\hat{x} = A\hat{x} dt + [P_{\infty} C^*](dz - C\hat{x} dt).$$

We derive the above equation, however, *without* the introduction of a Riccati equation for the error covariance. Our derivation is based on the classical spectral factorization approach, and provides an algorithm for the direct computation of the quantity represented as $[P_{\infty} C^*]$, the filter gains. This approach bypasses the theoretical and computational difficulties associated with finite dimensional approximations to a Riccati operator.

From the fact that the steady state solution to the Riccati equation is expected to satisfy

$$(13) \quad AP_{\infty} + P_{\infty} A^* - P_{\infty} C^* C P_{\infty} = -BQB^*,$$

and using a method parallel to that of [3] (and having roots in [4, § 26]) we produce an integral representation for the required filter gains. This produces an expression of the form

$$(14) \quad [P_{\infty} C^*] = \frac{1}{2\pi} \int_{-\infty}^{\infty} R(i\omega; A)BQB^*[R(i\omega; A)]^* C^* [F^-(\omega)]^{-1} d\omega.$$

Here $F^-(\omega)$ denotes the (anti-causal) factor of the power spectral density factorization

$$(15) \quad I + CR(i\omega; A)BQB^*[R(i\omega; A)]^* C^* = F^+(\omega)F^-(\omega).$$

The Riccati equation (13) should be regarded only as an heuristic guide to the derivation of (14). The right side of equation (14) is to be regarded as defining (initially) a bounded linear mapping $H_0 \rightarrow H$ of (under our hypotheses) finite dimensional range.

It remains to show, first, that the indicated integral in (14) is convergent, so that the filter gains are well defined. It is further required to show that the filter (12) is stable, and finally that the process generated by the filter is in fact an optimal estimate of the state in the usual sense.

3. Proof of the main result.

LEMMA 1. Suppose that the hypotheses of § 1 hold, and that the finite dimensional operator C has the form

$$(16) \quad C = \sum_{j=1}^p y_j h_j^*.$$

Then the integral

$$(17) \quad [P_\infty C^*] = \frac{1}{2\pi} \int_{-\infty}^{\infty} R(i\omega; A) B Q B^* [R(i\omega; A)]^* C^* [F^-(\omega)]^{-1} d\omega$$

is a strongly convergent operator-valued integral $H_0 \rightarrow H_1$.

Remarks. Since H_0 is a finite dimensional space, the above is essentially a vector H -valued integral. As a practical matter, it is convenient in many cases to evaluate the above as a Pettis integral. In the case of existence of a complete set of eigenvectors associated with an A with compact resolvent, this leads naturally to evaluation of the indicated integral by residues.

Proof. It is required to show the convergence of

$$(18) \quad [P_\infty C^*] \cdot y = \frac{1}{2\pi} \int_{-\infty}^{\infty} R(i\omega; A) B Q B^* [R(i\omega; A)]^* C^* [F^-(\omega)]^{-1} y d\omega,$$

where the H -valued integral on the right is convergent in the Bochner sense.

From the assumption of exponential decay of the semi-group generated by A we obtain the estimates

$$(19) \quad \|R(\cdot; A)v_i\|_{L_2} \leq K_1 \|v_i\|$$

and

$$(20) \quad \|v_i^* [R(\cdot; A)]^* h_j\|_{L_2} \leq K_2 \|v_i\|.$$

Using the spectral representation of BQB^* we obtain the estimates

$$\begin{aligned} & \|R(i\omega; A) B Q B^* [R(i\omega; A)]^* C^* [F^-(\omega)]^{-1} y\| \\ & \leq \sum_{i=1}^{\infty} \lambda_i \|R(i\omega; A)v_i\| \cdot \|v_i^* [R(i\omega; A)]^* C^* [F^-(\omega)]^{-1} y\| \\ & \leq \sum_{i=1}^{\infty} \lambda_i \|R(i\omega; A)v_i\| \sum_{j=1}^p |v_i [R(i\omega; A)]^* h_j y_j^* [F^-(\omega)]^{-1} y| \\ (21) \quad & \leq K \left[\sum_{i=1}^{\infty} \lambda_i \|R(i\omega; A)v_i\| \cdot \sum_{j=1}^p |v_i^* [R(i\omega; A)]^* h_j| \right] \cdot \|y\| \end{aligned}$$

which show integrability of the norm, (using the Schwarz inequality).

LEMMA 2. Let the "optimal feedback gain operator" $[P_\infty C^*]$ be defined as in Lemma 1. Then the closed operator $A = A - [P_\infty C^*]C$ generates an exponentially stable semi-group in H .

Proof. To establish the stability of the semi-group, we consider the abstract differential equation

$$(22) \quad \frac{dx}{dt} = \{A - [P_\infty C^*]C\}x,$$

with $[P_\infty C^*]$ the bounded operator defined by (14) above. From the usual perturbation theory for semi-group generators, it follows that

$$A - [P_\infty C^*]C$$

is a semi-group generator, and that the resolvents satisfy the relation [8]

$$(23) \quad R(s; A - [P_\infty C^*]C) = R(s; A) - R(s; A)[P_\infty C^*]C \cdot R(s; A - [P_\infty C^*]C).$$

Consequently, it follows that for an arbitrary $x_0 \in H$, the unique solution to the Cauchy problem

$$(24) \quad \frac{dx}{dt} = \{A - [P_\infty C^*]C\}x, \quad \lim_{t \rightarrow 0^+} x(t) = x_0$$

satisfies the integral equation

$$(25) \quad x(t) = S_t x_0 - \int_0^t S_{t-\tau} [P_\infty C^*]C x(\tau) d\tau, \quad t > 0,$$

where $\{S_t\}$ is the semi-group generated by A .

The "filter output" therefore satisfies the Volterra integral equation

$$(26) \quad Cx(t) = CS_t x_0 - \int_0^t CS_{t-\tau} [P_\infty C^*]C x(\tau) d\tau, \quad t > 0,$$

in the finite dimensional Hilbert space H_0 .

Assume for the moment that (26) (as will be shown below) has a unique square integrable solution. Since we have assumed strict stability of the semi-group S_t , the existence of a square-integrable solution $[Cx(\cdot)]$ to (26) for each x_0 guarantees square integrability of $x(\cdot)$ defined by (25). By the result of Datko [9], this will establish exponential decay of the semi-group generated by $A - [P_\infty C^*]C$, the desired result.

The equation (26) is a standard Wiener-Hopf equation in a finite dimensional space. By the results of Gohberg and Krein, this has an unique square integrable solution $Cx(\cdot)$ if and only if the index of the Fourier transform

$$I + CR(i\omega; A)[P_\infty C^*]$$

is zero. [Strictly speaking, we refer here to the index of any matrix function obtained from the above finite-dimensional operator through selection of a basis in the range of C .] Of course, the usual means of computing the index in stability problems of this sort is simply an application of the Nyquist criterion [10]. In this case, however, the fact that the index is zero follows directly from the integral representation of the filter gains (14).

By use of (14), the Laplace transform of the convolution kernel takes the form

$$(27) \quad G(s) = CR(s; A) \cdot \frac{1}{2\pi} \int_{-\infty}^{\infty} R(i\omega; A) BQB^* [R(i\omega; A)]^* C^* [F^-(\omega)]^{-1} d\omega.$$

Using the resolvent identity

$$(28) \quad R(s; A)R(i\omega; A) = -\frac{[R(s; A) - R(i\omega; A)]}{s - i\omega}$$

we obtain

$$(29) \quad G(s) = -\frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{C[R(s; A) - R(i\omega; A)]}{s - i\omega} \cdot BQB^* [R(i\omega; A)]^* C^* [F^-(\omega)]^{-1} d\omega.$$

The above is readily recognizable as the usual [11, p. 135] explicit formula for the “frequency domain” calculation of the projection operator P_+ . Using the fact that

$$(30) \quad \phi(\omega) = CR(s; A)BQB^*[R(i\omega; A)]^*[F^-(\omega)]^{-1}$$

under our hypotheses represents the Fourier transform of a (essentially matrix valued) function of negative support, we have

$$G(s) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{CR(i\omega; A)BQB^*[R(i\omega; A)]^*C^*[F^-(\omega)]^{-1}}{s - i\omega} d\omega.$$

This identifies $G(s)$ as the Laplace transform of the function with Fourier transform

$$P_+[CS_{xy}(\omega)[F^-(\omega)]^{-1}].$$

Passing to boundary values to identify the corresponding Fourier transform, we obtain

$$(31) \quad CR(i\omega; A)[P_{\infty}C^*] = P_+[CS_{xy}(\omega)[F^-(\omega)]^{-1}].$$

From the basic spectral factorization

$$(32) \quad I + S_{yy}(\omega) = F^+(\omega)F^-(\omega),$$

it follows that

$$[F^-(\omega)]^{-1} + S_{yy}(\omega)[F^-(\omega)]^{-1} = F^+(\omega).$$

Since (see [7]) the basic property of the spectral factorization is that

$$(33) \quad [F^-(\omega)]^{-1} = I + W^-(\omega),$$

where $W^-(\cdot)$ is the transform of a function of negative support, application of P_+ to (32) produces

$$(34) \quad I + P_+[S_{yy}(\omega)[F^-(\omega)]^{-1}] = F^+(\omega).$$

Combining this with (31) gives the (expected) result

$$(35) \quad I + CR(i\omega; A)[P_{\infty}C^*] = F^+(\omega)$$

from which it follows immediately that

$$I + CR(i\omega; A)[P_{\infty}C^*]$$

has zero index.

This establishes square integrability of the unique solution to (26), and hence exponential stability of the semi-group generated by $A - [P_{\infty}C^*]C$.

LEMMA 3. *Let the “optimal feedback gain operator” $[P_{\infty}C^*]$ be defined as in Lemma 1, and let x^* denote an arbitrary bounded linear functional on the Hilbert space H . Define the (bounded-input-bounded output stable) dynamical system by the forced evolution equation*

$$(36) \quad \frac{dx}{dt} = [A - [P_{\infty}C^*]C]x + [P_{\infty}C^*]u(t)$$

with output $y(t) = x^*x(t)$.

Define the weighting pattern of this system as

$$(37) \quad x^*T(t) = x^*S_*(t)[P_{\infty}C^*], \quad t \geq 0,$$

where $\{S_*(t)\}$ is the semi-group generated by $A - [P_{\infty}C^*]C$.

Then this weighting pattern satisfies the Wiener–Hopf equation

$$(38) \quad x^* T(\tau) + x^* [T^* R_{yy}(\tau)] = x^* R_{xy}(\tau), \quad \tau \geq 0.$$

Proof. From the relation between the resolvent of the perturbed and unperturbed semi-group, we obtain (for $\text{Re}(s) \geq 0$)

$$(39) \quad R(s; A - [P_\infty C^*]C) = R(s; A) - R(s; A)[P_\infty C^*]CR(s; A - [P_\infty C^*]C).$$

From this relation, and the fact established in the proof of Lemma 2 that $I + G(s)$ is the Laplace transform of the function whose Fourier transform is the positive spectral factor in the factorization

$$(40) \quad I + S_{yy}(\omega) = F^-(\omega)F^+(\omega),$$

we obtain the result that the Fourier transform of the weighting pattern is simply

$$x^* R(i\omega; A)[P_\infty C^*][F^+(\omega)]^{-1}.$$

Substitution of this into the transformed version of the Wiener–Hopf equation shows that the weighting pattern satisfies the equation if and only if

$$(41) \quad x^* R(i\omega; A)[P_\infty C^*] = P_+[x^* S_{xy}(\omega)[F^-(\omega)]^{-1}]$$

This equality is easily verified by an argument identical to that used above to establish the equality (31) of Lemma 2.

Combining the results above we obtain the following theorem.

THEOREM. *Consider the distributed Wiener-filtering model given by the Hilbert space process model*

$$(42) \quad dx = Ax \, dt + B \, dw_i(t)$$

and observation process

$$(43) \quad dz = Cx \, dt + dw_0(t)$$

under the hypotheses detailed in § 1 above. Then the feedback gains of the optimal filter are well defined by the integral representation

$$(44) \quad [P_\infty C^*] = \frac{1}{2\pi} \int_{-\infty}^{\infty} R(i\omega; A)BQB^*[R(i\omega; A)]^*C^*[F^-(\omega)]^{-1} \, d\omega.$$

Further, the closed operator $A - [P_\infty C^]C$ generates an exponentially stable semi-group in H , and the Itô equation*

$$(45) \quad d\hat{x} = \{A - [P_\infty C^*]C\}\hat{x} \, dt + [P_\infty C^*] \, dz$$

is a state representation of the optimal filter.

4. An example. The calculation method suggested above is based entirely on “frequency domain” representations of the systems involved. This amounts, of course, to requiring either explicit or accurate approximate calculations of the resolvent of the infinitesimal generator of the system model.

We outline below such a model which arises in the design of a controller for a multi-locomotive powered train. Such trains are subject to coupler failure, due largely to stresses generated by passage over terrain of uneven grade.

In [3] a distributed model was developed for longitudinal motion of such a train, and a mean square stress minimizing controller was developed. Reference [12] discusses a “plane-wave” disturbance model capable of incorporation as a grade force model in the train problem.

The governing equation is

$$(46) \quad d \begin{bmatrix} x_0 \\ v \\ u_x \\ u_t \end{bmatrix} = \begin{bmatrix} A_0 & 0 & 0 & 0 \\ 0 & -V_0 \frac{\partial}{\partial x} & 0 & 0 \\ 0 & 0 & 0 & \frac{\partial}{\partial x} \\ 0 & P & a \frac{\partial}{\partial x} & a \frac{\partial^2}{\partial x^2} \end{bmatrix} \begin{bmatrix} x_0 \\ v \\ u_x \\ u_t \end{bmatrix} dt + \begin{bmatrix} b_0 \\ 0 \\ 0 \\ 0 \end{bmatrix} dw$$

in the Hilbert space $R^n \times L_2[0, 1] \times L_2[0, 1] \times L_2[0, 1]/(0, 0, 0, 1)$. P in the above represents projection onto the subspace orthogonal to the constant function.

The domain of the infinitesimal generator associated with the above is given by

$$(47) \quad D(A) = \left\{ \begin{bmatrix} x_0 \\ v(x) \\ y(x) \\ z(x) \end{bmatrix} \mid v, y, z, \frac{dz}{dx} \text{ absolutely continuous,} \right.$$

$$\left. \frac{dv}{dx}, \frac{dy}{dx}, \frac{d^2z}{dx^2} \in L_2[0, 1], v(0) = c_0' x_0, \text{ and } y + \frac{dz}{dx} = 0 \text{ at } x = 0, 1 \right\}.$$

The interpretation of the above is that it represents the motion of a visco-elastic bar of zero mean velocity, subject to forces of the form

$$h_\omega(t - x/V_0),$$

where $h_\omega(\cdot)$ is a sample function of a stationary stochastic process with power spectral density

$$(48) \quad S_h(\omega) = |c_0'(Ii\omega - A_0)^{-1}b_0|^2.$$

It is a simple (but lengthy) calculation to produce an expression for the resolvent of the above operator, as well as to compute the spectrum. Explicit formulas for this problem in the case of an observation of the "inter locomotive distance" are given in [12].

5. Conclusion. A derivation of the optimal distributed filter for a class of stationary distributed estimation problems has been given. The derivation and resulting computational algorithm are based on an explicit representation for the optimal feedback gains, and are based on a spectral factorization technique.

As well as providing what appears to be a computationally attractive method for certain classes of problems, the derivation also provides simple proofs of stability and optimality of the resulting filter.

REFERENCES

- [1] R. J. CURTAIN, *Estimation theory for evolution equations excited by general white noise*, this Journal, 14 (1976), pp. 1124–1150.
- [2] R. B. VINTER, *Filter stability for stochastic evolution equations*, this Journal, 15 (1977), pp. 465–485.
- [3] J. H. DAVIS AND B. M. BARRY, *A distributed model for stress control in multiple locomotive trains*, Appl. Math. and Optimization, vol. 3, no. 2/3, pp. 163–190.
- [4] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.
- [5] J. C. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 621–634.
- [6] F. STENGER, *The approximate solution of Wiener–Hopf integral equations*, J. Math. Anal. Appl., 37, (1972), pp. 687–724.
- [7] I. C. GOHBERG AND M. G. KREIN, *Systems of integral equations with kernel depending on the difference of the arguments*, Amer. Math. Soc. Transl. 14 (1960), pp. 217–287.
- [8] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semi-groups*, A.M.S. Colloquium Series, vol. 31, American Mathematical Society, Providence, RI, 1957.
- [9] R. DATKO, *Extending a theorem of A. M. Liapunov to Hilbert space*, J. Math. Anal. Appl., 32 (1970), pp. 610–616.
- [10] J. H. DAVIS, *Stability conditions derived from spectral theory: discrete systems with periodic feedback*, this Journal, 10 (1972), pp. 1–13.
- [11] K. HOFFMAN, *Banach Spaces of Analytic Functions*, Prentice-Hall, Englewood Cliffs, NJ, 1962.
- [12] J. H. DAVIS, *Models and filters for travelling wave disturbance problems*, Queen's Math. Preprint, Queen's Univ., Kingston, Canada.

ON THE BOUNDARY BEHAVIOR OF SOLUTIONS TO ELLIPTIC AND PARABOLIC EQUATIONS—WITH APPLICATIONS TO BOUNDARY CONTROL FOR PARABOLIC EQUATIONS*

E. J. P. GEORG SCHMIDT† AND NORBERT WECK‡

Abstract. It is shown that for bounded domains with infinitely differentiable boundaries solutions of second order elliptic equations (involving infinitely differentiable coefficients), and of the associated parabolic equations, cannot vanish together with their normal derivatives on subsets of the boundary having positive measure. This fact has multiple applications to the boundary control problem for parabolic equations. The results obtained have previously required an analyticity assumption on the boundary and on the coefficients.

1. Introduction. In recent papers on boundary control for the heat equation by Fattorini [4] and Glashoff and Weck [6], the authors needed to assume that the data (i.e. the coefficients of the elliptic operator and the boundary of the domain under consideration) were analytic, in order to derive the bang-bang property of solutions to certain optimal control problems. Analyticity was used to justify the application of several results from the theory of partial differential equations. Fattorini needed the Cauchy–Kowalesky theorem, together with a theorem of Lewy ([7] and [8]; see also Garabedian [5]) which allows one to extend solutions of $\Delta u + \lambda u = 0$ across an analytic boundary. Glashoff and Weck used a theorem of Mizohata [10], which states that for solutions of second order parabolic equations the Dirichlet and Neumann data cannot vanish simultaneously on an open subset of the boundary, as well as a result of Tanabe [13] which is also concerned with analyticity.

The purpose of this paper is to point out that the bang-bang property of optimal controls, together with some closely related results on controllability and observability, can be proved also when the data is infinitely differentiable—the condition which is generally assumed for general treatments of the boundary control problem for parabolic equations. This involves applying a theorem of Weck [14] on the boundary values of solutions of elliptic equations (which in turn is based on a deep unique continuation theorem due to Aronszajn, Krzywicki and Szarski [2]), to show that the eigenfunctions of a second order elliptic operator (with self-adjoint boundary condition) cannot vanish together with their normal derivatives on a subset of the boundary having positive measure. This implies a result, related to that of Mizohata, for solutions of parabolic equations. That result has multiple applications to control theory.

2. The boundary behavior of solutions. Let Ω be a bounded open domain in R^n , $\bar{\Omega}$ be the closure of Ω , $\partial\Omega$ be the boundary, $x = (x_1, \dots, x_n)$ denote a point of R^n and $\partial_i = \partial/\partial x_i$. We consider uniformly elliptic operators of the form

$$Lu(x) = \sum_{i,j} \partial_i(a_{ij}(x)\partial_j u(x)) + a(x)u(x),$$

* Received by the editors July 20, 1977.

† Department of Mathematics, McGill University, Montreal, Quebec, Canada H3C 3G1. This author's work was supported by the National Research Council of Canada under Grant A7271, and by an Alexander von Humboldt Research Fellowship.

‡ Fachbereich 6, Mathematik, Universität Essen—Gesamthochschule, Essen, West Germany.

where each of the coefficient functions is in $C^\infty(\bar{\Omega})$, and the matrix $[a_{ij}(x)]$ is symmetric for each x , and satisfies $\sum a_{ij}(x)\xi_i\xi_j \geq c_0|\xi|^2$ for all x in Ω and ξ in $R^n - \{0\}$ (with c_0 a positive constant). At a boundary point x near which the boundary is of class C^1 , the outward pointing unit normal $n(x)$ is well defined. Let $m_i(x) = \sum_j a_{ij}(x)n_j(x)$, and $\partial/\partial m = \sum_j m_j(x)\partial/\partial x_j$. One associates with L the boundary condition $Bu(x) = a(\partial u/\partial m)(x) + bu(x) = 0$ (where a and b are nonnegative constants, not both zero); thus we obtain a self-adjoint operator.

The following theorem was proved by Weck [14], when $a = 0$ and $b = 1$.

THEOREM 2.1. *Let u be a nontrivial solution of $Lu(x) = 0$ in Ω . Suppose that for some open set \mathcal{O} in R^n*

(a) $\Gamma = \mathcal{O} \cap \partial\Omega$ *is an $n - 1$ dimensional C^∞ manifold;*

(b) $Bu(x) = 0$ *on Γ .*

Then

$$N = \left\{ x \in \Gamma \left| \frac{\partial u}{\partial m}(x) = 0 = u(x) \right. \right\} \quad \text{has measure zero.}$$

Proof. This involves only minor additions to the argument in [14]. Using the change of variables described there, one can assume that $\mathcal{O} \cap \partial\Omega$ lies in the half-space $x_n < 0$, that Γ is contained in the hyperplane $x_n = 0$, that for x in Γ and $i < n$, $a_{in}(x) = 0$ while $a_{nn}(x) \neq 0$ and that the boundary condition is $a\partial_n u(x) + bu(x) = 0$. Moreover there is a point x^* in Γ at which all derivatives of u vanish, so that for each integer $n > 0$ there is a positive constant with $|u(x)| \leq C_n |x - x^*|^n$. Let $x' = (x_1, \dots, x_{n-1})$. When $a = 0$ (or $b = 0$) one can extend u across Γ by setting $u(x', x_n) = -u(x', -x_n)$ (or $u(x', x_n) = u(x', -x_n)$). If one then extends the coefficients of L suitably one sees that the extended function u is a solution of an elliptic equation whose principal part has Lipschitz continuous coefficients; x^* is an interior point of the domain and an infinite zero of u . A theorem in [2] then implies that $u \equiv 0$. When both a and b are nonzero one introduces a new function $v(x) = \exp(-(b/a)x_n)u(x)$. This satisfies the boundary condition $\partial_n v(x) = 0$ on Γ , as well as an elliptic equation $\bar{L}v(x) = 0$, where \bar{L} has the same form as L . The previous argument then applies.

We remark that the theorem can also be proved for boundary operators of the form $Bu(x) = a(x)(\partial u/\partial m)(x) + b(x)u(x)$, where $a(x)$ and $b(x)$ are nonnegative infinitely differentiable functions on Γ , with $a(x)^2 + b(x)^2 \neq 0$, and satisfying the condition "for almost every x^0 in Γ there exists a neighborhood N such that either $a(x) \equiv 0$ in N , or $a(x) \neq 0$ in N ."

We have at once

COROLLARY 2.2. *Let Ω be a bounded domain in R^n whose boundary $\partial\Omega$ is a C^∞ manifold. Then for any eigenfunction φ of the self-adjoint operator in $L_2(\Omega)$ associated with L and the boundary condition $Bu \equiv 0$ on $\partial\Omega$ one has that*

$$N = \left\{ x \in \partial\Omega \left| \frac{\partial u}{\partial n}(x) = 0 = u(x) \right. \right\} \quad \text{has measure zero.}$$

This is what we need to prove

COROLLARY 2.3. *Let Ω be a bounded domain in R^n whose boundary $\partial\Omega$ is a C^∞ manifold. Let $u(x, t) \in C^\infty(\bar{\Omega} \times (0, T))$ (with $T > 0$) be a nontrivial solution of*

$$\frac{\partial u}{\partial t}(x, t) = Lu(x, t) \quad \text{for } (x, t) \text{ in } \Omega \times (0, T);$$

$$Bu(x, t) = 0 \quad \text{for } (x, t) \text{ in } \partial\Omega \times (0, T).$$

Then

$$N = \left\{ (x, t) \in \partial\Omega \times (0, T) \mid \frac{\partial u}{\partial n}(x, t) = 0 = u(x, t) \right\} \quad \text{has measure zero.}$$

Proof. We need facts concerning the elliptic eigenvalue problem. It is well known (see Agmon [1]) that one can find a complete orthonormal system $\{\varphi_k\}$ for $L_2(\Omega)$ satisfying

$$\begin{aligned} L\varphi_k(x) &= -\lambda_k\varphi_k(x) \quad \text{for } x \text{ in } \Omega; \\ B\varphi_k(x) &= 0 \quad \text{for } x \text{ in } \partial\Omega. \end{aligned}$$

One has $\lambda_k \geq 0$, and asymptotically, $\lambda_k = O(k^{2/n})$. Moreover each φ_k lies in $C^\infty(\bar{\Omega})$ and one has the estimate $|D\varphi_k(x)| \leq C_r \lambda_k^{m_r}$, where D is any partial derivative of order r and C_r, m_r are positive constants. Let $\{\mu_l\}$ denote the distinct eigenvalues, and $M_l = \{k \mid \lambda_k = \mu_l\}$.

Now suppose that N has positive measure. Pick $\varepsilon > 0$, such that $N_\varepsilon = \{(x, t) \in N \mid t > \varepsilon\}$ has positive measure. It is not difficult to see that for $t > \varepsilon$

$$u(x, t) = \sum_k e^{-\lambda_k(t-\varepsilon)} \langle u(\cdot, \varepsilon), \varphi_k \rangle \varphi_k(x) = \sum_l e^{-\mu_l t} \Psi_l(x),$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product on $L_2(\Omega)$, and $\Psi_l(x) = \sum_{k \in M_l} e^{\lambda_k \varepsilon} \langle u(\cdot, \varepsilon), \varphi_k \rangle \varphi_k(x)$ is an eigenfunction of L satisfying the boundary condition $B\Psi_l = 0$. We shall show that the set $\{x \in \partial\Omega \mid \Psi_l(x) = 0 = (\partial/\partial n)\Psi_l(x)\}$ has positive measure; it then follows from Corollary 2.2 that $\Psi_l = 0$, and consequently that $u \equiv 0$, contradicting the assumption that u is a nontrivial solution and thus completing the proof. Let S be the set of x in $\partial\Omega$ such that the corresponding section $\{t \in (\varepsilon, T) \mid (x, t) \in N\}$ of N_ε has positive measure. By Fubini's theorem S itself has positive measure. For fixed x in S , $\sum_l e^{-\mu_l t} \Psi_l(x)$ then vanishes on a subset of (ε, T) having positive measure; well known facts on Dirichlet series then imply that $\Psi_l(x) = 0$ for each l . Thus each Ψ_l vanishes on S , and one sees similarly that the normal derivatives also vanish on S . Since S has positive measure the proof is complete.

We remark that a similar method of proof was used in MacCamy, Mizel and Seidman [9].

We note finally that Corollary 2.3, together with the change of variable $t \rightarrow T - t$, yields the corresponding result for solutions of the adjoint equation

$$\begin{aligned} \frac{\partial w}{\partial t}(x, t) + Lw(x, t) &= 0 \quad \text{in } \Omega \times (0, T); \\ Bw(x, t) &= 0 \quad \text{in } \partial\Omega \times (0, T). \end{aligned}$$

This is what is needed in control theory.

3. Applications to control theory. Let L be an elliptic operator satisfying the conditions of the previous section, and set b (the constant occurring in B) equal to 1. Then one can show that given u_0 in $L_2(\Omega)$ and f in $U^\infty = L_\infty(\partial\Omega \times (0, \infty))$ the following parabolic initial boundary value problem has a unique weak solution:

$$\begin{aligned} \frac{\partial u}{\partial t}(x, t) &= Lu(x, t) \quad \text{for } (x, t) \text{ in } \Omega \times (0, \infty), \\ Bu(x, t) &= f(x, t) \quad \text{for } (x, t) \text{ in } \partial\Omega \times (0, \infty), \\ u(x, 0) &= u_0(x) \quad \text{for } x \text{ in } \Omega. \end{aligned}$$

That solution has the form

$$u(x, t) = V_t u_0(x) + S_t f(x),$$

where $\{V_t\}_{t \geq 0}$ is an analytic semi-group on $L_2(\Omega)$, with

$$V_t u_0(x) = \int_{\Omega} G(x, y; t) u_0(y) dy,$$

and $\{S_t\}_{t \geq 0}$ is a family of continuous linear operators from U^∞ (with the weak* topology) to $L_2(\Omega)$, given by

$$S_t f(x) = \int_{\partial\Omega} \int_0^t G^\partial(x, y; t-s) f(y, s) ds dS_y;$$

here

$$G(x, y; t) = \sum_k e^{-\lambda_k t} \varphi_k(x) \varphi_k(y),$$

and

$$G^\partial(x, y; t) = \sum_k e^{-\lambda_k t} \varphi_k(x) \varphi_k^\partial(y),$$

with $\varphi^\partial = (a^2 + 1)^{-1}(a\varphi - \partial\varphi/\partial m)$. When $a > 0$, $L_2(\Omega)$ can be replaced by $C(\bar{\Omega})$. This can all be proved along the lines of Fattorini [4], Glashoff/Weck [6] or the unified treatment in Schmidt [11]. With u_0 kept fixed it is convenient to denote the above solution by u_f . The boundary control f may be restricted to certain subclasses of U^∞ . Let M be a positive real number, and P be a measurable subset of $\partial\Omega \times (0, \infty)$. Then define

$$\begin{aligned} U^M &= \{f \in U^\infty \mid f(x, t) \leq M \text{ a.e.}\}; \\ U_P^\infty &= \{f \in U^\infty \mid f(x, t) = 0 \text{ a.e. outside } P\}; \\ U_P^M &= U^M \cap U_P^\infty. \end{aligned}$$

We now sketch the proofs of several results which are well known in case the coefficients of L and the boundary $\partial\Omega$ are analytic. For $U \subset U^\infty$ and $T > 0$, let

$$R_T(u_0; U) = \{u_f(\cdot, T) \mid f \in U\}.$$

THEOREM 3.1. *Suppose $P \subset \partial\Omega \times (0, T)$ has positive measure. Then $R_T(u_0; U_P^\infty)$ is dense in $L_2(\Omega)$ (in $C(\bar{\Omega})$ if $a > 0$).*

Proof. Suppose not. Then $R_T(0; U_P^\infty) = \{S_T f \mid f \in U_P^\infty\}$ likewise is not dense in $L_2(\Omega)$. Then there is a function v in $L_2(\Omega)$, $v \neq 0$, such that for each f in U_P^∞

$$\langle S_T^* v, f \rangle = \langle v, S_T f \rangle = 0.$$

Now one can verify, for the adjoint operator S_T^* , that

$$S_T^* v(x, t) = [V_{T-t} v]^\partial(x).$$

Thus, letting $w(x, t) = V_{T-t} v(x)$, we have that

$$\int_{\partial\Omega} \int_0^T w(x, t)^\partial f(x, t) dt dS_x = 0, \quad \text{for each } f \text{ in } U_P^\infty,$$

and hence that $w(x, t)^{\partial} = 0$ almost everywhere in P . But $w(x, t)$ is a solution of

$$\begin{aligned}\frac{\partial w}{\partial t} + Lw &= 0 \quad \text{in } \Omega \times (0, T) \\ Bw &= 0 \quad \text{in } \partial\Omega \times (0, T).\end{aligned}$$

Now $Bw = 0$ and $w(x, t)^{\partial} = 0$ on P imply together that $(\partial u / \partial n)(x, t) = 0 = u(x, t)$ on P . From the last remark of § 2 it then follows that $w \equiv 0$, contradicting the assumption that $v \neq 0$.

In the special case that $a > 0$ and $L_2(\Omega)$ is replaced by $C(\bar{\Omega})$, v is replaced by a measure of $\bar{\Omega}$, but the argument is otherwise unchanged.

THEOREM 3.2. *Let $\delta > 0$ be specified and u_1 be a given target function in $L_2(\Omega)$. Suppose that there exists $T > 0$ and f in U_P^M such that $\|u_f(\cdot, T) - u_1\| \leq \delta$. Let*

$$T^* = \inf \{T > 0 \mid \text{there exists } f \text{ in } U_P^M \text{ with } \|u_f(\cdot, T) - u_1\| \leq \delta\}.$$

Then there exists a unique f^ in U_P^M , such that $\|u_{f^*}(\cdot, T^*) - u_1\| = \delta$, and moreover, $|f^*(x, t)| = M$ almost everywhere in P .*

Proof. One can assume that P has positive measure. The existence of a minimizing control f^* is standard and depends on the continuity properties of the operators S_t . That $|f^*(x, t)| = M$ almost everywhere in P (from which the uniqueness also follows) is more subtle. Note that f^* is also a solution of the minimization problem: minimize $\|u_f(\cdot, T^*) - u_1\|$ for f in U_P^M . Hence applying a separation argument to the open ball of radius δ about u_1 and to the compact convex set $R_{T^*}(u_0; U_P^M)$, one finds v in $L_2(\Omega)$, $v \neq 0$, such that

$$\langle v, u_{f^*}(\cdot, T^*) \rangle \geq \langle v, u_f(\cdot, T^*) \rangle, \quad \text{for each } f \text{ in } U_P^M.$$

Consequently

$$\langle S_{T^*} v, (f^* - f) \rangle \geq 0 \quad \text{for each } f \text{ in } U_P^M.$$

If we let $w(x, t) = V_{T-t} v(x)$, this becomes

$$\int_{\partial\Omega} \int_0^T w(x, t)^{\partial} (f^*(x, t) - f(x, t)) dt dS_x \geq 0, \quad \text{for any } f \text{ in } U_P^M.$$

Since $w(x, t)$ is a nontrivial solution of the adjoint parabolic equation with $Bw \equiv 0$, it follows that $w(x, t)^{\partial} \neq 0$ almost everywhere in P , so that $f^*(x, t) = M \operatorname{sgn} w(x, t)^{\partial}$, from which the result follows.

We remark that this theorem also holds in $C(\bar{\Omega})$, if $a > 0$.

For the exact time optimal problem (i.e. the problem treated in Theorem 3.2, but with $\delta = 0$) the bang-bang property can also be proved under certain circumstances. Such a result is due to Fattorini [4], for the heat equation with an analytic boundary. We do not recall the details but simply state.

PROPOSITION 3.3. *Theorem 4-1 of Fattorini [4] holds if the boundary is infinitely differentiable.*

Finally we make a few comments concerning observability for parabolic equations (see also Dolecki and Russell [3] and Seidman [12]). This involves reconstructing, preferably in a continuous way, the solution of a parabolic equation from knowledge of the Dirichlet and Neumann boundary values on all or part of the boundary. Consider the following version of this problem. Let $u(x, t) = V_t u_0(x)$ denote the solution with u_0 given, $f \equiv 0$, and (for convenience) $a > 0$. From Corollary 2.3 it follows at once that the map $u_0 \rightarrow u(x, t)|_P$ is injective. Thus $u_0(x)$, and hence $u(x, t)$,

are uniquely determined by knowledge of $u(x, t)$ on any subset of $\partial\Omega \times (0, \infty)$ having positive measure. It would be nice to know that the map $u(x, t)|_P \rightarrow u(x, T)$ (with $T > 0$, fixed) was continuous. This seems difficult, as does the related question concerning exact controllability with controls in U_P^∞ .

Acknowledgment. E. J. P. Georg Schmidt acknowledges the hospitality of the Institute for Applied Mathematics and Statistics at the University of Würzburg.

REFERENCES

- [1] S. AGMON, *Lectures on Elliptic Boundary Value Problems*, Van Nostrand, Princeton, NJ, 1965.
- [2] N. ARONSZAJN, A. KRZYWICKI AND J. SZARSKI, *A unique continuation theorem for exterior differential forms on Riemannian manifolds*, Ark. Mat., 4 (1962), pp. 417–453.
- [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 boundary control of the heat equation*, Calculus of Variations and Control Theory, D. L. Russell, ed., Academic Press, New York, 1976, pp. 305–320.
- [5] P. R. GARABEDIAN, *Analyticity and reflection for plane elliptic systems*, Comm. Pure Appl. Math., 14 (1961), pp. 315–322.
- [6] K. GLASHOFF AND N. WECK, *Boundary control of parabolic differential equations in arbitrary dimensions: Supremum norm problems*, this Journal, 14 (1976), pp. 662–681.
- [7] H. LEWY, *Neuer Beweis des analytischen Charakters der Lösungen Differentialgleichungen*, Math. Ann., 101 (1929), pp. 601–619.
- [8] ———, *On the reflection laws of second order differential equations in two independent variables*, Bull. Amer. Math. Soc., 65 (1959), pp. 37–58.
- [9] R. C. MACCAMY, V. J. MIZEL AND T. I. SEIDMAN, *Approximate boundary controllability for the heat equation*, J. Math. Anal. Appl., 23 (1968), pp. 699–703.
- [10] S. MIZOHATA, *Unicité du prolongement des solutions pour quelques opérateurs différentiels paraboliques*, Mem. Coll. Sci. Kyoto Univ., A31 (1958), pp. 219–239.
- [11] G. SCHMIDT, *Boundary control for the heat equation with stationary targets*, submitted.
- [12] T. I. SEIDMAN, *A well-posed problem for the heat equation*, Bull. Amer. Math. Soc., 80 (1974), pp. 901–902.
- [13] H. TANABE, *On differentiability and analyticity of solutions of weighted elliptic boundary value problems*, Osaka J. Math., 2 (1965), pp. 163–190.
- [14] N. WECK, *Über das Prinzip der eindeutigen Fortsetzbarkeit in der Kontrolltheorie*, Optimization and Optimal Control, Lecture Notes in Mathematics 477, Springer-Verlag, Berlin, 1975, pp. 276–284.

FUNCTION SPACE CONTROLLABILITY OF LINEAR RETARDED SYSTEMS: A DERIVATION FROM ABSTRACT OPERATOR CONDITIONS*

A. MANITIUS† AND R. TRIGGIANI‡

Abstract. Controllability of linear retarded systems is investigated by using the abstract representation of such systems given by $\dot{x} = \tilde{A}x + \tilde{B}u$, where x belongs to Hilbert space $R^n \times L_2([-h, 0], R^n)$ denoted as M_2 , and \tilde{A} generates a C_0 -semigroup. It is shown that useful, practically verifiable conditions can be obtained by this approach.

The following problems are investigated: approximate controllability in the space M_2 and its subspace, L_2 , exact Euclidean (R^n) controllability, spectral controllability, feedback stabilizability and a relation between pointwise degeneracy and function space controllability. Starting from the abstract functional analytic framework, the analysis is carried down to the matrix theory level, through the crucial intermediate role of the theory of entire functions.

1. Introduction and preliminaries. Recent years have witnessed a good deal of research focused on abstract linear control systems, modeled by differential equations in a Banach space, within the framework of the theory of strongly continuous semigroups of bounded linear operators. One good reason for using such models is that they make it possible to treat, in a mathematically unified manner, a variety of different dynamical systems, including those governed by integrodifferential equations (IDE), partial differential equations (PDE) of both parabolic and hyperbolic type, as well as functional differential equations (FDE). This approach, however, could appear to be sterile unless it produced, for whatever problem under investigation, "useful" conditions in terms of the operators entering the abstract model.

We label a condition "useful" when, once specialized to particular classes of operators arising in physically significant dynamical systems, it yields verifiable tests. The translation from abstract results into verifiable results for concrete classes of dynamical systems is a point we wish to stress with a particular strength.

Regarding the controllability problem of such abstract systems, general results for approximate controllability extending the familiar finite dimensional rank condition are given in [61], [64] while results of negative type establishing lack of exact controllability in finite time are given in [61], [63], [66]. However, successful applications and/or illustrations of such general criteria for approximate controllability have been so far done only for some classes of IDE [61] and PDE of both parabolic and hyperbolic type [64], but not yet for FDE.

The present paper is intended to fill this gap and to carry out, for a class of retarded FDE, an analysis of controllability properties leading from general abstract conditions to practically verifiable conditions based on the matrices defining the original FDE.

The class of retarded FDE systems under study is described by

$$(1.1) \quad \dot{y}(t) = A_0 y(t) + A_1 y(t-h) + Bu(t)$$

* Received by the editors August 3, 1976, and in revised form July 14, 1977.

† Centre de Recherches Mathématiques, Université de Montréal, Montréal, Québec, Canada H3C 3J7. This research has been supported in part by the National Research Council of Canada under Grant A9240 (1975-76) and in part by the Ministère de l'Éducation du Gouvernement du Québec under Grant FCAC 1975-76.

‡ Department of Mathematics, Iowa State University, Ames, Iowa 50010. This research has been supported in part by the U.S. Air Force under Grant AFOSR-76-3038.

where $y \in R^n$, $u \in R^m$, A_0 and A_1 are $n \times n$ matrices, B is an $n \times m$ matrix. This relatively simple model was chosen to preserve clarity of exposition of the main ideas, but the methods used in the paper apply equally well to systems with several commensurate delays. An extension of our results to more general retarded systems is described in [45].

As is known, e.g. [9], [12], [4], [7], the equation (1.1) can be transformed into an abstract equation

$$(1.2) \quad \dot{x} = \tilde{A}x + \tilde{B}u$$

where x belongs to the Hilbert space $R^n \times L_2([-h, 0], R^n)$ denoted shortly as M_2 , or $M_2([-h, 0], R^n)$. The operator \tilde{A} is the infinitesimal generator of a strongly continuous semigroup of bounded linear operators, and \tilde{B} is a suitable bounded operator from R^m into M_2 . Details of this representation will follow later. Here we wish only to point out that if $y(t)$ denotes a solution of (1.1) with the standard notation [21] $y_t = y_t(\theta) = y(t + \theta)$, $\theta \in [-h, 0]$, and if $x(t)$ denotes a solution of (1.2), with $x = (x^0, x^1)$, $x^0 \in R^n$, $x^1 \in L_2([-h, 0], R^n)$ [12], then y_t and $x(t)$ are related by $x^0(t) = y(t) = y_t(0)$, $x^1(t)(\theta) = y(t + \theta)$, $\theta \in [-h, 0]$. Consequently, the task undertaken in this paper is to consider some controllability properties for the system (1.1) that can be analyzed by general methods applicable to (1.2) and then use the available results for systems of class (1.2) to obtain some verifiable conditions involving parameters of A_0 , A_1 , B and h of (1.1).

To date, nearly all of the work on controllability of FDE's concentrated on (1.1). Summaries of earlier work on controllability of retarded and/or neutral systems are contained in [5], [25], [39]. The concept of M_2 -approximate controllability in connection with the abstract equation analogous to (1.2) was stated in [12]. Zmood [69] analyzed the equation (1.1) directly, and derived some algebraic tests for L_2 -approximate controllability; the test obtained [69, Thm. 3.4.1] appears to be manageable only in the special case $A_0 = 0$; otherwise it is complicated to the extent that "even for quite a modest problem the matrices that have to be handled soon become enormous" [69, p. 92], and even simple examples with $n = m = 2$ are not practically solvable [69, Example 3.5.2]. Popov [57] derived a criterion for exact controllability to C^r functions on the interval $[-h + \varepsilon, 0]$ for any $\varepsilon > 0$ (a concept called C^r -reachability; see Remark 2.1); his arguments were an important factor in our investigations; it is, however, not clear a priori under what conditions the C^r reachability is preserved in the limit $\varepsilon \rightarrow 0$. A similar approach involving controllability to smooth functions on the interval $[-h + \varepsilon, -\varepsilon]$ was described by Korytowski [29], who obtained a new algebraic condition. Banks et al. [5], Kurcyusz and Olbrot [30], Jacobs and Langenhof [25] derived criteria for controllability in the space $W_2^{(1)}([-h, 0], R^n)$ for retarded and neutral systems respectively; work in the same direction was done by Pandolfi [55].¹ The controllability in eigensubspaces of the operator \tilde{A} was discussed by Osipov [53], Banks and Manitius [6] and others: recently, Pandolfi [54] and also Bhat and Koivo [8] derived a criterion for this type of controllability (an improvement of their criterion is contained in this paper). In spite of all the research effort mentioned above, to our knowledge there has been no significant attempt to obtain the controllability conditions directly from the abstract representation (1.2).

The price to pay for the mathematically unifying treatment of FDE's within the framework of semigroups of operators is that the system (1.2) can never be exactly

¹ In a recent paper [51] Olbrot discussed the relations between various approximate controllability concepts.

controllable in the space chosen (i.e. M_2 or even L_2) (see [63] and also [12]). Thus we are left with the concept of approximate controllability. However, this concept turns out to be satisfactory for many significant control problems in which the controllability plays a key role, among them for feedback stabilizability with an arbitrary preassigned exponential decay rate [14], [54] and for the infinite time linear-quadratic optimal control problem [14]. As a matter of fact, we believe that our results contain explicitly verifiable conditions for the solvability of both problems mentioned above.

Notation and terminology used in the paper are as follows. The letter \mathbb{C} denotes the set of complex numbers. R^n denotes the Euclidean n -dimensional space. The space M_2 (see also [12]) used in the paper is an abbreviation for the product space $R^n \times L_2([-h, 0], R^n)$, that is the space of pairs $(x^0, x^1) = x$, $x^0 \in R^n$, $x^1 \in L_2([-h, 0], R^n)$ endowed with the norm

$$(1.3) \quad \|x\|_{M_2} = (\|x^0\|_{R^n}^2 + \|x^1\|_{L_2}^2)^{1/2}$$

and the inner product

$$(1.4) \quad \langle x, z \rangle_{M_2} = x^{0T} z^0 + \int_{-h}^0 x^{1T}(\theta) z(\theta) d\theta$$

where the letter T denotes vector transposition. Symbols Π_0 and Π_1 will denote the projections of M_2 onto its component spaces R^n and $L_2([-h, 0], R^n)$, respectively. Thus $x^i = \Pi_i x$, $i = 0, 1$. Occasionally we will also use the Sobolev space $W_2^{(1)}([-h, 0], R^n)$ defined in the same way as in [4] or [25].

The operators \tilde{A} and \tilde{B} appearing in (1.2) are as follows (see e.g. [9], [4], [7]). The domain $D(\tilde{A}) \subset M_2$ is characterized by

$$(1.5) \quad D(\tilde{A}) = \{x \in M_2 | x^1 \in W_2^{(1)}([-h, 0], R^n) \text{ and } x^1(0) = x^0\}$$

and

$$(1.6) \quad \tilde{A}x = \left(A_0 x^0 + A_1 x^1(-h), \frac{d}{d\theta} x^1 \right),$$

$$(1.7) \quad \tilde{B}u = (Bu, 0).$$

The operator \tilde{A} is the infinitesimal generator of the strongly continuous semigroup $S(t)$, $t \geq 0$, on M_2 , which is defined by $(y(t), y_t) = S(t)(y(0), y_0)$, where $y(t)$ is the solution of (1.1) for $u = 0$. The semigroup $S(t)$ is differentiable and compact for $t \geq h$. More details on the use of M_2 space for FDE's can be found in [14], while the relationships between the FDE's and the semigroup theory in the space M_2 are discussed in [4], [9], [7], [67]. The original account on linear FDE's within the framework of strongly continuous semigroups in the space $C([-h, 0], R^n)$ is given in the monograph by Hale [21].

The unique strict solution of (1.2) initiating at $x = 0$ and $t = 0$ is

$$(1.8) \quad x(t, u) = \int_0^t S(t-\tau) \tilde{B}u(\tau) d\tau$$

at least when $u(\cdot)$ is smooth, say C^1 [28, p. 486]. For any $u(\cdot) \in L_{loc}^1$ we shall refer to the well defined Bochner integral at the right hand side of (1.8) as the *mild solution* of (1.2). See [4] for details on the connection between (1.2) and the solution of (1.1).

The symbol K_t will denote the set of attainability at time t , generated by (1.8) using either C^1 controls for the strict solution, or L_{loc}^1 controls for the mild solution.

By M_2 -approximate controllability at time t (respectively, in finite time) of (1.2)—or, equivalently, of (1.1)—we mean

$$(1.9) \quad \bar{K}_t = M_2([-h, 0], R^n) \quad \left(\text{respectively } \overline{\bigcup_{t>0} K_t} = M_2([-h, 0], R^n) \right)$$

where the closure is taken in the M_2 -topology.

Similarly, by L_2 -approximate controllability at time t (respectively, in finite time) of (1.2)—or, equivalently, of (1.1)—we mean

$$(1.10) \quad \overline{\Pi_1 K_t} = L_2([-h, 0], R^n) \quad \left(\text{respectively } \overline{\bigcup_{t>0} \Pi_1 K_t} = L_2([-h, 0], R^n) \right)$$

where the closure is taken in the L_2 -topology. Note that according to recent results [63], neither of the relations (1.9), (1.10) can be satisfied without the sign of closure, even for the mild solutions, since the operator \tilde{B} is compact.

By Euclidean controllability of (1.2) at time t —or, equivalently of (1.1)—we mean

$$\Pi_0 K_t = R^n.$$

Note that in the literature on finite dimensional systems these concepts of “controllability” would often be referred to as “reachability”.

We remark that according to recent results [5, Cor. 5.1] the set of attainability for (1.1) increases with t for $t \leq nh$ and remains constant thereafter. Consequently, the concepts of approximate (or Euclidean) controllability at time t and at finite time are equivalent for $t \geq nh$. We will therefore assume that $t \geq nh$ and refer succinctly to M_2 or L_2 approximate controllability and to Euclidean (or R^n) controllability.

The main results of this paper are: 1) criteria for M_2 - and L_2 -approximate controllability stated in terms of the system transfer function and a certain finite Laplace transform (Propositions 2.5, 2.6); 2) necessary conditions for L_2 - (and M_2 -) approximate controllability stated in terms of the rank of a certain polynomial matrix $P(\lambda)$, which is easily computable from the original system matrices (Theorem 3.1); 3) criteria for L_2 - and M_2 -approximate controllability that reduce to a question on whether a system of linear homogeneous equations has a nonzero solution (Corollary 4.6, Theorem 4.8); 4) sufficient conditions for M_2 -approximate controllability for all delays $h > 0$, stated directly in terms of the original system matrices (Theorems 5.3, 5.4); 5) simple criteria for spectral controllability and feedback stabilizability, which do not require knowledge of eigenvalues of \tilde{A} (Theorem 7.2, Remark 7.1); 7) an explanation of the role of pointwise degeneracy in the function space controllability problem (Theorem 8.1). The results are illustrated by several examples. We also point out relations between our results and those on Euclidean controllability and some other algebraic controllability conditions (Propositions 6.1, 6.3, Remark 6.1).

Further extensions of these results are given in [44], [45], [46]. Some of the details that were omitted in this text can be found in [43].

We remark that the results of this paper apply also to the observability problem. A known result on the duality between observability and approximate controllability (see e.g. [64]) says that the pair (\tilde{A}, \tilde{B}) is approximately controllable if and only if the pair $(\tilde{A}^*, \tilde{B}^*)$ (* denoting the dual operators) is observable. For a detailed discussion of observability and duality for retarded systems see [45].

2. Functional characterization of M_2 - and L_2 -approximate controllability. On the basis of the definitions given in § 1 it is obvious that the M_2 -approximate controllability implies both the R^n -controllability and the L_2 -approximate controllability.

That the converse is not true will be shown by examples in § 4 (Example 4.3). Also a doubt about whether the class of M_2 -approximately controllable systems is not empty will be dissipated by examples in this and following sections.

The general characterization of approximate controllability for abstract systems in Banach space, as applied to (1.2)–(1.7), is as follows:

THEOREM 2.1 ([17, Prop. 2.3]; [61, Thm. 3.1.1]). *The following statements are equivalent:*

$$(2.1) \quad \begin{aligned} & \text{(i)} \quad \overline{\bigcup_{0 < t} K_t} = M_2([-h, 0], R^n); \\ & \text{(ii)} \quad \begin{cases} \eta \in M_2, & \langle \eta, S(t)\tilde{B}U \rangle_{M_2} = 0, \quad \forall t \geq 0 \\ \Rightarrow \eta = 0; \end{cases} \\ & \text{(iii)} \quad \begin{cases} \eta \in M_2, & \langle \eta, R(\lambda, \tilde{A})\tilde{B}U \rangle_{M_2} = 0, \quad \forall \lambda \in \rho(\tilde{A}) \\ \Rightarrow \eta = 0; \end{cases} \\ & \text{(iv)} \quad \begin{cases} \eta \in M_2, & \langle \eta, R^k(\lambda_0, \tilde{A})\tilde{B}U \rangle_{M_2} = 0, \quad k = 0, 1, \dots, \lambda_0 \in \rho(\tilde{A}) \\ \Rightarrow \eta = 0; \end{cases} \end{aligned}$$

where $R(\cdot, \tilde{A})$ is the resolvent operator, $\rho(\tilde{A})$ the resolvent set of \tilde{A} , λ_0 a fixed arbitrary point in $\rho(\tilde{A})$ and $\tilde{B}U$ is the range of \tilde{B} .

A sketchy explanation follows: equivalence (i) \Leftrightarrow (ii) is a straightforward consequence of the solution formula (1.8), via a corollary of the Hahn–Banach theorem; equivalence (ii) \Leftrightarrow (iii) relies on the fundamental fact that the resolvent operator is the strong Laplace transform of the semigroup for

$$\operatorname{Re} \lambda \geq \omega_0 \stackrel{\text{def}}{=} \lim_{t \rightarrow \infty} \frac{\ln \|S(t)\|}{t} < +\infty$$

plus an analytic continuation argument carried over $\rho(\tilde{A})$ (we use here the known fact that the spectrum $\sigma(\tilde{A})$ of \tilde{A} is given by a countable set of eigenvalues [21]); equivalence (iii) \Leftrightarrow (iv) is an immediate consequence of the fact that the resolvent operator is an analytic function in $\rho(\tilde{A})$ and its derivatives are [60, p. 257]

$$(2.2) \quad \frac{d^k R(\lambda, \tilde{A})}{d\lambda^k} = (-1)^k k! R^{k+1}(\lambda, \tilde{A}), \quad k = 0, 1, \dots$$

Condition (iv) can be written in an equivalent way, reminiscent of the classical finite dimensional rank condition, as

$$(2.3) \quad \overline{\operatorname{span} \{R^k(\lambda_0, \tilde{A})\tilde{B}U, k = 0, 1, \dots\}} = M_2([-h, 0], R^n)$$

where the closure over the ‘span’ sign is taken in the M_2 -topology.

Similarly, for L_2 approximate controllability we have

COROLLARY 2.2. *The following statements are equivalent:*

$$(2.4) \quad \begin{aligned} & \text{(i)} \quad \overline{\bigcup_{0 < t} \Pi_1 K_t} = L_2([-h, 0], R^n); \\ & \text{(ii)} \quad \begin{cases} \eta \in L_2, & \langle \eta, \Pi_1 S(t)\tilde{B}U \rangle_{L_2} = 0, \quad \forall t \geq 0 \\ \Rightarrow \eta = 0; \end{cases} \\ & \text{(iii)} \quad \begin{cases} \eta \in L_2, & \langle \eta, \Pi_1 R(\lambda, \tilde{A})\tilde{B}U \rangle_{L_2} = 0, \quad \forall \lambda \in \rho(\tilde{A}) \\ \Rightarrow \eta = 0; \end{cases} \\ & \text{(iv)} \quad \begin{cases} \eta \in L_2, & \langle \eta, \Pi_1 R^k(\lambda_0, \tilde{A})\tilde{B}U \rangle_{L_2} = 0, \quad k = 0, 1, \dots, \lambda_0 \in \rho(\tilde{A}) \\ \Rightarrow \eta = 0; \end{cases} \end{aligned}$$

where, as before, λ_0 is a fixed arbitrary point in $\rho(\tilde{A})$. The analogue of (2.3) for L_2 -approximate controllability is now

$$(2.5) \quad \overline{\text{span}} \{ \Pi_1 R^k(\lambda_0, \tilde{A}) \tilde{B}U, k = 0, 1, \dots \} = L_2([-h, 0], R^n)$$

where the closure over the 'span' sign is taken in the L_2 -topology.

The present paper uses (2.3) and (2.5) as starting points for investigating the M_2 - and, respectively, L_2 -approximate controllability of (1.2), i.e. of (1.1). There are indeed [64] characterizations of approximate controllability in Banach spaces *directly* in terms of the operators appearing in the differential equations; i.e. the operator acting on the state (as opposed to its resolvent) as well as the operator acting on the control, but these apply [64, Remark 2.3] when either the semigroup is analytic for $t > 0$ (as for parabolic PDE) or the semigroup is a group (as for hyperbolic PDE). Either of these two conditions fails in the case of retarded FDE. Moreover, a general sufficient (only) condition [64, Thm. 2.1] with no extra assumption on the semigroup fails to hold for the operators \tilde{A} and \tilde{B} defined by (1.5)–(1.7). Thus, one has to resort to the characterizations involving the resolvent operator.

A natural program to investigate, say, the L_2 -approximate controllability would, therefore, be the following: (i) compute $\Pi_1 R(\lambda_0, \tilde{A}) \tilde{B}U$ at some arbitrarily chosen, convenient point $\lambda_0 \in \rho(\tilde{A})$; (ii) compute either the powers $\Pi_1 R^k(\lambda_0, \tilde{A}) \tilde{B}U$, $k = 0, 1, 2, \dots$ or, the derivatives

$$\left. \frac{d^k \Pi_1 R(\lambda, \tilde{A}) \tilde{B}U}{d\lambda^k} \right|_{\lambda=\lambda_0};$$

(iii) test for completeness in $L_2[-h, 0], R^n$ of either the powers or—equivalently by (2.2)—of the derivatives in terms of the original matrices A_0, A_1, B , and the delay h .

It is worth recalling that a similar program for normal operators with compact resolvent was employed successfully in [64], applications including parabolic and hyperbolic PDE on finite spatial domains and classical boundary conditions. In the present case, however, the outlined program runs into serious difficulties at stage (iii). To see this, compute the resolvent $R(\lambda, \tilde{A})$. Let $(f(0), f(\cdot)) \in D(\tilde{A})$, $\lambda \in \rho(\tilde{A})$, and $(g^0, g^1) \in M_2([-h, 0], R^n)$. Consider the relation $(\lambda I - \tilde{A})(f(0), f(\theta)) = (g^0, g^1)$, i.e. $(f(0), f(\theta)) = R(\lambda, \tilde{A})(g^0, g^1)$ $\theta \in [-h, 0]$, $\lambda \in \rho(\tilde{A})$. Using the definition of \tilde{A} given by (1.6) yields a differential equation which, once integrated, gives, after some standard computations

$$(2.6) \quad [\Pi_1 R(\lambda, \tilde{A})(g^0, g^1)](\theta) = f(\theta) = e^{\lambda\theta} \Delta^{-1}(\lambda) \left[g^0 + A_1 \int_{-h}^0 e^{\lambda(-h-s)} g^1(s) ds \right] - \int_0^\theta e^{\lambda(\theta-s)} g^1(s) ds$$

and

$$(2.7) \quad [\Pi_0 R(\lambda, \tilde{A})(g^0, g^1)] = f(0) = \Delta^{-1}(\lambda) \left[g^0 + A_1 \int_{-h}^0 e^{\lambda(-h-s)} g^1(s) ds \right]$$

where

$$(2.8) \quad \Delta(\lambda) = (\lambda I - A_0 - e^{-\lambda h} A_1).$$

In view of the definition (1.7) of $\tilde{B}u$, we finally obtain from (2.6) and (2.7)

$$(2.9) \quad [\Pi_1 R(\lambda, \tilde{A}) \tilde{B}u](\theta) = e^{\lambda\theta} \Delta^{-1}(\lambda) Bu,$$

$$(2.10) \quad [\Pi_0 R(\lambda, \tilde{A}) \tilde{B}u] = \Delta^{-1}(\lambda) Bu.$$

Now, successive differentiations of (2.9) in λ require computation of derivatives of $\Delta^{-1}(\lambda)$, which quickly become very complicated functions of A_0, A_1, B . For this reason formula (2.9) appears to be unsuitable for the purpose of deriving criteria for completeness in $L_2([-h, 0], R^n)$ of $d^k \Pi_1 R(\lambda_0, \tilde{A}) \tilde{B} u / d\lambda^k$, $k = 0, 1, 2, \dots, \lambda_0 \in \rho(\tilde{A})$.

We therefore resort to an alternative but equivalent route. Instead of using (2.4) (iv) ((2.1)(iv) respectively) which led to (2.12), we employ its equivalent formulation (2.4) (iii) ((2.1) (iii) respectively). In order to do this, we introduce the following:

DEFINITION 2.3. An n -vector valued function $q(\lambda)$ is said to be of class $\text{FLT}_2([0, h], R^n)$ in case

$$q(\lambda) = \int_0^h e^{-\lambda s} f(s) ds$$

where $f \in L_2([0, h], R^n)$ and $\lambda \in \mathbb{C}$ (complex plane). In other words, $q(\lambda)$ is a finite Laplace transform (FLT) of an n -vector valued L_2 -function, over $[0, h]$.

Note the following properties of $q(\lambda)$ of which we will make frequent use.

PROPOSITION 2.4. (i) $q(\lambda)$ can be alternatively written as

$$q(\lambda) = \int_{-h}^0 e^{\lambda \theta} \eta(\theta) d\theta, \quad \eta \in L_2([-h, 0], R^n),$$

where $\eta(\theta) = f(-\theta)$, $\theta \in [-h, 0]$;

(ii) $q(\lambda)$ is an entire transcendental function of the exponential type;

(iii) $q(\lambda)$ can be differentiated infinitely many times under the integral sign

$$\frac{d^k q(\lambda)}{d\lambda^k} = \int_{-h}^0 \theta^k e^{\lambda \theta} \eta(\theta) d\theta, \quad k = 0, 1, \dots;$$

(iv) in view of the completeness of the sequence $\{\theta^k\}$ in $L_2[-h, 0]$, $q(\lambda)$ can never be a polynomial, except for the trivial (null) polynomial;

(v) $q(\lambda) \equiv 0$ in \mathbb{C} if and only if $\eta(\theta) \equiv 0$ a.e. in $[-h, 0]$.

Example 2.1. For future reference, a few examples of FLT functions are given below:

$$\begin{aligned} f(\theta) &\equiv 1; & \int_{-h}^0 e^{\lambda \theta} d\theta &= \frac{1 - e^{-\lambda h}}{\lambda}, \\ f(\theta) &= \theta; & \int_{-h}^0 \theta e^{\lambda \theta} d\theta &= \frac{e^{-\lambda h} + \lambda h e^{-\lambda h} - 1}{\lambda^2}, \\ f(\theta) &= \theta^2; & \int_{-h}^0 \theta^2 e^{\lambda \theta} d\theta &= \frac{2 - \lambda^2 h^2 e^{-\lambda h} - 2\lambda h e^{-\lambda h} - 2 e^{-\lambda h}}{\lambda^3}. \end{aligned}$$

Note that the singularities at $\lambda = 0$ are canceled by the numerator. Other examples of finite Laplace transforms can be found in [1], [2].

Note also that given an entire function of the exponential type one can use the Paley–Wiener theorem [15], in its form given below as Theorem 2.8, to test whether a function belongs to the class FLT_2 .

Conditions (2.4) (iii) and (2.1) (iii) can now be rewritten in the following form:

PROPOSITION 2.5. The system (1.2)—i.e. (1.1)—is L_2 -approximately controllable if and only if

$$\begin{aligned} q(\lambda) &\in \text{FLT}_2([0, h], R^n), & q^T(\lambda) \Delta^{-1}(\lambda) B &\equiv 0, \quad \forall \lambda \in \rho(\tilde{A}) \\ &\Rightarrow q(\lambda) \equiv 0. \end{aligned}$$

PROPOSITION 2.6. *The system (1.2)—i.e. (1.1)—is M_2 -approximately controllable if and only if*

$$c \in R^n, \quad q(\lambda) \in \text{FLT}_2([0, h], R^n), \quad [c^T + q^T(\lambda)] \Delta^{-1}(\lambda) B \equiv 0, \quad \forall \lambda \in \rho(\tilde{A}) \\ \Rightarrow c = 0 \quad \text{and} \quad q(\lambda) \equiv 0.$$

Proof. Starting from (2.4) (iii) and using the form of $\Pi_1 R$, and $\Pi_0 R$, given by (2.9), (2.10) and the properties (i)–(v) above, we obtain Propositions 2.5 and 2.6 immediately. Q.E.D.

Remark 2.1. There is an interesting analogy between the characterization given by Proposition 2.5 and the result of Popov [57], on C^r reachability. We recall that a system is said [57] to have the C^r reachability property if there is an integer $r > 0$ and a time t such that for any $\varepsilon > 0$ ($\varepsilon < h$) and for any function $\varphi \in C^r([-h + \varepsilon, 0]; R^n)$ there is a control $u(\cdot)$ such that the corresponding solution of (1.1) satisfies $y_t(\theta) = \varphi(\theta)$, $\theta \in [-h + \varepsilon, 0]$. Now, according to [57] the system (1.1) with $m = 1$ is C^r reachable if and only if

$$c^T(\lambda) \Delta^{-1}(\lambda) B \equiv 0 \quad \text{implies} \quad c(\lambda) \equiv 0$$

for all n -vector polynomials $c(\lambda) = \sum_{i=0}^N c_i \lambda^i$ (N arbitrary).

A quick algebraic condition, exploiting the functional characterization of Proposition 2.6, is given next in a special case.

COROLLARY 2.7. *Let $\text{rank } B = n$. Then the system (1.2)—i.e. (1.1)—is M_2 -approximately controllable.*

Proof. Let \hat{B} be an $n \times n$ submatrix of B with $\text{rank } \hat{B} = n$. ($\hat{B} = B$, if $m = n$). Then the identity $[c^T + q^T(\lambda)] \Delta^{-1}(\lambda) \hat{B} \equiv 0$, $\forall \lambda \in \rho(\tilde{A})$, multiplied on the right by $\hat{B}^{-1} \Delta(\lambda)$ yields $c^T = 0$ and $q(\lambda) \equiv 0$, since $q(\lambda) \rightarrow 0$ as $\lambda \rightarrow +\infty$ for $q(\lambda) \in \text{FLT}_2([0, h], R^n)$. (A more complete description of the growth properties of an $\text{FLT}_2[0, h]$ function is given in the subsequent Lemma 4.2). Q.E.D.

Remark 2.2. The above result is hardly surprising in view of known results [5], showing that $\text{rank } B = n$ implies that for each $\varphi, \psi \in W_2^1([-h, 0], R^n)$ there is a control $u \in L_2([0, t], R^m)$ such that $x_t(\cdot, \varphi, u) = \psi$, $t > h$. (The important result that the converse holds has been also demonstrated in [5]). Therefore, if $\text{rank } B = n$, the evolution equation (1.2) in $R^n \times L_2([-h, 0], R^n)$ steers the initial zero vector $(0, 0)$ to any point of the subspace $\{(f(0), f(\cdot)), f(\cdot) \in W_2^1([-h, 0], R^n)\}$. But this subspace is precisely the domain $D(\tilde{A})$ of the generator \tilde{A} , which is dense in $R^n \times L_2([-h, 0], R^n)$ and so (1.2) is M_2 -approximately controllable. Notice that our derivation of Corollary 2.7 is independent of the results cited above, and will be considerably extended in §§ 4 and 5.

We now quote the Paley–Wiener theorem (mentioned before) in the form that will be useful later.

THEOREM 2.8 ([15, pp. 238, 241]; [27]). *Let $f(\lambda)$ be an entire function of exponential type, i.e. $|f(\lambda)| \leq a e^{b|\lambda|}$ for all $\lambda \in \mathbb{C}$, $a > 0$, $b > 0$. Then there exist constants H' , H and an L_2 function $F(t)$, $t \in R$, such that $F(t) = 0$ for $t \notin [H', H]$ and $f(\lambda) = \int_{H'}^H e^{-\lambda t} F(t) dt$, if and only if $\int_{-\infty}^{\infty} |f(i\omega)|^2 d\omega < \infty$. Moreover if we take*

$$H' = \limsup_{u \rightarrow \infty, u \in R} \frac{1}{u} \log |f(u)| \quad \text{and} \quad H = \limsup_{u \rightarrow \infty, u \in R} \frac{1}{u} \log |f(-u)|$$

then the interval $[H', H]$ cannot be replaced by a smaller one.

3. Algebraic necessary conditions for L_2 -approximate controllability. This section is devoted to deriving preliminary necessary conditions for L_2 - (hence also M_2 -)

approximate controllability, which are algebraic consequences of Proposition 2.5, stated in terms of matrices A, B . Let $\Delta(\lambda, \mu)$ denote the matrix $(\lambda I - A_0 - \mu A_1)$, while $\Delta(\lambda)$ will still denote $\Delta(\lambda, e^{-h\lambda})$ as in (2.8).

Similarly as in [57], [10], [42], [5], we introduce a special representation of $\Delta^{-1}(\lambda, \mu)$. Let $\text{adj } \Delta(\lambda, \mu)$ denote the algebraic adjoint of $\Delta(\lambda, \mu)$, that is the transpose of the corresponding matrix of cofactors. Obviously

$$(3.1) \quad \Delta^{-1}(\lambda, \mu)B = \frac{1}{\det \Delta(\lambda, \mu)} \text{adj } \Delta(\lambda, \mu)B$$

for all complex λ and μ for which the determinant does not vanish. Since $\Delta(\lambda, \mu) = \lambda I - A_0 - \mu A_1$, by the properties of determinants one has

$$\text{adj } \Delta(\lambda, \mu) = \sum_{j=0}^{n-1} Q_j(\lambda) \mu^j$$

where $Q_j(\lambda)$ denotes an $n \times n$ matrix polynomial in λ of degree at most $n-1-j$.

Thus, $\text{adj } \Delta(\lambda, \mu)B$ can be written as

$$(3.2) \quad \text{adj } \Delta(\lambda, \mu)B = \sum_{j=0}^{n-1} P_{n-1-j}(\lambda) \mu^j$$

where

$$P_{n-1-j}(\lambda) = Q_j(\lambda)B$$

is an $n \times m$ matrix polynomial in λ of degree at most $n-1-j$.

Define now the $n \times mn$ polynomial matrix

$$(3.3) \quad P(\lambda) = [P_{n-1}(\lambda), \dots, P_1(\lambda), P_0(\lambda)]$$

where $P_0(\lambda) = P_0 \equiv \text{const.}$, and the $mn \times m$ polynomial matrix

$$(3.4) \quad v(\mu) = \begin{bmatrix} I_m \\ I_m \mu \\ \vdots \\ I_m \mu^{n-1} \end{bmatrix} = \begin{bmatrix} 1 \\ \mu \\ \vdots \\ \mu^{n-1} \end{bmatrix} \otimes I_m$$

where I_m is the $m \times m$ identity matrix and the symbol \otimes denotes the Kronecker product. Formula (3.1) takes now the form

$$(3.5) \quad \Delta^{-1}(\lambda, \mu)B = \frac{1}{\det \Delta(\lambda, \mu)} P(\lambda) v(\mu).$$

For $m = 1$, this formula reduces to the one introduced in the paper by Popov [57]. An important aspect of the formula is that the numerator is factored into a product of polynomials in λ and in μ . The matrix $P(\lambda)$ will play a crucial role in our developments. Notice that $P(\lambda)$ depends only on A_0, A_1, B but not on delay h .

We say that a polynomial matrix $P(\lambda)$ has rank n if there is a $\lambda_0 \in \mathbb{C}$ such that the numerical matrix $P(\lambda_0)$ has rank n . This will be written subsequently as $\text{rank}_{\mathbb{C}} P(\lambda) = n$ to distinguish from $\text{rank } P(\lambda)$, which will denote the rank of the numerical matrix $P(\lambda)$ at a specific (fixed) point λ . Note that $\text{rank}_{\mathbb{C}} P(\lambda) = n$ implies that $\text{rank } P(\lambda) = n$ for all $\lambda \in \mathbb{C}$ except possibly a finite number of points (which in case $m = 1$ are the roots of the polynomial $\det P(\lambda)$). Equivalently, $\text{rank}_{\mathbb{C}} P(\lambda) = n$ means that $P(\cdot)$ has rank n over the ring of polynomials.

THEOREM 3.1. *A necessary condition for L_2 -approximate controllability of (1.2)—i.e. of (1.1)—is that the $n \times nm$ polynomial matrix $P(\lambda)$ satisfies $\text{rank}_{\mathbb{C}} P(\lambda) = n$.*

Remark 3.1. For $m = 1$, the condition $\text{rank}_{\mathbb{C}} P(\lambda) = n$ has been proved by Popov [57] to be necessary and sufficient for C^r reachability mentioned in § 1. Hence, in case $m = 1$, L_2 -approximate controllability implies C^r reachability. As will be seen later (Corollary 4.6), the converse is also true for $n \leq 2$ but need not be true for $n \geq 3$ (Examples 3.1 and 4.1). Also notice that for $n = 1$ one has $P(\lambda) \equiv B$.

Proof. The proof of Theorem 3.1 will be split into two lemmas of independent interest.

LEMMA 3.2. *Let $\text{rank } P(\lambda) < n$ for all $\lambda \in \mathbb{C}$. Then there exists an $n \times 1$ polynomial vector $p(\lambda)$ not identically zero (in fact $p(0) \neq 0$) such that $p^T(\lambda)P(\lambda) \equiv 0$, $\lambda \in \mathbb{C}$.*

Proof. The proof follows the one given for $m = 1$ by Popov in [57], and is reported here in detail mainly because Popov's report does not seem to have been published in the literature. We use the canonical form for a polynomial matrix [18], [3] according to which we have

$$N(\lambda)P(\lambda)M(\lambda) = D(\lambda)$$

where

$N(\lambda) = n \times n$ polynomial matrix with constant, nonzero determinant

$M(\lambda) = m \cdot n \times m \cdot n$ polynomial matrix with constant, nonzero determinant

$D(\lambda) = n \times mn$ canonical matrix of the form

$$D(\lambda) = \left[\begin{array}{ccc|cc} d_1(\lambda) & & 0 & 0 & 0 \\ & d_2(\lambda) & & & \\ & & \ddots & & \\ 0 & & & d_n(\lambda) & \\ \hline & & & & 0 & 0 \end{array} \right]$$

where $d_i(\lambda)$ are polynomials such that each of them is divisible by the preceding one.

From

$$\text{rank } N(\lambda) = n, \quad \text{rank } M(\lambda) = mn, \quad \text{rank } P(\lambda) < n, \quad \forall \lambda \in \mathbb{C},$$

it follows that $\text{rank } D(\lambda) < n$, for all $\lambda \in \mathbb{C}$, since the rank of the product of rectangular matrices does not exceed the rank of each factor [18, p. 12]. Hence, some of the $d_i(\lambda)$ must vanish identically. Then, as $p^T(\lambda)$ in the conclusion of the lemma one can take the row of $N(\lambda)$ corresponding to a zero row of $D(\lambda)$. This yields in fact $p^T(\lambda)P(\lambda)M(\lambda) \equiv 0$ and hence $p^T(\lambda)P(\lambda) \equiv 0$, $\forall \lambda \in \mathbb{C}$, since $M(\lambda)$ is nonsingular. Also $p^T(0) \neq 0$, otherwise one would have $\det N(0) = 0$, a contradiction. Q.E.D.

LEMMA 3.3. *Given an arbitrary $n \times 1$ polynomial vector $p(\lambda)$, $p(0) \neq 0$ such that*

$$p^T(\lambda)P(\lambda) \equiv 0, \quad \forall \lambda \in \mathbb{C},$$

there is a nonzero $n \times 1$ function $q(\lambda)$ of class $\text{FLT}_2([0, h], R^n)$, i.e. $q(\lambda) = \int_{-h}^0 e^{\lambda\theta} \eta(\theta) d\theta$ for some nonzero $\eta(\cdot) \in L_2([-h, 0], R^n)$, such that $q^T(\lambda)P(\lambda) \equiv 0$, $\forall \lambda \in \mathbb{C}$.

Proof. 1. Let $f(\cdot)$ be a scalar function in $L_2[-h, 0]$, and define $(Kf)(\theta) = \int_{-h}^{\theta} f(\sigma) d\sigma$, where K is a (Volterra) operator on $L_2[-h, 0]$. Integration by parts then yields the following implication:

$$(Kf)(0) = 0 \quad \text{implies} \quad \int_{-h}^0 e^{\lambda\theta} f(\theta) d\theta = -\lambda \int_{-h}^0 (Kf)(\theta) e^{\lambda\theta} d\theta.$$

By repeating the above argument one concludes that: if

$$(3.6) \quad (K^l f)(0) = 0, \quad l = 1, 2, \dots, j,$$

then

$$(3.7) \quad \int_{-h}^0 e^{\lambda \theta} f(\theta) d\theta = (-\lambda)^j \int_{-h}^0 (K^j f)(\theta) e^{\lambda \theta} d\theta.$$

Remark 3.2. Notice that, since [60, p. 291]

$$(K^l f)(\xi) = \frac{1}{(l-1)!} \int_{-h}^{\xi} (\xi - \theta)^{l-1} f(\theta) d\theta, \quad l = 1, 2, \dots,$$

the condition (3.6) for $f(\cdot)$ is equivalent to the solvability for $f(\cdot)$ of the following moment problem:

$$(3.8) \quad \int_{-h}^0 \theta^{l-1} f(\theta) d\theta = 0, \quad l = 1, 2, \dots, j.$$

The moment problem (3.8) always has a nonzero solution in $L_2([-h, 0], R)$.

2. Next, write an arbitrary $n \times 1$ polynomial $p(\lambda)$ as

$$p(\lambda) = \sum_{i=0}^j a_i \lambda^i, \quad a_i \in R^n,$$

and consider the $L_2([-h, 0], R^n)$ function $\eta(\cdot)$

$$(3.9) \quad \eta(\theta) = \sum_{i=0}^j (-1)^i a_i (K^{j-i} f)(\theta), \quad -h \leq \theta \leq 0,$$

for some nonzero $f(\cdot)$ satisfying (3.6), i.e., (3.8). Then one verifies, by a repeated application of the implication (3.6) \Rightarrow (3.7), that

$$(3.10) \quad \begin{aligned} q(\lambda) &= \int_{-h}^0 e^{\lambda \theta} \eta(\theta) d\theta = \sum_{i=0}^j (-1)^i a_i \int_{-h}^0 (K^{j-i} f)(\theta) e^{\lambda \theta} d\theta \\ &= \sum_{i=0}^j (-1)^i a_i (-1)^i \lambda^i \int_{-h}^0 (K^i f)(\theta) e^{\lambda \theta} d\theta \\ &= \left(\sum_{i=0}^j a_i \lambda^i \right) \int_{-h}^0 e^{\lambda \theta} g(\theta) d\theta \end{aligned}$$

with $(K^i f)(\theta) \stackrel{\text{def}}{=} g(\theta) \neq 0$ and $g(\cdot) \in L_2[-h, 0]$. We have thus constructively proved that, given an arbitrary $n \times 1$ polynomial vector, there is an $L_2([-h, 0], R^n)$ -function $\eta(\cdot)$ —given in fact by (3.9)—such that the corresponding $q(\lambda)$ contains such polynomial as a factor.

3. We are now in the position to prove the lemma. If $p(\lambda)$ is a polynomial as in the assumption, then

$$\left[\int_{-h}^0 e^{\lambda \theta} g(\theta) d\theta \right] p^T(\lambda) P(\lambda) = q^T(\lambda) P(\lambda) = 0, \quad \forall \lambda \in \mathbb{C},$$

with $g(\theta)$ and $q(\lambda)$ as in step 2, $q(\lambda) \neq 0$. Q.E.D.

*Proof of Theorem 3.1.*² If, by contradiction, $\text{rank } P(\lambda) < n$, $\forall \lambda \in \mathbb{C}$, then Lemmas 3.2 and 3.3 imply that there is a nonzero $n \times 1$ function $q(\lambda)$ of class $\text{FLT}_2([0, h], \mathbb{R}^n)$ —given in fact by (3.9)—such that $q^T(\lambda)P(\lambda) \equiv 0$, $\forall \lambda \in \mathbb{C}$. But then, because of (3.5) with $\mu = e^{-\lambda h}$,

$$q^T(\lambda)\Delta^{-1}(\lambda)B = 0, \quad \forall \lambda \in \rho(\tilde{A})$$

which contradicts Proposition 2.5, since $q(\lambda)$ is nonzero. Q.E.D.

The next theorem exhibits an algebraic condition directly in terms of A_0, A_1, B , for $\text{rank } P(\lambda_0)$ to be full for some λ_0 .

THEOREM 3.4. *The following conditions are equivalent:*

- (i) $\text{rank } P(\lambda_0) = n$ for some $\lambda_0 \in \mathbb{C}$
- (ii) $\text{rank } H(\lambda_1) = n$ for some $\lambda_1 \notin \sigma(A_0)$ (spectrum of A_0), where $H(\lambda)$ is an $n \times m(n-1)$ matrix defined by

$$(3.11) \quad \begin{aligned} H(\lambda) &= [G(\lambda), F(\lambda)G(\lambda), \dots, F^{n-1}(\lambda)G(\lambda)], \\ G(\lambda) &= (\lambda I - A_0)^{-1}B, \quad F(\lambda) = (\lambda I - A_0)^{-1}A_1, \quad \lambda \notin \sigma(A_0). \end{aligned}$$

In particular, if $\lambda_0 \notin \sigma(A_0)$, we can take $\lambda_1 = \lambda_0$.

Remark 3.3. Obviously we have: $\text{rank } H(\lambda) = \text{rank } \tilde{H}(\lambda)$, $\forall \lambda \notin \sigma(A_0)$, where

$$\tilde{H}(\lambda) = [\tilde{G}(\lambda), \tilde{F}(\lambda)\tilde{G}(\lambda), \dots, \tilde{F}^{n-1}(\lambda)\tilde{G}(\lambda)],$$

$$\tilde{G}(\lambda) = \text{adj } (\lambda I - A_0)B, \quad \tilde{F}(\lambda) = \text{adj } (\lambda I - A_0)A_1$$

and $\tilde{H}(\lambda)$ makes sense also for $\lambda \in \sigma(A_0)$. Notice, however, that $\text{rank } H(\lambda_1) = n$, $\lambda_1 \notin \sigma(A_0)$, need not imply $\text{rank } \tilde{H}(\lambda_2) = n$ for any $\lambda_2 \in \sigma(A_0)$

$$\left(\text{for example take } A_0 = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}, A_1 = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right).$$

Proof. We prove, equivalently that

$$(3.12) \quad \text{rank } P(\lambda) < n, \quad \forall \lambda \in \mathbb{C} \Leftrightarrow \text{rank } H(\lambda) < n, \quad \forall \lambda \notin \sigma(A_0).$$

\Rightarrow : There is a nonzero n -column vector p_λ , depending on λ such that $p_\lambda^T P(\lambda) \equiv 0$, $\forall \lambda \in \mathbb{C}$ (we know, in view of Lemma 3.2, that we could take for p_λ a polynomial vector $p(\lambda)$, with $p(0) \neq 0$, but this will not be needed). By (3.5) one obtains

$$(3.13) \quad p_\lambda^T \Delta^{-1}(\lambda, \mu)B = 0$$

for all λ, μ such that $\det \Delta(\lambda, \mu) \neq 0$. Let λ_1 be an arbitrary point not in $\sigma(A_0)$ so that $\det \Delta(\lambda_1, 0) \neq 0$. Then $\det \Delta(\lambda_1, \mu) \neq 0$ for all μ suitably small in modulus, say $|\mu| < \delta$, for some $\delta > 0$. Differentiating the identity $\Delta^{-1}(\lambda_1, \mu)\Delta(\lambda_1, \mu) = I$, one has

$$\left. \frac{d\Delta^{-1}(\lambda_1, \mu)B}{d\mu} \right|_{\mu=0} = -\Delta^{-1}(\lambda_1, \mu) \frac{d\Delta(\lambda_1, \mu)}{d\mu} \Delta^{-1}(\lambda_1, \mu)B \Big|_{\mu=0} = F(\lambda_1)G(\lambda_1).$$

By induction

$$(3.14) \quad \left. \frac{d^k \Delta^{-1}(\lambda_1, \mu)B}{d\mu^k} \right|_{\mu=0} = k! F^k(\lambda_1)G(\lambda_1), \quad k = 0, 1, 2, \dots$$

² It was pointed out by one of the referees that Theorem 3.1 can also be derived from Theorem 5.2 of [5].

Then (3.13) and (3.14) imply

$$p_{\lambda_1}^T F^k(\lambda_1) G(\lambda_1) = 0, \quad k = 0, 1, 2, \dots,$$

with $p_{\lambda_1} \neq 0$, and the right hand side of (3.12) follows.

\Leftarrow : There is a nonzero n -dimensional column vector p_λ , depending on λ , such that $p_\lambda^T H(\lambda) \equiv 0, \forall \lambda \notin \sigma(A_0)$. Let $\lambda = \lambda_1$ be an arbitrary point not in $\sigma(A_0)$ so that, as before, $\det \Delta(\lambda_1, \mu) \neq 0$ for all $|\mu| < \delta$. Then

$$(3.15) \quad p_{\lambda_1}^T F^k(\lambda_1) G(\lambda_1) = 0, \quad k = 0, 1, \dots, n-1,$$

and by the Caley–Hamilton theorem applied to the matrix $F(\cdot)$, (3.15) holds for all $k = 0, 1, \dots$. Then (3.14) implies

$$p_{\lambda_1}^T \Delta^{-1}(\lambda_1, \mu) B \equiv 0, \quad \forall |\mu| < \delta,$$

in view of the analyticity of $\Delta^{-1}(\lambda_1, \cdot)$ and, by (3.12) and (3.5)

$$(3.16) \quad p_{\lambda_1}^T P(\lambda_1) v(\mu) \equiv 0, \quad \forall \mu \in \mathbb{C},$$

where the extension to all $\mu \in \mathbb{C}$ is obtained by analytic continuation. Equation (3.16) repeated for n -distinct values $\mu_1, \mu_2, \dots, \mu_n$ yields

$$(3.17) \quad p_{\lambda_1}^T P(\lambda_1) V = 0$$

where

$$(3.18) \quad V = [v(\mu_1), v(\mu_2), \dots, v(\mu_n)] = \begin{bmatrix} 1 & 1 & \dots & 1 \\ \mu_1 & \mu_2 & \dots & \mu_n \\ \vdots & \vdots & & \vdots \\ \mu_1^{n-1} & \mu_2^{n-1} & \dots & \mu_n^{n-1} \end{bmatrix} \otimes I_m$$

and \otimes denotes the direct (Kronecker) product of two matrices [34], [56]. By the nonsingularity of the Vandermonde matrix for distinct μ_i and by the fact that for any two square nonsingular matrices A, B the direct product $A \otimes B$ is also nonsingular [34] one has that the $mn \times mn$ matrix V is nonsingular. Thus (3.17) implies

$$(3.19) \quad p_{\lambda_1}^T P(\lambda_1) = 0 \quad \text{while } p_{\lambda_1} \neq 0.$$

Now, suppose that $\text{rank } P(\lambda) = n$ at some $\lambda = \lambda_0 \in \mathbb{C}$. Then there exists a minor of $P(\lambda)$, of order n , which is nonzero at $\lambda = \lambda_0$. Since $P(\lambda)$ is a polynomial matrix, this minor is nonzero everywhere except at a finite number of isolated points. Thus, $\text{rank } P(\lambda) = n$ almost everywhere in the complex plane, but then one must have $p_\lambda^T = 0$ almost everywhere, which contradicts (3.19) (where λ_1 is arbitrary except that $\lambda_1 \notin \sigma(A_0)$), and proves the theorem. Q.E.D.

Remark 3.4. The relation between the matrix functions $\Delta^{-1}(\lambda, \mu)B, G(\lambda), F(\lambda)$ can also be obtained by interpreting $\Delta^{-1}(\lambda, \mu)$ as a perturbation of the resolvent $(I\lambda - A_0)^{-1}$. In fact, the matrix $A_0 + A_1\mu$ can be treated as a matrix operator A_0 perturbed by $A_1\mu$. Using the so called second Neumann series for the resolvent [28, p. 66–72] one obtains, for λ such that $\det \Delta(\lambda, 0) \neq 0$,

$$\begin{aligned} (I\lambda - A_0 - A_1\mu)^{-1} &= (I\lambda - A_0)^{-1} [I + \mu A_1 (I\lambda - A_0)^{-1} \\ &\quad + \mu^2 A_1 (I\lambda - A_0)^{-1} A_1 (I\lambda - A_0)^{-1} + \dots \text{etc.}] \end{aligned}$$

where the series converges for μ such that $|\mu|$ is sufficiently small [28, p. 67]. Thus

$$\Delta^{-1}(\lambda, \mu)B = G(\lambda) + \mu F(\lambda)G(\lambda) + \mu^2 F^2(\lambda)G(\lambda) + \dots,$$

which is an alternative way of stating (3.14).

Remark 3.5. Notice that, when A_0 and A_1 commute, we can write more simply

$$\text{rank } H(\lambda) = \text{rank } [A_1^{n-1}B, (\lambda I - A_0)A_1^{n-2}B, \dots, (\lambda I - A_0)^{n-2}A_1B, (\lambda I - A_0)^{n-1}B].$$

We explicitly note two important special cases.

COROLLARY 3.5. $\text{rank}_{\mathbb{C}} P(\lambda) = n$ is equivalent to:

- (i) $\text{rank } [B, A_1B, \dots, A_1^{n-1}B] = n$, when $A_0 = \alpha I$, for any constant α ;
- (ii) $\text{rank } [B, A_0B, \dots, A_0^{n-1}B] = n$, when $A_1 = \alpha I$, for any constant $\alpha \neq 0$.

A simple algebraic necessary condition for $\text{rank}_{\mathbb{C}} P(\lambda) = n$ follows.

THEOREM 3.6. *The condition $\text{rank}_{\mathbb{C}} P(\lambda) = n$ implies*

$$\text{rank } [A_1, B] = n.$$

Proof. Otherwise, there is a nonzero column vector $v \in R^n$ such that $v^T[A_1, B] = 0$. Let $\lambda_0 \in \rho(A_0)$ and define $\eta = (\lambda_0 I - A_0)^T v$. Then $\eta \neq 0$, $v^T = \eta^T(\lambda_0 I - A_0)^{-1}$ and $\eta^T(\lambda_0 I - A_0)^{-1}[A_1, B] = \eta^T[F(\lambda_0), G(\lambda_0)] = 0$ with $F(\lambda)$ and $G(\lambda)$ defined in Theorem 3.4. But then there is a ξ , in fact $\xi = 0$, such that

$$\text{rank } [\xi I - F(\lambda_0), G(\lambda_0)] < n, \quad \forall \lambda_0 \in \rho(A),$$

and a known result [23] implies $\text{rank } H(\lambda_0) < n$ with $H(\lambda)$ defined in Theorem 3.4. The conclusion now follows from Theorems 3.1 and 3.4. Q.E.D.

COROLLARY 3.7. *If the system (1.2)—i.e., (1.1)—is L_2 -approximately controllable, then $\text{rank } [A_1, B] = n$.*

We terminate this section with an example of a system for which $\text{rank}_{\mathbb{C}} P(\lambda) = n$, but for which the L_2 -approximate controllability does not obtain.

Example 3.1. Let $n = 3$, $m = 1$, arbitrary $h > 0$, and

$$A_0 = \begin{bmatrix} 0 & 0 & 0 \\ 2 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix}; \quad A_1 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix}; \quad B = \begin{bmatrix} 1 \\ -2 \\ -1 \end{bmatrix}.$$

A straightforward calculation shows that $P(\lambda)$, defined by (3.3)–(3.5), is given by

$$P(\lambda) = \begin{bmatrix} \lambda^2 & 2-2\lambda & -2 \\ 2\lambda-2\lambda^2 & -2\lambda & 0 \\ -2+2\lambda-\lambda^2 & 2 & 0 \end{bmatrix};$$

so $\det P(\lambda) = 4\lambda^3$ and the necessary condition $\text{rank}_{\mathbb{C}} P(\lambda) = 3$ of Theorem 3.1 is satisfied. We have, moreover, $\det \Delta(\lambda) = \lambda^3$ (independent on h !) and

$$\Delta^{-1}(\lambda)B = \frac{1}{\lambda^3} \begin{bmatrix} \lambda^2 + (2-2\lambda)e^{-\lambda h} - 2e^{-2\lambda h} \\ 2\lambda - 2\lambda^2 - 2\lambda e^{-\lambda h} \\ -2 + 2\lambda - \lambda^2 + 2e^{-\lambda h} \end{bmatrix}.$$

Now, notice that for $h = 1$ the second and third rows $r_2(\lambda)$ and $r_3(\lambda)$ of the above matrix are of class $\text{FLT}_2([0, 1], R)$, because by using Example 2.1 one obtains

$$r_2(\lambda) = \frac{2(\lambda - \lambda^2 - \lambda e^{-\lambda})}{\lambda^3} = \int_{-1}^0 e^{\lambda\theta} [-2(\theta+1)] d\theta,$$

$$r_3(\lambda) = \frac{-2 + 2\lambda - \lambda^2 + 2e^{-\lambda}}{\lambda^3} = \int_{-1}^0 e^{\lambda\theta} (-\theta^2 - 2\theta - 1) d\theta.$$

Hence, the vector function $q^T(\lambda) = [0, -r_3(\lambda), r_2(\lambda)]$ is a nonzero member of class

$\text{FLT}_2([0, 1], R^3)$ annihilating $\Delta^{-1}(\lambda)B: q^T(\lambda) \Delta^{-1}(\lambda)B \equiv 0$. In view of Proposition 2.5, the L_2 -approximate controllability does not obtain for the present system with $h = 1$. The system will be reexamined in more depth in Example 4.1.

This example shows that the necessary condition given by Theorem 3.1 (or, equivalently, by Theorem 3.4) is not always sufficient for L_2 - (let alone M_2 -) approximate controllability. This is hardly surprising since the matrix $P(\lambda)$ involves only the matrices A_0, A_1 , and B but not the delay h . This motivates further study of Propositions 2.5 and 2.6 in search of some additional conditions that together with $\text{rank}_C P(\lambda) = n$ would guarantee the sought-after controllability properties. Before proceeding to do that, however, we would like to point out that the example given above is a result of quite an artificial construction, in which we used some results on pointwise degenerate systems (this will be explained in § 8). One may, for instance, observe that the above functions $r_i(\lambda)$ will not be of class $\text{FLT}_2([0, h], R)$ unless h equals exactly 1. (See further analysis in Example 4.1.) The failure of $\text{rank}_C P(\lambda) = n$ to guarantee (at least) L_2 -approximate controllability will be shown later to be some sort of a singular case.

4. Necessary and sufficient conditions for M_2 - and L_2 -approximate controllability. In this section we show that the necessary and sufficient conditions for M_2 - and L_2 -approximate controllability given by Propositions 2.5 and 2.6 reduce to a purely algebraic test on whether a certain system of linear (algebraic) homogeneous equations has a nonzero solution.

The main idea underlying our considerations will be to characterize the elements $c \in R^n$ and $q(\cdot) \in \text{FLT}_2([0, h], R^n)$ which annihilate the transfer function $\Delta^{-1}(\lambda)B$. This characterization will be obtained by making use of an interesting, but not widely known, result of Hardy on entire functions, and of the Paley-Wiener theorem. As a result, the annihilating $q(\cdot)$ will be expressed by a simple formula with unknown coefficients which satisfy the system of linear equations mentioned above.

In view of Propositions 2.5 and 2.6 and of identity (3.5) with $\mu = e^{-\lambda h}$ we shall consider the relations

$$(4.1) \quad q^T(\lambda)P(\lambda)v(e^{-\lambda h}) \equiv 0, \quad \forall \lambda \in \mathbb{C},$$

and

$$(4.2) \quad [c^T + q^T(\lambda)]P(\lambda)v(e^{-\lambda h}) \equiv 0, \quad \forall \lambda \in \mathbb{C},$$

where $c \in R^n$, $q(\cdot) \in \text{FLT}_2([0, h], R^n)$. First observe that if one could prove that (4.1) implies, for $q(\cdot) \in \text{FLT}_2([0, h], R^n)$, that

$$(4.1') \quad q^T(\lambda)P(\lambda) \equiv 0, \quad \forall \lambda \in \mathbb{C},$$

then the necessary condition $\text{rank}_C P(\lambda) = n$ would also be sufficient for L_2 -approximate controllability, since (4.1') would imply $q(\lambda) \equiv 0$ as desired. Similarly, if one could prove that (4.2) implies

$$(4.2') \quad [c^T + q^T(\lambda)]P(\lambda) \equiv 0, \quad \forall \lambda \in \mathbb{C},$$

then $\text{rank}_C P(\lambda) = n$ would yield $c^T + q^T(\lambda) \equiv 0$, hence $c = 0$ (by letting λ real $\rightarrow +\infty$; see subsequent Lemma 4.2 (i)) and $q(\lambda) \equiv 0$. Notice that in case $n = 1$ both (4.1') and (4.2') are true. In general, however, the left hand sides of (4.1') and (4.2') are given by some nonzero expressions that will be derived below.

Remark 4.1. Note also that in case of C^r reachability investigated by Popov [57] (see Remark 2.1) the function $q(\cdot)$ in (4.1) and (4.2) is replaced by a vector *polynomial*,

in which case the implication $(4.1) \Rightarrow (4.1')$ is always true. Indeed, it is not difficult to prove by elementary means that for any polynomial in two variables $p(\lambda, \mu)$, the relation

$$\{p(\lambda, e^{-\lambda h}) = 0 \text{ for all } \lambda \text{ such that } \operatorname{Re} \lambda \geq \alpha, \text{ (with arbitrary } \alpha)\}$$

implies $p(\lambda, \mu) = 0$ for all λ and all μ . Consequently, (4.1), with $q(\lambda)$ being a polynomial vector, implies $q^T(\lambda)P(\lambda)v(\mu) \equiv 0, \forall \lambda, \mu \in \mathbb{C}$; this and the arguments leading from (3.17) to (3.19) yield (4.1') in the case of C^r reachability.

In the case of L_2 -approximate controllability, $q^T(\lambda)$ is an entire transcendental function which can never be a nonzero polynomial, so the arguments above cannot be applied. The main tool we shall employ instead is a result on entire functions due to Hardy.

LEMMA 4.1 (Hardy [22]). *Let $f(\lambda)$ be a scalar entire function satisfying*

(i) $f(\lambda) = O(|\lambda|^n e^{a|\lambda|}), a > 0$, *for large $|\lambda|$;*

(ii) $f(\xi) = O(\xi^n e^{-a\xi})$, *for large (real) positive ξ .*

Then $f(\lambda) = p_n(\lambda)e^{-a\lambda}$, where $p_n(\lambda)$ is a polynomial of degree at most n .

In order to apply Hardy's lemma, we compute some growth estimates for $q(\lambda)$, which are collected in the lemma below.

LEMMA 4.2. *Let $q_i(\lambda)$ denote the i th component of $q(\lambda)$, where*

$$q_i(\lambda) = \int_{-h}^0 e^{\lambda\theta} \eta_i(\theta) d\theta, \quad \eta_i(\cdot) \in L_2[-h, 0].$$

Then, for each $i = 1, \dots, n$, we have

$$(4.3) \quad \begin{cases} \text{(i)} & \left\{ \begin{aligned} &= O\left(\frac{1}{|\operatorname{Re} \lambda|^{1/2}}\right), & \operatorname{Re} \lambda > 0, \\ \text{(ii)} & |q_i(\lambda)| \leq \text{const. } \|\eta_i\|_{L_1}, & \operatorname{Re} \lambda = 0, \\ \text{(iii)} & \left\| \eta_i \right\|_{L_2} \left| \frac{1 - e^{-2\operatorname{Re} \lambda h}}{2 \operatorname{Re} \lambda} \right|^{1/2}, & \operatorname{Re} \lambda < 0. \end{aligned} \right. \end{cases}$$

Since $(1 - e^{-2\xi h})/2\xi \rightarrow h$ as $\xi \rightarrow 0$, the estimate for $\operatorname{Re} \lambda \leq 0$ can be replaced by $O(e^{|\lambda|h})$.

If $p_k(\lambda)$ is a polynomial of degree k , then we can write

$$(4.4) \quad \begin{cases} \text{(i)} & |p_k(\lambda)q_i(\lambda)| = \begin{cases} O(|\lambda|^k e^{|\lambda|h}) & \text{for } |\lambda| \text{ large, } \lambda \in \mathbb{C}, \\ O(\xi^k) & \text{for } \xi \text{ real } \rightarrow +\infty; \end{cases} \end{cases}$$

$$(4.5) \quad \begin{cases} \text{(i)} & |e^{-\lambda h} p_k(\lambda)q_i(\lambda)| = \begin{cases} O(|\lambda|^k e^{2|\lambda|h}) & \text{for } |\lambda| \text{ large, } \lambda \in \mathbb{C}, \\ O(\xi^k e^{-\xi h}) & \text{for } \xi \text{ real } \rightarrow +\infty. \end{cases} \end{cases}$$

We also have the following lemma.

LEMMA 4.3. *Let $q(\lambda)$ be as in Lemma 4.2 and suppose, moreover, that $|q(\xi)| = O(\xi^k e^{-h\xi})$ for large positive ξ and some nonnegative integer k . Then $q(\lambda) \equiv 0$.*

Proof. By Hardy's lemma, we have $q(\lambda) = p_k(\lambda)e^{-\lambda h}$, i.e.,

$$\int_0^h e^{\lambda\sigma} \eta(\sigma - h) d\sigma = p_k(\lambda).$$

Differentiate in λ the above identity infinitely many times (under the integral sign, as

in (iii) in Proposition 2.4), and use the completeness of $\{\theta^j\}$, $j = 0, 1, \dots$, in $L_2[0, h]$ to obtain that $\eta(\theta) \equiv 0$ a.e. in $[-h, 0]$. Q.E.D.

Recalling the definition of $P(\lambda)$ given by (3.3), equation (4.1) can be now written as

$$\sum_{i=1}^n [c^T + q^T(\lambda)] P_{n-i}(\lambda) (e^{-\lambda h})^{i-1} = 0, \quad \forall \lambda \in \mathbb{C},$$

or

$$(4.6) \quad [c^T + q^T(\lambda)] P_{n-1}(\lambda) = - \sum_{i=2}^n [c^T + q^T(\lambda)] P_{n-i}(\lambda) (e^{-\lambda h})^{i-1}.$$

The left hand side of identity (4.6) is $O(|\lambda|^{n-1} e^{h|\lambda|})$ for large $|\lambda|$, by (4.4) (i), while by (4.5) (ii), the right hand side of (4.6) with $\lambda = \xi$ is $O(\xi^{n-2} e^{-h\xi})$ for large positive ξ . These estimates and Hardy's lemma applied componentwise yield that

$$(4.7) \quad [c^T + q^T(\lambda)] P_{n-1}(\lambda) = p_{n-1}(\lambda) e^{-\lambda h}, \quad \lambda \in \mathbb{C},$$

where $p_{n-1}(\lambda)$ is an $1 \times m$ polynomial vector with entries all of degree at most $n-1$. Substituting (4.7) into (4.6) gives

$$(4.8) \quad P_{n-1}(\lambda) + [c^T + q^T(\lambda)] P_{n-2}(\lambda) = -e^{-\lambda h} [c^T + q^T(\lambda)] \sum_{i=3}^n P_{n-i}(\lambda) (e^{-\lambda h})^{i-3}.$$

Now, the above reasoning can be repeated. The left hand side of (4.8) is $O(|\lambda|^{n-2} e^{h|\lambda|})$ for large $|\lambda|$, while the right hand side of (4.8) with $\lambda = \xi$ is $(\xi^{n-3} e^{-h\xi})$ for large positive ξ . By using again the Hardy's lemma one obtains

$$(4.9) \quad p_{n-1}(\lambda) + [c^T + q^T(\lambda)] P_{n-2}(\lambda) = p_{n-2}(\lambda) e^{-\lambda h}$$

where $p_{n-2}(\lambda)$ is an $1 \times m$ polynomial vector with entries all of degree at most $n-2$.

Continuing in this manner, we obtain the last term from

$$(4.10) \quad p_2(\lambda) + [c^T + q^T(\lambda)] P_1(\lambda) = -e^{-\lambda h} [c^T + q^T(\lambda)] P_0, \quad \lambda \in \mathbb{C},$$

where $p_2(\lambda)$ is an $1 \times m$ polynomial vector with components all of degree at most 2. The left hand side of (4.10) is $O(|\lambda| e^{h|\lambda|})$ for $|\lambda|$ large, while the right hand side with $\lambda = \xi$ is $O(e^{-h\xi})$ for large positive ξ . Application of Hardy's lemma then gives

$$(4.11) \quad p_2(\lambda) + [c^T + q^T(\lambda)] P_1(\lambda) = p_1(\lambda) e^{-\lambda h}$$

with $p_1(\lambda)$ being an $1 \times m$ polynomial vector with components all of degree at most 1. Substituting (4.11) into (4.10) yields

$$(4.12) \quad [c^T + q^T(\lambda)] P_0 = -p_1(\lambda).$$

The results obtained through the above procedure are collected below:

$$(4.13) \quad \begin{aligned} [c^T + q^T(\lambda)] P_{n-1}(\lambda) &= p_{n-1}(\lambda) e^{-\lambda h} && \text{(from (4.7)),} \\ [c^T + q^T(\lambda)] P_{n-2}(\lambda) &= -p_{n-1}(\lambda) + p_{n-2}(\lambda) e^{-\lambda h} && \text{(from (4.9)),} \\ \vdots & && \vdots \\ [c^T + q^T(\lambda)] P_1(\lambda) &= -p_2(\lambda) + p_1(\lambda) e^{-\lambda h} && \text{(from (4.11)),} \\ [c^T + q^T(\lambda)] P_0 &= -p_1(\lambda) && \text{(from (4.12)).} \end{aligned}$$

This result can be further improved, namely it will be shown that:

Claim (i): in the case of M_2 -approximate controllability the maximal degrees of the polynomials p_j , $j = 1, \dots, n-1$ appearing in (4.13) are actually lower by one unit;

Claim (ii): in the case of L_2 -approximate controllability ($c^T = 0$ in (4.13)) these degrees are actually lower by two units ($p_{-1}(\lambda)$ is a null polynomial).

Consider first the identity

$$(4.14) \quad [c^T + q^T(\lambda)]P_{n-k}(\lambda) = -p_{n-k+1}(\lambda) + p_{n-k}(\lambda)e^{-\lambda h}, \quad k = 2, \dots, n-1,$$

where

$$P_{n-k}(\lambda) = \sum_{i=0}^{n-k} D_{k,i} \lambda^i \quad \text{and} \quad p_{n-k+1}(\lambda) = \sum_{i=0}^{n-k+1} d_{k,i} \lambda^i$$

and $D_{k,i}$, $d_{k,i}$ are $n \times m$ and $1 \times m$ matrices, respectively. To prove Claim (ii), set $c^T = 0$ and divide both sides of (4.14) by λ^{n-k} . This gives

$$(4.15) \quad \begin{aligned} q^T(\lambda) \left[\frac{1}{\lambda^{n-k}} D_{k,0} + \dots + \frac{1}{\lambda} D_{k,n-k+1} + D_{k,n-k} \right] \\ = - \left[\frac{1}{\lambda^{n-k}} d_{k,0} + \dots + d_{k,n-k} + \lambda d_{k,n-k+1} \right] + p_{n-k}(\lambda) e^{-\lambda h} / \lambda^{n-k}, \end{aligned}$$

$k = 2, \dots, n-1.$

For λ real positive, $\lambda \rightarrow +\infty$, (4.3) (i) yields that the left hand side of (4.15) tends to zero; similarly does the second term on the right hand side. This implies $d_{k,n-k} = d_{k,n-k+1} = 0$, $k = 2, \dots, n-1$. The cases $k = 1$ and $k = n$ are handled similarly. Thus, all the polynomials $p_j(\lambda)$ of degree j appearing in (4.13) with $c^T = 0$ can be replaced by the polynomials $p_{j-2}(\lambda)$, of degree $j-2$, $j = 2, \dots, n-1$, (and $p_1(\lambda)$ is replaced by $p_{-1}(\lambda) \equiv 0$) and Claim (ii) is established.

To prove Claim (i), one divides both sides of (4.14) by λ^{n-k+1} and proceeds similarly, to conclude that, in this case, $d_{k,n-k+1} = 0$.

With both claims taken into account, (4.13) becomes

$$(4.16) \quad [c^T + q^T(\lambda)]P(\lambda) = w_M(\lambda)$$

where $w_M(\lambda)$ is a $1 \times mn$ row vector function given by

$$(4.17) \quad w_M(\lambda) = [p_{n-2}(\lambda)e^{-\lambda h}, -p_{n-2}(\lambda) + p_{n-3}(\lambda)e^{-\lambda h}, \dots, -p_1(\lambda) + p_0e^{-\lambda h}, -p_0].$$

Similarly, in the case of L_2 -approximate controllability, (4.13) becomes

$$(4.18) \quad q^T(\lambda)P(\lambda) = w_L(\lambda)$$

where

$$(4.19) \quad w_L(\lambda) = [p_{n-3}(\lambda)e^{-\lambda h}, -p_{n-3}(\lambda) + p_{n-4}(\lambda)e^{-\lambda h}, \dots, -p_1(\lambda) + p_0e^{-\lambda h}, -p_0, 0].$$

To simplify the notation, let 0_m denote a row of m zeros, and define the $1 \times (n-1)m$ polynomial vectors

$$(4.20) \quad \gamma(\lambda) = [p_{n-2}(\lambda), p_{n-3}(\lambda), \dots, p_1(\lambda), p_0],$$

$$(4.21) \quad \hat{\gamma}(\lambda) = [p_{n-3}(\lambda), \dots, p_1(\lambda), p_0, 0_m].$$

Now, $w_M(\lambda)$ can be simply written as

$$(4.22) \quad w_M(\lambda) = [\gamma(\lambda), 0_m]e^{-\lambda h} - [0_m, \gamma(\lambda)]$$

and a similar notation can be obtained for $w_L(\lambda)$, with $\hat{\gamma}(\lambda)$ replacing $\gamma(\lambda)$.

We have proved above that (4.1) and (4.2) imply (4.18) and (4.16) respectively. Since, on the other hand, $w_M(\lambda)v(e^{-\lambda h}) \equiv 0$ and $w_L(\lambda)v(e^{-\lambda h}) \equiv 0$, the reverse implication also holds. We have, therefore, proved the following result.

THEOREM 4.4. *Let $c \in R^n$, $q(\cdot) \in \text{FLT}_2([0, h], R^n)$, $P(\lambda)$ and $v(\mu)$ be given by (3.3) and (3.4), respectively, $\gamma(\lambda)$ and $\hat{\gamma}(\lambda)$ be given by (4.20) and (4.21) respectively. Then*

(a) *the following two identities are equivalent:*

$$(4.23) \quad \begin{aligned} (i) \quad & [c^T + q^T(\lambda)]P(\lambda)v(e^{-\lambda h}) \equiv 0, & \forall \lambda \in \mathbb{C} \\ \text{and} \quad & \\ (ii) \quad & [c^T + q^T(\lambda)]P(\lambda) = [\gamma(\lambda), 0_m]e^{-\lambda h} - [0_m, \gamma(\lambda)], & \forall \lambda \in \mathbb{C}; \end{aligned}$$

(b) *similarly, the following two identities are equivalent:*

$$(4.24) \quad \begin{aligned} (i) \quad & q^T(\lambda)P(\lambda)v(e^{-\lambda h}) \equiv 0, & \forall \lambda \in \mathbb{C}, \\ \text{and} \quad & \\ (ii) \quad & q^T(\lambda)P(\lambda) = [\hat{\gamma}(\lambda), 0_m]e^{-\lambda h} - [0_m, \hat{\gamma}(\lambda)], & \forall \lambda \in \mathbb{C}. \end{aligned}$$

Putting together Propositions 2.5 and 2.6 as well as Theorem 4.4, one then has, via (3.5):

COROLLARY 4.5. *The system (1.2)—i.e., (1.1)—is:*

(i) *M_2 -approximately controllable if and only if (4.23)(i)—or equivalently, (4.23)(ii)—with $c \in R^n$ and $q(\lambda) \in \text{FLT}_2([0, h], R^n)$ implies $c = 0$ and $q(\lambda) \equiv 0$.*

(ii) *L_2 -approximately controllable if and only if (4.24)(i)—or, equivalently, (4.24)(ii)—with $q(\lambda) \in \text{FLT}_2([0, h], R^n)$ implies $q(\lambda) \equiv 0$.*

Note that under the assumption $\text{rank}_C P(\lambda) = n$, which implies the invertibility of an $n \times n$ minor of the matrix $P(\lambda)$, Theorem 4.4 characterizes the class of c^T and $q^T(\lambda)$ that annihilate $\Delta^{-1}(\lambda)B$. For instance, if $m = 1$ and $\det P(\lambda) \neq 0$, the statement (b) of Theorem 4.4 implies that any entire function $q(\lambda) \in \text{FLT}_2([0, h], R^n)$ that annihilates $\Delta^{-1}(\lambda)B$ is of the form $q^T(\lambda) = w_L(\lambda)P^{-1}(\lambda)$, where $w_L(\lambda)$ contains yet unspecified constants and can be null. Thus, at this stage, the test for L_2 -approximate controllability can consist of computing $w_L(\lambda)P^{-1}(\lambda)$ and checking—e.g., via the Paley–Wiener theorem reported as Theorem 2.8—whether there is a nonzero function $w_L(\lambda)P^{-1}(\lambda)$ of class $\text{FLT}_2([0, h], R^n)$. A similar technique can be used to test for M_2 -approximate controllability.

Example 4.1. Consider the system given in Example 3.1. Since $n = 3$, $m = 1$ one has $w_L(\lambda) = [p_0 e^{-\lambda h}, -p_0, 0]$ where p_0 is a scalar. The inverse of $P(\lambda)$ is

$$P^{-1}(\lambda) = \frac{1}{2\lambda^3} \begin{bmatrix} 0 & -2 & -2\lambda \\ 0 & -2 + 2\lambda - \lambda^2 & -2\lambda + 2\lambda^2 \\ -\lambda^3 & -\lambda^3 - 4\lambda^2 + 4\lambda - 2 & -3\lambda^3 + 4\lambda^2 - 2\lambda \end{bmatrix}.$$

For the product $w_L(\lambda)P^{-1}(\lambda)$ one obtains

$$w_L(\lambda)P^{-1}(\lambda) = \frac{p_0}{2} \left[0, -\frac{-2 + 2\lambda - \lambda^2 + 2e^{-\lambda h}}{\lambda^3}, \frac{2(\lambda - \lambda^2 - \lambda e^{-\lambda h})}{\lambda^3} \right],$$

which, except for the arbitrary constant p_0 , coincides with the $q^T(\lambda)$ given in Example 3.1. But for $h = 1$ that $q^T(\lambda)$ was shown to be of class $\text{FLT}_2([0, 1], R^3)$. This means that the present system is *not* L_2 -approximately controllable for $h = 1$. However, we may also observe that for all delays $h \neq 1$ the function $w_L(\lambda)P^{-1}(\lambda)$ written above is

not entire (unless $p_0 = 0$), thus not of class $\text{FLT}_2([0, h], R^3)$ (the triple zero at $\lambda = 0$ of the denominator is not canceled by the numerator, unless $h = 1$), so that the present system is indeed L_2 -approximately controllable for all delay values $h \neq 1$. So, the value $h = 1$ turns out to be an isolated pathological case for the system in question. Further analysis of this example is given in § 7.

Example 4.2. Consider the system (1.1) with $m = 1$, $A_0 = \alpha I$, $-\infty < \alpha < \infty$ and A_1, b in the controllable canonical form [35, p. 105]. Computing $P(\lambda)$, we find

$$(4.25) \quad P(\lambda) = \begin{bmatrix} 0 & . & . & . & . & 0 & 1 \\ . & . & . & . & . & \lambda - \alpha & 0 \\ . & . & . & . & . & . & . \\ 0 & . & . & (\lambda - \alpha)^{n-2} & . & . & . \\ (\lambda - \alpha)^{n-1} & 0 & . & . & . & . & 0 \end{bmatrix}$$

so that $\text{rank}_{\mathbb{C}} P(\lambda) = n$. Theorem 4.4(a) gives for this example

$$\begin{aligned} c_1 + q_1(\lambda) &= -p_0, \\ c_2 + q_2(\lambda) &= \frac{1}{\lambda - \alpha} (-p_1(\lambda) + p_0 e^{-\lambda h}), \\ &\vdots \\ c_{n-1} + q_{n-1}(\lambda) &= \frac{1}{(\lambda - \alpha)^{n-2}} (-p_{n-2}(\lambda) + p_{n-3}(\lambda) e^{-\lambda h}), \\ c_n + q_n(\lambda) &= \frac{1}{(\lambda - \alpha)^{n-1}} p_{n-2}(\lambda) e^{-\lambda h}. \end{aligned}$$

Considering the last row and taking λ real $\rightarrow +\infty$, we obtain $c_n = 0$; hence

$$q_n(\lambda) = \frac{p_{n-2}(\lambda)}{(\lambda - \alpha)^{n-1}} e^{-\lambda h}.$$

Since this function has a singularity at $\lambda = \alpha$ unless $p_{n-2}(\lambda) \equiv 0$, we conclude that $q_n(\lambda) \equiv p_{n-2}(\lambda) \equiv 0$. Proceeding to the next-to-last row we face the same situation with n replaced by $n - 1$. Thus $c_{n-1} \equiv q_{n-1}(\lambda) \equiv p_{n-3}(\lambda) \equiv 0$. After n steps we obtain $c^T = 0$ and $q^T(\lambda) \equiv 0$, which proves that the present system is M_2 -approximately controllable. Since M_2 -approximate controllability does not depend upon the choice of coordinates in R^n , we have in fact proved that: *every system (1.1) with $m = 1$, $A_0 = \alpha I$, $-\infty < \alpha < \infty$, and with the pair (A_1, b) controllable [35, p. 105] is M_2 -approximately controllable for any value of $h > 0$.* This result will be extended to more general classes of systems in § 5.

Conversely, if the system (1.1) with $A_0 = \alpha I$, $-\infty < \alpha < \infty$, and m arbitrary is L_2 -approximately controllable for some value of $h > 0$, then the pair (A_1, B) is controllable. This follows from Theorem 3.1 and Corollary 3.5(i).

The above examples illustrate that one can use directly Theorem 4.4 and Corollary 4.5 as a test for controllability. However, generally speaking, the inversion of the polynomial matrix $P(\lambda)$ may represent an unpleasant obstacle. In view of this, we will use Theorem 4.4 and Corollary 4.5 as a basis for subsequent investigations aimed at deriving simpler conditions. We begin with some immediate consequences of Theorem 4.4.

COROLLARY 4.6. *Let $n \leq 2$. Then the condition $\text{rank}_C P(\lambda) = n$ is sufficient (as well as necessary by Theorem 3.1) for L_2 -approximate controllability of the system (1.2)—i.e., (1.1)—, for all values of the delay $h > 0$.*

Proof. For $n \leq 2$, we have $q^T(\lambda)P(\lambda) \equiv w_L(\lambda) \equiv 0$. For $n = 1$, this was already observed at the beginning of § 4, after (4.2'). From $n = 2$, this follows from (4.18) and (4.19). Since there is a λ_0 such that $\text{rank } P(\lambda_0) = n$, one then has $\text{rank } P(\lambda) = n$ for $|\lambda - \lambda_0| < \delta$, for some $\delta > 0$, but then $q(\lambda) \equiv 0$ for $|\lambda - \lambda_0| < \delta$, and, by analytic continuation, $q(\lambda) \equiv 0$. Q.E.D.

Example 4.3. This is the example in which both the Euclidean and the L_2 -approximate controllability hold, but the M_2 -approximate controllability does not obtain.

Let $n = 2$, $m = 1$, and

$$A_0 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad A_1 = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 3 \end{bmatrix}.$$

We obtain

$$\begin{aligned} \Delta(\lambda) &= \begin{bmatrix} \lambda - 1 + e^{-\lambda h} & -1 \\ 0 & \lambda - 2 - e^{-\lambda h} \end{bmatrix}, \\ \det \Delta(\lambda) &= (\lambda - 1 + e^{-\lambda h})(\lambda - 2 - e^{-\lambda h}) \\ \Delta^{-1}(\lambda) &= \frac{1}{\det \Delta(\lambda)} \begin{bmatrix} \lambda - 2 - e^{-\lambda h} & 1 \\ 0 & \lambda - 1 + e^{-\lambda h} \end{bmatrix} \\ P(\lambda) &= \begin{bmatrix} \lambda + 1 & -1 \\ 3(\lambda - 1) & 3 \end{bmatrix}; \quad \det P(\lambda) = 6\lambda. \end{aligned}$$

The necessary condition $\text{rank}_C P(\lambda) = 2$ is satisfied. Since $n = 2$, by Corollary 4.6 we obtain that the system is L_2 -approximately controllable for all $h > 0$. It is also easy to verify that the system is Euclidean controllable (see, e.g., § 6).

Consider now the M_2 -approximate controllability property. By Theorem 4.4(a) we have ($m = 1$)

$$[c^T + q^T(\lambda)]P(\lambda) = [p_0 e^{-\lambda h}, -p_0] \quad \text{or} \quad c^T + q^T(\lambda) = p_0 [e^{-\lambda h}, -1] P^{-1}(\lambda),$$

which gives

$$c^T + q^T(\lambda) = \frac{p_0}{6} \left[3 - 3 \frac{1 - e^{-\lambda h}}{\lambda}, -1 - \frac{1 - e^{-\lambda h}}{\lambda} \right].$$

Since the function $(1 - e^{-\lambda h})/\lambda$ is of class $\text{FLT}_2([0, h], \mathbb{R})$ for all values of h , (see Example 2.1), then we can always choose $p_0 = 6$, $c^T = [3, -1]$ and

$$q^T(\lambda) = - \left[3 \frac{1 - e^{-\lambda h}}{\lambda}, \frac{1 - e^{-\lambda h}}{\lambda} \right].$$

This shows that even though the present system is both Euclidean controllable and L_2 -approximately controllable for all $h > 0$, it is not M_2 -approximately controllable for any value of $h > 0$.

COROLLARY 4.7. *Let the $n \times nm$ matrix $P(\lambda)$ contain an $n \times n$ submatrix $P_{n \times n}(\lambda)$ with determinant identically equal to a nonzero constant α , so that $\text{rank}_C P(\lambda) = n$. Then the system (1.2)—i.e., (1.1)—is M_2 -approximately controllable for all values of the delay $h > 0$.*

Remark 4.2. That the class of systems of type (1.1) satisfying the assumptions of Corollary 4.7 is nonempty can be seen from Examples 5.1, 5.3 and Theorem 5.4.

Proof of Corollary 4.7. From (4.16) we have

$$(4.26) \quad q^T(\lambda) = \frac{w'_M(\lambda) \operatorname{adj} P_{n \times n}(\lambda) - \alpha c^T}{\alpha}$$

where $w'_M(\lambda)$ denotes the $1 \times n$ subvector of $w_M(\lambda)$ in (4.17) corresponding to the submatrix $P_{n \times n}(\lambda)$. But the right hand side of (4.26) is a polynomial in λ and $e^{-\lambda h}$ and, by the Paley–Wiener Theorem, it must be identically zero in $\lambda \in \mathbb{C}$ in order to be an $\operatorname{FLT}_2([0, h], R^n)$ -function. So $q^T(\lambda) \equiv 0$. It then follows from (4.23)(i) with $q^T(\lambda) \equiv 0$ that $c^T P(\lambda) v(\mu) \equiv 0$, $\forall \lambda, \mu \in \mathbb{C}$ (see Remark 4.1). Hence $c^T = 0$, as it follows by employing the argument leading from (3.16) to (3.19). So M_2 -approximate controllability is established by virtue of Corollary 4.5(i). Q.E.D.

We now proceed to show that Theorem 4.4 and Corollary 4.5 reduce to a test involving a system of linear homogeneous equations. For simplicity of notation we consider first the case of $m = 1$; the extension to the case $m > 1$ is straightforward.

From (4.23)(ii) one obtains

$$(4.27) \quad c^T + q^T(\lambda) = w_M(\lambda) \operatorname{adj} P(\lambda) / \det P(\lambda).$$

Let $\Gamma(\lambda)$ and $\psi(\lambda)$ denote the $n \times n$ polynomial matrix and the scalar polynomial, respectively, obtained from $\operatorname{adj} P(\lambda)$ and $\det P(\lambda)$ by deleting all possible common divisors. Let N denote the degree of $\psi(\lambda)$. Obviously $N \leq \deg \det P(\lambda) \leq n(n-1)/2$ (see § 5) and $\operatorname{adj} P(\lambda) / \det P(\lambda) = \Gamma(\lambda) / \psi(\lambda)$. From (4.27) we now have

$$(4.28) \quad \begin{aligned} q^T(\lambda) &= \frac{1}{\psi(\lambda)} [w_M(\lambda) \Gamma(\lambda) - c^T \psi(\lambda)] \\ &= \frac{1}{\psi(\lambda)} [g_1^T(\lambda) + g_2^T(\lambda) e^{-\lambda h}] \end{aligned}$$

where, from (4.22), $g_i^T(\lambda)$ are $1 \times n$ row vector polynomials given by

$$(4.29) \quad g_1^T(\lambda) = -[0, \gamma(\lambda)] \Gamma(\lambda) - c^T \psi(\lambda),$$

$$(4.30) \quad g_2^T(\lambda) = [\gamma(\lambda), 0] \Gamma(\lambda).$$

In order that $q^T(\lambda)$ given by (4.28) be of class $\operatorname{FLT}_2([0, h], R^n)$ it must be (i) entire, that is all the zeros of $\psi(\lambda)$ must be canceled by zeros of $g_1^T(\lambda) + g_2^T(\lambda) e^{-\lambda h}$; also, by the Paley–Wiener Theorem 2.8, it must be (ii) square integrable on the imaginary axis, which implies that the polynomials $g_i(\lambda)$ have degrees at most $N-1$. Therefore, we must have

$$(4.31) \quad g_1(\lambda) = \sum_{i=0}^{N-1} \alpha_i \lambda^i,$$

$$(4.32) \quad g_2(\lambda) = \sum_{i=0}^{N-1} \beta_i \lambda^i$$

where $\alpha_i, \beta_i \in R^n$ are vectors of unknown coordinates, which must satisfy the cancellation conditions mentioned above.

Conversely, any entire function of the form given by (4.28), (4.31), (4.32) is of class $\operatorname{FLT}_2([0, h], R^n)$, because (i) it is $O(1/|\lambda|)$ on the imaginary axis for large $|\lambda|$; (ii) the constants H and H' referred to in Theorem 2.8 are, in this case $H' = 0$, $H = h$; (iii)

the inverse Laplace transform of (4.28) (see, e.g., [1]) for $t \in [0, h]$ is a combination of functions $t^l e^{\lambda_k t}$, $0 \leq l \leq l_k$, $k = 1, 2, \dots$ (where λ_k denote zeros of $\psi(\lambda)$ and l_k their multiplicities), so it certainly is in $L_2([0, h], R^n)$.

We have, therefore, obtained a general formula for $q(\lambda)$ with as yet unknown coefficients α_i, β_i . Similarly, the $n-1$ row vector polynomial $\gamma(\lambda)$ in (4.20) can be characterized by a finite number of unknown coefficients, for letting

$$p_k(\lambda) = \sum_{i=0}^k \gamma_{k,i} \lambda^i \quad k = 0, 1, \dots, n-2,$$

where $\gamma_{k,i}$ are scalar coefficients, one obtains from (4.20)

$$(4.33) \quad \gamma(\lambda) = \left[\sum_{i=0}^{n-2} \gamma_{n-2,i} \lambda^i, \dots, \sum_{i=0}^1 \gamma_{1,i} \lambda^i, \gamma_{0,0} \right].$$

Substituting (4.28) into (4.23)(ii) and making use of Remark 4.1, one obtains the following system of polynomial equations

$$(4.34) \quad \{c^T P(\lambda) + [0, \gamma(\lambda)]\} \psi(\lambda) + g_1^T(\lambda) P(\lambda) = 0, \quad \forall \lambda \in \mathbb{C}$$

$$(4.35) \quad [\gamma(\lambda), 0] \psi(\lambda) - g_2^T(\lambda) P(\lambda) = 0, \quad \forall \lambda \in \mathbb{C}.$$

In addition, the cancellation condition mentioned before gives, for a single root λ_k of $\psi(\lambda)$, the equation

$$(4.36) \quad g_1(\lambda_k) + g_2(\lambda_k) e^{-\lambda_k h} = 0$$

while for a root λ_k of $\psi(\lambda)$ of multiplicity l_k it gives the equations

$$(4.37) \quad \frac{d^l}{d\lambda^l} [g_1(\lambda) + g_2(\lambda) e^{-\lambda h}]_{\lambda=\lambda_k} = 0, \quad l = 0, 1, \dots, l_k - 1.$$

In the above equations, the unknowns are $c \in R^n$, $g_i(\lambda)$, $i = 1, 2$, characterized by (4.31), (4.32), and $\gamma(\lambda)$ characterized by (4.33), while the matrix $P(\lambda)$, the polynomial $\psi(\lambda)$ and its roots λ_k are presumed known.

Generalization of this procedure to the case $m > 1$ requires only minor modifications; in particular, $P(\lambda)$ should be replaced by any nonsingular $n \times n$ submatrix of $P(\lambda)$, and the unknown coefficients $\gamma_{k,i}$ become $1 \times m$ row vector valued.

We summarize our development below.

THEOREM 4.8. 1) A necessary and sufficient condition that the system (1.1) be M_2 -approximately controllable is that (i) $\text{rank}_{\mathbb{C}} P(\lambda) = n$, (ii) the system of equations (4.34)–(4.37) with unknown $c \in R^n$ and unknown polynomials $g_1(\lambda)$, $g_2(\lambda)$, $\gamma(\lambda)$ of the form (4.31) (4.32) (4.33), respectively, has only the null solution $c = 0$, $g_1(\lambda) \equiv g_2(\lambda) \equiv 0$, $\gamma(\lambda) \equiv 0$. 2) A necessary and sufficient condition that the system (1.1) be L_2 -approximately controllable is obtained following the same procedure described in 1), except that the vector $\gamma(\lambda)$ in (4.20) is replaced by vector $\hat{\gamma}(\lambda)$ in (4.21) and the vector c is set a priori equal to zero.

A few comments about using this theorem are in order. The system of equations (4.34) to (4.37) can be rewritten as a system of algebraic linear homogeneous equations with respect to unknown coefficients c_i , $i = 1, \dots, n$, $\alpha_i, \beta_i \in R^n$, $i = 1, \dots, N$, $\gamma_{k,i} \in R^m$, $k = 0, \dots, n-2$, $i = 0, \dots, k$. Therefore the necessary and sufficient condition given by Theorem 4.8 reduces to a test on the rank of the numerical matrix corresponding to the system of homogeneous equations. The system (1.1) is M_2 -approximately controllable if and only if the rank of this matrix is full.

The dimensions of such matrix in case $m = 1$ are as follows. The number of unknowns is $2nN + n(n+1)/2$ (n coefficients c_i , $1 + \dots + n - 1$ coefficients $\gamma_{k,i}$, $2nN$ coefficients of the vector α_i, β_i). The number of equations is $3nN + n^2$ ((4.34) is equivalent to $nN + n(n+1)/2$ algebraic equations, (4.35) and (4.36) give $nN + n(n-1)/2$ and nN equations respectively). The matrix is thus $(3nN + n^2) \times (2nN + n(n+1)/2)$. Since N ranges between 0 (case of Corollary 4.7) and $n(n-1)/2$ (see § 5), the dimensions of the matrix can be of order n^3 , especially if $N = n(n-1)/2$, which will be seen to correspond to the case where (A_1, b) is a controllable pair.

The procedure for testing M_2 -approximate controllability is summarized below (for $m = 1$)

Step 1. Compute $P(\lambda)$

Step 2. Check $\text{rank}_C P(\lambda) = n$ and compute $\det P(\lambda)$. Reduce $\det P(\lambda)$ to $\psi(\lambda)$, or set directly $\psi(\lambda) = \det P(\lambda)$ (the reduction is not mandatory).

Step 3. Compute the roots λ_k of $\psi(\lambda)$.

Step 4. Write the system of linear homogeneous equations corresponding to (4.34)–(4.37).

Step 5. Test whether the matrix of this system has full rank.

The same procedure in which the vector c is set a priori equal to zero and the vector $\gamma(\lambda)$ in (4.20) is replaced by the vector $\hat{\gamma}(\lambda)$ in (4.21) gives the rest for L_2 -approximate controllability.

Example 4.4. This example illustrates the use of Theorem 4.8 and also shows a nontrivial function $w_M(\lambda)$.

$$A_0 = \begin{bmatrix} 1 & 0 & 1 \\ 2 & 0 & 0 \\ 0 & 1 & 3 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

h is arbitrary.

Step 1. Computation of $P(\lambda)$ gives

$$P(\lambda) = \begin{bmatrix} \lambda & 0 & 0 \\ 2 & \lambda - 1 & -1 \\ \lambda(\lambda - 1) & -\lambda & 0 \end{bmatrix}.$$

Step 2. $\det P(\lambda) = -\lambda^2$. However, $\text{adj } P(\lambda)$ contains λ as a factor

$$\text{adj } P(\lambda) = \begin{bmatrix} -\lambda & 0 & 0 \\ -\lambda(\lambda - 1) & 0 & \lambda \\ -\lambda[2 + (\lambda - 1)^2] & \lambda^2 & \lambda(\lambda - 1) \end{bmatrix}.$$

Therefore, we take $\psi(\lambda) = \lambda$ (one can avoid the computation of $\text{adj } P(\lambda)$ and take $\psi(\lambda) = \det P(\lambda)$; this merely would increase the number of unknowns and equations).

Step 3. The only root of $\psi(\lambda)$ is $\lambda_1 = 0$.

Step 4. $N = 1$, therefore the maximum number of unknowns and equations is $2nN + n(n+1)/2 = 12$, $3nN + n^2 = 18$, respectively. Both numbers will turn out to be smaller, due to the sparsity of the equations.

$$g_1(\lambda) = \alpha_0 \in \mathbb{R}^3, \quad g_2(\lambda) = \beta \in \mathbb{R}^3;$$

$$q_i(\lambda) = \frac{\alpha_0^i + \beta_0^i e^{-\lambda h}}{\lambda}, \quad i = 1, 2, 3,$$

where α_0^i, β_0^i denote the coordinates of α_0, β_0 . Now, using (4.36) we obtain $\alpha_0^i +$

$\beta_0^i e^0 = 0$, $i = 1, 2, 3$, which gives $\beta_0^i = -\alpha_0^i$ and eliminates 3 unknowns. Since $\gamma(\lambda) = [\gamma_{10} + \gamma_{11}\lambda, \gamma_{00}]$, there are 9 unknowns: $c_1, c_2, c_3, \alpha_0^1, \alpha_0^2, \alpha_0^3, \gamma_{10}, \gamma_{11}, \gamma_{00}$. Writing (4.34) and (4.35) in detail and equating coefficients of powers of λ to zero, one obtains $\alpha_0^2 = c_3 = 0$ and the following system of equations:

$$\begin{bmatrix} 0 & 2 & 1 & -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & -1 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \alpha_0^1 \\ \alpha_0^3 \\ \gamma_{10} \\ \gamma_{11} \\ \gamma_{00} \end{bmatrix} = 0.$$

Step 5. The 8×7 matrix of this system has rank less than 7, so that there is a nonzero solution and the system is not M_2 -approximately controllable.

Computing a solution corresponding to $\gamma_{00} = 1$, one obtains

$$c^T + q^T(\lambda) = \left[-1 - \frac{1 - e^{-\lambda h}}{\lambda}, 1, \frac{1 - e^{-\lambda h}}{\lambda} \right]$$

and

$$w_M(\lambda) = [(2 - \lambda)e^{-\lambda h}, -(2 - \lambda) + e^{-\lambda h}, -1].$$

The latter formula shows that the degrees of polynomials $p_k(\lambda)$ appearing in $\gamma(\lambda)$ in Theorem 4.4 cannot be further reduced. The computations involved in solving this example also show that many of all possible $3nN + n^2$ equations are just identities $0 = 0$, so that the actual effort involved in using Theorem 4.8 is significantly smaller than the maximum number of equations would indicate.

One can check that the present system is L_2 -approximately controllable for all values of $h > 0$.

5. Algebraic sufficient conditions for M_2 - and L_2 -approximate controllability for all values of $h > 0$. The purpose of this section is to show that a further study of results obtained in § 4 leads to some sufficient conditions involving directly the matrices A_0, A_1, B , which guarantee the M_2 - or L_2 -approximate controllability for all values of the delay $h > 0$. These conditions, based on some triangularity properties of $P(\lambda)$, show an interesting role played by canonical forms and invariant subspaces of the matrices A_i in the function space controllability problem.

Since the reasoning leading to conditions that involve A_0, A_1, B is based on linear algebra methods and is quite technical, only a summary of the main results will be presented here, while the detailed proofs are deferred to another paper [44] (see also [43]).

At the beginning of this section we discuss the role of triangularity properties of $P(\lambda)$ [Theorems 5.1 and 5.2], and later we show that these properties are enjoyed by quite large classes of systems. For the sake of clarity of exposition we shall mostly consider the case $m = 1$.

Remark 5.1. Observe that a change of coordinates in R^n , $y = Tz$, with T an $n \times n$ nonsingular matrix, transforms the matrices A_0, A_1, B and $P(\lambda)$ into $TA_0T^{-1}, TA_1T^{-1}, TB$ and $TP(\lambda)$ respectively; this means that the coordinate transformations amount to constant (independent of λ) operations on rows of $P(\lambda)$ or, in other words,

to a *premultiplication* of $P(\lambda)$ by a constant nonsingular $n \times n$ matrix T . On the other hand, if we define $\hat{c}^T = c^T T$, $\hat{q}^T(\lambda) = q^T(\lambda) T$ the coordinate transformations amount to modifications of c and $q(\lambda)$ by a linear transformation. In particular, when we apply Corollary 4.5, row permutations in $P(\lambda)$ correspond to reindexing the components of $y_i(t)$, c_i and $q_i(\lambda)$. Note also that the properties of function space controllability are unaffected under coordinate transformation.

THEOREM 5.1. *Let $m = 1$ and assume $\text{rank}_{\mathbb{C}} P(\lambda) = n$. Suppose, further, that the $n \times n$ matrix $P(\lambda)$ can, by a premultiplication by a constant nonsingular $n \times n$ matrix T , be transformed into either (a) a right-triangular matrix or (b) a left-triangular matrix. Denote $TP(\lambda)$ by $\bar{P}(\lambda)$. Then in case (a) the system (1.1) is M_2 -approximately controllable for any value of the delay $h > 0$, and, in case (b), the system is L_2 -approximately controllable for any value of $h > 0$ and it is actually M_2 -approximately controllable for any value of $h > 0$ if, in addition, the diagonal elements of $\bar{P}(\lambda)$ are all constant.³*

Proof. We have, by assumption

$$\bar{P}(\lambda) = \begin{bmatrix} \beta_{n-1}(\lambda) & * & * \\ 0 & \beta_{n-2}(\lambda) & * \\ 0 & 0 & \beta_0 \end{bmatrix}$$

in case (a), and

$$\bar{P}(\lambda) = \begin{bmatrix} \alpha_{n-1}(\lambda) & 0 & 0 \\ * & \alpha_{n-2}(\lambda) & 0 \\ * & * & \alpha_0 \end{bmatrix}$$

in case (b).

Also, according to the general properties of $P(\lambda)$ the diagonal entries $\beta_{n-1}(\lambda), \dots, \beta_1(\lambda), \beta_0$ are scalar polynomials of degree at most $(n-1), \dots, 1, 0$ respectively. Similarly for the scalar polynomials $\alpha_{n-1}(\lambda), \dots, \alpha_1(\lambda), \alpha_0$.

Case (a): First, the assumption $\text{rank}_{\mathbb{C}} P(\lambda) = n$ implies $\beta_i(\lambda) \not\equiv 0$, $i = 0, 1, \dots, n-1$. The first column of $\bar{P}(\lambda)$, premultiplied by $c^T + q^T(\lambda) = [c_1 + q_1(\lambda), \dots, c_n + q_n(\lambda)]$ gives, in view of (4.16) and (4.17):

$$c_1 + q_1(\lambda) = \frac{p_{n-2}(\lambda)}{\beta_{n-1}(\lambda)} e^{-\lambda h}.$$

Letting λ be real $\rightarrow +\infty$ we obtain $c_1 = 0$ (Lemma 4.2(i)), and by Lemma 4.3, it follows that $q_1(\lambda) \equiv 0$. Hence $p_{n-2}(\lambda) \equiv 0$. In general, we have from (4.16)

$$(5.1) \quad c_i + q_i(\lambda) = \frac{p_{n-(i+1)}(\lambda)}{\beta_{n-i}(\lambda)} e^{-\lambda h}, \quad i = 2, \dots, n-1,$$

since $p_{n-i}(\lambda) \equiv 0$ from previous inductive step. Repetition of the above argument yields $c_i = 0$ and $q_i(\lambda) \equiv 0$ and hence $p_{n-(i+1)}(\lambda) \equiv 0$, $i = 2, \dots, n-1$. The last term is

$$c_n + q_n(\lambda) = \frac{-p_0}{\beta_0}$$

where $p_0 = 0$ from previous inductive step. Hence $c_n = 0$ and $q_n(\lambda) = 0$. We have thus

³ In agreement with Corollary 4.7: see also Remark 4.2.

shown that (4.16) implies $c = 0$ and $q(\lambda) \equiv 0$. In view of Corollary 4.5(i), M_2 -approximate controllability is established in case (a), for all $h > 0$.

Case (b): By the assumption $\text{rank}_C P(\lambda) = n$ we have $\alpha_i(\lambda) \neq 0$, $i = 0, \dots, n$. Let us compute now $q^T(\lambda)\bar{P}(\lambda)$: the last column of $\bar{P}(\lambda)$ premultiplied by $q^T(\lambda)$ gives, by (4.18) and (4.19): $q_n(\lambda)\alpha_0 = 0$ and hence $q_n(\lambda) \equiv 0$. The next to the last column in (4.18) then becomes

$$q_{n-1}(\lambda) = -\frac{p_0}{\alpha_1(\lambda)}.$$

Hence $q_{n-1}(\lambda) \equiv 0$ and so $p_0 = 0$, otherwise $q_{n-1}(\lambda)$ would fail to be of class $\text{FLT}_2([0, h], R)$. In general, from (4.18) we have

$$(5.2) \quad q_{n-i}(\lambda)\alpha_i(\lambda) = -p_{i-1}(\lambda), \quad \alpha_i(\lambda) \neq 0, \quad i = 2, \dots, n-2,$$

since $p_{i-2}(\lambda) \equiv 0$ from the previous inductive step, while the last term is

$$(5.3) \quad q_1(\lambda)\alpha_{n-1}(\lambda) \equiv 0, \quad \alpha_{n-1}(\lambda) \neq 0.$$

Identity (5.2) implies $q_{n-i}(\lambda) \equiv 0$ and $p_{i-1}(\lambda) \equiv 0$, since otherwise $q_{n-i}(\lambda)$ would have at least one pole and could not be of class $\text{FLT}_2([0, h], R)$, $i = 2, \dots, n-2$. Finally, (5.3) gives $q_1(\lambda) \equiv 0$. We have thus shown that (4.18) implies $q(\lambda) \equiv 0$. In view of Corollary 4.5(ii), L_2 -approximate controllability is established in case (b), for all $h > 0$. It remains to show that if, in addition, $\alpha_i(\lambda)$ are all constants $\alpha_i(\lambda) \equiv \alpha_i$, $i = 0, \dots, n-1$, then the system (1.2) is actual M_2 -approximately controllable for all values of $h > 0$. This can be done either directly, with the use of (4.23)(ii), Lemma 4.3' and a suitable adaptation of the above argument (see [43, Appendix B]), or it follows from Corollary 4.7. Q.E.D.

Remark 5.2. We wish to stress that Theorem 5.1 provides only a sufficient condition which may in fact fail if the 'off' triangular terms are present in the matrix $P(\lambda)$ (see Example 3.1 for $h = 1$).

Example 5.1. Consider the system (1.1) with $m = 1$, $A_1 = \alpha I$, $\alpha \neq 0$ and the pair (A_0, b) is in the controllable canonical form [35, p. 105]. We then compute only the last column of the matrix $\text{adj} (I\lambda - A_0 - A_1\mu)$, so that $P(\lambda)$ is given by

$$P(\lambda) = \begin{bmatrix} 1 & 0 & & & 0 \\ \lambda & (-\alpha) & & & \\ \lambda^2 & -2\lambda & (-\alpha)^2 & & \\ \vdots & & & & 0 \\ \lambda^{n-1} & & (-\alpha)^l \binom{n-1}{l} \lambda^{n-1-l} & & (-\alpha)^{n-1} \end{bmatrix}$$

and Theorem 5.1(b) applies. Since M_2 -approximate controllability as well as the above matrix A_1 do not depend upon the choice of coordinates in R^n , we have proved that:

every system (1.1) with $m = 1$, $A_1 = \alpha I$, $\alpha \neq 0$, and with the pair (A_0, b) controllable is M_2 -approximately controllable for any value of $h > 0$. Conversely, if (1.1) with $A_1 = \alpha I$, $\alpha \neq 0$, m arbitrary is L_2 -approximately controllable for some $h > 0$, then (A_0, B) is controllable.

This follows from Theorem 3.1 and Corollary 3.5(ii).

An extension of Theorem 5.1 to the case $m > 1$ is indeed possible, but cumbersome to treat in full generality. Therefore, for sake of clarity, we confine ourselves

here to only two concrete cases, which turn out to be typical of entire classes of systems. They will exhibit all the essential features needed in the sought after generalization. For more details see [43].

Example 5.2. An illustrative case for $m > 1$. Motivated by the case $m = 1$ in Example 4.2, we let

- (i) $A_0 = \alpha I$, $-\infty < \alpha < \infty$;
- (ii) the pair (A_1, B) be controllable.

Then, by operating a preliminary change of coordinates on R^n , which leaves A_0 unaltered, we may assume that A_1 and B are in the canonical form (7), (8), (9) of [38, p. 291]. For the sake of clarity of exposition, take:

- (iii) $n = 8$, $m = 3$ and the blocks occurring in the canonical form of size $l_1 = l_3 = 3$ and $l_2 = 2$.

Claim: The system (1.1) defined by the triple $\{A_0, A_1, B\}$ satisfying (i)–(iii) is M_2 -approximately controllable for any value of $h > 0$.

Indeed, we first compute the $n \times nm$ matrix $P(\lambda)$ corresponding to A_0, A_1, B with A_1, B in said canonical form from its definition: $P(\lambda)v(\mu) = \text{adj}(\lambda I - A_0 - \mu A_1)B$ (only the first, third and fifth column of $\text{adj}((\lambda - \alpha)I - \mu A_1)$ need be computed). Tedious but straightforward computations yield⁴ the matrix $P(\lambda)$ given by

$$P(\lambda) = \begin{bmatrix} (\lambda - \alpha)^7 + \text{l.o.t.} & 0 & 0 & * & & & & \\ 0 & 0 & 0 & (\lambda - \alpha)^6 + \text{l.o.t.} & & & & \\ 0 & 0 & 0 & 0 & & & & \\ \hline 0 & (\lambda - \alpha)^7 + \text{l.o.t.} & 0 & 0 & & & & \\ 0 & 0 & 0 & 0 & & & & \\ \hline 0 & 0 & (\lambda - \alpha)^7 + \text{l.o.t.} & 0 & & & & \\ 0 & 0 & 0 & 0 & & & & \\ 0 & 0 & 0 & 0 & & & & \\ \hline & 0 & 0 & * & 0 & 0 & & \\ & 0 & 0 & * & & & & \\ & 0 & 0 & (\lambda - \alpha)^5 + \text{l.o.t.} & 0 & 0 & & \\ \hline * & 0 & 0 & * & 0 & & & \\ (\lambda - \alpha)^6 + \text{l.o.t.} & 0 & 0 & * & 0 & & & \\ \hline 0 & * & 0 & 0 & * & & & \\ 0 & (\lambda - \alpha)^6 + \text{l.o.t.} & 0 & 0 & * & & & \\ 0 & 0 & 0 & 0 & 0 & (\lambda - \alpha)^5 + \text{l.o.t.} & & \end{bmatrix}$$

where only the first 9 columns of $P(\lambda)$ out of 24 have been reported. An * denotes a possibly nonzero term and l.o.t. stands for 'lower order term in λ '. It is plain that $\text{rank}_C P(\lambda) = n$. A suitable adaptation of the same argument used in the proof of Theorem 5.1(a) will be now given to substantiate the claim. Denote by $p_{i,1}(\lambda)$, $p_{i,2}(\lambda)$ and $p_{i,3}(\lambda)$ the three scalar components of the 1×3 polynomial vectors $p_i(\lambda)$ of degree i occurring in (4.17). Computation of the first three columns of (4.16) for $P(\lambda)$ gives

$$[c_j + q_j(\lambda)][(\lambda - \alpha)^7 + \text{l.o.t.}] = p_{6,k}(\lambda)e^{-\lambda h}$$

with $j = k = 1$; $j = 4$, $k = 2$; $j = 6$, $k = 3$. Hence, as in the proof of Theorem 5.1(a), it

⁴ A more rational method of computing $P(\lambda)$ is available and is based on the general structure of $P(\lambda)$ discussed later in this section.

follows that $c_1 = c_4 = c_6 = 0$ (by taking $\lambda \rightarrow +\infty$) and $q_1(\lambda) = q_4(\lambda) = q_6(\lambda) \equiv 0$ (by Lemma 4.3), so that $p_{6,k}(\lambda) \equiv 0$. The next three columns of (4.16) then become

$$[c_j + q_j(\lambda)][(\lambda - \alpha)^6 + \text{l.o.t.}] = p_{5,k}(\lambda)e^{-\lambda h}$$

with $j = 2, k = 1; j = 5, k = 2; j = 7, k = 3$. Again, this implies $c_2 = c_5 = c_7 = 0$, $q_2(\lambda) = q_7(\lambda) \equiv 0$ and hence $p_{5,k}(\lambda) \equiv 0$. Finally the seventh and ninth columns of (4.16) become

$$[c_j + q_j(\lambda)][(\lambda - \alpha)^5 + \text{l.o.t.}] = p_{4,k}(\lambda)e^{-\lambda h}$$

with $j = 3, k = 1; j = 8, k = 3$. Hence $c_3 = c_8 = 0$ and $q_3(\lambda) = q_8(\lambda) \equiv 0$ so that $c = 0$ and $q(\lambda) \equiv 0$. By Corollary 4.5(i), the claim is established. Q.E.D.

Example 5.3. Another illustrative case for $m > 1$. Motivated by the case $m = 1$ in Example 5.1, we let

- (i) $A_1 = \alpha I$, $\alpha \neq 0$;
- (ii) the pair (A_0, B) be controllable. Again, by operating a change of coordinates in R^n (which leaves A_1 unaltered), we may assume that A_0 and B are in the canonical form (7), (8), (9) of [38, p. 291]. Take again the concrete case:
- (iii) $n = 8, m = 3, l_1 = l_3 = 3$ and $l_2 = 2$, as before.

Claim: The system (1.1) defined by the triple $\{A_0, A_1, B\}$ satisfying (i)–(iii) is M_2 -approximately controllable for any value of $h > 0$.

Indeed, using the abovementioned canonical form for A_0 and B , compute the 8×24 matrix $P(\lambda)$ from $\text{adj}(\lambda I - A_0 - \mu A_1)B = P(\lambda)v(\mu)$ to arrive at the matrix $P(\lambda)$ given below, where only the last 9 columns of $P(\lambda)$ were reported.

$$P(\lambda) = \begin{bmatrix} \cdots & \binom{7}{5}(-\alpha)^5\lambda^2 & 0 & 0 & \binom{7}{6}(-\alpha)^6\lambda \\ \cdots & \binom{6}{5}(-\alpha)^5\lambda & 0 & 0 & (-\alpha)^6 \\ \cdots & (-\alpha)^5 & 0 & 0 & 0 \\ \cdots & 0 & \binom{7}{5}(-\alpha)^5\lambda^2 & 0 & 0 \\ \cdots & 0 & \binom{6}{5}(-\alpha)^5\lambda & 0 & 0 \\ \cdots & 0 & 0 & \binom{7}{5}(-\alpha)^5\lambda^2 & 0 \\ \cdots & 0 & 0 & \binom{6}{5}(-\alpha)^5\lambda & 0 \\ \cdots & 0 & 0 & (-\alpha)^5 & 0 \\ 0 & 0 & 0 & (-\alpha)^7 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \cdots & \binom{7}{6}(-\alpha)^6\lambda & 0 & 0 & (-\alpha)^7 & 0 \\ \cdots & (-\alpha)^6 & 0 & 0 & 0 & 0 \\ \cdots & 0 & \binom{7}{6}(-\alpha)^6\lambda & 0 & 0 & (-\alpha)^7 \\ \cdots & 0 & (-\alpha)^6 & 0 & 0 & 0 \\ \cdots & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

From this structure of $P(\lambda)$, it is plain that $P(\lambda)$ contains a *full* minor with constant determinant. Application of Corollary 4.7 then yields the claim. Q.E.D.

In the two illustrative examples with $m > 1$ given above, the structure of the corresponding matrix $P(\lambda)$ 'generalizes' respectively the right triangularity and left triangularity form needed in Theorem 5.1 for $m = 1$, in the sense that the idea used in the proof of such theorem still works. For more details see [43].

We now describe an algorithm generating the columns of $P(\lambda)$. An analysis of the structure of $P(\lambda)$ based on this algorithm and described in detail in [44] gives conditions for triangularization of $P(\lambda)$ stated directly in terms of A_0, A_1, B . The results are given in Theorems 5.2 and 5.4 below. By using the Faddeev [18] algorithm to compute $\text{adj}(I\lambda - A)$ one can obtain the following formulas characterizing $P(\lambda)$. Let $\Phi_{i,k}$ denote $n \times n$ matrices defined recursively by

$$(5.4) \quad \phi_{i+1,k} = A_0\Phi_{i,k} + A_1\Phi_{i,k-1} + \theta_{i,k}I \quad i = 1, \dots, n-1,$$

$$(5.5) \quad \theta_{i,k} = -\frac{1}{i} \text{tr} [A_0\Phi_{i,k} + A_1\Phi_{i,k-1}] \quad k = 0, \dots, n-1,$$

$$(5.6) \quad \Phi_{1,k} = \begin{cases} I, & k = 0, \\ 0, & \text{otherwise,} \end{cases}$$

and $\Phi_{i,k} = 0$ for $k < 0$. Define

$$(5.7) \quad Z_{i,k} = \Phi_{i,k}B \quad (n \times m \text{ matrix}).$$

Note that the formulas (5.4) to (5.7) are easily programmable on a computer.⁵ Once $\Phi_{i,k}$ are computed, $Z_{i,k}$ can be obtained via (5.7). Note also that, because of the latter formula, the matrices $Z_{i,k}$ satisfy

$$(5.8) \quad Z_{i+1,k} = A_0Z_{i,k} + A_1Z_{i,k-1} + \theta_{i,k}B, \\ i = 1, \dots, n-1, \quad k = 0, \dots, n-1$$

$$(5.9) \quad Z_{1,k} = \begin{cases} B, & k = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Now, the matrix $P(\lambda)$ is given by

$$(5.10) \quad P(\lambda) = \begin{bmatrix} \sum_{i=1}^n \lambda^{n-i} Z_{i,0}, & \sum_{i=2}^n \lambda^{n-i} Z_{i,1}, & \dots, & \sum_{i=n-1}^n \lambda^{n-i} Z_{i,n-2}, & Z_{n,n-1} \end{bmatrix}.$$

Remark 5.3. By using (5.10) we make the following observation. If $m = 1$ then the coefficient of the highest power of λ in $\det P(\lambda)$, that is the coefficient of λ^N , $N = n(n-1)/2$, is equal to $\det [Z_{1,0}, Z_{2,1}, \dots, Z_{n,n-1}]$ and, by (5.10) is equal zero if and only if $\det [b, A_1b, \dots, A_1^{n-1}b] = 0$. For $m > 1$ a similar observation is valid [43]. The controllability of the pair (A_1, B) is, therefore, sufficient (but not necessary, see Theorem 3.6) to have $\text{rank}_C P(\lambda) = n$.

Let $\text{Im } A_i B$, $i = 0, 1$, denote the image of $A_i B$. In the remaining part of this section we will state the controllability conditions in terms of $\text{Im } A_i B$, which can be then easily translated into rank-type conditions. Instead of doing this for each result, we just remind the reader that if A and B are arbitrary matrices having the same number of rows, the condition $\text{Im } B \subset \text{Im } A$ can be equivalently stated as rank

⁵ A computer program in language APL that generates columns of $P(\lambda)$ for given matrices A_{-1}, A_0, A_1, b of a neutral equation $\dot{x}(t) - A_{-1}\dot{x}(t-h) = A_0x(t) + A_1x(t-h) + Bu(t)$, with A_{-1}, b in control canonical form, was given in [24].

$A = \text{rank } [A, B]$. Note also that all the conditions given below in the present section are preserved under coordinate transformations.

We now give a result that completely characterizes the properties of $P(\lambda)$ required by the Theorem 5.1.

THEOREM 5.2. *Let $m = 1$, $B = b$. A necessary and sufficient condition that the matrix $P(\lambda)$ can, by a premultiplication by a constant nonsingular $n \times n$ matrix T , be transformed into a right triangular matrix, and that it satisfy $\text{rank}_{\mathbb{C}} P(\lambda) = n$ is*

$$(5.11) \quad \begin{aligned} (i) \quad & \text{rank } [b, A_1 b, \dots, A_1^{n-1} b] = n, \\ (ii) \quad & A_0 \text{Im } A_1^j B \subset \sum_{i=0}^j \text{Im } A_1^i B, \quad j = 0, 1, \dots, n-1. \end{aligned}$$

The interpretation of the condition (5.11)(ii) is that all the subspaces $\text{span } \{b, A_1 b, \dots, A_1^j b\}$, $j = 0, \dots, n-1$, be invariant under A_0 .

We now state our first main algebraic sufficient condition.

THEOREM 5.3. *Let $m \geq 1$. If the system (1.1) satisfies the conditions*

$$(5.12) \quad \begin{aligned} (i) \quad & \text{rank } [B, A_1 B, \dots, A_1^{n-1} B] = n, \\ (ii) \quad & A_0 \text{Im } A_1^j B \subset \sum_{i=0}^j \text{Im } A_1^i B, \quad j = 0, \dots, n-2, \end{aligned}$$

then it is M_2 -approximately controllable for any value of $h > 0$.

Remark 5.4. Theorem 5.3 is a substantial generalization of a result by Zmood [69, Thm. 3.12], who, in the special case of Theorem 5.3 with $A_0 = 0$, pair (A_1, B) controllable, $h = 1$, claims only L_2 -approximate controllability.

If the pair (A_1, B) is not controllable, but $\text{rank } [A_1, B] = n$, (the necessary condition of Theorem 3.6), one can try another sufficient condition based on Theorem 5.1. This will assume controllability of the pair (A_0, B) .

THEOREM 5.4. *Let $m = 1$. Consider the following conditions*

$$(5.13) \quad \begin{aligned} (i) \quad & \text{rank } [b, A_0 b, \dots, A_0^{n-1} b] = n, \\ (ii) \quad & A_1 A_0^k b \in \sum_{j=0}^k \text{Im } A_0^j b, \quad k = 0, \dots, n-1, \\ (iii) \quad & \text{rank } [b, A_1 A_0 b, \dots, A_1 A_0^{n-1} b] = n. \end{aligned}$$

Then (i) and (ii) imply that (a) the matrix $P(\lambda)$ can, by a premultiplication by a constant nonsingular $n \times n$ matrix T , be transformed into a left-triangular matrix and (b) $\det P(\lambda) = \alpha = \text{const}$. Furthermore, under assumptions (i) and (ii), $\alpha \neq 0$ if and only if condition (iii) holds. Consequently, if (i) (ii) and (iii) hold simultaneously, the system is M_2 -approximately controllable for any value of $h > 0$.

6. Euclidean and algebraic controllability conditions. In this section we discuss briefly the relations between the function space controllability conditions described in § 3 and some other types of controllability conditions encountered in the literature.

First of all, we remark that the Euclidean controllability (see, e.g., [20], [36], [39], [40], [41], [70]) is obviously necessary for both M_2 - and L_2 -approximate controllability. Criteria for Euclidean controllability can be obtained entirely via the abstract approach presented in this paper; by analogy to (2.1)(ii) and (2.4)(ii), a necessary and sufficient condition of Euclidean controllability is that for $\eta \in \mathbb{R}^n$ the statement $\eta^T \Pi_0 S(t) \tilde{B} U = 0$, $t \geq 0$, implies $\eta = 0$; this is equivalent to $\eta^T X(t) B = 0$, $t \geq 0$ implies $\eta = 0$, where $X(t)$ = fundamental matrix of (1.1). The latter condition is precisely the usual starting point [20], [40] to derive the Euclidean controllability

criteria via an analysis in time domain. On the other hand, by using the former condition and the Laplace transform one can characterize the Euclidean controllability by conditions stated in the λ -domain. Some of the conditions given below have not appeared before in the literature.

PROPOSITION 6.1. *Each of the following conditions is equivalent to Euclidean controllability for the system (1.1) (on the interval $[0, t_1]$, $t_1 \geq nh$):*

- (i) $\begin{cases} \eta \in R^n, & \eta^T \Delta^{-1}(\lambda)B \equiv 0, \quad \forall \lambda \in \rho(\tilde{A}) \\ \Rightarrow \eta = 0; \end{cases}$
- (ii) $\begin{cases} \eta \in R^n, & \eta^T P(\lambda)v(e^{-h\lambda}) \equiv 0, \quad \forall \lambda \in \mathbb{C} \\ \Rightarrow \eta = 0; \end{cases}$
- (iii) $\begin{cases} \eta \in R^n, & \eta^T P(\lambda)v(\mu) \equiv 0, \quad \forall \lambda, \mu \in \mathbb{C} \\ \Rightarrow \eta = 0; \end{cases}$
- (iv) $\begin{cases} \eta \in R^n, & \eta^T \Delta^{-1}(\lambda, \mu)B \equiv 0, \quad \forall \lambda, \mu \in \mathbb{C}, \text{ such that } \det \Delta(\lambda, \mu) \neq 0, \\ \Rightarrow \eta = 0; \end{cases}$
- (v) $\begin{cases} \eta \in R^n, & \eta^T P(\lambda) \equiv 0, \quad \forall \lambda \in \mathbb{C} \\ \Rightarrow \eta = 0. \end{cases}$

Proof. Condition (i) is—in view of (2.10)—equivalent to

$$\begin{cases} \eta^T \Pi_0 R^k(\lambda_0, \tilde{A}) \tilde{B} U = 0, & k = 0, 1, \dots, \eta \in R^n, \\ \Rightarrow \eta = 0 & \lambda_0 = \text{a fixed point in } \rho(\tilde{A}) \end{cases}$$

and is the analogue of Propositions 2.5 and 2.6 for L_2 - and M_2 -approximate controllability, respectively. The equivalences (i) \Leftrightarrow (ii) as well as (iii) \Leftrightarrow (iv) follow from (3.5). That the negation of (ii) implies the negation of (iii) can be seen in an elementary way, since $\bar{\eta}^T P(\lambda)v(e^{-h\lambda})$, for $0 \neq \bar{\eta} \in R^n$ is a *polynomial* row vector in λ and $e^{-h\lambda}$ —see also Remark 4.1. The implication: negation of (iii) \Rightarrow negation of (v) follows by choosing n distinct values μ_1, \dots, μ_n and arguing as in going from (3.16) to (3.19). Negation of (v) obviously implies negation of (ii). Q.E.D.

Remark 6.1. In view of (v) in Proposition 6.1, one sees that the condition $\text{rank}_{\mathbb{C}} P(\lambda) = n$ is sufficient (but generally not necessary, unless $n = 1$) for Euclidean controllability. An example for $n \geq 2$ of a Euclidean controllable system for which the condition $\text{rank}_{\mathbb{C}} P(\lambda) = n$ is violated is given by the scalar difference-differential equation written in vector notation. In view of Remark 3.1 and Theorem 3.1, it follows that L_2 -approximate controllability implies Euclidean controllability, in agreement with the very definitions.

By using Proposition 6.1(iv) one can obtain (see [43]) all the known Euclidean controllability conditions [20]; also one can obtain the following result (proof is omitted; see [43]).

THEOREM 6.2. *The system (1.1) is Euclidean controllable if and only if for any $\eta \in R^n$ (independent of μ) the relation*

$$(6.1) \quad \eta^T [B, (A_0 + A_1\mu)B, \dots, (A_0 + A_1\mu)^{n-1}B] = 0, \quad \forall \mu \in \mathbb{C}$$

implies $\eta = 0$.

In the recent literature on algebraic system theory, the controllability of retarded systems has also been studied by purely algebraic methods [32], [49], [59], [68]. By treating μ as a delay operator, and defining $\hat{A}(\mu) = A_0 + A_1\mu$ as a matrix over $\mathbb{R}[\mu]$, the ring of polynomials in μ with coefficients in the field of real numbers, the system (1.1) can be represented by

$$\dot{y}(t) = \hat{A}(\mu)y(t) + Bu(t).$$

The system is called "weakly controllable" (in the algebraic sense [49]) if

$$(6.2) \quad \text{rank}_{\mathbb{C}} [B, \hat{A}(\mu)B, \dots, \hat{A}^{n-1}(\mu)B] = \eta.$$

Denote $[B, \hat{A}(\mu)B, \dots, \hat{A}^{n-1}(\mu)B]$ by $\hat{H}(\mu)$. The failure of (6.2) implies that there is a nontrivial $n \times 1$ polynomial vector $d(\mu)$ such that

$$d^T(\mu)\hat{H}(\mu) = 0, \quad \forall \mu \in \mathbb{C}$$

(in general $d(\mu) \neq d = \text{const.}$). Thus (6.2) is stronger than the Euclidean controllability. We point out below a relationship between condition (6.2) and the approach adopted in this paper.

Define the $n \times mn$ polynomial matrix $M(\mu)$ by the following identity:

$$(6.3) \quad \text{adj } \Delta(\lambda, \mu)B = M(\mu)v(\lambda)$$

where $v(\lambda) = (1, \lambda, \dots, \lambda^{n-1})^T \otimes I_m$; compared to (3.5), equation (6.3) is just an alternative way of factoring $\text{adj } \Delta(\lambda, \mu)B$ into a product of polynomial matrices. We then have:

PROPOSITION 6.3. $\text{rank}_{\mathbb{C}} \hat{H}(\mu) = n$ if and only if $\text{rank}_{\mathbb{C}} M(\mu) = n$.

Proof. (Only if). If $\text{rank}_{\mathbb{C}} M(\mu) < n$, there is a nonzero n -vector p_μ , depending on μ such that

$$(6.4) \quad p_\mu^T M(\mu) = 0, \quad \forall \mu \in \mathbb{C}.$$

Multiplying by $v(\lambda)$, one obtains

$$(6.5) \quad p_\mu^T \Delta^{-1}(\lambda, \mu)B = 0$$

for all λ, μ such that $\det \Delta(\lambda, \mu) \neq 0$. Fixing μ , differentiating (6.16) with respect to λ , and using the relation [28, Chap. I, § 2]

$$\frac{d^k}{d\lambda^k} \Delta^{-1}(\lambda, \mu) = (-1)^k k! \Delta^{-(k+1)}(\lambda, \mu), \quad k = 0, 1, \dots,$$

one obtains

$$(6.6) \quad p_\mu^T \Delta^{-k}(\lambda, \mu)B = 0, \quad k = 1, 2, \dots, \quad \lambda, \mu \text{ such that } \det \Delta(\lambda, \mu) \neq 0$$

and hence

$$(6.7) \quad p_\mu^T \Delta^k(\lambda, \mu)B = 0, \quad k = 1, 2, \dots,$$

which for $\lambda = 0$ gives $\text{rank}_{\mathbb{C}} \hat{H}(\mu) < n$.

(If). There is a nonzero p_μ annihilating $\hat{H}(\mu)$, so (6.7) holds for all λ . Fixing μ and restricting λ to the set $\{\lambda \mid \det \Delta(\lambda, \mu) \neq 0\}$, all the steps from (6.7) to (6.5) can be reversed while the step from (6.5) to (6.4) follows by taking n distinct values $\lambda_1, \dots, \lambda_n$ and using arguments analogous to those in the last part of the proof of Theorem 3.4. Q.E.D.

COROLLARY 6.4. *The system (1.1) is Euclidean controllable on $[0, t_1]$, $t_1 \geq nh$ if and only if for any $\eta \in R^n$ (independent of μ)*

$$\begin{aligned} \eta^T M(\mu) &= 0, \quad \forall \mu \in \mathbb{C} \\ \Rightarrow \eta &= 0. \end{aligned}$$

Example 7.3 given in the next section is an example of a system which is Euclidean controllable but not controllable in the sense of condition (6.2). In fact, in that example each of $\hat{H}(\mu)$, $M(\mu)$ is annihilated by $p_\mu^T = (1, -\mu, 0)$.

Remark 6.2. $\text{rank}_{\mathbb{C}} M(\mu) = n$ does not imply $\text{rank}_{\mathbb{C}} P(\lambda) = n$. For instance, any system with (A_0, b) controllable has $\text{rank}_{\mathbb{C}} M(\mu) = n$ regardless of A_1 , i.e. even for $A_1 = 0$, when obviously $\text{rank}_{\mathbb{C}} P(\lambda) < n$ (if $m < n$). On the other hand, controllability of the pair (A_1, B) guarantees both $\text{rank}_{\mathbb{C}} P(\lambda) = n$ (Remark 5.3) and $\text{rank}_{\mathbb{C}} M(\mu) = n$. We have noticed for many other examples that $\text{rank}_{\mathbb{C}} M(\mu) = n$ whenever $\text{rank}_{\mathbb{C}} P(\lambda) = n$. One can prove for $n = 2$ that $\text{rank}_{\mathbb{C}} P(\lambda) = n$ implies $\text{rank}_{\mathbb{C}} M(\mu) = n$, but it is an open question whether this implication is true in general.

7. Spectral controllability and stabilizability. The conditions presented in the previous sections involved values of λ belonging to the resolvent set $\rho(\tilde{A})$. Further insight into the problem can be obtained by studying the spectral points $\lambda \in \sigma(\tilde{A})$.

The spectral analysis of retarded systems within the framework of the space $C([-h, 0], \mathbb{R}^n)$ has been developed by Hale [21]. It can be verified that most of these results have their counterparts in the framework of the space M_2 . In particular, since the resolvent operator $R(\lambda, \tilde{A})$ is compact, the spectrum of \tilde{A} reduces to point spectrum [28, Thm. 6.29], and is given by

$$(7.1) \quad \sigma(\tilde{A}) = \{\lambda \mid \det \Delta(\lambda) = 0\}.$$

This set coincides with the spectrum of the infinitesimal generator of the semigroup in space C . Properties of that spectrum are well known [21]; in particular, they satisfy the assumptions needed for the spectral decomposition [65, § 4] of the equation (1.2).

Let $\lambda_j \in \sigma(\tilde{A})$ and let $\text{Ker}(\lambda_j I - \tilde{A})^{k_j}$ denote the generalized eigenspace (of some dimension d_j) of the operator \tilde{A} , with the basis $\Phi_{\lambda_j} \in D(\tilde{A})$. Similarly as in [21], [6], the projection Π_{λ_j} of $x(t)$ onto $\text{Ker}(\lambda_j I - \tilde{A})^{k_j}$ is given by

$$(7.2) \quad \Pi_{\lambda_j} x(t) = \frac{1}{2\pi i} \int_{\Gamma_j} R(\lambda, \tilde{A}) x(t) d\lambda = \Phi_{\lambda_j} \xi(t)$$

where Γ_j is a rectifiable, simple, closed curve containing only the eigenvalue λ_j and $\xi(t)$ is an d_j dimensional vector of coefficients. As a function of t , $\xi(t)$ satisfies an ordinary differential equation

$$(7.3) \quad \dot{\xi}(t) = \Lambda_j \xi(t) + K_j u(t)$$

where Λ_j is a $d_j \times d_j$ matrix whose only eigenvalue is λ_j , and K_j is an $d_j \times m$ matrix such that $\Phi_{\lambda_j} K_j = \Pi_{\lambda_j} \tilde{B}$.

We shall say that a *spectral mode associated with λ_j is controllable*, if the equation (7.3) is Euclidean controllable in \mathbb{R}^{d_j} , i.e. if the pair (Λ_j, K_j) is controllable.

The system (1.2) will be called *spectrally controllable* if all its spectral modes are controllable.

PROPOSITION 7.1. *A necessary condition for the system (1.2) to be M_2 -approximately controllable is that the system be spectrally controllable.*

*Proof.*⁶ The statement

$$\forall \psi \in M_2, \quad \forall \varepsilon > 0 \quad \exists t \quad \exists u \in L_1([0, t], \mathbb{R}^m) \quad \text{such that}$$

$$\|\psi - x(t; u)\|_{M_2} < \varepsilon$$

implies

$$\|\Pi_{\lambda_j} \psi - \Pi_{\lambda_j} x(t; u)\|_{M_2} < \varepsilon \gamma, \quad \gamma = \|\Pi_{\lambda_j}\|.$$

⁶ Additional results on relations between spectral controllability and approximate controllability are contained in [45], [46]; it is shown in [45] that a property called *F*-controllability (weaker than M_2 -approximate controllability) implies spectral controllability.

Moreover, due to the invariance of $\text{Ker} (\Lambda_j - \tilde{A})^{k_j}$ under \tilde{A} , hence under the motion of (1.2), the above statement implies that the system (7.3) is approximately controllable in R^{d_j} . But since (7.3) is finite dimensional, the approximate controllability of (7.3) is equivalent to exact controllability. Q.E.D.

The matrices Λ_j and K_j obtained by using spectral projection in the space M_2 are identical to those obtained via projections in space C [6]. Direct testing of controllability of (Λ_j, K_j) as suggested by Osipov [53] is, in most cases, not practically feasible. Recently, Pandolfi [54], and also Bhat and Koivo [8], have proved that the controllability of (Λ_j, K_j) is equivalent to

$$(7.4) \quad \text{rank} [\Delta(\lambda_j); B] = n.$$

This is a generalization of a well known Hautus condition [23]. Since $\lambda \in \rho(\tilde{A})$ implies $\text{rank} \Delta(\lambda) = n$, one has that the system (1.2) is spectrally controllable if and only if

$$(7.5) \quad \text{rank} [\Delta(\lambda), B] = n \quad \text{for all } \lambda \in \mathbb{C}.$$

The conditions reported above⁷ have an advantage over those of Osipov in that they do not require computing of projections Π_{λ_p} , but they also have the essential shortcoming of requiring the computation of eigenvalues of \tilde{A} . Since \tilde{A} usually has an infinite spectrum, verification of condition (7.4) for all $\lambda_j \in \sigma(\tilde{A})$ usually is not feasible.

Our next result overcomes this difficulty.

THEOREM 7.2. *If rank $B = 1$ then (7.5) implies*

$$(7.6) \quad P(\lambda)v(e^{-\lambda h}) \neq 0 \quad \text{for all } \lambda \in \mathbb{C}.$$

The converse (7.6) \Rightarrow (7.5) holds for any rank $B \geq 1$.

Proof. 1. Suppose that $P(\lambda)v(e^{-\lambda h}) = 0$ at some $\lambda = \lambda_0$. By (3.5) this means that

$$(7.7) \quad \text{adj } \Delta(\lambda_0)B = 0.$$

Hence $\text{rank} [\text{adj } \Delta(\lambda_0)]$ is less than n (otherwise there would be no column of B orthogonal to all rows of $\text{adj } \Delta(\lambda_0)$), and so $\text{rank } \Delta(\lambda_0) < n$; i.e. by (7.1) λ_0 is an eigenvalue of \tilde{A} . There are now two possibilities

$$(7.8) \quad (a) \quad \text{rank } \Delta(\lambda_0) = n - 1 \Leftrightarrow \text{rank} [\text{adj } \Delta(\lambda_0)] = 1,$$

$$(7.9) \quad (b) \quad \text{rank } \Delta(\lambda_0) < n - 1 \Leftrightarrow \text{adj } \Delta(\lambda_0) = 0$$

(these equivalences follow from the properties of the determinants). In case (a) the rows of $[\text{adj } \Delta(\lambda_0)]$ are all aligned, and at least one of them, denoted by η^T , is nonzero, $\eta^T \neq 0$. Then from $[\text{adj } \Delta(\lambda_0)] \Delta(\lambda_0) = I \det \Delta(\lambda_0)$ one has $\eta^T \Delta(\lambda_0) = 0$, and, by (7.7), $\eta^T B = 0$, which gives $\text{rank} [\Delta(\lambda_0), B] < n$. In case (b) if $\text{rank } B = 1$, one can still conclude that $\text{rank} [\Delta(\lambda_0), B] < n$. If $\text{rank } B > 1$, one can only say that $\text{rank} [\Delta(\lambda_0), b_i] < n$ for each column b_i of B , $i = 1, \dots, m$. This proves the first part of the theorem.

2. Suppose that condition (7.5) fails at λ_0 . Thus, there is a nonzero $\eta \in R^n$, such that

$$(7.10) \quad \eta^T \Delta(\lambda_0) = 0$$

and

$$(7.11) \quad \eta^T B = 0.$$

⁷ These conditions were recently generalized to systems with delays in control [52].

(7.10) implies that $\text{rank } \Delta(\lambda_0) \leq n-1$, so that again either (a) or (b) holds. In case (b) the result obtains trivially. In case (a) all the rows of $\text{adj } \Delta(\lambda_0)$ are collinear, and the dimension of the nullspace of $\Delta^T(\lambda_0)$ is 1. Since one has both (7.10) and $[\text{adj } \Delta(\lambda_0)] \Delta(\lambda_0) = 0$, all the rows of $[\text{adj } \Delta(\lambda_0)]$ are collinear with η^T , so by (7.11) they annihilate B

$$0 = \text{adj } \Delta(\lambda_0)B = P(\lambda_0)v(e^{-\lambda_0 h}). \quad \text{Q.E.D.}$$

COROLLARY 7.3. *Let $m = 1$. A necessary and sufficient condition for spectral controllability is*

$$(7.12) \quad P(\lambda)v(e^{-\lambda h}) \neq 0, \quad \forall \lambda \in \mathbb{C}.$$

For $m > 1$ the condition is sufficient only.

COROLLARY 7.4. *Let $m = 1$. A necessary condition for M_2 -approximate controllability is*

$$(7.13) \quad \begin{aligned} & \text{(i) } \det P(\lambda) \neq 0 \\ & \text{(ii) } P(\lambda)v(e^{-\lambda h}) \neq 0, \quad \forall \lambda \in \mathbb{C}. \end{aligned}$$

We note here that a condition analogous to (7.12) was first stated in [24] as a necessary condition for exact controllability in the space $W_2^{(1)}([-h, 0], \mathbb{R}^n)$ of linear *neutral* differential-difference equations; that condition was also proved to be sufficient for $n = 2$, or under some other additional hypotheses; however, the condition was derived independently of spectral notions. The sufficient part of Corollary 7.3 for neutral systems is contained in Remark 5.1 of [25]. Also, a result analogous to Theorem 7.2 (but in a less general form) can be found in Corollary 5.1 of [25].

Remark 7.1. Observe that the condition (7.12) have a distinct advantage over (7.5). If $m = 1$ and $\text{rank}_{\mathbb{C}} P(\lambda) = n$, $\det P(\lambda)$ is a nonzero polynomial. The roots of this polynomial are the only points at which the rank $P(\lambda)$ is less than n , that is at which the condition

$$P(\lambda)v(e^{-\lambda h}) \neq 0$$

(hence, equivalently, the condition $\text{rank } [\Delta(\lambda), B] = n$) can possibly fail. By computing those roots and checking at them the condition (7.12) *one can check the spectral controllability without a prior knowledge of eigenvalues of \tilde{A} !*

In other words, in case $\text{rank}_{\mathbb{C}} P(\lambda) = n$ and $m = 1$ the roots of $\det P(\lambda) = 0$ are the only potential candidates for the eigenvalues associated with uncontrollable eigenmodes (all such eigenvalues must satisfy $\det P(\lambda) = 0$ and $P(\lambda)v(e^{-\lambda h}) = 0$).

Example 7.1. Consider again the system of Example 4.3.

$$P(\lambda) = \begin{bmatrix} \lambda - 1 & -1 \\ 3(\lambda - 1) & 3 \end{bmatrix}.$$

The polynomial $\det P(\lambda) = 6\lambda$ has only one root $\lambda_1 = 0$. Checking the condition (7.12) we find

$$P(\lambda)v(e^{-\lambda h}) = \begin{bmatrix} \lambda + 1 & -1 \\ 3(\lambda - 1) & 3 \end{bmatrix} \begin{bmatrix} 1 \\ e^{-\lambda h} \end{bmatrix}$$

which for $\lambda = 0$ equals

$$\begin{bmatrix} 1 & -1 \\ -3 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Now, the reason for which the system of Example 4.3 was not M_2 -approximately controllable has become clear: $\lambda = 0$ is an eigenvalue whose associated eigenmode is not controllable. This can be confirmed by inspecting

$$\Delta(\lambda) = \begin{bmatrix} \lambda - 1 + e^{-\lambda h} & -1 \\ 0 & \lambda - 2 - e^{-\lambda h} \end{bmatrix}$$

which, for $\lambda = 0$, gives $\det \Delta(0) = 0$ and

$$\text{rank} [\Delta(0), B] = \text{rank} \begin{bmatrix} 0 & -1 & 1 \\ 0 & -3 & 3 \end{bmatrix} = 1 < 2,$$

regardless of value of h . One can also verify that a slight perturbation of B (replace 3 by $3 + \varepsilon$, $\varepsilon > 0$) yields (7.12), thus the spectral controllability.

Example 7.2. The reader can verify that for the system of Example 3.1 and 4.1 one has

$$(7.14) \quad \begin{aligned} P(\lambda)v(e^{-\lambda}) &= 0 \quad \text{for } \lambda = 0 \quad (h = 1) \\ \frac{d}{d\lambda}[P(\lambda)v(e^{-\lambda})]_{\lambda=0} &= 0 \end{aligned}$$

and

$$\frac{d^2}{d\lambda^2}[P(\lambda)v(e^{-\lambda})]_{\lambda=0} = 0.$$

Since in this example $\det \Delta(\lambda) = \lambda^3$, one sees that

$$\Delta^{-1}(\lambda)B = \frac{P(\lambda)v(e^{-\lambda})}{\lambda^3}$$

has, in fact, no singularity at $\lambda = 0$, so that *all* the spectral modes of the system are *uncontrollable*. This extremely pathological situation is related to the phenomenon of pointwise degeneracy, as discussed in the next section.

A problem closely related to spectral controllability is that of stabilizability. The stabilizability is understood here in the same sense as in [65], that is that there exists a bounded linear operator $\tilde{K}: M_2 \rightarrow R^m$ such that the semigroup $S_{\tilde{F}}(t)$, $t \geq 0$ in M_2 , generated by $\tilde{F} = \tilde{A} + \tilde{B}\tilde{K}: D(\tilde{A}) = D(\tilde{F}) \rightarrow M_2$ satisfies

$$\|S_{\tilde{F}}(t)x_0\|_{M_2} \rightarrow 0 \quad \text{as } t \rightarrow \infty, \quad \forall x_0 \in M_2.$$

By [65, Cor. 6.2, plus § 7, item 3] the linear retarded systems are stabilizable if all the spectral modes corresponding to eigenvalues satisfying $\text{Re } \lambda \geq 0$ are controllable. By using the space decomposition as in [65, § 4] it turns out that this condition also is necessary. See the Appendix. Furthermore, for retarded systems, stabilizability is equivalent to exponential stabilizability

$$(7.15) \quad \|S_{\tilde{F}}(t)x_0\|_{M_2} \leq C e^{-\delta t} \|x_0\|, \quad \delta > 0, \quad t \geq 0.$$

From these remarks and from Theorem 7.1 we obtain

COROLLARY 7.5. *Let $m = 1$. A necessary condition for stabilizability of (1.2) is*

$$P(\lambda)v(e^{-\lambda h}) \neq 0, \quad \forall \lambda \text{ such that } \text{Re } \lambda \geq 0.$$

This condition is sufficient for any $m \geq 1$.

Example 7.3. Consider the system [49]

$$\begin{aligned}\dot{x}_1(t) &= x_1(t) + x_3(t-h), \\ \dot{x}_2(t) &= x_2(t) + x_3(t), \\ \dot{x}_3(t) &= u(t).\end{aligned}$$

An elementary computation gives

$$P(\lambda)v(e^{-\lambda h}) = \begin{bmatrix} (\lambda-1)e^{-\lambda h} \\ \lambda-1 \\ (\lambda-1)^2 \end{bmatrix},$$

which vanishes for $\lambda = 1$. The system is not stabilizable (which confirms the statement made in [49]), and is not L_2 -approximately controllable ($\text{rank}_{\mathbb{C}} P(\lambda) = 2 < 3$).

In the literature, the stabilizing feedback has often been described as an operator involving integrals of $y(t+\theta)$ over $\theta \in [-h, 0]$, which is believed to be inconvenient in implementation. We indicate below how a stabilizing feedback for the system (1.1) can be realized by using a lumped parameter system with $y(t)$ and $y(t-h)$ as its inputs.

Suppose that $\text{Re } \lambda_i \geq 0$, $i = 1, \dots, N$ and $\text{Re } \lambda_j < 0$ for $j > N$, and furthermore, that all the systems (7.3) corresponding to $j = 1, \dots, N$ are controllable. Then the system can be stabilized [53], [54], [65] by a feedback of the form

$$(7.16) \quad u(t) = \sum_{j=1}^N F_j \xi_j(t)$$

where F_j are $m \times d_j$ matrices, and, as is well known [21], [6], the $\xi_j(t)$ satisfy

$$(7.17) \quad \xi_j(t) = \Psi_{\lambda_j}(0)y(t) + \int_{-h}^0 \Psi_{\lambda_j}(h+\theta)A_1y(t+\theta) d\theta$$

where $\Psi_{\lambda_j}(0)$ is a $k_j \times n$ matrix function satisfying

$$\Psi_{\lambda_j}(\theta) = e^{-\Lambda_j \theta} \Psi_{\lambda_j}(0), \quad \psi_{\lambda_j}(0) \Delta(\lambda_j) = 0.$$

Define

$$v_j(t) = \int_{-h}^0 \Psi_{\lambda_j}(h+\theta)A_1y(t+\theta) d\theta = \int_{t-h}^t \Psi_{\lambda_j}(h+s-t)A_1y(s) ds.$$

Differentiating $v_j(t)$, one obtains

$$(7.18) \quad \dot{v}_j(t) = e^{-\Lambda_j h} \Psi_{\lambda_j}(0)A_1y(t) - \Psi_{\lambda_j}(0)A_1y(t-h) + \Lambda_j v_j(t);$$

the stabilizing feedback becomes

$$(7.19) \quad u(t) = F_0 y(t) + \sum_{j=1}^N F_j v_j(t); \quad F_0 = \sum_{j=1}^N F_j \psi_{\lambda_j}(0)$$

where v_j satisfy the differential equations (7.18). Theoretically, therefore, the realization of the stabilizing feedback could be done just by using the dynamical feedback given by (7.18), (7.19). Practically, a system with such a feedback might be sensitive to errors in parameters. Design of "insensitive" feedbacks is an interesting area for future study.

We note that using Theorem 4.4 one can, in some cases, prove that the spectral controllability along with the condition $\text{rank}_{\mathbb{C}} P(\lambda) = n$ implies M_2 -approximate

controllability. A discussion of such results is given in [46]. Here we report only the following result.

PROPOSITION 7.6. *Let $n = 2$, $m = 1$. A necessary and sufficient condition for M_2 -approximate controllability is*

$$\det P(\lambda) \neq 0 \quad \text{and} \quad P(\lambda) \begin{pmatrix} 1 \\ e^{-\lambda h} \end{pmatrix} \neq 0, \quad \forall \lambda \in \mathbb{C},$$

i.e., spectral controllability and $\text{rank}_{\mathbb{C}} P(\lambda) = n (= 2)$ are equivalent to M_2 -approximate controllability.

It is also interesting to note that for $n = 2$, $m = 1$ the condition (7.13)(ii) is equivalent to null function space controllability [26]. Therefore, for two dimensional systems with $\text{rank}_{\mathbb{C}} P(\lambda) = n$, $m = 1$ one has that M_2 -approximate controllability, spectral controllability and null function space controllability are all equivalent.

8. Relation between pointwise degeneracy and lack of L_2 -approximate controllability. In this section we show that there is an interesting relation between the pointwise degenerate systems and the systems which are not L_2 -approximately controllable.

Suppose that the system

$$(8.1) \quad \dot{z}(t) = A_0 z(t) + A_1 z(t-h), \quad z \in \mathbb{R}^n,$$

is pointwise degenerate with respect to $b \in \mathbb{R}^n$ at $t \geq lh$, l an integer ≥ 2 (see [58]), that is $b^T z(t) = 0$ $t \geq lh$ for any solution $z(\cdot)$.

The following observation of Kappel [27], which we now recast in the framework of C_0 semigroups, plays an important role. Let $\eta \in M^2$, $\eta = (\eta^0, 0)$, $\eta^0 \in \mathbb{R}^n$ arbitrary and consider for $\text{Re } \lambda > \omega_0$ [ω_0 defined above (2.2)]

$$\begin{aligned} b^T \Delta^{-1}(\lambda) \eta^0 &= b^T \Pi_0 R(\lambda, \tilde{A}) \eta \quad (\text{by (2.5)}) \\ &= \int_0^\infty e^{-\lambda t} b^T \Pi_0 S(t) \eta \, dt = \int_0^{lh} e^{-\lambda t} b^T X(t) \eta^0 \, dt, \quad \forall \eta^0 \in \mathbb{R}^n, \end{aligned}$$

where $X(t)$ = fundamental matrix; this shows that $b^T \Delta^{-1}(\lambda)$ is an $\text{FLT}_2([0, lh], \mathbb{R}^n)$ function. Suppose further that among the components of $b^T \Delta^{-1}(\lambda)$ there are at least two which belong to the class $\text{FLT}_2([0, h], \mathbb{R})$, so that $b^T \Delta^{-1}(\lambda)$ can be written as

$$b^T \Delta^{-1}(\lambda) = [* , * , \dots , * , r_j(\lambda) , * , \dots , * , r_k(\lambda) , *]$$

where $*$ denote irrelevant elements, and both $r_j(\lambda)$ and $r_k(\lambda)$ denote the components of class $\text{FLT}_2([0, h], \mathbb{R})$, appearing at j th and k th place respectively.

Consider now the system

$$(8.2) \quad \dot{y}(t) = A_0^T y(t) + A_1^T y(t-h) + bu(t).$$

Its transfer function is $\Delta^{T-1}(\lambda)b$, and one easily sees that it is annihilated by $q(\cdot) \in \text{FLT}_2([0, h], \mathbb{R}^n)$ given by

$$(8.3) \quad q^T(\lambda) = [0, 0, \dots, 0, -r_k(\lambda), 0, \dots, 0, r_j(\lambda), \dots]$$

where the nonzero elements $-r_k(\lambda)$, $r_j(\lambda)$ appear at j th and k th place respectively. Thus (8.2) is not L_2 -approximately controllable.

From now on, a system with matrices A_0 , A_1 , B such that the system with matrices A_0^T , A_1^T is pointwise degenerate with respect to B (all columns of B), will be called a *transpose of a pointwise degenerate system*.

The system described in Examples 3.1 and 4.1 is for $h = 1$ (and for this value only), a transpose of the pointwise degenerate system given by Popov [58]. A nontrivial fact in that example is that, for $h = 1$ the transfer function $\Delta^{-1}(\lambda)b$ does indeed have two components which are finite Laplace transforms over the interval $[0, 1]$ (the third one is an FLT_2 over $[0, 2]$). Furthermore, as said before, the system is L_2 -approximately controllable for all $h \neq 1$. (However, by application of Corollary 7.4 one obtains that the system is not M_2 -approximately controllable for any $h > 0$).

Example 8.1. Consider the system

$$A_0 = \begin{bmatrix} 0 & -2 & 0 \\ 0 & -3 & 2 \\ 0 & -1 & 0 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}.$$

Computing $\Delta^{-1}(\lambda)b$, we obtain

$$\Delta^{-1}(\lambda)b = \frac{1}{\lambda^3 + 3\lambda^2 + 2\lambda} \begin{bmatrix} \lambda^2 - \lambda + 2 - 2e^{-\lambda h} \\ 2\lambda^2 + \lambda e^{-\lambda h} \\ -2\lambda - e^{-\lambda h} + 2\lambda e^{-\lambda h} + e^{-2\lambda h} \end{bmatrix}.$$

Hence

$$(8.4) \quad P(\lambda) = \begin{bmatrix} \lambda^2 - \lambda + 2 & -2 & 0 \\ 2\lambda^2 & \lambda & 0 \\ -2\lambda & -1 + 2\lambda & 1 \end{bmatrix}.$$

Now, the necessary condition of Theorem 3.1 is fulfilled, because $\det P(\lambda) = \lambda^3 + 3\lambda^2 + 2\lambda \neq 0$ (we observe that in this example $\det P(\lambda) \equiv \det \Delta(\lambda)$). The roots of $\det P(\lambda)$ are 0, -1, -2. The first and the second elements of $\Delta^{-1}(\lambda)b$ are

$$r_1(\lambda) = \frac{\lambda^2 - \lambda + 2 - 2e^{-\lambda h}}{\lambda(\lambda + 1)(\lambda + 2)} \quad \text{and} \quad r_2(\lambda) = \frac{2\lambda^2 + \lambda e^{-\lambda h}}{\lambda(\lambda + 1)(\lambda + 2)}.$$

We easily find that all the zeros of the denominator are canceled by the zeros of both numerators, if and only if $h = \ln 2$, in which case the system is the transpose of Zverkin's pointwise degenerate example [71]. Thus for $h = \ln 2$ both $r_1(\lambda)$ and $r_2(\lambda)$ are entire functions (as they should, by the degeneracy of Zverkin's example) of exponential type. Computing the constants H and H' appearing in Theorem 2.8, we find $H' = 0$, $H = h$. Since both r_1 and r_2 are $O(|\lambda|^{-1})$ when $\lambda = i\omega$, $\omega \rightarrow \infty$, they indeed satisfy all the conditions of Theorem 2.8 and, therefore, belong to the class $\text{FLT}_2([0, h], R)$. Consequently, $q^T(\lambda) = [-r_2(\lambda), r_1(\lambda), 0] \in \text{FLT}_2([0, h], R^3)$ annihilates $\Delta^{-1}(\lambda)b$, which proves that the system is not L_2 -approximately controllable *only* for the isolated value of $h = \ln 2$. However, since $P(\lambda)v(e^{-h\lambda}) = 0$ for $\lambda = 0$, the system is not spectrally controllable, let alone M_2 -approximately controllable, for any value of $h > 0$.

We now show that the situation arising in the transposed Popov example and in the transposed Zverkin example is not accidental.

THEOREM 8.1. *Every system of the form*

$$(8.5) \quad \dot{y}(t) = A_0 y(t) + A_1 y(t - h) + Bu(t)$$

such that the system

$$(8.6) \quad \dot{z}(t) = A_0^T z(t) + A_1^T z(t - h)$$

is pointwise degenerate with respect to B on $[2h, \infty)$ is not L_2 -approximately controllable.

Before proving the Theorem, we make the following observation.

Remark 8.1. Let $X(t)$ denote the $n \times n$ fundamental matrix of system (8.5). By using the known property of commutativity of the semigroup operator $S(t)$ with its infinitesimal generator \tilde{A}

$$(8.7) \quad \tilde{A}S(t)x = S(t)\tilde{A}x, \quad \forall t \geq 0, \quad \forall x \in D(\tilde{A})$$

one can prove the following relation:

$$(8.8) \quad A_0X(t) + A_1X(t-h) = X(t)A_0 + X(t-h)A_1, \quad \forall t \geq 0.$$

Proof of (8.8) using (8.7), for systems more general than (1.1), can be found in [7]; the same relation can also be proved without using (8.7) [41].

Proof of Theorem 8.1. Let (8.6) be pointwise degenerate with respect to B (all columns of B), $t \geq 2h$. This means that $B^T z(t) \equiv 0$ for $t \geq 2h$ for every solution $z(t)$ of (8.6). By virtue of (8.8), $X^T(t)$ is a fundamental matrix for (8.6), i.e.

$$(8.9) \quad \frac{d}{dt}X^T(t) = A_0^T X^T(t) + A_1^T X^T(t-h), \quad X^T(0) = I.$$

Degeneracy implies, as seen before [27]:

$$(8.10) \quad B^T X^T(t) \equiv 0, \quad t \geq 2h.$$

Differentiating and substituting (8.9), we obtain

$$(8.11) \quad \begin{aligned} 0 &\equiv \frac{d}{dt}B^T X^T(t) = B^T A_0^T X^T(t) + B^T A_1^T X^T(t-h), & t \geq 2h, \\ &= B^T X^T(t)A_0^T + B^T X^T(t-h)A_1^T, & t \geq 2h, \end{aligned}$$

where the last step obtains by (8.8). Now (8.10) and (8.11) imply

$$B^T X^T(t-h)A_1^T \equiv 0, \quad t \geq 2h,$$

or

$$X(t)B \in \text{Ker } A_1, \quad \text{for } t \geq h.$$

At this point we recall the result of Popov [58, Cor. 1] saying that if the system (8.6) is degenerate, then $\text{rank } A_1 \geq 2$. Thus the dimension of $\text{Ker } A_1$ is not greater than $n-2$. By a suitable change of basis in R^n (which does not affect the L_2 -approximate controllability) one can always arrange that all vectors in $\text{Ker } A_1$ have form $(\alpha_1, \alpha_2, \dots, \alpha_{n-2}, 0, 0)^T$. In such a modified system of coordinates, related to the original one by some transformation $z_M = Mz$, $z = M^{-1}z_M$, we will have

$$M^{-1}X(t)MM^{-1}B \stackrel{\text{def}}{=} X_M(t)B_M = \begin{bmatrix} \alpha_{1,1}(t) & \cdots & \alpha_{n,1}(t) \\ \vdots & & \vdots \\ \alpha_{1,n-2}(t) & & \alpha_{n,n-2}(t) \\ 0 & & 0 \\ 0 & & 0 \end{bmatrix}, \quad t \geq h,$$

which means that (at least) two coordinates of $X_M(t)B_M$ go to zero $[0, h]$. Since for system (8.5) $\Delta^{-1}(\lambda)B$ is just the Laplace transform of $X(t)B$, in the modified system of coordinates we will have that $\Delta_M^{-1}(\lambda)B_M$ contains at least two components of the class $\text{FLT}_2([0, h], R)$. Q.E.D.

As the matrix $P(\lambda)$ is not involved in the proof above, the statement of Theorem 8.1 is independent of $\text{rank}_{\mathbb{C}} P(\lambda)$. Examples 3.1, 4.1 and 8.1 show that the intersection of the class of systems for which $\text{rank}_{\mathbb{C}} P(\lambda) = n$ with the class of transposes of pointwise degenerate systems with respect to B on $[2h, \infty)$ is not empty.

We also have the following result.

LEMMA 8.2. *Let $m = 1$ and $\text{rank}_{\mathbb{C}} P(\lambda) = n$. Suppose that $\Delta^{-1}(\lambda)b$ is an entire function. Then the spectrum of \tilde{A} is finite, and $\det \Delta(\lambda)$ is a polynomial that divides $\det P(\lambda)$.*

Proof. If

$$\Delta^{-1}(\lambda)b = \frac{1}{\det \Delta(\lambda)} [\text{adj } \Delta(\lambda)]b$$

is an entire function, the zeros of $[\text{adj } \Delta(\lambda)]b$ must coincide with those of $\det \Delta(\lambda)$. But $[\text{adj } \Delta(\lambda)]b = P(\lambda)v(e^{-\lambda h})$ can have zeros only at those points $\bar{\lambda}$ at which $\text{rank } P(\bar{\lambda}) < n$. The assumption implies that there are only a finite number of such points, each of them satisfying $\det P(\bar{\lambda}) = 0$. In other terms, all zeros of $\det \Delta(\lambda)$ are also zeros of $\det P(\lambda)$. Q.E.D.

Examples 3.1, 4.1 and 8.1 are of the type described in the lemma.

We next show that the class of systems having $\text{rank}_{\mathbb{C}} P(\lambda) = n$ and yet not L_2 -approximately controllable is not limited to transposes of pointwise degenerate systems on $[2h, \infty)$. Starting from the latter class, we shall construct a class of augmented systems which has the same features as before, i.e. $q^T(\lambda)\Delta^{-1}(\lambda)B \equiv 0$ with nonzero $q^T(\lambda)$, $\text{rank}_{\mathbb{C}} P(\lambda) = n$, except that now $\Delta^{-1}(\lambda)B$ will not be entire.

To be specific, consider a system of dimension n given by $A_0, A_1, b, h, (m = 1)$, such that

(i) it is a transpose of a pointwise degenerate system on $[2h, \infty)$, so that the function $\Delta^{-1}(\lambda)b = [r_1(\lambda), \dots, r_n(\lambda)]^T$ has its n coordinates $r_i(\lambda)$ of class $\text{FLT}_2([0, 2h], R)$;

(ii) at least two of the functions $r_1(\lambda), \dots, r_n(\lambda)$, say $r_i(\lambda)$ and $r_j(\lambda)$, $j > i$, are of class $\text{FLT}_2([0, h], R)$;

(iii) The matrix $P(\lambda)$ satisfies $\text{rank}_{\mathbb{C}} P(\lambda) = n$.

Note that systems given in Examples 3.1 and 8.1 satisfy the above requirements.

Next, construct an augmented system of dimension $(n + 1)$ defined by

$$(8.12) \quad \hat{A}_0 = \left[\begin{array}{c|c} A_0 & 0 \\ \hline 0 & 0 \end{array} \right], \quad \hat{A}_1 = \left[\begin{array}{c|c} A_1 & b \\ \hline 0 & 1 \end{array} \right], \quad \hat{b} = \left[\begin{array}{c} 0 \\ 1 \end{array} \right].$$

Define $\hat{\Delta}(\lambda) \stackrel{\text{def}}{=} (I\lambda - \hat{A}_0 - \hat{A}_1 e^{-\lambda h})$. Inverting $\hat{\Delta}(\lambda)$ by partitioning (e.g. [3, p. 56]), we obtain

$$\det \hat{\Delta}(\lambda) = \det \Delta(\lambda)(\lambda - e^{-\lambda h})$$

and

$$\hat{\Delta}^{-1}(\lambda)\hat{b} = \left[\begin{array}{c} r_1(\lambda) \frac{e^{-\lambda h}}{\lambda - e^{-\lambda h}} \\ \vdots \\ r_n(\lambda) \frac{e^{-\lambda h}}{\lambda - e^{-\lambda h}} \\ \hline 1 \\ \hline \lambda - e^{-\lambda h} \end{array} \right] = \left[\begin{array}{c} \frac{P(\lambda)v(e^{-\lambda h})}{\det \Delta(\lambda)} \frac{e^{-\lambda h}}{\lambda - e^{-\lambda h}} \\ \hline 1 \\ \hline \lambda - e^{-\lambda h} \end{array} \right].$$

Because of the factor $1/(1 - e^{-\lambda h})$ the transfer function $\hat{\Delta}^{-1}(\lambda)\hat{b}$ is not entire, so that the augmented system is not a transpose of a pointwise degenerate system. From the formula above it follows that the matrix $\hat{P}(\lambda)$ of the augmented system is given by

$$\hat{P}(\lambda) = \begin{bmatrix} \text{---} & P(\lambda) \\ \det \Delta(\lambda) & 0 \end{bmatrix}$$

so that $\det \hat{P}(\lambda) = (-1)^n \det \Delta(\lambda) \det P(\lambda) \neq 0$, hence $\text{rank}_c \hat{P}(\lambda) = n + 1$.

Now, the following nonzero function of the class $\text{FLT}_2([0, h], R^{n+1})$ annihilates $\hat{\Delta}^{-1}(\lambda)\hat{b}$:

$$q^T(\lambda) = [0, \dots, \underset{\substack{\uparrow \\ \text{ith place}}}{-r_j(\lambda)}, \dots, \underset{\substack{\uparrow \\ \text{jth place}}}{r_i(\lambda)}, \dots, 0].$$

Indeed

$$q^T(\lambda)\hat{\Delta}^{-1}(\lambda)\hat{b} = [-r_j(\lambda)r_i(\lambda) + r_i(\lambda)r_j(\lambda)] \frac{e^{-\lambda h}}{\lambda - e^{-\lambda h}} \equiv 0.$$

Consequently, the augmented system is not L_2 -approximately controllable; we summarize this result below.

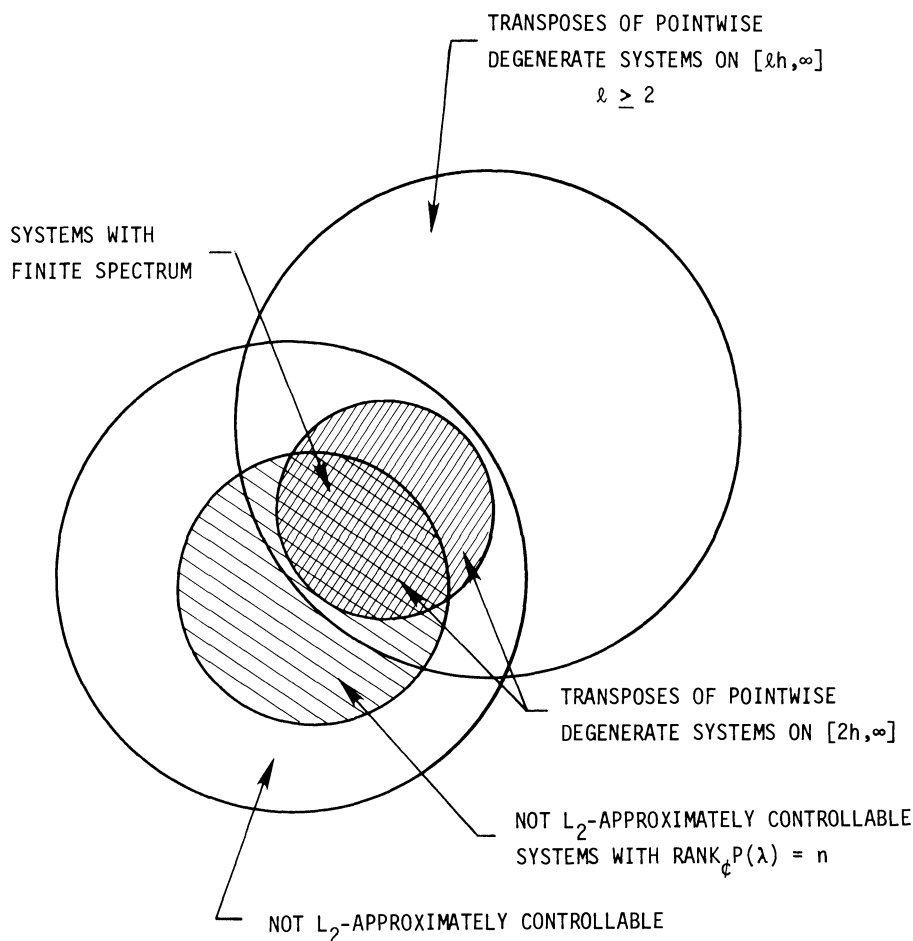


FIG. 1

THEOREM 8.3. *If the n -dimensional system (1.1) with $m = 1$ satisfies the conditions (i)–(iii) above, the augmented $(n + 1)$ -dimensional system defined by (8.12) satisfies $\text{rank}_C P(\lambda) = n + 1$, is not a transpose of a pointwise degenerate system, yet is not L_2 -approximately controllable.*

The facts obtained can be illustrated by Fig. 1.

Appendix. We now complement Corollary 6.2 of [65], with the following converse. In what follows, it will be expedient to refine the notation of [65], by adding the explicit dependence on the number $\delta > 0$ for which the space decomposition of § 4.1 in [65] holds. We shall therefore write $A_u(\delta)$, $A_s(\delta)$, $X_u(\delta)$, $X_s(\delta)$, $P(\delta)$, etc. for the quantities denoted in [65] simply by A_u , A_s , X_u , X_s , P , etc. corresponding to such δ .

THEOREM (Converse of Corollary 6.2 of [65]). *Let assumptions (i) and (iii) of Theorem 6.1 hold, as in Corollary 6.2 of [65], i.e., let the spectrum $\sigma(A)$ of A satisfy the spectrum decomposition assumption of § 4.1 in [65] for some $\delta > 0$, and let the corresponding operator $A_s(\delta)$ satisfy the spectrum determined growth assumption of § 2 in [65] on the subspace $X_s(\delta)$. Assume further that $X_u(\delta)$ is finite-dimensional. Then, if the pair $\langle A, B \rangle$ is stabilizable (as defined in § 1 in [65]), it follows that there exists a number ρ , $0 < \rho \leq \delta$, such that the corresponding pair $\langle A_u(\rho), B_u(\rho) \rangle$ with $B_u(\rho) = P(\rho)B$ is controllable on $X_u(\rho)$.*

Proof. The finite-dimensionality of $X_u(\delta)$ implies that the space decomposition holds also for all $0 < \rho \leq \delta$. The only nontrivial case to prove is when the set

$$\sigma(A) \cap \{\lambda : \text{Re } \lambda \geq 0\}$$

is nonempty. In this case assume, by contradiction, that for all ρ , $0 < \rho \leq \delta$, the corresponding pairs $\langle A_u(\rho), B_u(\rho) \rangle$ are not controllable. Since $X_u(\rho)$ is finite dimensional, it is known (e.g., [35, p. 93]) that $X_u(\rho)$ can be further split into two subspaces, $X_{u,1}(\rho)$ and $X_{u,2}(\rho)$, invariant for $A_u(\rho)$ and that the system $\langle A_u(\rho), B_u(\rho) \rangle$ can be written as

$$\begin{vmatrix} \dot{x}_{u,1} \\ \dot{x}_{u,2} \end{vmatrix} = \begin{vmatrix} A_{u,1}(\rho) & A_{u,2}(\rho) \\ 0 & A_{u,3}(\rho) \end{vmatrix} \begin{vmatrix} x_{u,1} \\ x_{u,2} \end{vmatrix} + \begin{vmatrix} B_u(\rho) \\ 0 \end{vmatrix} u$$

with $x_u = [x_{u,1}, x_{u,2}]$ (In fact $X_{u,1}(\rho)$ is given by

$$X_{u,1}(\rho) = \text{span} \{B_u(\rho)U, A_u(\rho)B_u(\rho)U, \dots, A_u^r(\rho)B_u(\rho)U\}$$

for some nonnegative integer $r = r(\rho)$, not greater than the dimension of $X_u(\rho)$). Since $\sigma(A_{u,3}(\rho) \subset \sigma(A_u(\rho))$, then for sufficiently small ρ there is a $\lambda \in \sigma(A_{u,3}(\rho))$ with $\text{Re } \lambda \geq 0$, which is an eigenvalue of $A_{u,3}(\rho)$ by the finite-dimensionality of $X_u(\rho)$. Hence

$$\|x_{u,2}(t)\| = \|e^{A_{u,3}(\rho)t}x_{u,2}(0)\|$$

does not converge to zero as $t \rightarrow \infty$. Since the control $u(t)$ does not influence $x_{u,2}(t)$, the pair $\langle A, B \rangle$ is not stabilizable, which is a contradiction. Q.E.D.

Remark. Applying Corollary 6.2 of [65], we see in fact that the stabilizability postulated in Theorem 2 above for the pair $\langle A, B \rangle$ is in fact exponential stabilizability. As for classes of physically significant dynamical systems to which Theorem 2 above is applicable we refer to [65, § 7].

REFERENCES

- [1] C. H. ANDERSON, *The linear differential-difference equation with constant coefficients*, J. Math. Anal. Appl. 40 (1972), pp. 122–130.
- [2] M. ABRAMOWITZ AND I. A. STEGUN, *Handbook of Mathematical Functions*, National Bureau of Standards, Washington, DC, 1964.

- [3] F. AYRES, JR., *Theory and Problems of Matrices*, Schaum's Outline Series, McGraw-Hill, New York, 1962.
- [4] H. T. BANKS AND J. A. BURNS, *Hereditary control problems: Numerical methods based on averaging approximation*, Res. rep. Lefschetz Center for Dynamical Systems, Brown University, Providence, RI, 1976; this Journal, to appear.
- [5] H. T. BANKS, M. Q. JACOBS AND C. E. LANGENHOP, *Characterization of the controlled states in $W_2^{(1)}$ of linear hereditary systems*, this Journal, 13 (1975), pp. 611–649.
- [6] 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.
- [7] C. BERNIER AND A. MANITIUS, *On semigroups in $R^n \times L^p$ corresponding to differential equations with delays*, Rep. CRM-665, Centre de Recherches Mathématiques, Université de Montréal, 1976.
- [8] K. P. M. BHAT AND H. N. KOIVO, *Modal characterizations of controllability and observability for time-delay systems*, IEEE Trans. Autom. Control, AC-21(1976), pp. 292–293.
- [9] J. G. BORISOVIC AND A. S. TURBABIN, *On the Cauchy problem for linear non-homogeneous differential equations with retarded argument*, Soviet Math. Dokl., 10 (1969), pp. 401–405.
- [10] P. CHARRIER, *Equations différentielles avec retard: applications de la dégénérescence à la théorie de la commande*, Thèse de 3^e cycle (Ph.D. thesis), Université de Bordeaux I, 1972.
- [11] A. K. CHOUDHURY, *A contribution to the controllability of time-lag systems*, Internat. J. Control, 17 (1973), no. 2.
- [12] M. C. DELFOUR AND S. K. MITTER, *Controllability, observability and optimal feedback control of affine hereditary differential systems*, this Journal, 10 (1972), pp. 298–328.
- [13] M. C. DELFOUR, *State theory of linear hereditary differential systems*, J. Math. Anal. Appl., 60 (1977), pp. 8–35.
- [14] M. C. DELFOUR, C. MCCALLA AND S. K. MITTER, *Stability and the infinite-time quadratic cost problem for linear hereditary differential systems*, this Journal, 13 (1975), pp. 48–88.
- [15] G. DOETSCH, *Handbuch der Laplace Transformation*, vol. III, Birkhauser Verlag, Basel, Switzerland, 1956.
- [16] H. O. FATTORINI, *On complete controllability of linear systems*, J. Differential Equations, 3 (1967), pp. 391–402.
- [17] ———, *Some remarks on complete controllability*, this Journal, 4 (1966), pp. 686–694.
- [18] F. R. GANTMACHER, *Theory of Matrices*, vol. 1, Chelsea, New York, 1957.
- [19] W. GREUB, *Linear Algebra*, Springer-Verlag, Berlin/New York, 1975.
- [20] R. GABASOV AND M. KIRILLOVA, *Qualitative Theory of Optimal Processes*, Nauka, Moscow, 1971. (In Russian.)
- [21] J. K. HALE, *Functional Differential Equations*, Springer-Verlag, New York, 1972.
- [22] G. H. HARDY, *A theorem concerning Fourier transforms*, J. London Math. Soc., 8 (1933), pp. 227–231.
- [23] M. L. J. HAUTUS, *Controllability and observability conditions of linear autonomous systems*, Indag. Math. 31 (1969), pp. 443–448.
- [24] M. Q. JACOBS AND C. E. LANGENHOP, *Criteria for function space controllability of linear neutral systems*, Res. rep. Department of Mathematics, University of Missouri–Columbia, 1975.
- [25] ———, *Criteria for function space controllability of linear neutral systems*, this Journal, 14 (1976), pp. 1009–1048.
- [26] ———, *Controllable two dimensional neutral systems*, Banach Center Publications, vol. I, Polish Scientific Publishers, Warsaw, 1976, pp. 107–113.
- [27] F. KAPPEL, *On degeneracy of functional differential equations*, J. Differential Equations, 22 (1976), pp. 250–267.
- [28] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York/Berlin, 1966.
- [29] A. KORYTOWSKI, *Functional controllability of a system with delay*, Arch. Automat. i Telemech., 20 (1975), pp. 19–28.
- [30] 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.
- [31] E. KAMEN, *On an algebraic theory of systems defined by convolution operators*, Math. System Theory, 9 (1975), pp. 57–74.
- [32] ———, *On an operator theory of linear systems with pure and distributed delays*, Proceedings of 1975 IEEE Conference on Decision and Control (Houston, TX).
- [33] F. KAPPEL, *Some remarks to the problem of degeneracy for functional differential equations*, Equations Différentielles et Fonctionnelles, Actes du Colloque "Equadiff" à Bruxelles, Sept. 1973, P. Jansens, J. Mawhin and N. Rouche, eds., Hermann, Paris, 1974.
- [34] P. LANCASTER *Theory of Matrices*, Academic Press, New York, 1969.
- [35] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

- [36] E. B. LEE, *Linear Hereditary Control Systems*, Calculus of Variations and Control Theory, Academic Press, New York, 1976, pp. 47–72.
- [37] C. E. LANGENHOP, *A row reduction of λ -matrices*, Linear Algebra Appl., 9 (1974), pp. 185–198.
- [38] D. G. LUENBERGER, *Canonical forms for linear multivariable systems*, IEEE Trans. Autom. Control, AC-12 (1967), pp. 290–293.
- [39] A. MANITIUS, *Optimal Control of Hereditary Systems*, Control Theory and Topics in Functional Analysis, vol. III, International Atomic Energy Agency, Vienna, 1976, pp. 43–178.
- [40] A. MANITIUS AND A. W. OLBROT, *Controllability conditions for linear systems with delayed states and control*, Arch. Automat. i Telemekh., 17 (1972), pp. 119–131.
- [41] A. MANITIUS, *On controllability conditions for systems with distributed delays in state and control*, Ibid., 17 (1972), pp. 363–377.
- [42] ———, *Optimal control of linear time-lag processes with quadratic performance indexes*, Proc. IV Congress of the International Federation of Automatic Control (Warsaw, Poland, 1969), vol. 13, pp. 16–28.
- [43] A. MANITIUS AND R. TRIGGIANI, *Function space controllability of linear retarded systems: A derivation from abstract operator conditions*, Rep. CRM-605, Centre de Recherches Mathématiques, Université de Montréal, 1976.
- [44] ———, *Sufficient conditions for function space controllability and feedback stabilizability of linear retarded systems*, Proc. 1976 IEEE Conference on Decision and Control, pp. 1209–1216.
- [45] A. MANITIUS, *Controllability, observability and stabilizability of retarded systems*, Proc. 1976 IEEE Conference on Decision and Control, pp. 752–758.
- [46] ———, *Function space controllability of retarded systems: some new algebraic conditions*, Proc. XIV Allerton Conference on Circuit and System Theory, University of Illinois at Urbana-Champaign, 1976, pp. 320–327.
- [47] S. A. MINJUK, *On complete controllability of linear controllable systems with delay*, Differencial'nye Uravnenija 8 (1972), pp. 254–259.
- [48] S. A. MINJUK AND N. N. STEPANJUK, *The theory of completely controllable linear systems with delay*, Ibid., 10 (1974), pp. 629–634.
- [49] A. S. MORSE, *Ring models for delay differential systems*, Automatica, 12 (1976), pp. 529–531.
- [50] A. W. OLBROT, *Algebraic criteria of controllability to zero function for linear constant time-lag systems*, Control and Cybernetics 2 (1973), pp. 59–77.
- [51] ———, *Control of retarded systems with function space constraints. Part 2: Approximate controllability*, Ibid., 6 (1977), no. 2.
- [52] ———, *Stabilizability, detectability and spectrum assignment for linear systems with general time delays*, Rep. CRM-712, Centre de Recherches Mathématiques, Université de Montréal, 1977.
- [53] YU. S. OSIPOV, *Stabilization of controlled systems with delays*, Differencial'nye Uravnenija 1 (1965), no. 5, pp. 605–618.
- [54] L. PANDOLFI, *On feedback stabilization of functional differential equations*, Boll. Un. Mat. Ital., 4, 11, supplemento al fascicolo 3, Giugno 1975, Serie IV, Vol. XI, pp. 626–635.
- [55] ———, *On the infinite dimensional controllability of differential-difference control processes*, Boll. Un. Mat. Ital., 10 (1974), pp. 114–123.
- [56] M. C. PEASE III, *Methods of Matrix Algebra*, Academic Press, New York, 1965.
- [57] V. M. POPOV, *On the property of reachability for some delay-differential equations*, Tech. Rep. R-70-08, University of Maryland, College Park, 1970.
- [58] ———, *Pointwise degeneracy of linear time-invariant, delay-differential equations*, J. Differential Equations, 11 (1972), pp. 541–561.
- [59] E. D. SONTAG, *Linear systems over commutative rings: A survey*, Ricerche Automatica, 7 (1976), pp. 1–34.
- [60] A. E. TAYLOR, *Introduction to Functional Analysis*, John Wiley, New York, 1958.
- [61] R. TRIGGIANI, *Controllability and observability in Banach space with bounded operators*, this Journal, 13 (1975), pp. 462–491.
- [62] ———, *Pathological asymptotic behavior of control systems in Banach space*, J. Math. Anal. Appl., 49 (1975), pp. 411–429.
- [63] ———, *On the lack of exact controllability for mild solutions in Banach space*, Ibid., 50 (1975), pp. 438–446.
- [64] ———, *Extensions of rank conditions for controllability and observability to Banach spaces and unbounded operators*, this Journal, 14 (1976), pp. 313–338.
- [65] ———, *On the stabilizability problem in Banach space*, J. Math. Anal. Appl., 52 (1975), pp. 383–403; Addendum, Ibid., 56 (1976).

- [66] ———, *A note on the lack of exact controllability for mild solutions in Banach spaces*, this Journal, 15 (1977), pp. 407–411.
- [67] R. VINTER, *On the evolution of the state of linear differential delay equations in M^2 : properties of the generator*, Tech. Rep. ESL-R-541, Mass. Inst. of Tech., Cambridge, 1974.
- [68] N. S. WILLIAMS AND V. ZAKIAN, *A ring of delay operators with applications to delay-differential systems*, this Journal, 15 (1977), pp. 247–255.
- [69] R. B. ZMOOD, *On Euclidean space and function space controllability of control systems with delay*, Doctoral Dissertation, University of Michigan, Ann Arbor, 1971.
- [70] ———, *The Euclidean space controllability of control systems with delay*, this Journal, 12 (1974), pp. 609–623.
- [71] A. M. ZVERKIN, *On pointwise completeness of systems with lag*, Conference of Lumumba University of Friendship of Peoples (Moscow, May 24–27). Also: *Differencial'nye Uravnenija* 9 (1973), no. 3, pp. 430–436.

NUMERICAL ASPECTS OF RECURSIVE REALIZATION ALGORITHMS*

LIEUWE SYTSE DE JONG†

Abstract. The known recursive algorithms for the minimal realization problem are numerically unstable. The reasons for the numerical instability are explained for Rissanen's algorithm. It is shown how the algorithm may be "stabilized". Finally, a numerically stable algorithm is assessed with respect to efficiency, capability (i.e. which problems can be dealt with on the available computer) and with respect to suitability for finding approximate minimal realizations.

1. Introduction. An internal description of a *discrete, linear, dynamical system* with p input and q output terminals is given by

$$\Sigma_n(A, B, C) \equiv \begin{cases} x(t+1) = Ax(t) + Bu(t), \\ y(t) = Cx(t). \end{cases}$$

The number n is called the *order* of the system. A , B and C are respectively $n \times n$, $n \times p$ and $q \times n$ matrices with elements from a field F (say). The vectors $x(t)$, $u(t)$ and $y(t)$ are respectively n -, p - and q -dimensional vectors with elements from F , called *state*, *input* and *output* vectors. The time is denoted by t .

We consider the system at times $t = 0, 1, 2, \dots$. If at $t = 0$ the system is at rest ($x(0) = 0$) and an impulse $u(0)$ is applied as an input, then the *impulse response* is

$$y(t) = CA^{t-1}Bu(0) \quad (t = 1, 2, \dots).$$

The (*complete*) *minimal realization problem* MRP reads: given a sequence of $q \times p$ matrices $(S_i)_{i=1}^{\infty}$ with elements from F , to determine a system $\Sigma_n(A, B, C)$ of minimal order that has the impulse response

$$y(t) = Su(0) \quad (t = 1, 2, \dots).$$

Hence, A , B and C are to be determined such that A has minimal dimensions and $S_t = CA^{t-1}B$ ($t = 1, 2, \dots$). A solution $\Sigma_n(A, B, C)$ of the MRP is called a *minimal realization* of the impulse response $(S_i)_{i=1}^{\infty}$.

The MRP is classical in the single-input, single-output case ($p = q = 1$) and with F the field of reals; for historical details we refer to de Jong [4]. Yet it lasted till the mid 1960's before algorithms were proposed capable of solving the MRP without unnecessary restrictions on the impulse response (Ho, Kalman; Youla, Tissi; Silverman [3]; [12]; [10]). All these algorithms determine a minimal realization from a decomposition or a factorization of the *Hankel matrix*

$$H_{k,l} = \begin{bmatrix} S_1 & S_2 & \cdots & S_l \\ S_2 & S_3 & \cdots & S_{l+1} \\ \vdots & \vdots & \ddots & \vdots \\ S_k & S_{k+1} & \cdots & S_{k+l-1} \end{bmatrix}$$

where k and l should be large enough; if α and β denote controllability and observability indices (assuming that these exist), then it is a precondition that $k \geq \beta$, $l \geq \alpha$, $k + l > \alpha + \beta$. The algorithms do not prescribe how the decomposition should be obtained.

In 1971 Rissanen proposed an algorithm where the decomposition is found in a recursive way: it is obtained by repeatedly updating the decomposition of a smaller

* Received by the editors March 7, 1977, and in revised form August 17, 1977.

† Computing Center, Eindhoven University of Technology, Eindhoven, The Netherlands.

Hankel matrix to the decomposition of a larger Hankel matrix [9]. The advantage of such an approach is that it is appropriate for solving the *partial* minimal realization problem (find a system that realizes a first finite part of $(S_i)_{i=1}^\infty$). Besides, this approach is suitable for computing approximate minimal realizations in case the given impulse response contains noise. Last but not least, recursive algorithms are by a factor n more efficient than nonrecursive algorithms in case the order n is not a priori known.

Prior to the work of Rissanen, Massey gave an efficient recursive solution for the single-input, single-output case [8]. He used an algorithm, which was developed by Berlekamp for the decoding of BCH codes [1]. The algorithm is not based upon a Hankel matrix approach but it may be formulated in that way. In 1974, Dickinson, Morf and Kailath generalized the Massey/Berlekamp algorithm to the multiple-input, multiple-output case [2].

If, however, F is an infinite field, then these recursive algorithms are numerically unstable in the sense that became customary in numerical analysis. The inexact minimal realization as it is supplied by a computing machine with finite arithmetic might not be a "numerical neighbor" of a minimal realization that corresponds with a "numerical neighbor" of the given $(S_i)_{i=1}^\infty$. This is caused by some malicious round-off errors that may gain in importance and destroy all accuracy. Because all recursive algorithms mentioned above are based upon the same principles (see de Jong [4]), we shall investigate only Rissanen's algorithm and show where such malicious round-off errors may occur. We shall elucidate how the algorithm may be stabilized and, eventually, assess a numerically stable algorithm.

A seeming simplification in this article is that only the single-input, single-output case is discussed. However, with respect to the numerical properties of the algorithm, this implies no restrictions.

2. Rissanen's algorithm. From now on we assume that $p = q = 1$ and that $F = \mathbb{R}$. We suppose that $(S_i)_{i=1}^\infty$ has a minimal realization of order n . Before formulating the algorithm we state two basic results due to Kalman et al. [6], [7].

LEMMA 1. *The impulse response $(S_i)_{i=1}^\infty$ has a minimal realization of order n if and only if*

$$\text{rank}(H_{n+1,j}) = \text{rank}(H_{n,j}) = n \quad (j = n, n+1, \dots).$$

LEMMA 2. *Any solution of the MRP is unique up to a similarity transformation: if $\Sigma_n(A, B, C)$ is a solution, then all other solutions are obtained by choosing a regular $n \times n$ matrix M and forming*

$$\Sigma_n(MAM^{-1}, MB, CM^{-1}).$$

The following corollary is an immediate consequence of Lemma 1.

COROLLARY 3. *If $(S_i)_{i=1}^\infty$ has a minimal realization of order n , then there exists a unique vector x such that $(x^T, -1)H_{n+1,j} = 0^T$ ($j = 1, 2, \dots$). A minimal realization of $(S_i)_{i=1}^\infty$ is then*

$$\Sigma_n \left(\begin{bmatrix} 0 & 1 & 0 & \cdot & 0 \\ 0 & 0 & 1 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & 1 \\ x_1 & x_2 & x_3 & \cdot & x_n \end{bmatrix}, \begin{bmatrix} S_1 \\ S_2 \\ \cdot \\ S_{n-1} \\ S_n \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}^T \right).$$

The condition of Lemma 1 is known as the "rank condition". It expresses that the last row of the Hankel matrix $H_{n+1,*}$ is a linear combination of the other rows of

$H_{n+1,*}$. The vector x of Corollary 3 contains the coefficients of this combination. Because $\text{rank}(H_{n+1,n}) = \text{rank}(H_{n,n}) = n$, the vector x is uniquely determined by $H_{n+1,n}$. Hence, a minimal realization of $(S_i)_{i=1}^{\infty}$ may be computed from the first $2n$ elements of $(S_i)_{i=1}^{\infty}$. Therefore, a realization algorithm needs only to investigate a *finite* number of elements of $(S_i)_{i=1}^{\infty}$. However, it is necessary that n or an upper bound for n is known. Otherwise, the algorithm cannot decide when *enough* elements were investigated.

Let us suppose that N is an upper bound for n . First, we give an outline of Rissanen's algorithm.

The algorithm starts out with a pair (k, l) such that

$$(2.1) \quad \text{rank}(H_{k+1,l}) = \text{rank}(H_{k,l}) = k$$

and a decomposition of $H_{k+1,l}$ from which $\text{rank}(H_{k+1,l})$ and the vector x such that $(x^T, -1)H_{k+1,l} = 0^T$ may be easily obtained. If $k + l \geq 2N$, then enough elements of $(S_i)_{i=1}^{\infty}$ were investigated and the algorithm may terminate, otherwise there are two possibilities. Either (2.1) holds for all values of $l \leq 2N - k$, which is verified by investigating the decomposition for $H_{k+1,2N-k}$ (obtained by gradually extending the decomposition of $H_{k+1,l}$) and in this case the algorithm may terminate. Or there exists an $l' > l$ such that $\text{rank}(H_{k+1,l'}) = k + 1$ and then $k' > k$ is determined such that

$$\text{rank}(H_{k'+1,l'}) = \text{rank}(H_{k',l'}) = k',$$

which is the starting point again and the procedure may be repeated. The decomposition for $H_{k'+1,l'}$ is obtained by first extending the decomposition for $H_{k+1,l}$ to $H_{k+1,l'}$ and, thereafter, to $H_{k'+1,l'}$. The process is finite because $k' + l' > k + l$. At the end a minimal realization based on Corollary 3 is constructed. The vector x is deduced from the final decomposition.

THE ALGORITHM (in pseudo-ALGOL).

0. {a decomposition for $H_{k+1,l}$ is known; (2.1) is satisfied};

1. **while** $k + l < 2 * N$ **do**

2. **begin** $l := l + 1$; {update the decomposition};

3. **while** $\text{rank}(H_{k+1,l}) = k + 1$ **do**

4. **begin** $k := k + 1$; {update the decomposition}

5. **end**

6. **end**;

7. {print a minimal realization}

We shall prove that the algorithm is correct; viz., if N is an upper bound for n , then the algorithm is finite and at the end it supplies in k the value of n . The finiteness follows from the observations that in each cycle the sum of k and l increases at least by one (line 2) and that the loop in line 3 due to the properties of the rank can be executed at most $(l - k)$ times.

Next, let us show that finally $k = n$. If just before execution of the loop in line 3 we have $\text{rank}(H_{k,l}) = k$, then this is also true just after execution of the loop. Using this, it is easy to verify that (2.1) is an invariant relation for the loop in line 1. Consequently, if at the beginning of the algorithm (2.1) holds, then (2.1) also holds at the end when $k + l \geq 2 * N$.

The final value of k cannot be greater than n , because then $\text{rank}(H_{k,l}) < k$ (Lemma 1). On the other hand, k cannot be smaller than n , since that would mean $l \geq N \geq n$; the rank condition of Lemma 1 then implies that $\text{rank}(H_{k+1,l}) = k + 1$, which contradicts (2.1). So at the end we must have $k = n$.

Let us now turn to the decomposition of $H_{k+1,l}$. Theorem 4 describes a decomposition for $H_{k+1,l}$ from which $\text{rank}(H_{k+1,l})$ and the vector x of Corollary 3 (if it exists) can be determined, and which can be readily updated in case k or l are increased.

THEOREM 4. *If $\text{rank}(H_{k,l}) = k$, then $H_{k+1,l}$ has a decomposition*

$$M_{k+1,k+1}H_{k+1,l}P = R_{k+1,l}$$

of the following kind:

$M_{k+1,k+1}$ is a $(k+1) \times (k+1)$ regular matrix, P is a permutation matrix and $R_{k+1,l}$ has upper trapezoidal form:

$$R_{k+1,l} = \begin{bmatrix} * & \cdot & * & \cdot & \cdot & * \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 0 & * & \cdot & \cdot & * \\ 0 & \cdot & 0 & * & \cdot & * \end{bmatrix}.$$

The last row of $R_{k+1,l}$ is zero if and only if $\text{rank}(H_{k+1,l}) = k$.

We shall not prove this elementary theorem. Instead we assume that k and l are known such that (2.1) holds and, moreover, that for $H_{k+1,l}$ a decomposition as in Theorem 4 is available; we shall show how the decomposition is updated whenever k or l increase as in the algorithm (note that in case $S_1 \neq 0$, one may take $k = l = 1$).

Updating the decomposition after $l := l + 1$; (line 2). A decomposition for $H_{k+1,l-1}$ is available. Since $\text{rank}(H_{k+1,l-1}) = k$, the last row of $R_{k+1,l-1}$ is zero. We have

$$(2.2) \quad M_{k+1,k+1}H_{k+1,l} \begin{bmatrix} P & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} R_{k+1,l-1} \\ \vdots \end{bmatrix} M_{k+1,k+1} \begin{bmatrix} S_l \\ \vdots \\ S_{k+1} \end{bmatrix}.$$

If the last row of the right-hand side matrix is zero, then the updating process finishes.

Otherwise the new permutation matrix is adapted such that $M_{k+1,k+1} \begin{bmatrix} S_l \\ \vdots \\ S_{k+l} \end{bmatrix}$ becomes

the $(k+1)$ th column of the right-hand side matrix.

Updating the decomposition after $k := k + 1$; (line 4). A decomposition for $H_{k,l}$ is available and we know that the last row of $R_{k,l}$ is not zero, since $\text{rank}(H_{k,l}) = k$. We have

$$M_{k,k}H_{k,l}P = R_{k,l}.$$

Let (m_1, m_2, \dots, m_k) denote the last row of $M_{k,k}$. Then

$$(2.3) \quad \begin{bmatrix} \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & m_1 & \dots & m_{k-1} & m_k \end{bmatrix} H_{k+1,l}P = \begin{bmatrix} R_{k,l} \\ c^T \end{bmatrix}.$$

Due to the Hankel structure of $H_{k+1,l}$, the use of the shifted last row brings about that at least $(l-1)$ components of c^T are identical to $(l-1)$ components of the last row of $R_{k,l}$. Hence c^T has at least $(k-2)$ components zero. Investigating the structure of P , it can even be shown that the first $(k-2)$ components of c^T are zero.

If the right-hand side matrix of (2.3) has not upper trapezoidal form then it is brought into that form by premultiplying (2.3) at the left and the right by *elementary*

transformation matrices

$$(2.4) \quad T_i := \begin{bmatrix} 1 & & & \\ & 0 & & \\ & 1 & & \\ 0 & \cdot & \cdot & \\ & \mu_i & \cdot & 1 \end{bmatrix} \leftarrow i \quad (1 \leq i \leq k).$$

The multipliers μ_i are chosen such that

$$T_i \cdots T_1 \begin{bmatrix} R_{k,l} \\ c^T \end{bmatrix} = \begin{bmatrix} \cdots R_{k,l} \cdots \\ 0 \cdots 0 \underbrace{* \cdots *}_i \end{bmatrix} \quad (1 \leq i \leq k).$$

Since the first $(k-2)$ components of c^T are zero, only μ_k and μ_{k-1} are nonzero. If $\text{rank}(H_{k+1,l}) = k$, then the last row of the transformed R is zero and the updating process finishes. Otherwise, the permutation matrix P is adapted such that the transformed matrix R has upper trapezoidal form and nonzero diagonal.

Remarks. 1. The presentation of the algorithm differs from Rissanen's presentation.

2. By analyzing the updating process in detail one can show that only $2Nn + \frac{1}{2}n^2$ operations (counting multiplications and divisions) are necessary. The main reason for this very small number is the use of the shifting trick. Due to this trick one needs only two transformation matrices in (2.3) and, moreover, one needs only to compute the last two components of

$$M_{k+1,k+1} \begin{bmatrix} S_l \\ \vdots \\ S_{k+1} \end{bmatrix} \quad \text{in (2.2).}$$

3. Without loss of generality we may assume that the $(k+1, k+1)$ st element of any matrix $M_{k+1,k+1}$ that occurs in the computational process is equal to 1. This follows from the observation that the transformation matrices T_i of (2.4) do not affect the last column of the matrix

$$\left[\begin{array}{c|c} M_{k,k} & 0 \\ \hline 0 & m_1 \cdots m_k \end{array} \right];$$

after premultiplication by the T_i this matrix still has m_k as $(k+1, k+1)$ st element.

4. If in (2.4) it holds that $|\mu_i| \leq 1$, then the elementary transformation matrix T_i is called *stabilized*, else *unstabilized* (compare Wilkinson [11]).

5. If a vector $\begin{bmatrix} a \\ b \end{bmatrix}$ has to be transformed into the form $\begin{bmatrix} * \\ 0 \end{bmatrix}$, as in the updating process of the algorithm of Rissanen, then there are two elementary transformation matrices with which this can be achieved:

$$\begin{bmatrix} 1 & 0 \\ -ba^{-1} & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} a \\ 0 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 0 & 1 \\ 1 & -ab^{-1} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}.$$

Either the left or the right transformation matrix is stabilized. Rissanen's algorithm always employs the left transformation matrix.

6. Another way to transform a vector $\begin{bmatrix} a \\ b \end{bmatrix}$ into the form $\begin{bmatrix} * \\ 0 \end{bmatrix}$ is by means of a plane rotation matrix or a Givens matrix. Let ϕ be the (rotation) angle with $0 \leq |\phi| \leq$

$\pi/2$ such that $\sin \phi = b/r$, $\cos \phi = a/r$, where $r = (a^2 + b^2)^{1/2}$. Then

$$S := \begin{bmatrix} \cos \phi & \sin \phi \\ -\sin \phi & \cos \phi \end{bmatrix}$$

is called a plane rotation matrix and

$$S \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} r \\ 0 \end{bmatrix}.$$

Note that S is an orthogonal matrix and that the product of any number of orthogonal transformation matrices is again an orthogonal matrix.

3. The numerical stability of Rissanen's algorithm. The computations (multiplications, additions) performed by a digital computer are afflicted with round-off errors. Since the real numbers are replaced by a finite set of machine numbers, it is a coincidence if the exact result of an operation (addition, multiplication) applied to machine numbers is again a machine number. If it is a "good" machine then either it supplies the machine number, which is the closest to the exact result, or it breaks the computational process down in case of underflow or overflow. However, even though the computer does the best one may expect for *one* operation, a machine number resulting from a *series* of operations may deviate prohibitively from the exact result. A much quoted example concerns the computation of the roots of a quadratic equation $ax^2 + 2bx + c = 0$ via the formula

$$y(\pm) = (-b \pm (b^2 - ac)^{1/2})/a.$$

If $|ac| \ll b^2$ and $b > 0$, then the computed approximation for $y(+)$ is very bad, due to a loss of accuracy when computing $-b + (b^2 - ac)^{1/2}$. Because $-b + (b^2 - ac)^{1/2}$ is much smaller than $(b^2 - ac)^{1/2}$, the round-off error made when computing $(b^2 - ac)^{1/2}$ gains very much in importance when $(b^2 - ac)^{1/2}$ is diminished with b . The obvious remedy is to compute $y(+)$ by means of the formula

$$y(+) = -c/(b + (b^2 - ac)^{1/2}).$$

One should be well on one's guard against this phenomenon of numerical instability. It is inadmissible that round-off errors made somewhere in the computational process can gain in importance and eclipse the final result.

Elaborating on this it is possible to give concrete definitions of numerical stability and instability and to prove that Rissanen's algorithm is numerically unstable (see [4]). Here we restrict ourselves to a heuristic approach.

The following demands should be met for numerical stability.

1. The last row of the computed $\bar{M}_{k+1,k+1}$ should deviate little from the last row of the exact $M_{k+1,k+1}$.

2. The computed matrix $\bar{R}_{k+1,l}$ should enable trustworthy statements concerning the rank of $H_{k+1,l}$.

The first demand is necessary because the minimal realization is based upon the last row of $\bar{M}_{k+1,k+1}$. The second demand requires some explanation. The rank of $H_{k+1,l}$ is determined by investigating $R_{k+1,l}$: if the bottom row is zero, then the rank is equal to k ; otherwise it is equal to $k + 1$. However, only a perturbation $\bar{R}_{k+1,l}$ is known. If $R_{k+1,l}$ has a zero bottom row then $\bar{R}_{k+1,l}$ will probably have a nonzero bottom row. It is essential then that this row is small compared to the rest of $\bar{R}_{k+1,l}$. Otherwise the wrong conclusion about rank ($H_{k+1,l}$) and, finally, about the order of a minimal realization is drawn. We shall show that both demands may be violated.

The first demand. Let us consider the multiplication of $\bar{M}_{k+1,k+1}$ and an elementary transformation matrix at line 4 of the algorithm. Omitting the insignificant parts of the transformation matrix and the irrelevant rows of $\bar{M}_{k+1,k+1}$ we have essentially the resulting transformation

$$(3.1) \quad \text{fl}\left(\begin{bmatrix} 1 & 0 \\ \mu & 1 \end{bmatrix} \begin{bmatrix} a^T & 0 \\ m^T & 1 \end{bmatrix}\right) = \begin{bmatrix} a^T & 0 \\ \text{fl}(\mu a^T + m^T) & 1 \end{bmatrix}.$$

The notation fl indicates that the operations are performed by a (floating point) computer. By $(a^T, 0)$ we denote one of the first k rows of $\bar{M}_{k+1,k+1}$ and $(m^T, 1)$ corresponds with the last row of $\bar{M}_{k+1,k+1}$. We have

$$\text{fl}(\mu a^T + m^T) = (\mu a^T(I + E_1) + m^T)(I + E_2) \sim (\mu a^T + m^T)(I + E_2) + \mu a^T E_1,$$

where E_1 and E_2 are diagonal matrices with elements in the order of the relative machine precision. The round-off error $\mu a^T E_1$ may eclipse the last row in case μ is large but $\mu a^T + m^T$ is not.

Given an impulse response sequence, it is impossible to give a priori bounds for the multipliers μ ; these numbers can be unpredictably large. So the first demand is not satisfied.

Remark 5 in § 2 indicates that the use of unstabilized elementary transformation matrices ($|\mu| > 1$) may be avoided. Doing so the $(k+1, k+1)$ st element of $\bar{M}_{k+1,k+1}$ may become arbitrarily small, because then it is the product of a number of multipliers, which all have modulus less than 1. So then the following situation may arise at line 4 of the algorithm:

$$\text{fl}\left(\begin{bmatrix} 0 & 1 \\ 1 & \mu \end{bmatrix} \begin{bmatrix} a^T & 0 \\ m^T & \varepsilon \end{bmatrix}\right) = \begin{bmatrix} m^T & \varepsilon \\ \text{fl}(a^T + \mu m^T) & \text{fl}(\mu \varepsilon) \end{bmatrix},$$

with $|\mu| \leq 1$ and $|\varepsilon| \ll 1$. By $(a^T, 0)$ we denote again one of the first rows of $\bar{M}_{k+1,k+1}$ and (m^T, ε) denotes the last row. We have

$$\text{fl}(a^T + \mu m^T) \sim (a^T + \mu m^T)(I + E_2) + \mu m^T E_1,$$

where as before E_1 and E_2 denote diagonal matrices with elements of the order of the relative machine precision. The round-off error $\mu m^T E_1$ may eclipse the last row of $\bar{M}_{k+1,k+1}$ in case a^T and m^T are of the order of 1 but $a^T + \mu m^T$ is of the order of ε . So the use of stabilized matrices does not help. If the shifting trick is eliminated then the $(k+1, k+1)$ st element of $\bar{M}_{k+1,k+1}$ is the product of at most k multipliers, which is better than before, but also then it is possible that a similar situation arises. The fundamental reason that all three variants fail is the following. If a matrix is premultiplied by elementary transformation matrices, then its *columns* are transformed, not its *rows*. Hence, all one may expect is that the columns have a good accuracy; there is no reason why the rows should have small relative errors. However, if it were a priori known that the transformations preserve the length of the rows of M (as it would be in the case of orthogonal transformations without employing a shifting trick), then one may also expect a good accuracy in the rows.

The conclusion is that either the use of elementary transformation matrices be abandoned or that the minimal realization be based upon another formula.

The second demand. Let us suppose that for the computed $\bar{M}_{k+1,k+1}$ and $\bar{R}_{k+1,l}$ we have

$$\bar{M}_{k+1,k+1} H_{k+1,l} P = \bar{R}_{k+1,l} + \Delta R_{k+1,l}.$$

Let $\sigma(A)$ denote¹ the distance of a matrix A to the nearest matrix of lower rank in the matrix norm that is subordinate to the Euclidean vector norm. Then we can show that (omitting the subindices and denoting this norm by $\|\cdot\|$)

$$(3.2) \quad \|\bar{M}\|^{-1}(\sigma(\bar{R}) - \|\Delta R\|) \leq \sigma(H) \leq \|\bar{M}^{-1}\|(\sigma(\bar{R}) + \|\Delta R\|)$$

and that both sides of this inequality may be sharp.

Estimates for both bounds should be computed. If the upper bound is of the order of the machine precision times $\|H\|$, then one should conclude that $\text{rank}(H) = k$. Otherwise, if the lower bound is large with respect to the machine precision times $\|H\|$, then one should conclude that $\text{rank}(H) = k + 1$. This strategy is successful provided that

1. the machine precision is small enough,
2. $\|\Delta R\|$ is of the order of the machine precision times $\|R\|$,
3. $\|\bar{M}\| \|\bar{M}^{-1}\| (= \text{Cond}(\bar{M}))$ can a priori be bounded

(see de Jong [4]). The first point is natural for all computational problems. As in (3.1) one may show that the second point may be violated due to the use of unstabilized elementary transformation matrices. Also the third point may be violated if unstabilized transformations are employed. If only stabilized transformations are used then the second point is not violated, but then the third point may be violated due to the shifting trick of (2.3):

$$\text{Cond}\left(\begin{bmatrix} 0 & 1 \\ 1 & \varepsilon \end{bmatrix}\right) \sim 1 \quad \text{but} \quad \text{Cond}\left(\begin{bmatrix} 0 & 1 & 0 \\ 1 & \varepsilon & 0 \\ 0 & 1 & \varepsilon \end{bmatrix}\right) \sim \varepsilon^{-1}$$

(ε is not bounded away from zero).

The conclusion is that the use of unstabilized transformation matrices as well as the shifting trick be abandoned.

Let us summarize. If we want to stabilize the algorithm of Rissanen maintaining the companion form realization, then the use of elementary transformation matrices as well as the shifting trick must be avoided. Instead, the decomposition of the successive Hankel matrices should be updated with orthogonal transformation matrices. On the other hand, if we are prepared to employ another formula for a minimal realization, e.g. based upon the matrix $R_{k+1,l}$, then we may use (stabilized) elementary transformation matrices provided that the shifting trick is abandoned. With respect to the number of operations (which is of the order of n^3) it is not possible to have a preference for one of the methods, even if in the orthogonal transformation method the minimal realization is based upon $R_{k+1,l}$.

The reason for this somewhat surprising fact is that in the elementary transformation method one needs to compute estimates for $\|M^{-1}\|$ whereas in the other method M is orthogonal implying that then $\|M^{-1}\| = 1$. Orthogonal transformation matrices are numerically more stable than elementary transformation matrices. Moreover, they lead to sharper bounds in (3.2). Therefore we prefer the orthogonal transformation method with a minimal realization, which is based upon $R_{k+1,l}$; a companion form matrix is a bad starting point for further computations, for instance, of the eigenvalues of the realization matrix.

4. Some aspects of a numerically stable realization algorithm. We shall from now on consider the minimal realization algorithm that employs orthogonal transformation matrices and that delivers a minimal realization based upon $R_{k+1,l}$. We shall

¹ $\sigma(A)$ is the smallest singular value of A .

not give proof of the numerical stability of the algorithm (which has been done in [4]), but discuss and assess some important features of the algorithm.

4.1. The numerical stability of the decomposition. The matrix $H_{k+1,l}$ of which the algorithm finally delivers a decomposition is transformed a number of times (to be precise: $\frac{1}{2}k(1+k)$ times) by orthogonal transformation matrices

$$T(2, 1); \quad T(3, 1), T(3, 2); \quad \cdots; \quad T(k+1, 1), \cdots, T(k+1, k).$$

A matrix $T(i, j)$ corresponds with a rotation in the (i, j) plane. Each $T(i, j)$ is determined by one parameter: the rotation angle ϕ . In practice it is more convenient to use the parameters $\cos(\phi)$ and $\sin(\phi)$ or possibly some other set of parameters depending on how the transformation is actually arranged (see de Jong [4]). Anyway, it is not necessary to evaluate the resulting transformation matrix $M_{k+1,k+1}$ explicitly; it can be stored in “product” form. In [4] it is shown that (for any occurring value of k and l)

$$(4.1) \quad M_{k+1,k+1}H_{k+1,l}P = R_{k+1,l} + \Delta R_{k+1,l}$$

where $M_{k+1,k+1}$ is the exact product of the exact plane rotations, $R_{k+1,l}$ is the matrix as it is obtained on a floating point computer with the following upper bound for $\Delta R_{k+1,l}$:

$$(4.2) \quad \|\Delta R_{k+1,l}\| \leq C \cdot k(1+k) \cdot \eta \cdot \|H_{k+1,l}\|.$$

C is a constant whose value depends on the actual arrangement of the transformation and η is the relative machine precision. Note that this result is the best one can expect.

4.2. The determination of the rank. We wish to determine rank $(H_{k+1,l})$ from the decomposition (4.1). We shall omit the subscripts $(k+1)$ and l in this section. From the regularity of M and P it follows that

$$\text{rank}(H) = \text{rank}(R + \Delta R),$$

and from the orthogonality of M and P it follows that

$$\sigma(H) = \sigma(R + \Delta R).$$

A singular value decomposition of R would supply $\sigma(R)$ with an accuracy of the order of $\eta\|R\|$. Because of (3.2) we have

$$\sigma(R) - \delta \leq \sigma(H) \leq \sigma(R) + \delta,$$

where $\delta := \|\Delta R\|$ is also of the order of $\eta\|R\|$. Hence we would find a value for $\sigma(H)$ that with respect to accuracy cannot essentially be improved upon by any other numerical method. This shows that a singular value decomposition of R enables a very trustworthy statement about rank (H) .

Such a decomposition however would involve $O(lk^2)$ operations, which would make the algorithm of Rissanen on the whole an $O(n^4)$ process; the rank of a matrix H has to be determined at least n times.

Another possibility would be to determine rank (H) directly from a singular value decomposition of the matrices H itself. However, the construction of a singular value decomposition in a recursive way is also an $O(n^4)$ process. Therefore, we investigate whether there is another way to determine rank (H) .

Let

$$R = R_{k+1,l} = \begin{bmatrix} R_{11} & R_{12} \\ 0^T & R_{22} \end{bmatrix},$$

where R_{11} denotes the $(k \times k)$ upper triangular part of R . If all computations are exact

then R_{22} is a zero row if and only if $\text{rank}(H) = k$. In the presence of round-off errors we expect that if $\text{rank}(H) = k$, the row R_{22} approaches zero if the relative machine precision η approaches zero. It may be shown that

$$\|R_{22}\| \left(\left\| \begin{bmatrix} -R_{11}^{-1}R_{12} \\ I \end{bmatrix} \right\| + \|R_{11}^{-1}\| \|R_{22}\| \right)^{-1} \leq \sigma(R) \leq \|R_{22}\|.$$

Hence we arrive at the following bounds for $\sigma(H)$:

$$(4.3) \quad \|R_{22}\| \left(\left\| \begin{bmatrix} -R_{11}^{-1}R_{12} \\ I \end{bmatrix} \right\| + \|R_{11}^{-1}\| \|R_{22}\| \right)^{-1} - \delta \leq \sigma(H) \leq \|R_{22}\| + \delta.$$

We shall say that $\text{rank}(H) < k + 1$ if the upper bound $\|R_{22}\| + \delta$ is small with respect to a tolerance ε and that $\text{rank}(H) = k + 1$ if the lower bound for $\sigma(H)$ is large with respect to ε .

In de Jong [4] it is shown that the upper and lower bound may be estimated in $O(k)$ operations. We shall show that this strategy delivers the correct rank provided that a good choice for ε is made and the relative machine precision η is small enough. Such a choice for ε is

$$(4.4) \quad \varepsilon := 2\delta \left(1 + \left\| \begin{bmatrix} -R_{11}^{-1}R_{12} \\ I \end{bmatrix} \right\| \right).$$

If H has full rank, then the lower bound in (4.3) is bounded away from zero, but ε approaches zero if η approaches zero. Hence, if η is less than η_0 (say), then ε is smaller than the lower bound in (4.3). Rank (H) is then correctly determined. If H has not full rank and η is small enough, then the last row of $R + \Delta R$ is linearly dependent on the remaining rows. Hence

$$-\Delta R_{21}(R_{11} + \Delta R_{11})^{-1}(R_{12} + \Delta R_{12}) + (R_{22} + \Delta R_{22}) = 0^T$$

or

$$R_{22} = (\Delta R_{21}, \Delta R_{22}) \begin{bmatrix} -(R_{11} + \Delta R_{11})^{-1}(R_{12} + \Delta R_{12}) \\ I \end{bmatrix}.$$

Consequently, if η is small enough,

$$\|R_{22}\| \leq 1.1\delta \left\| \begin{bmatrix} -R_{11}^{-1}R_{12} \\ I \end{bmatrix} \right\|.$$

We see that in this case the upper bound in (4.3) is by a factor two smaller than ε . Hence also then rank (H) is correctly determined. This analysis showed that the determination of the rank is an $O(n^2)$ process (globally) provided that the relative machine precision η is small enough.

Remark. The following definition of numerical rank is currently used by Golub et al.: if $\delta > \varepsilon > 0$ and if in an ε - as well as a δ -neighborhood of the matrix A the lowest occurring matrix rank is r then

$$\text{rank}(A, \delta, \varepsilon) = r.$$

The advantage of such a definition is that a neighborhood of A can be found in which this numerical rank is stable: Let $\delta > \varepsilon > 0$ such that $\text{rank}(A, \delta, \varepsilon) = r$. If A is perturbed by a matrix E with $\|E\| = \gamma < (\delta - \varepsilon)/2$, then there exists a matrix B with rank r in an $(\varepsilon + \gamma)$ -neighborhood of $A + E$. Moreover, in a $(\delta - \gamma)$ -neighborhood of $A + E$

there cannot be a matrix B with rank less than r . Hence, if $\text{rank}(A, \delta, \varepsilon) = r$ then also $\text{rank}(A + E, \delta - \|E\|, \varepsilon + \|E\|) = r$, where obviously $\|E\| < (\delta - \varepsilon)/2$.

In practice one chooses for ε a number that is related to the resolution of the computer and one is satisfied if δ/ε is greater than 2 (say). In the Jong [4] the ε -stable rank $r_\varepsilon(A)$ is introduced. One may show that $r_\varepsilon(A) = r(A, 2\varepsilon, \varepsilon)$.

If in (4.3) we denote the lower and upper bound by lb and ub respectively, then we have, if η is small enough,

$$\text{rank}(H_{k+1,l}, \varepsilon, \text{ub}) = k \quad \text{if} \quad \text{rank}(H_{k+1,l}) = k$$

and

$$\text{rank}(H_{k+1,l}, \text{lb}, \varepsilon) = k + 1 \quad \text{if} \quad \text{rank}(H_{k+1,l}) = k + 1.$$

In the first case we have $\varepsilon/(\text{ub}) \sim 2$ and in the second case $\text{lb}/\varepsilon \gg 2$. So, if η is small enough, we have

$$\text{rank}\left(H_{k+1,l}, \varepsilon, \frac{\varepsilon}{2}\right) = \text{rank}(H_{k+1,l}),$$

which shows that our definition of rank is stable in an $(\varepsilon/4)$ -neighborhood of $H_{k+1,l}$. The number ε is related to the distance of the matrix R from the matrix R that would have been obtained with exact computations.

4.3. Capability of the algorithm. From § 4.2, it follows that, if all computations may contain round-off errors of the order of η , the algorithm supplies the correct dimension of a given impulse response provided that η is small enough. Here we investigate how small η actually should be given a particular impulse response and ε given by (4.4). The number ε is obviously of the order of magnitude of $\eta \text{Cond}(H_{k,l}) \cdot \|H_{k+1,l}\|$; $\text{Cond}(H_{k,l})$ is here defined by $\text{Cond}(H_{k,l}) := \|H_{k,l}\| \|H_{k,l}^+\|$ where $H_{k,l}^+$ denotes the pseudo-inverse of $H_{k,l}$ ($H_{k,l}$ is not a square matrix). If $\text{rank}(H_{k+1,l}) = k$, then the lower bound in (4.3) is of the order of magnitude of $\text{Cond}^{-1}(H_{k,l}) \|H_{k+1,l}\|$. Hence the demand that ε should be substantially smaller than the lower bound in (4.3) leads to the condition

$$\eta \text{Cond}^2(H_{k,l}) \ll 1.$$

This condition should of course be true for all k and l such that $\text{rank}(H_{k,l}) = k$. Therefore, we find

$$(4.5) \quad \begin{aligned} \eta \cdot \max \text{Cond}^2(H_{k,l}) &\ll 1. \\ k + l &\leq 2N; \quad \text{rank}(H_{k,l}) = k. \end{aligned}$$

It is difficult to connect this condition to properties of the system to be realized. Let $\Sigma_a(A, B, C)$ be the system and let A have n different eigenvalues $\alpha_1, \alpha_2, \dots, \alpha_n$. Then $H_{n,n}$ has the decomposition

$$H_{n,n} = \begin{bmatrix} 1 & 1 & \cdot & 1 \\ \alpha_1 & \alpha_2 & \cdot & \alpha_n \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \alpha_1^{n-1} & \alpha_2^{n-1} & \cdot & \alpha_n^{n-1} \end{bmatrix} \begin{bmatrix} c_1 & 0 \\ c_2 & \cdot \\ 0 & c_n \end{bmatrix} \begin{bmatrix} 1 & \alpha_1 & \cdot & \alpha_1^{n-1} \\ 1 & \alpha_2 & \cdot & \alpha_2^{n-1} \\ \cdot & \cdot & \cdot & \cdot \\ 1 & \alpha_n & \cdot & \alpha_n^{n-1} \end{bmatrix}.$$

If $\max |c_i| \cdot \min^{-1} |c_i|$ is large, but the α_i are well separated, then $\text{Cond}(H_{n,n})$ will be large. On the other hand, if all c_i have the same order of magnitude then $\text{Cond}(H_{n,n})$

will also be large if for some i and j the eigenvalues α_i and α_j are close to each other. These arguments show that one may not expect a relation between the eigenvalues of A and $\text{Cond}(H_{k,l})$. Given a computing machine, then (4.5) determines which impulse responses can be treated successfully with the algorithm.

Remark. If the decomposition of the matrices $H_{k+1,l}$ is made with the singular value decomposition, then the algorithm is also numerically stable; it requires globally $O(n^4)$ operations, but the demand (4.5) is relaxed to

$$\eta \cdot \max \text{Cond}(H_{k,l}) \ll 1.$$

$$k + l \leq N; \quad \text{rank}(H_{k,l}) = k.$$

4.4. The construction of a minimal realization triple. Let us suppose that at the end of the algorithmic process we have

$$(4.6) \quad M_{k+1,k+1} H_{k+1,l} P = R_{k+1,l} + \Delta R_{k+1,l}$$

with $k = n$, $l > n$ and the estimate (4.2) for $\Delta R_{k+1,l}$. In this section we omit the subscripts $k + 1$ and l .

Let

$$H_1 := \begin{bmatrix} S_1 & \cdots & S_{l-1} \\ \vdots & & \vdots \\ S_k & \cdots & S_{k+l-2} \end{bmatrix} \quad \text{and} \quad H_2 := \begin{bmatrix} S_2 & \cdots & S_l \\ \vdots & & \vdots \\ S_{k+1} & \cdots & S_{k+l-1} \end{bmatrix}.$$

If C denotes the companion matrix of Corollary 3, then it holds that $CH_1 = H_2$. Since H_1 has full row rank and therefore has a right-inverse $H_1^{(r)}$, we find $C = H_2 H_1^{(r)}$. Consequently

$$\Sigma_n(H_2 H_1^{(r)}, (S_1, \dots, S_n)^T, e_1^T)$$

is an alternative formula for the minimal realization of Corollary 3 (e_1 is the first unit vector in \mathbb{R}^n). Taking $L := M^T T^{-1}$ and $U := T(R + \Delta R)P^T$ where T is defined by

$$T := \begin{bmatrix} I & 0 \\ -\Delta R_{21}(R_{11} + \Delta R_{11})^{-1} & 1 \end{bmatrix}$$

we obtain from (4.6) the decomposition for H :

$$H = LU$$

where L is a regular $(k + 1) \times (k + 1)$ matrix and U has zero bottom row.

Let L_{11} and U_{11} denote the matrices that are obtained by removing the last row and the last column of L and U ; let U_{12} denote the matrix that one obtains by removing the last row and the first column of U . Then it can be shown that

$$H_1 = L_{11} U_{11} \quad \text{with } L_{11} \text{ regular, } U_{11} \text{ of full row rank,}$$

$$H_2 = L_{11} U_{12},$$

$$H_2 H_1^{(r)} = L_{11} U_{12} U_{11}^{(r)} L_{11}^{-1}.$$

Using Lemma 2 we obtain the minimal realization

$$(4.7) \quad \Sigma_n(U_{12} U_{11}^{(r)}, U_{11} e_1, e_1^T L_{11}).$$

In practice we choose the matrices $L + \Delta L := M^T$ and $U + \Delta U := R P^T$. For $U_{11}^{(r)}$ we take the pseudo-inverse U_{11}^+ . The realization (4.7) is based upon R . We prefer this

realization to the one of Corollary 3, because it has no companion form and because it contains information of all elements of $\{S_i\}_{i=1}^{k+l}$.

The errors, which are introduced in (4.7) by choosing $L + \Delta L$ and $U + \Delta U$ are harmless. It can readily be shown that

$$\|(\Delta U)_{11}\| = \|(RP^T)_{11}\| \leq \text{Constant} \cdot \|H\| \eta,$$

$$\|(\Delta U)_{12}\| = \|(RP^T)_{12}\| \leq \text{Constant} \cdot \|H\| \eta,$$

$$\|(\Delta L)_{11}\| \leq \text{Constant} \cdot \text{Cond}(H) \cdot \eta,$$

where, as before, η denotes the relative machine precision.

4.5. The efficiency of the algorithm. Counting only multiplications one can show that the decomposition method as well as the computation of $U_{12}U_{11}^+$ need on the order of lk^2 operations. All other computations, including the determination of the rank, can be arranged such that they are asymptotically insignificant. If $N = n$ (N is the a priori given upper bound for n), then $\frac{7}{6}n^3$ operations are needed. This compares favorably with the n^3 operations that are necessary for the inversion of an $n \times n$ matrix. For details we refer to [4].

4.6. Approximate minimal realizations. In practice only a perturbed impulse response $(\tilde{S}_i)_{i=1}^\infty$ is known. The algorithm can then find the correct order of a minimal realization and an approximate minimal realization of the unperturbed impulse response if

$$(4.8) \quad \begin{aligned} W \cdot \max \text{Cond}^2(H_{k,l}) &\ll 1. \\ k + l &\leq 2N; \quad \text{rank}(H_{k,l}) = k. \end{aligned}$$

W is a measure for the noise on $(S_i)_{i=1}^\infty$. The effect of the round-off errors of the computer should of course be negligible with respect to W . The number ε of formula (4.4) should be adapted to account for the noise in $H_{k+1,l}$. If ε is situated between the lower and upper bound of (4.3), then the condition (4.8) is not satisfied and the algorithm cannot determine the order of the unperturbed impulse response.

One may wonder if (4.8) is not satisfied, whether the algorithm may be used to determine the minimum order of all impulse responses in a certain neighborhood of a given, noisy impulse response. For this to be possible, one should be able to compute for a given Hankel matrix $H_{k+1,l}$ of full row rank the nearest Hankel matrix $\tilde{H}_{k+1,l}$ such that $\text{rank}(\tilde{H}_{k+1,l}) < \text{rank}(H_{k+1,l})$. We shall restrict the discussion to a theorem from which such a computation may be derived.

THEOREM 5. *Let $H_{k+1,l} = (S_{i+j-1})$ have full row rank and let $\tilde{H}_{k+1,l} = (\tilde{S}_{i+j-1})$ be the Hankel matrix such that $\text{rank}(\tilde{H}_{k+1,l}) < k + 1$ and $\sum_{i=1}^{k+l} (S_i - \tilde{S}_i)^2$ is minimal. Then*

$$\left(\sum_{i=1}^{k+l} (S_i - \tilde{S}_i)^2 \right)^{1/2} = \min \{ \|X^+ H_{l,k+1} x\|, x \in \mathbb{R}^{k+1}, \|x\| = 1 \},$$

where X is the $l \times (k+l)$ matrix

$$X = \begin{bmatrix} x_1 & \cdots & x_{k+1} & & 0 \\ & \ddots & & \ddots & \\ 0 & & x_1 & \cdots & x_{k+1} \end{bmatrix}.$$

If x is the minimizing vector, then

$$(S_1, \cdots, \tilde{S}_{k+l}) = (S_1, \cdots, S_{k+l})(I - X^+ X).$$

Proof. Let $x \in \mathbb{R}^{k+1}$ have Euclidean length 1. Let $H'_{k+1,l}$ be a Hankel matrix such that $x^T H'_{k+1,l} = 0^T$ (such a matrix does exist; its rank is necessarily less than $k+1$). Define

$$F_{k+1,l} := H_{k+1,l} - H'_{k+1,l};$$

$$f_i := S_i - S'_i \quad (1 \leq i \leq k+l).$$

We have

$$x^T F_{k+1,l} = x^T H_{k+1,l} \quad \text{or} \quad Xf = Xs,$$

where X is defined as in the theorem. Considering $Xf = Xs$ as a system of linear equations in f , the solution with minimal Euclidean length is given by

$$f = X^+ Xs = X^+ H_{l,k+1} x.$$

Minimizing f over all $x \in \mathbb{R}^{k+1}$ supplies the matrix $\tilde{H}_{k+1,l}$. The final assertion of the theorem is now readily obtained. \square

REFERENCES

- [1] E. R. BERLEKAMP, *Algebraic Coding Theory*, McGraw-Hill, New York, 1968.
- [2] B. W. DICKINSON, M. MORF AND T. KAILATH, *A minimal realization algorithm for matrix sequences*, IEEE Trans. Automatic Control, AC-19 (1974), pp. 31–38.
- [3] B. L. HO AND R. E. KALMAN, *Effective construction of linear state variable models from input/output functions*, Regelungstechn., 14 (1966), pp. 545–548.
- [4] L. S. DE JONG, *Numerical aspects of realization algorithms in linear systems theory*, thesis, Eindhoven Univ. of Technology, The Netherlands, 1975.
- [5] R. E. KALMAN, *Mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152–192.
- [6] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1968.
- [7] R. E. KALMAN, *On minimal partial realizations of a linear input/output map*, Aspects of Network and System Theory, Holt, Rinehard and Winston, New York, 1971, pp. 385–407.
- [8] J. L. MASSEY, *Shift-register synthesis and BCH-decoding*, IEEE Trans. Information Theory, IT-15 (1969), pp. 122–127.
- [9] J. RISSANEN, *Recursive identification of linear systems*, this Journal, 9 (1971), pp. 420–430.
- [10] L. SILVERMAN, *Realization of linear dynamical systems*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 554–567.
- [11] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1964.
- [12] D. C. YOULA AND P. TISSI, *N-port synthesis via reactance extraction, Part I*, IEEE Internat. Conv. Rec., 14 (1966), pp. 183–205.

ESTIMATION AND FILTER STABILITY OF STOCHASTIC DELAY SYSTEMS*

RAYMOND H. KWONG† AND ALAN S. WILLSKY‡

Abstract. Linear and nonlinear filtering for stochastic delay systems are studied. A representation theorem for conditional moment functionals is obtained, which, in turn, is used to derive stochastic differential equations describing the optimal linear or nonlinear filter. A complete characterization of the optimal filter is given for linear systems with Gaussian noises. Stability of the optimal filter is studied in the case where there are no delays in the observations. Using the duality between linear filtering and control, asymptotic stability of the optimal filter is proved. Finally, the cascade of the optimal filter and the deterministic optimal quadratic control system is shown to be asymptotically stable as well.

1. Introduction. In recent years, the control of delay differential systems has received considerable attention. Optimal control problems for both linear as well as nonlinear delay systems have been studied intensively. In particular, there is a rather well-developed theory for the optimal control of linear delay systems with a quadratic criterion [1]–[4]. In contrast, optimal filtering for delay systems has not yet received an in-depth study. There is very little literature on the filtering of nonlinear stochastic delay systems which takes into account the structure of such systems. The linear filtering problem on a finite interval has been studied by Kwakernaak [5], Lindquist [6], Mitter and Vinter [7], and recently by Bagchi [8]. Kwakernaak's derivations in [5] were formal; Lindquist [6] did not characterize the covariance of the optimal filter; and Mitter and Vinter [7] restricted their considerations to time-invariant systems and excluded point delays in their observation equations. Bagchi [8] recently gave a rigorous derivation of the filter equations for linear systems with only point delays, using martingale theory and functional analytic methods very different from those in this paper. Stability of the linear filter was also studied recently by Vinter [9], independently of our work. He used infinite dimensional filtering methods, again quite different from our approach. In this paper, we shall study the filtering problem for both nonlinear and linear delay systems. We give a representation theorem which characterizes conditional moment functionals of nonlinear delay systems. Under certain conditions, stochastic differential equations for conditional moment functionals can be derived from the representation theorem. We then specialize these results to obtain the filtering equations for general linear delay systems. We study the stability of the optimal filter in the case of time-invariant systems with no delays in the observations. Under suitable stabilizability and detectability assumptions, we prove that the optimal filter is asymptotically stable. Finally, we combine the linear deterministic control results and the linear filtering results to show that the closed-loop linear stochastic control system is also asymptotically stable.

* Received by the editors June 14, 1976, and in revised form July 27, 1977. This research was performed at the M.I.T. Electronic Systems Laboratory and supported by the National Science Foundation Grant GK41647 and NASA Ames Grant NGL-22-009-124, and at McGill University by National Research Council of Canada under Grants A9067 and A3921, and at the Centre de Recherches Mathématiques, Université de Montréal under subvention FCAC du Ministère de l'Éducation du Québec.

† Department of Electrical Engineering, McGill University and Centre de Recherches Mathématiques, Université de Montréal, Montreal, Quebec, Canada. Now at Department of Electrical Engineering, University of Toronto, Toronto, Ontario, Canada.

‡ Electronic Systems Laboratory and Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

2. Stochastic delay differential systems. We shall study the filtering problem for stochastic delay differential systems of the form

$$(2.1) \quad \begin{aligned} dx(t) &= f(x_t, t) dt + F(t) dw(t), & t \in [0, T], \\ x(\theta) &= x_0(\theta), & \theta \in [-\tau, 0]. \end{aligned}$$

The observation equation is given by

$$(2.2) \quad \begin{aligned} dz(t) &= h(x_t, t) dt + N(t) dv(t), & t \in [0, T], \\ z(t) &= 0, & t \leq 0. \end{aligned}$$

All stochastic processes are defined relative to a given probability space (Ω, \mathcal{F}, P) and on an interval of the form $[0, T]$. The system process $x(t)$ takes values in R^n , the observation process $z(t)$ in R^p . The process x_t is a function on $[-\tau, 0]$ derived from $x(t)$ and is defined by

$$x_t(\theta) = x(t + \theta), \quad \theta \in [-\tau, 0].$$

Unless otherwise stated, we shall let the process x_t take values in \mathcal{C} , the space of R^n -valued continuous functions on $[-\tau, 0]$. For simplicity, we take $w(t)$ and $v(t)$ to be standard Wiener processes in R^m and R^p respectively, completely independent of each other. The initial function x_0 is taken to be some random function on $[-\tau, 0]$, independent of $w(t)$ and $v(t)$. The maps f and h are respectively R^n and R^p -valued functionals defined on $\mathcal{C} \times [0, T]$. The maps $F(t)$ and $N(t)$ are $n \times m$ and $p \times p$ matrix-valued continuous functions respectively. Furthermore, $N(t)$ is assumed to be nonsingular. We shall also write $F(t)F'(t) = Q(t)$ and $N(t)N'(t) = R(t)$.

In order for our estimation problem to be well defined, we need conditions which guarantee existence and uniqueness of solutions to the stochastic functional differential equations (2.1) and (2.2). Such questions have been studied by various authors [10]–[12]. Following their work, we assume that the following conditions are satisfied:

(A1) $f(\phi, t)$ is Borel measurable on $\mathcal{C} \times [0, T]$;

(A2) there exists a bounded measure Γ on $[-\tau, 0]$ and a positive constant K such that for ϕ and ψ in \mathcal{C}

$$|f(\phi, t) - f(\psi, t)| \leq K \int_{-\tau}^0 |\phi(s) - \psi(s)| d\Gamma(s)$$

and

$$|f(\phi, t)|^2 \leq K \left[1 + \int_{-\tau}^0 |\phi(s)|^2 d\Gamma(s) \right];$$

(A3) on the interval $[-\tau, 0]$, $x(t)$ is continuous with probability 1 with $E|x(\theta)|^4 < \infty$, $-\tau \leq \theta \leq 0$;

(A4) $h(\phi, t)$ is Borel measurable on $\mathcal{C} \times [0, T]$;

$$(A5) \quad \int_0^T E[h(x_t, t)h(x_t, t)] dt < \infty.$$

Under these assumptions, (2.1) and (2.2) can be shown ([10]–[12]) to have a solution which is continuous w.p.1 and has bounded second moment. Furthermore x_t is a Markov process.

Since linear stochastic delay systems admit a much more complete theory, we shall, for greater clarity in our exposition, use a different notation for such system. We

write

$$(2.3) \quad dx(t) = a(x_t, t) dt + F(t) dw(t),$$

$$(2.4) \quad dz(t) = c(x_t, t) dt + N(t) dv(t),$$

where $a(x_t, t)$ and $c(x_t, t)$ are given by the Lebesgue–Stieltjes integrals

$$a(x_t, t) = \int_{-\tau}^0 d_\theta A(t, \theta) x(t + \theta),$$

$$c(x_t, t) = \int_{-\tau}^0 d_\theta C(t, \theta) x(t + \theta).$$

Here $A(t, \theta)$ is a function on $R \times R$ jointly measurable in (t, θ) , continuous in t , of bounded variation in θ for each t , with $\text{Var}_{[-\tau, 0]} A(t, \cdot) \leq m(t)$, a locally integrable function on R^n . Furthermore $A(t, \theta) = 0$ for $\theta \geq 0$, $A(t, \theta) = A(t, -\tau)$ for $\theta \leq -\tau$, and it is continuous from the left in θ on $(-\tau, 0)$. The function $C(t, \theta)$ is assumed to satisfy similar conditions.

It is not difficult to show (see e.g. [6], [13]) that the linear stochastic delay system (2.3)–(2.4) has a unique solution (up to almost everywhere equivalence) given by the formula

$$(2.5) \quad x(t) = \Phi(t, 0)x_0(0) + \int_{-\tau}^0 d_\beta \left\{ \int_0^\tau \Phi(t, s) A(s, \beta - s) ds \right\} x_0(\beta) + \int_0^t \Phi(t, s) F(s) dw(s)$$

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

$$\dot{x}(t) = a(x_t, t)$$

(see e.g. [14]).

3. A representation theorem for conditional moment functionals. In filtering problems for stochastic differential systems, one is usually interested in estimating some function ϕ of the system process $x(t)$ given the observations $z(s)$, $0 \leq s \leq t$. It is well known that the optimal estimate with respect to a large class of criteria is the conditional expectation $E\{\phi(x(t))/z^t\}$ where z^t denotes the σ -algebra generated by the observations $z(s)$, $0 \leq s \leq t$. We shall also write $E\{\phi(x(t))/z^t\}$ as $E^t\{\phi[x(t)]\}$, and we shall omit the qualification of almost sure equivalence for conditional expectations. Fujisaki et al [15] have given a stochastic differential equation for the evolution of $E^t[\phi(x(t))]$ for rather general stochastic systems, which include our delay model. Specifically, they showed that for the delay system (2.1) and (2.2)

$$(3.1) \quad dE^t[\phi(x(t))] = E^t[\mathcal{L}_t \phi(x(t))] dt + \{E^t[\phi(x(t))h'(x_t, t)] - E^t[\phi(x(t))]E^t[h'(x_t, t)]\} \\ \cdot R^{-1}(t)[dz(t) - E^t(h(x_t, t)) dt]$$

where \mathcal{L}_t is a differential operator (see (4.7)). However, the right hand side of (3.1) contains terms of the form $E^t[g(x_t, t)]$, which we shall call conditional moment functionals. It is not clear how one can obtain an equation for these conditional moment functionals from (3.1). Equation (3.1), therefore, does not constitute a complete solution. It appears that the fundamental quantities we need to calculate are the conditional moment functionals $E^t[\phi(x_t)]$ (see also §§ 4 and 5). In this section, we will derive a representation theorem for $E^t[\phi(x_t)]$.

Our derivation is based on the work of Kunita [16]. Kunita obtained a representation theorem under the assumptions:

(i) The signal process is a stationary Markov process with compact state space, and the functional h is independent of time and continuous.

(ii) The functional $\phi: \mathcal{C} \rightarrow R$ is bounded.

We shall extend his results by making only the assumptions:

$$(A6) \quad E|\phi(x_t)|^2 < \infty,$$

$$(A7) \quad \int_0^T E|\phi(x_t)h(x_s, t)|^2 dt < \infty.$$

THEOREM 3.1. *Suppose (A1)–(A7) hold. Then we have the following representation for the conditional expectation of ϕ given z^t :*

$$(3.2) \quad E^t[\phi(x_t)] = E[\phi(x_t)] + \int_0^t E^s\{E[\phi(x_t)|x_s][h'(x_s, s) - E^s(h'(x_s, s))]\}R^{-1}(s) d\nu(s)$$

where $\nu(t) = z(t) - \int_0^t E^s[h(x_s, s)] ds$ is the innovations.

Proof. We follow the approach of Kunita [16]. First suppose that ϕ is bounded. Let \mathcal{G}^t denote the σ -algebra $\sigma\{x(s), v(s); s \leq t\}$. Clearly $z^t \subset \mathcal{G}^t$. Moreover, by the Markov property of x_t and the independence of the x and v processes, $E[\phi(x_t)|\mathcal{G}^s] = E[\phi(x_t)|x_s]$, for $t \geq s$.

By the assumptions of the theorem, all terms in (3.2) are in $L^2(\Omega, z^t, P)$. Thus, just as in [16], it is sufficient to verify that

$$(3.3) \quad E\{[E^t(\phi(x_t)) - E(\phi(x_t))]Y_t\} \\ = E\left\{\int_0^t E^s[E(\phi(x_t)|x_s)[h'(x_s, s) - E^s(h'(x_s, s))]]R^{-1}(s) d\nu(s)Y_t\right\}$$

for all Y_t represented as $\int_0^t g'_s d\nu(s)$, with g_s a jointly measurable and z^t -adapted process such that $\int_0^t E|g_s|^2 ds < \infty$.

Using the independence of the x and v processes, we conclude, on following the same argument as Kunita [16], that

$$(3.4) \quad E\left\{[E^t(\phi(x_t)) - E(\phi(x_t))]\int_0^t g'_s d\nu(s)\right\} \\ = \int_0^t E\{E(\phi(x_t)|x_s)[h'(x_s, s) - E^s(h'(x_s, s))]\} ds \\ = \int_0^t E\{E[\phi(x_t)|\mathcal{G}^s]g'_s[h'(x_s, s) - E^s(h'(x_s, s))]\} ds \\ = \int_0^t E\{E[\phi(x_t)|x_s]g'_s[h'(x_s, s) - E^s(h'(x_s, s))]\} ds.$$

Since

$$(3.5) \quad E\left\{\int_0^t E^s[E(\phi(x_t)|x_s)[h'(x_s, s) - E^s(h'(x_s, s))]]R^{-1}(s) d\nu(s) \int_0^t g'_s d\nu(s)\right\} \\ = \int_0^t E\{E[\phi(x_t)|x_s][h'(x_s, s) - E^s(h'(x_s, s))]\}g'_s ds$$

we obtain, on combining (3.4) and (3.5), the desired equation (3.3). Thus the theorem is true if ϕ is bounded.

In the general case, let $\phi_N(x_t) = \phi(x_t)\chi_N$, where $\chi_N = 1$ if $|\phi(x_t)| \leq N$, and $\chi_N = 0$ if $|\phi(x_t)| > N$. Clearly $E|\phi_N(x_t) - \phi(x_t)| \rightarrow 0$ as $N \rightarrow \infty$. Since ϕ_N is bounded, the above development yields

$$(3.6) \quad \begin{aligned} E'[\phi_N(x_t)] &= E[\phi_N(x_t)] \\ &+ \int_0^t E^s\{E[\phi_N(x_t)|x_s][h(x_s, s) - E^s(h(x_s, s))]\}R^{-1}(s) d\nu(s). \end{aligned}$$

By assumption (A7), the last term on the right hand side of (3.6) converges in probability to

$$\int_0^t E^s\{E[\phi(x_t)|x_s][h(x_s, s) - E^s(h(x_s, s))]\}R^{-1}(s) d\nu(s)$$

(see, for example, [17]). Hence, on letting $N \rightarrow \infty$ in (3.6), we finally obtain (3.2). The proof is completed.

The following corollary is immediate.

COROLLARY 3.1. *The smoothed estimate $E'[x(t+\theta)]$, $-\tau \leq \theta < 0$, is given by*

$$(3.7) \quad \begin{aligned} E'[x(t+\theta)] &= E^{t+\theta}[x(t+\theta)] \\ &+ \int_{t+\theta}^t E^s\{x(t+\theta)[h'(x_s, s) - E^s(h'(x_s, s))]\}R^{-1}(s) d\nu(s). \end{aligned}$$

Remark 3.1. Theorem 3.1 remains true if we merely assume that the signal process is a Markov process with state space a separable complete metric space. The same proof goes through for this more general case. Theorem 3.1 corresponds to the special situation where the signal process is the Markov process x_t generated by the stochastic delay equation (2.1). Similar remarks also apply to Theorem 4.1 in the next section on stochastic differential equations for the nonlinear filtering problem.

4. Stochastic differential equations for nonlinear filtering of delay systems. While Theorem 3.1 gives an abstract representation for the optimal estimates, it is completely nonrecursive in the sense that knowledge of $E[\phi(x_t)/z^t]$ is of no use in determining $E[\phi(x_{t+\Delta})/z^{t+\Delta}]$. In fact, for every t , we must completely reprocess our past observations. For implementation and approximation purposes, one would like to obtain a stochastic differential equation for the evolution of $E[\phi(x_t)/z^t]$. In this section, we shall give conditions on ϕ under which we can obtain a stochastic differential equation for $E[\phi(x_t)/z^t]$. As we shall see, these conditions are intimately related to the (extended) infinitesimal generator of the Markov process x_t [15].

DEFINITION. A family of linear operators A_t , $t \in [0, T]$ defined on the space of real-valued measurable functions on \mathcal{C} is called an (extended) infinitesimal generator if

$$(4.1) \quad E[\phi(x_t)/x_s] - \phi(x_s) = \int_s^t E[A_u \phi(x_u)/x_s] du$$

is satisfied for all $0 \leq s < t \leq T$. We use the notation $D(A)$ to denote the space of all functionals ϕ satisfying $E|\phi(x_t)|^2 < \infty$, $\int_0^T E|A_t \phi(x_t)|^2 dt < \infty$, and (4.1).

Define the process $e_h(t) = h(x_t, t) - E'[h(x_t, t)]$. Then we have

THEOREM 4.1. *Let the conditions of Theorem 3.1 be satisfied. In addition, let ϕ belong to $D(A)$ and suppose that*

$$\int_0^T E|[A_t \phi(x_t)]h(x_t, t)|^2 dt < \infty.$$

Then the functional $E[\phi(x_t)|z^t]$ satisfies the stochastic differential equation

$$(4.2) \quad dE[\phi(x_t)|z^t] = E[A_t\phi(x_t)/z^t] dt + E[\phi(x_t)e'_h(t)/z^t]R^{-1}(t) d\nu(t).$$

Proof. For any $t \in [0, T]$ and $\varepsilon > 0$, we have, by a simple calculation,

$$(4.3) \quad \begin{aligned} & E[\phi(x_{t+\varepsilon})/z^{t+\varepsilon}] - E[\phi(x_t)/z^t] \\ &= E[\phi(x_{t+\varepsilon}) - \phi(x_t)] + \int_0^t E\{E[\phi(x_{t+\varepsilon}) - \phi(x_t)/x_s]e'_h(s)/z^s\}R^{-1}(s) d\nu(s) \\ & \quad + \int_t^{t+\varepsilon} E\{E[\phi(x_{t+\varepsilon})/x_s]e'_h(s)/z^s\}R^{-1}(s) d\nu(s). \end{aligned}$$

Using (4.1), we get that

$$(4.4) \quad \begin{aligned} & \int_0^t E\{E[\phi(x_{t+\varepsilon}) - \phi(x_t)/x_s]e'_h(s)/z^s\}R^{-1}(s) d\nu(s) \\ &= \int_0^t E\left\{E\left\{\int_t^{t+\varepsilon} E[A_u\phi(x_u)/x_t] du/x_s\right\}e'_h(s)/z^s\right\}R^{-1}(s) d\nu(s) \\ &= \int_0^t E\left\{\int_t^{t+\varepsilon} E[A_u\phi(x_u)/x_s]e'_h(s)/z^s\right\}R^{-1}(s) du d\nu(s) \\ &= \int_t^{t+\varepsilon} \int_0^t E\{E[A_u\phi(x_u)/x_s]e'_h(s)/z^s\}R^{-1}(s) d\nu(s) du \end{aligned}$$

with the last equality justified in view of the assumptions of the theorem. Similarly

$$(4.5) \quad \begin{aligned} & \int_t^{t+\varepsilon} E\{E[\phi(x_{t+\varepsilon})/x_s]e'_h(s)/z^s\}R^{-1}(s) d\nu(s) \\ &= \int_t^{t+\varepsilon} E[\phi(x_s)e'_h(s)/z^s]R^{-1}(s) d\nu(s) \\ & \quad + \int_t^{t+\varepsilon} \int_s^{t+\varepsilon} E\{E[A_u\phi(x_u)/x_s]e'_h(s)/z^s\} du R^{-1}(s) d\nu(s). \end{aligned}$$

Finally, using the representation theorem for $A_t\phi(x_t)$, we find that

$$(4.6) \quad \begin{aligned} & E[\phi(x_{t+\varepsilon}) - \phi(x_t)] = \int_t^{t+\varepsilon} E[A_u\phi(x_u)] du \\ &= \int_t^{t+\varepsilon} E[A_u\phi(x_u)/z^u] du \\ & \quad - \int_t^{t+\varepsilon} \int_0^u E\{E[A_u\phi(x_u)/x_s]e'_h(s)/z^s\}R^{-1}(s) d\nu(s) du. \end{aligned}$$

Adding up (4.4) to (4.6) yields

$$\begin{aligned} & E[\phi(x_{t+\varepsilon})/z^{t+\varepsilon}] - E[\phi(x_t)/z^t] \\ &= \int_t^{t+\varepsilon} E[A_u\phi(x_u)/z^u] du + \int_t^{t+\varepsilon} E[\phi(x_u)e'_h(u)/z^u]R^{-1}(u) d\nu(u) \end{aligned}$$

which is precisely (4.2).

Remark 4.1. Theorem 4.1 is a generalization to systems with delays of the usual formula for conditional moments of ordinary diffusion processes. While the form of the stochastic differential equation is exactly the same as that for diffusion processes, here we need to know the structure of the infinitesimal generator of the x_t process. We know from (4.1) that functionals of the form $\phi[x(t+\theta)]$, $\theta \in (-\tau, 0)$ do not belong to $D(A)$, since $x(t)$ is not in general differentiable. Hence it is not possible to derive a stochastic differential equation for a functional of the form $\phi[x(t+\theta)]$. Indeed, a complete characterization of the operators A_t is not known, although certain special classes of functionals which are in the domain of A_t have been stated in Kushner [12]. We mention these results to illustrate Theorem 4.1.

Case 1. Suppose the functional $\phi(x_t) = \phi[x(t)]$, and is twice continuously differentiable in its argument. Then

$$(4.7) \quad A_t \phi[x(t)] \equiv \mathcal{L}_t \phi[x(t)] = f(x_t, t)' \phi_x[x(t)] + \frac{1}{2} \text{tr } Q(t) \phi_{xx}[x(t)]$$

where ϕ_x is the n -vector whose i th component is $(\partial \phi / \partial x_i)[x(t)]$. In particular

$$(4.8) \quad dE'[x(t)] = E'[f(x_t, t)] dt + \{E'[x(t)h'(x_t, t)] - E'[x(t)]E'[h'(x_t, t)]\}R^{-1}(t) d\nu(t)$$

In this case, (4.2) reduces to the well-known results of Fujisaki et al. [15].

Case 2. Let $\phi(x_t) = \int_{-\tau}^0 \psi(\theta)g[x(t+\theta), x(t)] d\theta$, where ψ is continuously differentiable on $[-\tau, 0]$, and g is twice continuously differentiable in its second argument. Then

$$(4.9) \quad \begin{aligned} A_t \phi(x_t) = & \psi(0)g[x(t), x(t)] - \psi(-\tau)g[x(t-\tau), x(t)] \\ & - \int_{-\tau}^0 \dot{\psi}(\theta)g[x(t+\theta), x(t)] d\theta + \int_{-\tau}^0 \psi(\theta)\mathcal{L}_t g[x(t+\theta), x(t)] d\theta \end{aligned}$$

where \mathcal{L}_t is the operator defined in Case 1 and acts on g as a function of $x(t)$ only.

Case 3. Let $\phi(x_t) = D[F(x_t)]$ where D is a twice continuously differentiable real-valued function, and $F(x_t) = \int_{-\tau}^0 \psi(\theta)g[x(t+\theta), x(t)] d\theta$ is the type of functional described in Case 2. Then

$$A_t \phi(x_t) = D_\alpha(\alpha)|_{\alpha=F(x_t)} A_t F(x_t) + \frac{1}{2} D_{\alpha\alpha}(\alpha)|_{\alpha=F(x_t)} \cdot G$$

where

$$G = \int_{-\tau}^0 \int_{-\tau}^0 \psi(\theta)\psi(\eta) \sum_{i,j} g_{\beta_i}[x(t+\theta), x(t)] g_{\beta_j}[x(t+\theta), x(t)] Q_{ij}(t) d\theta d\eta$$

and g_{β_i} denotes partial differentiation of g with respect to the i th component of the second argument.

From the above special cases, we can see that basically we need twice continuous differentiability of ϕ with respect to the dependence on $x(t)$, and Fréchet differentiability with respect to the dependence on the piece of the trajectory x_t . As discussed before, this rules out functionals of the form $\phi[x(t+\theta)]$, $\theta \in [-\tau, 0)$. Hence for nonlinear systems with point delays, any attempt in deriving stochastic differential equations for conditional moment functionals will have to face the difficulty of functionals not being in the domain of the generator of the Markov process x_t . For example, if the observation process is of the form

$$dz(t) = \{h_1[x(t)] + h_2[x(t-\tau)]\} dt + d\nu(t)$$

$\phi[x(t)]h_2[x(t-\tau)]$ will not be in the domain of A_t . On the other hand, in order to

analyze (4.2), we do need to calculate the conditional expectation for

$$\phi[x(t)]h_2[x(t-\tau)]$$

Of course, there are many physical problems (for example, radar problems with spread targets [23]) where the observations are of the form $h(x_i) = \int_{-\tau}^0 \psi(\theta) H[x(t+\theta), x(t)] d\theta$. Moreover, one can approximate point delays by distributed delays of the above form. This will allow us to write a stochastic differential equation for $\phi[x(t)]h(x_i)$. However, we will then get the unknown $A_i \phi[x(t)]h(x_i)$ in our equation for $\phi[x(t)]h(x_i)$. If $\psi(-\tau) \neq 0$, $A_i \phi[x(t)]h(x_i)$ will contain a term with point delay (see Case 2 above), and we are faced with the same problems as before. In general, if the functionals involved are in the domain of A_i^i , $i = 1, \dots, n$, we can write n coupled stochastic differential equations involving the moment functionals, just as in the diffusion process case. It should be clear from the above discussion that this puts rather severe restrictions on the functionals involved.

There is, however, one special case where the optimal filter can be completely specified even when there are point delays in the system. This is the linear case with Gaussian distributions and will be treated next.

5. Optimal filtering of linear stochastic delay systems. Consider the linear stochastic delay system defined by

$$(5.1) \quad \begin{aligned} dx(t) &= a(x_t, t) dt + F(t) dw(t), \\ x(\theta) &= x_0(\theta), \quad \theta \in [-\tau, 0]; \end{aligned}$$

$$(5.2) \quad dz(t) = c(x_t, t) dt + N(t) dv(t)$$

where $a(\cdot, \cdot)$ and $c(\cdot, \cdot)$ are as described in § 2. Also, we take x_0 to be a Gaussian process on $[-\tau, 0]$ with mean $\bar{x}_0(\theta)$ and cov $[x_0(\theta); x_0(\xi)] = \Sigma_0(\theta, \xi)$. By an argument similar to the case without delays, it is readily seen that the conditional distribution of $x(t+\theta)$, for any $\theta \in [-\tau, 0]$, given $z(s)$, $0 \leq s \leq t$, is Gaussian. We shall write $\hat{x}(t+\theta/t)$ to denote $E\{x(t+\theta)/z^t\}$, $\theta \in [-\tau, 0]$. Using (4.8), we immediately obtain the following stochastic differential equation for the conditional mean

$$(5.3) \quad \begin{aligned} d\hat{x}(t/t) &= \int_{-\tau}^0 d_\theta A(t, \theta) \hat{x}(t+\theta/t) dt \\ &+ \left[\int_{-\tau}^0 E^t(x(t)x(t+\theta)') d_\theta C'(t, \theta) \right. \\ &\quad \left. - \int_{-\tau}^0 \hat{x}(t/t) \hat{x}(t+\theta/t)' d_\theta C'(t, \theta) \right] R^{-1}(t) dv(t) \end{aligned}$$

with the innovations $\nu(t)$ given by

$$\nu(t) = z(t) - \int_0^t \int_{-\tau}^0 d_\theta C(s, \theta) \hat{x}(s+\theta/s) ds.$$

Define the "smoothed" conditional error covariance as

$$P(t, \theta, \xi) = E^t\{[x(t+\theta) - \hat{x}(t+\theta/t)][x(t+\xi) - \hat{x}(t+\xi/t)]'\}, \quad -\tau \leq \theta, \quad \xi \leq 0.$$

Then (5.3) can be rewritten as

$$(5.4) \quad d\hat{x}(t/t) = \int_{-\tau}^0 d_\theta A(t, \theta) \hat{x}(t+\theta/t) dt + \int_{-\tau}^0 P(t, 0, \theta) d_\theta C'(t, \theta) R^{-1}(t) dv(t).$$

To evaluate the unknown terms on the right hand side of (5.4), we use (3.16) to write the smoothed estimate as

$$(5.5) \quad \hat{x}(t+\theta/t) = \hat{x}(t+\theta/t+\theta) + \int_{t+\theta}^t \int_{-\tau}^0 P(s, t+\theta-s, \xi) d_{\xi} C(s, \xi)' R^{-1}(s) d\nu(s) \\ \text{for } \theta \in [-\tau, 0].$$

An inspection of (5.4) and (5.5) shows that the optimal linear filter is completely characterized by $\hat{x}(t+\theta/t)$, $-\tau \leq \theta \leq 0$, and the "smoothed" error covariance function $P(t, \theta, \xi)$. It remains only to derive appropriate equations for $P(t, \theta, \xi)$. Since the processes x and z are jointly Gaussian, the error process $x(t+\theta) - \hat{x}(t+\theta|t)$ is independent of $z(s)$, $s \leq t$ (see, for example, Bagchi [8]). Hence $P(t, \theta, \xi)$ is independent of the observations and equals $E\{[x(t+\theta) - \hat{x}(t+\theta|t)][x(t+\xi) - \hat{x}(t+\xi|t)]'\}$. This fact will enable us to simplify the derivations of the equations for $P(t, \theta, \xi)$. The next theorem summarizes the complete structure of the optimal filter.

THEOREM 5.1. *The optimal filter for the system (5.1)–(5.2) is characterized as follows:*

- (i) *The conditional mean $\hat{x}(t/t)$ satisfies (5.4).*
- (ii) *The smoothed estimate $\hat{x}(t+\theta/t)$ satisfies (5.5).*
- (iii) *The smoothed error covariance $P(t, \theta, \xi)$ satisfies the equations*

$$(5.6) \quad \frac{d}{dt} P(t, 0, 0) = \int_{-\tau}^0 P(t, 0, \theta) d_{\theta} A'(t, \theta) + \int_{-\tau}^0 d_{\theta} A(t, \theta) P(t, \theta, 0) \\ - \int_{-\tau}^0 \int_{-\tau}^0 P(t, 0, \theta) d_{\theta} C'(t, \theta) R^{-1}(t) d_{\xi} C(t, \xi) P(t, \xi, 0) + Q(t),$$

$$(5.7) \quad \sqrt{2} P_{\eta}(t, \theta, 0) = \int_{-\tau}^0 P(t, \theta, \xi) d_{\xi} A'(t, \xi) - \int_{-\tau}^0 \int_{-\tau}^0 P(t, \theta, \xi) d_{\xi} C'(t, \xi) R^{-1}(t) \\ \cdot d_{\alpha} C(t, \alpha) P(t, \alpha, 0),$$

$$(5.8) \quad \sqrt{3} P_{\sigma}(t, \theta, \xi) = - \int_{-\tau}^0 \int_{-\tau}^0 P(t, \theta, \beta) d_{\beta} C'(t, \beta) R^{-1}(t) d_{\alpha} C(t, \alpha) P(t, \alpha, \xi)$$

where η is the unit vector in the $(1, -1, 0)$ direction, σ the unit vector in the $(1, -1, -1)$ direction, and $P_{\eta}(t, \theta, 0)$ and $P_{\sigma}(t, \theta, \xi)$ are the directional derivatives of $P(t, \theta, 0)$ and $P(t, \theta, \xi)$ in the directions η and σ respectively. The initial conditions are given by

$$\hat{x}(\theta/0) = \bar{x}_0(\theta), \quad \theta \in [-\tau, 0], \\ P(0, \theta, \xi) = \Sigma_0(\theta, \xi), \quad -\tau \leq \theta, \quad \xi \leq 0.$$

Proof. See Appendix A.

Remark 5.1. Equations similar to those of (5.6)–(5.8) for $P(t, \theta, \xi)$ were formally derived by Kwakernaak in [5], and rigorously rederived by Bagchi in [8], for systems with point delays only. Instead of directional derivatives, they used partial derivations with respect to the variables θ and ξ . In the general case, however, $P(t, \theta, \xi)$ will not be continuously differentiable in (t, θ, ξ) . This is why directional derivatives have to be used. A similar situation has already been noted in the quadratic optimal control problem for linear delay systems [20].

In the special case where $x_0 \equiv 0$, $a(x_t, t) = Ax(t) + Bx(t-\tau)$, $c(x_t, t) = Cx(t)$, Q, R constant matrices, it can be shown directly or by exploiting the connection between linear optimal filtering and optimal control with quadratic criterion (see § 6) that

$P(t, \theta, \xi)$ is in fact continuously differentiable in (t, θ, ξ) . When we compare the solutions to the linear optimal control and optimal linear filtering problems in our study of filter stability, it will be helpful to use the notation $P_0(t) = P(t, 0, 0)$, $P_1(t, \theta) = P(t, \theta, 0)$, and $P_2(t, \theta, \xi) = P(t, \theta, \xi)$. In this case, the optimal filter is given by the equations

$$(5.9) \quad d\hat{x}(t/t) = A\hat{x}(t/t) dt + B\hat{x}(t-\tau/t) dt + P_0(t)C'R^{-1}[dz(t) - C\hat{x}(t/t) dt];$$

$$(5.10) \quad \hat{x}(t-\tau/t) = \hat{x}(t-\tau/t-\tau) + \int_{t-\tau}^t P_1(s, t-\tau-s)C'R^{-1}[dz(s) - \hat{C}x(s/s) ds],$$

$$\hat{x}(\theta/0) = 0, \quad -\tau \leq \theta \leq 0;$$

$$(5.11) \quad \frac{d}{dt}P_0(t) = AP_0(t) + P_0(t)A' - P_0(t)C'R^{-1}CP_0(t) + Q + BP_1(t, -\tau) + P_1'(t, -\tau)B';$$

$$(5.12) \quad \left(\frac{\partial}{\partial t} - \frac{\partial}{\partial \theta}\right)P_1(t, \theta) = P_1(t, \theta)[A' - C'R^{-1}CP_0(t)] + P_2(t, \theta, -\tau)B';$$

$$(5.13) \quad \left(\frac{\partial}{\partial t} - \frac{\partial}{\partial \theta} - \frac{\partial}{\partial \xi}\right)P_2(t, \theta, \xi) = -P_1(t, \theta)C'R^{-1}CP_1(t, \xi);$$

with

$$(5.14) \quad \begin{aligned} P_0(0) &= P_1(0, \theta) = P_2(0, \theta, \xi) = 0, \\ P_1(t, 0) &= P_0(t), \quad P_2(t, \theta, 0) = P_1(t, \theta), \\ P_0(t) &= P_0'(t), \quad P_2(t, \theta, \xi) = P_2'(t, \xi, \theta). \end{aligned}$$

Notice that in this special case where there are no delays in the observations, $\hat{x}(t/t)$ depends only on $\hat{x}(s/s)$, $t-\tau \leq s \leq t$, and from (5.9) and (5.10), we can obtain the following explicit stochastic delay equation for $\hat{x}(t/t)$:

$$(5.15) \quad \begin{aligned} d\hat{x}(t/t) &= [A\hat{x}(t/t) + B\hat{x}(t-\tau/t-\tau)] dt + P_0(t)C'R^{-1}[dz(t) - C\hat{x}(t/t) dt] \\ &\quad + \int_{t-\tau}^t BP_1(s, t-\tau-s)C'R^{-1}[dz(s) - C\hat{x}(s/s) ds] dt, \\ \hat{x}(\theta/0) &= 0, \quad -\tau \leq \theta \leq 0. \end{aligned}$$

This will not be true if we have delays in the observations (see the discussions in [18] and [24]).

6. Stability of linear optimal filters and control systems. In this section, we study the stability of optimal linear filters and stochastic control systems for linear delay systems. We shall concentrate on the filters defined by (5.9)–(5.13). The extension to the case with multiple delays in the system dynamics is straightforward. However, the situation for systems with delays in the observations is much more complicated and will be treated separately in a forthcoming paper. In our analysis, we make essential use of the duality between optimal filtering of linear stochastic delay systems and optimal control of linear delay systems with quadratic cost. These results complete the extension of the well-known linear quadratic Gaussian theory to systems with delays in the dynamics.

We begin by summarizing the results for the optimal control of linear delay systems with quadratic cost [1]–[4]. Consider the system

$$(6.1) \quad \begin{aligned} \frac{dx}{dt} &= Ax(t) + Bx(t-\tau) + Cu(t), \\ x(\theta) &= x_0(\theta), \quad \theta \in [-\tau, 0]. \end{aligned}$$

In previous sections, we have used the space \mathcal{C} as our state space. For this system, however, we may allow the initial function x_0 to lie in the larger space $R^n \times L^2$ (see [4] or [20]). The admissible control set U is the set of R^m -valued L_2 functions on $[0, T]$. The cost functional is given by

$$J_T(u, x_0) = \int_0^T [x'(t)Mx(t) + u'(t)Su(t)] dt$$

where M and S are symmetric matrices of appropriate dimensions, $M \geq 0$, $S > 0$. When $T < \infty$, the optimal control is given by

$$(6.2) \quad u^*(t) = -S^{-1}C'K_0(t)x(t) - S^{-1}C' \int_{-\tau}^0 K_1(t, \theta)x(t+\theta) d\theta.$$

The feedback gains satisfy the following coupled set of partial differential equations:

$$(6.3) \quad \frac{d}{dt}K_0(t) = -A'K_0(t) - K_0(t)A + K_0(t)CS^{-1}C'K_0(t) - M - K_1(t, 0) - K_1'(t, 0),$$

$$(6.4) \quad \left(\frac{\partial}{\partial t} - \frac{\partial}{\partial \theta} \right) K_1(t, \theta) = -[A' - K_0(t)CS^{-1}C']K_1(t, \theta) - K_2(t, 0, \theta),$$

$$(6.5) \quad \left(\frac{\partial}{\partial t} - \frac{\partial}{\partial \theta} - \frac{\partial}{\partial \xi} \right) K_2(t, \theta, \xi) = K_1'(t, \theta)CS^{-1}C'K_1(t, \xi),$$

with

$$(6.6) \quad \begin{aligned} K_0(T) &= K_1(T, \theta) = K_2(T, \theta, \xi) = 0 \quad -\tau \leq \theta, \quad \xi \leq 0, \\ K_1(t, -\tau) &= K_0(t)B, \\ K_2(t, -\tau, \theta) &= B'K_1(t, \theta), \\ K_0(t) &= K_0'(t), \quad K_2(t, \theta, \xi) = K_2(t, \xi, \theta)'. \end{aligned}$$

The optimal cost can be expressed as

$$(6.7) \quad \begin{aligned} J_T^*(x_0) &= x_0'(0)K_0(0)x_0(0) + \int_{-\tau}^0 x_0'(0)K_1(0, \theta)x_0(\theta) d\theta \\ &\quad + \int_{-\tau}^0 x_0'(\theta)K_1'(0, \theta)x_0(0) d\theta + \int_{-\tau}^0 \int_{-\tau}^0 x_0'(\theta)K_2(0, \theta, \xi)x_0(\xi) d\theta d\xi. \end{aligned}$$

We now consider the infinite time control problem, i.e., $T = \infty$. To discuss this problem, we need some condition to ensure that the optimal cost will be finite. The relevant concepts are those of stabilizability and detectability. These definitions for the case of delay systems are given below.

DEFINITION 6.1. The system

$$(6.8) \quad \dot{x}(t) = Ax(t) + Bx(t-\tau) + Cu(t)$$

is said to be *stabilizable* if there exist matrices L_0 , L_1 , and $L_2(\theta)$, $\theta \in [-\tau, 0]$, with L_2 strongly measurable and bounded, such that

$$(6.9) \quad \dot{x}(t) = (A + CL_0)x(t) + (B + CL_1)x(t - \tau) + \int_{-\tau}^0 CL_2(\theta)x(t + \theta) d\theta$$

is asymptotically stable. We then also say that (A, B, C) is stabilizable.

DEFINITION 6.2. The system

$$(6.10) \quad \begin{aligned} \dot{x}(t) &= Ax(t) + Bx(t - \tau), \\ z(t) &= Cx(t) \end{aligned}$$

is said to be *detectable* if there exist matrices K_0 , K_1 , and $K_2(\theta)$, $\theta \in [-\tau, 0]$, with K_2 strongly measurable and bounded, such that

$$(6.11) \quad \dot{x}(t) = Ax(t) + Bx(t - \tau) + K_0z(t) + K_1z(t - \tau) + \int_{-\tau}^0 K_2(\theta)z(t + \theta) d\theta$$

is asymptotically stable. We then also say (A, B, C) is detectable.

The following proposition can be easily proved from the above definitions [18].

PROPOSITION 6.1. *The system (6.10) is detectable if and only if the "adjoint" system (which runs backwards in time)*

$$(6.12) \quad \dot{y}(t) = -A'y(t) - B'y(t + \tau) - C'u(t)$$

is stabilizable.

The properties of stabilizability and detectability, and their relationships to controllability and observability, are further discussed in [18], [25], [26], to which the reader is referred.

We can now state the result concerning the infinite time quadratic control problem. Let $M = H'H$.

PROPOSITION 6.2 ([19], [20], [21]). *Assume that (A, B, C) is stabilizable and (A, B, H) is detectable. Then the gains $K_0(t)$, $K_1(t, \theta)$, and $K_2(t, \theta, \xi)$, for each fixed $t < T$, converge to K_0 , $K_1(\theta)$ and $K_2(\theta, \xi)$ respectively as $T \rightarrow \infty$ in the following sense:*

$$\begin{aligned} K_1(t, \cdot) &\rightarrow K_1(\cdot) \quad \text{strongly in } L_2[-\tau, 0], \\ K_2(t, \cdot, \cdot) &\rightarrow K_2(\cdot, \cdot) \quad \text{strongly in } L_2[-\tau, 0] \times L_2[-\tau, 0]. \end{aligned}$$

The optimal control law for the infinite time problem is given by

$$(6.13) \quad u^*(t) = -S^{-1}C'K_0x(t) - \int_{-\tau}^0 S^{-1}C'K_1(\theta)x(t + \theta) d\theta$$

where K_0 , $K_1(\theta)$ and $K_2(\theta, \xi)$ satisfy the following set of equations

$$(6.14) \quad A'K_0 + K_0A - K_0CS^{-1}C'K_0 + M + K_1'(0) + K_1(0) = 0,$$

$$(6.15) \quad \frac{d}{d\theta}K_1(\theta) = [A' - K_0CS^{-1}C']K_1(\theta) + K_2(0, \theta),$$

$$(6.16) \quad \left(\frac{\partial}{\partial \theta} + \frac{\partial}{\partial \xi} \right) K_2(\theta, \xi) = -K_1'(\theta)CS^{-1}C'K_1(\xi),$$

with

$$(6.17) \quad \begin{aligned} K_1(-\tau) &= K_0B, & K_2(\theta, -\tau) &= K_1'(\theta)B, \\ K_0 &= K_0', & K_2(\theta, \xi) &= K_2'(\xi, \theta). \end{aligned}$$

Furthermore, the optimal closed-loop system is asymptotically stable with the optimal

cost given by

$$\begin{aligned}
 J_{\infty}^*(x_0) = & x_0'(0)K_0x_0(0) + \int_{-\tau}^0 x_0'(0)K_1(\theta)x_0(\theta) d\theta \\
 (6.18) \quad & + \int_{-\tau}^0 x_0'(\theta)K_1'(\theta)x_0(0) d\theta + \int_{-\tau}^0 \int_{-\tau}^0 x_0'(\theta)K_2(\theta, \xi)x_0(\xi) d\theta d\xi.
 \end{aligned}$$

Remark 6.1. Proposition 6.2 is an extension of the result of [19] and [20] where the matrix M is assumed to be positive definite. One of the authors first proved in [18] that the condition $M > 0$ can be relaxed to (A, B, H) observable. Subsequently, the work of Zabczyk [21] became known to the authors and the present conclusions, assuming the still weaker condition of detectability, can be obtained from the results of [21].

To connect the optimal control result of Proposition 6.2 with those of optimal filtering, we need the following duality theorem which can be deduced from the work of Lindquist [13].

PROPOSITION 6.3. *Consider the optimal filtering problem over the interval $[0, T]$ for the system*

$$\begin{aligned}
 (6.19) \quad dx(t) = & [Ax(t) + Bx(t - \tau)] dt + F dw(t), \\
 & x(\theta) = 0, \quad \theta \leq 0;
 \end{aligned}$$

$$(6.20) \quad dz(t) = Cx(t) dt + N dv(t).$$

Define the dual control system by

$$(6.21) \quad \dot{y}(t) = -A'y(t) - B'y(t + \tau) - C'u(t)$$

with

$$(6.22) \quad y(T) = b, \quad y(s) = 0, \quad s > T.$$

The dual control problem is defined to be to minimize

$$(6.23) \quad J_T(b, u) = \int_0^T [y'(t)Qy(t) + u'(t)Ru(t)] dt$$

where $Q = FF' \geq 0$ and $R = NN' > 0$. Let the optimal linear least squares estimate of $x(T)$ be $\hat{x}(T/T)$, and let the optimal control for the dual problem be u_T . Then $b'\hat{x}(T/T)$ is related to u_T by

$$(6.24) \quad b'\hat{x}(T/T) = - \int_0^T u_T'(s) dz(s).$$

We now have two representations of $b'\hat{x}(T/T)$, one directly from (5.9)–(5.13), the other indirectly from (6.24). Our strategy is to compare the two representations and identify the control and filter gains appropriately. This will enable us to exploit the known results of the optimal control problem to conclude filter stability. We begin by stating the following lemma.

LEMMA 6.1. *The conditional mean of $x(T)$, denoted by $\hat{x}(T/T)$, is given either by*

$$\begin{aligned}
 (6.25) \quad \hat{x}(T/T) = & \int_0^T \left\{ \Phi(T, s)P_0(s)C'R^{-1} \right. \\
 & \left. + \int_0^{\min(\tau, T-s)} \Phi(T, s+\theta)BP_1(s, \theta-\tau)C'R^{-1} d\theta \right\} dz(s)
 \end{aligned}$$

or by

$$(6.26) \quad \hat{x}(T/T) = \int_0^T \left\{ Y'(t, T) \tilde{K}_0(t) C' R^{-1} + \int_0^{\min(\tau, T-t)} Y(t+\theta, T) \tilde{K}_1(t, \theta) C' R^{-1} d\theta \right\} dz(t).$$

Here $P_0(t)$, $P_1(t, \theta)$, and $P_2(t, \theta, \xi)$ are given by (5.11)–(5.14) and $\Phi(t, s)$ is the fundamental matrix [14] associated with the delay equation

$$(6.27) \quad \dot{x}(t) = [A - P_0(t) C' R^{-1} C] x(t) + Bx(t - \tau) - B \int_{-\tau}^0 P_1(t + \theta, -\theta - \tau) C' R^{-1} C x(t + \theta) d\theta.$$

The functions $\tilde{K}_0(t)$, $\tilde{K}_1(t, \theta)$ and $\tilde{K}_2(t, \theta, \xi)$ satisfy the equations

$$(6.28) \quad \dot{\tilde{K}}_0(t) = A \tilde{K}_0(t) + \tilde{K}_0(t) A' - \tilde{K}_0(t) C' R^{-1} C \tilde{K}_0(t) + Q + \tilde{K}_1(t, 0) + \tilde{K}_1'(t, 0),$$

$$(6.29) \quad \left(\frac{\partial}{\partial t} - \frac{\partial}{\partial \theta} \right) \tilde{K}_1(t, \theta) = [A - \tilde{K}_0(t) C' R^{-1} C] \tilde{K}_1(t, \theta) + \tilde{K}_2(t, 0, \theta),$$

$$(6.30) \quad \left(\frac{\partial}{\partial t} - \frac{\partial}{\partial \theta} - \frac{\partial}{\partial \xi} \right) \tilde{K}_2(t, \theta, \xi) = -\tilde{K}_1'(t, \theta) C' R^{-1} C \tilde{K}_1(t, \xi)$$

with

$$(6.31) \quad \begin{aligned} \tilde{K}_0(0) &= \tilde{K}_1(0, \theta) = \tilde{K}_2(0, \theta, \xi) = 0, \quad 0 \leq \theta, \quad \xi \leq \tau, \\ \tilde{K}_1(t, \tau) &= \tilde{K}_0(t) B', \\ \tilde{K}_2(t, \tau, \xi) &= B \tilde{K}_1(t, \xi), \\ \tilde{K}_0(t) &= \tilde{K}_0'(t), \quad \tilde{K}_2(t, \theta, \xi) = \tilde{K}_2'(t, \xi, \theta) \end{aligned}$$

and $Y(t, s)$ is the fundamental matrix associated with the system

$$(6.32) \quad \frac{dy}{dt} = -[A' - C' R^{-1} C \tilde{K}_0(t)] y(t) - B' y(t + \tau) + \int_0^\tau C' R^{-1} C \tilde{K}_1(t, \theta) y(t + \theta) d\theta.$$

Proof. This simply involves solving for $\hat{x}(T/T)$ explicitly from (5.9)–(5.13) and from (6.24) and making appropriate changes of variables. For details, the reader may consult [18].

Next, we relate the various quantities involved in (6.25) and (6.26) in

LEMMA 6.2. *The optimal filter gains (5.11)–(5.14) are related to the optimal control gains (6.28)–(6.31) for the dual problem by*

$$(6.33) \quad P_0(t) = \tilde{K}_0(t),$$

$$(6.34) \quad P_1'(t, \theta - \tau) B' = \tilde{K}_1(t, \theta),$$

$$(6.35) \quad B P_2(t, \theta - \tau, \xi - \tau) B' = \tilde{K}_2(t, \theta, \xi), \quad 0 \leq t \leq T, \quad 0 \leq \theta, \quad \xi \leq \tau.$$

The systems (6.27) and (6.32) are adjoints of each other [14] so that

$$(6.36) \quad \Phi(t, s) = Y'(s, t).$$

Proof. For proving (6.33)–(6.35), we simply verify that they satisfy the same equations and boundary conditions. By uniqueness of the optimal control and optimal filter, we conclude that (6.33)–(6.35) hold. Substituting these results into (6.32), we see that $Y(t, s)$ is the fundamental matrix of

$$(6.37) \quad \dot{y}(t) = -[A' - C'R^{-1}CP_0(t)]y(t) - B'y(t + \tau) + \int_{-\tau}^0 C'R^{-1}CP'_1(t, -\theta - \tau)B'y(t - \theta) d\theta.$$

But (6.37) is precisely the adjoint equation [14] to (6.27), and it is well-known [14] that $\Phi(t, s) = Y'(s, t)$.

We are now ready to prove asymptotic stability of the optimal filter.

THEOREM 6.1. *Consider the system defined by (6.19)–(6.20). Suppose (A, B, C) is detectable and (A, B, F) is stabilizable. Then the gains of the optimal filter defined by (5.7)–(5.11) converge, and the steady state optimal filter is asymptotically stable.*

Proof. Proposition 6.1 shows that the dual system (A', B', C') defined by (6.21) is stabilizable and (A', B', F') is detectable. Proposition 6.2 then shows that the gains $\tilde{K}_0(t)$, $\tilde{K}_1(t, \theta)$, $\tilde{K}_2(t, \theta, \xi)$ for the dual control problem, as given by (6.28)–(6.31) converge to \tilde{K}_0 , $\tilde{K}_1(\theta)$, $\tilde{K}_2(\theta, \xi)$ respectively as $t \rightarrow \infty$. By Lemma 6.2, we conclude that as $t \rightarrow \infty$, $P_0(t) \rightarrow P_0$, $BP_1(t, \theta) \rightarrow BP_1(\theta)$, and $BP_2(t, \theta, \xi)B' \rightarrow BP_2(\theta, \xi)B'$, where

$$(6.38) \quad AP_0 + P_0A' - P_0C'R^{-1}CP_0 + Q + BP_1(-\tau) + P'_1(-\tau)B' = 0,$$

$$(6.39) \quad \frac{d}{d\theta}BP_1(\theta) = -BP_1(\theta)[A' - C'R^{-1}CP_0] - BP_2(\theta, -\tau)B',$$

$$(6.40) \quad \left(\frac{\partial}{\partial\theta} + \frac{\partial}{\partial\xi}\right)BP_2(\theta, \xi)B' = BP_1(\theta)C'R^{-1}CP'_1(\xi)B'$$

with

$$(6.41) \quad \begin{aligned} P_1(0) &= P_0, & P_2(\theta, 0) &= P_1(\theta), \\ P_0 &= P'_0, & P_2(\theta, \xi) &= P'_2(\xi, \theta). \end{aligned}$$

In view of (5.9) and (5.10), stability of the steady state filter is then governed by the stability of the equation

$$(6.42) \quad \frac{d}{dt}x(t) = [A - P_0C'R^{-1}C]x(t) + Bx(t - \tau) - B \int_{-\tau}^0 P_1(-\theta - \tau)C'R^{-1}Cx(t + \theta) d\theta$$

But the adjoint to (6.42) is given by

$$(6.43) \quad \dot{y}(t) = -[A' - C'R^{-1}CP_0]y(t) - B'y(t + \tau) + \int_0^\tau C'R^{-1}CP'_1(\theta - \tau)B'y(t + \theta) d\theta.$$

By Lemma 6.2 again, this corresponds to the closed-loop optimal system for the dual control problem. Proposition 6.2 then shows that (6.43) is asymptotically stable. Hence, the system defined by (6.42), being the adjoint of that of (6.43), is asymptotically stable as well.

Remark 6.2. Theorem 6.1 is not the most general form of the filter stability result for delay systems. Generalization to cases where we can have delays in the observations, random initial conditions, etc., will be treated in a forthcoming paper (see also [24]).

Remark 6.3. Vinter [9] has independently obtained a similar filter stability result using infinite dimensional filtering methods quite different from ours. In addition to the conclusion given in Theorem 6.1, he also proved that $\|\tilde{T}_{t,0}\| \rightarrow 0$ as $t \rightarrow \infty$, where $\tilde{T}_{t,s}$ is the evolution operator connected with the error process for the *time-varying* filter (5.9)–(5.14). His arguments can be readily adapted to our setting to prove the same result.

7. Stochastic control of linear delay systems. We can now combine the results for optimal control with quadratic cost and optimal linear filtering to obtain a stochastic control scheme which is asymptotically stable. To that end, we define the stochastic control problem as that of minimizing the cost functional

$$(7.1) \quad J_T(u, x_0) = E \int_0^T [x'(t)Mx(t) + u'(t)Su(t)] dt$$

for u in some set of admissible control laws, subject to the constraint

$$(7.2) \quad dx(t) = [Ax(t) + Bx(t - \tau)] dt + Gu(t) dt + F dw(t),$$

$$x(\theta) = x_0(\theta), \quad \theta \in [-\tau, 0],$$

$$(7.3) \quad dz(t) = Cx(t) dt + N dv(t).$$

We can evidently write

$$(7.4) \quad z(t) = z_0(t) + \int_0^t \int_0^s \Phi(s, \sigma) Gu(\sigma) d\sigma ds.$$

Define the set U_0 consisting of the class of processes $u(t)$ satisfying the following conditions:

- (i) $u(t)$ is measurable with respect to $\sigma\{z(s), 0 \leq s \leq t\}$, i.e., there is a Borel measurable function Π such that $u(t) = \Pi(t; z(s), 0 \leq s \leq t)$.
- (ii) For each $u \in U_0$, the feedback system, obtained by using $\Pi(t; z(s), 0 \leq s \leq t)$ for $u(t)$ in (7.2) and (7.3), has a unique solution.
- (iii) $\int_0^T E|u(t)|^2 dt < \infty$.
- (iv) For each $u \in U_0$, $\sigma\{z(s), 0 \leq s \leq t\} = \sigma\{z_0(s), 0 \leq s \leq t\}$.

We shall take U_0 to be the set of admissible controls. For a discussion on this choice, see [18], [22].

The following result has been proved by Lindquist [6].

PROPOSITION 7.1. *The problem of determining $u \in U_0$ so as to minimize (7.1) has the following solution*

$$(7.5) \quad u^*(t) = -S^{-1}G'K_0(t)\hat{x}(t|t) - S^{-1}G' \int_{-\tau}^0 K_1(t, \theta)\hat{x}(t + \theta|t) d\theta$$

where $K_0(t)$ and $K_1(t, \theta)$ are the optimal gains for the deterministic optimal control problem and are given by (6.3)–(6.6), and $\hat{x}(s|t)$, $t - \tau \leq s \leq t$, is the conditional expectation of $x(s)$ given $z(\sigma)$, $0 \leq \sigma \leq t$.

We now give the expression for the optimal cost, obtained in [18].

LEMMA 7.1. *Corresponding to the optimal control (7.5), the optimal cost associated*

with the stochastic control problem (7.1)–(7.4) is given by

$$\begin{aligned}
 J^* = & EV(x_0) + \int_0^T \text{tr } FF'K_0(t) dt \\
 & + \int_0^T \text{tr} \left\{ K_0(t)GS^{-1}G'K_0(t)P_0(t) \right. \\
 (7.6) \quad & + \int_{-\tau}^0 K_1'(t, \theta)GS^{-1}G'K_0(t)P_1'(t, \theta) d\theta \\
 & + \int_{-\tau}^0 K_0(t)GS^{-1}G'K_1(t, \theta)P_1(t, \theta) d\theta \\
 & \left. + \int_{-\tau}^0 \int_{-\tau}^0 K_1'(t, \theta)GS^{-1}G'K_1(t, \xi)P_2(t, \xi, \theta) d\theta d\xi \right\} dt
 \end{aligned}$$

where

$$\begin{aligned}
 V(x_t) = & x'(t)K_0(t)x(t) + \int_{-\tau}^0 x'(t)K_1(t, \theta)x(t+\theta) d\theta \\
 (7.7) \quad & + \int_{-\tau}^0 x'(t+\theta)K_1'(t, \theta)x(t) d\theta + \int_{-\tau}^0 \int_{-\tau}^0 x'(t+\theta)K_2(t, \theta, \xi)x(t+\xi) d\theta d\xi.
 \end{aligned}$$

Proof. See Appendix B.

We turn our attention now to the stochastic control system defined by using the steady state version of (7.5). The behavior of the closed-loop system under this law is summarized in the following theorem.

THEOREM 7.1. *Let $M = H'H$. Suppose (A, B, G) and (A, B, F) are stabilizable, and (A, B, C) and (A, B, H) are detectable. Then the control law*

$$(7.8) \quad u(t) = -S^{-1}G'K_0\hat{x}(t|t) - S^{-1}G' \int_{-\tau}^0 K_1(\theta)\hat{x}(t+\theta|t) d\theta$$

where $\hat{x}(t+\theta|t)$, $-\tau \leq \theta \leq 0$, is generated by the steady state filter of Theorem 6.1, and $K_0, K_1(\theta)$ are given by the deterministic stationary control law of Proposition 6.2, gives rise to an asymptotically stable closed-loop system. Furthermore, the cost “rate”

$$(7.9) \quad J_r = \lim_{T \rightarrow \infty} \frac{1}{T} E \left\{ \int_0^T [x'(t)Mx(t) + u'(t)Su(t)] dt \right\}$$

associated with the above law is given by

$$\begin{aligned}
 J_r = & \text{tr } FF'K_0 + \text{tr} \left\{ K_0GS^{-1}G'K_0P_0 \right. \\
 (7.10) \quad & + \int_{-\tau}^0 K_1'(\theta)GS^{-1}G'K_0P_1'(\theta) d\theta \\
 & + \int_{-\tau}^0 K_0GS^{-1}G'K_1(\theta)P_1(\theta) d\theta \\
 & \left. + \int_{-\tau}^0 \int_{-\tau}^0 K_1'(\theta)GS^{-1}G'K_1(\xi)P_2(\xi, \theta) d\theta d\xi \right\}.
 \end{aligned}$$

Proof. Stabilizability of (A, B, G) and detectability of (A, B, H) ensure that K_0 , $K_1(\theta)$, $K_2(\theta, \xi)$ are well defined and that the solutions of the system

$$(7.11) \quad \dot{x}(t) = (A - GS^{-1}G'K_0)x(t) + Bx(t - \tau) - \int_{-\tau}^0 GS^{-1}G'K_1(\theta)x(t + \theta) d\theta$$

are asymptotically stable (see Proposition 6.2). Detectability of (A, B, C) and stabilizability of (A, B, F) guarantee that the steady state filter is well defined and asymptotically stable (see Theorem 6.1). The closed-loop system is defined by the coupled set of equations

$$(7.12) \quad \begin{aligned} dx(t) &= [Ax(t) + Bx(t - \tau)] dt - GS^{-1}G'K_0\hat{x}(t|t) dt \\ &\quad - \int_{-\tau}^0 GS^{-1}G'K_1(\theta)\hat{x}(t + \theta|t) d\theta dt + F dw(t), \\ d\hat{x}(t|t) &= A\hat{x}(t|t) dt + B\hat{x}(t - \tau|t - \tau) dt \\ &\quad + P_0C'R^{-1}[dz(t) - C\hat{x}(t|t) dt] \\ (7.13) \quad &\quad + B \int_{t-\tau}^t P_1(t-s-\tau)C'R^{-1}[dz(s) - C\hat{x}(s|s) ds] dt. \end{aligned}$$

Let the estimation error $e(t + \theta/t)$, $-\tau \leq \theta \leq 0$, be defined as $e(t + \theta/t) = x(t + \theta) - \hat{x}(t + \theta/t)$. We then get

$$(7.14) \quad \begin{aligned} dx(t) &= (A - GS^{-1}G'K_0)x(t) dt + Bx(t - \tau) dt \\ &\quad - \int_{-\tau}^0 GS^{-1}G'K_1(\theta)x(t + \theta) d\theta dt + F dw(t) \\ &\quad + GS^{-1}G'K_0e(t|t) dt + \int_{-\tau}^0 GS^{-1}G'K_1(\theta)e(t + \theta|t) d\theta dt \end{aligned}$$

and

$$(7.15) \quad \begin{aligned} de(t|t) &= (A - P_0C'R^{-1}C)e(t|t) dt + Be(t - \tau|t - \tau) dt \\ &\quad + \int_{-\tau}^0 BP_1(-\theta - \tau)C'R^{-1}Ce(t + \theta|t + \theta) d\theta dt \\ &\quad + F dw(t) - P_0C'R^{-1} dv(t) - B \int_{t-\tau}^t P_1(t-s-\tau)C'R^{-1} dv(s) dt. \end{aligned}$$

Since (7.15) is decoupled from (7.14), the stability properties of the closed-loop system are precisely those of (7.11) and the steady state optimal filter. Since both of these are asymptotically stable as a consequence of our assumptions, the closed-loop stochastic control system is asymptotically stable as well. The expression for J_r follows readily from Lemma 7.1.

8. Concluding remarks. We have treated the problem of filtering and control for stochastic delay systems. The general filtering problem is studied for both linear and nonlinear stochastic delay systems. A representation theorem for conditional moment functionals is given, which forms the basis for derivations of stochastic differential equations describing the optimal linear or nonlinear filter. For linear systems with Gaussian initial conditions and noises, the optimal filter is completely specified by the

equations derived for the conditional mean and covariance functions. The linear time-invariant case with delays in the system dynamics is investigated in detail, with particular emphasis on the stability of the optimal filter and stochastic control system. These results, together with those on deterministic optimal control or linear delay systems with quadratic cost, give a rather complete linear-quadratic-Gaussian theory for this class of delay systems. In a forthcoming paper, we will extend this theory to systems with delays in the control and delays in the observations.

Appendix A.

Proof of Theorem 5.1. It is only necessary to derive the equations for $P(t, \theta, \xi)$. For systems with point delays only, Bagchi [8] derived the equations for $P(t, \theta, \xi)$ using properties of the innovations and martingale theory. Although our approach is different, we shall use some of his results to simplify our derivations (a more complicated proof was given in [18]).

It is easy to see that

$$(A.1) \quad P(t, \theta, \xi) = E[x(t + \theta)x'(t + \xi)] - E[\hat{x}(t + \theta|t)\hat{x}(t + \xi|t)'].$$

Using (5.5), we obtain

$$(A.2) \quad \begin{aligned} & E[\hat{x}(t + \theta|t)\hat{x}(t + \xi|t)'] \\ &= E\left\{ \left[\hat{x}(t + \theta|t + \theta) + \int_{t+\theta}^t \int_{-\tau}^0 P(s, t + \theta - s, \beta) d_{\beta} C(s, \beta)' R^{-1}(s) d\nu(s) \right] \right. \\ & \quad \cdot \left. \left[\hat{x}(t + \xi|t + \xi) + \int_{t+\xi}^t \int_{-\tau}^0 P(\sigma, t + \xi - \sigma, \alpha) d_{\alpha} C(\sigma, \alpha)' R^{-1}(\sigma) d\nu(\sigma) \right] \right\}. \end{aligned}$$

For any ε such that $-\tau \leq \theta + \varepsilon \leq 0$, $-\tau \leq \xi + \varepsilon \leq 0$, we get, using (A.1) and (A.2), that

$$(A.3) \quad \begin{aligned} & P(t, \theta, \xi) - P(t - \varepsilon, \theta + \varepsilon, \xi + \varepsilon) \\ &= -E\left\{ \left[\int_{t-\varepsilon}^t \int_{-\tau}^0 P(s, t + \theta - s, \beta) d_{\beta} C(s, \beta)' R^{-1}(s) d\nu(s) \right] \hat{x}(t + \xi|t + \xi)' \right\} \\ & \quad - E\left\{ \hat{x}(t + \theta|t + \theta) \left[\int_{t-\varepsilon}^t \int_{-\tau}^0 P(\sigma, t + \xi - \sigma, \alpha) d_{\alpha} C(\sigma, \alpha)' R^{-1}(\sigma) d\nu(\sigma) \right]' \right\} \\ & \quad - E\left\{ \left[\int_{t+\theta}^t \int_{-\tau}^0 P(s, t + \theta - s, \beta) d_{\beta} C(s, \beta)' R^{-1}(s) d\nu(s) \right] \right. \\ & \quad \cdot \left. \left[\int_{t+\xi}^t \int_{-\tau}^0 P(\sigma, t + \xi - \sigma, \alpha) d_{\alpha} C(\sigma, \alpha)' R^{-1}(\sigma) d\nu(\sigma) \right]' \right\} \\ & \quad + E\left\{ \left[\int_{t+\theta}^{t-\varepsilon} \int_{-\tau}^0 P(s, t + \theta - s, \beta) d_{\beta} C(s, \beta)' R^{-1}(s) d\nu(s) \right] \right. \\ & \quad \cdot \left. \left[\int_{t+\xi}^{t-\varepsilon} \int_{-\tau}^0 P(\sigma, t + \xi - \sigma, \alpha) d_{\alpha} C(\sigma, \alpha)' R^{-1}(\sigma) d\nu(\sigma) \right]' \right\}. \end{aligned}$$

Using the fact that $E[\nu(t) - \nu(s)|z^s] = 0$ as in [8], we see that the first two terms in (A.3) vanish. By the same argument, the last two terms can be easily simplified to yield

$$\begin{aligned}
& P(t, \theta, \xi) - P(t - \xi, \theta + \varepsilon, \xi + \varepsilon) \\
&= -E \left\{ \left[\int_{t-\varepsilon}^t \int_{-\tau}^0 P(s, t + \theta - s, \beta) d_{\beta} C(s, \beta)' R^{-1}(s) d\nu(s) \right] \right. \\
&\quad \cdot \left. \left[\int_{t-\varepsilon}^t \int_{-\tau}^0 P(\sigma, t + \xi - \sigma, \alpha) d_{\alpha} C(\sigma, \alpha)' R^{-1}(\sigma) d\nu(\sigma) \right]' \right\} \\
&= - \int_{t-\varepsilon}^t \int_{-\tau}^0 P(s, t + \theta - s, \beta) d_{\beta} C(s, \beta)' R^{-1}(s) \int_{-\tau}^0 d_{\alpha} C(s, \alpha) P(s, \alpha, t + \xi - s) ds.
\end{aligned}
\tag{A.4}$$

Since $P(t, \theta, \xi)$ is clearly continuous in (t, θ, ξ) , we may divide (A.4) throughout by ε and let ε go to 0. This gives (5.8). The same arguments apply to the derivations of (5.6) and (5.7). Finally, the initial conditions follow immediately from the properties of conditional expectations.

Appendix B.

Proof of Lemma 7.1. We apply the Itô differential rule to the function $V(x_t)$ defined in (7.7). We calculate the first and second terms to illustrate the computations involved:

$$\begin{aligned}
d[x'(t)K_0(t)x(t)] &= [dx'(t)]K_0(t)x(t) \\
&\quad + x'(t)[dK_0(t)]x(t) + x'(t)K_0(t)[dx(t)] + \text{tr } FF'K_0(t) dt \\
&= x'(t - \tau)B'K_0(t)x(t) dt + u'(t)G'K_0(t)x(t) dt \\
&\quad + dw'(t)F'K_0(t)x(t) dt + x'(t)K_0(t)Bx(t - \tau) dt \\
&\quad + x'(t)K_0(t)Gu(t) dt + x'(t)K_0(t)F dw(t) - x'(t)Mx(t) dt \\
&\quad + x'(t)K_0(t)GR^{-1}G'K_0(t)x(t) dt - x'(t)K_1'(t, 0)x(t) dt \\
&\quad - x'(t)K_1(t, 0)x(t) dt + \text{tr } FF'K_0(t) dt, \\
d_t \left[\int_{-\tau}^0 x'(t)K_1(t, \theta)x(t + \theta) d\theta \right] \\
&= d_t \left[x'(t) \int_{t-\tau}^t K_1(t, \sigma - t)x(\sigma) d\sigma \right] \\
&= \{[x'(t)A' + x'(t - \tau)B' + u'(t)G'] dt + dw'(t)F'\} \int_{-\tau}^0 K_1(t, \theta)x(t + \theta) d\theta \\
&\quad + x'(t)K_1(t, 0)x(t) dt - x'(t)K_1(t, -\tau)x(t - \tau) dt \\
&\quad + x'(t) \int_{t-\tau}^t d_t K_1(t, \sigma - t)x(\sigma) d\sigma dt \\
&= [x'(t)A' + x'(t - \tau)B' + u'(t)G'] \int_{-\tau}^0 K_1(t, \theta)x(t + \theta) d\theta dt \\
&\quad + dw'(t)F' \int_{-\tau}^0 K_1(t, \theta)x(t + \theta) d\theta + x'(t)K_1(t, 0)x(t) dt \\
&\quad - x'(t)K_1(t, -\tau)x(t - \tau) dt + x'(t) \int_{-\tau}^0 \left[\frac{\partial}{\partial t} - \frac{\partial}{\partial \theta} \right] K_1(t, \theta) x(t + \theta) d\theta dt.
\end{aligned}$$

Similar calculations for the last two terms on the right hand side of (7.7) yield the following expression

$$\begin{aligned}
 (B.1) \quad dV(x_t) = & V_1(t) dt + dw'(t)F'K_0(t)x(t) dt + x'(t)K_0(t)F dw(t) \\
 & + \text{tr } FF'K_0(t) dt + dw'(t)F' \int_{-\tau}^0 K_1(t, \theta)x(t+\theta) d\theta \\
 & + \int_{-\tau}^0 x'(t+\theta)K_1'(t, \theta) d\theta F dw(t) - x'(t)Mx(t) - u'(t)Su(t) dt
 \end{aligned}$$

where

$$\begin{aligned}
 (B.2) \quad V_1(t) = & \left[u(t) + S^{-1}G'K_0(t)x(t) + \int_{-\tau}^0 S^{-1}G'K_1(t, \theta)x(t+\theta) d\theta \right]' \\
 & \cdot S \left[u(t) + S^{-1}G'K_0(t)x(t) + \int_{-\tau}^0 S^{-1}G'K_1(t, \xi)x(t+\xi) d\xi \right].
 \end{aligned}$$

Using the boundary conditions at T for $K_0(t)$, $K_1(t, \theta)$ and $K_2(t, \theta, \xi)$, we see that $V(x_T) = 0$. Therefore, integrating (B.1) from 0 to T and taking expectations, we get

$$(B.3) \quad E \int_0^T [x'(t)Mx(t) + u'(t)Su(t)] dt = EV(x_0) + E \int_0^T V_1(t) dt + \int_0^T \text{tr } FF'K_0(t) dt.$$

Now

$$E \int_0^T V_1(t) dt = \int_0^T EV_1(t) dt = \int_0^T E\{E[V_1(t)/z']\} dt$$

by the use of Fubini's theorem and properties of conditional expectations. Substituting the control law in (7.5) into (B.2), we get that

$$\begin{aligned}
 (B.4) \quad E[V_1(t)/z'] = & \text{tr} \left\{ K_0(t)G_1R^{-1}G'K_0(t)P_0(t) \right. \\
 & + \int_{-\tau}^0 K_1'(t, \theta)GR^{-1}G'K_0(t)P_1'(t, \theta) d\theta \\
 & + \int_{-\tau}^0 K_0(t)GR^{-1}G'K_1(t, \theta)P_1(t, \theta) d\theta \\
 & \left. + \int_{-\tau}^0 \int_{-\tau}^0 K_1'(t, \theta)GR^{-1}G'K_1(t, \xi)P_2(t, \theta, \xi) d\theta d\xi \right\}.
 \end{aligned}$$

$E\{V_1(t)/z'\}$ is now seen to be a deterministic function and hence equal to $EV_1(t)$. Substituting (B.4) into (B.3) yields the conclusion of the lemma.

Acknowledgments. We would like to thank the referees for drawing our attention to the work of Kunita [16] and Bagchi [8], and for helpful suggestions which led to general improvements in the presentation. The above references, in particular, enabled us to reduce the complicated proofs of Theorem 3.1 and 5.1 given in [18] to the simple forms here in this paper.

REFERENCES

- [1] H. J. KUSHNER AND D. I. BARNEA, *On the control of a linear functional differential equation with quadratic cost*, this Journal, 8 (1970), pp. 257–272.
- [2] Y. ALEKAL, P. BRUNOVSKY, D. H. CHYUNG AND E. B. LEE, *The quadratic problem for systems with time delays*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 673–688.
- [3] A. MANITIUS, *Optimal control of time-lag systems with quadratic performance indexes*, IV Congress of International Federation of Automatic Control, Warsaw, 1969, paper 13.2.
- [4] M. C. DELFOUR AND S. K. MITTER, *Controllability, observability and optimal feedback control of affine hereditary differential systems*, this Journal, 10 (1972), pp. 298–328.
- [5] H. KWAKERNAK, *Optimal filtering in linear systems with time delays*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 169–173.
- [6] A. LINDQUIST, *Optimal control of linear stochastic systems with applications to time lag systems*, Information Sci., 5 (1973), pp. 81–126.
- [7] S. K. MITTER AND R. B. VINTER, *Filtering for linear stochastic hereditary differential systems*, Intern. Symp. Control Theory, Numerical Methods, and Computer Systems Modelling, Institut de Recherche d'Informatique et d'Automatique, Rocquencourt, France, June 1974.
- [8] A. BAGCHI, *A martingale approach to state estimation in delay-differential systems*, J. Math. Anal. Appl., 56 (1976), pp. 195–210.
- [9] R. B. VINTER, *Filter stability for stochastic evolution equations*, this Journal, 15 (1977), pp. 465–485.
- [10] K. ITO AND M. NISIO, *On stationary solutions of a stochastic differential equation*, Kyoto J. Math., 4 (1964), pp. 1–75.
- [11] W. H. FLEMING AND M. NISIO, *On the existence of optimal stochastic controls*, J. Math. Mech., 15 (1966), pp. 777–794.
- [12] H. J. KUSHNER, *On the stability of processes defined by stochastic differential-difference equations*, J. Differential Equations, 4 (1968), pp. 424–443.
- [13] 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.
- [14] J. K. HALE, *Functional Differential Equations*, Springer-Verlag, New York, 1971.
- [15] M. FUJISAKI, G. KALLIANPUR AND H. KUNITA, *Stochastic differential equations for the nonlinear filtering problem*, Osaka J. Math., 9 (1972), pp. 19–40.
- [16] H. KUNITA, *Asymptotic behavior of the nonlinear filtering errors of Markov processes*, J. Multivariate Anal., 1 (1971), pp. 365–393.
- [17] I. I. GIKHMAN AND A. V. SKOROKHOD, *Introduction to the Theory of Random Processes*, W. B. Saunders, London, 1969.
- [18] R. H. KWONG, *Structural properties and estimation of delay systems*, Electronic Systems Lab. rep. ESL-R-614, Massachusetts Inst. of Tech., Cambridge, MA, Sept. 1975.
- [19] R. DATKO, *Unconstrained control problems with quadratic cost*, this Journal, 11 (1973), pp. 32–52.
- [20] M. C. DELFOUR, C. MCCALLA AND S. K. MITTER, *Stability and the infinite-time quadratic cost problem for linear hereditary differential systems*, this Journal, 13 (1975), pp. 48–88.
- [21] J. ZABCZYK, *Remarks on the algebraic Riccati equation in Hilbert space*, Appl. Math. and Optimization, 2 (1976), pp. 251–258.
- [22] A. LINDQUIST, *On feedback control of linear stochastic systems*, this Journal, 11 (1973), pp. 323–343.
- [23] H. L. VAN TREES, *Detection, Estimation, and Modulation Theory, Part III*, John Wiley, New York, 1971.
- [24] R. H. KWONG, *The linear quadratic Gaussian problem for systems with delays in the state, control and observations*, Proc. 14th Allerton Conf. Circuit and System Theory, Univ. of Illinois, Sept. 29–Oct. 1, 1976, pp. 545–549.
- [25] A. MANITIUS AND R. TRIGGIANI, *Function space controllability of linear retarded systems: a derivation from abstract operator conditions*, Centre de Recherches Mathématiques rep. CRM-605, Université de Montréal, Montréal, Québec, Canada, March 1976.
- [26] A. MANITIUS, *Controllability, observability and stabilizability of retarded systems*, Proc. IEEE Conf. Decision and Control (Clearwater, FL, Dec. 1976), IEEE, New York, 1977, pp. 752–758.

ON THE POLYHEDRALITY OF THE CONVEX HULL OF THE FEASIBLE SET OF AN INTEGER PROGRAM*

R. R. MEYER[†] AND M. L. WAGE[‡]

Abstract. Polyhedrality is established for convex hulls of sets defined by systems of *equations* in nonnegative integer variables. This property is useful for certain existence, duality, and sensitivity results in integer programming. The structural theorems obtained also shed light on the relationship between the convex hull and the relaxation obtained by deleting integrality constraints.

1. Introduction. A number of results dealing with existence [7], duality [1], and sensitivity analysis [2] for integer programming have been established for integer programs whose feasible sets have convex hulls that are polyhedral (i.e., the intersection of a finite number of closed half-spaces). This is due to the fact that, given a set $S \subseteq R^n$ and a linear function cx , if the convex hull of S (denoted $\text{conv } S$) is polyhedral, then the problem $\sup cx$ subject to $x \in S$ has the same optimal value as the *linear program* $\max cx$ subject to $x \in \text{conv } S$ (including the infeasible case in which the optimal value is set to $-\infty$, and the unbounded case in which the optimal value is taken as $+\infty$), and, moreover, every *optimal extreme point* of the linear program is an optimal solution of the problem over S . In this report, polyhedrality is established for the convex hull of an arbitrary set S of the form

$$(1) \quad S \equiv \{x | Ax = b, x \geq 0, x \text{ integer}\},^1 \quad \text{where } x = (x_1, \dots, x_n)^T \in R^n,$$

A is a given $m \times n$ matrix of *real* numbers, and b is a given element of R^m . While it might be thought that the polyhedrality of the convex hull of the feasible set of an integer program could be taken for granted, it has been shown that in the *inequality* constrained case, the convex hull may be quite complex [4], [5], [12] and, in fact, need not be polyhedral [3]. In the case of *rational* coefficients (for both inequalities and equations), polyhedrality was previously proved in [7]. Here we will show that this rationality hypothesis is *not* required in the equality-constrained case.

2. A rational representation. In this section it will be shown that the set S defined by (1) always has an *equivalent* representation as $\{x | A'x = b', x \geq 0, x \text{ integer}\}$, where A' and b' are *rational*. Once this result is established, polyhedrality of $\text{conv } S$ follows directly from Theorem 3.9 of [7]. (However, a more compact and geometrically motivated proof is possible due to absence of the continuous variables allowed in [7], and this alternative method of proof is given in § 3.)

Theorem 1 employs the concept of *rational independence*: a set of real numbers $\{\gamma_1, \dots, \gamma_k\}$ is said to be *rationally independent* if $\gamma_1 r_1 + \dots + \gamma_k r_k = 0$, where r_1, \dots, r_k are rational, implies $r_1 = \dots = r_k = 0$. Rational independence of a set of n -vectors is similarly defined. (It is easily seen that rational independence and integral independence, i.e., independence with respect to integer weights, are equivalent, but

* Received by the editors May 21, 1976, and in revised form October 24, 1977.

[†] Department of Computer Sciences, University of Wisconsin—Madison, Madison, Wisconsin 53706. The work of this author was supported in part by the United States Army under Contract DAAG29-75-C-0024 and in part by the National Science Foundation under Contract DCR74-20584.

[‡] Institute for Medicine and Mathematics, Ohio University, Athens, Ohio, and Mathematics Department, University of Wisconsin—Madison, Madison, Wisconsin. The work of this author was supported in part by the National Science Foundation under Contract MCS74-08550. Now at Department of Mathematics, Yale University, New Haven, Connecticut 06520.

¹ A vector or matrix is termed integer or rational if all its elements are respectively integer or rational.

due to the mechanics of the proofs to follow, rational independence is more convenient to work with. Note that while linear independence clearly implies rational independence, the converse is not true.)

THEOREM 1. *Let $S_e = \{x | Ax = b, x \text{ integer}\}$. Then there exists an $m' \times n$ matrix A' of rationals and a vector b' of rationals such that*

$$S_e = \{x | A'x = b', x \text{ integer}\}.$$

Proof. If $S_e = \emptyset$, then we may take $m' = 1$, $A = 0$, $b = 1$ and achieve the required rational representation, so we may assume $S_e \neq \emptyset$. We will first consider the case in which the system $Ax = b$ consists of a *single* equation, since a similar analysis performed *equation-by-equation* will yield the desired result in the general case. Denote the single equation by

$$(2) \quad \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n = \beta.$$

If all $\alpha_i = 0$, then feasibility implies $\beta = 0$, so no transformation is needed. Thus, we may assume that not all α_i are 0, and, for notational convenience, we also assume that the variables have been ordered so that $\alpha_1 \neq 0$ (in dealing with a system of equations, a different ordering might be required for each equation, but this causes no problems). If $n = 1$ then $\beta = \alpha_1 x_1^*$ for some integer x_1^* , and the data may be rationalized by dividing through by α_1 .

If $n \geq 1$, we replace coefficients by rational combinations of “previous”, “independent” coefficients whenever possible. Thus, if $\alpha_2 = \alpha_1 r_{1,2}$, where $r_{1,2}$ is rational, we rewrite (2) in the form

$$(3) \quad \alpha_1(x_1 + r_{1,2}x_2) + \cdots + \alpha_n x_n = \beta.$$

Continuing this procedure, we end up with an index set $I \subseteq \{1, \dots, n\}$ such that (2) is equivalent to

$$(4) \quad \sum_{i \in I} \left(\alpha_i \sum_{j=1}^n r_{i,j} x_j \right) = \beta,$$

where the $r_{i,j}$ are *rational*, and the α_i for $i \in I$ are rationally independent. Since $S_e \neq \emptyset$, $\sum_{i=1}^n \alpha_i x_i^* = \beta$ for some integers x_1^*, \dots, x_n^* , and by carrying out the same conversion as above we have

$$(5) \quad \sum_{i \in I} \left(\alpha_i \sum_{j=1}^n r_{i,j} x_j^* \right) = \beta.$$

By subtracting (5) from (4) we have

$$(6) \quad \sum_{i \in I} \left(\alpha_i \sum_{j=1}^n r_{i,j} (x_j - x_j^*) \right) = 0.$$

We will now show that

$$(7) \quad S_e = \left\{ x \mid \sum_{j=1}^n r_{i,j} x_j = \sum_{j=1}^n r_{i,j} x_j^* \text{ for } i \in I, x \text{ integer} \right\}.$$

Clearly, if x is integer and satisfies the equations of (7), then $x \in S_e$; so suppose that $x \in S_e$, but that $\sum_{j=1}^n r_{i,j} x_j \neq \sum_{j=1}^n r_{i,j} x_j^*$ for at least one i .

From (6) it would follow that the α_i with $i \in I$ were *not* rationally independent, which is a contradiction of the way in which they were constructed. Thus in the single equation case, the set S_e has an equivalent representation of the form (7).

When $Ax = b$ consists of more than one equation, an analogous procedure may be performed for *each* equation in the system, so that S_e may be represented in terms of the collection of the corresponding systems of the form (7). Alternatively, we could express A and b in terms of a "basis" of rationally independent columns of A , and carry out a proof analogous to that of the scalar case. \square

Example. Let $S_e \equiv \{x | -1 \cdot x_1 - \frac{1}{2}\sqrt{2}x_2 + \frac{4}{3} \cdot x_3 = \frac{5}{3} - \sqrt{2}, x \text{ integer}\}$. -1 and $-\frac{1}{2}\sqrt{2}$ are rationally independent, but $\frac{4}{3} = -1 \cdot -\frac{4}{3} - \frac{1}{2}\sqrt{2} \cdot 0$, so the equation in S_e may be written as $-1 \cdot (x_1 - \frac{4}{3}x_3) - \frac{1}{2}\sqrt{2} \cdot (x_2) = \frac{5}{3} - \sqrt{2}$. Since setting $x_1 = 1, x_2 = 2, x_3 = 2$, yields a point in S_e , we may write $\frac{5}{3} - \sqrt{2} = -1 \cdot (1) - \frac{1}{2}\sqrt{2} \cdot (2) + \frac{4}{3} \cdot (2)$. Substituting for the original RHS (right-hand side), we can write the equation in S_e as $-1 \cdot (x_1 - \frac{4}{3}x_3) - \frac{1}{2}\sqrt{2} \cdot (x_2) = -1 \cdot (1) - \frac{1}{2}\sqrt{2} \cdot (2) + \frac{4}{3} \cdot (2) = -1 \cdot (-\frac{5}{3}) - \frac{1}{2}\sqrt{2} \cdot (2)$. From the preceding theorem we conclude that S_e may be written as $\{x | x_1 - \frac{4}{3}x_3 = -\frac{5}{3}, x_2 = 2, x \text{ integer}\}$.

An example may be found in [9, p. 5] of the application of the conversion procedure in a 2-constraint case, along with a discussion of the difference between doing the conversion equation-by-equation and doing it for the entire system by working with the columns of A .

Note that if there exists a number γ such that all of the quotients $\alpha_1/\gamma, \dots, \alpha_n/\gamma, \beta/\gamma$ are rational, then (assuming $\alpha_1 \neq 0$) $(\alpha_i/\gamma)/(\alpha_1/\gamma) = \alpha_i/\alpha_1$ is rational for $j = 2, \dots, n$, so the conversion procedure in Theorem 1 yields $I = \{1\}$, and, in fact, that procedure is simply equivalent to dividing through by γ . (In fact, if there exists a γ such that the quotients $\alpha_1/\gamma, \dots, \alpha_n/\gamma$ are all rational, then there exists a γ' such that $\alpha_1/\gamma', \dots, \alpha_n/\gamma'$ are all integer, and in this case integrality of β/γ' is clearly a *necessary* condition for the existence of an integer solution. Under this divisibility assumption, a *necessary and sufficient* condition for the existence of an integer solution is that the "generalized greatest common divisor" (see [8]) of $\alpha_1, \dots, \alpha_n$ "divide" β in the sense of giving an integer quotient.) However, as the numerical example found in [9, p. 5] shows, the coefficients need not have this divisibility property, and in such a case $I \neq \{1\}$ and a single equation will be converted into an equivalent system of equations.

Corollary 1 below gives two additional results easily obtained from Theorem 1 and an analysis of its proof.

COROLLARY 1. *If $S_e = \{x | Ax = b, x \text{ integer}\}$ and $S_e^R = \{x | Ax = b, x \text{ rational}\}$, then there exist integer \hat{A} and \hat{b} such that $S_e = \{x | \hat{A}x = \hat{b}, x \text{ integer}\}$ and $S_e^R = \{x | \hat{A}x = \hat{b}, x \text{ rational}\}$.*

Proof. By use of Theorem 1, S_e may be written as $\{x | A'x = b', x \text{ integer}\}$, where A' and b' are rational, and by multiplying each equation of $A'x = b'$ by a suitable integer, conversion to integer data is achieved. Inspection of the proof of Theorem 1 shows that all of the steps go through if the x_i are assumed rational rather than integer. \square

In § 3 the rational representation of S is used to prove polyhedrality of $\text{conv } S$; however, it might also be noted that other useful structural properties (see [6]) can also be derived from the rational representation.

3. Structural properties. In establishing the polyhedrality of $\text{conv } S$, we may assume by Theorem 1 that S is represented in the form

$$(8) \quad S = \{x | A'x = b', x \geq 0, x \text{ integer}\},$$

where A' and b' are rational.

In order to state structural properties of $\text{conv } S$ in a compact form, we introduce

the following definitions:

$$\begin{aligned} E &\equiv \{x \mid x \text{ is an extreme point of } S\}, \\ K' &\equiv \{x \mid A'x = 0, x \geq 0\}, \\ K'_R &\equiv \{x \mid x \in K', x \text{ rational}\}, \\ K'_I &\equiv \{x \mid x \in K', x \text{ integer}\}, \\ Z^n &\equiv \{x \mid x \in R^n, x \text{ integer}\}. \end{aligned}$$

LEMMA 1 (dominance lemma). *Every infinite sequence from the nonnegative integer points of R^n admits an infinite subsequence for which each of the n infinite sequences of components is separately either strictly increasing or completely constant.*

Proof. Select a subsequence whose first components are strictly increasing or constant. Select a subsequence of the result whose second components are strictly increasing or constant, etc. \square

LEMMA 2. $|E|$ is finite.

Proof. If the result were false, by the dominance lemma, there would be two extreme points x' and x'' satisfying $x' \leq x''$ with $x' \neq x''$. However, $x' + 2(x'' - x')$ is easily seen to be in S , and the equation $x'' = \frac{1}{2}x' + \frac{1}{2}[x' + 2(x'' - x')]$ contradicts the hypothesis that x'' was an extreme point of S . \square

Lemma 2 was proved in [7], but the above proof is more compact and offers more geometric insight. The next result was also established in [7].

THEOREM 2. $S \subseteq \text{conv } E + K'_R$.

It might be noted that the structural result $S = (\text{conv } E + K'_R) \cap Z^n$ is easily obtained from Theorem 2 simply by checking that the linear constraints of S are satisfied. On the other hand the result *cannot* be sharpened to $S \subseteq \text{conv } E + K'_I$ as is seen from the following example: consider the set determined by the constraints

$$x_1 + x_2 \geq 1, \quad 1 \geq x_1 - x_2 \geq -1,$$

and x integer; by adding the appropriate slack variables it may be seen that $(1, 1, 1, 1, 1)$ is a feasible solution of the resulting system of equations but is not in $\text{conv } E + K'_I$.

The proof of the following result is straightforward and may be found in [9], which is an expanded version of this paper.

LEMMA 3. $K' = \text{conv } K'_R = \text{conv } K'_I$.

THEOREM 3. $\text{conv } S = \text{conv } E + K'$.

Proof. From Theorem 2 it follows that $S \subseteq \text{conv } E + K'_R \subseteq \text{conv } E + K'$. Since $\text{conv } E + K'$ is convex, $\text{conv } S \subseteq \text{conv } E + K'$. To get the opposite inequality, note that by Lemma 3, $\text{conv } E + K' = \text{conv } E + \text{conv } K'_I$, so that $\text{conv } E + K' = \text{conv } (E + K'_I) \subseteq \text{conv } S$, since $(E + K'_I) \subseteq S$. \square

Since the sets $\text{conv } E$ and K' are polyhedral, Theorem 3 establishes the *polyhedrality* of S .

Theorem 3 also allows an interesting comparison to be made between $\text{conv } S$ and the *linear programming relaxation* T' of S defined by $T' \equiv \{x \mid A'x = b', x \geq 0\}$. Since T' is polyhedral and line-free, we have (see [11]) $T = \text{conv } E' + K'$, where E' is the set of extreme points of T' . Comparing this with Theorem 2, we note that, roughly speaking, $\text{conv } S$ and T' "coincide" in their asymptotic parts and "differ" only in their extreme points. (From a computational point of view, however, this difference is crucial, since the extreme points of T' have a nice algebraic characterization (as basic feasible solutions) and have a cardinality that is bounded from above by $\binom{n}{m}$, whereas these

properties do not carry over to the extreme points of S .) This property is *not* the case, however, for the *linear programming relaxation* T defined in terms of the original constraints by $T = \{x | Ax = b, x \geq 0\}$, as may be seen by considering the following example: if the constraints $Ax = b$ are given by $x_1 - \sqrt{2}x_2 = 1 - \sqrt{2}$, then T consists of the ray $\{x | x_1 - \sqrt{2}x_2 = 1 - \sqrt{2}, x_1 \geq 0, x_2 \geq 0\}$, whereas $T' = \{x | x_1 = 1, x_2 = 1\} = \{(1, 1)\}$. Of course if A is rational, then T and T' coincide, but without hypotheses on A , it is only possible to conclude that $T \supseteq T'$ and that $K = \{x | Ax = 0, x \geq 0\} \supseteq K'$ (that $A'x = b'$ implies $Ax = b$ and that $A'x = 0$ implies $Ax = 0$ are easily seen from (7)). These results are summarized in the following theorem, where E^* denotes the set of *extreme points* of T .

THEOREM 4. $\text{conv } S = \text{conv } E + K' \subseteq \text{conv } E' + K' = T' \subseteq T = \text{conv } E^* + K$ and $K' \subseteq K$.

Proof. Since $S \subseteq T'$ and T' is convex, $\text{conv } S \subseteq T'$. The other relations have been previously discussed. \square

Note also that Theorem 4 implies that if T is a *bounded* set, then $K' = K = \{0\}$ and $\text{conv } E \subseteq \text{conv } E' \subseteq \text{conv } E^*$. However, if T and T' are *unbounded*, then no ordering relations need hold between E, E' , or E^* (or between their convex hulls), as may be seen by considering the example of § 2 in which the corresponding sets are $E = \{(1, 2, 2)\}$, $E' = \{(0, \frac{5}{4}, 2)\}$, and $E^* = \{(0, 0, \theta)\}$, where $\theta = (\frac{5}{3} - \sqrt{2})/(\frac{4}{3})$.

Polyhedrality of $\text{conv } S$ may also be demonstrated directly without resorting to the rational representation of Theorem 1. Defining $K_R = \{x | Ax = 0, x \geq 0, x \text{ rational}\}$ we may prove along the lines of the proof of Theorem 3 that $\text{conv } S = \text{conv } E + \text{conv } K_R$, so that polyhedrality of $\text{conv } S$ will follow from the polyhedrality of $\text{conv } K_R$. Polyhedrality of $\text{conv } K_R$ is established by consideration of $\text{span } K_R$, i.e. the set of all linear combinations of elements of K_R , and use of the following lemma:

LEMMA 4. If $x \in \text{span } K_R \cap R_+^n$, then there is a rational $\bar{x} \in \text{span } K_R \cap R_+^n$ with $\bar{x}_i = 0$ if and only if $x_i = 0$ ($i = 1, \dots, n$).

Proof. First note that since $\text{span } K_R$ is the span of rational vectors, any maximal independent subset of K_R forms a rational basis for $\text{span } K_R$. Fix such a maximal independent set and let B be the matrix whose i th column is the i th vector in this independent set. Then

$$\text{span } K_R = \{B\alpha | \alpha \in R^k\}.$$

Fix $x \in \text{span } K_R \cap R_+^n$. Then there exists an α such that $B\alpha = x$. Let \hat{B} be the matrix consisting of those rows, b_i , of B for which $x_i = 0$. Since \hat{B} is rational, $\{\beta | \hat{B}\beta = 0\}$ has a rational basis matrix, C . Hence $\alpha = C\gamma$ for some γ . Perturb γ slightly to get a rational γ' . Then $\alpha' = C\gamma'$ is rational, $\hat{B}\alpha' = 0$, and with a small enough perturbation α' is sufficiently close to α so that $B\alpha'$ has positive components where $B\alpha$ has positive components. Let $\bar{x} = B\alpha'$. \square

THEOREM 5. $\text{conv } K_R = \text{span } K_R \cap R_+^n$.

Proof. Since $\text{span } K_R \cap R_+^n$ is convex, it suffices to show that $\text{conv } K_R \supseteq \text{span } K_R \cap R_+^n$. Fix $x \in \text{span } K_R \cap R_+^n$. We simply drive each coordinate of x to zero by subtracting appropriate multiples of rationals in $\text{span } K_R \cap R_+^n$.

Specifically, using Lemma 4, choose a rational $r_1 \in \text{span } K_R \cap R_+^n$ that has the same zero coordinates as x . Then there exists a number, γ_1 , such that $x - \gamma_1 r_1$ is nonnegative and has more zero coordinates than x . Continue choosing r 's and γ 's so that at each step, j , $x - \sum_{i < j} \gamma_i r_i$ is nonnegative and has more zero coordinates than $x - \sum_{i < j-1} \gamma_i r_i$. This process must stop at some $j_0 \leq n$ with $x - \sum_{i < j_0} \gamma_i r_i = 0$, i.e. $x = \sum_{i < j_0} \gamma_i r_i$, where the weights γ_i are nonnegative. By adjusting the weights and r_i as in the proof of Lemma 3, it may be shown that $x \in \text{conv } K_R$. \square

In closing, it should be reiterated that in the *inequality* constrained case, if we define $S_I = \{x | Ax \leq b, x \geq 0, x \text{ integer}\}$, it need *not* be the case that $\text{conv } S_I$ is polyhedral. This may be seen from the following problem considered in [7]:

$$\begin{aligned}
 &\text{maximize} && -\alpha x_1 + x_2 \\
 &\text{subject to} && -\alpha x_1 + x_2 \leq 0 \\
 (9) \quad &&& x_1 \geq 1 \\
 &&& x_2 \geq 0 \\
 &&& x_1, x_2 \text{ integer.}
 \end{aligned}$$

It was shown that the problem (9) does not have an optimal solution if α is any positive irrational, even though it is feasible and not unbounded. This phenomenon could not occur if the convex hull of the feasible set were polyhedral, since that property would guarantee the existence of an optimal solution. (In this particular example, it may be shown that the corresponding S_I actually has an infinite number of extreme points (see [3] for related work). Moreover, by replacing α by rationals suitably close to α , and by replacing the variable x_1 by $x'_1 = x_1 - 1$, it may be shown that, in the equality-constrained case, the number of extreme points of S can be made arbitrarily large if $n \geq 3$, i.e., if the number of variables is at least 3. This contrasts with the equality-constrained cases in which $n = 1$ and 2, where from the geometry of S it is clear that maximum number of extreme points is 1 and 2 respectively. (For related complexity results in the inequality-constrained case see [3], [4], [5], [12].) However, if the matrix A is *rational*, then the constraints of S_I may be converted into an equivalent set of equations in integer variables, so that the results above may be applied to prove that $\text{conv } S_I$ is polyhedral in the rational coefficient case.

REFERENCES

- [1] A. M. GEOFFRION, *Lagrangean relaxation for integer programming*, Mathematical Programming Study 2, 1974.
- [2] A. M. GEOFFRION AND R. NAUSS, *Parametric and postoptimality analysis in integer linear programming*, Western Management Science Institute Working Paper, No. 246, Univ. of California, Los Angeles, January 1976.
- [3] S. HALFIN, *Arbitrarily complex corner polyhedra are dense in R^n* , SIAM J. Appl. Math., 23 (1972), pp. 157–163.
- [4] R. G. JEROSLOW, *On the unlimited number of faces in integer hulls of linear problems with two constraints*, Tech. Rpt. #67, Dept. of Opns. Res., Cornell Univ., Ithaca, NY, April 1969.
- [5] ———, *Comments on integer hulls of two linear constraints*, Operations Res., 19 (1971), pp. 1061–1069.
- [6] ———, *Some structure and basis theorems for integral monoids*, Man. Sci. Res. Rpt., No. 367, Carnegie-Mellon Univ., Pittsburgh, PA, July 1975.
- [7] R. R. MEYER, *On the existence of optimal solutions to integer and mixed-integer programming problems*, Math. Programming, 7 (1974), pp. 223–235.
- [8] R. R. MEYER AND J. M. FLEISHER, *Strong duality for a class of integer programs*, Computer Sci. Dept. Tech. Rpt. #256, Univ. of Wisconsin—Madison, May 1975.
- [9] R. R. MEYER AND M. L. WAGE, *On the polyhedrality of the convex hull of the feasible set of an integer program*, Tech. Sum. Rpt. #1653, Math. Research Center, University of Wisconsin—Madison, July 1976.
- [10] H. NOLTEMEIER, *Sensitivitätsanalyse bei diskreten linearen Optimierungsproblemen*, Lecture Notes in Operations Research and Mathematical Systems, M. Beckmann and H. P. Kunzi, eds., Springer-Verlag, New York, 1970.
- [11] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [12] D. S. RUBIN, *On the unlimited number of faces in integer hulls of linear programs with a single constraint*, Operations Res., 18 (1970), pp. 940–946.

RELAXED CONTROLS AND THE DYNAMICS OF CONTROL SYSTEMS*

ZVI ARTSTEIN†

Abstract. The relationship between relaxed controls and the family of processes or flows generated by ordinary controls is studied. We find that the flows generated by the relaxed controls form a completion of the space of flows generated by ordinary controls. With the aid of this completion we study the asymptotic and limiting behavior of the dynamics of the control system. Invariance properties of the ω -limiting sets of admissible solutions are established. Stability, eventual stability and finite time stability properties of solutions with respect to ordinary and relaxed controls are investigated.

1. Introduction. Relaxed controls were designed primarily to provide generalized solutions to optimal control problems, when ordinary solutions fail to exist, see Warga [13] and Young [14]. There are two basic schemes. One is to complete the space of ordinary controls by the relaxed ones; within the completion the existence of an optimal solution is guaranteed by compactness arguments. The second one is to complete the space of possible trajectories, thus obtaining generalized solutions, or generalized curves. Further study was devoted to the characterization of optimal solutions, necessary conditions, approximation by ordinary controls, etc., see Warga [13] and the references therein.

In the present paper we study ordinary and relaxed controls from a different angle—the relation to the dynamics of the system. The consideration of the dynamics of the control system is fundamental in the investigation of attainability and controllability. The dynamics also plays an essential role in the discussion of stability properties of the system and in its limiting behavior. We shall discuss the relation of relaxed controls and the global analogue of generalized curves, namely generalized processes, to these problems.

We use the relaxed controls and the generalized processes in the same way as was originally introduced by Warga and by Young; they form a completion of the corresponding spaces of ordinary controls and ordinary processes. To each ordinary control a process is associated, this process describes the dynamics of the system under that particular control. We introduce generalized processes as limits of ordinary processes. Under appropriate compactness assumptions the generalized processes are generated by relaxed controls.

We shall see how the limiting behavior of the control system is described in terms of the generalized processes and the relaxed controls. This includes invariance properties of ω -limit sets, stability and asymptotic stability properties. We are dealing in this paper only with open loop controls, i.e., controls which are functions of time only. Stabilization of systems via synthesis is not treated.

The paper is organized as follows. In § 2 we set our notations and state our assumptions.

The processes generated by ordinary controls are introduced in § 3. The space of processes is then completed, in an appropriate metric, and the generalized processes are formed. Under an appropriate compactness stated in this section we show how the generalized processes are generated by relaxed controls.

Invariance properties of ω -limit sets for autonomous and nonautonomous systems are treated in §§ 4 and 5 respectively. The invariance in the autonomous case

* Received by the editors August 3, 1976 and in revised form July 25, 1977.

† Department of Mathematics, The Weizmann Institute of Science, Rehovot, Israel. This research was supported in part by the United States-Israel Binational Science Foundation.

is basically a restatement of Peng's original work [9] in terms of generalized processes or relaxed controls. In the nonautonomous case we incorporate the recent developments by Wakeman [12], Dafermos [4], [5], and others (see the survey [2]) and modify them to suit the special structure of control systems.

Finite time stability (see Infante and Weiss [6]) is the subject of § 6. We discuss the preservation of finite time stability under perturbations in the control.

Stability is treated in § 7. We are interested to see how stability, uniform stability and attractivity are maintained under perturbations in the controls. In particular, the perturbations allowed show that if one of the properties holds for a general (relaxed) control, then it holds for a certain ordinary control as well.

2. Notations and assumptions. The control systems in this paper have the form

$$(2.1) \quad \dot{x} = f(x, t, u),$$

where $x \in R^n$ the n -dimensional Euclidean space, $t \in R$ the real line and u denotes an element in a fixed metric (not necessarily compact) space U .

Measurability is understood to be in the Lebesgue sense; equalities are always "almost everywhere" and solutions are absolutely continuous functions. The norm of $x \in R^n$ is denoted by $|x|$ and \dot{x} denotes differentiation with respect to time, i.e., $\dot{x} = (dx/dt)$.

A family \mathcal{U} of measurable functions u from R to U is given. These are the *admissible controls*.

We assume that f is continuous in x and u and measurable in t . We also assume that for every admissible control $u = u(t)$ in \mathcal{U} and every $(x_0, t_0) \in R^n \times R$ the initial value problem

$$\dot{x} = f(x, t, u(t)), \quad x(t_0) = x_0,$$

has a unique solution, which we denote by $x(t, t_0, x_0; u)$, and which is defined for all t . This solution is then continuous in (t, t_0, x_0) .

3. Generalized processes and relaxed controls. A study of processes and their relation to nonautonomous differential equations was done by Dafermos [4] and Sell [10], [11]. Following Dafermos [5] we have the following definition.

DEFINITION 3.1. A process on R^n is a mapping $\varphi: R \times R \times R^n \rightarrow R^n$ such that

- (1) φ is continuous,
- (2) $\varphi(0, t_0, x_0) = x_0$,
- (3) $\varphi(t + \tau, t_0, x_0) = \varphi(t, t_0 + \tau, \varphi(\tau, t_0, x_0))$

for all t, τ, t_0 and x_0 . A process φ is a *dynamical system* if φ does not depend on t_0 , i.e. $\varphi = \varphi(t, x_0)$.

We want to promote the following interpretation. The process φ describes the evolution of a (generally time-dependent) system. The pair (t_0, x_0) is the initial data, (initial time t_0 and initial state x_0). The mapping $\varphi(t, t_0, x_0)$ gives the motion as a function of time t , relative to t_0 , and the initial data. Notice that t measures the time passed from the initial time t_0 . Thus condition (2) states that if no time passed, $t = 0$, the motion is still at x_0 . An example of a process is given in the following proposition.

PROPOSITION 3.2. Let u be an admissible control. Let $x(t, t_0, x_0; u)$ be the solution of $\dot{x} = f(x, t, u(t))$, $x(t_0) = x_0$. Then $\varphi(t, t_0, x_0) = x(t + t_0, t_0, x_0; u)$ is a process.

Proof. Obvious from the continuity of x and from x being a solution.

The processes given in the preceding result are defined by admissible controls. The adjective "generalized" of the generalized processes indicates that these are not in general generated by admissible controls, but rather in a limiting procedure. As a

matter of fact, a generalized process might not be governed by an ordinary differential equation at all. (We shall see later how they are generated by relaxed controls.) Before introducing the generalized processes we need the following definition.

DEFINITION 3.3. The sequence φ_k of processes on R^n converges to the process φ if $\varphi_k(t, t_0, x_0) \rightarrow \varphi(t, t_0, x_0)$ in R^n uniformly on compact subsets of $R \times R \times R^n$.

Remark 3.4. The convergence of processes given in the previous definition is a metric convergence. A metric can be given as follows. Let B_j be an increasing sequence of compact subsets of $R \times R \times R^n$, with union equal to $R \times R \times R^n$. Define

$$\rho(\varphi, \psi) = \sum_{j=1}^{\infty} 2^{-j} \min(1, \sup\{|\varphi(t, t_0, x_0) - \psi(t, t_0, x_0)| : (t, t_0, x_0) \in B_j\}).$$

Then $\varphi_k \rightarrow \varphi$ if and only if $\rho(\varphi_k, \varphi) \rightarrow 0$.

DEFINITION 3.5. The process φ is a *generalized process* (relative to (2.1)) if there is a sequence \mathbf{u}_k of admissible controls in \mathcal{U} such that $\varphi_k = \varphi(t, t_0, x_0; \mathbf{u}_k)$ converges to φ .

A natural question arises, namely whether the family of processes generated by admissible controls has a compactness property. I.e., is it true that each sequence of such processes has a subsequence converging to a (generalized) process. If the system possesses this compactness property we shall say that it is *dynamically compact*.

We shall now introduce the relaxed controls and their relation to the generalized processes. For notational convenience we shall work under quite restrictive conditions.

On several occasions in the sequel we shall use the following set of assumptions (therefore we give it a name).

Assumption L. The function f is locally bounded and locally Lipschitz in x , i.e., for every bounded $B \subset R^n \times R$ a constant K exists such that $|f(x, t, u)| \leq K$ and $|f(x, t, u) - f(y, t, u)| \leq K|x - y|$ for every $(x, t), (y, t)$ in B and $u \in U$. The set U is compact and the family \mathcal{U} consists of all the measurable functions $\mathbf{u}: R \rightarrow U$.

We shall see now that under Assumption L the system is dynamically compact, and furthermore the generalized processes are generated by relaxed controls. To this end we merely have to employ standard material (which appears in detail in [13], [14]). Therefore our presentation will be brief.

Throughout the rest of this section we assume that Assumption L holds.

We denote by M the metric space of probability measures on U endowed with the metric generated by weak convergence. I.e., $\mu_k \rightarrow \mu$ in M if $\int_U g d\mu_k$ converge to $\int_U g d\mu$ for every continuous $g: U \rightarrow R$.

An *admissible relaxed control* is a measurable function $\mathbf{v}: R \rightarrow M$. The family of admissible relaxed controls will be denoted by \mathcal{RC} .

We consider \mathcal{RC} as a metric space. The formal definition of the metric can be found in Warga [13; Chap. IV]. What we actually need to know is that \mathcal{RC} is a compact space and the convergence of \mathbf{v}_k to \mathbf{v} in \mathcal{RC} is equivalent to

$$(3.6) \quad \int_T \int_U g(t, u) d\mathbf{v}_k dt \rightarrow \int_T \int_U g(t, u) d\mathbf{v} dt$$

for every bounded interval T and a real-valued function g which is continuous in u , measurable in t and such that $m(t) = \max\{|g(t, u)| : u \in U\}$ is integrable over T .

An ordinary control \mathbf{u} is considered as an element of \mathcal{RC} by viewing the value $u(t)$ as being the Dirac measure concentrated at the point $u(t) \in U$, (i.e., the measure of $\{u(t)\}$ is 1). An important information is (see Warga [13; Chap. IV]): *the ordinary controls are dense in \mathcal{RC}* .

It is also worth noting that the restriction of the convergence in $\mathcal{R}\mathcal{C}$ to ordinary controls is equivalent to the local convergence in measure, i.e., if \mathbf{u}_k and \mathbf{u}_0 are ordinary controls, then $\mathbf{u}_k \rightarrow \mathbf{u}_0$ in $\mathcal{R}\mathcal{C}$ if and only if for every bounded T and every $\varepsilon > 0$ the Lebesgue measure of $\{t \in T: d(u_k(t), u_0(t)) \geq \varepsilon\}$ tends to zero as $k \rightarrow \infty$ (here $d(\cdot, \cdot)$ denotes the metric in U). Equivalently, for every T the integrals $\int_T d(u_k(t), u_0(t)) dt$ converge to zero. The sufficiency of this is obvious. The necessity is provided by considering the special function g in (3.6) given by $g(t, u) = d(u, u_0(t))$.

Following Young [14] we define $f(x, t, v)$ for a measure v in M by

$$f(x, t, v) = \int_U f(x, t, u) dv.$$

(Clearly, $f(x, t, u)$ is not affected if $u \in U$ is viewed as the corresponding Dirac measure.) With an element $\mathbf{v} = v(t)$ in $\mathcal{R}\mathcal{C}$ we associate the ordinary differential equation

$$\dot{x} = f(x, t, v(t)).$$

Our assumptions imply that $f(x, t, v(t))$ is continuous in x and measurable in t ; moreover the Lipschitz condition is preserved by the integration. A classical result (see Young [14; § 33]) implies that for every (t_0, x_0) in $R \times R^n$ there exists a unique solution of $\dot{x} = f(x, t, v(t))$ satisfying $x(t_0) = x_0$. We denote this solution by $x(t, t_0, x_0; \mathbf{v})$ and assume that $x(t, t_0, x_0; \mathbf{v})$ is defined for all t , i.e., the solution exists on the entire line. Clearly, if \mathbf{v} is a relaxed control then $\varphi(\mathbf{v})$ defined by $\varphi(t, t_0, x_0; \mathbf{v}) = x(t_0 + t, t_0, x_0; \mathbf{v})$ is a process.

The following result establishes the relations between relaxed controls and the generalized processes, and summarizes the information we shall need in the sequel.

THEOREM 3.7. *The mapping $\mathbf{v} \rightarrow \varphi(\mathbf{v})$ is a continuous mapping from $\mathcal{R}\mathcal{C}$ to the space of processes (see Definition 3.3). Therefore the compactness of $\mathcal{R}\mathcal{C}$ implies that the system is dynamically compact. Furthermore, each generalized process is generated as $\varphi(\mathbf{v})$ for a certain relaxed control \mathbf{v} ; and the density of the ordinary controls in $\mathcal{R}\mathcal{C}$ implies that each relaxed control generates a generalized process.*

Proof. The key statement is the opening one. This is an old result on continuous dependence on parameters. It follows directly from the characterization of convergence in $\mathcal{R}\mathcal{C}$ given in (3.6). For details consult [13] or [1].

As we noted before there is a similarity between our construction of generalized processes and the construction of generalized solutions in Warga [13] or Young [14]. There is one procedure which does not generalize to processes. A well known result is that the solutions of the differential equation $\dot{x} = f(x, t, v(t))$ where $\mathbf{v} = v(t)$ is a relaxed control are identical with the solutions of the differential relation $\dot{x} \in F(x, t)$ where $F(x, t)$ is the convex hull of $\{f(x, t, u): u \in U\}$. The analogous statement for processes does not hold. A process φ can be generated by the differential relation, i.e., satisfy it, without being a generalized process. Berkovitz [3] introduces relaxed controls as functions $v(t)$ with values being purely atomic measures concentrated at $n + 1$ points in U . This definition does not suit our purposes either since not every generalized process is generated by such a relaxed control (although any generalized solution is generated in this way).

4. Invariance for time-independent systems. Among other invariance properties we shall reproduce Peng's invariance results [9] in terms of generalized processes and relaxed controls. Illustrative examples will be given at the end of this section; for various other applications see Peng [9], Wakeman [12] and LaSalle [7].

In this section we treat time-independent control systems

$$(4.1) \quad \dot{x} = f(x, u);$$

but notice that for a particular control function $u(t)$ the associated equation $\dot{x} = f(x, u(t))$ is nonautonomous. An *admissible trajectory* $\mathbf{x} = x(t)$ is a solution of (4.1) for an admissible control $\mathbf{u} = u(t)$, i.e., $\dot{x}(t) = f(x(t), u(t))$. Recall that $x(t, t_0, x_0; \mathbf{v})$ denotes the solution of $\dot{x} = f(x, v(t))$, $x(t_0) = x_0$.

Let $\mathbf{y} = y(t)$ be an R^n -valued continuous function defined on an infinite interval $[t_0, \infty)$. The ω -limit set $\Omega(\mathbf{y})$ of \mathbf{y} is the set of vectors z for which an increasing sequence of times $t_j \rightarrow \infty$ exists such that $z = \lim y(t_j)$.

We say that the family of admissible controls is *translation invariant* if for every $\mathbf{u} = u(t)$ in \mathcal{U} and for every σ the translation \mathbf{u}^σ of \mathbf{u} by σ , defined by $u^\sigma(t) = u(\sigma + t)$, is also in \mathcal{U} . Notice that under Assumption L the family \mathcal{U} is translation invariant.

THEOREM 4.2. *Assume that \mathcal{U} is translation invariant and that the system is dynamically compact. Let $\mathbf{x} = x(t)$ be an admissible trajectory of (4.1). Then $\Omega(\mathbf{x})$ has the following invariance property. For each $z \in \Omega(\mathbf{x})$ there is a generalized process φ such that $\varphi(t, 0, z) \in \Omega(\mathbf{x})$ for every t .*

Proof. Let $\mathbf{u} = u(t)$ be the admissible control for which \mathbf{x} is the solution of $\dot{x} = f(x, u(t))$. Let $z = \lim x(t_j)$ where $t_j \rightarrow \infty$. We shall write \mathbf{u}^j for \mathbf{u}^{t_j} , i.e., the translation of \mathbf{u} by t_j . Notice that $\varphi(t, t_j, x(t_j); \mathbf{u}) = \varphi(t, 0, x(t_j); \mathbf{u}^j)$. Since the system is dynamically compact the sequence $\varphi(\mathbf{u}^j)$ has a subsequence converging to a generalized process φ . We shall show that $\varphi(t, 0, z)$ is in $\Omega(\mathbf{x})$ for every t . To this end notice that $\varphi(t, 0, z)$ is the limit of $\varphi(t, 0, x(t_j); \mathbf{u}^j)$ which is equal to $\varphi(t, t_j, x(t_j); \mathbf{u}) = x(t + t_j)$. Since $t + t_j \rightarrow \infty$ for any fixed t the desired relation follows. This completes the proof.

COROLLARY 4.3. *Suppose that Assumption L holds. Let $\mathbf{x} = x(t)$ be an admissible solution for (4.1). Then for every $z \in \Omega(\mathbf{x})$ there exists a relaxed control \mathbf{v} such that $\varphi(t, 0, z; \mathbf{v}) \in \Omega(\mathbf{x})$ for every t .*

Proof. Follows from Theorem 4.2 together with Theorem 3.7.

COROLLARY 4.4 (Peng's quasi-invariance [9]). *Suppose that Assumption L holds. For each $z \in \Omega(\mathbf{x})$ there is a sequence \mathbf{u}_i of ordinary controls such that the uniform limit on bounded intervals of $x(t, 0, z; \mathbf{u}_i)$ exists and belongs to $\Omega(\mathbf{x})$ for all t .*

Proof. This follows from Corollary 4.3 and the density of the ordinary controls in \mathcal{RC} ; indeed, any sequence of ordinary controls \mathbf{u}_i which converges to \mathbf{v} will do.

COROLLARY 4.5 (Peng's invariance [9]). *Suppose that Assumption L holds. If $F(x) = \{f(x, u); u \in U\}$ is convex for every x , then for each $z \in \Omega(\mathbf{x})$ an ordinary control \mathbf{u} exists for which $\varphi(t, 0, z; \mathbf{u}) \in \Omega(\mathbf{x})$ for all t .*

Proof. This is implied by the Filippov lemma; see Warga [13] or Young [14].

Another type of invariance properties is concerned with changing the role of the quantifiers in Theorem 4.2 as following. (Compare [2; Theorem 7.3]).

THEOREM 4.6. *Assume that \mathcal{U} is translation invariant. Let $\mathbf{u} = u(t)$ be an admissible control and assume that $\Omega(\mathbf{x})$ is compact where $\mathbf{x} = x(t)$ is a solution of $\dot{x} = f(x, u(t))$. Then $\Omega(\mathbf{x})$ has the following invariance property: if φ is a generalized process obtained as a limit of $\varphi(\mathbf{u}^{t_j})$ for a certain $t_j \rightarrow \infty$ then a point $z \in \Omega(\mathbf{x})$ exists such that $\varphi(t, 0, z) \in \Omega(\mathbf{x})$ for every t .*

Proof. Since $\Omega(\mathbf{x})$ is compact it follows that $x(t_j)$ has a converging subsequence, say $x(t_k)$, with a limit $z \in \Omega(\mathbf{x})$. We claim that this is the desired z . Indeed, $\varphi(t, 0, z)$ is by definition the limit of $\varphi(t, 0, x(t_k); \mathbf{u}^{t_k})$ which is equal to $\varphi(t, t_k, x(t_k); \mathbf{u}) = x(t + t_k)$. Since $t + t_k$ tends to infinity it follows that $\varphi(t, 0, z) \in \Omega(\mathbf{x})$. This completes the proof.

COROLLARY 4.7. *Suppose that Assumption L holds. Let $\mathbf{u} = u(t)$ be an admissible control and let \mathbf{v} be the limit of \mathbf{u}^{t_j} for a certain $t_j \rightarrow \infty$. If for a solution $\mathbf{x} = x(t)$ of*

$\dot{x} = f(x, u(t))$ the ω -limit set $\Omega(\mathbf{x})$ is compact, then a point $z \in \Omega(\mathbf{x})$ exists such that $\varphi(t, 0, z; \mathbf{v})$ is in $\Omega(\mathbf{x})$ for every t .

We shall now discuss two examples illustrating the previous results and their possible applications.

Consider first the controlled damping harmonic oscillator

$$\ddot{x} + g(u)\dot{x} + x = 0$$

where u takes two values, u_1 and u_2 , and $g(u_1) = 1$ (positive damping), and $g(u_2) = -1$ (negative damping). Suppose that under the control $\mathbf{u}_0 = u_0(t)$ the eventual behavior of the oscillator is a periodic one, i.e., in the (x, \dot{x}) plane the motion tends to a circle, say $\Gamma = \{(x, \dot{x}) : x^2 + \dot{x}^2 = 1\}$. The relaxed controls \mathcal{RC} are functions $\mathbf{v}_0 = v_0(t)$ associating with each t a probability distribution on $\{u_1, u_2\}$. The only such control which generates a motion which leaves Γ invariant is the control which is concentrated equally on u_1 and u_2 . From Theorem 4.6 we can conclude that \mathbf{u}'_0 converges to \mathbf{v}_0 in \mathcal{RC} . The latter convergence means in this case that for any fixed τ the measure of $\{s : u_0(s) = u_1, t \leq s \leq t + \tau\}$ tends to $\tau/2$ as $t \rightarrow \infty$.

Consider now the system

$$\begin{aligned}\dot{x} &= y + u \\ \dot{y} &= -(u^2 - 1)^2 y - x\end{aligned}$$

with $u \in \{0, 1\}$. Suppose we want to locate those points $z_0 = (x_0, y_0)$ such that $z_0 = \lim_{t \rightarrow \infty} \varphi(t, 0, z; \mathbf{u}_0)$ for certain \mathbf{u}_0 and z . Our Theorem 4.2 tells us that $\{z_0\}$ is invariant with respect to a certain process generated by a relaxed control. A singleton being invariant means that it is a rest point of the process. Namely, the right hand side of the equation should vanish identically. The value of the relaxed control \mathbf{v}_0 at t is determined by a distribution between 0 and 1, say with weights μ and $1 - \mu$, where $0 \leq \mu \leq 1$. It is therefore easy to see that if (x_0, y_0) is a limit point as before, then for a certain $\mu \in [0, 1]$

$$\begin{aligned}y_0 &= -\mu \\ x_0 &= \mu(1 - \mu).\end{aligned}$$

The latter equations give in a parametrized form the location of the possible limit points. It is not hard to see (using phase portraits of the equations generated by the two controls 0 and 1) that each such point (x_0, y_0) is indeed obtained as a limit point for an ordinary control.

5. Limiting control systems and invariance for nonautonomous systems. The invariance properties of the ω -limit sets, given in the previous section, do not hold for time dependent systems $\dot{x} = f(x, t, u)$. (Obviously the key equality $\varphi(t, t_i, x_0; \mathbf{u}) = \varphi(t, 0, x_0; \mathbf{u}')$ uses strongly the time-independence of the equation.) A possible generalization is as follows. Given the admissible control \mathbf{u} we are facing a nonautonomous equation $\dot{x} = f(x, t, u(t))$. Limit sets of solutions of this equation have invariance properties with respect to the *limiting equations* of the system. Peng's results were generalized in this direction by Wakeman [12]. Further developments concerning nonautonomous equations are described in Artstein [2]. Here we want to discuss another type of limiting behavior and the related invariance properties, namely, the limiting behavior of the entire control system $\dot{x} = f(x, t, u)$ at infinity. A comparison with the former approach will be done below.

For convenience we will use in this section assumptions on f slightly stronger than Assumption L. We require that the Lipschitz constant in Assumption L is uniform in t ,

i.e., for every bounded $W \subset R^n$ a constant $K = K(W)$ exists such that $|f(x, t, u) - f(y, t, u)| \leq K|x - y|$ whenever $x, y \in W$ and for all t and u .

DEFINITION 5.1. The translate by τ of the function f is denoted by f^τ and given by $f^\tau(x, t, u) = f(x, t + \tau, u)$.

DEFINITION 5.2. A control system $\dot{x} = g(x, t, u)$ is a *limiting system* of $\dot{x} = f(x, t, u)$ if there exists a sequence $\tau_i \rightarrow \infty$ such that for every x, t and u

$$(5.3) \quad \int_0^t g(x, s, u) ds = \lim_{j \rightarrow \infty} \int_0^t f^{\tau_j}(x, s, u) ds.$$

We then say that f^{τ_j} converges to g .

Before presenting the applications we want to describe some structure properties of the limiting systems.

PROPOSITION 5.4. *The family of limiting systems is closed under translations, i.e., if $\dot{x} = g(x, t, u)$ is a limiting system then $\dot{x} = g^\tau(x, t, u)$ is a limiting system for every τ .*

Proof. Notice that f^{τ_i} converges to g implies $f^{\tau + \tau_i}$ converges to g^τ .

PROPOSITION 5.5. *For any sequence $\tau_i \rightarrow \infty$ there is a subsequence τ_j for which f^{τ_j} converges.*

Proof. For a triplet (x, t, u) the sequence of mappings $f(x, s + \tau_i, u): [-t, t] \rightarrow R^n$ is weakly precompact in $L_1([-t, t], R^n)$ and has a weak limit $g(x, s, u)$. Let (x_k, t_k, u_k) be a dense sequence in $R^n \times R \times U$. A simple diagonal argument enables us to find a subsequence τ_j of τ_i such that for each k , $f(x_k, s + \tau_j, u_k)$ converge weakly to $g(x_k, s, u_k)$ on $[-t_k, t_k]$. In particular

$$\int_0^t g(x_k, t, u_k) = \lim_{j \rightarrow \infty} \int_0^t f(x_k, s + \tau_j, u_k) ds$$

for every t . We claim that this g , so far defined on a dense sequence, has an extension to the entire space $R^n \times R \times U$. To this end notice that on the dense sequence g is Lipschitzian in x and uniformly continuous in u , so an extension exists. Furthermore it is not hard to see (compare Artstein [1; § 4]) that the extension, which we again denote by g , satisfies the same estimations for boundedness and the Lipschitz constant as f . Having these estimations it is easily seen that (5.3) holds. This will complete the proof.

We want to analyze the limiting behavior of the processes under translations. Since we compare processes generated by different systems (the original system, its translations and its limiting systems), we will add one argument to the notations, namely, we denote by

$$\varphi(t, t_0, x_0; \mathbf{v}, g)$$

the process generated by applying the control \mathbf{v} to the system $\dot{x} = g(x, t, u)$. We denote by \mathbf{v}^τ and φ^τ the translations by τ of the functions \mathbf{v} and φ respectively, i.e.,

$$v^\tau(t) = v(t + \tau) \quad \text{and} \quad \varphi^\tau(t, t_0, x_0) = \varphi(t, t_0 + \tau, x_0).$$

The following equality is easily checked

$$\varphi^\tau(t, t_0, x_0; \mathbf{v}, g) = \varphi(t, t_0, x_0; \mathbf{v}^\tau, g^\tau).$$

LEMMA 5.6. *Suppose that \mathbf{u}^{τ_i} converge to \mathbf{v} in \mathcal{RC} and that f^{τ_i} converge to g . Then $\varphi^{\tau_i}(\mathbf{u}, f)$ converge as $i \rightarrow \infty$ to $\varphi(\mathbf{v}, g)$.*

Proof. This is again (see the proof of Theorem 3.7) a result on continuous dependence on parameters, here the parameters (\mathbf{u}, f) . It is easy to check that under

the assumptions on f the convergence f^{τ_i} to g and \mathbf{u}^{τ_i} to \mathbf{v} imply that $f^{\tau_i}(x, t, u(t))$ converge weakly in L_1 to $g(x, t, v(t))$ for any fixed x and on any bounded interval of times. Then we can apply the continuous dependence results, for instance in Artstein [1] or one of the references therein.

THEOREM 5.7. *Let \mathbf{u} be an admissible control and let $\mathbf{x} = x(t)$ be a solution of $\dot{x} = f(x, t, u(t))$. Then for every $z \in \Omega(\mathbf{x})$ there is a limiting system $\dot{x} = g(x, t, u)$ and a relaxed control \mathbf{v} such that $\varphi(t, 0, z; \mathbf{v}, g) \in \Omega(\mathbf{x})$ for every t .*

Proof. z is the limit of $x(\tau_i)$ for a certain sequence $\tau_i \rightarrow \infty$. Let τ_j be a subsequence such that f^{τ_j} has a limit g , and \mathbf{u}^{τ_j} has a limit \mathbf{v} . By the lemma $\varphi(t, 0, z; \mathbf{v}, g)$ is the limit of $\varphi(t, 0, x(\tau_j); \mathbf{u}^{\tau_j}, f^{\tau_j})$. The latter is equal to $\varphi(t, \tau_j, x(\tau_j); \mathbf{u}, f) = x(t + \tau_j)$. Since $t + \tau_j \rightarrow \infty$ it follows that $\varphi(t, 0, z; \mathbf{v}, g) \in \Omega(\mathbf{x})$. This completes the proof.

The previous theorem supplies a class of equations— $\dot{x} = g(x, t, v(t))$ for limiting systems $\dot{x} = g$ and all relaxed controls \mathbf{v} —with respect to which $\Omega(\mathbf{x})$ has the invariance property. If the control \mathbf{u} which is used in generating \mathbf{x} is known, we can limit the class of equations above to those generated by the limits $\mathbf{v} = \lim \mathbf{u}^{\tau_i}$ for a sequence $\tau_i \rightarrow \infty$. (This can be seen in the proof of the theorem.) The family can be even further limited by considering the *limiting equations* of $\dot{x} = f(x, t, u(t))$, i.e., the equations obtained as limits

$$h(x, t) = \frac{\partial}{\partial t} \lim \int_0^t f(x, s + \tau_i, u(s + \tau_i)) ds$$

for every fixed pair (x, t) . Compare Artstein [2] or Wakeman [12]. The conceptual disadvantage of the latter approach is that limiting systems of $\dot{x} = f(x, t, u)$ can be computed a priori, while for the computation of the limiting equations we need the precise control function which was used.

6. Stability considerations and finite time stability. We fix initial state x_1 , initial time t_1 and a time interval $[t_1, t_1 + T]$. Let φ be a process. We want to examine the stability of the trajectory through (t_1, x_1) over the finite interval with length T . Following Infante and Weiss [6], we define

DEFINITION 6.1. The trajectory $\varphi(t, t_1, x_1)$ is *stable* (over $[t_1, t_1 + T]$ and with respect to φ) *with respect to* (α, β) , $0 < \alpha \leq \beta$, if $|y_1 - x_1| \leq \alpha$ implies $|\varphi(t, t_1, y_1) - \varphi(t, t_1, x_1)| \leq \beta$ for $0 \leq t \leq T$. The following result follows directly from the definition of the convergence in the space of processes.

PROPOSITION 6.2. *If $\varphi(t, t_1, x_1)$ is stable with respect to (α, β) then $\psi(t, t_1, x_1)$ is stable with respect to (α', β) if $\alpha' < \alpha$ and ψ is close to φ in the sense provided by Definition 3.3 (see Remark 3.4).*

The preceding result can be applied to control systems. It implies that if a certain generalized process, say generated by a relaxed control, possesses a certain modulus (α, β) of finite time stability then this modulus can almost be obtained by using an ordinary control. Compare Proposition 6.2 to Definition 3.5 and Theorem 3.7. This might affect the choice of control when stability of the solution enters into the considerations, as the following example shows.

Example 6.3. We want to illustrate the significance of maintaining the finite time stability estimates under perturbations in the control. Consider the following optimization problem:

Minimize $\int_0^T |x(t)| dt$ where $x(t)$ is an admissible solution of the control system

$$\dot{x} = u + (1 - u^2)x, \quad x(0) = 0,$$

(x and u are scalars) and the control u takes values in the set $U = \{-1, 0, 1\}$. It is easy to see that there is a unique optimal ordinary control, namely $\mathbf{u}_0 = 0$. However, there are other optimal controls which are relaxed; for instance the constant-valued control \mathbf{v}_0 concentrated equally on 1 and -1 . It is easy to see that the stability estimates are as follows: $\varphi(t, 0, 0; \mathbf{u}_0)$ is stable with respect to $(\alpha, \alpha e^T)$ while $\varphi(t, 0, 0; \mathbf{v}_0)$ is stable with respect to (α, α) for any $\alpha > 0$. This immediately affects the optimal cost. If there is a perturbation up to α in the initial condition then if \mathbf{u}_0 is used, the maximum cost guaranteed is up to $\alpha(e^T - 1)$ while if \mathbf{v}_0 is used the cost does not exceed αT . Thus, although the performance of \mathbf{v}_0 cannot be realized by an ordinary control it might be preferable (if perturbations are expected and T is large) to use an ordinary approximation of \mathbf{v}_0 and maintain almost the same stability, rather than use the less stable control \mathbf{u}_0 .

7. Stability. We assume that $f(0, t, u) = 0$ for all t and u . We are interested in the preservation of stability properties of the origin under perturbations in the controls. In particular we will be pleased to find that stability properties of 0 which hold when a certain generalized process or a relaxed control is used, are valid also for close enough ordinary controls. As before $\varphi(t, t_0, x_0; \mathbf{v})$ denotes the process generated by the control \mathbf{v} (see § 3). The assumption above implies that $\varphi(t, t_0, 0; \mathbf{v}) = 0$ for all t, t_0 and \mathbf{v} . We are interested in the stability over the positive ray $[0, \infty)$.

The origin 0 is *stable* with respect to a process φ if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that $|y_0| \leq \delta$ implies $|\varphi(t, 0, y_0)| \leq \varepsilon$ for all $t \geq 0$. (The stability is stated with respect to initial time 0, but it clearly implies stability for any initial time.) The origin is *uniformly stable* with respect to φ if for every $\varepsilon > 0$ a $\delta > 0$ exists such that $|y_0| \leq \delta$ implies $|\varphi(t, t_0, y_0)| \leq \varepsilon$ for all $t, t_0 \geq 0$. The origin is a *uniform attractor* with respect to φ if there is a $\delta_0 > 0$ and a positive function $p(t)$ on $[0, \infty)$ such that $p(t) \rightarrow 0$ as $t \rightarrow \infty$ and $|\varphi(t, t_0, y_0)| \leq p(t)$ whenever $|y_0| \leq \delta_0$ and for all $t, t_0 \geq 0$. We shall say that the origin is *stable* (or *uniformly stable*, or a *uniform attractor*) with respect to a control \mathbf{v} if the origin is *stable* (resp. *uniformly stable* or a *uniform attractor*) with respect to the process generated by \mathbf{v} .

In Remark 3.4 we presented the metric which generates the concept of convergence of processes, which has the uniform convergence on compact sets. But stability is a global property of a process. We shall therefore modify the consideration of a distance between processes. For each $n > 0$ we define

$$\rho_n(\varphi, \psi) = \sum_{j=1}^{\infty} 2^{-j} \min(1, \sup\{|\varphi(t, n, x_0) - \psi(t, n, x_0)| : |x_0| \leq j, 0 \leq t \leq 1\}).$$

Thus ρ_n measure the distance between φ and ψ operating on the interval $[n, n+1]$. In particular φ_k converges to φ in the meaning of Definition 3.3 if and only if $\rho_n(\varphi_k, \varphi) \rightarrow 0$ for each fixed n .

In a similar way we modify the consideration of a distance in \mathcal{RC} . Recall that the restriction of the functions in \mathcal{RC} to a unit interval $I = [n, n+1]$ belong to a bounded set in $L_{\infty}(I, C(U)^*)$ and that the weak* topology on the latter induces a metric on the restrictions (see § 3). We denote this metric by $d_n(\cdot, \cdot)$, this for every integer n . In this way with any two relaxed controls \mathbf{v} and \mathbf{u} a sequence of distances $d_n(\mathbf{u}, \mathbf{v})$ is associated. Notice that $\mathbf{v}_k \rightarrow \mathbf{v}$ in \mathcal{RC} if and only if $d_n(\mathbf{v}, \mathbf{v}_k) \rightarrow 0$ as $k \rightarrow \infty$, for any fixed n . (In fact, an explicit form for the metric in \mathcal{RC} could have been given by $\sum 2^{-|n|} d_n(1 + d_n)^{-1}$.)

Even a uniform bound on $d_n(\mathbf{u}, \mathbf{v})$ is not enough to preserve stability. An example is the scalar equation $\dot{x} = |u|x$. Here 0 is stable with respect to the identically zero control, but it is unstable for any $u(t) = c$ if $c \neq 0$.

We shall state and prove the following result in terms of processes. The applications to control system are given as corollaries.

THEOREM 7.1. *Suppose that 0 is stable with respect to the process φ . Then there is a sequence of positive numbers b_0, b_1, b_2, \dots , such that 0 is stable with respect to any process ψ provided 0 is a rest point of ψ and $\rho_n(\varphi, \psi) \leq b_n$ for $n = 0, 1, \dots$.*

Proof. We proceed in two steps. First let $\varepsilon > 0$ be fixed. There is a $\delta > 0$ such that $|y_0| \leq \delta$ implies $|\varphi(t, 0, y_0)| \leq \varepsilon$ for all $t \geq 0$. Let $B(t) = \{\varphi(t, 0, y_0) : |y_0| \leq \delta\}$ and let $S(t)$ be the boundary of $B(t)$. Notice that $B(t)$ is homeomorphic to the ball $\{y_0 : |y_0| \leq \delta\}$, and that $S(t)$ is the image of $\{y_0 : |y_0| = \delta\}$ by the mapping $y_0 \rightarrow \varphi(t, 0, y_0)$. We shall define by induction a set-valued mapping $C(t)$. Let $C(0) = \{y_0 : |y_0| \leq \delta_1\}$ where $0 < \delta_1 < \delta$ is fixed. Suppose now that $C(n)$ is defined, compact, included in $B(t)$ and disjoint from $S(t)$. Let $C(t)$ for $n \leq t < n+1$ be given by $C(t) = \{\varphi(t-n, n, y_0) : y_0 \in C(n)\}$. Since $C(n)$ is disjoint from $S(n)$ there is a positive number η_n such that $z \in C(t)$ implies $\inf\{|z-y| : y \in S(t)\} \geq 2\eta_n$. The definition of the distance ρ_n implies that a number a_n exists such that if $\rho_n(\varphi, \psi) \leq a_n$, $y_0 \in C(n)$ and $n \leq t \leq n+1$ then $|\varphi(t-n, n, y_0) - \psi(t-n, n, y_0)| \leq \eta_n$. We define $C(n+1)$ to be the closed η_n -neighborhood of $\{\varphi(1, n, y_0) : y_0 \in C(n)\}$. The estimation above shows that $C(n+1)$ is included in $B(n+1)$ and disjoint from $S(n+1)$; this completes the induction step. The construction implies that if $y_0 \in C(n)$ and $\rho_i(\varphi, \psi) \leq a_i$ for $i \geq n$ then $|\psi(t, n, y_0)| \leq \varepsilon$ for all $t \geq 0$.

Let now $\varepsilon_k = k^{-1}$ for $k = 1, 2, \dots$. For each k the construction above gives us a sequence $a_{n,k}$ of positive numbers and a set-valued mapping $C_k(t)$ such that if $y_0 \in C_k(n)$ and $\rho_i(\varphi, \psi) \leq a_{i,k}$ for $i \geq n$, then $|\psi(t, n, y_0)| \leq \varepsilon_k$ for all $t \geq 0$. Define $b_n = \min\{a_{n,k} : k \leq n\}$. We claim that b_n is the desired sequence of positive numbers.

In order to prove the claim let ψ be such that $\rho_n(\varphi, \psi) \leq b_n$ for $n \geq 0$. Let ε_k be fixed. The continuity of φ in the initial conditions and $\psi(k, 0, 0) = 0$ imply that a $\delta_1 > 0$ exists such that $|y_0| \leq \delta_1$ implies $\psi(k, 0, y_0) \in C_k(k)$. Now $\rho_n(\varphi, \psi) \leq b_n \leq a_{n,k}$ for $n \geq k$ implies that $|\psi(t, 0, y_0)| \leq \varepsilon_k$ for all $t \geq 0$. This completes the proof.

COROLLARY 7.2. *Suppose that Assumption L holds. Suppose that 0 is stable with respect to a control \mathbf{v} in \mathcal{RC} . Then there is a sequence of positive numbers b_0, b_1, \dots , such that 0 is stable with respect to any control \mathbf{u} in \mathcal{RC} provided $d_n(\mathbf{v}, \mathbf{u}) \leq b_n$ for $n = 1, 2, \dots$.*

COROLLARY 7.3. *If the origin is stable with respect to a certain generalized process, then it is stable also with respect to a certain ordinary control.*

Proof. Given a generalized process φ and positive numbers b_n for $n \geq 0$, there is no problem in constructing an ordinary control \mathbf{u} such that $\rho_n(\varphi, \varphi(\mathbf{u})) \leq b_n$ for $n \geq 0$.

Although the details of the construction in the proof of Theorem 7.1 might appear complicated, the geometrical idea is simple. The same is true with respect to the proof of the next two results.

THEOREM 7.4. *Suppose that 0 is uniformly stable with respect to φ . Then a sequence of positive numbers b_0, b_1, \dots exists, such that $\rho_n(\varphi, \psi) \leq b_n$ for all $n \geq 0$ and 0 being a rest point of ψ imply 0 is uniformly stable with respect to ψ .*

Proof. We proceed in two steps. First let $\varepsilon > 0$. Then a $\delta > 0$ exists such that $|y_0| < \delta$ implies $|\varphi(t, t_0, y_0)| \leq \varepsilon$ for any $t, t_0 \geq 0$. Let $B_k(t) = \{\varphi(t, k, y_0) : |y_0| \leq \delta\}$ and let $S_k(t)$ be its boundary. Let δ_1 be $0 < \delta_1 < \delta$. We successively define sets $C_k(t)$. We let $C_k(0) = \{y_0 : |y_0| \leq \delta_1\}$. Suppose that $C_k(n)$ is defined, compact, included in $B_k(n)$ but disjoint from $S_k(n)$. Let $C_k(t) = \{\varphi(t-n, n+k, y_0) : y_0 \in C_k(n)\}$ for $n \leq t \leq n+1$. Since $C_k(n)$ is disjoint from $S_k(n)$ it follows that a positive number $\eta_{k,n}$ exists such that $z \in C_k(t)$ for $n \leq t < n+1$ implies that $\inf\{|z-y| : y \in S_k(t)\} \geq 2\eta_{k,n}$. By the continuity of φ in the initial data a number $a_{k,n} > 0$ exists such that $\rho_{k+n}(\varphi, \psi) \leq a_{k,n}$ implies that

$|\varphi(t-n, n+k, y_0) - \psi(t-n, n+k, y_0)| \leq \eta_{k,n}$. Also let $C_k(n+1)$ be the closed $\eta_{k,n}$ -neighborhood of $\{\varphi(1, n+k, y_0) : y_0 \in C_k(n)\}$. Then $C_k(n+1)$ is compact, contained in $B_k(n+1)$ but disjoint from $S_k(n+1)$. This completes the induction step. It also follows that if $|y_0| \leq \delta_1$ and $\rho_{k+i}(\varphi, \psi) \leq a_{k,i}$ for $i \geq 0$ then $|\psi(t, k, y_0)| \leq \varepsilon$ for all t . The uniform stability of 0 with respect to φ also implies that a $\delta_2 > 0$ exists such that $|z_0| \leq \delta_2$ implies $|\varphi(t, t_0, z_0)| \leq (1/2)\delta_1$ for all t . Then the continuity of φ implies that for each interval $[n, n+1]$ a positive number c_n exists such that $\rho_n(\varphi, \psi) \leq c_n$; $n \leq t_0 \leq n+1$ and $|z_0| \leq \delta_2$ imply that $|\psi(n+1-t_0, t_0, z_0)| \leq \delta_1$. Define now $a_n = \min\{c_n, a_{n,0}, a_{n-1,1}, \dots, a_{0,n}\}$. It is easy to check that $\rho_n(\varphi, \psi) \leq a_n$ for $n \geq 0$ and $|z_0| \leq \delta_2$ imply that $|\psi(t, t_0, z_0)| \leq \varepsilon$ for all $t, t_0 \geq 0$.

Let now $\varepsilon_k = k^{-1}$. For each k a $\delta_k \geq 0$ exists and a sequence of positive numbers $b_{k,0}, b_{k,1}, \dots$, such that $\rho_n(\varphi, \psi) \leq b_{k,n}$ for all n , and $|z_0| \leq \delta_k$ imply $|\psi(t, t_0, z_0)| \leq \varepsilon_k$. Define $b_n = \min\{b_{k,n} : k \leq n\}$. We claim that this is the desired sequence. In order to prove it let ψ be such that $\rho_n(\varphi, \psi) \leq b_n$ for all n . Let ε_k be fixed. Then $\rho_n(\varphi, \psi) \leq b_{k,n}$ for $n \geq k$, which means that $|z_0| \leq \delta_k$ implies $|\psi(t, t_0, z_0)| \leq \varepsilon_k$ for all $t \geq 0$ and $t_0 \geq k$. For the finite interval $[0, k]$ a number δ'_k exists such that $|z_0| \leq \delta'_k$ and $t_0 \leq k$ imply $|\psi(k-t_0, t_0, z_0)| \leq \delta_k$. (Only now we use the assumption that 0 is a rest point of ψ .) The min (δ_k, δ'_k) will establish the uniform stability of ψ . This completes the proof.

COROLLARY 7.5. *Suppose that Assumption L holds. If 0 is uniformly stable with respect to a control \mathbf{v} in \mathcal{RC} , then a sequence of positive numbers b_0, b_1, \dots exists such that $d_n(\mathbf{v}, \mathbf{u}) \leq b_n$ for all $n \geq 0$ implies that 0 is uniformly stable with respect to \mathbf{u} .*

COROLLARY 7.6. *If 0 is uniformly stable with respect to a certain generalized process then an ordinary control exists with respect to which 0 is uniformly stable.*

It is implicit in the proofs of Theorems 7.1 and 7.4 that if $\delta(\varepsilon)$ is the modulus of stability of φ then for a fixed ε almost the same modulus can be maintained by the approximation ψ . In terms of the corollaries this means that if $\delta(\varepsilon)$ is the modulus of stability of a generalized process, then for a fixed ε , almost the same modulus can be maintained by using an ordinary control.

Another interesting problem is how to construct the sequences b_k and how to verify whether an ordinary control, or the generated process, satisfies the inequalities required in the results. In principle the proofs are constructive. Namely, the estimates b_k , and the estimates used in constructing b_k , are concerned with estimates of continuous dependence on parameters, and can all be checked and computed. This however might be very complicated if the geometry of the solutions funnel is complicated. If there are good estimates concerning the solutions funnel of the process, the estimates can be determined rather easily. We shall illustrate this using an example.

Consider the equation

$$\dot{x} = u|x| + (1-u^2)x$$

where, say, $u \in \{-1, 0, 1\}$. The origin is a rest point for any admissible control, but it is not a stable rest point for any constant control. The origin is a stable rest point if the constant relaxed control \mathbf{v}_0 , which is distributed identically on 1 and -1 , is used: Then the equation is simply $\dot{x} = 0$. It is clear that an approximation to \mathbf{v}_0 is any ordinary control \mathbf{u}_0 which takes values 1 and -1 on alternating small intervals. Given such an ordinary control $\mathbf{u}_0 = u_0(t)$, the distance d_n between \mathbf{v}_0 and \mathbf{u}_0 has the representation

$$(7.7) \quad d_n(\mathbf{v}_0, \mathbf{u}_0) = \max_{n \leq t \leq s \leq n+1} \int_t^s u_0(\sigma) d\sigma.$$

It is also easy to check then that $|\varphi(t, s, x_0; \mathbf{u}_0)| \leq |x_0| + d_n(\mathbf{v}_0, \mathbf{u}_0)$ if $|x_0| \leq 1$. This estimate is the desired estimate of the continuous dependence needed for the construction of the sequence b_n . Indeed, if b_n is summable, i.e., $\sum b_n < \infty$, and if \mathbf{u}_0 takes only the values 1 and -1 and $d_n(\mathbf{u}_0, \mathbf{v}_0) \leq b_n$ for every n , then 0 is uniformly stable respect to \mathbf{u}_0 .

THEOREM 7.8. *Suppose that 0 is a uniform attractor with respect to the process φ . Then there is a sequence of positive numbers b_0, b_1, \dots , such that $\rho_n(\varphi, \psi) \leq b_n$ for $n \geq 0$ implies that 0 is a uniform attractor with respect to ψ .*

Proof. Let δ_0 and $p(t)$ be given by the definition of a uniform attractor. Let $0 < \delta_1 < \delta_0$. We proceed in two steps. First we fix an integer $k \geq 0$. For $k \leq t_0 \leq k+1$ and $t \geq 0$ we define $B(t, t_0) = \{\varphi(t, t_0, y_0) : |y_0| \leq \delta_0\}$. Then $z \in B(t, t_0)$ implies $|z| \leq p(t)$. We inductively define sets $C(t, t_0)$ as follows. Let $C(0, t_0) = \{y : |y| \leq \delta_1\}$. Suppose that $C(n, t_0)$ are defined ($k \leq t_0 \leq k+1$), contained in $B(n, t_0)$ but disjoint from its boundary. Let $C(t, t_0)$ be defined for $n \leq t < n+1$ by $C(t, t_0) = \{\varphi(t-n, t_0+n, y_0) : y_0 \in C(n, t_0)\}$. Then $C(t, t_0)$ is included in $B(t, t_0)$ but disjoint from its boundary. Let $\eta_n > 0$ be a number such that an η_n -neighborhood of $C(t, t_0)$ is still disjoint from the boundary of $B(t, t_0)$. Then two positive numbers a_n, c_{n+1} exist such that $\rho_{k+n}(\varphi, \psi) \leq a_n$ and $\rho_{k+n+1}(\varphi, \psi) \leq c_{n+1}$ imply that $|\varphi(t-n, n+t_0, y_0) - \psi(t-n, n+t_0, y_0)| \leq \eta_n$ for $y_0 \in C(n, t_0)$. Let $C(n+1, t_0)$ be the closed η_n -neighborhood of $\{\varphi(1, n+t_0, y_0) : y_0 \in C(n, t_0)\}$. This completes the induction step. In particular we get that if $b_n = \min(a_n, c_n)$ and $\rho_{n+k}(\varphi, \psi) \leq b_n$ for $n \geq 0$ then $|z_0| \leq \delta_1$ implies $|\psi(t, t_0, z_0)| \leq p(t)$ for $t \geq 0$ and $k \leq t_0 \leq k+1$.

In order to complete the proof notice that the numbers b_n obtained in the first step really depend on k , so denote them by $b_{n,k}$. Define $b_n = \min\{b_{n-k,k} : k \leq n\}$. This sequence b_0, b_1, \dots is clearly the desired one.

COROLLARY 7.9. *Suppose that Assumption L holds. Suppose that 0 is a uniform attractor with respect to a control \mathbf{v} in \mathcal{RC} . Then there is a sequence of positive numbers b_0, b_1, \dots such that $d_n(\mathbf{u}, \mathbf{v}) \leq b_n$ for $n \geq 0$ implies that 0 is a uniform attractor with respect to \mathbf{u} .*

COROLLARY 7.10. *If 0 is a uniform attractor with respect to a certain generalized process, then an admissible ordinary control exists with respect to which 0 is too a uniform attractor.*

Our previous remark concerning the preservation of the range of stability is valid also for the range of attractivity. Namely, the proof of Theorem 7.8 shows that if δ_0 determines the region of attraction of the process φ then for any $\delta_1 < \delta_0$, the estimates b_n can be chosen so that δ_1 determines a region of attraction of a process φ if $\rho_n(\varphi, \psi) \leq b_n$ for every n . In particular, the region of attraction of a generalized process, or a relaxed control, can almost be maintained by an ordinary control. The same sort of preservation holds with respect to the rate of convergence to 0, which is estimated by the function $p(t)$. This is also clear from the proof.

How the rate of convergence enters into the considerations of choosing the controls is demonstrated by the following example. Consider the scalar equation

$$\dot{x} = u + (1 - u^2)x - ax,$$

where $a > 0$ and $|u(t)| \leq 1$. It is quite easy to choose a control $\mathbf{u}_0 = u_0(t)$ which stabilizes the origin, i.e., which makes the origin a uniform attractor. Furthermore, it is easy to make the rate of convergence an exponential one. However, the best convergence rate will be achieved by the relaxed control \mathbf{v}_0 , which is constant and equally distributed between 1 and -1 . (The reason is clear, such a control annihilates the unstable part $(1 - u^2)x$ and makes the asymptotically stable part $-ax$ dominant.) In

order to maintain almost the same rate by an ordinary control, the latter should be close in the sense of Corollary 7.9 to v_0 . If this control takes only 1 and -1 as values, then (7.7) is valid and can be used as the basis of estimating the rate of convergence.

Acknowledgment. I want to thank the referee for his constructive criticism and for suggesting the examples closing §§ 4 and 7.

REFERENCES

- [1] Z. ARTSTEIN, *Topological dynamics of an ordinary differential equation*, J. Differential Equations, 23 (1977), pp. 216–223.
- [2] ———, *Limiting equations and stability of nonautonomous ordinary differential equations*, an appendix in LaSalle [7].
- [3] L. D. BERKOVITZ, *Optimal Control Theory*, Springer-Verlag, New York, 1974.
- [4] C. M. DAFERMOS, *An invariance principle for compact processes*, J. Differential Equations, 9 (1971), pp. 239–252.
- [5] ———, *Semiflows associated with compact and uniform processes*, Math. Systems Theory, 8 (1974), pp. 142–149.
- [6] E. T. INFANTE AND L. WEISS, *On the stability of systems defined over a finite interval*, Proc. Nat. Acad. Sci. U.S.A., 54 (1965), pp. 44–48.
- [7] J. P. LASALLE, *The Stability of Dynamical Systems*, CBMS Regional Conference Series in Applied Mathematics, 25, Society for Industrial and Applied Mathematics, Philadelphia, 1976.
- [8] E. J. MCSHANE, *Relaxed controls and variational problems*, this Journal, 5 (1967), pp. 438–485.
- [9] T. K. C. PENG, *Invariance and stability for bounded uncertain systems*, this Journal, 10 (1972), pp. 679–690.
- [10] G. R. SELL, *Non-autonomous differential equations and topological dynamics*, I and II. Trans. Amer. Math. Soc., 127 (1967), pp. 241–283.
- [11] ———, *Lectures on Topological Dynamics and Differential Equations*, Van Nostrand Reinhold, London, 1971.
- [12] D. R. WAKEMAN, *An application of topological dynamics to obtain a new invariance property for non-autonomous ordinary differential equations*, J. Differential Equations, 17 (1975), pp. 259–295.
- [13] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
- [14] L. C. YOUNG, *Calculus of Variations and Optimal Control Theory*, W. B. Saunders Co., Philadelphia 1969.

A STOCHASTIC CONTROL MODEL WITH CHANCE CONSTRAINTS*

NORBERT CHRISTOPEIT†

Abstract. In this paper a control system governed by a linear stochastic differential equation is considered. In the class of all feedback controls which are linear in the state the expected value of an integral performance index is to be minimized subject to the additional condition that with a prescribed minimum probability the terminal point lies in some fixed target set. An equivalent nonlinear deterministic control problem is derived together with an existence result and necessary optimality conditions.

1. Introduction. The system dynamics are represented by the linear stochastic differential equation

$$(1.1) \quad dx = A(t)x(t) dt + B(t)u(t) dt + C(t) dw, \quad 0 \leq t \leq T,$$

with initial condition

$$(1.2) \quad x(0) = x_0.$$

Here w is a d -dimensional Brownian motion defined on some probability space (Ω, \mathcal{E}, P) , and x_0 is a random or deterministic n -vector.

Our problem is to find a control u in some class of functions to be specified below that minimizes the performance index

$$(1.3) \quad J(x, u) = E \left\{ \int_0^T l(t, x(t), u(t)) dt \right\}$$

subject to the constraints (1.1), (1.2) and

$$(1.4) \quad P(x(T) \in S) \geq 1 - \alpha.$$

Here S is a given set in \mathbb{R}^n and α is a fixed number between 0 and 1 determining the significance (or confidence) level. In applications α will be chosen near 0. This last constraint requires that the target set is attained at least with probability $1 - \alpha$.

The model bears some relation to the so-called chance constrained programming problem of Charnes and Cooper [4], [5]; in our case, however, the variable x is itself stochastic rather than the data as in chance constrained programming.

Results concerning the existence of optimal controls in problems like (1.1)–(1.3) have been obtained in [2], [7] and [10]. In our case, however, the sets in which the controls u are allowed to vary—see (1.5) below—will not be compact and in general not convex. An additional difficulty comes from the constraint (1.4), which cannot be expressed as the expected value of some continuous function of the final state.

As feasible controls we admit all feedback control laws u that are linear in the state variable, i.e.

$$(1.5) \quad u(t) = U(t)x(t) + v(t)$$

with nonstochastic $m \times n$ -matrices $U(t)$ and m -vectors $v(t)$ taking on values in fixed sets $\mathcal{U}(t)$ and $\mathcal{V}(t)$, respectively, for almost all t . Inserting (1.5) into (1.1) we get

$$(1.6) \quad dx = [A(t) + B(t)U(t)]x(t) dt + B(t)v(t) dt + C(t) dw.$$

* Received by the editors February 1, 1977.

† Institut für Ökonometrie und Operations Research, University of Bonn, D-53 Bonn, West Germany. This work was supported by the Sonderforschungsbereiche 21 and 72 at the University of Bonn.

If the initial data x_0 is normally distributed (possibly degenerated) the solution of (1.6) is itself a Gaussian process, its mean and covariance satisfying the ordinary differential equations

$$(1.7) \quad \dot{\mu}(t) = [A(t) + B(t)U(t)]\mu(t) + B(t)v(t)$$

and

$$(1.8) \quad \dot{\Sigma}(t) = [A(t) + B(t)U(t)]\Sigma(t) + \Sigma(t)[A(t)' + U(t)'B(t)'] + C(t)C(t)'$$

with initial data

$$(1.9) \quad \mu(0) = \mu_0 = E(x_0)$$

and

$$(1.10) \quad \Sigma(0) = \Sigma_0 = E((x_0 - \mu_0)(x_0 - \mu_0)'),$$

respectively. Since $x(t)$ follows a $N(\mu(t), \Sigma(t))$ -distribution, (1.3) and (1.4), too, are uniquely determined by the means and covariances. With the normal density

$$\nu(\mu, \Sigma; \xi) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp(-\frac{1}{2}(\xi - \mu)' \Sigma^{-1}(\xi - \mu))$$

we can write

$$(1.11) \quad P(x(T) \in S) = \int_S \nu(\mu(T), \Sigma(T); \xi) d\xi = g(\mu(T), \Sigma(T)) \geq 1 - \alpha$$

and

$$(1.12) \quad \begin{aligned} J(x, u) &= \int_0^T E\{l(t, x(t), u(t))\} dt \\ &= \int_0^T dt \int l(t, \xi, U(t)\xi + v(t)) \nu(\mu(t), \Sigma(t); \xi) d\xi \\ &= \int_0^T L(t, \mu(t), \Sigma(t), U(t), v(t)) dt = F(\mu, \Sigma, U, v), \end{aligned}$$

provided that $\Sigma(t)$ is nonsingular for all $t > 0$ and the interchange of the integrals is justified. It should be noted that the functions $g(\mu, \Sigma)$ and $L(t, \mu, \Sigma, U, v)$ are defined only for regular matrices Σ .

Let us introduce the following classes of functions:

$$\mathcal{F}_1 = \{U: [0, T] \rightarrow \mathbb{R}^{m \times n}: U \text{ measurable, } U(t) \in \mathcal{U}(t) \text{ a.e.}\},$$

$$\mathcal{F}_2 = \{v: [0, T] \rightarrow \mathbb{R}^m: v \text{ measurable, } v(t) \in \mathcal{V}(t) \text{ a.e.}\},$$

$$\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2,$$

$$\mathcal{A} = \{(\mu, \Sigma): [0, T] \rightarrow \mathbb{R}^n \times \mathbb{R}^{n \times n}: (\mu, \Sigma) \text{ is a solution to (1.7)–(1.10) for some pair } (U, v) \in \mathcal{F}\}.$$

Control laws $(U, v) \in \mathcal{F}$ will be called *feasible controls*, corresponding responses *feasible paths or trajectories*.

We can summarize the above considerations by stating that minimizing (1.3) subject to the constraints (1.1), (1.2), (1.4) and (1.5) with $(U, v) \in \mathcal{F}$ —which will be referred to as problem (P)—is equivalent to finding the minimum of (1.12) among all feasible pairs (U, v) subject to the constraints (1.7)–(1.11)—which will be referred to as problem (P'). Thus a deterministic equivalent to the original stochastic control problem has been found, the optimal solutions (U, v) and u , respectively, being related by (1.5).

2. Existence of optimal solutions. Let us make the following *assumptions*:

- (A) The matrix functions $A(t)$, $B(t)$ and $C(t)$ are continuous. $C(t)$ has rank n for almost all $t \in [0, T]$.
- (B) The initial value x_0 is normally distributed (possibly degenerated) and independent of $w_t - w_0$ for all $t \in [0, T]$.
- (C) $\mathcal{U}(t)$ and $\mathcal{V}(t)$ are compact convex sets for all $t \in [0, T]$ varying continuously with t (in the Hausdorff metric).

Assumptions (A) and (B) guarantee that for every pair $(U, v) \in \mathcal{F}$, (1.6), (1.2) has a unique solution which is a Gaussian process with mean and covariance function satisfying (1.7)–(1.10). To make the relationship clear we shall sometimes write $(\mu_{U,v}, \Sigma_{U,v})$ for the solution of (1.7)–(1.8) corresponding to (U, v) . Further it is well known (compare Arnold [1]) that the solution of (1.8) can be written in the form

$$(2.1) \quad \Sigma(t) = \Phi(t) \left(\Sigma_0 + \int_0^t \Phi(s)^{-1} C(s) C(s)' (\Phi(s)^{-1})' ds \right) \Phi(t)'$$

where $\Phi(t)$ is the matrix of fundamental solutions to the homogeneous equation

$$(2.2) \quad \dot{\mu}(t) = [A(t) + B(t)U(t)]\mu(t).$$

For an arbitrary nonzero n -vector a it follows that

$$\begin{aligned} a' \Sigma(t) a &\geq \int_0^t a' \Phi(s) \Phi(s)^{-1} C(s) C(s)' (\Phi(s)^{-1})' \Phi(t)' a ds \\ &= \int_0^t [C(s)' (\Phi(s)^{-1})' \Phi(t)' a]' [C(s)' (\Phi(s)^{-1})' \Phi(t)' a] ds. \end{aligned}$$

By the rank condition in (A) the integrand is positive for almost all s , hence $a' \Sigma(t) a > 0$ for all $t > 0$. So we find that for solutions of (1.6) and (1.12) $\Sigma(t)$ is positive definite for $t > 0$ and thus g and L are well defined along feasible paths.

A feasible control law (U, v) will be called *admissible* if $(\mu_{U,v}(T), \Sigma_{U,v}(T))$ satisfies (1.11). We require:

- (D) The class \mathcal{F}' of admissible controls is nonempty.
- (E) The real-valued function $l(t, x, u)$ is defined on $I \times \mathbb{R}^n \times \mathbb{R}^m$, where I is some open interval containing $[0, T]$. It is continuous in (t, x, u) , convex in u and satisfies the polynomial growth condition

$$|l(t, x, u)| \leq C(1 + |x|^{2p} + |u|^{2p})$$

uniformly in $t \in I$ for some constant C and some integer p .

Condition (E) ensures that $J(x, u)$ is finite for all feasible (x, u) and that the integrals may be interchanged.

Our existence proof will be based on the following result from [3, Thm. 5.1], which we shall state here in a specialized version adapted to our problem (P').

PROPOSITION 1. *Suppose that assumptions (A), (C) and (D) hold. Assume further that there exists a region (=open connected set) \mathcal{R} in (t, μ, Σ) -space together with a compact subset $\mathcal{R}_0 \subset \mathcal{R}$ such that $(t, \mu(t), \Sigma(t)) \in \mathcal{R}_0$ for all $(\mu, \Sigma) \in \mathcal{A}$, $t \in [0, T]$, and such that the following conditions hold:*

- (i) *The target set $\mathcal{T} = \{(\mu, \Sigma); \Sigma \text{ positive definite, } g(\mu, \Sigma) \geq 1 - \alpha\}$ is compact.*
- (ii) *The function L is lower semicontinuous on $\mathcal{G} = \mathcal{R} \times \mathbb{R}^{n \times m} \times \mathbb{R}^m$.*

Then problem (P') has an optimal solution $(\mu^, \Sigma^*, U^*, v^*)$.*

It is plain to see that (i) is not true for problem (P') as it stands. But it will be shown in the next three lemmas that \mathcal{T} may be replaced by a compact set \mathcal{T}' without changing the problem.

LEMMA 1. Let \mathcal{W} be a bounded set in $\mathbb{R}^{n \times n}$ and let \mathcal{G} be the family of all measurable matrix functions $W: [0, T] \rightarrow \mathbb{R}^{n \times n}$ with values in \mathcal{W} a.e. Denote by ξ_W the unique solution of the homogeneous differential equation

$$\dot{\xi}(t) = W(t)\xi(t), \quad \xi(0) = \xi_0.$$

Then for all $t \in [0, T]$ the correspondence $W \mapsto \xi_W(t)$ defines a mapping $\mathcal{G} \rightarrow \mathbb{R}^n$ which is continuous with respect to the weak topology on \mathcal{G} as a subset of $L_2^{n \times n}$ (the space of square integrable $n \times n$ -matrix functions on $[0, T]$).

Proof. The matrices W admitting a uniform bound

$$|W(t)| \leq M_1 \quad \text{a.e. on } [0, T] \quad \text{for all } W \in \mathcal{G}$$

we obtain the estimate

$$(2.3) \quad |\xi_W(t)| \leq |\xi_0| + \int_0^t |W(s)\xi_W(s)| ds \leq |\xi_0| + nM_1 \int_0^t |\xi_W(s)| ds,$$

from which we get

$$|\xi_W(t)| \leq |\xi_0| e^{nM_1 t} \leq |\xi_0| e^{nM_1 T} = M_2 \quad \text{for all } t \in [0, T] \text{ and all } W \in \mathcal{G}$$

by Gronwall's inequality. Now take a sequence of matrices $\{W_k\}$ from \mathcal{G} converging weakly to some $W \in \mathcal{G}$. This means in particular that the functions

$$h_k(t) = \left| \int_0^t (W(s) - W_k(s))\xi_W(s) ds \right|$$

converge to zero for every $t \in [0, T]$ as k goes to infinity. Moreover they are bounded by the constant $2nM_1M_2T$ uniformly in k and t . We have the estimate

$$\begin{aligned} |\xi_W(t) - \xi_{W_k}(t)| &= \left| \int_0^t (W(s)\xi_W(s) - W_k(s)\xi_{W_k}(s)) ds \right| \\ &\leq \left| \int_0^t (W(s) - W_k(s))\xi_W(s) ds \right| + \left| \int_0^t W_k(s)(\xi_W(s) - \xi_{W_k}(s)) ds \right| \\ &\leq \left| \int_0^t (W(s) - W_k(s))\xi_W(s) ds \right| + nM_1 \int_0^t |\xi_W(s) - \xi_{W_k}(s)| ds \end{aligned}$$

for all $t \in [0, T]$. By the Gronwall-Bellman inequality (Fleming and Rishel [8])

$$|\xi_W(t) - \xi_{W_k}(t)| \leq h_k(t) + M_1 \int_0^t h_k(s) e^{M_1(t-s)} ds.$$

By Lebesgue's theorem on dominated convergence the last integral approaches zero as k tends to infinity. \square

LEMMA 2. Under assumption (C) the class \mathcal{F} is weakly closed in $L_2^{m \times n} \times L_2^m$.

Proof. \mathcal{F} is strongly closed and convex, hence weakly closed. \square

Define the attainable set $K(T) = \{(\mu(T), \Sigma(T)) : (\mu, \Sigma) \in \mathcal{A}\}$. Then we have the following result:

LEMMA 3. Under the assumptions (A) and (C), $K(T)$ is compact.

Proof. By the continuity assumption in (C) the sets $\mathcal{U}(t)$ and $\mathcal{V}(t)$, $0 \leq t \leq T$, are contained in some bounded set of $\mathbb{R}^{m \times n}$ and \mathbb{R}^m , respectively. Hence \mathcal{F} is bounded in

the L_2 -norm and by Lemma 2 weakly compact. With Φ_U denoting the fundamental matrix of (2.2) it results from Lemma 1 that the correspondence $U \mapsto \Phi_U(t)$ is weakly continuous for all t . By formula (2.1) and the variation of parameters formula for the solution of (1.7) this is also true for the correspondence $(U, v) \mapsto (\mu_{U,v}(T), \Sigma_{U,v}(T))$. Hence $K(T)$ is the continuous image of a compact set. \square

The constraint (1.11) may be written in the form

$$(\mu(T), \Sigma(T)) \in \mathcal{T}$$

with the target set $\mathcal{T} = \{(\mu, \Sigma) : \Sigma \text{ positive definite, } g(\mu, \Sigma) \geq 1 - \alpha\}$. \mathcal{T} is closed in the cone $\mathcal{K} = \{(\mu, \Sigma) : \Sigma \text{ positive definite}\}$. From the relation $K(T) \subset \mathcal{K}$, Lemma 3 and assumption (D) it follows that

$$\mathcal{T}' = \mathcal{T} \cap K(T)$$

is a nonempty compact set. But obviously problem (P') remains unchanged when (1.11) is replaced by the condition

$$(1.11') \quad (\mu(T), \Sigma(T)) \in \mathcal{T}'.$$

So far we have made no assumptions concerning Σ_0 . In order to obtain continuity properties of L let us first deal with the case that Σ_0 is positive definite, i.e. x_0 follows a nondegenerate normal distribution. Then for all trajectories $(\mu, \Sigma) \in \mathcal{A}$ the point $(t, \mu(t), \Sigma(t))$ lies in the open region $\mathcal{R} = I \times \mathcal{K}$ for all $t \in [0, T]$. In fact the following is true.

LEMMA 4. *Suppose that Σ_0 is positive definite. Then under assumptions (A) and (C) there exists a compact subset \mathcal{R}_0 of \mathcal{R} such that $(t, \mu(t), \Sigma(t)) \in \mathcal{R}_0$ for all $t \in [0, T]$ and all $(\mu, \Sigma) \in \mathcal{A}$.*

Proof. An estimate similar to (2.3) and application of Gronwall's inequality show that $\mu(t)$ is uniformly bounded for all $t \in [0, T]$ and all feasible μ , i.e. $\mu(t) \in \mathcal{K}_1$ for some compact set \mathcal{K}_1 .

Next, consider a sequence $\{\Sigma_k(t_k)\}$, where Σ_k is the response to a feasible U_k and the t_k 's are points in $[0, T]$. Since the controls U_k belong to \mathcal{F}_1 we can select a subsequence $\{U_{k'}\}$ converging weakly to some $U \in \mathcal{F}_1$ with response Σ . Denoting by $\{t_{k''}\}$ a subsequence of $\{t_{k'}\}$ converging to some $t \in [0, T]$, we arrive at a sequence $\{\Sigma_{k''}(t_{k''})\}$ which will be shown to converge to $\Sigma(t)$. For notational simplicity let us use the old index k for this sequence. Then Σ_k converges pointwise to Σ , as was shown in the proof of Lemma 3. On the other hand the Σ_k are uniformly bounded on $[0, T]$ (this may be shown in the same way as for the μ), hence

$$\begin{aligned} |\Sigma_k(t) - \Sigma_k(t')| &\leq 2 \int_{t'}^t [\|A\| + \|B\| \|U_k\|] \|\Sigma_k\| ds + \int_{t'}^t \|C\|^2 ds \\ &\leq \text{const.} \cdot |t - t'| \quad \text{for all } k. \end{aligned}$$

Here $|\cdot|$ denotes Euclidean matrix norm and $\|\cdot\|$ the corresponding essential supremum on $[0, T]$. Combining these two facts we obtain

$$\begin{aligned} |\Sigma_k(t_k) - \Sigma(t)| &\leq |\Sigma_k(t_k) - \Sigma_k(t)| + |\Sigma_k(t) - \Sigma(t)| \\ &\leq \text{const.} \cdot |t_k - t| + |\Sigma_k(t) - \Sigma(t)|, \end{aligned}$$

which shows that $\Sigma_k(t_k) \rightarrow \Sigma(t)$. Hence the set $\mathcal{K}_2 = \{\Sigma(t) : \Sigma \text{ feasible, } t \in [0, T]\}$ is compact. The set $\mathcal{R}_0 = [0, T] \times \mathcal{K}_1 \times \mathcal{K}_2$ then satisfies our requirements. \square

Next we will show that the function L satisfies the assumption (ii) of Proposition 1.

LEMMA 5. Let $q(t, \mu, \Sigma, U, v; \xi)$ be a function from $\mathcal{R} \times \mathbb{R}^{n \times m} \times \mathbb{R}^m \times \mathbb{R}^n$ to the reals which is continuous in (t, μ, Σ, U, v) and satisfies the growth condition

$$(2.4) \quad |q(t, \mu, \Sigma, U, v; \xi)| \leq C(1 + |\mu|^{2p} + |\xi|^{2p} + |\mu|^{2p}|\xi|^{2p})r(\Sigma)$$

for all (t, U, v) in a bounded set. Hence C is a constant, p a positive integer and r a continuous function defined on the halfspace of positive definite matrices.

Then the function

$$(2.5) \quad Q(t, \mu, \Sigma, U, v) = \int_{\mathcal{S}} q(t, \mu, \Sigma, U, v; \xi) \nu(\mu, \Sigma; \xi) d\xi$$

is continuous on $\mathcal{G} = \mathcal{R} \times \mathbb{R}^{n \times m} \times \mathbb{R}^m$.

Proof. Consider a sequence $\{(t_k, \mu_k, \Sigma_k, U_k, v_k)\}$ in \mathcal{G} converging to $(t, \mu, \Sigma, U, v) \in \mathcal{G}$. Using the abbreviating notations $Q_k = Q(t_k, \mu_k, \Sigma_k, U_k, v_k)$, $Q = Q(t, \mu, \Sigma, U, v)$ and similarly $q_k(\xi)$, $q(\xi)$, $\nu_k(\xi)$, $\nu(\xi)$, we obtain

$$\begin{aligned} |Q_k - Q| &\leq \left| \int_{\mathcal{S}} [q_k(\xi) - q(\xi)] \nu(\xi) d\xi \right| + \left| \int_{\mathcal{S}} q_k(\xi) [\nu_k(\xi) - \nu(\xi)] d\xi \right| \\ &= I_1 + I_2. \end{aligned}$$

As to the first term, $q_k(\xi)$ converges pointwise to $q(\xi)$ and is bounded uniformly in k by the function

$$C(1 + \bar{\mu}^{2p} + |\xi|^{2p} + \bar{\mu}^{2p}|\xi|^{2p})\bar{r}$$

with $\bar{\mu} = \sup_k |\mu_k|$, $\bar{r} = \sup_k (r(\Sigma_k))$, which is integrable with respect to the $N(\mu, \Sigma)$ -distribution. Hence, by Lebesgue's theorem, $I_1 \rightarrow 0$.

To obtain an estimate for I_2 , let us first show that the family $\{q_k \nu_k\}$ is uniformly integrable, i.e. for every $\varepsilon > 0$ there exists a $c > 0$ such that

$$(2.6) \quad \int_{|\xi| > c} |q_k(\xi)| \nu_k(\xi) d\xi < \varepsilon \quad \text{for all } k.$$

To this end, write N_k for the distribution $N(\mu_k, \Sigma_k)$, P_k and E_k for the corresponding probabilities and expectations, respectively, and consider the estimate

$$\begin{aligned} &\int_{|\xi| > c} |q_k(\xi)| dN_k(\xi) \\ &\leq \int_{|\xi| > c} C(1 + \bar{\mu}^{2p} + |\xi|^{2p} + \bar{\mu}^{2p}|\xi|^{2p})\bar{r} dN_k(\xi) \\ &\leq \int_{|\xi| > c} C \left(1 + \bar{\mu}^{2p} + |\xi|^{2p} \left| \frac{\xi}{c} \right|^2 + \bar{\mu}^{2p} |\xi|^{2p} \left| \frac{\xi}{c} \right|^2 \right) \bar{r} dN_k(\xi) \\ &\leq C\bar{r} \left[(1 + \bar{\mu}^{2p})P_k(|\xi| > c) + \frac{1}{c^2} (1 + \bar{\mu}^{2p})E_k(|\xi|^{2(p+1)}) \right] \\ &= C\bar{r}(1 + \bar{\mu}^{2p}) \left[P_k(|\xi| > c) + \frac{1}{c^2} \sum_{i=1}^n m_{k,i}^{2(p+1)} \right], \end{aligned}$$

where $m_{k,i}^q$ stands for the q th moment of the i th component of the N_k -distribution, i.e. $m_{k,i}^q = \int \xi_i^q dN_k(\xi)$. Since the moments of a multivariate normal distribution are polynomials in the elements of μ and Σ , convergence of (μ_k, Σ_k) implies convergence of

the moments. Hence the sum of the moments is bounded uniformly in k . By Chebyshev's inequality

$$P_k(|\xi| > c) \leq \frac{1}{c^2} \sum_{i=1}^n m_{k,i}^2.$$

So we find that (2.6) converges to zero uniformly in k as c tends to infinity.

The same argument shows that the family $\{q_k \nu\}$ is uniformly integrable. Now we obtain the following estimate for I_2 :

$$I_2 \leq \left| \int_{S \cap \{|\xi| > c\}} q_k(\xi) dN_k(\xi) \right| + \left| \int_{S \cap \{|\xi| > c\}} q_k(\xi) dN(\xi) \right| + \left| \int_{S \cap \{|\xi| \leq c\}} q_k(\xi) \times [\nu_k(\xi) - \nu(\xi)] d\xi \right|.$$

By uniform integrability the first two terms can be made arbitrarily small for all k if c is chosen large enough. The last term tends to zero for $k \rightarrow \infty$ by Lebesgue's theorem on bounded convergence. \square

COROLLARY 1. *Under assumption (E), $L(t, \mu, \Sigma, U, v)$ is continuous on \mathcal{G} .*

Proof. The only thing to show is that l satisfies the growth condition (2.4). But this is easily derived from the growth condition in (E). \square

Collecting all the results obtained so far, we find that the hypotheses of the existence Theorem 5.1 in [3] hold.

THEOREM 1. *Suppose that the distribution of the initial value is nondegenerate normal. Then under assumptions (A)–(E) problems (P) and (P') have an optimal solution (x^*, u^*) and $(\mu^*, \Sigma^*, U^*, v^*)$, respectively, the control laws u^* and (U^*, v^*) being related by $u^*(t) = U^*(t)x^*(t) + v^*(t)$.*

For degenerate initial data x_0 the covariance Σ_0 is singular, and feasible paths $(t, \mu(t), \Sigma(t))$ start from the boundary of the region \mathcal{R} . We are faced with the problem of finding some open region \mathcal{R}' containing $[0, T] \times \text{cl}(\mathcal{K})$ —all of the points in this set may a priori lie on some feasible path—such that L can be extended continuously to all of $\mathcal{G}' = \mathcal{R}' \times \mathbb{R}^{m \times n} \times \mathbb{R}^m$. Working with the distributions $N(\mu, \Sigma)$ themselves instead of the densities does not lead any further because $N(\mu, \Sigma)$ is defined only for positive semidefinite matrices Σ . We shall indicate how to overcome these difficulties in the special case where $l(t, x, u)$ is a polynomial in the components of x and u with time varying coefficients: $l(t, x, u) = P(t, x, u)$. If P is of degree r in x and of degree s in u then for $(t, \mu, \Sigma) \in [0, T] \times \text{cl}(\mathcal{K})$

$$L(t, \mu, \Sigma, U, v) = \int P(t, \xi, U\xi + v) dN(\mu, \Sigma; \xi)$$

is a linear function of the mixed moments of $N(\mu, \Sigma)$ up to the $(r+s)$ th order with coefficients depending on t and (U, v) . As these moments are in turn polynomials in the components of μ and Σ , it turns out that L is a polynomial in (μ, Σ, U, v) with time varying coefficients. If the coefficients vary continuously with time we find that $L(t, \mu, \Sigma, U, v)$ is defined and continuous on \mathcal{G}' with $\mathcal{R}' = I \times \mathbb{R}^n \times \mathbb{R}^{n \times n}$. Requiring $l(t, x, u)$ to be convex in u the existence result obtained in Theorem 1 is still valid under assumptions (A)–(D).

Let us finally remark that the case just considered covers quadratic performance criteria $l(t, x, u) = x'M(t)x + u'N(t)u$ with nonnegative definite symmetric matrices $M(t), N(t)$ varying continuously in time.

We have dealt here only with the case of a fixed time interval $[0, T]$. Free initial and end time problems can be included in our analysis if the feasible time intervals

$[t_1, t_2]$ are constrained to lie in some compact interval and to satisfy the condition $t_2 \geq t_1 + \delta$ for some fixed positive number δ . However, the formulation of the assumptions and the results becomes a bit more complicated.

3. Necessary optimality conditions. In order to obtain necessary optimality conditions in form of a maximum principle we shall convert problem (P') into a control problem in standard form by substituting an appropriate vector valued variable for the matrix valued variable Σ . For an arbitrary $k \times n$ -matrix M with column vectors m_i let $\text{vec}(M)$ denote the kn -vector $(m'_1, \dots, m'_n)'$. For two matrices M and N define the Kronecker product

$$M \otimes N = \begin{pmatrix} m_{11}N & \cdots & m_{1n}N \\ \vdots & & \vdots \\ m_{k1}N & \cdots & m_{kn}N \end{pmatrix}$$

(compare [12]). It is easy to verify that (1.8) is equivalent to

$$(1.8') \quad \text{vec}(\dot{\Sigma}) = [I \otimes (A + BU) + (A + BU) \otimes I] \text{vec}(\Sigma) + \text{vec}(CC'),$$

where I denotes the $n \times n$ unit matrix.

Next we introduce functions H and G defined on the space $X = C^n \times C^{n^2} \times L_\infty^{nm} \times L_\infty^m$ (where $C^k [L_\infty^k]$ denotes the space of continuous [essentially bounded measurable] functions from $[0, T]$ to \mathbb{R}^k) with values in $C^n \times C^{n^2}$ and \mathbb{R} , respectively, by setting

$$H(\mu, \text{vec}(\Sigma), \text{vec}(U), v)(t)$$

$$= \begin{pmatrix} \mu(t) - \mu_0 - \int_0^t [(A(s) + B(s)U(s))\mu(s) + B(s)v(s)] ds \\ \text{vec}(\Sigma(t)) - \text{vec}(\Sigma_0) - \int_0^t \{ [I \otimes (A(s) + B(s)U(s)) \\ + (A(s) + B(s)U(s)) \otimes I] \text{vec}(\Sigma(s)) \\ + \text{vec}(C(s)C(s')) \} ds \end{pmatrix}$$

and

$$G(\mu, \text{vec}(\Sigma), \text{vec}(U), v) = (1 - \alpha) - g(\mu(T), \Sigma(T)),$$

where g is the same as in (1.11) for $(\mu, \Sigma) \in \mathcal{K}$ and defined arbitrarily—say zero—elsewhere. Further, by abuse of notation, we shall consider F (see (1.12)) as defined on X by setting it equal to some constant for all (μ, Σ) for which $\Sigma(t)$ is not positive definite for all $t > 0$.

Now (P') can be written as an abstract optimization problem:

$$\text{minimize } F(x)$$

$$\text{subject to}$$

$$G(x) \leq 0, \quad H(x) = 0,$$

$$x = (\mu, \text{vec}(\Sigma), \text{vec}(U), v) \in X, \quad (U, v) \in \mathcal{F}.$$

Necessary optimality conditions will be derived on the basis of a multiplier rule proved in [6] which involves convex approximations of the functions F , G and H at the optimal point. It can be shown that for l satisfying (E) and Σ_0 positive definite or for l

which is a polynomial in x and u with continuously time dependent coefficients these convex approximations are given by the Fréchet differentials of G and H at the optimal point $x^* = (\mu^*, \text{vec}(\Sigma^*), \text{vec}(U^*), v^*)$ and by the function

$$\begin{aligned} \hat{F}^*(x) = & \int_0^T [L_{\mu, \text{vec}(\Sigma)}^*(t)(\mu(t)', \text{vec}(\Sigma(t))')' \\ & + L(t, \mu^*(t), \Sigma^*(t), U^*(t) + U(t), v^*(t) + v(t)) \\ & - L(t, \mu^*(t), \Sigma^*(t), U^*(t), v^*(t))] dt, \end{aligned}$$

where $L_{\mu, \text{vec}(\Sigma)}^*(t)$ is the Fréchet differential with respect to $(\mu, \text{vec}(\Sigma))$ evaluated along $x^*(t)$, $x = (\mu, \text{vec}(\Sigma), \text{vec}(U), v)$. To see what the differentials of G and L look like we calculate the partial derivatives of $\nu(\mu, \Sigma; \xi)$ with respect to the components of μ and $\text{vec}(\Sigma)$ at points $(\mu, \Sigma) \in \mathcal{H}$. Making use of the formulas

$$\begin{aligned} \frac{\partial |A|}{\partial \text{vec}(A)} &= |A|(\text{vec}(A^{-1}))', \\ \frac{\partial a^{ij}}{\partial a_{kl}} &= -a^{ik}a^{lj}, \quad A = (a_{kl})_{k,l}, \quad A^{-1} = (a^{ij})_{i,j} \end{aligned}$$

for the differentiation of the determinant and the inverse of a nonsingular matrix (compare [12]) we obtain

$$\begin{aligned} \nu_\mu(\mu, \Sigma; \xi) &= \nu(\mu, \Sigma; \xi)(\xi - \mu)' \Sigma^{-1}, \\ \nu_{\text{vec}(\Sigma)}(\mu, \Sigma; \xi) &= \frac{1}{2} \nu(\mu, \Sigma; \xi)(\text{vec}(\Sigma^{-1}(\xi - \mu)(\xi - \mu)' \Sigma^{-1} - \Sigma^{-1}))'. \end{aligned}$$

It is easy to see that for μ and Σ ranging in a bounded subset of \mathcal{H} the components of ν_μ , $\nu_{\text{vec}(\Sigma)}$ and of $l(t, \xi, U\xi + v)\nu_\mu(\mu, \Sigma; \xi)$, $l(t, \xi, U\xi + v)\nu_{\text{vec}(\Sigma)}(\mu, \Sigma; \xi)$ are bounded uniformly in μ and Σ , respectively, by integrable functions of ξ . Hence integration and differentiation may be interchanged, and we obtain the following vectors of partial derivatives:

$$(3.1) \quad (g_\mu, g_{\text{vec}(\Sigma)}) = \int_S (\nu_\mu, \nu_{\text{vec}(\Sigma)})(\xi) d\xi,$$

$$(3.2) \quad (L_\mu, L_{\text{vec}(\Sigma)}) = \int l(t, \xi, U\xi + v)(\nu_\mu, \nu_{\text{vec}(\Sigma)})(\xi) d\xi.$$

The components of (3.1) and (3.2) are of the form (2.5) for fixed t and (U, v) with

$$q_i = \sum_{j=1}^n (\xi - \mu)_j \sigma^{ji} \quad \text{and} \quad q_{kl} = \frac{1}{2} \left[\sum_{i,j=1}^n (\xi - \mu)_i (\xi - \mu)_j \sigma^{ik} \sigma^{lj} - \sigma^{kl} \right],$$

respectively, and are easily seen to satisfy (2.4). Hence the partial derivatives are continuous throughout \mathcal{H} , and (3.1) and (3.2) are in fact Fréchet differentials of g and L with respect to μ and $\text{vec}(\Sigma)$. The Fréchet differential of G at the optimal point x^* applied to $x = (\mu, \text{vec}(\Sigma), \text{vec}(U), v)$ is then given by

$$\begin{aligned} \nabla G(x^*)(x) = & - \int_S [(\xi - \mu^*(T))' \Sigma^*(T)^{-1} \mu(T) \\ & + \frac{1}{2} (\text{vec}(\Sigma^*(T)^{-1}(\xi - \mu^*(T))(\xi - \mu^*(T))' \Sigma^*(T)^{-1} \\ & - \Sigma^*(T)^{-1})' \text{vec}(\Sigma(T))] dN(\mu^*(T), \Sigma^*(T); \xi), \end{aligned}$$

and

$$\begin{aligned}
 & L_{\mu, \text{vec}(\Sigma)}^*(t)(\mu(t)', \text{vec}(\Sigma(t))')' \\
 &= \int l(t, \xi, U^*(t)\xi + v^*(t))[(\xi - \mu^*(t))'\Sigma^*(t)^{-1}\mu(t) \\
 &\quad + \frac{1}{2}(\text{vec}(\Sigma^*(t)^{-1}(\xi - \mu^*(t))(\xi - \mu^*(t))'\Sigma^*(t)^{-1} \\
 &\quad - \Sigma^*(t)^{-1})'\text{vec}(\Sigma(t))]\,dN(\mu^*(t), \Sigma^*(t); \xi).
 \end{aligned}$$

Applying the multiplier rule cited above we obtain right continuous row vector valued functions $p(t) = (p_1(t), \dots, p_n(t))$ and $q(t) = (q_1(t), \dots, q_{n^2}(t))$ of bounded variation, $p(T) = 0$, $q(T) = 0$, and nonnegative numbers r, s such that $(p, q, r, s) \neq 0$, $s(1 - \alpha) = s$, $g(\mu^*(T), \Sigma^*(T))$ and

$$\begin{aligned}
 & \int_0^T dp(t) \left\{ \mu(t) - \int_0^t [(A(s) + B(s)U^*(s))\mu(s) + B(s)U(s)\mu^*(s) + B(s)v(s)]\,ds \right\} \\
 & + \int_0^T dq(t) \left\{ \text{vec}(\Sigma(t)) - \int_0^t [[I \otimes (A(s) + B(s)U^*(s)) \right. \\
 & \quad + (A(s) + B(s)U^*(s)) \otimes I] \text{vec}(\Sigma(s)) + (I \otimes B(s)U(s) \\
 & \quad \left. + B(s)U(s) \otimes I) \text{vec}(\Sigma^*(s))]\,ds \right\} \\
 (3.3) \quad & + r \int_0^T [L_{\mu, \text{vec}(\Sigma)}^*(t)(\mu(t)', \text{vec}(\Sigma(t))')' \\
 & \quad + L(t, \mu^*(t), \Sigma^*(t), U^*(t) + U(t), v^*(t) + v(t)) \\
 & \quad - L(t, \mu^*(t), \Sigma^*(t), U^*(t), v^*(t))]\,dt \\
 & + s \nabla G(\mu^*, \text{vec}(\Sigma^*), \text{vec}(U^*), v^*)(\mu, \text{vec}(\Sigma), \text{vec}(U), v) \geq 0
 \end{aligned}$$

for all $(\mu, \text{vec}(\Sigma), \text{vec}(U), v)$ with $(U, v) \in \mathcal{F} - (U^*, v^*)$. Integrating by parts and setting $(U, v) = 0$ we obtain

$$\begin{aligned}
 & \int_0^T \left\{ \left[p(t)(A(t) + B(t)U^*(t)) + r \int l(t, \xi, U^*(t)\xi + v^*(t)) \right. \right. \\
 (3.4) \quad & \left. \left. \times (\xi - \mu^*(t))'\Sigma^*(t)^{-1} dN(\mu^*(t), \Sigma^*(t); \xi) \right] \mu(t) - p(t)\dot{\mu}(t) \right\} dt = 0
 \end{aligned}$$

for all absolutely continuous μ satisfying $\mu(0) = \mu(T) = 0$ and

$$\begin{aligned}
 & \int_0^T \left\{ [q(t)(I \otimes (A(t) + B(t)U^*(t)) + (A(t) + B(t)U^*(t)) \otimes I) \right. \\
 (3.5) \quad & \left. + \frac{r}{2} \int l(t, \xi, U^*(t)\xi + v^*(t))(\text{vec}(\Sigma^*(t)^{-1}(\xi - \mu^*(t))(\xi - \mu^*(t))'\Sigma^*(t)^{-1} \right. \\
 & \left. - \Sigma^*(t)^{-1})'\,dN(\mu^*(t), \Sigma^*(t); \xi)] \text{vec}(\Sigma(t)) - q(t) \text{vec}(\dot{\Sigma}(t)) \right\} dt = 0
 \end{aligned}$$

for all absolutely continuous Σ satisfying $\Sigma(0) = \Sigma(T) = 0$. By the fundamental lemma

in the calculus of variation (see [10]), (3.4) and (3.5) imply

$$(3.6) \quad \begin{aligned} \dot{p}(t) = & -p(t)[A(t) + B(t)U^*(t)] \\ & - r \int l(t, \xi, U^*(t)\xi + v^*(t))(\xi - \mu^*(t))' \Sigma^*(t)^{-1} dN(\mu^*(t), \Sigma^*(t); \xi) \end{aligned}$$

and

$$(3.7) \quad \begin{aligned} \dot{q}(t) = & -q(t)[I \otimes (A(t) + B(t)U^*(t)) + (A(t) + B(t)U^*(t)) \otimes I] \\ & - \frac{r}{2} \int l(t, \xi, U^*(t)\xi + v^*(t))(\text{vec}(\Sigma^*(t)^{-1}(\xi - \mu^*(t)) \\ & \times (\xi - \mu^*(t))' \Sigma^*(t)^{-1} - \Sigma^*(t)^{-1}))' dN(\mu^*(t), \Sigma^*(t); \xi). \end{aligned}$$

If we go back to matrix notation, (3.7) may be written in the form

$$(3.8) \quad \begin{aligned} \dot{Q}(t) = & -[A(t)' + U^*(t)'B(t)']Q(t) - Q(t)[A(t) + B(t)U^*(t)] \\ & - \frac{r}{2} \int l(t, \xi, U^*(t)\xi + v^*(t))[\Sigma^*(t)^{-1}(\xi - \mu^*(t)) \\ & \times (\xi - \mu^*(t))' \Sigma^*(t)^{-1} - \Sigma^*(t)^{-1}] dN(\mu^*(t), \Sigma^*(t); \xi), \end{aligned}$$

where $Q(t)$ is the matrix such that $q(t)' = \text{vec}(Q(t))$. Setting $(\mu, \Sigma) = 0$ in (3.3), we find that

$$\begin{aligned} & \int_0^T \{p(t)B(t)[U(t)\mu^*(t) + v(t)] + q(t)[I \otimes B(t)U(t) + B(t)U(t) \otimes I] \\ & \quad \times \text{vec}(\Sigma^*(t)) + rL(t, \mu^*(t), \Sigma^*(t), U(t), v(t))\} dt \\ & \cong \int_0^T \{p(t)B(t)[U^*(t)\mu^*(t) + v^*(t)] + q(t)[I \otimes B(t)U^*(t) + B(t)U^*(t) \otimes I] \\ & \quad \times \text{vec}(\Sigma^*(t)) + rL(t, \mu^*(t), \Sigma^*(t), U^*(t), v^*(t))\} dt \end{aligned}$$

for all $(U, v) \in \mathcal{F}$. If we assume in addition that the sets $\mathcal{U}(t)$ and $\mathcal{V}(t)$ do not depend on time, i.e. that $\mathcal{U}(t) = \mathcal{U}$ and $\mathcal{V}(t) = \mathcal{V}$ for all $t \in [0, T]$, a pointwise maximum principle can be obtained (compare [3]). Performing some straightforward calculations on the Kronecker product we find that

$$(3.9) \quad \begin{aligned} & p(t)B(t)[U\mu^*(t) + v] \\ & \quad + \text{tr}[Q(t)(B(t)U\Sigma^*(t) + \Sigma^*(t)U'B(t)')] \\ & \quad + r \int l(t, \xi, U\xi + v) dN(\mu^*(t), \Sigma^*(t); \xi) \\ & \cong p(t)B(t)[U^*(t)\mu^*(t) + v^*(t)] \\ & \quad + \text{tr}[Q(t)(B(t)U^*(t)\Sigma^*(t) + \Sigma^*(t)U^*(t)'B(t)')] \\ & \quad + r \int l(t, \xi, U^*(t)\xi + v^*(t)) dN(\mu^*(t), \Sigma^*(t); \xi) \end{aligned}$$

a.e. on $[0, T]$ for all $(U, v) \in \mathcal{U} \times \mathcal{V}$. Finally, inserting functions μ_j^ε of the form

$$\mu_{j,i}^\varepsilon(t) = \begin{cases} \frac{1}{\varepsilon}(t - T + \varepsilon) & \text{for } T - \varepsilon \leq t \leq T, \\ 0 & \text{for } 0 \leq t \leq T - \varepsilon, \end{cases}$$

$$\mu_{j,i}^\varepsilon \equiv 0 \quad \text{for } i \neq j$$

and similar functions for $\text{vec}(\Sigma)$ into (3.3) and passing to the limit $\varepsilon \searrow 0$, it turns out that

$$(3.10) \quad p(T) = -s \int_S (\xi - \mu^*(T))' \Sigma^*(T)^{-1} dN(\mu^*(T), \Sigma^*(T); \xi),$$

$$(3.11) \quad Q(T) = -s \int_S [\Sigma^*(T)^{-1} (\xi - \mu^*(T)) \times (\xi - \mu^*(T))' \Sigma^*(T)^{-1} - \Sigma^*(T)^{-1}] dN(\mu^*(T), \Sigma^*(T); \xi),$$

where p and Q have been redefined to be equal to the left-handed limit at T .

Let us collect these results in

THEOREM 2. *Assume that*

- (i) (A) holds,
- (ii) \mathcal{U} and \mathcal{V} are convex sets with nonempty interior,
- (iii) l satisfies condition (E) and Σ_0 is positive definite or
- (iii') l is a polynomial in x and u which is convex in u and has coefficients varying continuously with time.

Then a necessary condition for $(\mu^, \Sigma^*, U^*, v^*)$ to be an optimal solution of (P') is the existence of absolutely continuous functions p, Q and of nonnegative numbers r, s , $(p, Q, r, s) \neq 0$, $s(1 - \alpha) = s$, $g(\mu^*(T), \Sigma^*(T))$, such that the adjoint equations (3.6), (3.8), the maximum principle (3.9) and the transversality conditions (3.10), (3.11) hold.*

If A, B, C and l are continuously differentiable with respect to t it can be shown by means of a time transformation (compare [9]) that condition (ii) is dispensable.

Under assumptions (A), (B) and (E), problems (P) and (P') are equivalent, hence the above conditions are also necessary for (P).

Acknowledgment. The author wishes to thank the referees whose comments helped to reduce some of the proofs and to improve the paper.

Note added in proof. If Σ_0 is positive the rank condition on $C(t)$ can be dispensed with.

REFERENCES

- [1] L. ARNOLD, *Stochastische Differentialgleichungen*, Oldenbourg Verlag, Munich, W. Germany, 1973.
- [2] V. E. BENES, *Existence of optimal stochastic control laws*, this Journal, 9 (1971), pp. 446–472.
- [3] L. D. BERKOVITZ, *Optimal Control Theory*, Springer-Verlag, New York, 1974.
- [4] A. CHARNES AND W. W. COOPER, *Chance-constrained programming*, Management Sci., 6 (1959/60), pp. 73–79.
- [5] ———, *Deterministic equivalents for optimizing and satisficing under chance-constraints*, Operations Res., 11 (1963), pp. 18–39.
- [6] N. CHRISTOPEIT, *Necessary optimality conditions with application to a variational problem*, this Journal, 15 (1977), pp. 683–698.
- [7] T. DUNCAN AND P. VARAIYA, *On the solutions of a stochastic control system*, this Journal, 9 (1971), pp. 354–371.

- [8] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.
- [9] I. V. GIRSANOV, *Lectures on Mathematical Theory of Extremum Problems*, Springer-Verlag, New York, 1972.
- [10] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.
- [11] H. J. KUSHNER, *Existence Results for Optimal Stochastic Controls*, J. Optimization Theory Appl., 15 (1976), pp. 347–360.
- [12] H. THEIL, *Principles of Econometrics*, John Wiley, London, 1971.

LIE ALGEBRAS OF VECTOR FIELDS AND LOCAL APPROXIMATION OF ATTAINABLE SETS*

HENRY HERMES†

Abstract. Consider an analytic, n -dimensional control system described by $dx/dt = X(x) + u(t)Y(x)$, $x(0) = p$ and let $\mathcal{A}(t, p)$ denote the set of states attainable at time t by use of all admissible control functions u , which we take as measurable functions with values $|u(t)| \leq 1$. Our goal is to derive second order conditions to determine if the reference solution, $(\exp tX)(p)$, corresponding to $u(t) \equiv 0$, lies on the boundary or interior of $\mathcal{A}(t, p)$ for small $t > 0$. If $t_1 > 0$ and $p^1 = (\exp t_1 X)(p)$ the approach is to use the Campbell-Baker-Hausdorff formula to obtain second order tangent vectors to $\mathcal{A}(t_1, p)$ at p^1 . These involve elements of the Lie algebra generated by X and Y having an arbitrary number of X factors and two Y factors. With certain hypotheses, for $n = 2, 3$ relatively complete results are obtained.

Introduction. Let M be a real, analytic, n -dimensional manifold, TM_p its tangent space at p , and V the linear space of analytic vector fields on M considered as a Lie algebra with product the Lie product $[X, Y]$, $X, Y \in V$. For $\mathcal{C} \subset V$ we let $L(\mathcal{C})$ denote the Lie subalgebra generated by the elements of \mathcal{C} and $L(\mathcal{C})(p) \subset TM_p$ denote the elements of $L(\mathcal{C})$ evaluated at p . For real s , $(\exp sX)(p)$ denotes the solution, evaluated at $t = s$, of the differential equation $dx/dt = X(x)$, $x(0) = p$, or equivalently, the solution at time $t = 1$, of $dx/dt = sX(x)$, $x(0) = p$. One may also view $(\exp sX)(\cdot) : M \rightarrow M$ as a diffeomorphism.

For l a positive integer, q_i a positive rational and c_i real, an expression of the form

$$(1) \quad \sum_{i=1}^l c_i t^{q_i}$$

will be called a *rational polynomial*. If all $c_i \geq 0$ we will call this a *positive rational polynomial*. To illustrate the type of questions considered, let $\varepsilon > 0$, and $X, Y \in V$ be given. Define $\mathcal{E} \subset L(X, Y)$ to consist of those vector fields W for which there exists an integer k and rational polynomials $r_1, \dots, r_k, s_1, \dots, s_k : [0, \varepsilon) \rightarrow \mathbb{R}^1$ such that

$$(2) \quad (\exp r_k(t)X) \circ (\exp s_k(t)Y) \circ (\exp r_{k-1}(t)X) \circ \dots \circ (\exp r_1(t)X) \\ \circ (\exp s_1(t)Y) = \exp(tW + o(t))$$

where $o(t)$ denotes terms which approach zero faster than t , as $t \rightarrow 0$. The problem is to characterize \mathcal{E} . Since the rational polynomials r_j, s_j may take both positive and negative values, it is not difficult to show (and will be shown in § 1) that $\mathcal{E} = L(X, Y)$. This may loosely be interpreted as stating that $L(X, Y)$ may be considered as the Lie algebra of the group of diffeomorphisms with elements given by the left side of (2) if this group is provided with the proper analytic structure.

A more interesting, and difficult, question is to characterize $\mathcal{E}_X^+ \subset L(X, Y)$ where $W \in \mathcal{E}_X^+$ implies that for some integer k , there exist positive rational polynomials $\sigma_{2k}, r_{2k}, \dots, \sigma_1, r_1$ such that

$$(3) \quad (\exp \sigma_{2k}(t)X) \circ (\exp r_{2k}(t)(X + Y)) \circ (\exp \sigma_{2k-1}(t)X) \circ (\exp r_{2k-1}(t)(X - Y)) \\ \circ \dots \circ (\exp \sigma_1(t)X) \circ (\exp r_1(t)(X - Y)) \circ \exp \left(- \sum_{j=1}^{2k} (\sigma_j(t) + r_j(t))X \right) \\ = \exp(tW + o(t)).$$

* Received by the editors May 16, 1977 and in revised form August 22, 1977.

† Department of Mathematics, University of Colorado, Boulder, Colorado 80309. This research was supported by the National Science Foundation under Grant MCS 76-04419.

To see the application of this to control, consider a control system on M of the form

$$(4) \quad dx/dt = X(x) + u(t)Y(x), \quad x(0) = p \in M$$

where an admissible control u is piecewise continuous with $|u(t)| \leq 1$. For $t_1 > 0$ we let $\mathcal{A}(t_1, p)$ denote the "attainable set" at time t_1 , i.e., $y \in \mathcal{A}(t_1, p)$ implies there is a control u for which the corresponding solution of (4), evaluated at time t_1 , is y . In controllability problems (optimization problems) one must know if the reference solution, which we assume corresponds to $u \equiv 0$, has value $p^1 = (\exp t_1 X)(p)$ in the interior of $\mathcal{A}(t_1, p)$, denoted $\text{int } \mathcal{A}(t_1, p)$ (on the boundary of $\mathcal{A}(t_1, p)$, denoted $\partial \mathcal{A}(t_1, p)$). Suppose $W \in \mathcal{E}_X^+$, i.e., say (3) holds, and denote the left side of (3) by $q(t)$. Then a moments reflection shows that $q(t)p^1 \in \mathcal{A}(t_1, p)$ if $t \geq 0$ and satisfies $0 \leq \sum_1^{2k} (\sigma_j(t) + r_j(t)) < t_1$. Since $q(0)p^1 = p^1$ and

$$\lim_{t \rightarrow 0} \frac{dq(t)p^1}{dt} = W(p^1),$$

we can interpret $W(p^1)$ as a tangent vector to $\mathcal{A}(t_1, p)$ at p^1 . We shall show \mathcal{E}_X^+ is convex, (a result obtained by Krener in [1]) hence if elements of the form $\sum \alpha_i W^i(p^1)$, $\alpha_i \geq 0$, $W^i \in \mathcal{E}_X^+$ generate \mathbb{R}^n , one can conclude $p^1 \in \text{int } \mathcal{A}(t_1, p)$. In general $W \in \mathcal{E}_X^+$ does not mean $-W \in \mathcal{E}_X^+$. Also, it is important to note that since the "perturbation data" $\sigma_j(t)$, $r_j(t) \rightarrow 0$ as $t \rightarrow 0$, the perturbation described in (3) can be made at any point $\exp t_1 X$, $t_1 > 0$, along the X trajectory. The resulting tangent vector W may be viewed as a tangent vector based at $(\exp t_1 X)p$ resulting from a perturbation made at time t_1 .

Notationally, let $(\text{ad } X, Y) = [X, Y]$ and inductively $(\text{ad}^k X, Y) = [X, (\text{ad}^{k-1} X, Y)]$. In § 2 we give a simple proof that

$$(5) \quad \mathcal{S}^1 = \{\pm(\text{ad}^j X, Y) : j = 0, 1, 2, \dots\}$$

is contained in \mathcal{E}_X^+ . This result was obtained by Krener, [1], in an ingenious (but more difficult) manner. The elements in $\text{span } \mathcal{S}^1(p^1)$ therefore are tangent vectors to $\mathcal{A}(t_1, p)$ at p^1 and are, in fact, those tangent vectors which are obtained for the system (4) by use of "Pontryagin–McShane variations", i.e. $\text{span } \mathcal{S}^1(p^1)$ is the "first order" tangent cone to $\mathcal{A}(t_1, p)$ at p^1 as shown in [2].

In § 3 we make the simplifying assumption that $X(p) = 0$ and $\dim \text{span } \mathcal{S}^1(p) = n - 1$. The main result is Theorem 1 which yields a computable sufficient condition that $p \in \text{int } \mathcal{A}(t, p)$ for all $t > 0$ and a necessary condition that $p \in \partial \mathcal{A}(t_1, p)$ for small $t_1 > 0$. The condition is "second order" and related to the high order maximal principle of [1] but obtained in a different manner. For $\dim n = 2$ computations are simple and results agree with those obtained in [3]. With $n = 3$, computational difficulties increase; however several general results are obtained. For the system (4) with $n = 3$, $X(p) = 0$ and $\dim \text{span } \mathcal{S}^1(p) = 2$, we show (see Propositions 3.5, 3.7 and 3.8 for specific hypotheses) that if the smallest integer r such that $[(\text{ad}^r X, Y), Y](p) \notin \text{span } \mathcal{S}^1(p)$ is 1 or 3, the necessary condition that $p \in \partial \mathcal{A}(t, p)$ for small $t > 0$ is satisfied. However, if $r = 5$, $p \in \text{int } \mathcal{A}(t, p) \forall t > 0$. These results illustrate the applications of Theorem 1 and also the difficulty which can be expected in a general theory.

1. Characterization of \mathcal{E} . For X, W analytic vector fields on M , the Baker–Campbell–Hausdorff formulae state

$$(1.1) \quad (\exp t_1 X)(\exp t_2 W)(\exp -t_1 X) = \exp \xi(X, W),$$

where

$$\begin{aligned} \xi(X, W) &= t_2 \sum_{\nu=0}^{\infty} \left(\frac{(t_1)^\nu}{\nu!} \right) (\text{ad}^\nu X, W), \\ (1.2) \quad (\exp t_1 X)(\exp t_2 W) &= \exp \eta(X, W), \end{aligned}$$

where

$$\begin{aligned} \eta(X, W) &= t_1 X + t_2 W + \left(\frac{(t_1 t_2)}{2} \right) [X, W] + \left(\frac{(t_1 t_2^2)}{12} \right) [[X, W], W] \\ &\quad - \left(\frac{(t_1^2 t_2)}{12} \right) [[X, W], X] - \left(\frac{(t_1^2 t_2^2)}{24} \right) [X, [W, [X, W]]] + \dots \end{aligned}$$

To illustrate the use of (1.1), (1.2), we show $[X, Y] \in \mathcal{E}$. Indeed, in the expression (2) choose $k = 2$, $r_2(t) = \sqrt{t}$, $s_2(t) = \sqrt{t}$, $r_1(t) = -\sqrt{t}$, $s_1(t) = -\sqrt{t}$ obtaining $(\exp \sqrt{t} X)(\exp \sqrt{t} Y)(\exp -\sqrt{t} X)(\exp -\sqrt{t} Y) = \exp(t[X, Y] + o(t))$, a well known result, which shows $[X, Y] \in \mathcal{E}$.

To further illustrate how (1.1) and (1.2) will be used, we shall show $\mathcal{E} = L(X, Y)$. This is an equivalent statement to Theorem 1 of Brockett, [4], and as remarked in this reference, closely related to a basic theorem of Chow.

PROPOSITION 1.1. \mathcal{E} is convex.

Proof. Let $\alpha \in [0, 1]$ and $W^1, W^2 \in \mathcal{E}$, say $(\exp r_{k_1}^1(t)X) \circ \dots \circ (\exp s_1^1(t)Y) = \exp(tW^1 + o(t))$ and $(\exp r_{k_2}^2(t)X) \circ \dots \circ (\exp s_1^2(t)Y) = \exp(tW^2 + o(t))$ with the $r_j^1, s_j^1, r_j^2, s_j^2$ all being rational polynomials. Then

$$\begin{aligned} (\exp r_{k_1}^1(\alpha t)X) \circ \dots \circ (\exp s_1^1(\alpha t)Y) \circ \exp(r_{k_2}^2((1-\alpha)t)X) \circ \dots \circ \exp(s_1^2((1-\alpha)t)Y) \\ = \exp(\alpha t W^1 + o(t)) \circ \exp((1-\alpha)t W^2 + o(t)) \\ = \exp(\alpha t W^1 + (1-\alpha)t W^2 + o(t)); \end{aligned}$$

hence $(\alpha W^1 + (1-\alpha)W^2) \in \mathcal{E}$. \square

PROPOSITION 1.2. $\mathcal{E} = L(X, Y)$.

Proof. First note that for any real α, β , $(\exp \alpha t X)(\exp \beta t Y) = \exp(\alpha t X + \beta t Y + o(t))$; hence $\alpha X + \beta Y \in \mathcal{E}$, specifically $-X, -Y \in \mathcal{E}$.

We will show $W \in \mathcal{E}$ implies $[X, W] \in \mathcal{E}$, and hence also $-[X, W], \pm[Y, W] \in \mathcal{E}$. This means $W \in \mathcal{E}$ implies any product of the form $\pm(\text{ad}^{k_n} Y, (\text{ad}^{k_{n-1}} X, (\dots, (\text{ad}^{k_2} Y, (\text{ad}^{k_1} X, W) \dots)))$ belongs to \mathcal{E} . Dynkin, [5], shows elements of this form span $L(X, Y)$; hence from proposition 1.1, we obtain the desired result.

Suppose $W \in \mathcal{E}$, i.e. there exist rational polynomials $r_1, \dots, r_k, s_1, \dots, s_k$ such that $(\exp r_k(t)X) \circ \dots \circ (\exp s_1(t)Y) = \exp(tW + O(t^\alpha))$ where we may assume $\alpha > 1$ is rational. Choose β rational, $0 < \beta < 1$, $\alpha\beta > 1$. Then

$$\begin{aligned} (\exp t^{1-\beta} X) \exp(t^\beta W + O(t^{\alpha\beta})) \circ \exp(-t^{1-\beta} X) \circ \exp(-t^\beta W + O(t^{\alpha\beta})) \\ = \exp(t^\beta \sum_{\nu=0}^{\infty} ((t^{1-\beta})^\nu / \nu!) (\text{ad}^\nu X, W) + O(t^{\alpha\beta})) \circ \exp(-t^\beta W + O(t^{\alpha\beta})) \\ = \exp(t^\beta W + t[X, W] + (t^{2-\beta}/2)(\text{ad}^2 X, W) + \dots) \circ \exp(-t^\beta W + O(t^{\alpha\beta})) \\ = \exp(t[X, W] + (t^{2-\beta}/2)(\text{ad}^2 X, Y) + \dots + (t^2/2)[W, (\text{ad}^2 X, W)] + \dots + O(t^{\alpha\beta})) \\ = \exp(t[X, W] + O(t^\gamma)) \end{aligned}$$

where $\gamma > 1$, specifically $\gamma = \min\{\alpha, \beta, 2-\beta\}$ showing $[X, W] \in \mathcal{E}$. \square

An argument of the type used to show $W \in \mathcal{E}$ implies $[X, W] \in \mathcal{E}$ will be needed several times. When convenient, we shall merely refer to the method used above, rather than repeat it.

2. The set \mathcal{E}_X^+ . As mentioned in the Introduction, the set \mathcal{E}_X^+ has implications on problems of controllability and optimality associated with the control system (4). Notationally, for k an arbitrary integer, $\sigma_{2k}, \dots, \sigma_1, r_{2k}, \dots, r_1$ positive rational polynomials, we denote by π the “perturbation data” $\{\sigma_{2k}, \dots, \sigma_1, r_{2k}, \dots, r_1\}$ and let $\alpha(\pi, t)$ denote the corresponding left side of (3). We shall call α an *admissible variation* and if $W \in \mathcal{E}_X^+$, there will exist an admissible variation, denoted $\alpha_W(\pi, s)$ such that $\alpha_W(\pi, s) = \exp(sW + o(s))$.

Let $p^1 = T^X(t_1)p$ where $t_1 > 0$. Since all $r_i(s), \sigma_i(s) \rightarrow 0$ as $s \rightarrow 0$, for sufficiently small $s \geq 0$, $\alpha_W(\pi, s)p^1 \in \mathcal{A}(t_1, p)$ and $\lim_{s \rightarrow 0} (d/ds)\alpha_W(\pi, s)p^1 = W(p^1)$ showing $W(p^1)$ is a tangent vector to $\mathcal{A}(t_1, p)$ at p^1 . We shall refer to $\alpha_W(\pi, s)$ as an *admissible variation with tangent vector W* .

Much of this section amounts to rederiving results obtained by Krener in [1]. Our method is to use fractional powers of the independent variable (usually denoted s), the Campbell–Baker–Hausdorff formulae and first order derivatives of variations. Krener uses integer powers of the independent variable and higher derivatives. The methods differ only in computational ease.

LEMMA 2.1 (Lemma 3.3, [1]). *Let $W \in \mathcal{E}_X^+$ and $\alpha_W(\pi, s) = \exp(sW + o(s))$. Then for any positive rational polynomial $\mu(s)$,*

$$(*) \quad (\exp \mu(s)X)\alpha_W(\pi, s)(\exp -\mu(s)X) = \exp(sW + o(s)).$$

Proof. From (1.1)

$$\begin{aligned} (\exp \mu(s)X) \circ \exp(sW + o(s)) \circ (\exp -\mu(s)X) &= \exp\left(s \sum_{\nu=0}^{\infty} (\mu(s))^\nu / \nu! (\text{ad}^\nu X, W) + o(s)\right) \\ &= \exp(sW + o(s)). \quad \square \end{aligned}$$

In words, this states that if the perturbation data π , in $\alpha_W(\pi, s)$, was $\{\sigma_j, r_j : j = 1, \dots, 2k\}$ and we define new perturbation data π' by merely replacing σ_{2k} with $\sigma'_{2k} = \sigma_{2k} + \mu$, then the left side of (*) is merely $\alpha(\pi', s)$ and $\alpha(\pi', s)$ is an admissible variation which also has W as a tangent vector. The introduction of $\mu(s)$ will allow us to form the composition of admissible variations.

LEMMA 2.2. *If $W^1, W^2 \in \mathcal{E}_X^+$ and $\beta, \gamma \geq 0$ then $(\beta W^1 + \gamma W^2) \in \mathcal{E}_X^+$.*

Proof. Let $\alpha_{W^1}(\pi, s)$ be an admissible variation with tangent vector W^1 ; $\alpha_{W^2}(\pi^2, s)$ an admissible variation with tangent vector W^2 and $\pi^2 = \{\sigma_j^2, r_j^2 : j = 1, \dots, 2k\}$ the perturbation data. Define $\mu(s) = \sum_{i=1}^{2k} (r_i^2(s) + \sigma_i^2(s))$. If $\pi = \{\sigma_j, r_j : j = 1, \dots, 2h\}$ is the perturbation data of $\alpha_{W^1}(\pi, s)$, let π^1 be obtained from π by replacing $\sigma_{2h}(s)$ with $\sigma_{2h}(s) + \mu(s)$. By the remarks following lemma 2.1, $\alpha(\pi^1, s)$ is an admissible variation with tangent vector W^1 , which we now denote $\alpha_{W^1}(\pi^1, s)$. Then $\alpha_{W^2}(\pi^2, s) \circ (\exp \mu(s)X) \circ \alpha_{W^1}(\pi, s) \circ \exp(-\mu(s)X) = \alpha_{W^2}(\pi^2, s) \circ \alpha_{W^1}(\pi^1, s)$ is an admissible variation, i.e., the $(\exp \mu(s)X)$ assures $\alpha_{W^2}(\pi^2, s) \circ (\exp \mu(s)X)$ does not involve traversing a trajectory “backwards” in time (equivalently the perturbation data of the composition $\alpha_{W^2}(\pi^2, s) \circ \alpha_{W^1}(\pi^1, s)$ consists of positive, rational polynomials). Now $\alpha_{W^2}(\pi^2, s) \circ \alpha_{W^1}(\pi^1, s) = \exp(sW^2 + o(s)) \circ \exp(sW^1 + o(s)) = \exp(s(W^1 + W^2) + o(s))$ while $\alpha_{W^2}(\pi^2, \gamma s) \circ \alpha_{W^1}(\pi^1, \beta s) = \exp(s(\beta W^1 + \gamma W^2) + o(s))$ showing $(\beta W^1 + \gamma W^2) \in \mathcal{E}_X^+$. \square

COROLLARY 2.1 (Lemma 3.4, [1]). \mathcal{E}_X^+ is convex.

PROPOSITION 2.1. *If $\pm W \in \mathcal{E}_X^+$ then $\pm(\text{ad}^j X, W) \in \mathcal{E}_X^+$ for $j = 0, 1, \dots$.*

Proof. (Induction on j .) For $j = 0$ we have hypothesized $\pm W \in \mathcal{E}_X^+$. Now assume $\pm(\text{ad}^m X, W) \in \mathcal{E}_X^+$; for notational ease let W^+, W^- denote respectively, $(\text{ad}^m X, W)$,

$-(\text{ad}^m X, W)$ and let $\alpha_{W^+}(\pi^+, s)$, $\alpha_{W^-}(\pi^-, s)$ be admissible variations with tangent vectors W^+ , W^- respectively. Then there exists a rational $\alpha > 1$ such that $\alpha_{W^\pm}(\pi^\pm, s) = \exp(\pm s(\text{ad}^m X, W) + O(s^\alpha))$. By Lemma 2.1, we may with no loss of generality, assume the composition $\alpha_{W^+}(\pi^+, s) \circ \alpha_{W^-}(\pi^-, s)$ is an admissible variation. Now let β be a rational such that $0 < \beta < 1$, $\alpha\beta > 1$. Then

$$\begin{aligned}
 & \alpha_{W^+}(\pi^+, s^\beta) \exp(s^{1-\beta} X) \alpha_{W^-}(\pi^-, s^\beta) \exp(-s^{1-\beta} X) \\
 &= \exp(s^\beta (\text{ad}^m X, W) + O(s^{\alpha\beta})) \circ \exp(s^{1-\beta} X) \\
 (2.1) \quad & \circ \exp(-s^\beta (\text{ad}^m X, W) + O(s^{\alpha\beta})) \circ \exp(-s^{1-\beta} X) \\
 &= \exp(s(\text{ad}^{m+1} X, Y) + o(s)),
 \end{aligned}$$

with the last step following as in the proof of Proposition 1.2. Since α_{W^+} , α_{W^-} compose, clearly the left side of (2.1) is an admissible variation with tangent vector $(\text{ad}^{m+1} X, W)$. Interchanging α_{W^+} , α_{W^-} in (2.1) gives $-(\text{ad}^{m+1} X, W) \in \mathcal{E}_X^+$ and the induction is complete. \square

COROLLARY 2.2 [1, pp. 270–272]. $\pm(\text{ad}^j X, Y) \in \mathcal{E}_X^+$ for $j = 0, 1, \dots$.

Proof. We note $\exp(s(X \pm Y)) \exp(-sX) = \exp(\pm sY + o(s))$. The left side is an admissible variation showing $\pm Y \in \mathcal{E}_X^+$. The result now follows from Proposition 2.1. \square

COROLLARY 2.3. With \mathcal{S}^1 as defined in (5), $\text{span } \mathcal{S}^1 \subset \mathcal{E}_X^+$.

Example 2.1. $[[X, Y], Y] \in \mathcal{E}_X^+$ but $-[[X, Y], Y] \notin \mathcal{E}_X^+$.

From (1.1) and (1.2) one obtains

$$\begin{aligned}
 (2.2) \quad & \exp t(X - Y) \exp t(X + Y) \exp(-2tX) \\
 &= \exp\left(t^2[X, Y] + \left(\frac{t^3}{3}\right)[[[X, Y], Y] + 2t^3[X, [X, Y]] + o(t^3)\right).
 \end{aligned}$$

Combining (2.2) with the similar result obtained by replacing Y with $-Y$ and using Lemma 2.1 to insert $\exp 2t^{1/3}X$ and $\exp -t^{1/3}X$ without changing tangent vectors generated gives

$$\begin{aligned}
 & \exp t^{1/3}(X - Y) \exp t^{1/3}(X + Y) \exp(-2t^{1/3}X) \exp(2t^{1/3}X) \\
 & \exp t^{1/3}(X + Y) \exp t^{1/3}(X - Y) \exp(-2t^{1/3}X) \exp(-2t^{1/3}X) \\
 &= \exp((2t/3)[[X, Y], Y] + o(t))
 \end{aligned}$$

showing $[[X, Y], Y] \in \mathcal{E}_X^+$.

To see that $-[[X, Y], Y] \notin \mathcal{E}_X^+$ it suffices to produce an example with attainable set to which this cannot be a tangent vector. Let $M = \mathbb{R}^2$, $p = 0$, $X(x) = (x_1, x_2^2)$, $Y = (1, 0)$ where all vectors are written as row vectors for notational ease. Since $\dot{x}_2(t) = x_2^2(t)$, $\dot{x} = dx/dt$, it follows that $\mathcal{A}(t, p)$ lies in the half space $x_2 \geq 0$ for all $t \geq 0$. Pick any $t_1 > 0$, then $p^1 = (\exp t_1 X)(p) = p$. Computing shows $[X, Y](x) = (1, 2x_1)$ so $[X, Y](p^1) = (1, 0)$ is a tangent vector to $\mathcal{A}(t_1, p)$ at p^1 . Next, $-[[X, Y], Y] = (0, -2)$ hence since $\mathcal{A}(t_1, p)$ lies in the half space $x_2 \geq 0$, there cannot be a smooth map $\gamma: [-1, 1] \rightarrow \mathcal{A}(t_1, p)$ with $\gamma(0) = p^1$ and $\dot{\gamma}(0) = (0, -2)$. Since any $W \in \mathcal{E}_X^+$ is a tangent vector to $\mathcal{A}(t_1, p)$ at p^1 , $-[[X, Y], Y] \notin \mathcal{E}_X^+$.

Remarks. 1. An interesting problem, not pursued here, would be to characterize \mathcal{E}_X^+ .

2. One can translate elements of \mathcal{E}_X^+ along the X trajectory, again getting tangent vectors to the attainable set. Specifically, let $0 < t_1 < t_2$, $\gamma = t_2 - t_1$, $p^2 = (\exp t_2 X)(p)$ and $\alpha_W(\pi, s)$ an admissible variation with tangent vector W . Define $\beta(\gamma, \alpha_W, s) = (\exp \gamma X) \circ \alpha_W(\pi, s) \circ \exp(-\gamma W)p^2$. Then $\beta(\gamma, \alpha_W, 0) = p^2$ while for small $s \geq 0$, $\beta(\gamma, \alpha_W, s) \in \mathcal{A}(t_2, p)$; hence, using (1.1), generates a tangent vector

$\lim_{s \rightarrow 0} (d/ds)\beta(\gamma, \alpha_w, s) = \sum_{\nu=1}^{\infty} (\gamma^\nu/\nu!)(\text{ad}^\nu X, W)(p^2)$ to $\mathcal{A}(t_2, p)$ at p^2 . One may view this tangent vector as a “ γ translate” along $\exp tX$, of W .

Let $t_2 > 0$, $p^2 = (\exp t_2 X)(p)$, $0 \leq \gamma_k \leq \gamma_{k-1} \leq \dots \leq \gamma_1 < t_2$, $\alpha_{w^i}(\pi^i, s)$ be an admissible variation with tangent vector W^i and $\beta(\gamma_i, \alpha_{w^i}, s)$ be defined as above. One may easily show that for any constant $c_1, \dots, c_k \geq 0$, $\beta(\gamma_k, \alpha_{w^k}, c_k s) \circ \dots \circ \beta(\gamma_1, \alpha_{w^1}, c_1 s)p^2 \in \mathcal{A}(t_2, p)$ for sufficiently small $s \geq 0$ and generates the tangent vector

$$\sum_{j=1}^k c_j \sum_{\nu=0}^{\infty} \left(\frac{(\gamma_j)^\nu}{\nu!} \right) (\text{ad}^\nu X, W^j)(p^2).$$

This was used in [2, § 1] with $W^j = \pm Y$ to show $\text{span } \mathcal{S}^1(p^2)$ is contained in the tangent cone to $\mathcal{A}(t_2, p)$ at p^2 . In the next section we shall have occasion to make γ translates along $\exp tX$.

3. Higher order tangent vectors. If $\alpha(\pi, s)$ is an admissible variation, $p^1 = (\exp t_1 X)(p)$ with $t_1 > 0$ and $\lim_{s \rightarrow 0} (d/ds)\alpha(\pi, s)(p^1) = 0$ then $\lim_{s \rightarrow 0} (d^2/ds^2) \cdot \alpha(\pi, s)(p^1)$ is a tangent vector to $\mathcal{A}(t_1, p)$ at p^1 . This tangent vector can also be obtained by reparametrizing in α and taking a first derivative, i.e. consider $\alpha(\pi, \sqrt{s})(p^1)$, etc., as illustrated in Example 2.1.

In this section, we will deal with the n -dimensional system (4) with the additional assumption

$$(3.1) \quad X(p) = 0.$$

This assumption, mainly, simplifies computations. Our first goal will be to construct variations $\alpha(\pi, s)$ such that $\lim_{s \rightarrow 0} (d/ds)\alpha(\pi, s)(p) = 0$ thereby making possible the generation of “higher order” tangent vectors. Our notation will be to use \mathcal{S}^2 to consist of elements which are a product of a pair of elements in \mathcal{S}^1 ; thus $W \in \mathcal{S}^2$ implies $W = [(\text{ad}^j X, Y), (\text{ad}^k X, Y)]$, i.e. two factors Y . If $j + k = r$, we may also write $W \in \mathcal{S}^{2,r}$; thus r determines the number of X factors.

Define

$$\begin{aligned} \xi^i(s) &\equiv \xi(s, \gamma_i, d_i, e_i) \equiv (\exp \gamma_i X) \exp (sd_i(X + e_i Y)) (\exp - \gamma_i X) \\ &= \exp \left(sd_i X + sd_i e_i \sum_{\nu=0}^{\infty} \left(\frac{(\gamma_i)^\nu}{\nu!} \right) (\text{ad}^\nu X, Y) \right) \end{aligned}$$

where $s, \gamma_i, d_i \geq 0$, $e_i \in \{-1, 0, 1\}$. Let $t_1 > 0$, k be a positive integer and

$$t_1 > \gamma_1 > \gamma_2 > \dots > \gamma_{3k} > 0.$$

Then if $0 = \sum_{i=1}^k (d_i s + d_{i+k} s^2 + d_{i+2k} s^3) \leq t_1 - \gamma_1$ it follows that

$$\begin{aligned} (3.2) \quad q(s) &= \xi^{3k}(s^3) \circ \dots \circ \xi^{2k+1}(s^3) \circ \xi^{2k}(s^2) \circ \dots \circ \xi^{k+1}(s^2) \circ \xi^k(s) \circ \dots \circ \xi^1(s) \\ &\quad \circ \exp \left(t_1 - \sum_{i=1}^k (d_i s + d_{i+k} s^2 + d_{i+2k} s^3) X \right) p \end{aligned}$$

belongs to $\mathcal{A}(t_1, p)$ for sufficiently small $s \geq 0$ while $q(0) = p$. The assumption $X(p) = 0$ leads to the simplification

$$(3.2)' \quad q(s) = \xi^{3k}(s^3) \circ \dots \circ \xi^{2k+1}(s^3) \circ \xi^{2k}(s^2) \circ \dots \circ \xi^1(s)p.$$

For notational ease, when the product $d_i e_i$ occurs, we will denote it c_i noting that c_i may assume any value in \mathbb{R}^1 .

Let

$$V^i = \left(d_i X + c_i \sum_{\nu=0}^{\infty} \left(\frac{(\gamma_i)^\nu}{\nu!} \right) (\text{ad}^\nu X, Y) \right)$$

so $\xi^i(s) = \exp(sV^i)$. Then, using (1.2),

$$\begin{aligned} q(s) &= \exp(s^3 V^{3k}) \circ \dots \circ \exp(s^3 V^{2k+1}) \circ \exp(s^2 V^{2k}) \circ \dots \\ &\quad \circ \exp(sV^k) \circ \dots \circ \exp(sV^1) p \\ &= \exp \left\{ s \sum_{i=1}^k V^i + s^2 \left(\sum_{i=k+1}^{2k} V^i + \frac{1}{2} \sum_{1 \leq i < j \leq k} [V^i, V^j] \right. \right. \\ (3.2)'' \quad &\quad \left. \left. + s^3 \left(\sum_{i=2k+1}^{3k} V^i + \frac{1}{2} \sum_{1 \leq i, j \leq k} [V^{k+i}, V^j] \right. \right. \right. \\ &\quad \left. \left. + \frac{1}{12} \sum_{1 \leq i < j \leq k} ([V^i, V^j], V^i) - [[V^i, V^j], V^i] \right) \right. \\ &\quad \left. \left. + \frac{1}{4} \sum_{1 \leq i < j < l \leq k} [V^l, [V^i, V^j]] \right) + o(s^3) \right\} p. \end{aligned}$$

Now $q'(0) \equiv \lim_{s \rightarrow 0} dq(s)/ds$ is a tangent vector to $\mathcal{A}(t_1, p)$ at p . If $q'(0) = 0$, then $q''(0)$ is a tangent vector. Our goal is to expand q in powers of s and choose the constants c_i, γ_i to make certain terms (for example the coefficient of s) vanish. Expanding $q(s)$, as given in (3.2)'', to order two in s gives

$$\begin{aligned} q(s) &= \exp \left\{ s \left(\sum_{i=1}^k d_i X + \sum_{i=1}^k c_i \sum_{\nu=0}^{\infty} \left(\frac{(\gamma_i)^\nu}{\nu!} \right) (\text{ad}^\nu X, Y) \right. \right. \\ &\quad \left. \left. + s^2 \left(\sum_{i=k+1}^{2k} d_i X + \sum_{i=k+1}^{2k} c_i \sum_{\nu=0}^{\infty} \left(\frac{(\gamma_i)^\nu}{\nu!} \right) (\text{ad}^\nu X, Y) \right. \right. \right. \\ (3.3) \quad &\quad \left. \left. + \frac{1}{2} \sum_{1 \leq i < j \leq k} d_j c_i \sum_{\nu=0}^{\infty} \left(\frac{(\gamma_i)^\nu}{\nu!} \right) (\text{ad}^{\nu+1} X, Y) \right. \right. \\ &\quad \left. \left. + \frac{1}{2} \sum_{1 \leq i < j \leq k} c_i c_j \sum_{\nu_1=0}^{\infty} \sum_{\nu_2=0}^{\infty} \left(\frac{(\gamma_i)^{\nu_1}}{\nu_1!} \right) \right. \right. \\ &\quad \left. \left. \left(\frac{(\gamma_j)^{\nu_2}}{\nu_2!} \right) [(\text{ad}^{\nu_2} X, Y), (\text{ad}^{\nu_1} X, Y)] + o(s^2) \right) \right\} p. \end{aligned}$$

Our next goal is to simplify the last term on the right in eq. (3.3), specifically the coefficients of elements in \mathcal{S}^2 .

PROPOSITION 3.1. *Let X, W be analytic vector fields and $p \in M$ (we need not have $X(p) = 0$). If $W((\exp tX)(p)) \in \text{span } \mathcal{S}^1((\exp tX)(p))$ for $0 \leq t \leq t_1$, $t_1 > 0$, then $(\text{ad}^\nu X, W)(p) \in \text{span } \mathcal{S}^1(p)$ for $\nu = 0, 1, 2, \dots$.*

Proof. For any $t \in [0, t_1]$, $(\exp tX): M \rightarrow M$ is a diffeomorphism, let $(\exp tX)_*$ denote the induced tangent space isomorphism. First note that

$$(*) \quad (\exp -tX)_*: \text{span } \mathcal{S}^1((\exp tX)(p)) \rightarrow \text{span } \mathcal{S}^1(p).$$

Indeed, for any integer $j \geq 0$,

$$(\exp -tX)_*(\text{ad}^j X, Y)((\exp tX)(p)) = \sum_{\nu=0}^{\infty} \left(\frac{t^\nu}{\nu!} \right) (\text{ad}^{\nu+j} X, Y)(p).$$

Thus $(\exp -tX)_* W((\exp tX)(p)) = \sum (t^\nu/\nu!)(\text{ad}^\nu X, W)(p) \in \text{span } \mathcal{S}^1(p)$ since $W((\exp tX)(p)) \in \text{span } \mathcal{S}^1((\exp tX)(p))$ by hypothesis. Since $t \in [0, t_1]$ is arbitrary, it follows that $(\text{ad}^\nu X, W)(p) \in \text{span } \mathcal{S}^1(p)$, $\nu = 0, 1, \dots$. \square

PROPOSITION 3.2 ([1, pp. 279–280]). *For any integers ν_1, ν_2 with $\nu_1 + \nu_2 = r$ $[(\text{ad}^{\nu_2} X, Y), (\text{ad}^{\nu_1} X, Y)](p) = (-1)^{\nu_1} [(\text{ad}^{\nu_1} X, Y), Y](p) + V(p)$ where $V(p) \in \text{span } \mathcal{S}^1(p)$ and r is the smallest integer such that $\dim \text{span } (\mathcal{S}^1(p) \cup \mathcal{S}^{2,r}(p)) = n$.*

Proof. If $\nu_1 = 0$ there is nothing to prove so assume $\nu_1 \geq 1$. The Jacobi identity yields $[(\text{ad}^{\nu_2} X, Y), (\text{ad}^{\nu_1} X, Y)] = [(\text{ad}^{\nu_2+1} X, Y), (\text{ad}^{\nu_1-1} X, Y)] + [X, [(\text{ad}^{\nu_2} X, Y), (\text{ad}^{\nu_1-1} X, Y)]]$. From the definition of r and Proposition 1, we have $[X, [(\text{ad}^{\nu_2} X, Y), (\text{ad}^{\nu_1-1} X, Y)]](p) \in \text{span } \mathcal{S}^1(p)$. We can now proceed inductively on the term $[(\text{ad}^{\nu_2+1} X, Y), (\text{ad}^{\nu_1-1} X, Y)]$, above, to obtain the result. \square

Returning to the last term on the right in equation (3.3), using Proposition 3.2, and $V(p)$ to represent an arbitrary element in $\text{span } \mathcal{S}^1(p)$, we obtain

$$\begin{aligned}
 & \frac{1}{2} \sum_{1 \leq i < j \leq k} c_i c_j \sum_{m=0}^{\infty} \sum_{\nu_1 + \nu_2 = m} \left(\frac{(\gamma_i)^{\nu_1}}{\nu_1!} \right) \left(\frac{(\gamma_j)^{\nu_2}}{\nu_2!} \right) [(\text{ad}^{\nu_2} X, Y), (\text{ad}^{\nu_1} X, Y)](p) \\
 (3.4) \quad &= \frac{1}{2} \sum_{1 \leq i < j \leq k} c_i c_j \sum_{m=0}^{\infty} \sum_{l=0}^m \left(\frac{(-\gamma_i)^l}{l!} \right) \left(\frac{(\gamma_j)^{m-l}}{(m-l)!} \right) [(\text{ad}^m X, Y), Y](p) + V(p) \\
 &= \left(\frac{1}{2} \sum_{m=1}^{\infty} (1/m!) [(\text{ad}^m X, Y), Y](p) \sum_{1 \leq i < j \leq k} c_i c_j (\gamma_j - \gamma_i)^m \right) + V(p).
 \end{aligned}$$

Using (3.4) to simplify the right side of (3.3) we are now in a position to examine derivatives of the function $q(s)$ as given in (3.3). We first make an additional

Assumption: $\dim \text{span } \mathcal{S}^1(p) = n - 1$ while $\dim \text{span } (\mathcal{S}^1 \cup \mathcal{S}^2)(p) = n$. The integer r will be used throughout to denote the smallest integer such that $\dim \text{span } (\mathcal{S}^1 \cup \mathcal{S}^{2,r})(p) = n$.

Again using $X(p) = 0$ we compute

$$(3.5) \quad q'(0) = \sum_{i=1}^k c_i \sum_{\nu=0}^{\infty} \left(\frac{(\gamma_i)^\nu}{\nu!} \right) (\text{ad}^\nu X, Y)(p).$$

Since $\dim \text{span } \mathcal{S}^1(p) = n - 1$ we may choose, in succession, integers $\nu_1 < \nu_2 < \dots < \nu_{n-1}$, each as small as possible, so that the set $\{(\text{ad}^{\nu_1} X, Y)(p), \dots, (\text{ad}^{\nu_{n-1}} X, Y)(p)\}$ is linearly independent. With this choice equation (3.5) can be rewritten as

$$(3.6) \quad q'(0) = \sum_{i=1}^k c_i \sum_{j=1}^{n-1} \left(\frac{(\gamma_i)^{\nu_j}}{\nu_j!} + o(\gamma_i)^{\nu_j} \right) (\text{ad}^{\nu_j} X, Y)(p).$$

Let $M(\gamma)$ denote the $k \times (n-1)$ matrix with entries

$$(3.7) \quad m_{ij} = (\gamma_i)^{\nu_j} / \nu_j!, \quad i = 1, \dots, k; \quad j = 1, \dots, n-1,$$

and $\hat{M}(\gamma)$ the $k \times (n-1)$ matrix with entries

$$\hat{m}_{ij} = \left(\frac{(\gamma_i)^{\nu_j}}{\nu_j!} + o(\gamma_i)^{\nu_j} \right).$$

Then $M(\gamma)$ has rank $(n-1)$ and for $t_1 > 0$ sufficiently small (hence all $\gamma_i > 0$ are small) $\text{rank } \hat{M}(\gamma) = n - 1$. Thus

PROPOSITION 3.3. *For $k > n - 1$ there exists a nonzero vector $c = (c_1, \dots, c_k)$ such that $c\hat{M}(\gamma) = 0$ (notationally $c \in \mathcal{N}(\hat{M}(\gamma))$, the null space of $\hat{M}(\gamma)$), i.e. such that $q'(0) = 0$.*

Since k represents the number of “switches” from the reference control, without loss of generality we may consider $k > n - 1$ and c chosen so $q'(0) = 0$. Returning to equation (3.3) and using (3.4) we have

$$q''(0) = \sum_{i=k+1}^{2k} c_i \sum_{\nu=0}^{\infty} \left(\frac{(\gamma_i)^\nu}{\nu!} \right) (\text{ad}^\nu X, Y)(p) + V(p; c_1, \dots, c_k, \gamma_1, \dots, \gamma_k) \\ + \frac{1}{2} \sum_{m=1}^{\infty} (1/m!) [(\text{ad}^m X, Y), Y](p) \sum_{1 \leq i < j \leq k} c_i c_j (\gamma_j - \gamma_i)^m$$

where $V(p; c_1, \dots, c_k, \gamma_1, \dots, \gamma_k) \in \text{span } \mathcal{S}^1(p)$. Again, let r be the smallest integer such that $[(\text{ad}^r X, Y), Y](p) \notin \text{span } \mathcal{S}^1(p)$. One may then choose (it will be assumed done) c_{k+1}, \dots, c_{2k} so that for any choice of $c_1, \dots, c_k, \gamma_1, \dots, \gamma_k$,

$$\sum_{i=k+1}^{2k} c_i \sum_{\nu=0}^{\infty} \left(\frac{(\gamma_i)^\nu}{\nu!} \right) (\text{ad}^\nu X, Y)(p) + V(p; c_1, \dots, c_k, \gamma_1, \dots, \gamma_k) \\ + \frac{1}{2} \sum_{m=1}^{r-1} (1/m!) [(\text{ad}^m X, Y), Y](p) \sum_{1 \leq i < j \leq k} c_i c_j (\gamma_j - \gamma_i)^m = 0.$$

Then for t_1 (and hence all γ_i) sufficiently small

$$(3.8) \quad q''(0) = \frac{1}{2} [(\text{ad}^r X, Y), Y](p) \sum_{1 \leq i < j \leq k} (1/r!) c_i c_j (\gamma_j - \gamma_i)^r + o(t_1^r).$$

Let

$$\alpha_{ij} = \begin{cases} \frac{1}{r!} (\gamma_j - \gamma_i)^r, & 1 \leq i < j \leq k \\ 0, & j \leq i \leq k \end{cases}$$

so the coefficient of $[(\text{ad}^r X, Y), Y](p)$ in (3.8) is a quadratic form involving a nonsymmetric matrix with elements α_{ij} . Form the $k \times k$ symmetric matrix $A = A(\gamma)$ with entries

$$(3.9) \quad a_{ij} = \begin{cases} (1/r!) (\gamma_j - \gamma_i)^r, & 1 \leq i < j \leq k \\ (1/r!) (\gamma_i - \gamma_j)^r, & 1 \leq j < i \leq k \end{cases}$$

and let $c = (c_1, \dots, c_k)$ be such that $c\hat{M}(r) = 0$. Then

$$q''(0) = (1/4)(cA(\gamma)c^T)[(\text{ad}^r X, Y), Y](p) + o(t_1^r).$$

THEOREM 1. Assume, for system (4), that $X(p) = 0$, $\dim \text{span } \mathcal{S}^1(p) = n - 1$ and r is the smallest integer such that $\dim \text{span } (\mathcal{S}^1(p) \cup \mathcal{S}^{2,r}(p)) = n$. A sufficient condition that $(\exp tX)(p) = p \in \text{int } \mathcal{A}(t, p) \forall t > 0$ is that given any $t_1 > 0$ there exist $\gamma_1, \dots, \gamma_k$ with $t_1 > \gamma_1 > \gamma_2 > \dots > \gamma_k > 0$ and k -dimensional vectors $c^1, c^2 \in \mathcal{N}(M(\gamma))$ such that $c^1 A(\gamma)(c^1)^T > 0$, $c^2 A(\gamma)(c^2)^T < 0$.

Remark. With the assumptions of Theorem 1, if $(\exp tX)(p) \in \partial \mathcal{A}(t, p)$ for $0 \leq t \leq t_1$ one cannot have both $[(\text{ad}^r X, Y), Y](\exp tX)(p)$ and its negative as tangent vectors to $\mathcal{A}(t, p)$ at $(\exp tX)(p)$. Thus one can establish the existence of a nonzero vector valued function, $\lambda(t)$, which geometrically can be considered as an outer normal to $\mathcal{A}(t, p)$ at $(\exp tX)(p)$, such that the pair $(\exp tX)(p)$, $\lambda(t)$, satisfy the Pontryagin maximum principle while the inner product $\lambda(t)[(\text{ad}^r X, Y), Y](\exp tX)(p) \leq 0$ (or ≥ 0) holds for $0 \leq t \leq t_1$. This is one form of the high order maximal principle; see [1, Theorem 5.1].

Proof. It is well known that if $p \in \text{int } \mathcal{A}(t_1, p)$ for some $t_1 > 0$ then this holds for all $t \geq t_1$. We first show that for $\varepsilon > 0$ and sufficiently small the hypotheses imply the existence of $\varepsilon\gamma = (\varepsilon\gamma_1, \dots, \varepsilon\gamma_k)$ with $\varepsilon t_1 > \varepsilon\gamma_1 > \dots > \varepsilon\gamma_k > 0$ and $\hat{c}^1, \hat{c}^2 \in \mathcal{N}(\hat{M}(\varepsilon\gamma))$ such that $\hat{c}^1 A(\varepsilon\gamma)(\hat{c}^1)^T > 0$, $\hat{c}^2 A(\varepsilon\gamma)(\hat{c}^2)^T < 0$. Assume for given $t_1 > 0$, γ and c^1 (say $|c^1| = 1$) are as stated in the theorem. Thus for any $\varepsilon > 0$, $\varepsilon t_1 > \varepsilon\gamma > \dots > \varepsilon\gamma_k > 0$ while $c^1 \in \mathcal{N}(M(\gamma))$ implies $c^1 \in \mathcal{N}(M(\varepsilon\gamma))$. Also $A(\varepsilon\gamma) = \varepsilon' A(\gamma)$. Now the Hausdorff metric distance between the closed unit discs in $\mathcal{N}(M(\varepsilon\gamma))$ and $\mathcal{N}(\hat{M}(\varepsilon\gamma))$ is $o(\varepsilon)$ as $\varepsilon \rightarrow 0$; hence there exists a k dimensional vector $v(\varepsilon)$ with $|v(\varepsilon)| = o(\varepsilon)$ such that if $\hat{c}^1(\varepsilon) = c^1 + v(\varepsilon)$ then $\hat{c}^1(\varepsilon) \in \mathcal{N}(\hat{M}(\varepsilon\gamma))$. Computing, $\hat{c}^1(\varepsilon) A(\varepsilon\gamma)(\hat{c}^1(\varepsilon))^T = c^1 A(\varepsilon\gamma)(c^1)^T + 2c^1 A(\varepsilon\gamma)v^T(\varepsilon) + v(\varepsilon)A(\varepsilon\gamma)v^T(\varepsilon) = \varepsilon' c^1 A(\gamma)(c^1)^T + o(\varepsilon'^{r+1})$ which is positive for sufficiently small $\varepsilon > 0$. A similar argument applies for \hat{c}^2 .

Corresponding to \hat{c}^1 and \hat{c}^2 there are maps, as in (3.3), denoted respectively $q^+(s)$, $q^-(s)$ with $q^\pm(0) = 0$, $q^{\pm n}(0) = \pm[(\text{ad}^r X, Y), Y](p)$. Then if $\mu(s) = q^+(\sqrt{s})$, $\mu'(0) = (1/2)q^{+n}(0)$.

Now choose integers ν_1, \dots, ν_{n-1} with $\nu_i < \nu_{i+1}$, and each as small as possible, so that $(\text{ad}^{\nu_i} X, Y)(p), \dots, (\text{ad}^{\nu_{n-1}} X, Y)(p)$ are linearly independent. From Corollary 2.3, for each i there exist admissible variations denoted $\alpha_\pm^i(s)$ such that $\lim_{s \rightarrow 0} (d/ds) \alpha_\pm^i(s) = \pm(\text{ad}^{\nu_i} X, Y)(p)$. Define

$$\alpha^i(s) = \begin{cases} \alpha_+^i(s) & \text{if } s \geq 0, \\ \alpha_-^i(|s|) & \text{if } s < 0, \end{cases} \quad q(s) = \begin{cases} q^+(\sqrt{s}) & \text{if } s \geq 0, \\ q^-(\sqrt{|s|}) & \text{if } s < 0. \end{cases}$$

Modifying variations by use of Lemma 2.1 to form compositions if necessary, but retaining the above notation, $\alpha^1(s_1) \circ \dots \circ \alpha^{n-1}(s_{n-1}) \circ q(s_n) \in \mathcal{A}(t_1, p)$ if $\sum s_i^2$ is sufficiently small. Furthermore the map $(s_1, \dots, s_n) \rightarrow \alpha^1(s_1) \circ \dots \circ \alpha^{n-1}(s_{n-1}) \circ q(s_n)$ takes zero to p and is differentiable with differential of rank n at $s_1 = \dots = s_n = 0$ since $(\text{ad}^{\nu_1} X, Y)(p), \dots, (\text{ad}^{\nu_{n-1}} X, Y)(p), [(\text{ad}^r X, Y), Y](p)$ are linearly independent. Thus the image of a neighborhood of $0 \in \mathbb{R}^n$ under this map covers a neighborhood of p in $\mathcal{A}(t_1, p)$. \square

For ease in later use we state the contrapositive of Theorem 1 as

THEOREM 1'. Assume, for the system (4), that $X(p) = 0$, $\dim \text{span } \mathcal{S}^1(p) = n - 1$ and r is the smallest integer such that $\dim \text{span } (\mathcal{S}^1(p) \cup \mathcal{S}^{2,r}(p)) = n$. A necessary condition that \exists a $t_1 > 0$ such that $p = T^X(t)p \in \partial \mathcal{A}(t, p)$ for $0 \leq t \leq t_1$ is that for all $\gamma_1, \dots, \gamma_k$ with $t_1 > \gamma_1 > \dots > \gamma_k > 0$ and all $c \in \mathcal{N}(M(\gamma))$, $cA(\gamma)c^T$ be semi-definite.

Remark. $A(\gamma)$ is a symmetric matrix with zero trace thus has both positive and negative real eigenvalues. Thus $cA(\gamma)c^T$ cannot be definite on \mathbb{R}^k but it may well be definite on $\mathcal{N}(M(\gamma))$ as will be illustrated in examples to follow.

The case $n = 2$.

PROPOSITION 3.4. If $n = 2$, $X(p) = 0$ and $\dim \text{span } \mathcal{S}^1(p) = 1$ then $Y(p) \neq 0$ and $\dim \text{span } (\mathcal{S}^1(p) \cup \mathcal{S}^2(p)) = 2$ if and only if $(\text{ad}^2 Y, X)(p) \notin \text{span } \mathcal{S}^1(p)$.

Proof. If both X and Y vanish at p so do all elements in $L(X, Y)$; hence one could not have $\dim \text{span } \mathcal{S}^1(p) = 1$. Thus $Y(p) \neq 0$.

If $(\text{ad}^2 Y, X)(p) \notin \text{span } \mathcal{S}^1(p)$ clearly $\dim \text{span } (\mathcal{S}^1(p) \cup \mathcal{S}^2(p)) = 2$. For the converse we show $(\text{ad}^2 Y, X)(p) \in \text{span } \mathcal{S}^1(p)$ implies all elements in $\mathcal{S}^2(p)$ belong to $\text{span } \mathcal{S}^1(p)$. This will be done by induction on k , the number of X factors in an element. For $k = 1$ we have $(\text{ad}^2 Y, X)(p) \in \text{span } \mathcal{S}^1(p)$. Now assume $[(\text{ad}^i X, Y), (\text{ad}^j X, Y)](p) \in \text{span } \mathcal{S}^1(p)$ if $i + j \leq k - 1$. Then for $i + j = k$ we note $(\text{ad}^i X, Y)(p) = \alpha_i Y(p)$, $(\text{ad}^j X, Y)(p) = \beta_j Y(p)$ for some α_i, β_j ; hence $\beta_j(\text{ad}^i X, Y)(p) - \alpha_i(\text{ad}^j X, Y)(p) = 0$. Using this, computation shows

$$(**) \quad [(\text{ad}^i X, Y), (\text{ad}^j X, Y)](p) = \beta_j [(\text{ad}^i X, Y), Y](p) - \alpha_i [(\text{ad}^j X, Y), Y](p).$$

If both $i, j < k$ the induction hypothesis shows the right side of (**), hence the left side, is in $\text{span } \mathcal{S}^1(p)$. If, say, $i = k$, then $j = 0$ and we consider $[Y, (\text{ad}^k X, Y)] = [[X, Y], (\text{ad}^{k-1} X, Y)](p) + [X, [Y, (\text{ad}^{k-1} X, Y)]](p)$. But $[[X, Y], (\text{ad}^{k-1} X, Y)](p)$ has $i = 1, j = k - 1$ hence is in $\text{span } \mathcal{S}^1(p)$ from (**), i.e. the case $i, j < k$. Also, $[Y, (\text{ad}^{k-1} X, Y)](p) \in \text{span } \mathcal{S}^1(p)$ by the induction hypothesis; thus by Proposition 3.1, since $X(p) = 0$, $[X, [Y, (\text{ad}^{k-1} X, Y)]](p) \in \text{span } \mathcal{S}^1(p)$ hence so is $[Y, (\text{ad}^k X, Y)](p)$. \square

Example 3.1. If $n = 2$, $X(p) = 0$, $\dim \text{span } \mathcal{S}^1(p) = 1$ and $\dim \text{span } (\mathcal{S}^1(p) \cup \mathcal{S}^2(p)) = 2$, we shall show the necessary condition of Theorem 1' is always satisfied (i.e. the sufficient condition of Theorem 1 that $p \in \text{int } \mathcal{A}(t, p)$ is never satisfied. This is expected, i.e. see [3, Theorem 1 and Ex. 1].).

With the above assumptions, from Proposition 3.4 we must have $[\text{ad} X, Y](p) \notin \text{span } \mathcal{S}^1(p)$, i.e. $r = 1$ in Theorem 1'. Since $Y(p) \neq 0$ (Proposition 3.4), the integer $\nu_1 = 0$ in the definition of $M(\gamma)$ and for any k (i.e. any number of switches), $M(\gamma)$ is the $k \times 1$ matrix $M(\gamma) = \text{col}(1, \dots, 1)$. It follows that an arbitrary element in $\mathcal{N}(M(\gamma))$ has the form $c(\alpha) = (\alpha_1, \alpha_2 - \alpha_1, \alpha_3 - \alpha_2, \dots, \alpha_{k-1} - \alpha_{k-2}, \alpha_{k-1})$ where $\alpha = (\alpha_1, \dots, \alpha_{k-1}) \in \mathbb{R}^{k-1}$ is arbitrary. Since $r = 1$,

$$A(\gamma) = \begin{pmatrix} 0 & (\gamma_2 - \gamma_1) & (\gamma_3 - \gamma_1) & \cdots & (\gamma_k - \gamma_1) \\ (\gamma_2 - \gamma_1) & 0 & (\gamma_3 - \gamma_2) & \cdots & (\gamma_k - \gamma_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (\gamma_k - \gamma_1) & \cdots & \cdots & (\gamma_k - \gamma_{k-1}) & 0 \end{pmatrix}.$$

Computation now shows with $c(\alpha)$ as above, $c(\alpha)A(\gamma)c^T(\alpha) = -2\alpha_1^2(\gamma_2 - \gamma_1) - 2\alpha_2^2(\gamma_3 - \gamma_2) - \cdots - 2\alpha_{k-1}^2(\gamma_k - \gamma_{k-1}) \geq 0$ since $\gamma_1 > \gamma_2 > \cdots > \gamma_k > 0$. This shows that with the assumptions of this example the necessary conditions of Theorem 1', i.e. that $p \in \partial \mathcal{A}(t, p)$ for small $t > 0$, are satisfied. In [3], it is shown that in this case, $p \in \partial \mathcal{A}(t, p)$ for small $t > 0$.

Extending the ideas in Example 3.1 gives

PROPOSITION 3.5. *Let $n \geq 2$ be an arbitrary integer, $X(p) = 0$ and $\dim \text{span } \mathcal{S}^1(p) = n - 1$. If $[[X, Y], Y](p) \notin \text{span } \mathcal{S}^1(p)$ the necessary condition that $p \in \partial \mathcal{A}(t, p)$ for small $t > 0$ of Theorem 1' is satisfied.*

Proof. Again, as in the proof of Proposition 2.4, we conclude $Y(p) \neq 0$ so $\nu_1 = 0$ and the first column of $M(\gamma) = \text{col}(1, \dots, 1)$. Denote this $k \times 1$ matrix by $M^1(\gamma)$. Then $\mathcal{N}(M(\gamma)) \subset \mathcal{N}(M^1(\gamma))$ which has arbitrary element $c(\alpha)$ as given in Example 2.1. The hypothesis $[[X, Y], Y](p) \notin \text{span } \mathcal{S}^1(p)$ implies $r = 1$ and $A(\gamma)$ is as in example 3.1. Since $c(\alpha)A(\gamma)c^T(\alpha) \geq 0$ for all $c(\alpha) \in \mathcal{N}(M^1(\gamma))$, the same holds for $c \in \mathcal{N}(M(\gamma))$. \square

Example 3.2. Let $X(x) = (x_2, 0, 0)$, $Y(x) = (0, 1, x_1)$, $p = 0$. Then $[X, Y](x) = (1, 0, -x_2)$, $(\text{ad}^2 X, Y)(x) = 0$ so $\dim \text{span } \mathcal{S}^1(p) = 2$. Next, $[[X, Y], Y](p) = (0, 0, -2) \notin \text{span } \mathcal{S}^1(p)$ and the second order necessary condition of Theorem 1' is satisfied.

For the general system (4), a necessary and sufficient condition that $\text{int } \mathcal{A}(t, p) \neq \emptyset \forall t > 0$ is $\dim L(\mathcal{S}^1)(p) = n$. (See [6, Th. 3.2] or [2, prop. 1.5].) If $n = 2$, $X(p) = 0$ and $\dim L(\mathcal{S}^1)(p) = 2$ so $Y(p) \neq 0$, a necessary and sufficient condition that $p \in \text{int } \mathcal{A}(t, p) \forall t > 0$ is that the smallest positive integer r such that $(\text{ad}^r Y, X)(p) \notin \text{span } \mathcal{S}^1(p)$ be odd.

The case $n > 2$. Throughout this section we shall assume $n > 2$, $X(p) = 0$, $\dim \text{span } \mathcal{S}^1(p) = n - 1$ and $\dim \text{span } (\mathcal{S}^1(p) \cup \mathcal{S}^2(p)) = n$. This means $\dim L(\mathcal{S}^1)(p) = n$ so $\text{int } \mathcal{A}(t, p) \neq \emptyset$ if $t > 0$ and also that there exists a smallest positive integer,

denoted r , such that $\dim \text{span}(\mathcal{S}^1(p) \cup \mathcal{S}^{2,r}(p)) = n$. We next show that r must be odd which is analogous to a result obtained in [1, Theorem 5.1].

PROPOSITION 3.6. *If r is even, say $r = 2m$, then $[(\text{ad}^{2m}X, Y), Y](p) \in \text{span } \mathcal{S}^1(p)$ if $[(\text{ad}^jX, Y), Y](p) \in \text{span } \mathcal{S}^1(p)$ for $j < 2m$.*

Proof. $[(\text{ad}^{2m}X, Y), Y](p) = [[X, (\text{ad}^{2m-1}X, Y)], Y](p) = [[X, Y], (\text{ad}^{2m-1}X, Y)](p) + [X, [(\text{ad}^{2m-1}X, Y), Y]](p)$. The second term on the right is in $\text{span } \mathcal{S}^1(p)$ by Proposition 3.1 and the assumption $[(\text{ad}^{2m-1}X, Y), Y](p) \in \text{span } \mathcal{S}^1(p)$. Thus we need only compute, for the first term on the right, $[[X, Y], (\text{ad}^{2m-1}X, Y)](p) = [[X, Y], [X, (\text{ad}^{2m-2}X, Y)]](p) = -[(\text{ad}^2X, Y), (\text{ad}^{2m-2}X, Y)](p) + [X, [[X, Y], (\text{ad}^{2m-2}X, Y)]](p)$. Again, the second term on the right belongs to $\text{span } \mathcal{S}^1(p)$ and we continue with the first term on the right as before. Inductively, since $r = 2m$, at the $(m-1)$ st step the process yields

$$\begin{aligned} [(\text{ad}^{m-1}X, Y), (\text{ad}^{m+1}X, Y)](p) &= [(\text{ad}^{m-1}X, Y), [X, (\text{ad}^mX, Y)]](p) \\ &= [(\text{ad}^mX, Y), (\text{ad}^mX, Y)](p) + W(p) = W(p) \end{aligned}$$

where $W(p) \in \text{span } \mathcal{S}^1(p)$. \square

Proposition 3.6 shows that the first case in which one might expect an interesting result would be $n = 3$, the number of switches $k \geq 3 > n - 1$ and $r = 3$, i.e. $[(\text{ad}^3X, Y), Y](p) \notin \text{span } \mathcal{S}^1(p)$. To illustrate the computations involved, we take $k = 4$ and assume $Y(p), (\text{ad}X, Y)(p)$ are linearly independent so $\nu_1 = 0, \nu_2 = 1$ while

$$(3.10) \quad M(\gamma) = \begin{pmatrix} 1 & \gamma_1 \\ 1 & \gamma_2 \\ 1 & \gamma_3 \\ 1 & \gamma_4 \end{pmatrix}.$$

Then

$$(3.11) \quad \mathcal{N}(M(\gamma)) = \left\{ \left(\frac{\alpha(\gamma_2 - \gamma_3)}{(\gamma_1 - \gamma_3)}, -\alpha + \frac{\beta(\gamma_3 - \gamma_4)}{(\gamma_2 - \gamma_4)}, -\beta \right. \right. \\ \left. \left. + \frac{\alpha(\gamma_1 - \gamma_2)}{(\gamma_2 - \gamma_4)}, \frac{\beta(\gamma_2 - \gamma_3)}{(\gamma_2 - \gamma_4)} \right) : \alpha, \beta \in \mathbb{R} \right\}.$$

Let $c(\alpha, \beta)$ denote the element in $\mathcal{N}(M(\gamma))$ obtained for parameter choices α, β . We wish to compute

$$(3.12) \quad c(\alpha, \beta) \begin{pmatrix} 0 & (\gamma_2 - \gamma_1)^3 & (\gamma_3 - \gamma_1)^3 & (\gamma_4 - \gamma_1)^3 \\ (\gamma_2 - \gamma_1)^3 & 0 & (\gamma_3 - \gamma_2)^3 & (\gamma_4 - \gamma_2)^3 \\ (\gamma_3 - \gamma_1)^3 & (\gamma_3 - \gamma_2)^3 & 0 & (\gamma_4 - \gamma_3)^3 \\ (\gamma_4 - \gamma_1)^3 & (\gamma_4 - \gamma_2)^3 & (\gamma_4 - \gamma_3)^3 & 0 \end{pmatrix} c^T(\alpha, \beta)$$

and see if this could change sign for some $\alpha, \beta \in \mathbb{R}$ and $\gamma_1 > \gamma_2 > \gamma_3 > \gamma_4 > 0$ with γ_1 small. If m is a large integer and one chooses $\gamma_1 = 4/m, \gamma_2 = 3/m, \gamma_3 = 2/m, \gamma_4 = 1/m$, i.e. evenly spaced switches, for arbitrary $\alpha, \beta \in \mathbb{R}$, the form (3.12) has value $-4(2\alpha^2 + \alpha\beta + 2\beta^2) \leq 0$. We next resolve the sign of this form in the case $k \geq 3$ is arbitrary and the spacing of the k switching times is also arbitrary.

PROPOSITION 3.7. *Let $n = 3, X(p) = 0, \dim \text{span } \mathcal{S}^1(p) = 2$ (assume $Y(p), [X, Y](p)$ are linearly independent) and $[(\text{ad}^3X, Y), Y](p) \notin \text{span } \mathcal{S}^1(p)$ while $[(\text{ad}^jX, Y), Y](p) \in \text{span } \mathcal{S}^1(p)$ for $j = 1$ (and hence for $j = 1, 2$ by Prop. 3.6). Then for arbitrary $k \geq 3$ and $0 < \gamma_k < \dots < \gamma_1, cA(\gamma)c^T \leq 0$ for all $c \in \mathcal{N}(M(\gamma))$.*

Proof. We have shown, above, that with four equally spaced "switches" (i.e. $k = 4$) we can generate $-[(\text{ad}^3 X, Y), Y](p)$. It suffices to produce a single example satisfying our hypotheses for which $[(\text{ad}^3 X, Y), Y](p)$ is not a tangent vector to $\mathcal{A}(t, p)$ at p for small $t > 0$. This implies, for this example and hence for all such systems, $cA(\gamma)c^T \leq 0$ for $c \in \mathcal{N}(M(\gamma))$. Consider $X(x) = (x_2, 0, -x_1^2)$, $Y = (0, 1, 0)$, $p = 0$. Clearly $\dot{x}_3(t) = -x_1^2(t) \leq 0$ so $p \in \partial\mathcal{A}(t, p)$ for all $t \geq 0$. Computation shows that $(\text{ad } X, Y)(p) = (1, 0, 0)$, $(\text{ad}^j X, Y)(p) = 0$ if $j \geq 2$ so $Y(p)$, $(\text{ad } X, Y)(p)$ generate $\text{span } \mathcal{S}^1(p)$ which has dimension 2. Next, $[(\text{ad } X, Y), Y](p) = [(\text{ad}^2 X, Y), Y](p) = 0$ while $[(\text{ad}^3 X, Y), Y](p) = (0, 0, 2) \notin \text{span } \mathcal{S}^1(p)$. This completes the proof. \square

Geometrically, above, we see that all points in $\mathcal{A}(t, p)$ have third coordinate nonpositive; hence since $[(\text{ad}^3 X, Y), Y](p) = (0, 0, 2)$ we would expect, as is the case, that $-[(\text{ad}^3 X, Y), Y](p)$ is a tangent vector but $[(\text{ad}^3 X, Y), Y](p)$ is not.

PROPOSITION 3.8. *Let $n = 3$, $X(p) = 0$, $\dim \text{span } \mathcal{S}^1(p) = 2$ (assume $Y(p)$, $(\text{ad } X, Y)(p)$ are linearly independent) and $[(\text{ad}^5 X, Y), Y](p) \notin \text{span } \mathcal{S}^1(p)$ while $[(\text{ad}^j X, Y), Y](p) \in \text{span } \mathcal{S}^1(p)$ for $j = 1, 3$ (and hence for $j = 1, 2, 3, 4$). Then $p \in \text{int } \mathcal{A}(t, p) \forall t > 0$.*

Proof. Here we have $r = 5$, $n = 3$, $\nu_1 = 0$, $\nu_2 = 1$. Let $t_1 > 0$ and m be such that $0 < 4/m < t_1$. We will show that with $k = 4$, and switching times $\gamma_1 = 4/m$, $\gamma_2 = 3/m$, $\gamma_3 = 1/m$ we can generate both $[(\text{ad}^5 X, Y), Y](p)$ and its negative as tangent vectors to $\mathcal{A}(t_1, p)$ at p .

With $k = 4$, $M(\gamma)$ is as in (3.10) while for $\gamma_1, \dots, \gamma_4$ as above, the general element in $\mathcal{N}(M(\gamma))$, as given in (3.11), has the form $c(\alpha, \beta) = (1/(2m))(\alpha, -2\alpha + \beta, \alpha - 2\beta, \beta)$. Also for our choice of the γ_i , and $r = 5$;

$$A(\gamma) = \frac{1}{5!m^5} \begin{pmatrix} 0 & -1 & -2^5 & -3^5 \\ -1 & 0 & -1 & -2^5 \\ -2^5 & -1 & 0 & -1 \\ -3^5 & -2^5 & -1 & 0 \end{pmatrix}.$$

Computing shows $5!c(\alpha, \beta)A(\gamma)c^T(\alpha, \beta) = (1/(4m^7))(-56\alpha^2 - 244\alpha\beta - 56\beta^2)$. If one chooses $\alpha > 0$ and $\beta = \alpha$ this yields $356\alpha^2/(4m^7) > 0$ while should we choose $\beta = -\alpha$ we obtain $-132\alpha^2/(4m^5) < 0$. Theorem 1 applies to show $p \in \text{int } \mathcal{A}(t, p) \forall t > 0$. \square

Example 3.3. Let $M = \mathbb{R}^3$, $X(x) = (x_2^2 + x_1, x_1, x_1x_2)$, $Y(x) = (1, 0, 0)$, $p = 0$. Then $[X, Y](p) = (1, 1, 0)$, $\dim \text{span } \mathcal{S}^1(p) = 2$ and $Y(p)$, $[X, Y](p)$ generate $\text{span } \mathcal{S}^1(p)$. Calculation shows $[(\text{ad}^j X, Y), Y](p) \in \text{span } \mathcal{S}^1(p)$ if $1 \leq j \leq 4$ while $[(\text{ad}^5 X, Y), Y](p) \notin \text{span } \mathcal{S}^1(p)$; hence Proposition 2.8 shows $p \in \text{int } \mathcal{A}(t, p) \forall t > 0$.

REFERENCES

- [1] A. J. KRENER, *The high order maximal principle and its applications to singular extremals*, this Journal, 15 (1977), pp. 256–293.
- [2] H. HERMES, *Local controllability and sufficient conditions in singular problems*, J. Differential Equations, 20, (1976) pp. 213–232.
- [3] ———, *Controlled stability*, Ann. Mat. Pura Appl., to appear.
- [4] R. BROCKETT, *System theory on group manifolds and coset spaces*, this Journal, 10, (1972) pp. 265–284.
- [5] E. B. DYNKIN, *Normed Lie algebras and analytic groups*, Amer. Math. Soc. Transl. Ser. 1, vol. 9 (1962), pp. 3–66.
- [6] H. J. SUSSMAN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.

SURVEY OF MEASURABLE SELECTION THEOREMS: RUSSIAN LITERATURE SUPPLEMENT*

A. D. IOFFE†

Abstract. This is a brief survey of results on measurable selection obtained in the U.S.S.R. and published mainly in Russian journals.

1. Introduction. These comments were intended to be an addendum to the survey of D. H. Wagner (D. H. Wagner, *Survey of measurable selection theorems*, this Journal, 15 (1977), pp. 859–903). Unfortunately, the idea of including such an addendum came up too late to be added in proof to the survey.

It was Wagner's suggestion that the comments should be published separately. Conversations with and materials presented by I. Evstigneev and especially V. Levin were very helpful for this work. R. T. Rockafellar and D. H. Wagner made many useful remarks concerning the first version of the comments. I have pleasure in expressing sincere thanks to all of them.

I have tried to retain as many usages of Wagner's comments as possible. This concerns terminology and notation as well as coding references. Thus (T, \mathcal{M}) will be a measurable space with σ -algebra \mathcal{M} ; X, Y, Z will denote the range spaces for set-valued mappings on T ; $\mathcal{S}(F)$ will denote the collection of \mathcal{M} -measurable selections for the set-valued mapping F . The few discrepancies in bibliographic coding with that of Wagner's survey will be specially stipulated. Where necessary, we shall refer to the survey of Wagner as "Wagner."

All the results presented are divided into two groups, the first containing general results on measurable selection and certain related results and the second dealing with Borel measurable selection theorems. The presentation is substantially chronological within each group. As far as achievements of a particular national school are considered here, this way to present the results seems to be natural though it differs from that accepted in Wagner.

2. General results. The problem of uniformization was being discussed in the U.S.S.R., mainly in the Moscow descriptive set theory school headed by Lusin, since the early thirties. The discussion was initiated by Lusin's work [LS] (cited in Wagner's addendum as [LS2]; see also Sierpinski, [SP1] in Wagner, for the same result) which appeared as a reply to a question posed by J. Hadamard in connection with the Zermelo choice axiom.

In works which followed, the situation $T = X = R$ and $\text{Gr } F$ being complementary Souslin was mainly considered. (In fact, in the pre-War literature no set-valued mappings were explicitly mentioned, only sets in the product space.) Sometimes additional assumptions were imposed on F such as F being closed-valued or finitely-many-valued, etc. (see Novikov [NO1], Lyapunov [LP1]–[LP3]). The main interest lay in studying projective properties of selections (uniformizations), and the typical result was that under certain conditions, there is a selection belonging to certain projective class, usually CA or PCA.

It is in the work of Yankov [YN] ([JN] in Wagner) that the case where $\text{Gr } F$ is Souslin was first considered and other properties of selections were investigated. His

* Received by the editors November 7, 1977.

† c/o Professor R. T. Rockafellar, Department of Mathematics, University of Washington, Seattle, Washington 98195.

result, which became known in Russian literature as the Lusin–Yankov, or merely Yankov, theorem, is stated in modern terms as follows.

THEOREM 1. *Let $T = X = R$, \mathcal{M} being the σ -algebra of Lebesgue measurable sets, and let $\text{Gr } F$ be Souslin. Then there is $f \in \mathcal{S}(F)$ whose graph has the Baire property and belongs to $(A_\rho)_{\sigma\delta}$, where A_ρ is the collection of differences of Souslin sets.*

The result of Yankov is known in the West mainly by the French translation ([JN] in Wagner). As noted by Wagner, a confusing discrepancy between the original and translated versions is that in the theorem statement of [YN] “measurable curve” appears, whereas “measurable” is omitted in the French translation. A particular consequence of this error was that Wagner (and probably many other Western mathematicians) concluded wrongly that Yankov did not in his theorem obtain a measurable function selection, although Wagner recognized that Yankov did so in his proof.

Another minor confusion may arise from the fact (also indicated by Wagner in the discussion preceding this supplement) that Yankov uses the term which translates into “unification” in English while the French translation uses the convenient word “uniformisation”. Wagner thinks that this discrepancy results from a mistranslation by Yankov of French works (like [LS]), which opinion I would rather share.

Several years after the work of Yankov had appeared, Rokhlin [RK1] observed that Yankov’s theorem immediately implies the following result: *if T is a Lebesgue space (i.e., (T, \mathcal{M}) is isomorphic to $(0, 1) \cup N$, N being the natural series, with the σ -algebra of Lebesgue measurable sets of $(0, 1)$ plus points of N as atoms), X is a Polish space and the graph of F is Souslin (mod 0) (which is to say, $\text{Gr } F$ can be turned into a Souslin set after changing a set whose projection on T has measure zero), then $\mathcal{S}(F) \neq \emptyset$.*

In [RK1], Rokhlin also stated and proved (though not absolutely correctly) a result which became later known as the theorem of Kuratowski and Ryll–Nardzewski. (For a thorough discussion, see Wagner’s addendum (iii).) A couple of months later, this fact with the same proof was repeated in [RK2].

The Yankov theorem, chiefly in Rokhlin’s form, gained diverse and numerous applications: ergodic theory [RK1], [RK2], integral representation of linear operators [NA], calculus of variations and control theory [AK], [IT1], [ARL3], Markov processes [DY] and others.

It is quite probable that, to a certain extent, the fact that Yankov’s theorem had appeared sufficient for so many applications, resulted in the underestimation of the second result of Rokhlin as well as the works of Kuratowski and Ryll–Nardzewski and Castaing. The most characteristic for the attitude is that Fillippov’s implicit function lemma, which proved to be so influential, attracted little attention in the U.S.S.R. surely because it is a direct corollary of Yankov’s theorem. Anyway, a new interest in the subject arose in the U.S.S.R. several years later. First brief surveys on the theory of measurable set-valued mappings were presented in [IL] and [IT2]; these surveys contain some new proofs but no new results. Among the works where new selection theory was applied in an effective way, we mention [DY], [DE], [EV1], [EV2] (probability and dynamic programming), [ER] (measure theory), [IL] (convex analysis), [VL] (partial differential equations).

Work contributing to a selection theory itself was resumed quite recently. Independently of Leese, Levin established the following result.

THEOREM 2 (Levin [LV2]). *Assume that X is weakly Souslin and $\text{Gr } F$ belongs to the Souslin family generated by $\mathcal{M} \otimes \mathcal{B}(X)$. Then there is a selection f of F which is measurable relative to the minimal σ -algebra containing all sets $F^-(A)$, where A is a Souslin subset of X .*

Levin also proved that *the projection on T of any element of the Souslin family generated by $\mathcal{M} \otimes \mathcal{B}(X)$ belongs to the Souslin family generated by \mathcal{M}* (this fact, in a somewhat less general setting can be found in the book of Dellacherie and Meyer [DM]; Levin's proof is essentially the same); hence Theorem 2 implies a result close to that of Leese (Theorem 13.1 in Wagner).

Assuming the continuum hypothesis, Evstigneev [EV3] extended Levin's result to the case where X is a compact Hausdorff space (not necessarily metrizable) and $\mathcal{B}(X)$ is replaced by the σ -algebra of Baire subsets of X .

Decomposition (or representation) theorems were studied by Ioffe ([IF2]). It was shown that in certain important cases when a measurable selection exists, it is possible to define the set-valued mapping by way of a Carathéodory mapping. More precisely, the following two theorems are true.

THEOREM 3. *Let X be a Polish space, and let F be measurable closed-valued with $F(t) \neq \emptyset$ for all $t \in T$. Then there are a Polish space Z and a Carathéodory mapping $f: T \times Z \rightarrow X$ such that $f(t, Z) = F(t)$ for any $t \in T$.*

The structure of Z and f can be further specified. For instance if X is linear and F is convex-valued, then Z can be considered a closed subset in a Frechét space and f can be taken linear in z .

THEOREM 4. *Assume that X is a (weak) Souslin space and $\text{Gr } F$ belongs to the Souslin family generated by $\mathcal{M} \otimes \mathcal{B}(X)$. Then there are a Polish space Z and a mapping $f: T \times Z \rightarrow X$ such that*

- (a) *f is a Carathéodory mapping with respect to the minimal σ -algebra containing the Souslin family generated by \mathcal{M} ;*
- (b) *$f(t, Z) = F(t)$ for every $t \in T$.*

3. Borel measurable selection theorems. All of the foregoing results are more or less connected, technically or conceptually, with Yankov's theorem. Meanwhile, another remarkable result was proved shortly before World War II.

THEOREM 5 (Novikov [NO2]). *Assume that $T = X = \mathbb{R}$ and $\text{Gr } F$ is Borel. Then the set $\{t \in T | F(t) \text{ is closed and nonempty}\}$ is complementary Souslin. In particular if F is closed-valued then $F^-(A)$ is Borel whenever A is closed. Hence in this case there is a Castaing representation of F by Borel functions.*

This statement is compiled from Novikov's theorem and corollary. (Certainly, Novikov used different terminology.) Note also that in the French translation " B " is omitted in the title.

Novikov's theorem generalized an earlier result of Kunugui, who required that $\text{Gr } F$ be F_5 . Novikov, however, gave a new proof based on a construction which appeared to be an essential element in proving further extensions of his theorem.

Arsenin [AS1] proved that *F has a Borel measurable selection if $\text{Gr } F$ is a Borel set and every $F(t)$ is congruent to one of countably many Borel sets E_1, E_2, \dots* . It is interesting that in this theorem "congruent" cannot be replaced by "similar" (i.e. equivalent under certain subsequent congruence and homothety), as seen from an example also given by Arsenin. In another paper [AS2], Arsenin proved that *under the assumption of Novikov's theorem, the set $\{t \in T | F(t) \text{ is a nonempty } F_\sigma\}$ also is complementary Souslin*. But he did not prove the corresponding selection theorem. This was done by Stschegolkov [ST1]. (Coban, [CB3] in Wagner, found an error in the proof of Stschegolkov's theorem given in [AL] in Wagner and produced a counterexample. But this error can be corrected by slightly redefining what is meant by an elementary CA set.) Later Stschegolkov strengthened this result: [ST2], [ST3]: *F has a Borel measurable selection if $\text{Gr } F$ is Borel and each $F(t)$ has a portion which is an F_σ* . (A portion of a linear set is the intersection of the set with an open interval.)

All these results can be extended to the case when T is a Polish space and X is a union of countably many compact metrizable spaces (a remark of V. Levin).

A very general result was announced by Čoban ([CB3] in Wagner). But the proof of this result seems not to be correct (the application of Baire's theorem is not sufficiently justified). Dellacherie [DL] proved a less general result, in fact, equivalent to that of Brown and Purves ([BP] in Wagner), and attributed this result to Stsčegolkov. According to a personal communication of Levin, the result of Dellacherie can indeed be deduced from Stsčegolkov's theorem.

Arkin and Levin [ARL2] (see also [ARL1]) studied convex-compact-valued mappings F from a Polish space into R^n . They showed that $\text{Gr } \tilde{F}$ is Borel if $\text{Gr } F$ is Borel. (\tilde{F} is the set of extreme points of F .) Together with the theorem of Novikov, this result enabled them to describe the set of extreme points of certain sets of functions of many variables and to establish new convexity theorems on integrals.

The role of Novikov's theorem is in particular that it allows one to prove graph-conditioned selection theorems without assuming that the measurable space is complete or a Souslin family, etc. (see [ARL3], [IF1]). A theorem of this sort was also established by Evstigneev [EV1], without invoking Novikov's theorem, as follows.

THEOREM 6. *Let X and Y be Polish spaces, and let $f: T \times X \times Y \rightarrow R$ be $\mathcal{M} \otimes \mathcal{B}(X) \otimes \mathcal{B}(Y)$ -measurable and such that the level set $\{(x, y) \in X \times Y \mid f(t, x, y) \leq c\}$ is compact for each $t \in T$ and $c \in R$. Then there is a mapping $u = u(t, x)$, from $T \times X$ into Y which is $\tilde{\mathcal{M}} \otimes \mathcal{B}(X)$ -measurable and such that*

$$\min_{y \in Y} f(t, x, y) = f(t, x, u(t, x)).$$

Here $\tilde{\mathcal{M}}$ denotes the completion of \mathcal{M} .

REFERENCES

- [AK] V. I. ARKIN, *On an infinite dimensional analogy to the nonconvex programming problem*, Kibernetika, 2 (1967), pp. 87–93.
- [ARL1] V. I. ARKIN AND V. L. LEVIN, *Extreme points of certain sets of measurable vector-functions of several variable and convexity of ranges of certain integrals*, Dokl. Akad. Nauk USSR, 199 (1971), pp. 1223–1226.
- [ARL2] ———, *Convexity of ranges of convex integrals, measurable choice theorems and variational problems*, Uspehi Mat. Nauk, 27 (1972), no. 3, pp. 21–77.
- [ARL3] ———, *Variational problems with functions of many variables and a resource distribution model*, in *Mathematical Economics and Functional Analysis*, B. S. Mityagin, Ed., Nauka, Moscow, 1974, pp. 7–34.
- [AS1] V. YA. ARSENIN, *On projections of B-sets*, Izv. Akad. Nauk USSR, (1939), no. 2, pp. 233–240.
- [AS2] ———, *The nature of projections of certain B-sets*, Izv. Akad. USSR Serija Mat., 4 (1940), pp. 403–410.
- [DL] C. DELLACHERIE, *Ensembles analytiques: Théorèmes de séparation et application*, Sémin. de Probabilité IX, Lecture Notes in Mathematics, Springer, New York, 1975.
- [DM] C. DELLACHERIE ET P.-A. MEYER, *Probabilités et Potential*, Hermann, Paris, 1975.
- [DN] E. B. DYNKIN, *Probability and concave dynamic programming*, Mat. Sb., 87 (1972), pp. 490–503.
- [DE] E. B. DYNKIN AND I. V. EVSTIGNEEV, *Regular conditional mathematical expectation of correspondences*, Teor. Verojatnost. i Priloz., 21 (1976), pp. 334–343.
- [DY] E. B. DYNKIN AND A. A. YUSHKEVITCH, *Markov Control Processes and their Applications*, Nauka, Moscow, 1975.
- [ER] M. P. ERSHOV, *Extensions of measures and stochastic equations*, Teor. Verojatnost. i Priloz., 19 (1974), pp. 457–471.
- [EV1] I. V. EVSTIGNEEV, *Measurable choice and dynamic programming*, Mathematics in Operations Research, 1 (1976), pp. 267–272.

- [EV2] ———, *The space 2^X and Markov fields*, Dokl. Akad. Nauk USSR, 230 (1976), pp. 22–25.
- [EV3] ———, *Methods of random sets*, Second Vilnius Conference on Probability and Statistics Abstracts, Vilnius, 1977.
- [IF1] A. D. IOFFE, *On lower semicontinuity of integral functions II*, this Journal, 15 (1977), pp. 991–1000.
- [IF2] ———, *Representation theorems for multifunctions and analytic sets*, Bull. Amer. Math. Soc., 84 (1978), pp. 142–144.
- [IL]' A. D. IOFFE AND V. L. LEVIN, *Subdifferentials of convex functions*, Trudy Mosk. Mat. Obšč., 26 (1972), pp. 3–73 = Trzns. Moscow Math. Soc., 26 (1972), pp. 1–72.
- [IT1]" A. D. IOFFE AND V. M. TIKHOMIROV, *Duality of convex functions and extremal problems*, Uspehi Mat. Nauk, 23 (1968), no. 6, pp. 51–116 = Russian Math. Surveys, 23 (1968), no. 6, pp. 53–125.
- [IT2]' ———, *Theory of Extremal Problems*, Nauka, Moscow, 1974; English translation, North-Holland, to appear 1978.
- [LV1] V. L. LEVIN, *On subdifferentials and continuous extensions preserving measurable dependence on parameter*, Funktsional. Anal. i prilož., 10 (1976), pp. 84–85.
- [LV2] ———, *Measurable selections of set-valued mappings and projections of measurable sets*, Ibid., to appear.
- [LP1] A. A. LYAPUNOV, *On some uniform analytical complements*, Izv. Akad. Nauk USSR, (1937), no. 2, pp. 286–304.
- [LP2] ———, *On the uniformization of plain CA - and A'_2 -sets*, Ibid., (1937), no. 1, pp. 41–52.
- [LP3] ———, *On the uniformization of analytical complements*, Mat. Sb., 3 (1938), pp. 219–223.
- [LS] N. N. LUSIN, *Sur la problème de J. Hadamard d'uniformisation des ensembles*, Matematica (Cluj), 4 (1930), pp. 54–56.
- [NA] M. A. NAIMARK, *Normed Rings*, Nauka, Moscow, 1968.
- [NO1] P. S. NOVIKOV, *Les projections des complémentaires analytiques uniformes*, Mat. Sb. 2 (1937), pp. 3–16.
- [NO2] ———, *On projections of certain B -sets*, Dokl. Akad. Nauk USSR, 23 (1939), pp. 863–866.
- [RK1] V. A. ROKHLIN, *Decomposition of a dynamical system into transitive components*, Mat. Sb., 25 (1949), pp. 235–249.
- [RK2] ———, *Selected topics from the metric theory of dynamical systems*, Uspehi Mat. Nauk., 4 (1949), no. 2, pp. 57–128 = Amer. Math. Soc. Transl., 49 (1966), pp. 171–240.
- [ST1] E. A. STSHEGOLKOV, *On the uniformization of certain B -sets*, Dokl. Akad. Nauk USSR, 59 (1943), pp. 1065–1068.
- [ST2] ———, *On the uniformization and decomposition of certain sets*, Ibid., 124 (1959), pp. 783–785. pp. 783–785.
- [ST] ———, *Uniformization of sets of certain classes*, Trudy Mat. Inst. Akad. Nauk USSR, 133 (1973), pp. 251–262.
- [VL]" A. M. VERSIK AND O. A. LADYZENSKAJA, *On the evolution of measures defined by the Navier–Stokes equation and on the solvability of the Cauchy problem for Hopf's statistical equation*, Dokl. Akad. Nauk USSR, 226 (1976) = Soviet Math. Dokl. 17 (1976), pp. 18–22.
- [YN] V. YANKOV, *On the unification of A -sets*, Ibid., 30 (1941), pp. 591–592.

SOME REMARKS ON CONTROLLABILITY FOR DISTRIBUTED PARAMETER SYSTEMS*

TOSHIHIRO KOBAYASHI†

Abstract. The purpose of this paper is to present a constructive procedure to obtain an approximate optimal control of the controllability problem for a distributed parameter system of parabolic type. The problem of optimal control determination is formulated as the problem of minimizing a functional $J(f)$ which measures the distance between the terminal state $u(T)$ and a given one u_T . Controllability is not sufficient for the existence of a minimizing solution of $J(f)$. It only assures $\inf_f J(f) = 0$. For a given $\eta > 0$ the elements which satisfy $J(f) \leq \eta$ generate the closed convex subset X_η . It is natural that the optimal control sequence is defined by the sequence such that each element has minimum norm in X_η for each η . An approximation method by regularization is presented to obtain constructively the optimal control sequence. The method gives uniquely the elements of the optimal control sequence depending continuously on the initial data u_0 and the terminal one u_T .

1. Introduction. Before entering into a discussion of a distributed parameter system, it may be useful to review the known facts for a finite dimensional system:

$$(1.1) \quad \frac{du}{dt} = Au(t) + Bf(t), \quad u(t) \in E^n, \quad f(t) \in E^r$$

The controllability problem is the following [4]: given $T > 0$ and points $u_0, u_1 \in E^n$, find $f \in L_2^r(0, T)$ such that the solution of (1.1) satisfies the initial condition

$$(1.2) \quad u(0) = u_0$$

and also satisfies the terminal condition

$$(1.3) \quad u(T) = u_T.$$

The solution of (1.1) is given by

$$(1.4) \quad u(t) = e^{At}u_0 + \int_0^t e^{A(t-s)}Bf(s) ds.$$

When such a control exists, it is not unique. It is natural that we seek the optimal control f_0 having minimum norm in $L_2^r(0, T)$. Let us define an operator W by

$$(1.5) \quad Wf = \int_0^T e^{A(T-t)}Bf(t) dt.$$

Then $W \in \mathcal{L}(L_2^r(0, T); E^n)$ (the space of continuous linear operators from $L_2^r(0, T)$ to E^n). We denote the adjoint operator of W by W^* . Solvability of the controllability problem is equivalent to an $n \times n$ matrix WW^* being positive. In this case, we can determine the unique optimal control f_0 having minimum norm by

$$(1.6) \quad f_0 = W^*(WW^*)^{-1}(u_T - e^{AT}u_0).$$

In the case of a distributed or infinite-dimensional system, the range of W is in general not closed [5]. Thus we cannot explicitly determine the unique optimal control by (1.6) [6]. Our purpose in this paper is to investigate a constructive method which gives an approximate optimal control depending continuously on the data (u_0, u_T) .

* Received by the editors March 9, 1977, and in revised form November 16, 1977.

† Department of Control Engineering, Kyushu Institute of Technology, Tobata, Kitakyushu, Japan.

2. Distributed parameter systems. We consider the class of distributed parameter systems described by parabolic partial differential equations [5]. Let V and H be Hilbert spaces with V, H and V' (the dual space of V) satisfying the inclusion relation

$$(2.1) \quad V \subset H = H' \subset V'$$

with each space dense in the following with continuous injection. We denote by $(\cdot, \cdot)_V$ (respectively, $(\cdot, \cdot)_H$) and $\|\cdot\|_V$ (respectively, $\|\cdot\|_H$) the scalar product in V (respectively H) and the norm on V (respectively, H). If $f \in V'$, $v \in V$, (f, v) denotes their scalar product; if $f \in H$, it coincides with the scalar product in H .

We are given a continuously bilinear form $a(u, v)$ on V such that

$$(2.2) \quad |a(u, v)| \leq L \|u\|_V \cdot \|v\|_V,$$

where L is a constant independent of u and v . For fixed u in V , the linear form

$$v \rightarrow a(u, v)$$

is continuous on V ; therefore it can be written

$$(2.3) \quad a(u, v) = (Au, v), \quad Au \in V'.$$

We also deduce from (2.2) that

$$(2.4) \quad \|Au\|_{V'} \leq L \|u\|_V \quad \text{for any } u \in V,$$

where $\|\cdot\|_{V'}$ is the dual norm of $\|\cdot\|_V$. Let the family of operators $A \in \mathcal{L}(V; V')$ be coercive; that is,

(2.5) there exists $\alpha > 0$ such that

$$a(u, u) \geq \alpha \|u\|_V^2, \quad u \in V.$$

If F , the control space, is a Hilbert space, we may define the following differential equation on V :

$$(2.6) \quad \frac{du(t)}{dt} + Au(t) = B(t)f(t), \quad t \in (0, T), \quad f \in L_2(0, T; F),$$

with initial condition

$$(2.7) \quad u(0) = u_0 \in H,$$

where $u(t)$ is the state vector, $f(t)$ is the control vector, and $B(t)$ is a continuous linear operator from F into V' . $u' = du/dt$ is taken in the sense of a distribution on $(0, T)$.

Remark 1. $L_2(0, T; F)$ denotes the space (equivalence class) of functions f defined on $[0, T]$ with values in a Hilbert space F such that

$$\int_0^T \|f(t)\|_F^2 dt < \infty.$$

Remark 2. In physical situations, the control space F is finite dimensional.

For the equation (2.6) with (2.7) we have the following existence and uniqueness lemma.

LEMMA 1 [5]. *Under the assumptions (2.2) and (2.5), the system (2.6) and (2.7) has a unique solution u such that $u \in L_2(0, T; V)$ and $u' \in L_2(0, T; V')$. Furthermore, the solution u depends continuously on u_0 and f .*

From Lemma 1, there exists an operator $U(t) \in \mathcal{L}(H; H)$ and the solution of the system (2.6) and (2.7) is given by

$$(2.8) \quad u(t; f) = U(t)u_0 + \int_0^t U(t-s)B(s)f(s) ds.$$

Remark 3 [5]. The family of operators $U(t)$ constitutes a semigroup of H . $-A$ is the infinitesimal generator of a semigroup. However, there exist operators A with $-A$ being the infinitesimal generator of a semigroup in a Hilbert space H without (2.5) being true.

3. Controllability. In this section, we investigate the controllability of the dynamical system described by (2.6) and (2.7) [1], [2], [5], [8].

We start with the following definition.

DEFINITION 1. The system (2.6) and (2.7) is said to be *controllable* at time T if $u(T; f)$ generates a dense subspace, $R(T)$, of the space H as f is varied without any constraints.

Remark 4. As the space $R(T)$ is not closed in the parabolic case, the above definition is natural. The definition says that for every $h \in H$, there exists a control f which steers the system arbitrarily close to h .

Now let us define an operator $W \in \mathcal{L}(L_2(0, T; F); H)$ by

$$(3.1) \quad Wf = \int_0^T U(T-t)B(t)f(t) dt, \quad f \in L_2(0, T; F).$$

The following theorem is immediate.

THEOREM 1. *The following are equivalent:*

- (i) *the system (2.6) and (2.7) is controllable at time T ;*
- (ii) *the null space of W^* is $\{0\}$;*
- (iii) *WW^* is positive.*

Proof. For any $h \in H$

$$(3.2) \quad (Wf, h)_H = (f, W^*h)_{L_2(0, T; F)}.$$

From this, the subspace which $u(T; f)$ generates is dense in H if and only if the null space of W^* is $\{0\}$.

On the other hand, since

$$(3.3) \quad \|W^*h\|_{L_2(0, T; F)}^2 = (WW^*h, h)_H, \quad h \in H.$$

the null space of W^* is $\{0\}$ if and only if the self adjoint operator WW^* on H is positive, that is, for any $h \in H$

$$(3.4) \quad (WW^*h, h)_H \geq 0 \quad \text{and} \quad (WW^*h, h)_H = 0 \quad \text{implies} \quad h = 0.$$

We have proved the theorem.

The explicit conditions of controllability were given by Kobayashi [3] for various types of systems with distributed, pointwise, scanning and boundary controls.

Now we can state the controllability problem for the distributed parameter system (2.6) and (2.7) as follows: given $T > 0$ and a terminal state $u_T \in H$, find an optimal control sequence $\{f_{0\eta}\}$ such that for any $\eta > 0$

$$(3.5) \quad \|u(T; f_{0\eta}) - u_T\|_H^2 \leq \eta,$$

and for each fixed η each element $f_{0\eta}$ has minimum norm of all elements satisfying (3.5).

We suppose that the system (2.6) and (2.7) is controllable at time T . By virtue of the definition, for any $\eta > 0$ there exists a control f_η such that $\|u(T; f_\eta) - u_T\|_H^2 \leq \eta$.

We may formulate our controllability problem as that of minimizing a functional $J(f)$ defined by

$$(3.6) \quad J(f) = \|u(T; f) - u_T\|_H^2$$

with respect to f , subject to the constraints (2.6) and (2.7). From (2.8) and (3.1), $J(f)$ becomes

$$(3.7) \quad \begin{aligned} J(f) &= \|U(T)u_0 + Wf - u_T\|_H^2 \\ &= \|Wf - u_d\|_H^2, \end{aligned}$$

where $u_d = U(T)u_0 - u_T$ and $u_d \in H$. If the system (2.6) and (2.7) is controllable at time T ,

$$(3.8) \quad \inf_f J(f) = 0, \quad f \in L_2(0, T; F).$$

However the infimum may never be realized.

Let X_η be the set of elements $f \in L_2(0, T; F)$ satisfying $J(f) \leq \eta$, $\eta > 0$. The set X_η is a closed subset of $L_2(0, T; F)$. Since $J(f)$ is a convex functional, for $f_1, f_2 \in X_\eta$ and $q \in (0, 1)$

$$J((1-q)f_1 + qf_2) \leq (1-q)J(f_1) + qJ(f_2) \leq \eta,$$

from which we get $(1-q)f_1 + qf_2 \in X_\eta$. Therefore for any $\eta > 0$ X_η is a closed convex subset of $L_2(0, T; F)$. Thus for each $\eta > 0$ there exists a unique element $f_{0\eta}$ having minimum norm in X_η . The existence of an optimal control sequence $\{f_{0\eta}\}$ has been shown.

The problem of concern here is to seek the optimal control sequence $\{f_{0\eta}\}$. For a finite-dimensional system, we can obtain the desired optimal control by (1.6) in the case of $\eta = 0$. This is because the range of the operator W is closed. For the distributed parameter system even if the system is controllable at time T , the range of W is not closed in H . Therefore we should investigate a method which constructively gives the optimal control sequence $\{f_{0\eta}\}$.

4. Construction of the optimal control sequence. This chapter presents a constructive method using regularization [5], [7] which explicitly determines the elements of the optimal control sequence. From a different point of view, this method corresponds to approximating the nonnegative operator $G = W^*W$ by a family of positive definite ones.

Let us introduce a regularized functional, $J_\varepsilon(f)$, corresponding to $J(f)$ defined by

$$(4.1) \quad J_\varepsilon(f) = J(f) + \varepsilon \|f\|_{L_2(0, T; F)}^2, \quad \varepsilon > 0.$$

We can see that there exists a unique minimizing solution f_ε of $J_\varepsilon(f)$. Since the operator W is continuous, $J_\varepsilon(f)$ is differentiable and convex. Hence the necessary condition for optimality is that

$$(4.2) \quad J'_\varepsilon(f_\varepsilon) \cdot f = 0 \quad \text{for every } f \in L_2(0, T; F).$$

That is,

$$(4.3) \quad (W^*(Wf_\varepsilon - u_d), f)_{L_2(0, T; F)} + \varepsilon (f_\varepsilon, f)_{L_2(0, T; F)} = 0$$

for every $f \in L_2(0, T; F)$. From this it follows that

$$(4.4) \quad ((W^*W + \varepsilon I)f_\varepsilon - W^*u_d, f)_{L_2(0, T; F)} = 0.$$

Here I is the identity operator on $L_2(0, T; F)$. Consequently the unique minimizing solution f_ε is determined by

$$(4.5) \quad f_\varepsilon = (W^*W + \varepsilon I)^{-1} W^*u_d = G_\varepsilon^{-1} W^*u_d.$$

Since the operator G_ε is positive definite, its inverse G_ε^{-1} is continuous. Thus f_ε depends continuously on the data u_d .

Under the assumption that the system (2.6), (2.7) is controllable at time T we shall first prove the following theorem.

THEOREM 2. *Let $J(f_\varepsilon) = \eta$, $\eta > 0$. Then f_ε is an element with minimum norm in X_η . That is, $\|f_\varepsilon\| \leq \|f\|$, $f \in X_\eta$.*

Proof. For every $f \in X_\eta$

$$J_\varepsilon(f_\varepsilon) \leq J_\varepsilon(f),$$

that is,

$$J(f_\varepsilon) + \varepsilon \|f_\varepsilon\|^2 \leq J(f) + \varepsilon \|f\|^2.$$

Since $J(f) \leq J(f_\varepsilon) = \eta$ and $\varepsilon > 0$,

$$(4.6) \quad \|f_\varepsilon\| \leq \|f\| \quad \text{for any } f \in X_\eta.$$

It follows from this theorem that f_ε is an element of the optimal control sequence $\{f_{0\eta}\}$.

LEMMA 2. *Suppose the system (2.6), (2.7) is controllable at time T . For any given $\eta > 0$, there exists $\varepsilon(\eta)$ such that $J(f_\varepsilon) \leq \eta$ for $\varepsilon \leq \varepsilon(\eta)$.*

Proof. It is sufficient to show that we can determine $\varepsilon(\eta)$ such that $J_\varepsilon(f_\varepsilon) \leq \eta$. Since the system (2.6), (2.7) is controllable at time T , we can choose $g \in L_2(0, T; F)$ so that $J(g) \leq \frac{1}{2}\eta$ for any $\eta > 0$. Then

$$J_\varepsilon(f_\varepsilon) \leq J_\varepsilon(g) = J(g) + \varepsilon \|g\|^2,$$

from which we obtain $J(f_\varepsilon) \leq \eta$ by taking

$$(4.7) \quad \varepsilon(\eta) \|g\|^2 \leq \frac{1}{2}\eta.$$

The lemma has been proved.

From this lemma we may use the sequence $\{f_\varepsilon\}$ as the optimal control sequence $\{f_{0\eta}\}$. Moreover, from the proof of Lemma 2 it is sufficient to choose one element $g \in L_2(0, T; F)$ for the determination of $\varepsilon(\eta)$. The element g must satisfy the condition $J(g) \leq \frac{1}{2}\eta$. This becomes

$$\|Wg\|_H - \|u_d\|_H \leq \|Wg - u_d\|_H \leq \frac{1}{\sqrt{2}} \sqrt{\eta}.$$

Since we may take $g \in L_2(0, T; F)$ such that $\|Wg\|_H \leq \|u_d\|_H$, we obtain

$$\|u_d\|_H - \|Wg\|_H \leq \frac{1}{\sqrt{2}} \sqrt{\eta},$$

from which

$$\|u_d\|_H - \frac{1}{\sqrt{2}} \sqrt{\eta} \leq \|Wg\|_H \leq \|W\| \cdot \|g\|.$$

Thus it is sufficient to choose the element g satisfying

$$(4.8) \quad \frac{\|u_d\|_H - (1/\sqrt{2})\sqrt{\eta}}{\|W\|} \leq \|g\|.$$

Next we only determine $\varepsilon(\eta)$ so that

$$(4.9) \quad \varepsilon(\eta)\|g\|^2 \leq \frac{1}{2}\eta.$$

If $\|u_d\|_H > (1/\sqrt{2})\sqrt{\eta}$, it follows from (4.8) that

$$\varepsilon(\eta) \left(\frac{\|u_d\|_H - (1/\sqrt{2})\sqrt{\eta}}{\|W\|} \right)^2 \leq \varepsilon(\eta)\|g\|^2 \leq \frac{1}{2}\eta.$$

Then we can determine $\varepsilon(\eta)$ such that

$$(4.10) \quad \varepsilon(\eta) \leq \frac{\frac{1}{2}\eta\|W\|^2}{(\|u_d\|_H - (1/\sqrt{2})\sqrt{\eta})^2}.$$

If $\|u_d\|_H \leq (1/\sqrt{2})\sqrt{\eta}$, we can arbitrarily take $\varepsilon(\eta) > 0$. Consequently we have the following theorem.

THEOREM 3. *Suppose the system (2.6), (2.7) is controllable at time T . If we take $\varepsilon(\eta)$ which satisfies (4.10) for any $\eta > 0$, then we have $J(f_\varepsilon) \leq \eta$ for all $\varepsilon \leq \varepsilon(\eta)$.*

This theorem shows that we can explicitly seek the control f_ε which satisfies

$$\|u(T; f_\varepsilon) - u_T\|_H^2 \leq \eta$$

for any given $\eta > 0$.

The problem which remains is to investigate the limiting properties of the optimal sequence $\{f_\varepsilon\}$.

LEMMA 3 [7]. (i) *For any $\varepsilon > 0$ the function $\phi(\varepsilon) = J_\varepsilon(f_\varepsilon)$ is continuous, increases monotonically and*

$$\lim_{\varepsilon \rightarrow \infty} \phi(\varepsilon) = \|u_d\|_H^2.$$

(ii) *For any $\varepsilon > 0$ the function $\gamma(\varepsilon) = \|f_\varepsilon\|^2$ is continuous and decreases monotonically and*

$$\lim_{\varepsilon \rightarrow \infty} \gamma(\varepsilon) = 0.$$

(iii) *For any $\varepsilon > 0$ the function $\rho(\varepsilon) = J(f_\varepsilon)$ is continuous, increases monotonically and its values cover the interval $[0, \|u_d\|_H^2]$, $0 < \varepsilon < \infty$.*

The proof of this lemma will be given in the Appendix. By virtue of Lemma 2 and Lemma 3, we obtain the following theorem.

THEOREM 4. *If the system (2.6), (2.7) is controllable at time T , then*

$$(4.11) \quad \lim_{\varepsilon \rightarrow 0} J(f_\varepsilon) = 0.$$

Since f_ε depends continuously on the data u_d and satisfies the condition (4.11), the sequence $\{f_\varepsilon\}$ is the desired optimal control sequence. We can use some f_ε as an approximate solution of our controllability problem.

Now we notice that the data (u_0, u_T) are known only through measurement. Hence the data u_d has errors which may be very small. Suppose u_d^* be the true data with

We should minimize the functional $J^*(f)$ defined by

$$(4.12) \quad J^*(f) = \|Wf - u_d^*\|_H^2.$$

We have the following theorem.

THEOREM 5. *Suppose the system (2.6), (2.7) is controllable at time T . Then*

$$(4.13) \quad \lim_{\varepsilon, \delta \rightarrow 0} J^*(f_\varepsilon) = 0.$$

Proof. It follows from (4.12) that

$$\begin{aligned} J^*(f_\varepsilon) &= \|Wf_\varepsilon - u_d^*\|_H^2 \leq (\|Wf_\varepsilon - u_d\|_H + \|u_d - u_d^*\|_H)^2 \\ &\leq (\|Wf_\varepsilon - u_d\|_H + \delta)^2. \end{aligned}$$

If the system (2.6), (2.7) is controllable at time T , Theorem 4 gives

$$\lim_{\varepsilon \rightarrow 0} \|Wf_\varepsilon - u_d\|_H^2 = 0.$$

Consequently we get

$$\lim_{\varepsilon, \delta \rightarrow 0} J^*(f_\varepsilon) = 0.$$

Moreover let f_ε^* be the minimizing solution of the functional $J_\varepsilon^*(f)$ such that

$$(4.14) \quad J_\varepsilon^*(f) = \|Wf - u_d^*\|_H^2 + \varepsilon \|f\|^2.$$

Then we have

THEOREM 6. *Suppose the system (2.6), (2.7) is controllable at time T . Then for any fixed ε*

$$(i) \quad \lim_{\delta \rightarrow 0} \|f_\varepsilon - f_\varepsilon^*\| = 0$$

If ε and δ are chosen with $\delta = o(\sqrt{\varepsilon})$ (δ has a higher order than $\sqrt{\varepsilon}$), then

$$(ii) \quad \lim_{\varepsilon, \delta \rightarrow 0} \|f_\varepsilon - f_\varepsilon^*\| = 0.$$

Proof. We obtain from (4.5)

$$(4.15) \quad G_\varepsilon f_\varepsilon = W^* u_d.$$

Similarly we have

$$(4.16) \quad G_\varepsilon f_\varepsilon^* = W^* u_d^*.$$

It follows from (4.15) and (4.16) that

$$(4.17) \quad G_\varepsilon (f_\varepsilon - f_\varepsilon^*) = W^* (u_d - u_d^*).$$

This equation means that the element $f_\varepsilon - f_\varepsilon^*$ realizes the lower bound of the following functional

$$(4.18) \quad I(f) = \|Wf - u_d + u_d^*\|_H^2 + \varepsilon \|f\|^2.$$

Thus

$$\inf_f I(f) = I(f_\varepsilon - f_\varepsilon^*),$$

from which

$$(4.19) \quad I(f_\varepsilon - f_\varepsilon^*) \leq I(0) = \|u_d - u_d^*\|_H^2 \leq \delta^2.$$

That is,

$$\varepsilon \|f_\varepsilon - f_\varepsilon^*\|^2 \leq \delta^2.$$

Therefore we have

$$(4.20) \quad \|f_\varepsilon - f_\varepsilon^*\| \leq \frac{\delta}{\sqrt{\varepsilon}}.$$

This inequality gives (i) and (ii) immediately.

Lastly we are going to give the following theorem corresponding to Theorem 3.

THEOREM 7. *Suppose the system (2.6), (2.7) is controllable at time T . We have*

$$J^*(f_\varepsilon) \leq \eta \quad \text{for all } \varepsilon \leq \varepsilon^*(\eta),$$

if we choose $\varepsilon^*(\eta)$ to satisfy the inequality

$$(4.21) \quad \varepsilon^*(\eta) \leq \frac{\frac{1}{2}\eta \cdot \|W\|^2}{(\|u_d\|_H + \delta - (1/\sqrt{2})\sqrt{\eta})^2}$$

for any $\eta > 0$.

Proof. From Theorem 3, if we take $\varepsilon^*(\eta)$ such that

$$(4.22) \quad \varepsilon^*(\eta) \leq \frac{\frac{1}{2}\eta \cdot \|W\|^2}{(\|u_d^*\|_H - (1/\sqrt{2})\sqrt{\eta})^2},$$

$J^*(f_\varepsilon) \leq \eta$ holds for all $\varepsilon \leq \varepsilon^*(\eta)$. We now have the inequality

$$(4.23) \quad |\|u_d\|_H - \|u_d^*\|_H| \leq \|u_d - u_d^*\|_H \leq \delta.$$

It follows from the statement (iii) of Lemma 3 that

$$(4.24) \quad J^*(f_\varepsilon) \leq \|u_d^*\|_H^2.$$

Therefore it is sufficient to take η such that $\sqrt{\eta} \leq \|u_d^*\|_H$. Since from (4.23)

$$\|u_d^*\|_H \leq \|u_d\|_H + \delta,$$

we obtain

$$(4.25) \quad \frac{\frac{1}{2}\eta \cdot \|W\|^2}{(\|u_d\|_H + \delta - (1/\sqrt{2})\sqrt{\eta})^2} \leq \frac{\frac{1}{2}\eta \cdot \|W\|^2}{(\|u_d^*\|_H - (1/\sqrt{2})\sqrt{\eta})^2}.$$

Therefore the theorem has been proved.

It follows from the preceding discussions that we may use some f_ε as an approximate optimal control for our controllability problem.

Appendix. *Proof of Lemma 3 [7].*

(i) Let ε and ε_1 be such that $0 < \varepsilon_0 \leq \varepsilon$, $0 < \varepsilon_0 < \varepsilon_1$. Then obviously $\phi(\varepsilon) = J_\varepsilon(f_\varepsilon) \leq J_\varepsilon(f_{\varepsilon_1})$ and, consequently

$$(1) \quad \phi(\varepsilon) - \phi(\varepsilon_1) \leq J_\varepsilon(f_{\varepsilon_1}) - J_{\varepsilon_1}(f_{\varepsilon_1}) = (\varepsilon - \varepsilon_1) \|f_{\varepsilon_1}\|^2.$$

If $\varepsilon \leq \varepsilon_1$ it follows from (1) that $\phi(\varepsilon) \leq \phi(\varepsilon_1)$, i.e. the function $\phi(\varepsilon)$ increases. Also

$$\phi(\varepsilon) \leq J_\varepsilon(0) = \|u_d\|_H^2, \quad \varepsilon > 0,$$

and consequently

$$(2) \quad \|f_\varepsilon\|^2 \leq \frac{\|u_d\|_H^2}{\varepsilon} \rightarrow 0, \quad \varepsilon \rightarrow \infty,$$

that is $\lim_{\varepsilon \rightarrow \infty} f_\varepsilon = 0$.

But then

$$\|u_d\|_H^2 \leq \lim_{\varepsilon \rightarrow \infty} \|Wf_\varepsilon - u_d\|_H^2,$$

and since

$$\|Wf_\varepsilon - u_d\|_H^2 \leq \phi(\varepsilon) \leq \|u_d\|_H^2, \quad \varepsilon > 0,$$

we have

$$\lim_{\varepsilon \rightarrow \infty} \|Wf_\varepsilon - u_d\|_H^2 = \lim_{\varepsilon \rightarrow \infty} \phi(\varepsilon) = \|u_d\|_H^2.$$

It remains to prove the continuity of the function $\phi(\varepsilon)$. It follows from (2) that for $\varepsilon \geq \varepsilon_0 > 0$

$$\|f_\varepsilon\|^2 \leq \frac{\|u_d\|_H^2}{\varepsilon_0} = c_0^2.$$

Interchanging ε and ε_1 in (1) and taking into account the evaluation obtained we have

$$|\phi(\varepsilon) - \phi(\varepsilon_1)| \leq |\varepsilon - \varepsilon_1| c_0^2 \rightarrow 0, \quad \varepsilon_1 \rightarrow \varepsilon,$$

i.e. the function $\phi(\varepsilon)$ is continuous for any $\varepsilon > 0$.

(ii) From (2) it follows that $\lim_{\varepsilon \rightarrow \infty} \gamma(\varepsilon) = 0$. We now show that the function $\gamma(\varepsilon)$ decreases. Interchanging ε and ε_1 in (1) and assuming that $\varepsilon > \varepsilon_1$, we obtain

$$(3) \quad \gamma(\varepsilon) = \|f_\varepsilon\|^2 \leq \frac{\phi(\varepsilon) - \phi(\varepsilon_1)}{\varepsilon - \varepsilon_1} \leq \|f_{\varepsilon_1}\|^2 = \gamma(\varepsilon_1),$$

which was to be proved. The continuity of $\gamma(\varepsilon)$ follows from the calculations below, in which we make use of the convexity of the functional $J(f)$;

$$\begin{aligned} \varepsilon \left\| \frac{f_\varepsilon - f_{\varepsilon_1}}{2} \right\|^2 &= \frac{\varepsilon}{2} \|f_\varepsilon\|^2 + \frac{\varepsilon}{2} \|f_{\varepsilon_1}\|^2 - \varepsilon \left\| \frac{f_\varepsilon + f_{\varepsilon_1}}{2} \right\|^2 \\ &= \frac{1}{2} \phi(\varepsilon) + \frac{1}{2} \phi(\varepsilon_1) - J_\varepsilon \left(\frac{f_\varepsilon + f_{\varepsilon_1}}{2} \right) + \frac{\varepsilon - \varepsilon_1}{2} \|f_{\varepsilon_1}\|^2 - \frac{1}{2} J(f_\varepsilon) \\ &\quad - \frac{1}{2} J(f_{\varepsilon_1}) + J \left(\frac{f_\varepsilon + f_{\varepsilon_1}}{2} \right) \\ &\leq \frac{1}{2} \phi(\varepsilon) + \frac{1}{2} \phi(\varepsilon_1) - J_\varepsilon \left(\frac{f_\varepsilon + f_{\varepsilon_1}}{2} \right) + \frac{\varepsilon - \varepsilon_1}{2} \|f_{\varepsilon_1}\|^2. \end{aligned}$$

Since

$$J_\varepsilon \left(\frac{f_\varepsilon + f_{\varepsilon_1}}{2} \right) \geq J_\varepsilon(f_\varepsilon) = \phi(\varepsilon),$$

we have

$$\varepsilon \left\| \frac{f_\varepsilon - f_{\varepsilon_1}}{2} \right\|^2 \leq \frac{1}{2} \phi(\varepsilon_1) - \frac{1}{2} \phi(\varepsilon) + \frac{\varepsilon - \varepsilon_1}{2} \|f_{\varepsilon_1}\|^2.$$

If $\varepsilon \geq \varepsilon_0 > 0$, $\varepsilon_1 \geq \varepsilon_0 > 0$ it follows from the inequality obtained that

$$\|f_\varepsilon - f_{\varepsilon_1}\|^2 \leq \frac{4}{\varepsilon_0} \left\{ \frac{1}{2} [\phi(\varepsilon_1) - \phi(\varepsilon)] + \frac{1}{2} (\varepsilon - \varepsilon_1) c_0^2 \right\} \rightarrow 0, \quad \varepsilon_1 \rightarrow \varepsilon$$

(by virtue of the continuity of the function $\phi(\varepsilon)$). But then

$$|\gamma(\varepsilon) - \gamma(\varepsilon_1)| = \|\|f_\varepsilon\|^2 - \|f_{\varepsilon_1}\|^2\| \leq 2c_0 \|f_\varepsilon - f_{\varepsilon_1}\|,$$

and it also approaches zero as $\varepsilon_1 \rightarrow \varepsilon$.

(iii) The continuity of the function $\rho(\varepsilon)$ follows from the continuity of the functions $\phi(\varepsilon)$ and $\gamma(\varepsilon)$ and the relation

$$\rho(\varepsilon) = \phi(\varepsilon) - \varepsilon \gamma(\varepsilon).$$

Also if $\varepsilon \geq \varepsilon_1$, using the statement (ii) we obtain

$$\begin{aligned} J_{\varepsilon_1}(f_{\varepsilon_1}) &= \|Wf_{\varepsilon_1} - u_d\|_H^2 + \varepsilon_1 \|f_{\varepsilon_1}\|^2 \leq J_{\varepsilon_1}(f_\varepsilon) \\ &= \|Wf_\varepsilon - u_d\|_H^2 + \varepsilon_1 \|f_\varepsilon\|^2 \\ &\leq \|Wf_\varepsilon - u_d\|_H^2 + \varepsilon_1 \|f_{\varepsilon_1}\|^2, \end{aligned}$$

and consequently $\rho(\varepsilon_1) \leq \rho(\varepsilon)$, $\varepsilon_1 \leq \varepsilon$. Moreover, since

$$\rho(\varepsilon) \leq \phi(\varepsilon) = J_\varepsilon(f_\varepsilon) \leq J_\varepsilon(0) = \|u_d\|_H^2,$$

we get

$$0 < \rho(\varepsilon) \leq \|u_d\|_H^2.$$

REFERENCES

- [1] H. O. FATTORINI, *Some remarks on complete controllability*, this Journal, 4 (1966), pp. 686–694.
- [2] ———, *On complete controllability of linear systems*, J. Differential Equations, 3 (1967), pp. 391–402.
- [3] T. KOBAYASHI, *Controllability and observability of distributed parameter systems*, Mem. Kyushu Inst. Tech. Engrg., (1975), no. 5, pp. 11–29.
- [4] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
- [5] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.
- [6] D. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.
- [7] V. A. MOROZOV, *The error principle in the solution of operational equations by the regularization method*, Ž. Vyčisl. Mat. i Mat. Fiz., 8 (1968), no. 2, pp. 295–309.
- [8] Y. SAKAWA, *Controllability for partial differential equations of parabolic type*, this Journal, 12 (1974), pp. 389–400.

APPLICATIONS OF ALGEBRAIC GEOMETRY TO SYSTEMS THEORY: THE McMILLAN DEGREE AND KRONECKER INDICES OF TRANSFER FUNCTIONS AS TOPOLOGICAL AND HOLOMORPHIC SYSTEM INVARIANTS*

CLYDE MARTIN† AND ROBERT HERMANN‡

Abstract. It is shown that every rational transfer function determines a mapping from the sphere S^2 into the Grassman manifold $G^m(\mathcal{C}^{m+p})$. Based on this embedding, it is proved that the McMillan degree of a multivariable rational transfer function can be defined using mixed algebro-geometric and algebro-topological methods. The pullback of the map from S^2 into $G^m(\mathcal{C}^{m+p})$ associates a vector bundle on S^2 with each such transfer function. The Grothendieck invariants of this bundle are shown to be feedback invariants of the transfer function. A complete systems theoretic interpretation of these invariants is obtained by relating the pullback bundle to the kernel bundle of a pencil of matrices associated with a minimal realization of the rational transfer function.

1. Introduction. Consider two linear, time-invariant multivariable systems, written in state space-form:

$$(1.1) \quad \frac{dx}{dt} = Ax + Bu, \quad y = Cx,$$

$$(1.2) \quad \frac{dx}{dt} = A'x + B'u, \quad y = C'x.$$

Here $u \in \mathcal{C}^m$, $y \in \mathcal{C}^p$, $x \in \mathcal{C}^n$: A, B, C, A', B', C' are complex matrices of the appropriate degree. The *transfer functions* of (1.1) and (1.2) are:

$$(1.3) \quad T(s) = C(sI - A)^{-1}B,$$

$$(1.4) \quad T'(s) = C'(sI - A')^{-1}B'.$$

These two systems are said to be *feedback equivalent* if there is a matrix F and invertible matrices α, γ, β such that

$$(1.5) \quad \begin{aligned} A' &= \alpha(A - B\gamma F\beta C)\alpha^{-1}, \\ B' &= \alpha B\gamma, \quad C' = \beta C\alpha^{-1}. \end{aligned}$$

The matrix F alters the system by feedback and the matrices α, γ , and β correspond to changes of basis in the states, input and output spaces, respectively. One can verify that, if the relation is satisfied, the transfer functions are related as follows

$$(1.6) \quad T' = \beta T \gamma (I - F \beta T \gamma)^{-1}.$$

Now, the transformation $T \rightarrow T'$ given by 1.6 is a *linear fractional transformation* [8]. This suggests that the natural setting in which to study feedback is a Grassman

* Received by the editor June 14, 1976, and in final revised form November 22, 1977.

† Division of Engineering and Applied Physics, Harvard University, Cambridge, Massachusetts 02138. The work of this author was supported in part by a National Science Foundation Energy Related Postdoctoral Fellowship and in part by NASA Contract No. A-35046-B. Current address: Ames Research Center, NASA, Moffett Field, California 94035.

‡ Department of Physics, Harvard University, Cambridge, Mass. 02138. The work of this author was supported in part by a National Research Council Senior Associateship at Ames Research Center, NASA, Moffett Field, California 94035.

manifold, since the Grassman manifolds are the homogeneous spaces of the group of linear fractional transformations. In fact we will show that one can associate to each T a map

$$(1.7) \quad \phi: S^2 \rightarrow G^m(\mathcal{C}^{m+p})$$

which is holomorphic. $G^m(\mathcal{C}^{m+p})$ is the Grassman manifold of m -dimension subspaces of $m+p$ dimensional complex space. We shall see that feedback equivalence of two systems means that the mappings ϕ and ϕ' are transformable into each other under the action of the group $Gl(p+m, \mathcal{C})$ acting on $G^m(\mathcal{C}^{m+p})$.

One would like to attach "invariants" to the mapping ϕ , which will determine when two of them are in the same feedback class. In this paper we will study two kinds of invariants. The first is related to the transfer function and its relation to its state-space realization.

Consider the general case: let s denote a complex variable and let $T(s)$ be a $p \times m$ matrix of complex valued rational functions of s such that

$$(1.8) \quad \lim_{s \rightarrow \infty} T(s) = 0.$$

Then, the *McMillan Degree* $\delta(t)$ is an integer with the following properties:

- (a) It is defined by algorithms, involving the "pure" algebra of rational functions.
- (b) It is equal to the state-space dimension of a minimal realization of T as the transfer function of a linear, time-invariant system with m inputs and p outputs.
- (c) If T and T' are transfer functions of two linear, time-invariant systems that are in the same feedback equivalence class, then

$$\delta(T) = \delta(T').$$

Each of these properties reflects fundamental systems-theoretic phenomena, which are discussed in great detail in the standard treatises, e.g., Brockett [2]; Kalman, Arbib and Falb [11]; and Rosenbrock [14].

Let r be the binomial coefficient $\binom{p+m}{p}$. Let \mathcal{C}^r be the space of r complex variables, and let $P_{r-1}(\mathcal{C})$ be the projective space for which \mathcal{C}^r is the space of homogeneous coordinates. Our main result about the McMillan degree can be described as follows:

Let $T(s)$ be a transfer function. It determines a rational map

$$S^2 \rightarrow P_{r-1}(\mathcal{C}),$$

i.e., an "algebraic curve" in $P_{r-1}(\mathcal{C})$, such that

$$(1.9) \quad \delta(T) = \text{intersection number of this algebraic curve with a hyperplane of } P_{r-1}(\mathcal{C}).$$

This property relates the McMillan degree, in a very fundamental way, to *both* topology and algebraic geometry. The embedding is obtained by using the classical Plücker embedding of $G^m(\mathcal{C}^{m+p})$ into projective space.

The second type of invariant with which we are concerned involves the definition and properties of certain *holomorphic* vector bundles associated with linear systems. Since $G^m(\mathcal{C}^{m+p})$ has a canonical vector bundle structure, we can use ϕ to construct a vector bundle, the pullback, on S^2 . We show that feedback equivalent systems

determine isomorphic vector bundles. Since a complete classification of holomorphic vector bundles on the sphere is known, we obtain a set of feedback invariants. We also give a complete interpretation of what the vector bundle isomorphism means in terms of the transfer functions. In so doing we show that the complete set of isomorphism invariants correspond to the well-known Kronecker invariants of the minimal realization of a transfer function.

2. Transfer functions. In this section we construct the basic object of this paper: the mapping induced by the transfer function from the Riemann sphere S^2 to the Grassman manifold $G^m(V)$. Let $T(s)$ be a $p \times m$ matrix of rational functions of a complex variable s and assume $T(s)$ satisfies (1.8). The following theorem (which is basically found in Rosenbrock [14]) is essential to this section. The theorem has an algebraic proof that we are unable to present here because of lack of space.

THEOREM 2.1. *Let $T(s)$ be a matrix of rational functions of a complex variable s . Then there exist matrices $N(s)$ and $D(s)$ of polynomial functions of s such that*

1. $T(s) = N(s)D(s)^{-1}$,
2. *there exist matrices of polynomials X and Y such that $X(s)N(s) + Y(s)D(s) = I$,*
3. $N(s)$ and $D(s)$ *are unique up to multiplication on the right by a unit of the ring of polynomial matrices.*

We can proceed with the construction of the desired map. Define, for each complex number s ,

$$(2.1) \quad \phi_T(s) = \{(T(s)u, u) : u \in \mathcal{C}^m\}.$$

Now the function ϕ is defined whenever the matrix $T(s)$ is defined, except at the poles of s . Using Theorem 2.1, we can extend the definition of ϕ_T as follows:

$$(2.2) \quad \phi_T(s) = \{(N(s)u, D(s)u) : u \in \mathcal{C}^m\}.$$

Now ϕ_T is defined for all $s \in \mathcal{C}$ and (2.2) agrees with (2.1) whenever $D(s)^{-1}$ exists. The set $\phi_T(s)$ is a linear subspace of $\mathcal{C}^m \times \mathcal{C}^p$ and when $D(s)^{-1}$ exists the dimension is m . The dimension is constant for all s unless $N(s)u = 0$ and $D(s)u = 0$ has a nonzero solution. However, by part 2 of Theorem 2.1, the only solution is $u = 0$; thus, the dimension is constant. By (1.8) we can extend the definition to $s = \infty$ by

$$(2.3) \quad \phi_T(\infty) = \{(0, u) : u \in \mathcal{C}^m\}.$$

Thus, ϕ_T is defined on all s in the Riemann sphere and takes values in $G^m(\mathcal{C}^{m+p})$. The transfer function T is then identified with the function

$$(2.4) \quad \phi_T: S^2 \rightarrow G^m(\mathcal{C}^{m+p}).$$

The general linear group acts naturally on $G^m(\mathcal{C}^{m+p})$ since a nonsingular linear transformation preserves dimension of subspaces. The action in local coordinates is given by linear fractional transformation (1.6). This motivates the use of Grassman manifolds. Consider the matrix

$$(2.5) \quad \alpha = \begin{bmatrix} I & 0 \\ F & I \end{bmatrix}$$

acting on $\phi_T(s)$. A calculation shows that

$$(2.6) \quad \alpha(\phi_T(s)) = \{(T(s)u, (FT(s) + I)u) : u \in U\}.$$

Whenever $(I + FT(s))^{-1}$ exists, we have $\alpha(\phi_T(s)) = \phi_{T'}(s)$ where

$$(2.7) \quad T' = T(I + FT)^{-1}.$$

Feedback is determined by a linear action on the Grassman manifold. Actually, change of basis in the input and output space can also be realized by such a linear action. Section 3 studies properties of ϕ_T that are invariant under continuous changes in T and § 4 studies properties of ϕ_T that are invariant under action by the general linear group acting on $G^m(\mathcal{C}^{m+p})$.

3. McMillan degree. In this section we show that the McMillan degree of a rational transfer function is the intersection number of $\phi(S^2)$ with an appropriately chosen submanifold of $G^m(\mathcal{C}^{p+m})$. The main tools used are the theory of intersections of manifolds (see Appendix A), the classical projective embedding of the Grassman space constructed by Plücker and Grassman, and the identification of the transfer function with a subset of the Grassman manifold.

Let α be the classical Plücker [9] embedding of $G^m(\mathcal{C}^{m+p})$ into the projective space of one-dimensional subspaces of $\wedge^m(\mathcal{C}^{m+p})$ where $\wedge^m(V)$ is the space of elements of degree m of the exterior algebra of V .

We can now define a hyperplane of $PS[\wedge^m(\mathcal{C}^{m+p})]$ by using exterior algebra. Let $\omega_1, \dots, \omega_m$ be elements of V^* , the dual space to \mathcal{C}^{m+p} . It defines a linear map: $\wedge_p(\mathcal{C}^{m+p}) \rightarrow \mathcal{C}$ as follows

$$(3.1) \quad (\omega_1 \wedge \dots \wedge \omega_m)(v_1 \wedge \dots \wedge v_m) = \begin{vmatrix} \omega_1(v_1) & \dots & \omega_1(v_m) \\ \vdots & & \vdots \\ \omega_m(v_1) & \dots & \omega_m(v_m) \end{vmatrix}.$$

Notice that this is multilinear and skew-symmetric in v_1, \dots, v_m (which is in turn a fundamental property of determinants), hence defines a *linear* function on $\wedge_m(V)$. Setting (3.1) to zero determines a hyperplane N in $PS[\wedge_m(V)]$.

Now apply this to $\mathcal{C}^p \oplus \mathcal{C}^m$. Let ϕ_T be the map from S^2 into $G^m(\mathcal{C}^{m+p})$ as constructed in § 2. Let $\omega_1, \dots, \omega_m$ be the elements of V^* defined as follows:

$$(3.2) \quad \omega_i(u \oplus y) = u_i \quad \text{for } 1 \leq i \leq m,$$

i.e., $\omega_i(u \oplus y)$ is the i th component of the vector u . Thus, we have

$$(3.3) \quad (\omega_1 \wedge \dots \wedge \omega_m)|\phi(s) = \det[D(s)].$$

In particular, $\phi(s)$ touches the hyperplane N if, and only if,

$$(3.4) \quad \det[D(s)] = 0.$$

Here is the main result of this section.

THEOREM 3.1. *The intersection of $\phi[P_1(\mathcal{C})]$ with the hyperplane N is equal to the degree of the determinant of $D(s)$.*

Proof. We proceed by first assuming that $\det[D(s)]$ has no multiple roots. It follows from (3.4) that at each intersection of $\phi[P_1(\mathcal{C})]$ with N they are in general position. Hence, there are exactly δ intersections where δ equals the degree of $\det[D(s)]$. Hence, the intersection number of $\phi[P_1(\mathcal{C})]$ with N is equal to δ , which proves Theorem 3.1.

To handle the general case, i.e., the case where the roots of $\det[D(s)]$ are multiple, we can perturb $D(s)$ to $D_\epsilon(s)$, with $D_\epsilon(s)$ having no multiple roots, and $\det[D_\epsilon(s)]$ having the same degree. The proof for this case again follows by the invariance of the intersection number under continuous deformations.

THEOREM 3.2. *Let (A, B, C) be a linear, time-invariant system (i.e., (A, B, C) are $(n \times n, n \times m, p \times n)$ complex matrices), and let*

$$T(s) = C(s - A)^{-1}B$$

be its transfer function. Suppose that the system is controllable and observable. Let $\phi: P_1(\mathbb{C}) \rightarrow G^m(\mathbb{C}^{n+m})$ be the rational map constructed above using T , and let δ be the intersection number of ϕ with the hyperplane N . Then, $\delta = n = \text{dimension of the state-space}$. In particular, δ is the McMillan degree of T , i.e., the dimension of a minimal realization of T as a transfer function.

Proof. Again, we can reduce to the case where: A has simple eigenvalues. (The general case can be reduced to this by deformations of A , under which δ and n remain unchanged.)

Continue to identify \mathbb{C}^{p+m} with $\mathbb{C}^m \oplus \mathbb{C}^p$, i.e., the set of ordered pairs (u, y) with $u \in \mathbb{C}^m \equiv \text{input vectors}$, $y \in \mathbb{C}^p \equiv \text{output vectors}$. Let

$$\gamma_0 = \{u \oplus y \in \mathbb{C}^{p+m} : u = 0\}.$$

γ_0 is an n -dimensional linear subspace of \mathbb{C}^{p+m} . It is readily seen that:

$$N = \{\gamma \in G^m(\mathbb{C}^{p+m}) : \gamma_0 \cap \gamma \neq (0)\}.$$

In words, the intersection of $G^m(\mathbb{C}^{m+n})$ with the hyperplane in the projective space in which it is embedded is the set of p -dimensional linear subspaces which do *not* meet the m -dimensional subspace γ_0 in general position. In particular, $T(s)\phi(s_0)$ meets N , for a value $s_0 \in \mathbb{C}$, if (and only if) $T(s)$ has a *pole* at $s = s_0$.

Let us now prove that *the poles of $T(s)$ are the eigenvalues of A* . To prove this fact let us write A in terms of its eigenvectors. This means writing

$$A = s_1 A_1 + \cdots + s_n A_n,$$

where A_1, \dots, A_n are the projective matrices onto the eigenvectors and s_1, \dots, s_n are the eigenvalues of A . Then,

$$\begin{aligned} A_i^2 &= A_i \quad \text{for } 1 \leq i \leq n, \\ A_i A_j &= 0 \quad \text{for } 1 \leq i \neq j \leq n. \end{aligned}$$

In particular, we have

$$\begin{aligned} T(s) &= C(s - A)^{-1}B \\ &= \frac{CA_1 B}{s - s_1} + \cdots + \frac{CA_n B}{s - s_n}. \end{aligned}$$

It is obvious from this formula that the poles of $T(s)$ are eigenvalues of A . Let us suppose that an eigenvalue of A —say s_1 —is *not* a pole of T . This requires that

$$CA_1 B = 0.$$

Hence, also

$$CA_1^k B = 0$$

for all integers k . Let x_1 be the eigenvector of A with eigenvalue s_1 . Then,

$$Ax_1 = s_1 x_1 = A_1 x_1$$

since the system (A, B, C) is controllable, there are input vectors $(u, u_1, \dots, u_r) \in \mathcal{C}^m$ such that

$$x_1 = Bu_0 + ABu_1 + \dots + A^r Bu_r.$$

Thus, for each integer k ,

$$\begin{aligned} CA_1^k x_1 &= s^k Cx_1 = CA_1^k Bu_0 + CA_1^k ABu_1 + \dots \\ &= 0 + CA_1^k ABu_1 + \dots. \end{aligned}$$

Now,

$$\begin{aligned} CA_1^k ABu_1 &= CA_1^k (s_1 A_1 + \dots) Bu, \\ &= CA_1^{k+1} s_1 Bu_1 = 0. \end{aligned}$$

Continuing this way, we see that

$$CA^k x_1 = 0$$

for all integers k .

Observability of the system (A, B, C) now implies that $x_1 = 0$, which is a contradiction.

Thus, we have shown that the points of intersection of $\phi(s)$ with the submanifold N are the eigenvalues of A . In particular, there are n of them, and $\delta \geq n$. Suppose that

$$\delta > n.$$

Again write

$$T(x) = N(s)D(s)^{-1},$$

where $N(s)$, $D(s)$ are relatively prime. We know that δ is the degree of the determinant of $D(s)$. We can also write

$$T(s) = \frac{CA_1 B(s-s_2) \cdots (s-s_n) + \dots}{(s-s_1) \cdots (s-s_n)}.$$

This is a contradiction to the way D is constructed [14], and finishes the proof of Theorem 3.2.

Remarks. That the McMillan degree is equal to the degree of the determinant of $D(s)$ was pointed out to us by M. Clark. He also remarked that the result can essentially be considered as being proved in Rosenbrock's book [14], although it does not seem to be explicitly stated there.

4. Vector bundles. In § 2 we have shown how a transfer function $T(s)$ induces a function from S^2 into $G^m(\mathcal{C}^{m+p})$. In this section we show that there is a natural vector bundle on S^2 associated with $T(s)$. We also show that a minimal realization of T determines a vector bundle on S^2 and we show that the two bundles are isomorphic. Certain known results about this vector bundle allow us to give a systems theoretic interpretation of bundle isomorphism classes of transfer functions. We make considerable use of the following theorem, which is basically found in [5].

THEOREM 4.1. *Every holomorphic vector bundle on S^2 is isomorphic to the direct sum of line bundles. The isomorphism classes are in one-to-one correspondence with sets of positive integers $k_1 \leq \dots \leq k_r$, and the k_i 's are the degrees of the line bundles in the decomposition.*

Let ϕ_T be the map induced from S^2 to $G^m(\mathcal{C}^{m+p})$ by T . Recall that the canonical vector bundle on $G^m(\mathcal{C}^{m+p})$ is the set of points $\{(\gamma, v); \gamma \in G^m(\mathcal{C}^{m+p}), v \in \gamma\}$. Define a vector bundle E_T on S^2 by the pullback of ϕ_T as follows: $(s, \zeta) \in E_T$ iff $\zeta \in \phi_T(s)$. It is well-known that the resulting bundle is holomorphic. We can alternatively describe E_T as

$$(4.1) \quad \{(s, y, u): y \in \mathcal{C}^p, u \in \mathcal{C}^m, y = T(s)u\}.$$

We have immediately the following result.

THEOREM 4.2. *Let $T(s)$ be a rational transfer function and let δ be its McMillan degree. Then there are positive integers*

$$\delta_1 \geq \cdots \geq \delta_p \geq 0, \quad \delta = \delta_1 + \cdots + \delta_p$$

which characterize the complex analytic isomorphism class of the vector bundle E_T .

Now consider vector bundles associated with a controllable system

$$(4.2) \quad \dot{x} = Ax + Bu.$$

Let $\alpha(s)$ be the pencil of matrices

$$(4.3) \quad \alpha(s) = (A - sI, B).$$

Let $V(s) = \{(x, u): (A - sI)x + Bu = 0\}$; if $s = \infty$ let $V(s) = \{(0, u): u \in \mathcal{C}^m\}$. Thus $V(s)$ consists of the kernel of $\alpha(s)$. We can now define a bundle $E_{A,B}$, which we call the kernel bundle of $\alpha(s)$ on S^2 , as $(s, x, u) \in E_{A,B}$ iff $(x, u) \in V(s)$. The following theorem connects bundles of the type E_T and $E_{A,B}$.

THEOREM 4.3. *Let (A, B, C) be a controllable, observable realization of a rational transfer function $T(s)$. Then E_T is isomorphic to $E_{A,B}$.*

Proof. Define a map C^* from $E_{A,B}$ to E_T as follows. C^* on the fiber at s is defined as

$$(4.4) \quad C^*(x, u) = (Cx, u).$$

We prove first that C^* is one-to-one. Suppose $C^*(x, u) = 0$; then $Cx = 0$ and $u = 0$. However, this implies by the definition of $E_{A,B}$, that $(A - sI)x = 0$. Multiplying by C gives $Cx = 0$ and, inductively, $CA^{k+1}x = 0$. Since (A, C) is an observable pair, $x = 0$ and C^* is one-to-one. Now, if s is not an eigenvalue of A , it is obvious that the dimension of the fiber at s in $E_{A,B}$ is the same as the dimension of the fiber in E_T ; so, C^* is an isomorphism, except possibly at those points. We show that the dimension of the fibers of $E_{A,B}$ is constant. For, suppose that $x'\alpha(s) = 0$ for some fixed s . Then $x'(A - sI) = 0$ and $x'B = 0$. Multiplying on the left by B we have

$$(4.5) \quad x'AB - sx'B = x'AB = 0.$$

Continuing, we have $x'A^k B = 0$ for all k and hence, by controllability, $x' = 0$. Thus, the rank of $\alpha(s)$ is n and the dimension of the kernel is m for all s . Therefore, the map C^* is a bundle isomorphism.

Now two systems (A, B) and (A', B') are said to be feedback equivalent if they are related as follows. There exist nonsingular matrices P_1 and P_2 and a matrix F such that

$$(4.6) \quad A' = P_1(A + BF)P_1^{-1}, \quad B' = P_1BP_2.$$

Now, Kalman [10] and Wonham and Morse [16] have given canonical forms and a complete set of invariants for systems under this group. Recall that these invariants

are sets of integers such that $k_1 \geq \cdots \geq k_m \geq 0$ and

$$k_1 + \cdots + k_m = n.$$

The following theorem relates feedback equivalence to bundle equivalence.

THEOREM 4.4. $E_{A,B}$ is isomorphic to $E_{A',B'}$ iff (A, B) is feedback equivalent to (A', B') .

Proof. If (A, B) is feedback equivalent to (A', B') then there exist P_1, P_2 and F such that

$$(4.7) \quad P_1(As - I, B) \begin{bmatrix} P_1^{-1} & 0 \\ FP_1^{-1} & P_2 \end{bmatrix} = (A' - sI, B').$$

Let

$$P = \begin{bmatrix} P_1^{-1} & 0 \\ FP_1^{-1} & P_2 \end{bmatrix},$$

and suppose $(A - sI, B)z = 0$; then $(A' - sI, B')P^{-1}z = P_1(A - sI, B)PP^{-1}z = 0$ and P^{-1} is the desired bundle isomorphism.

The converse is more subtle. Let (A, B) be given in canonical form with invariants k_1, \cdots, k_m . Then

$$(4.8) \quad A = \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ 0 & \cdots & 0 & A_m \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_m \end{bmatrix}$$

where

$$(4.9) \quad A_i = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & \cdots & 0 & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

has k_i rows and k_i columns. B_i has k_i rows and m columns and is zero except for a 1 in the i th column and last row. We show that the canonical form induces a decomposition of $E_{A,B}$ as the direct sum of line bundles. It is clear that $\alpha(s)$ is the direct sum of pencils $\alpha_i(s)$ where $\alpha_i(s)$ is determined by the pencil $(A_i - sI, B'_i)$. B'_i is unit vector with 1 in the last position, 0 elsewhere. It suffices to show that the kernel of $\alpha_i(s)$ is one dimensional. Since (A_i, B_i) is controllable, the rank of the pencil is k_i ; hence, the dimension of the kernel is 1. The kernel is given by

$$(4.10) \quad v(s) = e_1 + se_2 + \cdots + s^{k_i}e_{k_i+1},$$

and hence the degree of line bundle is k_i . Thus, we have a decomposition of $E_{A,B}$ as the direct sum of line bundles. Since the decomposition into line bundles is unique (up to ordering) [5] the feedback invariants are the same as the Grothendieck invariants. Now, if $E_{A,B}$ is isomorphic to $E_{A',B'}$, then the two bundles have the same Grothendieck invariants; hence, (A, B) and (A', B') have the same feedback invariants and are feedback equivalent. This finishes the proof of the theorem.

This theorem gives a means of determining when two transfer functions have isomorphic bundles.

COROLLARY 4.5. E_T is isomorphic to $E_{T'}$ iff for a minimal realization (A, B, C) of T and a minimal realization (A', B', C') of T' , (A, B) is feedback equivalent to (A', B') .

Corollary 4.5 reveals a lack of symmetry in this theory with respect to the way controllability and observability have been treated. It is not understood how this can be effectively done in this context.

In particular, if T and T' have isomorphic bundles, we can construct an explicit isomorphism by factoring through the realization. Consider $E_{A,B}$ and E_T . The isomorphism is given by a bundle map that, restricted to fibers, is

$$\begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix}.$$

We construct its inverse. Given $(u, y) \in E_T(s)$, we need to determine $(x, u) \in E_{A,B}(s)$. We have $Cx = y$ and $Ax = sx - Bu$; multiplying by C we have $CAX = sy + CBu$; continuing we find $CA^k x = s^k y - \sum_{i=0}^{k-1} s^i CA^{k-i-1} Bu$. By observability, the matrix

$$(4.11) \quad \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^n \end{bmatrix}$$

is left invertible and hence

$$(4.12) \quad x = C_1(s)y + C_2(s)u.$$

Thus, the inverse bundle isomorphism is given by

$$\begin{bmatrix} C_1(s) & C_2(s) \\ 0 & 1 \end{bmatrix}.$$

Also note that $C_1(s)(T)(s) + C_2(s) = (s - A)^{-1}B$ and so has the same McMillan degree as $T(s)$. We have the following corollary.

COROLLARY 4.6. If E_T is isomorphic to $E_{T'}$, then there is an isomorphism given by

$$(4.13) \quad \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} P_1 & 0 \\ FP_1 & P_2 \end{bmatrix} \begin{bmatrix} C_1(s) & C_2(s) \\ 0 & I \end{bmatrix}$$

where all three maps are isomorphisms. The matrices are constant except for C_1 and C_2 .

Although the Grothendieck results require complex isomorphisms, we have the stronger result that if T, T' are real transfer functions and E_T is isomorphic to $E_{T'}$, then there is a real isomorphism from E_T to $E_{T'}$. Thus, complex feedback is *not* necessary to achieve equivalence.

5. The McMillan degree as a Chern number. We have assigned to each transfer function $T(s)$ a holomorphic vector bundle E whose base space is S^2 . One can, of course, describe the basic topological invariants of the bundle in terms of characteristic classes [13]. Since these bundles have complex vector spaces as fibers, the relevant objects are called *Chern classes* [3]. Since the base is the two-dimensional sphere, the only nonzero one is the fiber Chern class C_1 , an element of $H^2(S^2, \mathbb{Z})$, the second cohomology group of S^2 with integer coefficients. The value of C_1 on the generator of $H_2(S^2, \mathbb{Z})$ is an integer called the *Chern number*. In fact, it can be proved

using the fact that C_1 is the pullback of a cohomology class on the Grassman manifold, that the Chern number is equal to the integration number as defined above, i.e., to the McMillan degree.

We have seen that the bundle E can be split up as a direct sum $E_1 \oplus \cdots \oplus E_m$ of "line bundles," i.e., bundles with one-dimensional complex vector spaces as fibers. By Whitney duality, the first Chern class of E is the sum of the Chern class of these line bundles. Again, it can be proved, by going back to the definitions [13], [3] that the Chern numbers of these line bundles are equal to the Kronecker indices.

Appendix. The theory of intersections of manifolds. Our basic reference for topology is Greenberg [4]; for manifold theory, Boothby [1], Hermann [6], and Loomis and Sternberg [12]. A sketch of how differential forms can be used to define the needed topological ideas is given in [7, Chap. 4].

Let M be a connected compact manifold. Assume it is orientable. For each integer n , let

$$H^n(M, R), \quad H_n(M, R)$$

denote the cohomology and homology vector spaces (with the real numbers R as coefficients). They are *dual vector spaces*. It follows from the orientability hypothesis that $H^m(M, R)$ ($m = \dim M$) is a one-dimensional vector space. The orientation determines a *generator* of $H^m(M, R)$, which enables it to be identified with R itself. This can be seen in two ways. First, in the usual topological setting—for example, the homology defined by singular chains, cohomology by Čech cochains—one can prove that $H_m(M, Z)$ (the homology with *integer* coefficients) is isomorphic to Z itself. ($Z \equiv$ the integers.) Then, there is a unique element of $H_m(M, Z)$, which generates it and is positively oriented. The image of the generator in $H_m(M, R)$, defined by the coefficient inclusion map $Z \rightarrow R$, is the distinguished generator of $H_m(M, R)$, which enables it to be identified with R . The second method is to work with cohomology defined in the manner of de Rham, i.e., with differential forms as cochains. Chains and homology are defined as the duals of the cochains. If an orientation is chosen for M , differential forms of degree m can be integrated over M . This defines an m -chain, which is the distinguished generator of $H_m(M, R)$.

For each pair (j, k) of integers, there is a bilinear mapping

$$(A.1) \quad H^j(M, R) \times H^k(M, R) \rightarrow H^{j+k}(M, R)$$

called the *cup product*. If $\omega_1 \in H^j(M, R)$, $\omega_2 \in H^k(M, R)$, the image of (ω_1, ω_2) in $H^{j+k}(M, R)$ under this map is denoted by

$$\omega_1 \cup \omega_2.$$

In particular, for $k = m - j$, it maps

$$(A.2) \quad H^j(M, R) \times H^{m-j}(M, R) \rightarrow H^m(M, R) = R.$$

POINCARÉ DUALITY THEOREM. *The bilinear mapping (A.2) is nondegenerate. In particular, it identifies $H^{m-j}(M, R)$ with the dual vector space of $H^j(M, R)$, and identifies $H^{m-j}(M, R)$ with $H_j(M, R)$.*

The cup-product (A.1) on cohomology then transforms (under this Poincaré duality isomorphism between homology and cohomology) into an algebraic operation on homology—this is the *intersection* operation. It defines a bilinear map:

$$(H_j \approx H^{m-j}) \times (H_k \approx H^{m-k}) \rightarrow (H^{2m-j-k} \approx H_{j+k-m}).$$

In particular, if

$$j + k = m,$$

and $H_0(M)$ is identified with R also, the *intersection* operation defines a bilinear map

$$H_j(M, R) \times H_k(M, R) \rightarrow R.$$

For $\alpha \in H_j(M, R)$, $\beta \in H_k(M, R)$, the real number

$$\iota(\alpha, \beta)$$

assigned to (α, β) by the operation is called the *intersection number* of the two homology classes α, β . We have shown in the main text that the McMillan degree can be obtained by specializing M to be a complex projective space; $j = 2$; α a homology class determined by a transfer function; and β the homology class determined by a hyperplane of complex projective space.

The above definitions of “intersection number” are conceptually very simple, once one understands basic homology theory. To be useful, it must be supplemented by a method of computing it in more familiar geometric terms, for a suitably “generic” situation. Differentiable manifold theory offers such a possibility. (Note that everything we have dealt with up to now holds for topological manifolds; the differentiable structure has not been used.)

Let N, N' be compact orientable manifolds, such that

$$\dim M = \dim N + \dim N'.$$

Fix orientation of N and N' . This determines generators of $H_n(N, R)$, $H_{n'}(N', R)$ ($n = \dim N$, $n' = \dim N'$), which are called the *fundamental homology classes* of the manifolds, denoted by $h_N, h_{N'}$. Let

$$\phi: N \rightarrow M, \quad \phi': N' \rightarrow M$$

be two continuous maps. Let

$$\phi_*(h_N) \in H_n(M, R), \quad \phi'_*(h_{N'}) \in H_{n'}(M, R),$$

be the image of these fundamental cycles in the homology of M (ϕ_* denotes the induced linear map on homology). The intersection

$$\iota[\phi_*(h_N), \phi'_*(h_{N'})]$$

is called the *intersection number* of the maps ϕ, ϕ' , denoted by

$$\iota(\phi, \phi').$$

Now suppose that ϕ, ϕ' are C^∞ maps. Let $p \in N, p' \in N'$ be two points such that

$$\phi(p) = \phi'(p'),$$

i.e., $\phi(N)$ and $\phi'(N')$ intersect at the point $\phi(p)$. The maps are said to *intersect in general position* at this point if

$$(A.3) \quad M_{\phi(p)} = d\phi(N_p) \oplus d\phi'(N'_{p'}).$$

(M_q denotes the tangent vector space to M at q ; $d\phi$ denotes the induced linear maps on tangent vectors.)

Now, fixing an orientation for N means that it makes sense when a basis for each tangent space is “positively” or “negatively” oriented. Let us say that $\phi(N)$ and $\phi'(N')$ meet at $\phi(p)$ in a positive way if (A.3) is satisfied, and if putting together a

positively oriented basis for N_p and N'_p provides a positively oriented basis for $M_{\phi(p)}$. Otherwise (and if they meet in general position) they are said to meet at $\phi(p)$ in a *negative* way.

Now, suppose that $\phi(N)$ and $\phi'(N')$ meet in general position at each point of intersection. Then, we have:

THEOREM A.1:

$$(A.4) \quad \iota(\phi, \phi') = \sum_{p \in \phi(N) \cap \phi'(N')} \pm 1.$$

Here, the sign $+$ or $-$ is chosen according to whether the submanifolds meet in a positive or negative way.

The left side of (A.4) involves topology; the right involves differential geometry. Their identity is a major link between differential geometry and topology.

Determining the orientations of the intersections is often an obstacle to determining the intersection number using formula (A.4). Working in the categories of *complex analytic* instead of *real* manifold removes this obstacle. The manifold M has a *complex manifold structure* if a set of coordinate charts is given, setting up coordinates in \mathcal{C}^m , with the transition maps between the charts given by complex analytic functions. A map $\phi: N \rightarrow M$ between complex manifolds is complex if it is given, in terms of complex charts, by complex analytic functions. A submanifold $\phi: N \rightarrow M$ is said to be complex if the map is complex.

Such a complex structure on manifold M determines an orientation for the manifold M . In terms of this orientation, two complex submanifolds always *meet with positive orientation*. Thus, the sum on the right-hand side of (A.4) *only involves plus signs*. In particular, $\iota(\phi, \phi')$ is equal to the number of intersections of the submanifolds $\phi(N)$, $\phi'(M')$, provided they meet in general position.

Here is the situation of greatest importance in algebraic geometry.

$$M = P_n(\mathcal{C}),$$

the complex projective space, of *real* dimension $2n$. It is the quotient of $\mathcal{C}^{n+1} - (0)$ under the dilatation group. $\phi(N)$, $\phi(N')$ are subsets determined by *nonsingular*, irreducible algebraic subsets of M . $P_n(\mathcal{C})$ is a complex manifold, and the algebraic subsets are complex submanifolds. The intersection number can, in this case, be defined by purely algebraic methods [15]. These algebraic methods also extend to algebraic subsets with singularities, although we make no use of them here.

Acknowledgment. We are indebted to Professor W. Fulton and A. Meyer for suggestions about mathematical tools. We would also like to thank R. Brockett for helpful conversations in the course of our work. We are grateful to M. Hazewinkel for pointing out a gap in an earlier version of this paper, whose correction led to the insight that completed this topic.

REFERENCES

- [1] W. M. BOOTHBY, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press, New York, 1975.
- [2] R. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1969.
- [3] S. S. CHERN, *On the multiplication in the characteristic ring of a sphere bundle*, Ann. of Math., 49 (1948), pp. 362–372.
- [4] M. GREENBERG, *Lectures on Algebraic Topology*, Benjamin, New York, 1967.

- [5] R. GUNNING, *Lectures on Vector Bundles over Riemann Surfaces*, Princeton Univ. Press, Princeton, NJ, 1967.
- [6] R. HERMANN, *Differential Geometry and the Calculus of Variations*, Academic Press, New York, 1968. (First edition is out of print. Second edition: Math.-Sci. Press, Brookline, MA, to appear.)
- [7] ———, *Vector Bundles in Mathematical Physics*, vol. II, Benjamin, Reading, MA, 1972.
- [8] ———, *Algebraic Topics in Systems Theory*, Interdisciplinary Mathematics, vol. 3, Math.-Sci. Press, Brookline, MA, 1973.
- [9] ———, *Linear Systems Theory and Introductory Algebraic Geometry*, Interdisciplinary Mathematics, vol. 8, Math.-Sci. Press, Brookline, MA, 1974.
- [10] R. E. KALMAN, *Kronecker Invariants and Feedback*, Conference on Ordinary Differential Equations, NRL Mathematics Research Center, June 14–23, 1971.
- [11] R. KALMAN, M. ARBIB AND P. FALB, *Topics in Mathematical Systems Theory*, McGraw-Hill, New York, 1969.
- [12] L. LOOMIS AND S. STERNBERG, *Advanced Calculus*, Addison-Wesley, Reading, MA, 1968.
- [13] J. W. MILNOR AND J. D. STANSHEFF, *Characteristic Classes*, Princeton Univ. Press, Princeton, NJ, 1974.
- [14] H. ROSENBROCK, *State Space and Multivariable Theory*, John Wiley, New York, 1970.
- [15] I. SHAFAREVITCH, *Basic Algebraic Geometry*, Springer-Verlag, Heidelberg, 1974.
- [16] W. M. WONHAM AND A. S. MORSE, *Feedback invariants of linear multivariable systems*, Automatica, 8 (1972), pp. 93–100.

NEAR-OPTIMAL FEEDBACK STABILIZATION OF A CLASS OF NONLINEAR SINGULARLY PERTURBED SYSTEMS*

JOE H. CHOW† AND PETAR V. KOKOTOVIĆ†

Abstract. A new series expansion method is developed for a class of nonlinear singularly perturbed optimal regulator problems. The resulting feedback control is near-optimal and can stabilize essentially nonlinear systems when linearized models provide no stability information. The stability domain is shown to include large initial conditions of the fast variables. The control law is implemented in two-time-scales, with the feedback from the fast state variables depending on slow state variables as parameters. The coefficients of the formal expansions of the optimal value function are obtained from equations involving only the slow variables.

1. Introduction. Compared with the rich literature on linear regulator theory, publications dealing with feedback design of nonlinear systems are a small minority. Realistic approaches to the difficult nonlinear feedback control problem usually exploit properties of special classes of systems to develop approximate methods [1], [2]. The approach in this paper exploits multiple time scale properties of a class of nonlinear singularly perturbed systems [3], [4] to achieve stabilization and near-optimality. The stabilization results obtained are essentially nonlinear in the sense that they also apply to the critical case when linearized models provide no stability information. Due to a separation of time scales, the proposed design procedure is applicable to higher order systems.

The problem considered is to optimally control the nonlinear system

$$(1a) \quad \dot{x} = a_1(x) + A_1(x)z + B_1(x)u, \quad x(0) = x_0,$$

$$(1b) \quad \mu \dot{z} = a_2(x) + A_2(x)z + B_2(x)u, \quad z(0) = z_0,$$

with respect to the performance index

$$(2) \quad J = \int_0^\infty [p(x) + s'(x)z + z'Q(x)z + u'R(x)u] dt,$$

where $\mu > 0$ is the small singular perturbation parameter, x, z are n -, m - dimensional states, respectively, u is an r -dimensional control and the prime denotes a transpose. It is assumed that there exists a domain $D \subset R^n$ containing the origin such that for all $x \in D$ and $z \in R^m$ the problem satisfies the following assumptions:

- I. The functions $a_1, a_2, A_1, A_2, B_1, B_2, p, s, q$ and R are differentiable with respect to x a sufficient number of times and a_1, a_2, p and s are all zero only at $x = 0$.
- II. The matrices $Q(x)$ and $R(x)$ are positive definite, that is, $Q(x) > 0, R(x) > 0$. Furthermore, the scalar function $p + s'z + z'Qz$ of x and z is positive definite in both x and z .
- III. For every fixed $x \in D$

$$(3) \quad \text{rank } [B_2, A_2B_2, \dots, A_2^{m-1}B_2] = m$$

* Received by the editors December 28, 1976, and in revised form September 12, 1977. This work was supported in part by the National Science Foundation under Grant ENG 74-20091, in part by the Energy Research and Development Administration under Contract U.S. ERDA E(49-18)-2088, and in part by the U.S. Air Force under Grant AFOSR 73-2570.

† Coordinated Science Laboratory and Department of Electrical Engineering, University of Illinois, Urbana, Illinois 61801.

and hence $A_2(x)$ is assumed to be nonsingular. (If not, then using $u = \hat{u} + K(x)z$ such that $A_2 + B_2K$ is nonsingular we redefine the problem.)

Assumptions I and II establish that the origin is the desired equilibrium of (1). Assumption III and $Q(x) > 0$ simplify the derivations. Alternatively a less restrictive stabilizability-detectability condition can be used.

Finite time trajectory optimization problems for the same class of systems have been treated in [3], [4] via singularly perturbed two point boundary value problems originating from necessary optimality conditions. The resulting controls are open-loop and require boundary layer correction terms at both ends of the interval. For the infinite time regulator problem considered here the Hamilton–Jacobi–Bellman sufficiency condition is more suitable since it readily incorporates stability requirements and leads to feedback solutions. Using this condition we obtain near-optimal stabilizing controls in feedback form and avoid explicit treatment of boundary layer phenomena.

Our procedure is based on a nested power series expansion of the optimal value function in z and μ . An advantage of this procedure is that it uses lower order equations involving only the slow variable x . In applications truncated series are of interest. Stabilizing properties of various truncated designs are discussed and an explicit estimate of the stability domain is given. It is of practical importance that this domain encompasses large initial disturbances of $z(0)$. Furthermore, near-optimality of these truncated designs is established in terms of $O(\mu)$, $O(\mu^2)$, etc. A particularly useful result is that an $O(\mu)$ near optimal feedback control can be implemented without knowing the value of the small parameter μ .

The paper is organized as follows. In § 2 a reduced order problem is formulated for the slow variable x . The crucial assumption is that the properties of its solution are known. Using a truncated expansion of the optimal value function the so-called composite control is introduced in § 3. Since the leading term in the series is the optimal value function of the reduced problem, the original problem is well posed. In § 4 it is shown that the composite control guarantees a finite domain of stability for the resulting feedback system. In § 5, a formal expansion of the optimal value function is proposed and near-optimality results are discussed. An example is discussed in § 6.

2. The reduced control. In singular perturbation techniques [5], a problem for the full order system (1) where $\mu > 0$ is interpreted as a perturbation of a reduced problem

$$(4a) \quad \dot{x} = a_1(x) + A_1(x)z + B_1(x)u, \quad x(0) = x_0$$

$$(4b) \quad 0 = a_2(x) + A_2(x)z + B_2(x)u,$$

in which $\mu = 0$. Due to assumption III, z can be solved from (4b) and eliminated from (4a) and (2). Then the reduced problem is to optimally control the system

$$(5) \quad \dot{x} = a_0(x) + B_0(x)u, \quad x(0) = x_0$$

with respect to

$$(6) \quad J_0 = \int_0^\infty [p_0(x) + 2s'_0(x)u + u'R_0(x)u] dt,$$

where

$$(7) \quad \begin{aligned} a_0 &= a_1 - A_1A_2^{-1}a_2, & B_0 &= B_1 - A_1A_2^{-1}B_2, \\ p_0 &= p - s'A_2^{-1}a_2 + a'_2A_2'^{-1}QA_2^{-1}a_2, & s_0 &= B'_2A_2'^{-1}(QA_2^{-1}a_2 - \tfrac{1}{2}s), \\ R_0 &= R + B'_2A_2'^{-1}QA_2^{-1}B_2. \end{aligned}$$

The origin $x = 0$ is the desired equilibrium of the optimally controlled reduced system (5) for all $x \in D$, since, in view of assumption II, $a_0(0) = 0$ and

$$(8) \quad p_0(x) + 2s'_0(x)u + u'R_0(x)u$$

is positive definite in x and u .

The reduced problem (5), (6) is considerably simpler than the original problem (1), (2) because of the elimination of the fast variables and the reduction of the system order. One of the tasks of the singular perturbation analysis is to establish whether the full problem is well-posed in the sense that its solution tends to the solution of the reduced problems as $\mu \rightarrow 0$. If so, then the next task is to deduce the properties of the original problem from the properties of the reduced problem. Finally these properties are to serve as a basis for a simplified design procedure.

To formulate our basic assumption about the properties of the solution of the reduced problem we use the optimality principle

$$(9) \quad 0 = \min_u [p_0(x) + 2s'_0(x)u + u'R_0(x)u + L_x(a_0(x) + B_0(x)u)],$$

where L is the optimal value function and L_x is its partial derivative with respect to x . This yields the minimizing control

$$(10) \quad u_0 = -R_0^{-1}(s_0 + \frac{1}{2}B'_0L'_x)$$

whose elimination from (9) results in the Hamilton-Jacobi equation

$$(11) \quad 0 = (p_0 - s'_0R_0^{-1}s_0) + L_x(a_0 - B_0R_0^{-1}s_0) - \frac{1}{4}L_xB_0R_0^{-1}B'_0L'_x, \quad L(0) = 0.$$

Note that, due to (8), $p_0 - s'_0R_0^{-1}s_0$ is positive definite in D . Our crucial assumption is then stated as follows.

IV. The unique positive definite solution $L(x)$ of (11) exists in D and is differentiable with respect to x a sufficient number of times. Furthermore the level surface $L = c_0 = \text{constant}$ is taken to be the boundary of the set D .

In the special case considered in [1], where the linearization of (5) at $x = 0$ is stabilizable and its states are observable in the quadratic approximation of J_0 , our assumption IV is automatically satisfied for all x near the origin. It follows from assumption IV that u_0 is the unique optimal feedback control for the reduced problem and L is a Lyapunov function of the optimally controlled reduced system

$$(12) \quad \dot{x} = a_0 - B_0R_0^{-1}(s_0 + \frac{1}{2}B'_0L'_x) = \bar{a}_0(x),$$

establishing that the origin is asymptotically stable and the set D belongs to its domain of attraction.

3. The composite control. The optimal value function $V(x, z, \mu)$ of the full problem (1), (2) satisfies the equation

$$(13) \quad 0 = \min_u \left[p + s'z + z'Qz + u'Ru + V_x(a_1 + A_1z + B_1u) + \frac{1}{\mu}V_z(a_2 + A_2z + B_2u) \right],$$

where V_x , V_z denote the partial derivatives of V with respect to the variables x , z , respectively. The minimizing control of (13) is

$$(14) \quad u = -\frac{1}{2}R^{-1}\left(B'_1V'_x + \frac{1}{\mu}B'_2V'_z\right),$$

and its substitution into (13) yields the Hamilton–Jacobi equation

$$(15) \quad 0 = p + s'z + z'Qz + V_x(a_1 + A_1z) + \frac{1}{\mu} V_z(a_2 + A_2z) \\ - \frac{1}{4} \left(V_x B_1 + \frac{1}{\mu} V_z B_2 \right) R^{-1} \left(B_1' V_x' + \frac{1}{\mu} B_2' V_z' \right), \quad V(0, 0, \mu) = 0.$$

Since system (1) is linear in z and J is quadratic in z , and since \dot{z} is multiplied by μ , we seek a solution of (15) in the form

$$(16) \quad V(x, z, \mu) = \bar{V}_0(x) + \mu \bar{V}_1'(x)z + \mu z' \bar{V}_2(x)z + \mu q(x, z, \mu) \\ \equiv \bar{V}(x, z, \mu) + \mu q(x, z, \mu), \quad \bar{V}_0(0) = 0$$

where

$$(17) \quad \partial q / \partial x = O(1), \quad \partial q / \partial z = O(\mu).$$

We shall investigate the expansion of q in a later section. The partial derivatives of V with respect to x, z are

$$(18) \quad V_x = \bar{V}_{0x} + O(\mu), \\ V_z = \mu \bar{V}_1' + 2\mu z' \bar{V}_2 + O(\mu^2).$$

Substituting (18) into (15) and neglecting the μ -dependent terms, we obtain the equation

$$(19) \quad 0 = p + \bar{V}_{0x}a_1 + \bar{V}_1'a_2 - \frac{1}{4}(\bar{V}_{0x}B_1 + \bar{V}_1'B_2)R^{-1}(B_1'\bar{V}_{0x} + B_2'\bar{V}_1) \\ + [s' + 2a_2'\bar{V}_2 + \bar{V}_{0x}(A_1 - B_1R^{-1}B_2'\bar{V}_2) + \bar{V}_1'(A_2 - B_2R^{-1}B_2'\bar{V}_2)]z \\ + z'(Q + \bar{V}_2A_2 + A_2'\bar{V}_2 - \bar{V}_2B_2R^{-1}B_2'\bar{V}_2)z.$$

In order to satisfy (19) identically for all z , we require that

$$(20) \quad 0 = p + \bar{V}_{0x}a_1 + \bar{V}_1'a_2 - \frac{1}{4}(\bar{V}_{0x}B_1 + \bar{V}_1'B_2)R^{-1}(B_1'\bar{V}_{0x} + B_2'\bar{V}_1), \quad \bar{V}_0(0) = 0,$$

$$(21) \quad 0 = s' + 2a_2'\bar{V}_2 + \bar{V}_{0x}(A_1 - B_1R^{-1}B_2'\bar{V}_2) + \bar{V}_1'(A_2 - B_2R^{-1}B_2'\bar{V}_2),$$

$$(22) \quad 0 = Q + \bar{V}_2A_2 + A_2'\bar{V}_2 - \bar{V}_2B_2R^{-1}B_2'\bar{V}_2.$$

At each fixed value of x , (22) is an algebraic Riccati equation for \bar{V}_2 . In view of (3) and $Q(x) > 0$, the unique positive definite solution \bar{V}_2 exists such that for all $x \in D$, the real parts of the eigenvalues of $\bar{A}_2 = A_2 - B_2R^{-1}B_2'\bar{V}_2$, denoted by $\text{Re}(\lambda(\bar{A}_2))$, are less than a negative constant. Thus \bar{A}_2 is nonsingular and \bar{V}_1 can be expressed in terms of \bar{V}_{0x} and \bar{V}_2 as

$$(23) \quad \bar{V}_1' = -[s' + 2a_2'\bar{V}_2 + \bar{V}_{0x}(A_1 - B_1R^{-1}B_2'\bar{V}_2)]\bar{A}^{-1}.$$

It is of crucial importance that the elimination of \bar{V}_1 from (21) results in an equation involving only \bar{V}_{0x} . For the well-posedness of the full problem it is necessary that the leading term \bar{V}_0 of (16) be identical to the solution L of the reduced problem.

LEMMA 1. *If assumptions III and IV are satisfied, then the unique positive definite solution $\bar{V}_0(x)$ of (20)–(22) exists in D and is identical to the solution $L(x)$ of the reduced problem (5), (6).*

Proof. It is shown in the appendix that eliminating \bar{V}_1 from (20), we obtain the Hamilton–Jacobi equation (11) with \bar{V}_{0x} in place of L_x , and hence $\bar{V}_0(x) \equiv L(x)$ with properties as in assumption IV.

By virtue of Lemma 1, \bar{V}_0 and \bar{V}_2 are solved independently from (11) and (22). This is the separation of time scales in the design of nonlinear regulators, analogous to the linear time-invariant design in [7].

Using \bar{V} , we derive the control

$$\begin{aligned}
 \bar{u} &= -\frac{1}{2}R^{-1}\left(B_1'\bar{V}'_x + \frac{1}{\mu}B_2'\bar{V}'_z\right) \\
 (24) \quad &= -\frac{1}{2}R^{-1}[B_1'\bar{V}'_{0x} + B_2'(\bar{V}_1 + 2\bar{V}_2z)] + O(\mu) \\
 &\equiv u_c + O(\mu),
 \end{aligned}$$

whose main part u_c is defined as *the composite control*. Eliminating \bar{V}_1 from (24) using (23) and following the derivation in [7], we can write u_c as

$$\begin{aligned}
 (25) \quad u_c &= -R_0^{-1}(s_0 + \frac{1}{2}B_0'\bar{V}'_{0x}) - R^{-1}B_2'\bar{V}_2[z + A_2^{-1}(a_2 - B_0R_0^{-1}(s_0 + \frac{1}{2}B_0'\bar{V}'_{0x}))] \\
 &= u_0 - R^{-1}B_2'\bar{V}_2(z + \bar{A}_2^{-1}\bar{a}_2),
 \end{aligned}$$

where

$$(26a) \quad \bar{A}_2(x) = A_2 - B_2R^{-1}B_2'\bar{V}_2,$$

$$(26b) \quad \bar{a}_2(x) = a_2 - \frac{1}{2}B_2R^{-1}(B_1'\bar{V}'_{0x} + B_2'\bar{V}_1), \quad \bar{a}_2(0) = 0.$$

Hence the composite control u_c consists of a slow control u_0 which optimizes the reduced system (5) and a fast control $-R^{-1}B_2'\bar{V}_2(z + \bar{A}_2^{-1}\bar{a}_2)$ which optimizes the fast part $(z + \bar{A}_2^{-1}\bar{a}_2)$ of z in the sense that \bar{V}_2 satisfies (22). Note that when z is not penalized in (2), that is when $Q(x) = 0$, but $\text{Re}\{\lambda(A_2)\} < 0$, then \bar{V}_2 is identically zero and u_c reduces to u_0 of (10). Stabilizing properties of the composite control u_c are established in the next section.

4. Stabilizing properties. System (1) controlled by u_c is

$$\begin{aligned}
 (27) \quad \dot{x} &= a_1 + A_1z + B_1u_c \equiv \bar{a}_1(x) + \bar{A}_1(x)z, & x(0) &= x_0, \\
 \mu\dot{z} &= a_2 + A_2z + B_2u_c \equiv \bar{a}_2(x) + \bar{A}_2(x)z, & z(0) &= z_0,
 \end{aligned}$$

where

$$\begin{aligned}
 (28) \quad \bar{a}_1 &= a_1 - \frac{1}{2}B_1R^{-1}(B_1'\bar{V}'_{0x} + B_2'\bar{V}_1), & \bar{a}_1(0) &= 0, \\
 \bar{A}_1 &= A_1 - B_1R^{-1}B_2'\bar{V}_2.
 \end{aligned}$$

With the change of variables

$$(29) \quad \eta = z + \bar{A}_2^{-1}\bar{a}_2$$

exhibiting η as the fast part of z , system (27) becomes

$$(30a) \quad \dot{x} = \bar{a}_0 + \bar{A}_1\eta, \quad x(0) = x_0,$$

$$\begin{aligned}
 (30b) \quad \mu\dot{\eta} &= \mu(\bar{A}_2^{-1}\bar{a}_2)_x\bar{a}_0 + [\bar{A}_2 + \mu(\bar{A}_2^{-1}\bar{a}_2)_x\bar{A}_1]\eta \\
 &\equiv \mu f(x) + [\bar{A}_2(x) + \mu F(x)]\eta, & \eta(0) &= z_0 + \bar{A}_2^{-1}(x_0)\bar{a}_2(x_0).
 \end{aligned}$$

Since the right-hand side of (30b) is an $O(\mu)$ perturbation of $\bar{A}_2(x)\eta$ and $\text{Re}\{\lambda(\bar{A}_2)\} < 0$ in D we expect that η will rapidly decay to an $O(\mu)$ quantity. This motivates the introduction of

$$(31) \quad U(x, \eta; \mathcal{E}) = \bar{V}_0(x) + \mathcal{E}\eta'\bar{V}_2(x)\eta.$$

as a tentative Lyapunov function for (30). Here \mathcal{E} is a small positive scalar to be determined. From assumptions III and IV, $\bar{V}_0(x)$ is positive definite and $\bar{V}_2(x) > 0$ in D . Hence U is positive definite for all $x \in D$ and $\eta \in R^m$. Furthermore, since $\bar{V}_0(x) = c_0 > 0$ for all x on the boundary of D , the surface

$$(32) \quad S(x, \eta; \mathcal{E}) = \{x, \eta : U(x, \eta; \mathcal{E}) = c_0\}$$

is closed in the $(n+m)$ -dimensional domain $x \in D$, $\eta \in R^m$. We define S_{in} to be the domain in the interior of S .

Let D_1 be a set strictly in the interior of D , that is, the boundary of D_1 does not intersect the boundary of D , and let E be a bounded set in R^m . The presence of \mathcal{E} in U extends S to encompass (x, z) for all $x \in D_1$ and for z in any prescribed set E . This crucial result is stated as follows.

LEMMA 2. *If assumption III and IV are satisfied, then there exists an $\mathcal{E} > 0$ such that the domain S_{in} contains all $x \in D_1$, $\eta \in E$.*

Proof. At each point $\hat{x} \in D_1$, the projection S onto the η subspace is the ellipsoid

$$(33) \quad \eta' \bar{V}_2(\hat{x}) \eta = (c_0 - \bar{V}_0(\hat{x}))/\mathcal{E},$$

implying that η extends to $O(1/\sqrt{\mathcal{E}})$. Hence for every \hat{x} , there exists an $\mathcal{E}(\hat{x})$ sufficiently small such that the ellipsoid (33) includes all $\eta \in E$. (Note that we must exclude the boundary of D because from (33) the projection of S at any point on the boundary of D is a single point $\eta = 0$.) Hence if we choose \mathcal{E}^* to be the smallest of such $\mathcal{E}(\hat{x})$, the domain S_{in} contains all $x \in D_1$, $\eta \in E$ for any $\mathcal{E} \in (0, \mathcal{E}^*]$.

By virtue of Lemma 2, the initial condition $\eta(0)$ of (30b), and hence $z(0)$ of (27), can be as far away from zero as $O(1/\sqrt{\mathcal{E}})$ and still be enclosed by S . We now examine the relationship between \mathcal{E} and μ .

Using (11), (22) and rearranging, we obtain the time derivative of U with respect to (30) as

$$(34) \quad \dot{U} = -g(x, \mathcal{E}, \mu) - \frac{\mathcal{E}}{2\mu} \xi' Q(x) \xi - \frac{\mathcal{E}}{\mu} \eta' M(x, \eta, \mathcal{E}, \mu) \eta$$

where

$$(35) \quad \begin{aligned} g &= g_1 - \frac{\mu}{2\mathcal{E}} y' Q^{-1} y, & g_1 &= p_0 - s_0' R_0^{-1} s_0 + \frac{1}{4} \bar{V}_{0x} B_0 R_0^{-1} B_0' \bar{V}_{0x}, \\ y &= \bar{A}_1' \bar{V}_{0x} + 2\mathcal{E} \bar{V}_2 f, & \xi &= \eta - \frac{\mu}{\mathcal{E}} Q^{-1} y, \\ M &= \frac{Q}{2} + \bar{V}_2 B_2 R^{-1} B_2' \bar{V}_2 - \mu (\bar{V}_2 \bar{F} + \bar{F}' \bar{V}_2) - \mu \dot{\bar{V}}_2. \end{aligned}$$

Since $\bar{V}_2 F_2 + \bar{F}' \bar{V}_2$ and $\dot{\bar{V}}_2$ are bounded for all x, η in S_{in} , and since $Q(x) > 0$ in D , it follows that there exists a $\mu_1^* > 0$ such that $M > 0$ for all x, η in S_{in} and for $\mu \in (0, \mu_1^*]$. Thus the last two terms in \dot{U} are positive definite. To ensure that $g(x, \mathcal{E}, \mu)$ is positive definite, we assume that the reduced problem also satisfies

V. The limit

$$(36) \quad \lim_{|x| \rightarrow 0} \frac{y' Q^{-1} y}{g_1} = k(\mathcal{E}) < \infty$$

exists for all fixed $\mathcal{E} > 0$.

Note that $k \geq 0$ because $y'Q^{-1}y$ is positive semidefinite and g_1 is positive definite. The limit (36) implies that there exists a domain \tilde{D} about $x = 0$ such that

$$(37) \quad y'Q^{-1}y \leq (1+k)g_1,$$

that is, such that for $\mu < 2\mathcal{E}/(1+k)$, g is positive definite in \tilde{D} ; see (35). Let $\bar{k}(\mathcal{E}) > 0$ be the minimum value of g_1 on the boundary of \tilde{D} . Hence in the domain

$$(38) \quad \tilde{D}_1(x) = \{x : g_1(x) < \bar{k}\},$$

g is positive definite. On the other hand, since D is bounded, there exists a $k_1(\mathcal{E}) > 0$ such that $y'Q^{-1}y < k_1$ for all $x \in D$, that is, such that g is positive definite when x is not in the domain

$$(39) \quad \bar{D}(x) = \{x : g_1(x) < \mu k_1/2\mathcal{E}\}$$

about the origin. But for $\mu < 2\mathcal{E}\bar{k}/k_1$, $\bar{D} \subset \tilde{D}_1$, implying that g is positive definite in D . Thus \dot{U} is negative definite for all x, η contained in S_{in} . We now conclude that U is a Lyapunov function for (30) guaranteeing that $x = 0, \eta = 0$ is asymptotically stable for all $x \in D_1, \eta \in E$ and for $\mu \in (0, \mu^*]$, where

$$(40) \quad \mu^* = \min\left(\frac{2\mathcal{E}}{1+k}, \frac{2\mathcal{E}\bar{k}}{k_1}, \mu_1^*\right).$$

Returning from the η variable to the z variable via $z = \eta - \bar{A}_2^{-1}\bar{a}_2$, we obtain for all $x \in D_1, \eta \in E$ a corresponding bounded domain E_1 for z . We summarize the above discussions on the asymptotic stabilizing property of u_c in (24) as follows.

THEOREM 1. *If assumptions I–V are satisfied, then there exists a $\mu^* > 0$ such that for all $\mu \in (0, \mu^*]$ and for all $x \in D_1$ and z in any prescribed bounded set E_1 , the origin $x = 0, z = 0$ of the feedback system (1) controlled by the composite control u_c is asymptotically stable.*

Theorem 1 can be applied in two different directions. As outlined above, for any given D_1 and E_1 , we first find \mathcal{E}^* such that S_{in} of (32) contains all $x \in D_1, z \in E_1$. Then we find μ^* from (40). This direction is suitable when μ is a parameter at the designer's disposal, such as a gain factor [9]. In the other direction, if μ represents some given physical parameters, such as time constants, we use its value to determine the smallest \mathcal{E} such that \dot{U} of (34) is negative definite, that is we find the largest D_1 and E_1 .

As a special case of assumption V, consider that the origin $x = 0$ of the reduced system (12) is exponentially stable. Then near the origin, $p_0 - s_0'R_0^{-1}s_0, \bar{V}_0$ grow as $|x|^2$, and $|\bar{V}_{0x}|, |\bar{a}_0|$ grow as $|x|$, and we can find positive constants k_2, \dots, k_9 and δ such that

$$(41) \quad \begin{aligned} k_2|x|^2 &\leq p_0 - s_0'R_0^{-1}s_0 \leq k_3|x|^2, & k_4|x|^2 &\leq \bar{V}_0 \leq k_5|x|^2 \\ k_6|x| &\leq |\bar{V}_{0x}| \leq k_7|x|, & k_8|x| &\leq |\bar{a}_0| \leq k_9|x| \end{aligned}$$

for all $|x| < \delta$. It follows from (41) that there exists a fixed $k_{10}(\mathcal{E}) > 0$ such that

$$(42) \quad y'Q^{-1}y \leq k_{10}|x|^2$$

and the limit (36) is bounded by

$$(43) \quad \lim_{|x| \rightarrow 0} \frac{y'Q^{-1}y}{g_1} \leq \lim_{|x| \rightarrow 0} \frac{k_{10}|x|^2}{k_2|x|^2} = \frac{k_{10}}{k_2}$$

satisfying assumption V.

In this case a claim stronger than Theorem 1 can be made.

COROLLARY 1. *If assumptions I–IV are satisfied and the origin $x = 0$ of the reduced system is exponentially stable, then the conclusion of Theorem 1 holds and moreover the origin $x = 0, z = 0$ of (27) is exponentially stable.*

Proof. The first part of the corollary follows from Theorem 1. The second part follows from the linearization of (27) at the origin

$$(44) \quad \begin{bmatrix} \delta \dot{x} \\ \delta \dot{z} \end{bmatrix} = \begin{bmatrix} \frac{\partial \bar{a}_1(0)}{\partial x} & \bar{A}_1(0) \\ \frac{1}{\mu} \frac{\partial \bar{a}_2(0)}{\partial x} & \frac{1}{\mu} \bar{A}_2(0) \end{bmatrix} \begin{bmatrix} \delta x \\ \delta z \end{bmatrix}.$$

The system matrix of (44) has one group of n small eigenvalues $O(\mu)$ close to those of $(\partial \bar{a}_1 / \partial x) - \bar{A}_1 \bar{A}_2^{-1} (\partial \bar{a}_2 / \partial x)|_{x=0}$ and another group of m large eigenvalues $O(1)$ close to those of $(1/\mu) \bar{A}_2(0)$ [8]. But $\bar{a}_1 - \bar{A}_1 \bar{A}_2^{-1} \bar{a}_2 = \bar{a}_0$ and $(\partial \bar{a}_0 / \partial x)|_{x=0} = (\partial \bar{a}_1 / \partial x) - \bar{A}_1 \bar{A}_2^{-1} (\partial \bar{a}_2 / \partial x)|_{x=0}$ as $\bar{a}_2(0) = 0$. Thus the real parts of the eigenvalues of the system matrix of (44) are all negative and $x = 0, z = 0$ is exponentially stable.

If the origin $x = 0$ of the reduced system is only asymptotically stable but not exponentially stable, then in general g need not be positive definite for all $x \in D$. This situation includes the critical case when the linearized model does not provide any stability information as clarified by the example in § 6. For this situation the system is now shown to possess a weaker stability property, that is, its trajectories tend to a small sphere around the origin. Define the domain in R^n

$$(45) \quad \rho(x) = \{x : g(x; \mathcal{E}, \mu) \leq 0\},$$

which is contained in the domain \bar{D} of (39). Due to the presence of μ in (34), \dot{U} may be positive only if $x \in \rho(x)$ and $\eta = O(\mu)$. Otherwise, \dot{U} is negative. If we define the surface

$$(46) \quad \pi(x, z) = \{x, z : x \in \rho(x; \mu), z = -\bar{A}_2^{-1}(x) \bar{a}_2(x)\}$$

about the origin in R^{m+n} , u_c defined by (24) is a stabilizing control in the following sense.

THEOREM 2. *If assumptions I–IV are satisfied, then there exists a $\mu^* > 0$ such that for all $\mu \in (0, \mu^*]$, the feedback control (24) steers all $x \in D_1, z \in E_1$ of the full system $O(\mu)$ close to the surface $\pi(x, z)$.*

Proof. Since $U > 0$ and $\dot{U} < 0$ except for $x \in \rho(x)$ and $\eta = O(\mu)$, x converges to $\rho(x)$ and η decays to an $O(\mu)$ quantity. Thus in the x, z variables, (x, z) converges to an $O(\mu)$ neighborhood of the surface $\pi(x, z)$.

In the case where the fast transients of z in (1) are exponentially stable, that is, $A_2(x)$ is stable for all $x \in D$, and we are only concerned with the optimality of the reduced system (5), then the z -independent reduced control u_0 of (10) stabilizes the full system (1) with essentially the same stabilizing properties as u_c of (24). We shall not repeat the argument.

An attractive feature of the controls u_c and u_0 is that they do not require the knowledge of the actual value of μ provided that it is sufficiently small. When appropriately implemented, these controls stabilize the full system (1) and achieve optimality of the reduced system, and in the case of u_c , also optimality of the fast part of z . The above results also answer the question of well-posedness by giving the conditions under which the same optimal reduced order system is obtained when μ is set equal to zero either when system (1) is uncontrolled or when it is controlled by u_c or u_0 . In contrast to many other singular perturbation results which require μ to be

sufficiently small, this section provides a method to compute an estimate of allowable values of μ given a stability domain or vice versa.

5. A formal expansion and near-optimality. The equation (16) only satisfies the Hamilton–Jacobi equation (15) to $O(\mu)$ order. We now propose to solve (15) by expanding V formally as a nested infinite power series. If this power series is convergent, then the optimal solution V of (15) exists. For x, z near the origin, it has been shown in [1] that the optimal solution exists and possesses a power series expansion when system (1) after linearization at the origin is stabilizable and the state in the quadratic approximation of J is observable. Here we are interested in a power series of V which satisfies (15) to any order of μ .

Since system (1) is linear in z and J is quadratic in z , the optimal value function can be expanded as a power series in the components of z [2]. In addition, since z is the fast variable, the z terms in the optimal value function are multiplied by appropriate powers of μ [5]. In view of these two characteristics, we seek a solution of (15) in the form

$$\begin{aligned}
 V(x, z, \mu) = & V_0(x, \mu) + \mu \sum_{j=1}^m V_{1j}(x, \mu) z_j + \mu \sum_{j=1}^m \sum_{k=1}^m V_{2jk}(x, \mu) z_j z_k \\
 (47) \quad & + \mu^2 \sum_{j=1}^m \sum_{k=1}^m \sum_{q=1}^m V_{3jkq}(x, \mu) z_j z_k z_q + \cdots \\
 & + \mu^{i-1} \sum_{j_1=1}^m \sum_{j_2=1}^m \cdots \sum_{j_i=1}^m V_{ij_1 j_2 \cdots j_i}(x, \mu) z_{j_1} z_{j_2} \cdots z_{j_i} + \cdots, \\
 & V_0(0, \mu) = 0,
 \end{aligned}$$

where $V_{ij_1 j_2 \cdots j_i}$ is the (j_1, j_2, \dots, j_i) element of the completely symmetric generalized matrix¹ V_i of dimension m^i and z_j is the j th component of z . The summation signs in (47) and in other equations in the paper will be omitted when there is no confusion as to which indices j_1, j_2, \dots, j_i are being summed. The partial derivatives $V_x, V_{z_1}, \dots, V_{z_m}$ expressed in terms of the vector x and the scalars z_1, \dots, z_m are

$$(48a) \quad V_x = V_{0x} + \mu V_{1jx} z_j + \mu V_{2j k x} z_j z_k + \cdots$$

$$(48b) \quad V_{z_i} = \mu V_{1i} + 2\mu V_{2ij} z_j + 3\mu^2 V_{3ijk} z_j z_k + \cdots, \quad i = 1, 2, \dots, m,$$

where the summation signs over j, k are omitted.

For the series (47) to satisfy (15) as an identity, we first rewrite (15) in terms of the vector x and the scalars z_1, \dots, z_m ,

$$\begin{aligned}
 0 = & p + s_i z_i + Q_{ij} z_i z_j + V_x (a_1 + A_{1i} z_i) + \frac{1}{\mu} V_{z_i} (a_{2i} + A_{2ij} z_j) \\
 (49) \quad & - \frac{1}{4} \left(V_x B_1 + \frac{1}{\mu} V_{z_i} B_{2i} \right) R^{-1} \left(B_1' V_x' + \frac{1}{\mu} B_{2i}' V_{z_i} \right),
 \end{aligned}$$

where s_i, a_{2i} are the i th components of the vectors s, a_2 , respectively, A_{1i} is the i th column of the matrix A_1 , B_{2i} is the i th row of B_2 , Q_{ij}, A_{2ij} are the (i, j) elements of Q, A_2 , respectively, and the summation signs over the indices i, j are omitted. Then, upon substituting (48) into (49) and equating the coefficients of the like powers of z_i , we

¹ The (j_1, j_2, \dots, j_i) elements of V_i are identical for all permutations of the indices j_1, j_2, \dots, j_i [6].

obtain

$$(50a) \quad 0 = p + V_{0x}a_1 + V_{1i}a_{2i} - \frac{1}{4}(V_{0x}B_1 + V_{1i}B_{2i})R^{-1}(B_1'V_{0x}' + B_{2i}'V_{1i}'),$$

$$V_0(0, \mu) = 0,$$

$$(50b) \quad 0 = s_i + V_{0x}A_{1i} + \mu V_{1ix}a_1 + V_{1j}A_{2ji} + 2V_{2ij}a_{2j} - \frac{1}{2}(V_{0x}B_1 + V_{1j}B_j)$$

$$R^{-1}(\mu B_1'V_{1ix}' + 2B_{2j}'V_{2ji}'), \quad i = 1, 2, \dots, m,$$

$$(50c) \quad 0 = Q_{ij} + \mu V_{2ijx}a_1 + \mu(V_{1ix}A_{1j})_s + 2(V_{2ik}A_{2kj})_s + 3\mu V_{3ijk}a_{2k}$$

$$- \frac{1}{2}(V_{0x}B_1 + V_{1k}B_{2k})R^{-1}(\mu B_1'V_{2ijx}' + 3\mu B_{2k}'V_{3kij}')$$

$$- \frac{1}{4}(\mu V_{1ix}B_1 + 2V_{2ik}B_{2k})R^{-1}(\mu B_1'V_{1jx}' + 2B_{2k}'V_{2kj}'),$$

$$i, j = 1, 2, \dots, m,$$

$$(50d)^2 \quad 0 = \mu^2 V_{3ijkx}a_1 + \mu(V_{2ijx}A_{1k})_s + 4\mu^2 V_{4ijkq}a_{2q} + 3\mu(V_{3ijq}A_{2qk})_s$$

$$- \frac{1}{2}(V_{0x}B_1 + V_{1q}B_{2q})R^{-1}(\mu^2 B_1'V_{3ijkx}' + 4\mu^2 B_{2q}'V_{4ijkq}')$$

$$- \frac{1}{2}((\mu V_{1ix}B_1 + 2V_{2iq}B_{2q})R^{-1}(\mu B_1'V_{2ijkx}' + 3\mu B_{2q}'V_{3qjk}'))_s,$$

$$\vdots$$

$$i, j, k = 1, 2, \dots, m,$$

where the right-hand sides of (50a), (50b), (50c), (50d), \dots , are the coefficients of the z -independent terms and of the $z_i, z_i z_j, z_i z_j z_k, \dots$, terms, respectively. Because of symmetry, there are $m(m+1)/2$ equations in (50c), $m(m+1)(m+2)/6$ equations in (50d) and in general, $(\prod_{k=0}^{i-1} (m+k))/i!$ equations when the coefficients of $z_{i_1} z_{i_2} \dots z_{i_i}$, $j_1, j_2, \dots, j_i = 1, 2, \dots, m$, are equated.

For a simplified treatment of these equations we now exploit the presence of the small singular perturbation parameter μ . We expand each coefficient of (47) as a power series in μ :

$$(51) \quad V_i(x, \mu) = \sum_{j=0}^{\infty} \mu^j V_i^j(x), \quad i = 0, 1, 2, \dots,$$

where the boundary condition of V_0^j is $V_0^j(0) = 0$, $j = 0, 1, 2, \dots$. The expressions (51) substituted into equations (50) are to satisfy them as identities in μ . Equating the coefficients of the like powers in μ , we generate sets of equations for V_i^j , $i, j = 0, 1, 2, \dots$. The first set of equations obtained by equating the μ -independent parts in (50a), (50b), (50c), are precisely equations (20), (21), (22), respectively. Hence from the uniqueness of solutions to (20), (21), (22), conclude that

$$(52) \quad V_0^0 = \bar{V}_0 = L, \quad V_1^0 = \bar{V}_1, \quad V_2^0 = \bar{V}_2,$$

and \bar{V} thus consists of the leading terms of V .

The second set of equations in matrix form

$$(53a) \quad 0 = V_{0x}^1 \bar{a}_1 + V_1^{1'} \bar{a}_2, \quad V_0^1(0) = 0,$$

$$(53b) \quad 0 = V_{0x}^1 \bar{A}_1 + \bar{a}_1' V_{1x}^{0'} + V_1^{1'} \bar{A}_2 + 2\bar{a}_2' V_2^1,$$

$$(53c) \quad 0 = V_{2x}^0 \bar{a}_1 + \frac{1}{2}(V_{1x}^0 \bar{A}_1 + \bar{A}_1' V_{1x}^{0'}) + V_2^1 \bar{A}_2 + \bar{A}_2' V_2^1 + 3(V_3^0 \bar{a}_2),$$

² The subscript s denotes the symmetrization operation of generalized matrices [6]. For example,

$$(V_{2ik}A_{2kj})_s = \frac{1}{2}(V_{2ik}A_{2kj} + V_{2jk}A_{2ki})$$

$$(V_{3ijq}A_{2qk})_s = \frac{1}{6}(V_{3ijq}A_{2qk} + V_{3jiq}A_{2qk} + V_{3ikq}A_{2qj} + V_{3kjq}A_{2qi} + V_{3ijk}A_{2qi} + V_{3kji}A_{2qk}).$$

$$(53d) \quad 0 = 3(V_3^0 \bar{A}_2)_s + (V_{2x}^0 \bar{A}_1)_s,$$

obtained by equating the μ terms in (50a), (50b), (50c), (50d), respectively, involve only the unknown terms V_{0x}^1 , V_1^1 , V_2^1 and V_3^0 . In (53) the multiplication of an $n_1 \times n_2 \times n_3$ matrix by an $n_3 \times n_4$ matrix results in an $n_1 \times n_2 \times n_4$ matrix. For convenience we suppress the last dimension of the $m \times m \times 1$ matrices $(V_{2x}^0 \bar{a}_1)$ and $(V_3^0 \bar{a}_2)$ and regard them as $m \times m$ matrices. Since \bar{A}_2 is stable, (53d) and (53c) can be solved sequentially for V_3^0 and V_2^1 , respectively. Then V_1^1 can be solved from (53b) and its substitution into (53a) results in the partial differential equation

$$(54) \quad 0 = V_{0x}^1 \bar{a}_0 - (\bar{a}_1' V_{1x}^0 + 2\bar{a}_2' V_2^1) \bar{A}_2^{-1} \bar{a}_2, \quad V_0^1(0) = 0.$$

In general, in equating the μ^i terms we obtain the $(i+1)$ st set of equations involving the unknown terms V_{0x}^i , V_1^i , V_2^i , V_3^{i-1} , \dots , V_{i+2}^0 . The terms V_{i+1}^0 , V_i^1 , \dots , V_2^{i-1} are solved for sequentially and then V_0^{i-1} is to be solved from an equation similar to (41).

The main accomplishment of the nested expansions is that the first set of equations (20)–(22) can be solved independently for the first three zeroth order terms V_0^0 , V_1^0 , and V_2^0 . Similarly, (53) and the subsequent sets of equations can be solved independently for V_0^i , V_1^i , \dots , V_{i+2}^0 . These equations are dependent only on x and not on z or μ . A further simplifying property is that at the first stage the equations (11), (22) for V_0^0 and V_2^0 are decoupled.

The approximation obtained by expanding V of (47), (51) to the i th set of equations is stated in the following theorem.

THEOREM 3. *Suppose that the solutions to the i -th set of equations of V exist and let V^i be the truncated series of (47), (51) including all the terms V_j^i up to the i -th set. Then the control*

$$(55) \quad u_i = -\frac{1}{2} R^{-1} \left(B_i' V_x^{i'} + \frac{1}{\mu} B_2' V_z^{i'} \right)$$

is near-optimal in the sense that V^i satisfies the Hamilton–Jacobi equation (15) to an $O(\mu^i)$ error.

Proof. Substituting the V_j^i terms into (15) and using the first i set of equations of V , the coefficients of μ^k terms, $k < i$, in the resulting equation vanish, implying $O(\mu^i)$ near-optimality.

Thus Theorem 3 implies that u_c of (24) is an $O(\mu)$ near-optimal control because it is an $O(\mu)$ approximation of u_1 which achieves $O(\mu)$ near-optimality. In general, retaining only the μ^j terms, $k < i$, in u_i , the resulting control also is $O(\mu^i)$ near-optimal in the sense of Theorem 3.

Repeating the derivation in § 4, we can show that u_i stabilizes the full system (1) with similar stabilizing properties as u_c of (24). We first introduce the x , $\eta = z + \bar{A}_2^{-1} \bar{a}_2$ variables and consider U in (31) as a tentative Lyapunov function. The analysis is more cumbersome but results similar to Theorems 1 and 2 and Corollary 1 can be established.

6. Discussion and example. The computational advantage of the proposed procedure is that all the terms of V in (47), (51) are obtained from equations involving the slow variable x only. Moreover V_0^0 and V_2^0 are solved for independently. Explicit consideration of the initial boundary layer is avoided and it is optimally stabilized by the z variable feedback. Furthermore using the x , η variables an estimate of the domain of stability is easily obtained. Alternatively, for a stability domain to encompass a prescribed bounded set $\eta \in E \subset R^m$ a bound for μ can be determined.

Several aspects of the design procedure and the stability properties of the resulting feedback system are now illustrated by considering the optimal control problem of the second order system

$$(56) \quad \dot{x} = xz, \quad \mu \dot{z} = -z + u,$$

with respect to the performance index

$$(57) \quad J = \int_0^\infty (x^4 + \frac{1}{2}z^2 + \frac{1}{2}u^2) dt.$$

Solving the reduced problem we obtain $L = V_0^0 = x^2$ and $u_0 = -x^2$. The optimally controlled reduced system (12) is $\dot{x} = -x^3$ and its unique asymptotically stable equilibrium is $x = 0$. Note that the linearization of the reduced system fails to provide any stability information at $x = 0$. Let D be the interval $[-1, 1]$, that is, $L = c_0 = 1$ at $x = \pm 1$ by assumption IV.

The pair $(A_2, B_2) = (-1, 1)$ satisfies (3) and we can solve (22) for $V_2^0 = \frac{1}{2}(\sqrt{2} - 1)$ such that $\bar{A}_2 = -\sqrt{2}$. Then the substitution of $V_0^0 = L = x^2$ and V_2^0 into (23) yields the following expressions for (24) and (16):

$$(58) \quad u_c = -(\sqrt{2}x^2 + (\sqrt{2} - 1)z),$$

$$(59) \quad \bar{V} = x^2 + \mu\sqrt{2}x^2z + \mu\frac{1}{2}(\sqrt{2} - 1)z^2.$$

The resulting feedback system is

$$(60) \quad \dot{x} = xz, \quad \mu \dot{z} = -\sqrt{2}x^2 - \sqrt{2}z.$$

This result is essentially nonlinear since the linearization of (60) at $x = 0, z = 0$ does not provide any stability information. After the change of variables $\eta = z + x^2$, system (60) becomes

$$(61) \quad \dot{x} = -x^3 + x\eta, \quad \mu \dot{\eta} = -2\mu x^4 - (\sqrt{2} - 2\mu x^2)\eta.$$

Since we require $|x| \leq 1$, μ is restricted to be less than $1/\sqrt{2}$. The tentative Lyapunov function (31) is

$$(62) \quad U(x, \eta; \mathcal{E}) = x^2 + \frac{1}{2}(\sqrt{2} - 1)\mathcal{E}\eta^2.$$

If we require that the initial conditions of (61) be in $|x| \leq .8, |\eta| \leq 5$, then we must set \mathcal{E} to be less than .0695 in order for the ellipse

$$(63) \quad S(x, \eta; \mathcal{E}) = \{x, \eta : U = x^2 + \frac{1}{2}(\sqrt{2} - 1)\mathcal{E}\eta^2 = 1\}$$

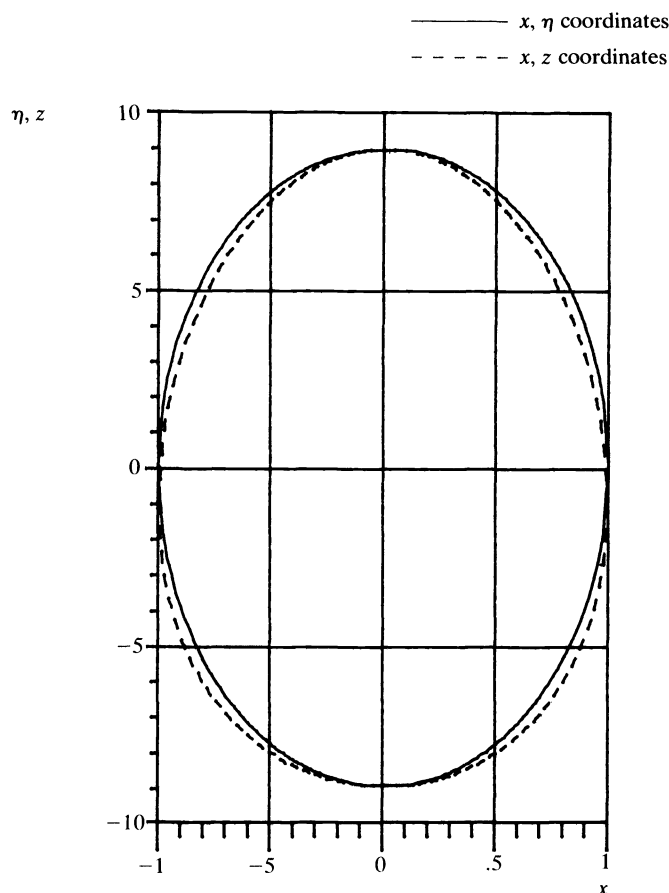
to enclose these initial conditions. Plots of S in the x, η coordinates and the x, z coordinates for $\mathcal{E} = .06$ are shown in Fig. 1. The time derivative of U with respect to (61) is

$$(64) \quad \dot{U} = -\left(g_1 - \frac{\mu}{\mathcal{E}}y^2\right) - \frac{\mathcal{E}}{4\mu}\xi^2 - \frac{\mathcal{E}}{\mu}M\eta^2,$$

where

$$(65) \quad \begin{aligned} g_1 &= 2x^4, & y &= 2(1 - \mathcal{E}(\sqrt{2} - 1)x^2)x^2, \\ \xi &= \eta - \frac{2\mu}{\mathcal{E}}y, & M &= \frac{7}{4} - \sqrt{2} - 2\mu(\sqrt{2} - 1)x^2. \end{aligned}$$

Since $\lim_{x \rightarrow 0} y^2/g_1 = 2$, assumption V is satisfied. For all x, η in the interior of S and

FIG. 1. Plot of S in (63).

$\mathcal{E} = .06$, \dot{U} is negative definite for all $\mu \in (0, .03]$. Hence $x = 0$, $z = 0$ is asymptotically stable for all $|x| \leq .8$, $|z + x^2| \leq 5$ and $\mu \in (0, .03]$. Furthermore, \bar{V} satisfies the Hamilton-Jacobi equation (15) with an error of $\mu 2\sqrt{2}x^2z^2$.

If we are only interested in the optimality of the reduced problem and consider the z -part as due to "system parasitics," we can apply the reduced control u_0 to (56) as $A_2 = -1$ is stable. System (56) controlled by u_0 is

$$(66) \quad \dot{x} = xz, \quad \mu \dot{z} = -x^2 - z.$$

Transforming z to $\eta = z + x^2$, system (66) becomes

$$(67) \quad \dot{x} = -x^3 + x\eta, \quad \mu \dot{\eta} = -2\mu x^2 - (1 - 2\mu x^2)\eta.$$

We use U in (62) as a Lyapunov function for (67) and the time derivative of U with respect to (67) is

$$(68) \quad \begin{aligned} \dot{U} = & - \left[2 - \frac{\mu}{\mathcal{E}} 2(\sqrt{2} - 1)(\sqrt{2} + 1 - \mathcal{E}x^2)^2 \right] x^4 - \frac{\mathcal{E}}{\mu} \frac{\sqrt{2} - 1}{2} \left[\eta - \frac{\mu}{\mathcal{E}} 2(\sqrt{2} + 1 - \mathcal{E}x^2)x^2 \right] \\ & - \frac{\mathcal{E}}{\mu} (\sqrt{2} - 1) \left(\frac{1}{2} - 2\mu x^2 \right) \eta^2. \end{aligned}$$

Thus for all x, η enclosed in S and $\mathcal{E} = .06$, \dot{U} is negative definite for all $\mu \in (0, .02]$. Hence $x = 0, z = 0$ of (66) is asymptotically stable for all $|x| \leq .8, |z + x^2| \leq 5, \mu \in (0, .02]$.

To obtain an $O(\mu^2)$ approximation of V in the sense of Theorem 3, we solve (53) for higher order terms of V_i^j and obtain

$$(69) \quad u_2 = u_c - \mu 2x^2 z,$$

$$(70) \quad V^2 = \bar{V} + \mu \frac{x^4}{\sqrt{2}} + \mu^2 x^2 z^2.$$

System (56) controlled by u_2 becomes

$$(71) \quad \dot{x} = xz, \quad \mu \dot{z} = -\sqrt{2}x^2 - (\sqrt{2} + \mu 2x^2)z,$$

or, in the $x, \eta = z + x^2$ variables,

$$(72) \quad \dot{x} = -x^3 + x\eta, \quad \mu \dot{\eta} = -\sqrt{2}\eta,$$

which is globally asymptotically stable for all $\mu > 0$. Furthermore, V^2 satisfies (15) with an error of $\mu^2(8x^4 z^2 + 2x^2 z^3)$.

7. Conclusions. A nested power series expansion method has been proposed for solving the optimal control problem of a class of nonlinear singularly perturbed systems. The terms in the expansion V are obtained from equations involving only the slow variable x . In addition, V_0^0 and V_2^0 are solved for independently. Explicit consideration of the initial boundary layer is avoided and it is optimized by the z variable feedback. Sufficient conditions are obtained such that feedback controls using truncated series stabilize the nonlinear systems and the stability domain can encompass large initial conditions of z . These truncated controls can achieve near-optimality of $O(\mu)$, $O(\mu^2)$, etc. In particular, an $O(\mu)$ near-optimal feed-back control can be implemented without knowing the value of the small parameter μ . The results apply to essentially nonlinear problems.

Appendix. Substituting (23) into (20) and rearranging yields

$$0 = X_1 + V_{0x}X_2 - \frac{1}{4}V_{0x}X_3 V_{0x}',$$

where

$$X_1 = p - (s' + 2a_2' V_2) \bar{A}_2^{-1} a_2 - (\frac{1}{2}s' + a_2' V_2) \bar{A}_2^{-1} B_2 R^{-1} B_2' \bar{A}_2^{-1} (\frac{1}{2}s + V_2 a_2),$$

$$X_2 = \tilde{a}_2 + \tilde{B}_0 R^{-1} B_2' \bar{A}_2^{-1} (\frac{1}{2}s + V_2 a_2), \quad X_3 = \tilde{B}_0 R^{-1} \tilde{B}_0',$$

$$\tilde{a}_0 = a_1 - (A_1 - B_1 R^{-1} B_2' V_2) \bar{A}_2^{-1} a_2, \quad \tilde{B}_0 = B_1 - (A_1 - B_1 R^{-1} B_2' V_2) \bar{A}_2^{-1} B_2,$$

$$\bar{A}_2 = A_2 - B_2 R^{-1} B_2 V_2,$$

and the superscript 0 in V_{0x} and V_2^0 has been dropped. Let $H = I + R^{-1} B_2' V_2 \bar{A}_2^{-1} B_2$. Then $H^{-1} = I - R^{-1} B_2' V_2 \bar{A}_2^{-1} B_2$ and $H'^{-1} R H^{-1} = R + B_2' \bar{A}_2^{-1} Q \bar{A}_2^{-1} B_2 = R_0$. Thus $\tilde{B}_0 = B_1 H - A_1 \bar{A}_2^{-1} B_2 = B_0 H$. Hence $X_3 = B_0 R_0^{-1} B_0'$. Also,

$$\begin{aligned} X_2 &= a_0 + B_0 R_0^{-1} [(R + B_2' \bar{A}_2^{-1} Q \bar{A}_2^{-1} B_2) R^{-1} B_2' V_2 \bar{A}_2^{-1} + B_2' \bar{A}_2^{-1} V_2] a_2 \\ &\quad + \frac{1}{2} B_0 R_0^{-1} B_2' \bar{A}_2^{-1} s \\ &= a_0 + B_0 R_0^{-1} B_2' \bar{A}_2^{-1} (A_2' V_2 + Q \bar{A}_2^{-1} B_2 R^{-1} B_2' V_2 + V_2 A_2 - V_2 B_2 R^{-1} B_2' V_2) \bar{A}_2^{-1} \\ &\quad + \frac{1}{2} B_0 R_0^{-1} B_2' \bar{A}_2^{-1} s \\ &= a_0 - B_0 R_0^{-1} s_0. \end{aligned}$$

Furthermore, $\bar{A}_2^{-1}B_2R^{-1}B_2'\bar{A}_2'^{-1} = A_2^{-1}B_2HR^{-1}H'B_2'A_2'^{-1} = A_2^{-1}B_2R_0^{-1}B_2'A_2'^{-1}$ and

$$\begin{aligned}\bar{A}_2^{-1} &= A_2^{-1} + A_2^{-1}B_2R^{-1}B_2'V_2\bar{A}_2'^{-1} \\ &= A_2^{-1} + A_2^{-1}B_2R_0^{-1}B_2'(V_2 + A_2'^{-1}QA_2^{-1}B_2R^{-1}B_2'V_2)\bar{A}_2'^{-1} \\ &= A_2^{-1} - A_2^{-1}B_2R_0^{-1}B_2'A_2'^{-1}QA_2^{-1} - A_2^{-1}B_2R_0^{-1}B_2'A_2'^{-1}V_2.\end{aligned}$$

Thus X_1 becomes

$$\begin{aligned}X_1 &= p - s'A_2^{-1}a_2 + s'A_2^{-1}B_2R_0^{-1}B_2'A_2'^{-1}QA_2^{-1} - \frac{1}{4}s'A_2^{-1}B_2R_0^{-1}B_2'A_2'^{-1}s \\ &\quad + a_2'V_2A_2^{-1}B_2R_0^{-1}B_2'A_2'^{-1}V_2a_2 - a_2'(V_2\bar{A}_2^{-1} + \bar{A}_2'^{-1}V_2)a_2.\end{aligned}$$

But

$$\begin{aligned}V_2\bar{A}_2^{-1} + \bar{A}_2'^{-1}V_2 &= -V_2A_2^{-1} - A_2'^{-1}V_2 + V_2A_2^{-1}B_2R_0^{-1}B_2'A_2'^{-1}QA_2^{-1} \\ &\quad + A_2'^{-1}QA_2^{-1}B_2R_0^{-1}B_2'A_2'^{-1}V_2 \\ &\quad + 2V_2A_2^{-1}B_2R_0^{-1}B_2'A_2'^{-1}V_2 \\ &= A_2'^{-1}QA_2^{-1} - A_2'^{-1}V_2B_2R^{-1}B_2'V_2A_2^{-1} \\ &\quad + (V_2 + A_2'^{-1}Q)A_2^{-1}B_2R_0^{-1}B_2'A_2'^{-1}(V_2 + QA_2^{-1}) \\ &\quad + V_2A_2^{-1}B_2R_0^{-1}B_2'A_2'^{-1}V_2 - A_2'^{-1}QA_2^{-1}B_2R_0^{-1}B_2'A_2'^{-1}QA_2^{-1},\end{aligned}$$

and

$$\begin{aligned}A_2'^{-1}V_2B_2R^{-1}B_2'V_2A_2^{-1} &= [-(V_2 + A_2'^{-1}Q_2)A_2^{-1} \\ &\quad + A_2'^{-1}V_2B_2R^{-1}B_2'V_2A_2^{-1}]B_2R^{-1}B_2'V_2A_2^{-1},\end{aligned}$$

that is,

$$\begin{aligned}A_2'^{-1}V_2B_2R^{-1}B_2'V_2A_2^{-1} &= -(V_2 + A_2'^{-1}Q_2)A_2^{-1}B_2R^{-1}B_2'V_2\bar{A}_2'^{-1} \\ &= (V_2 + A_2'^{-1}Q)A_2^{-1}B_2R^{-1}B_2'A_2'^{-1}(QA_2^{-1} + V_2),\end{aligned}$$

implying $X_1 = p_0 - s_0'R_0^{-1}s_0$. Hence elimination of V_1 from (20) yields the Hamilton-Jacobi equation (11) of the reduced problem.

REFERENCES

- [1] D. L. LUKES, *Optimal regulation of nonlinear dynamical systems*, this Journal, 7 (1969), pp. 75–100.
- [2] Y. NISHIKAWA, N. SANNOMIYA AND H. ITAKURA, *A method for suboptimal design of nonlinear feedback systems*, Automatica, 7 (1971), pt. 703–712.
- [3] P. SANNUTI, *Asymptotic series solution of singularly perturbed optimal control problems*, Ibid., 10 (1974), pp. 183–194.
- [4] R. E. O'MALLEY, JR., *Boundary layer methods for certain nonlinear singularly perturbed optimal control problems*, J. Math. Anal. Appl., 45 (1974), pp. 468–484.
- [5] P. V. KOKOTOVIC, R. E. O'MALLEY, JR. AND P. SANNUTI, *Singular perturbations and order reduction in control theory—An overview*, Automatica, 12 (1976), pp. 123–132.
- [6] R. L. BISHOP AND S. I. GOLDBERG, *Tensor Analysis on Manifolds*, Macmillan, New York, 1968.
- [7] J. H. CHOW AND P. V. KOKOTOVIC, *A decomposition of near-optimum regulators for systems with slow and fast modes*, IEEE Trans. Automatic Control, AC-21 (1976), pp. 701–705.
- [8] P. V. KOKOTOVIC AND A. H. HADDAD, *Singular perturbation of a class of time-optimal controls*, Ibid., AC-20 (1975), pp. 163–164.
- [9] K.-K. D. YOUNG, P. V. KOKOTOVIC AND V. I. UTKIN, *A singular perturbation analysis of high gain feedback systems*, Ibid., AC-22 (1977), pp. 931–938.

DUALITY METHODS IN THE CONTROL OF DENSITIES*

JEAN-MICHEL BISMUT†

Abstract. The purpose of this paper is to show the relation between the Davis and Varaiya maximum principle for controlled diffusions and the maximum principle previously obtained by the author.

1. Introduction. The purpose of this paper is to prove that the extremality relations obtained by Davis and Varaiya in [6] may be derived, under certain conditions, from the maximum principle previously obtained by the author in [2] (Theorem V-1).

Let us consider the stochastic differential equation:

$$(1.1) \quad \begin{aligned} dx &= b(\omega, t, u(\omega, t)) dt + \sigma(t, x_t) d\beta, \\ x(0) &= x \end{aligned}$$

where (Ω, \mathcal{F}, P) is the space of continuous functions endowed with the increasing family of σ -fields $\mathcal{F}_t = \mathcal{B}(x_s | 0 \leq s \leq t)$, U is a compact metrizable space, $b(\omega, t, u)$ a bounded measurable function on $\Omega \times R^+ \times U$, nonanticipating in (ω, t) versus \mathcal{F}_t and continuous in u , and β is a Brownian motion. We want to find a nonanticipating function u_0 defined on $\Omega \times R^+$ with values in U minimizing:

$$(1.2) \quad E \int_0^T L(\omega, t, u(\omega, t)) dt$$

where L satisfies the same assumptions as b .

To simplify the exposition we assume first that L does not depend on u . Then

$$\int_0^T L(\omega, t) dt$$

is a bounded \mathcal{F}_T -measurable function, which we call A_T .

If σ is sufficiently regular-continuous on $R^+ \times R^d$, bounded and invertible, we know by [10] that there is one and only one probability measure $Q_x^{b_u}$ on the space Ω such that on $(\Omega, Q_x^{b_u})$, (1.1) holds.

Moreover, on each \mathcal{F}_t , $Q_x^{b_u}$ has a density $Z_t^{b_u}$ relative to Q_x^0 given by:

$$(1.3) \quad \begin{aligned} Z_t^{b_u} &= \exp \left\{ \int_0^t \langle \sigma^{-1}(t, x_t) b(\omega, t, u(\omega, t)), d\beta_t \rangle \right. \\ &\quad \left. - \frac{1}{2} \int_0^t \langle b(\omega, t, u(\omega, t)), a^{-1}(t, x_t) b(\omega, t, u(\omega, t)) \rangle dt \right\} \end{aligned}$$

where β^0 is the Brownian motion generating x in (1.1) for $b=0$, and a is the function $\sigma\sigma^*$.

Then (1.2) may be written as:

$$(1.4) \quad E(A_T Z_T^{b_u}).$$

But Z^{b_u} is the unique solution of:

$$(1.5) \quad \begin{aligned} dZ &= Z \langle \sigma^{-1}(t, x_t) b(\omega, t, u(\omega, t)), d\beta \rangle, \\ Z(0) &= 1. \end{aligned}$$

* Received by the editors July 21, 1975, and in final revised form June 23, 1977.

† 191 rue d'Alésia, 75014 Paris, France.

Then (1.5) is an equation which has a unique solution in the sense of Ito. Moreover, in (1.1), u was a function of the past values of x , i.e. in closed-loop form relative to x . Now in (1.5), u is in stochastic open-loop relative to ω .

This fact is extremely important to the understanding of the power of the Girsanov transformation and to the use of the maximum principle.

We can apply the maximum principle given in [2, Thm. V-1] to equation (1.5) and criterion (1.4). Let \mathcal{H} be defined by:

$$(1.6) \quad \mathcal{H} = Z \langle H, \sigma^{-1} b \rangle.$$

The maximum principle can then be written:

$$(1.7) \quad \begin{aligned} dp &= -\frac{\partial \mathcal{H}}{\partial Z} dt + H \cdot d\beta + dM \\ p_T &= -A_T \\ \max_{u \in U} \mathcal{H} \end{aligned}$$

where p_0 is square integrable and \mathcal{F}_0 -measurable, where H is a predictable process such that

$$(1.8) \quad E \int_0^T |H_u|^2 du < +\infty$$

and where M is a square-integrable martingale such that $M\beta_1^0, \dots, M\beta_m^0$ are martingales. Moreover (1.7) may be written as:

$$(1.9) \quad \begin{aligned} dp &= -\langle H_b, \sigma^{-1}(t, x_t) b(\omega, t, u(\omega, t)) \rangle dt + H_t \cdot d\beta + dM \\ p_T &= -A_T \\ \max_{u \in U} \langle H_b, \sigma^{-1}(t, x_t) b(\omega, t, u) \rangle. \end{aligned}$$

Let us remember that, as indicated in [2], the process p_t is adapted to the σ -fields $\{\mathcal{F}_t\}_{t \geq 0}$, and that (1.9) is a functional equation, which must be solved by finding p_0, H, M . General backward stochastic equations with a terminal condition have been studied in [4].

In this paper we prove that it is feasible to apply the maximum principle because of the deep convex structure of the problem, and we prove the relation of (1.9) to Davis and Varaiya's necessary and sufficient conditions given in [6].

In § 2, the problem is rigorously defined. In § 3, the problem is put in the standard form of [2], and a dual problem is defined. In § 4, sufficient conditions are written for the solutions of the primal problem and of the dual problem. In § 5, using a result of Beneš, existence results are given for both problems. In § 6, a convexity assumption used in the previous sections is removed.

2. Definition of the problem. In what follows (Ω, \mathcal{F}, P) is a complete probability space endowed with an increasing family $\{\mathcal{F}_t\}_{t \geq 0}$ of complete sub- σ -fields of \mathcal{F} which is right-continuous, i.e., for any t , $\mathcal{F}_t = \bigcap_{t' > t} \mathcal{F}_{t'}$. We also assume that β is a m -dimensional Brownian motion defined on (Ω, \mathcal{F}, P) and adapted to \mathcal{F}_t .

Furthermore, \mathcal{T} is the σ -field of well measurable sets in $\Omega \times [0, +\infty[$ (see [7, VIII, D14]), \mathcal{T}^* is its completion relative to $dP \otimes dt$, and T is a positive constant.

The space L_{21} (resp. L_{22}) consists of the equivalence classes for the measure $dP \otimes dt$ of the \mathcal{T}^* -measurable processes x such that:

$$(2.1) \quad E\left(\int_0^T |x| dt\right)^2 < +\infty$$

(resp. $E(\int_0^T |x|^2) dt < +\infty$); L_{21} and L_{22} are endowed with the corresponding norms.

Also L_2^s is the space of square integrable \mathcal{F}_s -measurable random variables, and W^\perp is the space of square-integrable martingales orthogonal to β , null at time 0, and stopped at time T ; W^\perp may be identified as a subspace of L_2^T .

The set-valued function K is defined on $\Omega \times [0, T]$ with values in R^m , and has the following properties:

- (i) K has nonempty compact values,
- (ii) K is \mathcal{T}^* -measurable,
- (iii) K is uniformly bounded.

The set \mathcal{M} consists of the \mathcal{T}^* -measurable selections of K . Theorem 1 of [8] proves that \mathcal{M} is nonempty.

For m in \mathcal{M} , let Z be the solution of

$$(2.2) \quad \begin{aligned} dZ &= Zm \cdot d\beta, \\ Z(0) &= 1. \end{aligned}$$

By the Appendix of [4], (2.2) has a unique solution, and Z_t is a square-integrable martingale. It can be written as:

$$(2.3) \quad Z_t = \exp\left\{\int_0^t m_s \cdot d\beta_s - \frac{1}{2} \int_0^t |m_s|^2 ds\right\}.$$

Suppose $A_T \in L_2^T$.

DEFINITION 2.1. Problem (P) is the minimization of

$$(2.4) \quad E(A_T Z_T)$$

on \mathcal{M} .

3. General duality formulation. We assume in §§ 3, 4 and 5 that K has *convex values*. This assumption will be removed in § 6.

Let Γ be the set-valued function:

$$(3.1) \quad (\omega, t) \rightarrow \{(x, s, xk) \in R \times R \times R^m; x \geq 0, s = 0, k \in K(\omega, t)\}.$$

By Theorem 2 of [8], Γ is \mathcal{T}^* -measurable. It has, moreover, convex values, because K has convex values.

Let L be the normal convex integrand (in the sense of [8]) which is the indicator of Γ , i.e.,

$$(3.2) \quad L(\omega, t, x, y, H) = \begin{cases} 0 & \text{if } (x, y, H) \in \Gamma(\omega, t), \\ +\infty & \text{elsewhere.} \end{cases}$$

The functionals l_0 and l_T are defined respectively on L_2^0 and L_2^T by:

- a) l_0 is the indicator of 1, i.e. $l_0(Z)$ is 0 if $Z = 1$, $+\infty$ elsewhere,
- b) l_T is the linear function $Z_T \rightarrow E(A_T Z_T)$.

Now, let Z be written as a general process

$$(3.3) \quad Z_t = Z_0 + \int_0^t \dot{Z}_s ds + \int_0^t H_s \cdot d\beta_s$$

with

$$(Z_0, \dot{Z}, H) \in L_2^0 \times L_{21} \times L_{22}.$$

We then define $\Phi_{l,L}(Z)$ by:

$$(3.4) \quad \Phi_{l,L}(Z) = E \int_0^t L(\omega, t, Z, \dot{Z}, H) dt + l_0(Z_0) + l_T(Z_T).$$

Then $\Phi_{l,L}(Z) < +\infty$ if and only if $\dot{Z} = 0$ and $H = Zm$ with $m \in \mathcal{M}$. We then have:

PROPOSITION 3.1. *Problem (P) is equivalent to the minimization of $\Phi_{l,L}$.*

Problem (P) is then in the form given in [2, Definition II.2].

Let L^* be the dual function of L , i.e., L^* is defined by:

$$(3.5) \quad L^*(\omega, t, s, p, H') = \sup_{(x, y, H) \in R \times R \times R^m} \{ \langle s, x \rangle + \langle p, y \rangle + \langle H', H \rangle - L(\omega, t, x, y, H) \}.$$

We then have:

$$(3.6) \quad \begin{aligned} L^*(\omega, t, s, p, H') &= \sup_{\substack{x \geq 0 \\ k \in K(\omega, t)}} \{ \langle s, x \rangle + \langle H', xk \rangle \} \\ &= \sup_{x \geq 0} x \left(s + \sup_{k \in K(\omega, t)} \langle H', k \rangle \right). \end{aligned}$$

Let φ be the function defined by

$$(3.7) \quad \varphi(\omega, t, H') = \sup_{k \in K(\omega, t)} \langle H', k \rangle.$$

Then

$$(3.8) \quad L^*(\omega, t, s, p, H') = \begin{cases} 0 & \text{if } s + \varphi(\omega, t, H') \leq 0, \\ +\infty & \text{elsewhere.} \end{cases}$$

Similarly, we have:

$$(3.9) \quad \begin{aligned} l_0^*(p_0) &= E(p_0), \\ l_T^*(p_T) &= \begin{cases} 0 & \text{if } p_T = A_T, \\ +\infty & \text{elsewhere.} \end{cases} \end{aligned}$$

We now define the dual problem (P') associated with problem (P) by Definition II.2 of [2].

PROPOSITION 3.2. *The dual problem (P') consists of the minimization of $E(p_0)$ over all processes p which satisfy*

$$(3.10) \quad p = p_0 + \int_0^t \dot{p}_s ds + \int_0^t H'_s \cdot d\beta_s + M_t;$$

$$(3.11) \quad \begin{aligned} (p_0, \dot{p}, H', M) &\in L_2^0 \times L_{21} \times L_{22} \times W^\perp, \\ \dot{p} + \varphi(\omega, t, H') &\leq 0 \quad (dP \otimes dt \text{ a.e.}), \\ p_T &= -A_T. \end{aligned}$$

Proof. By [2, Definition II.2], the dual problem consists of the minimization of

$$E \int_0^T L^*(\omega, t, \dot{p}_t, p_t, H'_t) dt + l_0^*(p_0) + l_T^*(-p_T).$$

subject to (3.10).

By (3.8)–(3.9), we must then have (3.11) and the criterion to minimize is $E(p_0)$.

Remark. It must be noted that the process p is adapted to the family of σ -fields $\{\mathcal{F}_t\}_{t \geq 0}$, and is nonanticipating.

4. Coextremality conditions. We are now going to write sufficient conditions for Z to be a solution of problem (P) and for p to be a solution of problem (P').

THEOREM 4.1. *A sufficient condition for Z (associated with $m \in \mathcal{M}$ by (2.2)) to be a solution of problem (P) and for p (given by (3.10)) to be a solution of problem (P') is that:*

$$\begin{aligned} p_T &= -A_T, \\ (4.1) \quad \dot{p}_t &= -\varphi(\omega, t, H'_t) \quad (dP \otimes dt \text{ a.e.}), \\ \langle H'_t, m_t \rangle &= \varphi(\omega, t, H'_t) \quad (dP \otimes dt \text{ a.e.}). \end{aligned}$$

Moreover, if such a couple (Z, p) exists, then if Z' is any solution of problem (P), and p' any solution for problem (P'), they verify the corresponding relations.

Proof. We are to verify that (4.1) is nothing else than the coextremality relations given in [2, Definition IV.1], which are:

$$\begin{aligned} (\dot{p}_t, p_t, H'_t) &\in \partial L(\omega, t, Z_t, \dot{Z}_t, H_t) \quad (dP \otimes dt \text{ a.e.}), \\ (4.2) \quad p_0 &\in \partial l_0(Z_0), \\ p_T &\in -\partial l_T(Z_T). \end{aligned}$$

This is equivalent to

$$L(\omega, t, Z_t, \dot{Z}_t, H_t) dt + L^*(\omega, t, \dot{p}_t, p_t, H'_t) = \langle Z_t, \dot{p}_t \rangle + \langle \dot{Z}_t, p_t \rangle + \langle H'_t, H_t \rangle \quad (dP \otimes dt \text{ a.e.})$$

$$p_T = -A_T.$$

(Z, Zm) and (\dot{p}, H) must be such that:

$$(4.3) \quad Z\dot{p} + Z\langle H', m \rangle = 0.$$

But by (2.3), $Z > 0$. Then, (4.3) implies

$$(4.4) \quad \dot{p} = -\langle H', m \rangle.$$

The definition of φ , (4.4) and (3.11) imply (4.1). The remaining part of the theorem follows from Theorem IV.2 in [2].

5. Existence results. We give now existence results for problems (P) and (P').

THEOREM 5.1. *If K has convex values, problem (P) has a solution.*

Proof. This is a result of Beneš [1].

THEOREM 5.2. *If K has convex values, problem (P') has a unique solution p , and p is the unique adapted solution of*

$$\begin{aligned} (5.1) \quad dp &= -\varphi(\cdot, \cdot, H') dt + H' \cdot d\beta + dM, \\ p_T &= -A_T \end{aligned}$$

with $(p_0, H', M) \in L_2^0 \times L_{22} \times W^\perp$.

For Z to be a solution of problem (P), it is necessary and sufficient that

$$(5.2) \quad \langle H'_t, m_t \rangle = \varphi(\omega, t, H'_t) \quad (dP \otimes dt \text{ a.e.}),$$

and then

$$(5.3) \quad p_t = -\frac{E^{\mathcal{F}_t} A_T Z_T}{Z_t}.$$

Proof. Let Z be a solution of problem (P) and m the associated element in \mathcal{M} . By Theorem 2.2 of [4], equation

$$(5.4) \quad \begin{aligned} d\tilde{p} &= -\langle \tilde{H}, m \rangle dt + \tilde{H} \cdot d\beta + d\tilde{M}, \\ \tilde{p}_T &= -A_T \end{aligned}$$

constrained by the condition $(\tilde{p}_0, \tilde{H}, \tilde{M}) \in L_2^0 \times L_{22} \times W^\perp$, has a unique solution. By Proposition I.1 in [2], $\tilde{p}_t Z_t$ is a martingale. Then

$$\tilde{p}_t = -\frac{E^{\mathcal{F}_t} A_T Z_T}{Z_t}.$$

In particular,

$$(5.5) \quad E(\tilde{p}_0) + E(A_T Z_T) = 0.$$

If on a $dP \otimes dt$ nonnegligible set:

$$(5.6) \quad \langle \tilde{H}_t, m_t \rangle < \varphi(\omega, t, \tilde{H}_t),$$

one modifies m into m' on this set in order to have

$$(5.7) \quad \langle \tilde{H}_t, m'_t \rangle = \varphi(\omega, t, \tilde{H}_t).$$

Let Z' be the process corresponding to $m' \in \mathcal{M}$ by (2.2). Then, by Proposition I.1 in [2],

$$(5.8) \quad \tilde{p}_t Z'_t - \tilde{p}_0 Z_0 + \int_0^t Z'_u \langle \tilde{H}_u, m_t \rangle dt - \int_0^t Z'_u \langle \tilde{H}_u, m'_t \rangle dt$$

is a martingale null at the origin. This implies:

$$(5.9) \quad E(\tilde{p}_0) + E(A_T Z'_T) - E\left(Z'_T \int_0^T \langle \tilde{H}_t, m_t - m'_t \rangle dt\right) = 0.$$

But $Z'_T > 0$, so (5.5)–(5.9) imply

$$E(A_T Z'_T) < E(A_T Z_T).$$

Then Z would not be optimal. Hence

$$(5.10) \quad \langle \tilde{H}_t, m_t \rangle = \varphi(\omega, t, \tilde{H}_t) \quad (dP \otimes dt \text{ a.e.}).$$

If p is a solution of (5.1), it is possible to find Z'' such that Z'' and p are coextremal: it is sufficient to choose m'' in \mathcal{M} such that

$$\langle H'_t, m''_t \rangle = \varphi(\omega, t, H'_t) \quad (dP \otimes dt \text{ a.e.}),$$

and this is possible by the compactness of $K(\omega, t)$. But then Z and p are coextremal, and p is also a solution of (5.4). Hence $p = \tilde{p}$.

Remark. In this case where the dual state has dimension 1, equation (5.4) may be interpreted via the Girsanov transformation. If Z is the solution of (2.2), then under the probability measure $dP' = Z_T dP$, $\beta_t - \int_0^t m ds$ is a Brownian motion β^m .

Equation (5.4) may then be rewritten as

$$\begin{aligned} d\tilde{p} &= \tilde{H} \cdot d\beta^m + d\tilde{M}, \\ \tilde{p}_T &= -A_T. \end{aligned}$$

However A_T is generally only integrable relative to the new measure P' and not

square-integrable. In the case where we may write

$$(5.11) \quad E^{\mathcal{F}} A_T = \tilde{p}_0 + \int_0^t \tilde{H} \cdot d\beta^m + \tilde{M}_t,$$

where H is a predictable process such that $\int_0^T |\tilde{H}|^2 ds < +\infty$ a.s. and where \tilde{M} is a martingale for P' such that $M\beta_1^m, \dots, M\beta_n^m, \dots$ are local martingales, it is not obvious whether H is in L_{22} . We must then use Theorem 2.2 of [4] to give a precise meaning to (5.4). Only if A_T is in L_∞ can we easily pass from one approach to the other.

6. The general result. In this section we remove the convexity assumption on the values of K .

THEOREM 6.1. *Problem (P) has a solution, even if K is not convex-valued.*

Proof. Let $\hat{K}(\omega, t)$ be the closed convex hull of $K(\omega, t)$. By Corollary 3.3 of [8], \hat{K} is \mathcal{T}^* -measurable.

Let $\varphi(\omega, t, \cdot)$ be the support function of $K(\omega, t)$, which is also the support function of $\hat{K}(\omega, t)$. By Theorem 5.2, equation (5.1) has a unique solution. The sets $K(\omega, t)$ and $\hat{K}(\omega, t)$ have the same extremal points, so that by Theorem 2 of [8], m can be taken in \mathcal{M} such that

$$\langle H'_t, m_t \rangle = \varphi(\omega, t, H'_t) \quad (dP \otimes dt \text{ a.e.}).$$

By Theorem 5.2, problem (P) has then an optimal solution.

We reobtain partly the result of Davis [5], when L does not depend on u . In the case where L depends on u , we can use the method of Beneš [1], which gives then an optimal randomized strategy, by going back to the problem previously solved. To obtain a nonrandomized optimal control, we would have to use the method of Davis [5].

As noted in [3], convexity is generally an intermediary step in the problem of control of diffusions. Conditions (4.1) may be satisfied even when K does not have convex values.

7. Conclusion. Duality methods are not as strong as potential theory methods as used in [3]. Beneš' existence result gives us only "mixed" optimal control. Nevertheless, it is interesting to see how, in a loose sense, general problems of optimal stochastic control are problems of control of the martingale of densities.

REFERENCES

- [1] V. E. BENEŠ, *Existence of optimal stochastic control laws*, this Journal, 9 (1971), pp. 446–472.
- [2] J. M. BISMUT, *Conjugate convex functions in optimal stochastic control*, J. Math. Anal. Appl., 44 (1973), pp. 384–404.
- [3] ———, *Théorie probabiliste du contrôle des diffusions*, Mem. Amer. Math. Soc., 4 (Jan. 1976), no. 167, pp. 1–130.
- [4] ———, *Linear quadratic optimal stochastic control with random coefficients*, this Journal, 14 (1976), pp. 419–444.
- [5] M. H. A. DAVIS, *On the existence of optimal stochastic policies in stochastic control*, this Journal, 11 (1973), pp. 587–594.
- [6] M. H. A. DAVIS AND P. P. VARAIYA, *Dynamic programming conditions for partially observable stochastic systems*, this Journal, 11 (1973), pp. 226–261.
- [7] P. A. MEYER, *Probabilités et Potentiel*, Hermann, Paris; Blaisdell, Boston, 1966.
- [8] R. T. ROCKAFELLAR, *Measurable dependence of convex sets and functions on parameters*, J. Math. Anal. Appl., 28 (1969), pp. 4–25.
- [9] ———, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [10] D. W. STROOCK AND S. R. S. VARADHAN, *Diffusion process with continuous coefficients*, Comm. Pure Appl. Math., 22 (1969), pp. 345–400, 479–530.

FUNCTIONALS DEFINED ON FUNCTIONS OF BOUNDED VARIATION IN \mathbb{R}^n AND THE LEBESGUE AREA*

FRANCESCO FERRO†

Abstract. Let $BV_b(\mathbb{R}^n)$ be the space of all functions T such that $T \in L^1_{\text{loc}}(\mathbb{R}^n)$ and ∇T is a vector-valued measure of finite total variation on \mathbb{R}^n . In [1] we considered the functional

$$J_L(T) = \min \left\{ \liminf_{\alpha} \int_{\Omega} L(x, u_{\alpha}(x), \nabla u_{\alpha}(x)) \, dx : u_{\alpha} \xrightarrow{w^*} T \right\}, \quad T \in BV_b(\mathbb{R}^n),$$

where L is a proper normal integrand; $u_{\alpha} \in L^1_{\text{loc}}(\mathbb{R}^n)$, $\nabla u_{\alpha} \in (L^1(\mathbb{R}^n))^n$ and $u_{\alpha} \xrightarrow{w^*} T$ if and only if $\int_{\Omega} u_{\alpha} \rightarrow \int_{\Omega} T$ and $\int_{\mathbb{R}^n} G \nabla u_{\alpha} \rightarrow \int_{\mathbb{R}^n} G \nabla T$ for every continuous vector-valued function G with compact support in \mathbb{R}^n . In this paper we give some results about the w^* topology defined on $BV_b(\mathbb{R}^n)$ and derive new properties of the functional J_L . Afterward we carry on a comparison with the functional defining the Lebesgue area of a nonparametric surface.

1. Introduction. Variational problems concerning functions of bounded variation in one dimension have been studied in [2], [3] and [4]. In [2] results are obtained about the characterization of optimal arcs in terms of a so called “generalized Hamiltonian condition”, while in [3] and [4] sufficient conditions for the existence of an optimal arc are given. As to the n -dimensional case we recall that in [1], given an integral functional I_L defined on the Sobolev space $W^{1,1}(\Omega)$, we introduced a functional J_L on the space $BV_b(\mathbb{R}^n)$ (see § 2 for a precise definition) of the functions of bounded variation in \mathbb{R}^n , so that I_L and J_L have the same infimum; moreover we gave sufficient conditions for the existence of a minimum for J_L .

Section 2 of the present paper is devoted to summarize the definitions and main results given in [1].

In § 3 we study some properties of the suitable topology defined on $BV_b(\mathbb{R}^n)$, which lead to a deeper understanding of the structure of J_L .

In § 4 we consider the functional J_L corresponding to some special types of integrands L and prove that its values may be obtained by a limit of values of I_L on certain integral averages.

Finally in § 5 we carry on a comparison between J_L and the functional defining the Lebesgue area ([5] and [6]) for nonparametric surfaces and improve an existence theorem contained in [6].

2. Functional spaces, integrands and variational functionals. Throughout this paper Ω is an open, bounded and connected subset of the Euclidean space \mathbb{R}^n whose boundary $\partial\Omega$ verifies the local Lipschitz condition (in the sense of [7]).

In what follows, we often use the nonreflexive Banach spaces $L^1(\Omega)$, $L^1(\mathbb{R}^n)$, $W^{1,1}(\Omega)$, $W^{1,1}(\mathbb{R}^n)$; their definition and main properties are well-known [8], [9], [7]. $C_0(\mathbb{R}^n)$ is the space of all continuous functions which have compact support in \mathbb{R}^n ; if $f \in C_0(\mathbb{R}^n)$ we set

$$\|f\|_{C_0(\mathbb{R}^n)} = \max \{|f(x)| : x \in \mathbb{R}^n\}.$$

So $C_0(\mathbb{R}^n)$ is a normed space; let $\bar{C}_0(\mathbb{R}^n)$ be its completion. The dual space of $C_0(\mathbb{R}^n)$ (and of $\bar{C}_0(\mathbb{R}^n)$) will be denoted $M_b(\mathbb{R}^n)$ and consists of all measures which have a

* Received by the editors July 20, 1977.

† Istituto di Matematica, Università di Genova, Italy. This work was supported in part by Laboratorio per la Matematica Applicata del C.N.R., Italy.

finite total variation on \mathbb{R}^n [10]. If $v = (v_1, \dots, v_n) \in (M_b(\mathbb{R}^n))^n$ we define

$$\|v\|_{(M_b(\mathbb{R}^n))^n} = \sum_{i=1}^n \|v_i\|_{M_b(\mathbb{R}^n)},$$

where

$$\|v_i\|_{M_b(\mathbb{R}^n)} = \sup \left\{ \int_{\mathbb{R}^n} f(x) v_i(dx) : f \in C_0(\mathbb{R}^n), |f(x)| \leq 1 \right\}.$$

Now we consider the space

$$BV_b(\mathbb{R}^n) = \{T : T \in L^1_{\text{loc}}(\mathbb{R}^n), \nabla T \in (M_b(\mathbb{R}^n))^n\},$$

where $L^1_{\text{loc}}(\mathbb{R}^n)$ is the space of all functions which are locally summable in \mathbb{R}^n . $BV_b(\mathbb{R}^n)$ is a subspace of the space $BV(\mathbb{R}^n)$ of the functions of bounded variation in the sense of Cesari (see [11]). If $T \in BV_b(\mathbb{R}^n)$ we set

$$(1) \quad \|T\|_{BV_b(\mathbb{R}^n)} = \left| \int_{\Omega} T \right| + \|\nabla T\|_{(M_b(\mathbb{R}^n))^n};$$

We proved in [1] that $BV_b(\mathbb{R}^n)$, endowed with the norm (1), is a Banach space.

Let us consider the map

$$i : BV_b(\mathbb{R}^n) \rightarrow \mathbb{R} \oplus (M_b(\mathbb{R}^n))^n$$

such that

$$i(T) = \left(\int_{\Omega} T, \nabla T \right).$$

Obviously i establishes an isometric isomorphism between $BV_b(\mathbb{R}^n)$ and a subspace of $\mathbb{R} \oplus (M_b(\mathbb{R}^n))^n$. Since $\mathbb{R} \oplus (M_b(\mathbb{R}^n))^n$ is the dual space of $\mathbb{R} \oplus (\bar{C}_0(\mathbb{R}^n))^n$, it may be endowed with the weak topology of the dual space, that is with the so-called w^* topology. We may identify $BV_b(\mathbb{R}^n)$ with its image in the following sense: if $\{T_\alpha\}$ is a net (i.e. a generalized sequence) in $BV_b(\mathbb{R}^n)$ we shall say that $\{T_\alpha\}$ w^* -converges to $T \in BV_b(\mathbb{R}^n)$ (and we shall write $T_\alpha \xrightarrow{w^*} T$) if

$$\lim_{\alpha} \int_{\Omega} T_\alpha = \int_{\Omega} T$$

and

$$\lim_{\alpha} \int_{\mathbb{R}^n} f \nabla T_\alpha = \int_{\mathbb{R}^n} f \nabla T, \quad \text{for every } f \in (C_0(\mathbb{R}^n))^n;$$

this convergence defines a topology in $BV_b(\mathbb{R}^n)$ which will be called the w^* topology of $BV_b(\mathbb{R}^n)$. In [1] it is proved that the image of i is a w^* -closed subspace of $\mathbb{R} \oplus (M_b(\mathbb{R}^n))^n$; therefore, recalling the well-known Alaoglu theorem, we may assert that all closed balls of $BV_b(\mathbb{R}^n)$ are w^* -compact. Also, since $\mathbb{R} \oplus (\bar{C}_0(\mathbb{R}^n))^n$ is separable, the w^* topology of $BV_b(\mathbb{R}^n)$ is metrizable on all closed balls [12]. As in [1] we put

$$W^{1,1}_{s\text{-loc}}(\mathbb{R}^n) = \{u : u \in L^1_{\text{loc}}(\mathbb{R}^n), \nabla u \in L^1(\mathbb{R}^n)\}.$$

$W^{1,1}_{s\text{-loc}}(\mathbb{R}^n)$ is strongly closed in $BV_b(\mathbb{R}^n)$, but it is w^* -dense in $BV_b(\mathbb{R}^n)$ [1, Theorem 1.10].

In what follows let

$$L: \Omega \times \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$$

be a proper normal integrand [13]; that is:

- (i) $L(x, \cdot, \cdot)$ is lower semicontinuous for every $x \in \Omega$,
- (ii) $L(x, \cdot, \cdot) \not\equiv +\infty$ for every $x \in \Omega$,
- (iii) $E_L(x) = \{(u, v, \alpha): L(x, u, v) \leq \alpha\}$ is a measurable multifunction.

Condition (iii) means that

$$E_L^{-1}(C) = \{x: E_L(x) \cap C \neq \emptyset\}$$

is Lebesgue measurable for every $C \subset \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}$, C closed. It is known (see e.g. [13]) that if L is a proper normal integrand and if u and v are measurable then $L(x, u(x), v(x))$ is measurable. Let us define

$$\tilde{L}(x, u, v) = \begin{cases} L(x, u, v), & x \in \Omega, \\ 0, & x \in \mathbb{R}^n - \Omega. \end{cases}$$

\tilde{L} is a proper normal integrand [1].

Given the integrand L we define the functional

$$(2) \quad I_L(u) = \int_{\Omega} L(x, u(x), \nabla u(x)) dx, \quad u \in W^{1,1}(\Omega).$$

$I_L(u)$ is well-defined by (2) if $L(x, u(x), \nabla u(x))$ is a summable function; otherwise we put $I_L(u) = -\infty$ if there is a summable function greater than $L(x, u(x), \nabla u(x))$ and $I_L(u) = +\infty$ in every other case. Throughout this paper we always suppose that there exists $u \in W^{1,1}(\Omega)$ such that $I_L(u) \in \mathbb{R}$. We may define also

$$I_{\mathcal{L}}(u) = I_L(r(u)), \quad u \in W_{s\text{-loc}}^{1,1}(\mathbb{R}^n),$$

$r(u)$ being the restriction of u to Ω . Finally, if $T \in \text{BV}_b(\mathbb{R}^n)$ we put

$$S(T) = \{\{u_{\alpha}\}: \{u_{\alpha}\} \subset W_{s\text{-loc}}^{1,1}(\mathbb{R}^n), \{u_{\alpha}\} \text{ is a net, } u_{\alpha} \xrightarrow{w^*} T\}$$

and define

$$(3) \quad J_{\mathcal{L}}(T) = \min \left\{ \liminf_{\alpha} I_{\mathcal{L}}(u_{\alpha}): \{u_{\alpha}\} \in S(T) \right\}, \quad T \in \text{BV}_b(\mathbb{R}^n).$$

The functional $J_{\mathcal{L}}$ has been defined and studied in [1]. However we think it is useful to recall what follows:

$J_{\mathcal{L}}$ is w^* -lower semicontinuous;

$$J_{\mathcal{L}}(u) \leq I_{\mathcal{L}}(u), \quad u \in W_{s\text{-loc}}^{1,1}(\mathbb{R}^n);$$

$$\inf \{J_{\mathcal{L}}(T): T \in \text{BV}_b(\mathbb{R}^n)\} = \inf \{I_{\mathcal{L}}(u): u \in W_{s\text{-loc}}^{1,1}(\mathbb{R}^n)\};$$

$$(4) \quad J_{\mathcal{L}}(T_1) = J_{\mathcal{L}}(T_2) \text{ if } T_1 - T_2 \in E,$$

where E is the w^* -closure of all functions in $\text{BV}_b(\mathbb{R}^n)$ vanishing in a neighborhood of $\bar{\Omega}$. Moreover in [1] we proved that if the level sets of I_L (i.e. the sets of the form $\{u: u \in W^{1,1}(\Omega), I_L(u) \leq \alpha\}$) are bounded then $J_{\mathcal{L}}$ attains a minimum.

The space E in (4) will be characterized in the next section.

3. w^* -convergence and the functional $J_{\mathcal{L}}$. It will be useful to have the following lemma.

LEMMA 3.1. *Let G be an open set in \mathbb{R}^n and v a function in $L^1(G)$; then*

$$\|v\|_{L^1(G)} = \sup \left\{ \int_G fv : f \in C_0(G), |f(x)| \leq 1, x \in G \right\},$$

where $C_0(G)$ is the space of all continuous functions in G vanishing in a neighborhood of ∂G .

Proof. We have

$$\begin{aligned} \|v\|_{L^1(G)} &= \sup \left\{ \int_G fv : f \in L^\infty(G), \|f\|_{L^\infty(G)} \leq 1 \right\} \\ &\geq \sup \left\{ \int_G fv : f \in C_0(G), |f(x)| \leq 1, x \in G \right\}. \end{aligned}$$

Now we consider a sequence $\{f_m\} \subset C_0^\infty(G)$ (=space of all infinitely differentiable functions with compact support in G) such that

$$\lim_{m \rightarrow +\infty} f_m(x) = (\operatorname{sgn} v)(x) \quad \text{a.e. in } G,$$

and

$$|f_m(x)| \leq 1, \quad x \in G;$$

we have

$$\|v\|_{L^1(G)} = \lim_{m \rightarrow +\infty} \int_G f_m v \leq \sup \left\{ \int_G fv : f \in C_0(G), |f(x)| \leq 1, x \in G \right\}. \quad \square$$

The following proposition emphasizes the relation between w^* -convergence in $BV_b(\mathbb{R}^n)$ and strong convergence in $L^1(\Omega)$.

PROPOSITION 3.1. *Let $\{T_m\}$ be a sequence in $BV_b(\mathbb{R}^n)$; if $T_m \xrightarrow{w^*} T \in BV_b(\mathbb{R}^n)$ then $\lim_{m \rightarrow +\infty} r(T_m) = r(T)$ in $L^1(\Omega)$.*

Proof. From [1, Theorem 1.10] it is enough to consider the case when $\{T_m\} \subset W_{s\text{-loc}}^{1,1}(\mathbb{R}^n)$. By the uniform boundedness theorem, there exists a constant c_1 such that

$$(5) \quad \|T_m\|_{BV_b(\mathbb{R}^n)} \leq c_1, \quad \text{for every } m.$$

By Lemma 3.1 and (5) we obtain

$$\sum_{i=1}^n \|(T_m)_{x_i}\|_{L^1(\mathbb{R}^n)} = \|\nabla T_m\|_{(\mathcal{M}_b(\mathbb{R}^n))^n} \leq \|T_m\|_{BV_b(\mathbb{R}^n)} \leq c_1$$

and, by Poincaré's inequality [7], there exists a constant c_2 such that

$$\|T_m\|_{L^1(\Omega)} \leq c_2.$$

Hence there exist a subsequence of $\{T_m\}$ (we shall write $\{T_m\}$ again) and $T_0 \in L^1(\Omega)$ such that

$$\lim_{m \rightarrow +\infty} r(T_m) = T_0 \quad \text{in } L^1(\Omega);$$

hence

$$\lim_{m \rightarrow +\infty} \nabla(r(T_m)) = \nabla T_0 \quad \text{in } \mathcal{D}'(\Omega) \quad (= \text{distributions in } \Omega).$$

But we have also

$$\lim_{m \rightarrow +\infty} \nabla(r(T_m)) = \nabla(r(T)) \quad \text{in } \mathcal{D}'(\Omega);$$

so we obtain $T_0 = r(T)$ a.e. in Ω . Now it is easy to deduce that the whole sequence $\{r(T_m)\}$ converges to $r(T)$ in $L^1(\Omega)$. \square

It is well-known (see [9]) that there exists a map

$$p: L^1(\Omega) \rightarrow L^1(\mathbb{R}^n)$$

which is linear, continuous and such that $p(u) \in W^{1,1}(\mathbb{R}^n)$ whenever $u \in W^{1,1}(\Omega)$; moreover in this case we have

$$(6) \quad \|p(u)\|_{W^{1,1}(\mathbb{R}^n)} \leq c \|u\|_{W^{1,1}(\Omega)}$$

for a suitable constant c independent of u .

PROPOSITION 3.2. *Let $\{T_m\}$ be a sequence in $BV_b(\mathbb{R}^n)$; if $T_m \xrightarrow{w^*} T \in BV_b(\mathbb{R}^n)$ then $p(r(T_m)) \xrightarrow{w^*} p(r(T))$.*

Proof. As in the proof of Proposition 3.1 we may assume $\{T_m\} \subset W_{s\text{-loc}}^{1,1}(\mathbb{R}^n)$. We put $T'_m = p(r(T_m))$ and $T' = p(r(T))$; by Proposition 3.1 we obtain $\lim_{m \rightarrow +\infty} r(T_m) = r(T)$ in $L^1(\Omega)$; hence $\lim_{m \rightarrow +\infty} T'_m = T'$ in $L^1(\mathbb{R}^n)$ and

$$(7) \quad \lim_{m \rightarrow +\infty} T'_m = T' \quad \text{in } \mathcal{D}'(\mathbb{R}^n).$$

By (1), (6) and Lemma 3.1 there exist two constants $c_1, c_2 \in \mathbb{R}$ such that

$$\|T'_m\|_{BV_b(\mathbb{R}^n)} \leq c_1 \|T_m\|_{BV_b(\mathbb{R}^n)} \leq c_1 c_2.$$

Then there exists a w^* -convergent subsequence of $\{T'_m\}$; by (7) its limit will be T' . As in the proof of Proposition 3.1 it is easy to deduce that the whole sequence $\{T'_m\}$ w^* -converges to T' . \square

LEMMA 3.2. *Let T be a function in $BV_b(\mathbb{R}^n)$; then*

$$(8) \quad T - p(r(T)) \in E$$

and

$$(9) \quad J_L(T) = J_L(p(r(T))).$$

Proof. Let $\{u_m\} \subset W_{s\text{-loc}}^{1,1}(\mathbb{R}^n)$ and $u_m \xrightarrow{w^*} T$. By Proposition 3.2 we have

$$u_m - p(r(u_m)) \xrightarrow{w^*} T - p(r(T));$$

now (8) and (9) follow by [1, Proposition 2.6]. \square

Now we can prove the following theorem which gives a simple characterization of the space E (see (4) and what follows) and an interesting property of J_L .

THEOREM 3.1. *Let T_1 and T_2 be functions in $BV_b(\mathbb{R}^n)$. Then*

$$(10) \quad T_1 - T_2 \in E \text{ if and only if } T_1 - T_2 = 0 \text{ a.e. in } \Omega,$$

$$(11) \quad T_2 - T_2 = 0 \text{ a.e. in } \Omega \text{ implies } J_L(T_1) = J_L(T_2).$$

Proof. If $T_1 - T_2 \in E$, then $T_1 - T_2 = 0$ a.e. in Ω by the definition of E and

Proposition 3.1. Conversely if $T_1 - T_2 = 0$ a.e. in Ω we may write

$$\begin{aligned} T_1 - T_2 &= T_1 - p(r(T_1)) + p(r(T_1)) - p(r(T_2)) + p(r(T_2)) - T_2 \\ &= T_1 - p(r(T_1)) + p(r(T_2)) - T_2. \end{aligned}$$

Hence we obtain $T_1 - T_2 \in E$ by (8) and $J_L(T_1) = J_L(T_2)$ by (4). So (10) and (11) are completely proved. \square

Let us define

$$\begin{aligned} A_z(I_L) &= \{u: u \in W^{1,1}(\Omega), I_L(u) \leq z\}, \\ A'_z(I_L) &= \left\{u: u \in W^{1,1}(\Omega), I_L(u) \leq z, \int_{\Omega} u = 0\right\}, \\ \bar{S}(T) &= \left\{\{u_m\}: \{u_m\} \subset W^{1,1}_{s\text{-loc}}(\mathbb{R}^n), \{u_m\} \text{ is a sequence, } u_m \xrightarrow{w^*} T\right\}, \end{aligned}$$

where $T \in \text{BV}_b(\mathbb{R}^n)$.

The following theorem states that in certain cases the functional J_L does not change if in (3) we write $\bar{S}(T)$ instead of $S(T)$ (i.e. if we consider only sequences instead of nets).

THEOREM 3.2. *If for every $z \in \mathbb{R}$ the level set $A_z(I_L)$ is bounded (or empty), then*

$$(12) \quad J_L(T) = \min \left\{ \liminf_{m \rightarrow +\infty} I_L(u_m): \{u_m\} \subset \bar{S}(T) \right\}, \quad T \in \text{BV}_b(\mathbb{R}^n).$$

The same is true if $L = L(x, v)$ and $A'_z(I_L)$ is bounded (or empty) for every $z \in \mathbb{R}$.

Proof. Let us denote $\bar{J}_L(T)$ the right hand side of (12); since $\bar{S}(T) \subset S(T)$ we have $J_L(T) \leq \bar{J}_L(T)$. It follows that if $J_L(T) = +\infty$ (12) holds. Now we consider $T \in \text{BV}_b(\mathbb{R}^n)$ such that $J_L(T) < +\infty$. Let $\{u_\alpha\}$ be a net in $S(T)$ such that [1]

$$J_L(T) = \lim I_L(u_\alpha).$$

If $L = \dot{L}(x, v)$ we may assume $\int_{\Omega} u_\alpha = 0$.

In every case, if z is a constant such that $z > J_L(T)$ there exists a subnet of $\{u_\alpha\}$, which we denote $\{u_\beta\}$, such that $I_L(u_\beta) \leq z$. Hence

$$I_L(r(u_\beta)) \leq z$$

and by hypothesis there exists a constant $M > 0$ such that

$$\|r(u_\beta)\|_{W^{1,1}(\Omega)} \leq M.$$

Moreover we have

$$\|p(r(u_\beta))\|_{\text{BV}_b(\mathbb{R}^n)} \leq \|p(r(u_\beta))\|_{W^{1,1}(\mathbb{R}^n)} \leq c \|r(u_\beta)\|_{W^{1,1}(\Omega)} \leq cM,$$

for a suitable constant c . We have also

$$J_L(T) = \lim_{\beta} I_L(p(r(u_\beta))).$$

Therefore the value $J_L(T)$ does not change if in (3) we consider only nets contained in a suitable (depending on T) closed ball in $\text{BV}_b(\mathbb{R}^n)$; but, as we observed in § 2, the w^* topology is metrizable on all closed balls; then (12) holds. \square

COROLLARY 3.1. *If there exists $\theta \in L^1(\Omega)$ and $K > 0$ such that*

$$L(x, u, v) \geq K(|u| + |v|) - \theta(x), \quad \text{a.e. in } \Omega,$$

then $A_z(I_L)$ is bounded (or empty) for every $z \in \mathbb{R}$ and (12) holds. Moreover if $L = L(x, v)$ and there exist $\theta \in L(\Omega)$ and $K > 0$ such that

$$L(x, v) \geq K|v| - \theta(x), \quad \text{a.e. in } \Omega,$$

then $A'_z(I_L)$ is bounded (or empty) for every $z \in \mathbb{R}$ and (12) holds.

Proof. The proof is trivial; however see [1]. \square

4. Approximation by integral averages. In this section we prove that, if some conditions of regularity are fulfilled by L then $J_L(T) = \lim_{h \rightarrow 0^+} I_L(r(T_h))$ where T_h are the integral averages of T [7], [9], [1], [14].

Let λ and μ be nonnegative continuous functions defined on $[0, +\infty)$ such that

$$\lambda(0) = \mu(0) = 0;$$

moreover we suppose that there exists a constant c such that $\mu(t) \leq ct$ for large t .

LEMMA 4.1. Let $L: \Omega \times \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}$ satisfy

$$L \in C(\Omega \times \mathbb{R} \times \mathbb{R}^n),$$

$$L(x, u, \cdot) \text{ is convex for every } (x, u) \in \Omega \times \mathbb{R},$$

$$(13) \quad L(x, u, v) \geq -\phi(x), \text{ for every } (x, u, v) \in \Omega \times \mathbb{R} \times \mathbb{R}^n,$$

where $\phi(x) \geq 0$ and $|\phi(x) - \phi(x_1)| \leq \lambda(|x - x_1|)$ if $x, x_1 \in \Omega$,

$$|L(x, u, v) - L(x_1, u_1, v)| \leq \lambda(|x - x_1|)[1 + L(x, u, v)] + \mu(|u - u_1|).$$

If $\{u_m\}$ is a sequence in $W^{1,1}(\Omega)$, $T \in L^1(\Omega)$ and

$$\lim_{m \rightarrow +\infty} u_m = T \quad \text{in } L^1(\Omega),$$

then

$$(14) \quad \limsup_{h \rightarrow 0^+} \int_{\Omega_{(h)}} L(x, T_h(x), \nabla T_h(x)) \, dx \leq \liminf_{m \rightarrow +\infty} I_L(u_m),$$

where $\Omega_{(h)} = \{x: x \in \Omega, d(x, \partial\Omega) > h\}$.

Proof. We set $f(x, u, v) = L(x, u, v) + \phi(x)$. Then f satisfies the hypothesis of [14, Lemma 2] (see also [14, Lemma 1]) and (14) is proved. \square

Remark 4.1. The hypothesis of Lemma 4.1 assures that L is a proper normal integrand. Moreover if $L: \mathbb{R}^n \rightarrow \mathbb{R}$ (i.e. $L = L(v)$) is convex and nonnegative then (13) holds; in this case we have [14, Lemma 1]

$$\int_{\Omega_{(h)}} L(\nabla T_h(x)) \, dx \leq \liminf_{m \rightarrow +\infty} I_L(u_m). \quad \square$$

In what follows, if $T \in \text{BV}_b(\mathbb{R}^n)$ we denote $|\nabla T|$ the total variation measure of the vector-valued measure ∇T .

THEOREM 4.1. Let L satisfy the hypothesis of Lemma 4.1; let us suppose also that (12) holds and that there exist a constant A and $g \in L^1(\Omega)$ such that

$$(15) \quad L(x, u, v) \leq A(g(x) + |u| + |v|).$$

If $T \in \text{BV}_b(\mathbb{R}^n)$ and

$$(16) \quad |\nabla T|(\partial\Omega) = 0,$$

then

$$(17) \quad J_{\mathcal{L}}(T) = \lim_{h \rightarrow 0^+} I_L(r(T_h)).$$

Proof. We put

$$\tau(h) = I_L(r(T_h)) - \int_{\Omega(h)} L(x, T_h(x), \nabla T_h(x)) \, dx.$$

Recalling that $T_h \xrightarrow{w^*} T$ and, using Proposition 3.1, $\lim_{h \rightarrow 0^+} r(T_h) = r(T)$ in $L^1(\Omega)$, we obtain from Lemma 4.1

$$\begin{aligned} & \liminf_{h \rightarrow 0^+} I_L(r(T_h)) \\ & \leq \limsup_{h \rightarrow 0^+} I_L(r(T_h)) \\ & \leq \limsup_{h \rightarrow 0^+} \tau(h) + \limsup_{h \rightarrow 0^+} \int_{\Omega(h)} L(x, T_h(x), \nabla T_h(x)) \, dx \\ & \leq \limsup_{h \rightarrow 0^+} \tau(h) + J_{\mathcal{L}}(T) \\ & \leq \limsup_{h \rightarrow 0^+} \tau(h) + \liminf_{h \rightarrow 0^+} I_L(r(T_h)). \end{aligned}$$

If we prove $\lim_{h \rightarrow 0^+} \tau(h) = 0$ then (17) holds. By (13) and (15) we have

$$(18) \quad - \int_{\Omega - \Omega(h)} \phi(x) \, dx \leq \tau(h) \leq A \left(\int_{\Omega - \Omega(h)} g(x) \, dx + \int_{\Omega - \Omega(h)} |T_h(x)| \, dx + \int_{\Omega - \Omega(h)} |\nabla T_h(x)| \, dx \right).$$

The limit of the left hand side of (18) is zero; then we use the properties of integral averages and obtain

$$(19) \quad \tau(h) \leq A \left(\int_{\Omega - \Omega(h)} g(x) \, dx + \int_{\Omega^{(h)} - \Omega(2h)} |T(x)| \, dx + \int_{\Omega^{(h)} - \Omega(2h)} |\nabla T|(dx) \right),$$

where $\Omega^{(h)} = \{x: d(x, \bar{\Omega}) < h\}$. Since $|\nabla T|$ is a regular measure and by (16) the left hand side of (19) tends to zero. \square

We recall that the elements of $BV_b(\mathbb{R}^n)$ are class of equivalent functions. In what follows we shall say that $T \in BV_b(\mathbb{R}^n)$ is continuous at a point (or on a set) if in the equivalence class of T there is a function which is continuous at such a point (or at each point of such a set).

THEOREM 4.2. *Let L satisfy the hypothesis of Theorem 4.1. If $T \in BV_b(\mathbb{R}^n)$ and T is continuous on $\partial\Omega$ then (16) and (17) hold; the same is true if $T \in W^{1,1}(G)$, G being an open neighborhood of $\bar{\Omega}$.*

Proof. In the former case we have $\int_{\partial\Omega} |\nabla T|(dx) = 0$ (see [15, Thm. 2.6] and [6, § III]); hence (16) and (17) hold. In the latter case we know that ∇T is an absolutely continuous measure in G and the theorem is proved. \square

Let $T \in BV_b(\mathbb{R}^n)$; in what follows we shall say that $r(T)$ is continuous on $\partial\Omega$ if in the equivalence class of T there is a function whose restriction to $\bar{\Omega}$ is continuous on

$\partial\Omega$. As in § 3 we put $T' = p(r(T))$ and let T'_h be the integral averages of T' .

THEOREM 4.3'. *Let L satisfy the hypothesis of Theorem 4.1. If $T \in \text{BV}_b(\mathbb{R}^n)$ and either $r(T)$ is continuous on $\partial\Omega$ or $r(T) \in W^{1,1}(\Omega)$ then*

$$J_L(T) = \lim_{h \rightarrow 0^+} I_L(r(T'_h)).$$

Proof. We observe that either T' is continuous on $\partial\Omega$ or $T' \in W^{1,1}(\mathbb{R}^n)$; then by (9) and Theorem 4.2 we have

$$J_L(T) = J_L(T') = \lim_{h \rightarrow 0^+} I_L(r(T'_h)). \quad \square$$

5. A comparison with the Lebesgue area. Let J be the functional defining the Lebesgue area of a surface (see [5], [6]). We recall that

$$(20) \quad J(T) = \min \left\{ \liminf_{m \rightarrow +\infty} \int_{\Omega} (1 + |\nabla u_m|^2)^{1/2} \right\},$$

where the minimum is taken over all sequences $\{u_m\}$ of piecewise linear functions which converge to T uniformly in Ω . $J(T)$ is defined and finite if and only if $T \in \text{BV}_b(\Omega) \cap C(\bar{\Omega})$, $\text{BV}_b(\Omega)$ being the space of all functions T whose derivatives are measures such that $|\nabla T|(\Omega) < +\infty$. If $T \in \text{BV}_b(\mathbb{R}^n)$ we have $r(T) \in \text{BV}_b(\Omega)$; if $T \in \text{BV}_b(\Omega) \cap C(\bar{\Omega})$ there exists $T_1 \in \text{BV}_b(\mathbb{R}^n)$ such that $r(T_1) = T$ [6, § III].

THEOREM 5.1. *If $L(v) = (1 + |v|^2)^{1/2}$, $T \in \text{BV}_b(\mathbb{R}^n)$ and $r(T) \in C(\bar{\Omega})$, then*

$$J_L(T) = J(r(T)).$$

Proof. We observe that L satisfies the hypothesis of Lemma 4.1 and Theorem 4.3. Let $\{u_m\}$ be a sequence of piecewise linear functions in Ω ; if $T \in \text{BV}_b(\mathbb{R}^n)$ verifies $r(T) \in C(\bar{\Omega})$ and $\lim_{m \rightarrow +\infty} u_m = r(T)$ uniformly in Ω we have $\lim_{m \rightarrow +\infty} u_m = r(T)$ in $L^1(\Omega)$ and, by Theorem 4.3 and the proof of Theorem 4.1, we obtain

$$J_L(T) = \lim_{h \rightarrow 0^+} I_L(r(T'_h)) \leq \liminf_{m \rightarrow +\infty} I_L(u_m);$$

hence

$$J_L(T) \leq J(r(T)).$$

Since J is lower semicontinuous with respect to uniform convergence and $\lim_{h \rightarrow 0^+} r(T'_h) = r(T)$ uniformly in Ω , we have

$$J(r(T)) \leq \lim_{h \rightarrow 0^+} I_L(r(T'_h)) = J_L(T)$$

and the theorem is proved. \square

In [6] the following definition is given.

DEFINITION 5.1. Ω is *uniformly convex* if there exist a positive constant k and for every $x \in \partial\Omega$ a hyperplane π_x in \mathbb{R}^n such that $x \in \pi_x$, $\pi_x \cap \Omega$ is empty and

$$\sup \{|x - y|^2 / d(y, \pi_x) : y \in \Omega\} \leq k. \quad \square$$

The following theorem is proved in [6].

THEOREM 5.2. *Let Ω be uniformly convex and $\phi \in C(\partial\Omega)$. Then there exists $T_0 \in \text{BV}_b(\Omega) \cap C(\bar{\Omega})$ such that*

$$T_0 = \phi \quad \text{on } \partial\Omega, \quad J(T_0) \leq J(T),$$

for every $T \in \text{BV}_b(\Omega) \cap C(\bar{\Omega})$ such that $T = \phi$ on $\partial\Omega$. \square

Now we shall prove a simple generalization of Theorem 5.2. In what follows we denote $\gamma(T)$ the value of $r(T)$ on $\partial\Omega$ if $r(T)$ is continuous on $\partial\Omega$ and the trace of $r(T)$ on $\partial\Omega$ if $r(T) \in W^{1,1}(\Omega)$. It will always be $L = (1 + |v|^2)^{1/2}$.

THEOREM 5.3. *Let Ω be uniformly convex and $\phi \in C(\partial\Omega)$. Then there exists $T_0 \in \text{BV}_b(\mathbb{R}^n)$ such that $r(T_0) \in C(\bar{\Omega})$, $\gamma(T_0) = \phi$ and*

$$J_L(T_0) \leq J_L(T),$$

where $T \in \text{BV}_b(\mathbb{R}^n)$, $\gamma(T) = \phi$ and either $r(T)$ is continuous on $\partial\Omega$ or $r(T) \in W^{1,1}(\Omega)$.

Proof. By [6, Thm. 7.1], if $r(T)$ is continuous on $\partial\Omega$, for $h > 0$ there exists $T_{h,0} \in \text{BV}_b(\mathbb{R}^n) \cap C(\mathbb{R}^n)$ such that $J(r(T_{h,0}))$ is the minimum of J in the class of the Lipschitz functions whose trace on $\partial\Omega$ is $\gamma(r(T'_h))$; moreover there exists \bar{T}_0 in $\text{BV}_b(\Omega) \cap C(\bar{\Omega})$ such that $\lim_{h \rightarrow 0^+} r(T_{h,0}) = \bar{T}_0$ uniformly in $\bar{\Omega}$. Hence we have by Theorem 4.3

$$J_L(T) = \lim_{h \rightarrow 0^+} J_L(T'_h) = \lim_{h \rightarrow 0^+} J(r(T'_h)) \geq \lim_{h \rightarrow 0^+} J(r(T_{h,0})) \geq J(\bar{T}_0) = J_L(T_0),$$

where $T_0 = p(\bar{T}_0)$. If $r(T) \in W^{1,1}(\Omega)$, by [8] there exists $\bar{u} \in W^{1,1}(\Omega) \cap C(\bar{\Omega})$ such that $\gamma(\bar{u}) = \phi$; we have $\gamma(r(T) - \bar{u}) = 0$. Then there exists a sequence $\{v_m\} \subset C_0^1(\Omega)$ such that

$$\lim_{m \rightarrow +\infty} v_m = r(T) - \bar{u} \quad \text{in } W^{1,1}(\Omega).$$

Therefore we have

$$\begin{aligned} \lim_{m \rightarrow +\infty} (v_m + \bar{u}) &= r(T) \quad \text{in } W^{1,1}(\Omega), & \gamma(v_m + \bar{u}) &= \phi, \\ p(v_m + \bar{u}) &\in \text{BV}_b(\mathbb{R}^n), & v_m + \bar{u} &\in C(\bar{\Omega}). \end{aligned}$$

Hence [6, Prop. 8.1]

$$J_L(T_0) = J(r(T_0)) \leq \lim_{m \rightarrow +\infty} J(v_m + \bar{u}) = J(r(T)) = J_L(T). \quad \square$$

We note that the result of Theorem 5.3 is contained in [16, Thm. 1], and [17] (see also [18]).

If Ω verifies the local Lipschitz condition, it is proved in [6] that

$$(21) \quad J(T) = \sup \left\{ \int_{\Omega} (T \operatorname{div} g + g_{n+1}), g = (g_1, \dots, g_n), \sum_{i=1}^{n+1} g_i^2 \leq 1, g_i \in C_0^1(\Omega) \right\}.$$

The right hand side of (21) is finite if and only if $T \in \text{BV}_b(\Omega)$; then J may be defined by (21) for every $T \in \text{BV}_b(\Omega)$. The following theorem improves the result of Theorem 5.1.

THEOREM 5.4. *If $T \in \text{BV}_b(\mathbb{R}^n)$ we have*

$$(22) \quad J_L(T) = J(r(T))$$

where J is defined by (21).

Proof. If $\{T_m\} \subset W_{s\text{-loc}}^{1,1}(\mathbb{R}^n)$ and $T_m \xrightarrow{w^*} T$ then $r(T_m) \rightarrow r(T)$ in $L^1(\Omega)$ and by

(21) we have

$$\begin{aligned}
 J(r(T)) &= \sup \left\{ \lim_{m \rightarrow +\infty} \int_{\Omega} \left(- \sum_{i=1}^n T_m(g_i)_{x_i} + g_{n+1} \right), g_i \text{ as in (21)} \right\} \\
 &\leq \liminf_{m \rightarrow +\infty} \sup \left\{ \int_{\Omega} \left(- \sum_{i=1}^n T_m(g_i)_{x_i} + g_{n+1} \right), g_i \text{ as in (21)} \right\} \\
 &= \liminf_{m \rightarrow +\infty} \sup \left\{ \int_{\Omega} \left(\sum_{i=1}^n g_i(T_m)_{x_i} + g_{n+1} \right), g_i \text{ as in (21)} \right\} \\
 &= \liminf_{m \rightarrow +\infty} \int_{\Omega} (1 + |\nabla T_m|^2)^{1/2}.
 \end{aligned}$$

Hence

$$J(r(T)) \leq J_{\mathcal{L}}(T).$$

Now let μ be the $(n+1)$ -dimensional measure whose last component is the Lebesgue measure and the other components are the components of ∇T ; we obtain (see e.g. [11])

$$\int_{\Omega} (1 + |\nabla T_h|^2)^{1/2} \leq \int_{\Omega^{(h)}} |\mu|,$$

and so

$$J_{\mathcal{L}}(T) \leq \int_{\Omega} |\mu|;$$

if (16) holds we have

$$\int_{\Omega} |\mu| = \int_{\Omega} |\mu| = J(r(T)),$$

and (22) is proved. Otherwise we may consider an extension \bar{T} of $r(T)$ such that $\bar{T} \in \text{BV}_b(\mathbb{R}^n)$ and $|\nabla \bar{T}|(\partial\Omega) = 0$; hence we obtain (22) because $J_{\mathcal{L}}(\bar{T}) = J_{\mathcal{L}}(T)$. We remark that \bar{T} exists; e.g. we may take $\bar{T} \in W^{1,1}(\mathbb{R}^n - \bar{\Omega})$ in such a way that the trace (in the sense of Sobolev spaces) of \bar{T} on $\partial\Omega$ is equal to the inner trace (in the sense of BV functions; see [15], [16], [17], [18], [19]) of T . \square .

A more general definition of the trace of a function of bounded variation is given in [20].

REFERENCES

- [1] F. FERRO, *Sul minimo di funzionali definiti sullo spazio delle funzioni di variazione limitata in n dimensioni*, Ann. Mat. Pura Appl., to appear.
- [2] R. T. ROCKAFELLAR, *Dual problems of Lagrange for arcs of bounded variation*, Calculus of Variations and Control Theory, D. L. Russel ed., Academic Press, New York, 1976; Proc. of a Conference of the Mathematical Research Center, University of Wisconsin, Madison, September 1975.
- [3] C. CALIGARIS, F. FERRO AND P. OLIVA, *Sull'esistenza del minimo per problemi di calcolo delle variazioni relativi ad archi di variazione limitata*, Boll. Un. Mat. Ital. 14-B, 5 (1977), pp. 340–369.
- [4] O. CALIGARIS AND P. OLIVA, *Problemi di Bolza per archi di variazione limitata ed estensioni di funzionali variazionali*, Ibid., to appear.
- [5] J. SERRIN, *On the area of curved surfaces*, Amer. Math. Monthly, 68 (1961), pp. 435–440.

- [6] M. MIRANDA, *Un teorema di esistenza e unicità per il problema dell'area minima in n variabili*, Ann. Scuola Norm. Sup. Pisa Sci. Fis. Mat. Serie III, 19 (1965), pp. 233–249.
- [7] J. NEČAS, *Les Méthodes Directes en Théorie des Equations Elliptiques*, Masson et Cie, Paris. Academia, Editeurs, Prague, 1967.
- [8] E. GAGLIARDO, *Caratterizzazioni delle tracce sulla frontiera relative ad alcune classi di funzioni in n variabili*, Rend. Sem. Mat. Univ. Padova, 27 (1957), pp. 284–305.
- [9] R. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
- [10] N. BOURBAKI, *Intégration*, Elements de Mathématique, livre VI, Hermann, Paris, 1952.
- [11] W. H. FLEMING, *Functions whose partial derivatives are measures*, Illinois J. Math. 4 (1960), pp. 452–478.
- [12] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, part I, Interscience, New York, 1958.
- [13] R. T. ROCKAFELLAR, *Integral functionals, normal integrands and measurable selections*, Non Linear Operators and the Calculus of Variations, Lecture Notes in Mathematics, no. 543, Springer-Verlag, Berlin-Heidelberg-New York, 1976.
- [14] J. SERRIN, *On the definition and properties of certain variational integrals*, Trans. Amer. Math. Soc., 101 (1961), pp. 139–167.
- [15] M. MIRANDA, *Distribuzioni aventi derivate misure. Insiemi di perimetro localmente finito*, Ann. Scuola Norm. Sup. Pisa Sci. Fis. Mat. serie III, 19 (1964), pp. 27–56.
- [16] ———, *Un principio di massimo forte per le frontiere minimali e una sua applicazione alla risoluzione del problema al contorno per l'equazione delle superfici di area minima*, Rend. Sem. Mat. Univ. Padova, 45 (1971), pp. 355–366.
- [17] ———, *Boundaries of Caccioppoli sets in the calculus of variations*, from C.I.M.E. III ciclo 1972, Geometric measure theory and minimal surfaces, Coordinatore Prof. E. Bombieri, ed. Cremonese, Roma, 1973.
- [18] ———, *Dirichlet problem with L^1 data for the non-homogeneous minimal surface equation*, Indiana Univ. Math. J., 24 (1974), pp. 227–241.
- [19] ———, *Comportamento delle successioni convergenti di frontiere minimali*, Rend. Sem. Mat. Univ. Padova, 38 (1967), pp. 238–257.
- [20] T. GROMARD, *Valeurs frontière des fonctions à variation bornée sur un ouvert de \mathbb{R}^n* , Université Paris XI, U.E.R. Mathématique, 91-Orsay (France) n. 45, octobre 1973.

A SUFFICIENT CONDITION FOR LOCAL CONTROLLABILITY*

HÉCTOR J. SUSSMANN†

Abstract. A system S of vector fields is locally controllable at point p if, for every positive time t , the set of points reachable from p by an S -trajectory in time $\leq t$ contains p in its interior. Let K be the convex hull of the values $X(p)$ of those $X \in S$ for which $X(p) \neq 0$. It is well known that S is l.c. at p if $0 \in \text{interior}(K)$, and that S is not l.c. at p if $0 \notin K$. We prove that these are the only cases in which it is possible to determine if S is l.c. at p by just looking at the values at p of the elements of S . We prove a sufficient condition for local controllability which gives new information for the case when $0 \in K$ but $0 \notin \text{interior}(K)$.

1. Introduction. This paper is a first step towards the determination of high order conditions for local controllability at a point. Consider a finite set S of vector fields. An S -trajectory is a continuous curve which is a finite concatenation of integral curves of vector fields in S . A point q is said to be S -reachable from p if there is an S -trajectory γ , defined on the closed interval $[0, t]$, for some $t \geq 0$, such that $\gamma(0) = p$, $\gamma(t) = q$. If $t \leq T$, then we say that q is S -reachable from p in time $\leq T$. We call S locally controllable (l.c.) at p if, for every $T > 0$, the set of points S -reachable from p in time $\leq T$ contains p in its interior.

We emphasize that, in the definition of “ S -trajectory”, an integral curve of a vector field X is a mapping γ defined on some time interval I , such that $\dot{\gamma}(t) = X(\gamma(t))$ for $t \in I$. In particular, an integral curve of X run backwards in time is no longer an integral curve of X . Hence an S -trajectory run in reverse is not an S -trajectory except in the special case when S is symmetric (i.e. when S has the property that, if $X \in S$, then $-X \in S$).

For symmetric real analytic S , it is easy to give a necessary and sufficient condition for local controllability. Let $L(S)$ be the smallest Lie algebra of vector fields that contains S . Then S is locally controllable at p if and only if the set $L(S)(p)$ of values at p of the elements of $L(S)$ linearly spans the set of all tangent vectors at p (cf. for instance Lobry [3]). For nonsymmetric S , purely Lie algebraic conditions no longer suffice, and the relevant conditions involve inequalities and convexity arguments. Yet, in any case, it is clear that local controllability at a point p is a property of the germs at p of the elements of S . In the real analytic case, these germs are characterized by their Taylor series at p . Hence it is natural to seek to characterize local controllability at p in terms of conditions involving the values at p of the vector fields in S , and of their derivatives. One should expect a situation somewhat resembling what happens in the search for conditions for a point to be a local minimum of a function, i.e. that, for each k , one would be able to partition the possible sets S into three classes C^k , N^k , D^k where

- a) the decision whether S belongs to C^k or to N^k or to D^k only depends on the derivatives at p of order $\leq k$ of the elements of S ,
- b) if S is in C^k then it is locally controllable at p ,
- c) if S is in N^k then it is not l.c. at p , and
- d) if S is in D^k then it is not possible to tell whether or not S is l.c. at p by just looking at the derivatives of order $\leq k$ at p .

* Received by the editors July 8, 1977, and in revised form November 14, 1977.

† Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903. This work was partially supported by the National Science Foundation under Grant MPS73-08524 A03.

Let us call S *k-determined for local controllability at p* if it is possible to tell whether or not S is l.c. at p by just looking at the derivatives at p or order $\leq k$ of the elements of S . In this paper, we describe the zero-determined sets, and give a sufficient condition A^1 for local controllability which involves first derivatives and is of interest when S is not zero-determined for l.c. at p . It turns out that the important object to look at is the convex hull K of the values $X(p)$ of those $X \in S$ for which $X(p) \neq 0$. If 0 is in the interior of K , then S is l.c. at p . If $0 \notin K$, then S is not l.c. at p . Finally, if 0 is in K but not interior to K , we are in the nonzero-determined case, and we need derivatives to decide if S is l.c. at p .

Remarks. 1) That $0 \in \text{interior}(K)$, $0 \notin K$ imply, respectively, that S is l.c. at p and that S is not l.c. at p , are well known facts.

2) Our definition of local controllability is one of many possible "controllability" conditions that can be studied. Another interesting condition is "controllability at p " defined as follows: S is controllable at p if the set of points that are S -reachable from p contains p in its interior. The study of this condition is likely to be much harder, since the condition does not only depend on the germs at p of the $X \in S$.

3) The technique utilized here is inspired by Krener's work on the high order maximum principle (Krener [2]). Problems concerning controllability along a reference trajectory have been studied by Hermes [1].

2. Zeroth order conditions. If S is a set of vector fields on an open region Ω of R^n , and if $p \in \Omega$, we use $S(p)$ to denote the set of all values at p of the elements of S , i.e.

$$S(p) = \{X(p) : X \in S\}.$$

The convex hull of a set A of vectors is denoted by $\text{co } A$. If W is a linear subspace of R^n , and A a subset of W , then the interior of A relative to W will be denoted by $\text{int}_W A$. The absolute interior of a subset A of R^n (i.e. $\text{int}_{R^n} A$) will be denoted by $\text{int } A$.

We say that a set S of vector fields *satisfies condition C^0* at p if $0 \in \text{int co } S(p)$. We say that S *satisfies condition N^0* at p if the origin does not belong to the convex hull of the values $X(p)$ that correspond to those $x \in S$ for which $X(p) \neq 0$.

THEOREM 1. *Let S be a finite set of vector fields on an open set $\Omega \subseteq R^n$, and let $p \in \Omega$. Then:*

- (1) *If C^0 holds at p , the set S is locally controllable at p .*
- (2) *If N^0 holds at p , then S is not l.c. at p .*

Proof. Part (1) is well known (and follows easily from, e.g., the lemma of § 4 below).

Now suppose that N^0 holds at p . Let $S = S' \cup S''$ where

$$S' = \{X : X \in S, X(p) = 0\}, \quad S'' = \{X : X \in S, X(p) \neq 0\}.$$

Then $S''(p)$ is a compact convex set which does not contain 0 . Hence there is an affine function $f: R^n \rightarrow R$ such that $f(p) = 0$ and $f(q) > 0$ for all $q \in S''(p)$. It follows that Xf —the directional derivative of f in the direction of X —is positive at p , for each X in S'' . Since S'' is finite, we can find a neighborhood U of p and a constant $\alpha > 0$ such that $Xf(q) \geq \alpha$ for all $q \in U$, $X \in S''$.

Moreover, each $X \in S'$ satisfies $\|X(q)\| \leq c\|q - p\|$ for q in some neighborhood of p , where c is some positive constant. Hence we can, by shrinking U if necessary, assume that both bounds $Xf(q) \geq \alpha$ for $X \in S''$, and $\|X(q)\| \leq c\|q - p\|$ for $X \in S'$, hold

throughout U . Also, we may assume that $\|X(q)\| \leq D$ for $X \in S''$, $q \in U$, and that $|Xf(q)| \leq E\|q - p\|$ for $q \in U$, $X \in S'$.

Now let $t > 0$ and let $s \rightarrow x(s)$, $0 \leq s \leq t$ be an S -trajectory, with $x(0) = p$. If $S' = \{X_1, \dots, X_r\}$, $S'' = \{Y_1, \dots, Y_{r'}\}$, then the curve $s \rightarrow x(s)$ is the solution of a system

$$\frac{dx}{ds} = \sum_{i=1}^r u_i(s)X_i(x) + \sum_{j=1}^{r'} v_j(s)Y_j(x), \quad x(0) = p$$

where u_i, v_j are suitable characteristic functions of unions of intervals. Then we have the bound

$$\left\| \frac{dx}{ds} \right\| \leq c \left(\sum_{i=1}^r u_i(s) \right) \|x - p\| + D \left(\sum_{j=1}^{r'} v_j(s) \right)$$

i.e.

$$\left\| \frac{dy}{ds} \right\| \leq cu(s)\|y(s)\| + Dv(s),$$

where $u = \sum u_i$, $v = \sum v_j$, $y = x - p$.

Since $y(0) = 0$, Gronwall's inequality gives

$$\|y(s)\| \leq D e^{c \int_0^s u(\sigma) d\sigma} \int_0^s v(\sigma) d\sigma.$$

Now $\int_0^s v(\sigma) d\sigma = \tau(s)$ is the time between 0 and s during which our curve is an S'' -trajectory, whereas $\int_0^s u(\sigma) d\sigma = \eta(s)$ is the time during which it is an S' -trajectory.

Let $h(s) = f(x(s))$. Then, because $Xf(q) \geq \alpha$ for $X \in S''$, we have $dh/ds \geq \alpha$ as long as s is a point where our curve is an S'' -trajectory. On the other hand, we have

$$\left| \frac{dh}{ds} \right| \leq E\|y(s)\|,$$

as long as s is a point where the curve is an S' -trajectory. Hence, if s is such a point, we have

$$\left| \frac{dh}{ds} \right| \leq DE e^{c\eta(s)} \tau(s),$$

and therefore

$$\frac{dh}{ds} \geq -DE e^{c\eta(s)} \tau(s),$$

Since $h(0) = 0$, we conclude that

$$\begin{aligned} h(s) &\geq \alpha \tau(s) - DE e^{c\eta(s)} \tau(s) \eta(s) \\ &= [\alpha - DE \eta(s) e^{c\eta(s)}] \tau(s). \end{aligned}$$

If t is chosen so small that

$$DEt e^{ct} < \alpha$$

then $DE \eta(s) e^{c\eta(s)} \leq DEt e^{ct} < \alpha$ for $0 \leq s \leq t$, so $h(s) \geq 0$. Hence, as long as t is small enough, all S -trajectories defined for $0 \leq s \leq t$ stay in the half-space $\{q: f(q) \geq f(p)\}$. This shows that S is not locally controllable at p . Q.E.D.

The preceding theorem does not completely solve the problem of deciding when S is l.c. at p , since there is a gap between conditions C^0 and N^0 . However, Theorem 1 is the best result that can be obtained by using only the values at p of the components of the elements of S . To see this, let us call a finite set S *zero-determined for local controllability at p* if, whenever S' is any other finite set such that $S'(p) = S(p)$, then S is locally controllable at p if and only if S' is. (Intuitively, S is zero determined for l.c. at p if it is possible to decide whether S is l.c. at p by looking at the values at p of the elements of S .)

THEOREM 2. *Let S be a finite set of vector fields on Ω , and let $p \in \Omega$. Then S is zero-determined for local controllability at p if and only if S satisfies C^0 or N^0 at p .*

Proof. If S satisfies C^0 at p , then every S' such that $S'(p) = S(p)$ will also satisfy C^0 at p . Hence every such S' is locally controllable at p , so S is zero-determined for local controllability at p . A similar reasoning applies if S satisfies N^0 at p .

Now assume that neither C^0 nor N^0 hold for S at p . We have to find sets S' , S'' such that $S'(p) = S''(p) = S(p)$, and that S' is l.c. at p , but S'' is not l.c. at p . The construction of S' is easy. For each $X \in S$, let \tilde{X} be the constant vector field on R^n whose value is $X(p)$. Let $f: R^n \rightarrow R$ be a nonzero linear functional such that $f(v) \geq 0$ for all $v \in \text{co } S(p)$ (such an f exists, because $0 \notin \text{int co } S(p)$). It is easy to see that, if V is a constant vector field, and if $V(x) \equiv v$, for $x \in R^n$, then $Vf \equiv f(v)$ (where Vf is the directional derivative of f in the direction of V). Therefore $Vf \geq 0$ for all $V \in S'$, so that $f(q) \geq f(p)$ for all q that are reachable from p by a trajectory of S' . It follows that all such q are contained in a half-space determined by a hyperplane through p . Hence S' is not l.c. at p .

Now let us construct S'' . It is clearly no loss of generality to assume that p is the origin of R^n . Let $S = \{X_1, \dots, X_r\}$. Let $b_i = X_i(0)$, $i = 1, \dots, r$. Since we are assuming that N^0 does not hold, we can express 0 as a convex combination of the nonzero b_i . By reordering the X_i , if necessary, we can assume that, for some r' such that $1 \leq r' \leq r$:

$$0 = \sum_{i=1}^{r'} a_i b_i,$$

$b_i \neq 0$, $a_i > 0$ for $i = 1, \dots, r'$.

Since the b_i are nonvanishing, it follows that $r' \geq 2$. So $b_1 \neq 0$, $b_2 \neq 0$. Now choose a square matrix A such that the linear system

$$\dot{x} = Ax + ub_1$$

is completely controllable, i.e. that $b_1, Ab_1, \dots, A^{n-1}b_1$ span R^n .

Define vector fields \tilde{X}_i by

$$\tilde{X}_i(x) = Ax + b_i, \quad i = 1, \dots, r.$$

Let $S'' = \{\tilde{X}_1, \dots, \tilde{X}_r\}$. It is clear that $S''(0) = S(0)$. We will show that S'' is locally controllable at 0 by proving that the (possibly) smaller set

$$S''' = \{\tilde{X}_1, \dots, \tilde{X}_{r'}\}$$

is l.c. at 0.

Let σ be the canonical simplex of dimension $r' - 1$, i.e. the set of those points $(s_1, \dots, s_{r'})$ in $R^{r'}$ such that $0 \leq s_i$ for $i = 1, \dots, r'$, and that $s_1 + \dots + s_{r'} = 1$.

Let $R(t)$ denote the set of points that are S''' -reachable from 0 in time $\leq t$. Then $x \in R(t)$ if and only if

$$(*) \quad x = \int_0^t e^{(t-\tau)A} \left(\sum_{i=1}^{r'} u_i(\tau) b_i \right) d\tau$$

with $u = (u_1, \dots, u_r) \in PC_t(E(\sigma))$, where $E(\sigma)$ is the set of vertices of the simplex σ and, for any set F , $PC_t(F)$ denotes the set of piecewise constant F -valued functions defined on the closed interval $[0, t]$.

Let $\tilde{R}(t)$ denote the set of all x that are of the form $(*)$, with $u \in BM_t(\sigma)$, where we use the notation $BM_t(F)$ to denote the set of all bounded measurable F -valued functions defined on $[0, T]$. Let $\hat{R}(t)$ be the set of all x of the form $(*)$ with $u \in BM_t(R')$. Then $\hat{R}(t)$ is a linear subspace of R^n . By taking $u(s) = (1, 0, \dots, 0)$ for $0 \leq s \leq \tau$, τ being any number $\leq t$, we see that $e^{tA} \int_0^\tau e^{-sA} b_1 ds$ is in $\hat{R}(t)$. If we differentiate k times with respect to τ , and then set $\tau = 0$, we conclude that $e^{tA} A^k b_1 \in \hat{R}(t)$. Since this is true for all k , we conclude that $\hat{R}(t) = R^n$. If $x \in \hat{R}(t)$, it is clear that $\alpha x \in \hat{R}(t)$ for some $\alpha > 0$. Hence the compact convex set $\tilde{R}(t)$ contains 0 in its interior.

Let S^* be the set $\{-\tilde{X}_1, \dots, -\tilde{X}_r\}$, i.e. S^* is obtained by replacing the elements of S''' by their negatives or, equivalently, by replacing the $b_{i_j} A$ by $-b_{i_j} A$. Define $R^*(t)$, $\tilde{R}^*(t)$ in a way similar to the definition of $R(t)$, $\tilde{R}(t)$. Fix $t > 0$, $0 < \tau < t$. It follows from [4, Thm. 3.3] that $R(\tau) \subseteq \text{closure (interior } R(\tau))$, the interior being taken relative to the maximal integral manifold I of S''' through 0. Since $\tilde{R}(\tau) \subseteq I$, we conclude that I is n -dimensional, so the interior of $R(\tau)$ referred to above is the absolute interior of $R(\tau)$. Let $Z = \text{interior } (R(\tau))$. Since the closure of $R(\tau)$ is $\tilde{R}(\tau)$, it follows that the open set Z is dense in $\tilde{R}(\tau)$. Hence Z meets every set that contains a neighborhood of 0, since $0 \in \tilde{R}(\tau)$. In particular, Z meets $\tilde{R}^*(t - \tau)$. Since Z is open and $R^*(t - \tau)$ is dense in $\tilde{R}^*(t - \tau)$, it follows that Z meets $R^*(t - \tau)$. Let $p \in Z \cap R^*(t - \tau)$. Then p is reachable from 0 in time $t - \tau$ by an S^* trajectory. So 0 is reachable from p by an S''' -trajectory in time $t - \tau$. Let u be the corresponding controller. Then u steers the open set Z to an open set containing 0, in time $t - \tau$. Since $Z \subseteq R(\tau)$, it follows that $R(t)$ contains a neighborhood of 0. Q.E.D.

3. A sufficient condition for local controllability. Let S be a finite set of vector fields on an open set $\Omega \subseteq R^n$. For each point $p \in \Omega$, if the convex hull $\text{co } S(p)$ contains the zero vector, then there is a unique linear subspace $L(S, p) \subseteq R^n$ of maximum dimension with the property that

$$0 \in \text{int}_{L(S,p)} [\text{co } S(p) \cap L(S, p)].$$

Let S_p^0 denote the set of all vector fields $x \in S$ such that $x(p) \in L(S, p)$. We can characterize S_p^0 as the largest subset T of S such that 0 can be written as a linear combination of the $X(p)$, $X \in T$, with strictly positive coefficients. Let S_p^1 denote the set of all vector fields of the form $[X, Y]$, with $X \in S_p^0$, $Y \in S_p^0$.

We say that S satisfies condition A_p^1 if the zero vector belongs to the interior of the convex hull of $S(p) \cup S_p^1(p)$.

We are now ready to state our sufficient condition:

THEOREM 3. Assume that S satisfies condition A_p^1 . Then S is locally controllable at p .

Example 1. Let S consist of two vector fields X and Y in the plane. At a point p , the set S_p^0 is clearly empty if $X(p)$ and $Y(p)$ are linearly independent, or if one of these two vectors vanishes and the other one does not, or if both are nonzero and multiples of each other with a positive coefficient. In the remaining cases, $S_p^0 = \{X, Y\}$. So $S_p^0 \neq \emptyset$ if and only if a) X and Y both vanish at p , or b) $X(p)$ and $Y(p)$ are both nonzero and $X(p) = -cY(p)$, with $c > 0$.

In case a) the Lie bracket of X and Y will also vanish at p . In case b), condition A_p^1 will be satisfied if and only if $[X, Y](p)$ is linearly independent from $X(p)$. So we conclude that A_p^1 holds if and only if b) holds and the vectors $X(p)$, $[X, Y](p)$ are linearly independent. As an example, let X and Y be given, in coordinates, by $(1, 0)$

and $(-1, x)$, respectively. Then $L(S, 0) = \{X, Y\}$, $[X, Y] = (0, 1)$ and therefore condition A_0^1 holds. Hence S is locally controllable at 0.

Condition b) has a simple geometric meaning. It says that the integral curves of X and Y at p are tangent, and that their tangent vectors point in opposite directions. When this is so, and X and Y are analytic, it is not hard to completely determine the necessary and sufficient condition for local controllability at p . Defining, as usual, $(\text{ad } X)(Z) = [X, Z]$, we find that S is locally controllable at p if and only if b) holds and if there is an *odd* k such that $(\text{ad } X)^k(Y)(p)$ is linearly independent from $X(p)$ but $(\text{ad } X)^j(Y)(p)$ is a multiple of $X(p)$ for $j < k$. (If $(\text{ad } X)^k(Y)(p)$ is a multiple of $X(p)$ for all k , and if b) holds, then the integral curves of X and Y through p coincide as sets, so S is not locally controllable at p . Otherwise, let k be the first positive integer such that $(\text{ad } X)^k(Y)(p)$ is not a multiple of $X(p)$. Then k is even if the curves cross each other, and odd if they do not.) Condition A_p^1 holds when the local controllability condition stated above is satisfied with $k = 1$.

Example 2. In R^3 , with the usual x, y, z coordinates, let the four vector fields A, B, C, D be given by $(1, 0, 0)$, $(-1, x, 0)$, $(0, 1, 1)$ and $(0, 1, -1)$, respectively. Then, if p is the origin of R^3 , and $S = \{A, B, C, D\}$, it is easy to see that $S_p^0 = \{A, B\}$. So S_p^1 consists of the vector fields E and F , where $E = [A, B] = (0, 1, 0)$ and $F = [B, A] = (0, -1, 0)$. The convex hull of the values at 0 of the six vector fields A, B, C, D, E and F contains the origin in its absolute interior. Therefore the system S is locally controllable at p .

Remark. In the statement of condition A_p^1 , the set S_p^1 consists only of the brackets $[X, Y]$, for X and Y in S_p^0 . No higher order brackets such as $[X, [X, Y]]$ are included. To see why this restriction is necessary, let us suppose that, in analogous fashion to what has been done above, we were to define S_p^2 to be the set of all triple brackets of elements of S_p^0 , and then let $A_p^2 = S(p) \cup S_p^1(p) \cup S_p^2(p)$. Then Theorem 3 would not be true if condition A_p^1 is replaced in its statement by condition A_p^2 . To see this, consider the system S of three vector fields A, B, C in the plane, given in local coordinates x, y by $(1, 0)$, $(-1, x^2)$, and $(0, -1)$, respectively. Then $[A, B] = (0, x)$ and $[A, [A, B]] = (0, 1)$. Therefore, A_p^1 is not satisfied, but A_p^2 holds, if p is the origin of R^2 . But it is easy to see that S is not locally controllable at p , and therefore A_p^2 does not imply local controllability at p .

4. Proof of Theorem 3. For the purpose of the proof, we shall assume that $p = 0$. Therefore S is a finite set of vector fields, defined in a neighborhood of $0 \in R^n$. We can think of a vector field X on $\Omega \subseteq R^n$ as a smooth vector-valued function $X: \Omega \rightarrow R^n$. Then $\partial X / \partial x$ is a well defined matrix-valued function. The rows of $\partial X / \partial x$ are the gradients of the components of X . We use Φ_t^X to denote the flow of X , so that

$$t \rightarrow \Phi_t^X(x)$$

is the integral curve of X which goes through x when $t = 0$. It follows that

$$(1) \quad \frac{\partial}{\partial t}(\Phi_t^X(x)) = X(\Phi_t^X(x))$$

and therefore

$$(2) \quad \frac{\partial^2}{\partial t^2}(\Phi_t^X(x)) = \frac{\partial X}{\partial x}(\Phi_t^X(x))X(\Phi_t^X(x)).$$

In particular

$$(3) \quad \frac{\partial}{\partial t}(\Phi_t^X(x))_{t=0} = X(x),$$

and

$$(4) \quad \frac{\partial^2}{\partial t^2}(\Phi_t^X(x))_{t=0} = \frac{\partial X}{\partial x}(x)X(x).$$

Therefore

$$(5) \quad \Phi_t^X(x) = x + tX(x) + \frac{t^2}{2} \frac{\partial X}{\partial x}(x)X(x) + o(t^2),$$

the o being uniform in x as x varies in a compact set.

If X and Y are two vector fields, we recall that the *Lie bracket* of X and Y has the expression

$$(6) \quad [X, Y](x) = \frac{\partial Y}{\partial x}(x)X(x) - \frac{\partial X}{\partial x}(x)Y(x).$$

(The reader who is not familiar with other definitions of the Lie bracket may take (6) to be the definition of $[X, Y]$.)

Successive applications of (5) give:

$$\begin{aligned} \Phi_t^X \Phi_t^Y(x) &= \Phi_t^X\left(x + tY(x) + \frac{t^2}{2} \frac{\partial Y}{\partial x}(x)Y(x) + o(t^2)\right) \\ &= x + tY(x) + \frac{t^2}{2} \frac{\partial Y}{\partial x}(x)Y(x) + tX(x + tY(x)) + \frac{t^2}{2} \frac{\partial X}{\partial x}(x)X(x) + o(t^2) \\ &= x + t(X + Y)(x) + \frac{t^2}{2} \left(\frac{\partial Y}{\partial x}(x)Y(x) + \frac{\partial X}{\partial x}(x)X(x) + 2 \frac{\partial X}{\partial x}(x)Y(x) \right) + o(t^2). \end{aligned}$$

The identity

$$(7) \quad \left(\frac{\partial X}{\partial x} + \frac{\partial Y}{\partial x} \right)(X + Y) = \frac{\partial X}{\partial x}X + \frac{\partial Y}{\partial x}Y + 2 \frac{\partial X}{\partial x}Y + [X, Y]$$

enables us to conclude that

$$(8) \quad \Phi_t^X \Phi_t^Y(x) = x + tZ(x) + \frac{t^2}{2} \frac{\partial Z}{\partial x}(x)Z(x) + \frac{t^2}{2} [Y, X](x) + o(t^2)$$

where $Z = X + Y$.

More generally, suppose that $\bar{X} = (X^1, \dots, X^k)$ is a finite sequence of vector fields, and let $Z = X^1 + \dots + X^k$. A reasoning similar to the one used to establish (8) enables us to conclude that

$$(9) \quad \Phi_t^{X^k} \Phi_t^{X^{k-1}} \dots \Phi_t^{X^1}(x) = x + tZ(x) + \frac{t^2}{2} \frac{\partial Z}{\partial x}(x)Z(x) + \frac{t^2}{2} U^{\bar{X}}(x) + o(t^2)$$

where

$$(10) \quad U^{\bar{X}} = \sum_{1 \leq i < j \leq k} [X^i, X^j].$$

Indeed, we have already proved (9) when $k=2$. The general case follows by induction. If (9) holds for an integer k , let $Z^1 = Z + X^{k+1}$, and let $U^{\bar{X}^1}$ be defined in the same way as $u^{\bar{X}}$, with k replaced by $k+1$. Then $U^{\bar{X}^1} = U^{\bar{X}} + [Z, X^{k+1}]$. Applying (9) and then (5) we get

$$\begin{aligned}
 \Phi_t^{X^{k+1}} \Phi_t^{X^k} \cdots \Phi_t^{X^1}(x) &= \Phi_t^{X^{k+1}} \left(x + tZ(x) + \frac{t^2}{2} \left(\frac{\partial Z}{\partial x} Z \right)(x) + \frac{t^2}{2} U^{\bar{X}}(x) + o(t^2) \right) \\
 &= x + tZ(x) + \frac{t^2}{2} \left(\frac{\partial Z}{\partial x} Z \right)(x) + \frac{t^2}{2} U^{\bar{X}}(x) \\
 &\quad + tX^{k+1}(x + tZ(x)) + \frac{t^2}{2} \left(\frac{\partial X^{k+1}}{\partial x} X^{k+1} \right)(x) + o(t^2) \\
 &= x + tZ(x) + \frac{t^2}{2} \left[\frac{\partial Z}{\partial x} Z + \frac{\partial X^{k+1}}{\partial x} X^{k+1} \right](x) + \frac{t^2}{2} U^{\bar{X}}(x) \\
 &\quad + tX^{k+1}(x) + t^2 \left(\frac{\partial X^{k+1}}{\partial x} Z \right)(x) + o(t^2) \\
 &= x + tZ^1(x) + \frac{t^2}{2} \left\{ \left(\frac{\partial Z^1}{\partial x} Z^1 \right)(x) + [Z, X^{k+1}](x) + U^{\bar{X}}(x) \right\} + o(t^2),
 \end{aligned}$$

where in the last step, we have used (7). Since $U^{\bar{X}} + [Z, X^{k+1}] = U^{\bar{X}^1}$, it follows that (9) is indeed valid for $k+1$.

Now suppose that

$$(11) \quad (X^1 + \cdots + X^k)(0) = 0.$$

Let π be a permutation of $\{1, \dots, k\}$. Define

$$(12) \quad U_{\pi}^{\bar{X}} = \sum_{1 \leq i < j \leq k} [X^{\pi(i)}, X^{\pi(j)}],$$

and

$$(13) \quad \Phi_t^{\bar{X}, \pi} = \Phi_t^{X^{\pi(k)}} \Phi_t^{X^{\pi(k-1)}} \cdots \Phi_t^{X^{\pi(1)}}.$$

Then (9) gives

$$(14) \quad \Phi_t^{\bar{X}, \pi}(0) = \frac{t^2}{2} U_{\pi}^{\bar{X}}(0) + o(t^2).$$

Now suppose that π_1 and π_2 are two permutations. Put $h = \Phi_t^{\bar{X}, \pi_2}(0)$. Then by (9)

$$\begin{aligned}
 \Phi_t^{\bar{X}, \pi_1} \Phi_t^{\bar{X}, \pi_2}(0) &= \Phi_t^{\bar{X}, \pi_1}(h) \\
 (15) \quad &= h + tZ(h) + \frac{t^2}{2} \frac{\partial Z}{\partial x}(h) Z(h) + \frac{t^2}{2} U_{\pi_2}^{\bar{X}}(h) + o(t^2).
 \end{aligned}$$

On the other hand, by (14), $h = (t^2/2)U_{\pi_1}^{\bar{X}}(0) + o(t^2)$. Moreover, $Z(0) = 0$. Therefore

$$Z(h) = O(t^2),$$

so that

$$tZ(h) = o(t^2).$$

The third term of the right side of (15) is also $o(t^2)$. Moreover

$$U_{\pi_2}^{\bar{X}}(h) = U_{\pi_2}^{\bar{X}}(0) + o(1)$$

so that

$$\frac{t^2}{2} U_{\pi_2}^{\bar{X}}(h) = \frac{t^2}{2} U_{\pi_2}^{\bar{X}}(0) + o(t^2).$$

Therefore, we can conclude from (15) that

$$\Phi_t^{\bar{X}, \pi_1} \Phi_t^{\bar{X}, \pi_2}(0) = h + \frac{t^2}{2} U_{\pi_2}^{\bar{X}}(0) + o(t^2)$$

so that

$$(16) \quad \Phi_t^{\bar{X}, \pi_1} \Phi_t^{\bar{X}, \pi_2}(0) = \frac{t^2}{2} [U_{\pi_1}^{\bar{X}}(0) + U_{\pi_2}^{\bar{X}}(0)] + o(t^2).$$

It follows easily by induction that, if $\sigma = (\pi_1, \dots, \pi_m)$ is a sequence of permutations of $\{1, \dots, k\}$, and if we define

$$(17) \quad \Psi_t^{\bar{X}, \sigma} = \Phi_t^{\bar{X}, \pi_1} \Phi_t^{\bar{X}, \pi_2} \dots \Phi_t^{\bar{X}, \pi_m},$$

then

$$(18) \quad \Psi_t^{\bar{X}, \sigma}(0) = \frac{t^2}{2} \sum_{i=1}^m U_{\pi_i}^{\bar{X}}(0) + o(t^2).$$

We will apply (18) for a particular choice of the sequence σ . Let A_1 be the set of all the permutations π such that $\pi(1) = 1$ and $\pi(2) = 2$. Similarly, let A_2 be the set of those π for which $\pi(k-1) = 1$ and $\pi(k) = 2$. Let $A = A_1 \cup A_2$. Let the sequence $\bar{\sigma}$ consist of the elements of A (in any order). To apply (18), we must compute the sum

$$\sum_{\pi \in A} U_{\pi}^{\bar{X}}.$$

Now $U_{\pi}^{\bar{X}} = \sum_{i < j} \varepsilon_{\pi}^{ij} [X^i, X^j]$ where $\varepsilon_{\pi}^{ij} = 1$ if $\pi^{-1}(i) < \pi^{-1}(j)$ and $\varepsilon_{\pi}^{ij} = -1$ if $\pi^{-1}(i) > \pi^{-1}(j)$. Therefore

$$\sum_{\pi \in A} U_{\pi}^{\bar{X}} = \sum_{i < j} \alpha^{ij} [X^i, X^j]$$

where

$$\alpha^{ij} = \sum_{\pi \in A} \varepsilon_{\pi}^{ij}.$$

If $i > 2$, let τ be the transposition (ij) . Then $\pi \in A$ iff $\tau\pi \in A$, and $\varepsilon_{\tau\pi}^{ij} = -\varepsilon_{\pi}^{ij}$. Therefore $\alpha^{ij} = 0$. If $i = 2$ then $\varepsilon_{\pi}^{ij} = 1$ for $\pi \in A_1$ and $\varepsilon_{\pi}^{ij} = -1$ for $\pi \in A_2$. Since A_1 and A_2 each have $(k-2)!$ elements, it follows that $\alpha^{ij} = 0$.

Similarly, if $i = 1, j > 2$, then $\alpha^{ij} = 0$.

Finally, if $i = 1, j = 2$, then $\varepsilon_{\pi}^{ij} = 1$ for all $\pi \in A$, so that $\alpha^{ij} = c_k$, where

$$(19) \quad \begin{aligned} c_k &= 2 \cdot (k-2)! \quad \text{if } k > 2, \\ c_2 &= 1. \end{aligned}$$

Therefore we have shown that

$$\sum_{\pi \in A} U_{\pi}^{\bar{X}} = c_k[X^1, X^2].$$

Then (18) gives:

$$(20) \quad \Psi_t^{\bar{X}, \bar{\sigma}}(0) = \frac{c_k t^2}{2} [X^1, X^2](0) + o(t^2).$$

Now let E be a finite set of vector fields, with the property that

$$(21) \quad \sum_{X \in E} X(0) = 0.$$

Let $X \in E$, $Y \in E$. Order the elements of E in an arbitrary fashion into a sequence $\mathbf{X} = (X^1, X^2, \dots, X^k)$ in such a way that $X^1 = X$ and $X^2 = Y$. Then define $\bar{\sigma}$ as above. Write $\eta_t^{X, Y}$ for the map $\Psi_t^{\bar{X}, \bar{\sigma}}$ defined above. We then have

$$(22) \quad \eta_t^{X, Y}(0) = \frac{c_k t^2}{2} [X, Y](0) + o(t^2).$$

Now let $(X_1, Y_1), (X_2, Y_2)$ be two pairs of elements of E . Then

$$(23) \quad \eta_{t_1}^{X_1, Y_1}(h) = \eta_{t_1}^{X_1, Y_1}(0) + \left[\frac{\partial}{\partial x} (\eta_{t_1}^{X_1, Y_1}) \right](0) \cdot h + o(|h|).$$

Let $J(t_1, x)$ denote the Jacobian matrix $\partial/\partial x (\eta_{t_1}^{X_1, Y_1})(x)$. Since $\eta_0^{X_1, Y_1}(x) \equiv x$, it follows that $J(0, 0)$ is the identity matrix I . Then $J(t_1, 0) = I + o(1)$ as $t_1 \rightarrow 0$. Substituting this into (23), we get

$$(24) \quad \eta_{t_1}^{X_1, Y_1}(h) = \eta_{t_1}^{X_1, Y_1}(0) + h + o(|h|) + o(1) \cdot O(|h|).$$

Now let $h = \eta_{t_2}^{X_2, Y_2}(0)$, and apply (22) twice. The result is:

$$\eta_{t_1}^{X_1, Y_1} \eta_{t_2}^{X_2, Y_2}(0) = \frac{c_k}{2} (t_1^2 [X_1, Y_1] + t_2^2 [X_2, Y_2])(0) + o(t_1^2) + o(t_2^2) + o(1) \cdot O(t_2^2)$$

so that

$$(25) \quad \eta_{t_1}^{X_1, Y_1} \eta_{t_2}^{X_2, Y_2}(0) = \frac{c_k}{2} (t_1^2 [X_1, Y_1] + t_2^2 [X_2, Y_2])(0) + o(t_1^2 + t_2^2).$$

Formula (25) can be generalized, by an obvious induction, to

$$(26) \quad \eta_{t_1}^{X_1, Y_1} \eta_{t_2}^{X_2, Y_2} \dots \eta_{t_r}^{X_r, Y_r}(0) = \frac{c_k}{2} \left(\sum_{i=1}^r t_i^2 [X_i, Y_i] \right)(0) + o\left(\sum_{i=1}^r t_i^2 \right).$$

For $s \geq 0$, let $t(s) = \sqrt{2s/c_k}$, and put $\zeta_s^{X, Y} = \eta_{t(s)}^{X, Y}$. Then we get from (26):

$$(27) \quad \zeta_{s_1}^{X_1, Y_1} \zeta_{s_2}^{X_2, Y_2} \dots \zeta_{s_r}^{X_r, Y_r}(0) = \sum_{i=1}^r s_i [X_i, Y_i](0) + o(s_1 + \dots + s_r).$$

Formula (27) is valid for $s_1 \geq 0, \dots, s_r \geq 0$ and $X_1, \dots, X_r; Y_1, \dots, Y_r$ arbitrary sequences of elements of a set E for which (21) holds.

We are now ready to complete the proof of our theorem. We are given a finite set S of vector fields defined near 0, and we are assuming that condition (A_0^1) holds. Let \mathcal{F} denote the family of all subsets $F \subseteq S$ such that 0 can be expressed as a linear

combination $\sum_{X \in F} \alpha_X X(0)$ with $\alpha_X > 0$ for all $X \in F$. It is clear that, if F_1 and F_2 are in \mathcal{F} , so is $F_1 \cup F_2$. Therefore, if we let $S_0^0 = \bigcup_{F \in \mathcal{F}} F$ it follows that $S_0^0 \in F$. It is clear that this definition of S_0^0 is equivalent to the one given in § 3. We have

$$(28) \quad 0 = \sum_{X \in S_0} \alpha_X X(0),$$

where $\alpha_X > 0$ for all $X \in S_0^0$.

Clearly, both conditions (A_0^1) and local controllability at 0 are unchanged if each $X \in S$ is replaced by a multiple λX , where $\lambda > 0$ is a constant which may depend on X . After making such a change, we may assume, instead of (28), that

$$(29) \quad \sum_{X \in S_0^0} X(0) = 0.$$

Let S_0^1 be the set of all brackets $[X, Y]$, $X \in S_0^0$, $Y \in S_0^0$. Condition (A_0^1) implies that there exist elements V_1, \dots, V_d of S and pairs $(X_1, Y_1), \dots, (X_r, Y_r)$ of elements of S_0^0 such that the convex hull of the $V_i(0)$ and the $[X_j, Y_j](0)$ contains 0 in its absolute interior.

We now apply the theory developed above. Take E to be S_0^0 . Let $b = d + r$, and let R_+^b denote the set of points $(t_1, \dots, t_b) \in R^b$ such that $t_1 \geq 0, \dots, t_b \geq 0$. Define

$$\mu(s_1, \dots, s_b) = \Phi_{s_1}^{V_1} \Phi_{s_2}^{V_2} \dots \Phi_{s_d}^{V_d} \zeta_{s_{d+1}}^{X_1, Y_1} \dots \zeta_{s_{d+r}}^{X_r, Y_r}(0).$$

Then μ is defined provided s_1, \dots, s_b are nonnegative and sufficiently small. Therefore

I) μ maps $U \cap R_+^b$ into R^n , for some neighborhood U of 0 in R^b .

Moreover, it is clear from the definition that

II) μ is continuous and $\mu(0) = 0$. Formula (27) tells us that, if we put

$$\rho_0(s_{d+1}, \dots, s_b) = \zeta_{s_{d+1}}^{X_1, Y_1} \dots \zeta_{s_b}^{X_r, Y_r}(0)$$

then

$$\rho_0(s_{d+1}, \dots, s_b) = \sum_{i=1}^e s_{d+i} [X_i, Y_i](0) + o(s_{d+1} + \dots + s_b).$$

Now let

$$\rho_1(s_d, \dots, s_b) = \Phi_{s_d}^{V_d} \rho_0(s_{d+1}, \dots, s_b).$$

Then

$$(30) \quad \rho_1(s_d, \dots, s_b) = \Phi_{s_d}^{V_d}(0) + \frac{\partial}{\partial x} (\Phi_{s_d}^{V_d})(0) \cdot \rho_0(s_{d+1}, \dots, s_b) + o(\rho_0(s_{d+1}, \dots, s_b)).$$

By (5), we have

$$(31) \quad \Phi_{s_d}^{V_d}(0) = s_d V_d(0) + o(s_d).$$

Also, the Jacobian matrix $(\partial/\partial x)(\Phi_{s_d}^{V_d})(0)$ equals $(\partial/\partial x)(\Phi_0^{V_d})(0) + o(1)$ as $s_d \rightarrow 0$. Moreover, $(\partial/\partial x)(\Phi_0^{V_d})(0)$ is the identity matrix. Since $\rho_0(s_{d+1}, \dots, s_b)$ is $O(s_{d+1} + \dots + s_b)$, we get

$$(32) \quad \frac{\partial}{\partial x} (\Phi_{s_d}^{V_d})(0) \cdot \rho_0(s_{d+1}, \dots, s_b) = \rho_0(s_{d+1}, \dots, s_b) + o(s_{d+1} + \dots + s_b).$$

Finally we remark that

$$(33) \quad o(\rho_0(s_{d+1}, \dots, s_b)) = o(s_{d+1} + \dots + s_b).$$

If we substitute into (30) the results of (31), (32) and (33) we get

$$(34) \quad \rho_1(s_d, \dots, s_b) = s_d V_d(0) + \sum_{i=1}^r s_{d+i} [X_i, Y_i](0) + o(s_d + \dots + s_b).$$

More generally, we can define ρ_k recursively by

$$\rho_k(s_{d+1-k}, \dots, s_b) = \Phi_{s_{d+1-k}}^{V_{d+1-k}} \rho_{k-1}(s_{d+2-k}, \dots, s_b)$$

and prove by induction that

$$(35) \quad \rho_k(s_{d+1-k}, \dots, s_b) = \sum_{j=d+1-k}^d s_j V_j(0) + \sum_{i=1}^r s_{d+i} [X_i, Y_i](0) + o(s_{d+1-k} + \dots + s_b).$$

Define $\mu_0: R^b \rightarrow R^n$ by

$$\mu_0(s_1, \dots, s_b) = \sum_{j=1}^d s_j V_j(0) + \sum_{i=1}^r s_{d+i} [X_i, Y_i](0).$$

Then (35) can be applied for $k = d$ (so that $\rho_d = \mu$) and we conclude that

$$\text{III)} \quad \mu(s_1, \dots, s_b) = \mu_0(s_1, \dots, s_b) + o(s_1 + \dots + s_b).$$

Let e_1, \dots, e_b be the canonical basis of R^b . Then $\mu_0(e_i) = V_i(0)$ for $1 \leq i \leq d$, and $\mu_0(e_{d+j}) = [X_j, Y_j](0)$ for $1 \leq j \leq r$. It follows from the way the V_i, X_j, Y_j were chosen that

IV) the convex hull of $\mu_0(e_1), \dots, \mu_0(e_b)$ contains 0 in its absolute interior.

Conditions (I) to (IV) imply (by a well known result which is stated as a lemma and proved below) that the image of $U \cap R_+^b$ under μ contains a neighborhood of 0. Moreover, the construction of μ is such that $\mu(s_1, \dots, s_b)$ is in the positive orbit of S from 0, for all $(s_1, \dots, s_b) \in R_+^b$ such that $\mu(s_1, \dots, s_b)$ is defined. Since U can be shrunk as much as desired, we conclude that S is locally controllable at 0, completing the proof.

LEMMA 4. Let U be a neighborhood of 0 in R^b , and let $\mu: U \cap R_+^b \rightarrow R^n$ be a continuous map such that $\mu(0) = 0$. Assume that there is a linear map $\mu_0: R^b \rightarrow R^n$ such that

i) $\mu(t_1, \dots, t_b) = \mu_0(t_1, \dots, t_b) + o(t_1 + \dots + t_b)$ for (t_1, \dots, t_b) converging to zero in $U \cap R_+^b$.

ii) the convex hull of the vectors $\mu_0(e_1), \dots, \mu_0(e_b)$ contains 0 in its absolute interior (where e_1, \dots, e_b is the canonical basis of R^b).

Then $\mu(U \cap R_+^b)$ contains a neighborhood of 0 in R^n .

Proof. Select $n+1$ vectors v_0, \dots, v_n in R^n , such that each v_i is a convex combination of the $\mu_0(e_j)$, and that the v_i are the vertices of an n -simplex σ in R^n which contains 0 in its interior. Then select f_0, \dots, f_n in R^b , convex combinations of the e_j , such that $\mu_0(f_i) = v_i$. Then the f_i are affinely independent (i.e. if $\sum \lambda_i f_i = 0$ and $\sum \lambda_i = 0$ then $\lambda_0 = \dots = \lambda_n = 0$) because the v_i are. Therefore f_0, \dots, f_n are the vertices of an n -simplex $\tau \subseteq R_+^b$, and μ_0 maps τ homeomorphically onto σ . For $\varepsilon > 0$ define a map $\mu^\varepsilon: \tau \rightarrow R^n$ by $\mu^\varepsilon(f) = \varepsilon^{-1} \mu(\varepsilon f)$. Then μ^ε is well defined for ε sufficiently

small (e.g. for ε such that $\varepsilon\tau \subseteq U$). Moreover

$$\begin{aligned}\mu^\varepsilon(f) &= \varepsilon^{-1}\mu(\varepsilon f) = \varepsilon^{-1}(\mu_0(\varepsilon f) + o(\|\varepsilon f\|)) \\ &= \mu_0(f) + \varepsilon^{-1}o(\varepsilon) = \mu_0(f) + o(1),\end{aligned}$$

where $\|\cdot\|$ denotes any norm in R^b and the o 's are uniform in f for $f \in \tau$.

Therefore $\mu^\varepsilon(f) \rightarrow \mu_0(f)$ as $\varepsilon \rightarrow 0$, uniformly for $f \in \tau$. Then the map $\mu^\varepsilon \mu_0^{-1}: \sigma \rightarrow R^n$ converges uniformly to the inclusion map from σ to r^n , as $\varepsilon \rightarrow 0$. By a standard argument using degree theory, it follows that $\mu^\varepsilon \mu_0^{-1}(\sigma)$ contains a neighborhood of 0, if ε is small enough. Therefore $\mu^\varepsilon(\tau)$ contains a neighborhood of 0, and so does $\varepsilon \mu^\varepsilon(\tau)$. But $\mu(\varepsilon\tau) = \varepsilon \mu^\varepsilon(\tau)$ and for small ε , $\varepsilon\tau \subseteq U$. Therefore $\mu(U \cap R_+^b)$ contains a neighborhood of 0. Q.E.D.

REFERENCES

- [1] H. HERMES, *Local controllability and sufficient conditions in singular problems*, J. Differential Equations, 20 (1976), pp. 213–232.
- [2] A. J. KRENER, *The high order maximal principle and its applications to singular extremals*, this Journal, 15 (1977), pp. 256–293.
- [3] C. LOBRY, *Contrôlabilité des Systèmes non Linéaires*, this Journal, 8 (1970), pp. 573–605.
- [4] H. J. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.

CONTROLLABILITY AND A MULTIPLIER RULE FOR NONDIFFERENTIABLE OPTIMIZATION PROBLEMS*

J. WARGA†

Abstract. Let K be a compact subset of a normed vector space \mathcal{X} , C a convex body in a Banach space \mathcal{Y} , $k_0 \in K$, $(\phi, \Phi): K \rightarrow \mathbb{R}^m \times \mathcal{Y}$ continuous, and $\Phi(k_0) \in C$. We introduce the concept of a "directional derivate container" $\Lambda^e(\phi, \Phi)(k_0)$ for (ϕ, Φ) at k_0 whose definition is equivalent to that of a directional derivative in the special case where (ϕ, Φ) is "finitely C^1 ," that is, all the restrictions of (ϕ, Φ) to finite-dimensional convex subsets of K are continuously differentiable. In general, (ϕ, Φ) admits a (nonunique) directional derivate container at k_0 if it can be uniformly approximated by finitely C^1 functions (ϕ_i, Φ_i) whose directional derivatives in some "neighborhood" of k_0 in K , viewed as functions on K to $\mathbb{R}^m \times \mathcal{Y}$, form a bounded and equicontinuous subset of $C(K, \mathbb{R}^m \times \mathcal{Y})$. For a given $\Lambda^e(\phi, \Phi)(k_0)$ we define the corresponding "scalar directional derivate container" $\mathcal{L}\Lambda(\phi, \Phi, C)(k_0)$ as the collection of all (l, λ) such that $l = (l_1, l_2)$ is a weak star limit of $l^i = (l_1^i, l_2^i) \in \mathbb{R}^m \times \mathcal{Y}^*$, $|l_1^i| + |l_2^i| = 1$, $l_2^i y \leq 0$ if the closed ball in \mathcal{Y} of center y and radius $1/i$ is contained in $C - \Phi(k_0)$, $l \neq 0$, and λ is a pointwise limit of functions $l^i \circ M^i: K \rightarrow \mathbb{R}$ with $M^i \in \Lambda^{1/i}(\phi, \Phi)(k_0)$. We then prove a "controllability-multiplier rule" alternative which states (defining $S^F(0, \kappa)$ as the closed ball of center 0 and radius κ) that either there exist $\kappa > 0$ and a finite-dimensional subset K^* of K such that $k_0 \in K^*$ and $\{\phi(k) | k \in K^*, \Phi(k) + S^F(0, \kappa) \subset C\}$ contains a neighborhood of $\phi(k_0)$ in \mathbb{R}^m or there exists $(l_1, l_2, \lambda) \in \mathcal{L}\Lambda(\phi, \Phi, C)(k_0)$ such that $\lambda k_0 = \text{Min}_{k \in K} \lambda k$, $l_2 \Phi(k_0) = \text{Max}_{c \in C} l_2 c$. These results will be used elsewhere to study optimal control problems defined by hereditary functional-integral equations involving nondifferentiable functions of state variables.

1. Introduction. We shall consider a general optimization problem defined by continuous functions

$$\phi^0: K \rightarrow \mathbb{R}, \quad \phi^1: K \rightarrow \mathbb{R}^m, \quad \Phi: K \rightarrow \mathcal{Y}$$

and a convex body $C \subset \mathcal{Y}$, where $m \in \{1, 2, \dots\}$, K is a convex and compact subset of a real normed vector space \mathcal{X} , and \mathcal{Y} is a real Banach space. We shall say that a point $k_0 \in K$ is a *minimizing point* if it yields the minimum of ϕ^0 on the set

$$\{k \in K | \phi^1(k) = 0, \Phi(k) \in C\}.$$

We shall say that a function $h: K \rightarrow \mathbb{R}^p$ is (Φ, C) -controllable at k_0 if there exists $\kappa > 0$ such that

$$S^F(h(k_0), \kappa) \subset \{h(k) | k \in K, \Phi(k) + S^F(0, \kappa) \subset C\},$$

where $S^F(a, \kappa)$ denotes the closed ball of center a and radius κ in the appropriate space.

A large class of optimization problems, including those of optimal control theory, can be described in terms of the above problem. For example, K may represent an appropriate collection of relaxed control functions combined with control parameters, $\Phi(k)$ may be the solution of a differential or functional-integral equation involving the control k , ϕ^0 may be the cost functional, $\phi^1(k) = 0$ may describe the "isoperimetric" restrictions, and $\Phi(k) \in C$ may describe unilateral or other functional restrictions. Because K is compact and (ϕ^0, ϕ^1, Φ) continuous, the existence of a minimizing point k_0 is assured if the *admissible set* $\{k \in K | \phi^1(k) = 0, \Phi(k) \in C\}$ is nonempty. If the function (ϕ^0, ϕ^1, Φ) is $(m+1)$ -differentiable at k_0 , that is, if the restrictions of (ϕ^0, ϕ^1, Φ) to all sufficiently small convex $(m+1)$ -dimensional neighborhoods of k_0 in

* Received by the editors August 18, 1977.

† Department of Mathematics, Northeastern University, Boston, Massachusetts 02115. This work was supported in part by the National Science Foundation under Grant MCS 76-06756.

K are differentiable at k_0 (as is the case in “standard” problems of optimal control), then we have the “multiplier rule” (see, e.g. [4, V.2.3, p. 303]): for every minimizing point k_0 there exists a nonzero functional $l = (l_0, l_1, l_2) \in \mathbb{R} \times \mathbb{R}^m \times \mathcal{Y}^*$ such that $l_0 \geq 0$ and

$$(1) \quad lD(\phi^0, \phi^1, \Phi)(k_0; k - k_0) \geq 0 \quad (k \in K), \quad l_2\Phi(k_0) = \max_{c \in C} l_2c,$$

where \mathcal{Y}^* is the topological dual of \mathcal{Y} and $D\chi(k_0; k - k_0)$ denotes the directional derivative $\lim_{\alpha \rightarrow 0+} \alpha^{-1}[\chi(k_0 + \alpha[k - k_0]) - \chi(k_0)]$.

The arguments used in deriving this rule can be modified to show that if $k_0 \in K$, (ϕ^0, ϕ^1, Φ) is n -differentiable at k_0 for $n = 1, 2, \dots$ and there exists no $l \neq 0$ satisfying (1) then (ϕ^0, ϕ^1) is (Φ, C) -controllable at k_0 . In the context of optimal control, relation (1) yields various generalizations of the Pontryagin maximum principle and of the transversality relations that are applicable, in particular, to the optimal control of functional-integral equations with unilateral restrictions [4, Chaps. VII, VIII] (and thus also to ordinary differential and functional-differential equations and certain partial differential equations).

If the assumption that (ϕ^0, ϕ^1, Φ) is $(m+1)$ -differentiable no longer holds (as is the case with optimal control problems involving differential or functional-integral equations whose defining functions are not differentiable with respect to the state variable) then the multiplier rule cannot even be formulated because $D(\phi^0, \phi^1, \Phi)(k_0; k - k_0)$ need not exist. However, new types of multiplier rules (generalizing Pontryagin's principle) have been derived in the last few years by Kugušev [3], Clarke [1], [2], and Warga [5]–[8] for different categories of optimal control problems involving ordinary differential equations and defined by functions that are Lipschitz continuous in the state variable. In these extended multiplier rules, the (nonexistent) derivatives with respect to the state variable are replaced by elements of certain set-valued “derivatives” (such as Clarke's “generalized derivative” or our “derivate container”).

In the present paper, we shall extend the “controllability-multiplier rule alternative” [either (ϕ^0, ϕ^1) is (Φ, C) -controllable at k_0 or there exists $l \neq 0$ satisfying an analogue to (1)] to the case where (ϕ^0, ϕ^1, Φ) need not be n -differentiable at k_0 for any n but admits a “directional derivate container at k_0 ”. The latter is the case, essentially, if (ϕ^0, ϕ^1, Φ) can be uniformly approximated by continuous functions χ_i whose restrictions to sufficiently small convex finite-dimensional “neighborhoods” of k_0 in K are continuously differentiable and such that the collection of functions $k \rightarrow D\chi_i(k'; k - k')$, for k' “near” k_0 , is a bounded and equicontinuous subset of $C(K, \mathbb{R} \times \mathbb{R}^m \times \mathcal{Y})$. We shall show elsewhere [9] that our present results are applicable, in particular, to optimal control problems involving hereditary functional-integral equations and defined by functions that are Lipschitz-continuous with respect to the state variable.

2. The controllability-multiplier rule alternative. We denote by $M|_K$ the restriction of a function M to K , by \mathcal{T}_N the simplex

$$\left\{ \theta = (\theta^1, \dots, \theta^N) \in \mathbb{R}^N \mid \theta^j \geq 0, \sum_{j=1}^N \theta^j \leq 1 \right\},$$

by $\mathcal{L}(U, V)$ the collection of linear operators from a vector space U to a vector space V , and by $C(K, \mathcal{X})$ the Banach space of continuous functions on K to a Banach space \mathcal{X} with the sup norm. (We use the terms “vector (Banach) space” to mean “vector

(Banach) space over \mathbb{R} and “linear” in the algebraic sense; thus elements of $\mathcal{L}(U, V)$ need not be continuous even if U and V are normed.) We write \triangleq for equal by definition, A° for the interior and $\text{co } A$ for the convex hull of a set A , $d[b, A]$ for the distance from a point b to a set A , and $S^F(A, \kappa)$ for $\{b | d[b, A] \leq \kappa\}$. The term “convex body” means “a closed convex set with a nonempty interior”. For a function χ defined on a convex subset P of \mathbb{R}^N we use the term “differentiable at p ” to mean “differentiable at p relative to P ”; specifically, if $\chi: P \rightarrow \mathcal{X}$ then $\chi'(p)$ is a linear operator on \mathbb{R}^N to \mathcal{X} such that

$$\lim_{q \rightarrow p} |q - p|^{-1} |\chi(q) - \chi(p) - \chi'(p)(q - p)| = 0 \quad \text{as } q \rightarrow p, \quad q \in P - \{p\}.$$

DEFINITION 2.1. *Directional derivate containers.* Let $(\phi, \Phi): K \rightarrow \mathbb{R}^m \times \mathcal{Y}$ be continuous. A collection $\{\Lambda^\varepsilon(\phi, \Phi)(k_0) | \varepsilon > 0\}$ of nonempty subsets of $\mathcal{L}(\mathcal{X}, \mathbb{R}^m \times \mathcal{Y})$, also referred to as $\Lambda^\varepsilon(\phi, \Phi)(k_0)$, is a *directional derivate container* for (ϕ, Φ) at k_0 if there exist continuous functions $(\phi_i, \Phi_i): K \rightarrow \mathbb{R}^m \times \mathcal{Y}$ ($i = 1, 2, \dots$) such that

- (1) $\lim_i (\phi_i, \Phi_i) = (\phi, \Phi)$ uniformly;
- (2) $\Lambda^\varepsilon(\phi, \Phi)(k_0) \subset \Lambda^{\varepsilon'}(\phi, \Phi)(k_0) \quad (\varepsilon' > \varepsilon)$;
- (3) for every $\varepsilon > 0$ the set $\{M|_K | M \in \Lambda^\varepsilon(\phi, \Phi)(k_0)\}$ is a bounded and equicontinuous subset of $C(K, \mathbb{R}^m \times \mathcal{Y})$;
- (4) for every choice of $N \in \{1, 2, \dots\}$, $k^1, \dots, k^N \in K$ and $\varepsilon > 0$ there exist $\delta > 0$, $i^* \in \{1, 2, \dots\}$ and a corresponding set

$$K^* \triangleq \left\{ k_0 + \sum_{j=1}^N \omega^j (k^j - k_0) \mid \omega = (\omega^1, \dots, \omega^N) \in \delta \mathcal{T}_N \right\}$$

such that the functions

$$\omega \rightarrow (\phi_i, \Phi_i)(k_0 + \sum_{j=1}^N \omega^j (k^j - k_0)): \delta \mathcal{T}_N \rightarrow \mathbb{R}^m \times \mathcal{Y} \quad (i = 1, 2, \dots)$$

are continuously differentiable, and for every $k' \in K^*$ and $i \geq i^*$ there exists $M \in \Lambda^\varepsilon(\phi, \Phi)(k_0)$ satisfying

$$D(\phi_i, \Phi_i)(k'; k - k') = M(k - k') \quad (k \in K).$$

For every directional derivate container $\Lambda^\varepsilon(\phi, \Phi)(k_0)$ we define a corresponding set $\mathcal{L}\Lambda(\phi, \Phi, C)(k_0)$, the *scalar directional derivate container* for (ϕ, Φ, C) at k_0 , as the collection of all triplets (l_1, l_2, λ) such that $l = (l_1, l_2) \in \mathbb{R}^m \times \mathcal{Y}^*$, $l \neq 0$, $\lambda \in \mathcal{L}(\mathcal{X}, \mathbb{R})$, $\lambda|_K$ is continuous, and there exist sequences $(l^i) = ((l_1^i, l_2^i))$ in $\mathbb{R}^m \times \mathcal{Y}^*$ and (M^i) with $M^i \in \Lambda^{1/i}(\phi, \Phi)(k_0)$ satisfying

$$\lim_i l_1^i = l_1, \quad \lim_i l_2^i z = l_2 \quad (z \in \mathcal{Y}), \quad |l_1^i| + |l_2^i| = 1,$$

$$\lim_i l^i M^i k = \lambda k \quad (k \in K),$$

$$l_2^i y \leq 0 \quad \text{if } y + \Phi(k_0) + S^F(0, 1/i) \subset C.$$

THEOREM 2.2. *Let $\mathcal{L}\Lambda(\phi, \Phi, C)(k_0)$ be a scalar directional derivate container for (ϕ, Φ, C) at k_0 . Then either there exists $(l_1, l_2, \lambda) \in \mathcal{L}\Lambda(\phi, \Phi, C)(k_0)$ such that*

$$\lambda k_0 = \text{Min}_{k \in K} \lambda k, \quad l_2 \Phi(k_0) = \text{Max}_{c \in C} l_2 c$$

or there exist $\kappa, \delta > 0$, a positive integer N , points $k^1, \dots, k^N \in K$ and a corresponding set

$$K^* \triangleq \left\{ k_0 + \sum_{j=1}^N \omega^j (k^j - k_0) \mid \omega^j \geq 0, \sum_{j=1}^N \omega^j \leq \delta \right\}$$

such that

$$S^F(\phi(k_0), \kappa) \subset \{\phi(k) \mid k \in K^*, \Phi(k) + S^F(0, \kappa) \subset C\}.$$

As an easy corollary of Theorem 2.2, we shall derive

THEOREM 2.3. *Let k_0 yield the minimum of $\phi^0(k)$ on the set $\{k \in K \mid \phi^1(k) = 0, \Phi(k) \in C\}$, and let $\mathcal{L}\Lambda((\phi^0, \phi^1), \Phi, C)(k_0)$ be a scalar derivate container for $((\phi^0, \phi^1), \Phi, C)$ at k_0 . Then there exists $((l_0, l_1), l_2, \lambda) \in \mathcal{L}\Lambda((\phi^0, \phi^1), \Phi, C)(k_0)$ such that*

$$l_0 \geq 0, \quad \lambda k_0 = \min_{k \in K} \lambda k, \quad l_2 \Phi(k_0) = \max_{c \in C} l_2 c.$$

3. Proofs. For elements $x = (x^1, \dots, x^a)$ and $y = (y^1, \dots, y^a)$ of \mathbb{R}^a , we shall write $|x|$ for the Euclidean norm, $|x|_1$ for $\sum_{i=1}^a |x^i|$ and $x \cdot y$ for the scalar product. We use the superscript T to denote transposition. All elements of \mathbb{R}^a are treated as column vectors.

In order to prove Theorem 2.2, we first establish an auxiliary result.

THEOREM 3.1. *Let $k_0 \in K$, $\Phi(k_0) \in C$ and assume that there exist continuous functions $(\phi_i, \Phi_i): K \rightarrow \mathbb{R}^m \times \mathcal{Y}$ ($i = 1, 2, \dots$), a number $\gamma \in (0, 1]$, a positive integer N and points $k^1, \dots, k^N \in K$ such that, setting*

$$K^* \triangleq \left\{ k_0 + \sum_{j=1}^N \omega^j (k^j - k_0) \mid (\omega^1, \dots, \omega^N) \in \gamma \mathcal{T}_N \right\},$$

$$\hat{K} \triangleq \text{co} \{k_0, k^1, \dots, k^N\},$$

we have

$$(1) \quad \lim_i (\phi_i, \Phi_i) = (\phi, \Phi) \quad \text{uniformly};$$

$$(2) \quad \text{for every } i \in \{1, 2, \dots\}, \text{ the function}$$

$$\omega \rightarrow (\phi_i, \Phi_i)(k_0 + \sum_{j=1}^N \omega^j (k^j - k_0)): \gamma \mathcal{T}_N \rightarrow \mathbb{R}^m \times \mathcal{Y}$$

is continuously differentiable;

$$(3) \quad \text{for every } i \in \{1, 2, \dots\} \text{ and } k' \in K^*,$$

$$S^F(0, \gamma) \subset \{D\phi_i(k'; k - k') \mid k \in \hat{K}, D\Phi_i(k'; k - k') + S^F(0, \gamma) \subset C - \Phi(k_0)\}.$$

Then there exists $\kappa > 0$ such that

$$S^F(\phi(k_0), \kappa) \subset \{\phi(k) \mid k \in K^*, \Phi(k) + S^F(0, \kappa) \subset C\}.$$

Proof. Step 1. For $\omega = (\omega^1, \dots, \omega^N) \in \mathcal{T}_N$, we shall write

$$h(\omega) \triangleq k_0 + \sum_{j=1}^N \omega^j (k^j - k_0).$$

Since Φ is uniformly continuous on the compact set K and $\lim_i \Phi_i = \Phi$ uniformly, we

can determine $r \in (0, \gamma/2]$ such that

$$|\Phi_i(k_0) - \Phi_i(h(\omega'))| < \gamma/2 \quad (i = 1, 2, \dots, |\omega'|_1 \leq r).$$

Thus, in view of assumption (3),

$$(4) \quad S^F(0, r) \subset \{D\phi_i(k'; k - k') | k \in \hat{K}, S^F(D\Phi_i(k'; k - k'), r) \subset C - \Phi_i(k')\} \\ (k' = h(\omega'), |\omega'|_1 \leq r).$$

Let $c_1 \triangleq 2mr^{-1}N$. We shall first show that, whenever

$$(5) \quad i \in \{1, 2, \dots\}, \quad \mathbf{n} \in \mathbb{R}^m, \quad |\mathbf{n}| = 1, \quad \tilde{\omega} \in r\mathcal{T}_N, \quad \tilde{k} = h(\tilde{\omega}), \quad \tilde{r} = r - |\tilde{\omega}|_1, \quad q_1, q_2 \geq 0, \\ S^F(\Phi_i(\tilde{k}), \text{Min}(q_1, r/4)) \subset S^F(C, q_2)$$

then for every $b \in [0, c_1^{-1}\tilde{r}]$ there exist $k \in K^*$ and $\omega \in r\mathcal{T}_N$ satisfying the relations

$$(6) \quad |\omega - \tilde{\omega}|_1 \leq c_1 b, \quad k = h(\omega),$$

$$(7) \quad \phi_i(k) = \phi_i(\tilde{k}) + b\mathbf{n},$$

$$(8) \quad S^F(\Phi_i(k), \text{Min}(q_1 + b/8, r/4)) \subset S^F(C, q_2).$$

Let (5) be satisfied, and let \mathcal{B} denote the collection of all $b' \in [0, c_1^{-1}\tilde{r}]$ such that for every $b \in [0, b']$ there exist $k \in K^*$ and $\omega \in r\mathcal{T}_N$ satisfying relations (6)–(8). The number $b = 0$ belongs to \mathcal{B} and corresponds to $(k, \omega) = (\tilde{k}, \tilde{\omega})$. The set \mathcal{B} is closed because C is closed, $r\mathcal{T}_N$ compact and (ϕ_i, Φ_i) continuous. Now let $\bar{b} \in \mathcal{B}$, $0 \leq \bar{b} < c_1^{-1}\tilde{r}$, and let $(b, k, \omega) = (\bar{b}, \bar{k}, \bar{\omega})$ satisfy (6)–(8). We shall show that there exists $\alpha > 0$ such that $\bar{b} + \alpha \in \mathcal{B}$ which will prove that $\mathcal{B} = [0, c_1^{-1}\tilde{r}]$.

Since $|\bar{\omega} - \tilde{\omega}|_1 \leq c_1 \bar{b} < \tilde{r} = r - |\tilde{\omega}|_1$, we have $\bar{r} \triangleq r - |\bar{\omega}|_1 > 0$, and relation (4) is valid for each $k' = h(\omega')$ with $|\omega' - \bar{\omega}|_1 \leq \bar{r}$. We can select points $\xi_1, \dots, \xi_m \in \mathbb{R}^m$ such that

$$(9) \quad |\xi_j| = r, \quad \xi_j \cdot \xi_l = 0 \quad (j \neq l), \quad m^{-1/2}r\mathbf{n} = m^{-1} \sum_{j=1}^m \xi_j.$$

Then, by (4), there exist points $\omega_1, \dots, \omega_m \in \mathcal{T}_N$ and $k_1 = h(\omega_1), \dots, k_m = h(\omega_m)$ such that

$$(10) \quad D\phi_i(\bar{k}; k_j - \bar{k}) = \xi_j, \quad S^F(D\Phi_i(\bar{k}; k_j - \bar{k}), r) \subset C - \Phi_i(\bar{k}) \quad (j = 1, \dots, m).$$

Now let

$$(\psi, \Psi)(a) \triangleq (\phi_i, \Phi_i) \left(\bar{k} + \sum_{j=1}^m a^j (k_j - \bar{k}) \right) \quad (a \in \mathcal{T}_m),$$

and let $[\xi_1, \dots, \xi_m]$ denote the matrix with columns ξ_1, \dots, ξ_m . Then, by assumption (2), (ψ, Ψ) is continuously differentiable near $a = 0$ and, by (9) and (10),

$$(11) \quad \psi'(0) = [\xi_1, \dots, \xi_m], \quad \psi'(0)^{-1} = r^{-2} [\xi_1, \dots, \xi_m]^T, \\ |\psi'(0)| = r, \quad |\psi'(0)^{-1}| = r^{-1},$$

and there exists $\delta > 0$ such that, for $\Delta \triangleq \delta\mathcal{T}_m$, we have

$$(12) \quad (2m)^{-1/2}r\delta\mathbf{n} \in \psi'(\theta)\Delta, \quad |\psi'(\theta)^{-1}| < 2r^{-1}, \\ S^F(\Psi'(\theta)a, |a|_1 r/2) \subset |a|_1 [C - \Psi(0)] \quad (\theta, a \in \Delta).$$

For each $x \in \mathbb{R}^m$, let $s(x)$ denote the unique point in Δ that is closest to x (in the sense of the Euclidean distance), and let $\chi(x) \triangleq \psi'(s(x))$. Then χ is continuous and

nonsingular, and the differential equation

$$\dot{u}(s) = \chi(u(s))^{-1} \mathbf{n}, \quad u(0) = 0$$

has a continuously differentiable solution $u = (u^1, \dots, u^m): [0, \alpha] \rightarrow \mathbb{R}^m$ for some $\alpha > 0$. By (9) and (11), we have

$$|\dot{u}(0)| = |\psi'(0)^{-1} \mathbf{n}| = r^{-1}.$$

We may choose α sufficiently small so that (in view of the first relation in (12)),

$$(13) \quad (2r)^{-1}s \leq |u(s)| \leq 2r^{-1}s, \quad |u(s)|_1 \leq 1, \quad u(s) \in \Delta \quad (0 \leq s \leq \alpha).$$

We thus have

$$(14) \quad \psi(u(s)) = \psi(0) + \int_0^s \psi'(u(\sigma)) \dot{u}(\sigma) d\sigma = \psi(0) + s\mathbf{n} \quad (0 \leq s \leq \alpha).$$

Next we observe that, for $0 \leq s \leq \alpha$,

$$\Psi(u(s)) = \Psi(0) + \int_0^1 \Psi'(\sigma u(s)) u(s) d\sigma$$

and, by (12) and (13),

$$\Psi'(\sigma u(s)) u(s) + S^F(0, |u(s)|_1 r/2) \subset |u(s)|_1 [C - \Psi(0)] \quad (\sigma \in [0, 1]).$$

Since C is closed and convex, these last two relations imply that

$$(15) \quad \Psi(u(s)) - \Psi(0) + S^F(0, |u(s)|_1 r/2) \subset |u(s)|_1 [C - \Psi(0)].$$

We set

$$\mu \triangleq \text{Min}(q_1 + \bar{b}/8, r/4), \quad t \triangleq |u(s)|_1,$$

and recall that (\bar{b}, \bar{k}) satisfies (8); hence

$$\Psi(0) + S^F(0, \mu) \subset C + S^F(0, q_2)$$

and therefore

$$(1-t)\Psi(0) + S^F(0, (1-t)\mu) \subset (1-t)C + S^F(0, (1-t)q_2) \subset (1-t)C + S^F(0, q_2).$$

Combining this last relation with (15), we obtain

$$\begin{aligned} \Psi(u(s)) + S^F(0, tr/2 + (1-t)\mu) &\subset (1-t)\Psi(0) + tC \\ &+ S^F(0, (1-t)\mu) \subset tC + (1-t)C + S^F(0, q_2) \subset S^F(C, q_2) \end{aligned}$$

and, since $\mu \leq r/4$, this yields

$$(16) \quad \Psi(u(s)) + S^F(0, \mu + tr/4) \subset S^F(C, q_2).$$

By (13),

$$t \triangleq |u(s)|_1 \geq |u(s)| \geq (2r)^{-1}s$$

and therefore

$$\mu + tr/4 \geq \mu + s/8 \geq \text{Min}(q_1 + (\bar{b} + s)/8, r/4).$$

Thus, by (16),

$$(17) \quad S^F(\Psi(u(s)), \text{Min}(q_1 + (\bar{b} + s)/8, r/4)) \subset S^F(C, q_2).$$

If, for $b = \bar{b} + s \in [\bar{b}, \bar{b} + \alpha]$, we set

$$\omega = \bar{\omega} + \sum_{j=1}^m u^j(s)(\omega_j - \bar{\omega}), \quad k = h(\omega),$$

and observe that $\|\omega' - \omega''\|_1 \leq N$ for all $\omega', \omega'' \in \mathcal{T}_N$, then it follows from (13) that

$$\begin{aligned} \|\omega - \tilde{\omega}\|_1 &\leq \|\omega - \bar{\omega}\|_1 + \|\bar{\omega} - \tilde{\omega}\|_1 \leq N|u(s)|_1 + c_1 \bar{b} \leq 2mr^{-1}Ns + c_1 \bar{b} \\ &= c_1(\bar{b} + s) = c_1 b \end{aligned}$$

and thus k and ω satisfy relation (6). Furthermore, relations (14) and (17), both valid for $s \in [0, \alpha]$, show that

$$k = h(\omega) = \bar{k} + \sum_{j=1}^m u^j(s)(k_j - \bar{k})$$

satisfies (7) and (8). Thus $\bar{b} + \alpha \in \mathcal{B}$, and therefore $\mathcal{B} = [0, c_1^{-1}\bar{r}]$.

Step 2. Let $i \in \{1, 2, \dots\}$, $z \in \mathbb{R}^m$ be the endpoint of a polygon originating at $\phi_i(k_0)$ and of (Euclidean) length $c_1^{-1}r$, and $q_2 \triangleq d[\Phi_i(k_0), C]$. We initially set $q_1 = 0$, $\tilde{k} = k_0$, $\tilde{\omega} = 0$ and choose for \mathbf{n} a unit vector in \mathbb{R}^m directed as the first side of the polygon. Then relations (5) are satisfied. If the length of the first side is L_1 then $L_1 \leq c_1^{-1}r$ and, for $b = L_1$, there exist $\omega = \tilde{\omega}_1$ and $k = \tilde{k}_1 = h(\tilde{\omega}_1)$ satisfying relations (6)–(8). Thus $\phi_i(\tilde{k}_1)$ is the endpoint of the first side of the polygon and

$$S^F(\Phi_i(\tilde{k}_1), \text{Min}(L_1/8, r/4)) \subset S^F(C, q_2).$$

We now repeat the process for each side, choosing \mathbf{n} as the unit vector in the direction of that side, setting $q_1 = L/8$, where L is the sum of the lengths of the preceding sides, and selecting $\tilde{\omega}$ and $\tilde{k} = h(\tilde{\omega})$ so that $\phi_i(\tilde{k})$ be the endpoint of the previously considered side and

$$S^F(\Phi_i(\tilde{k}), \text{Min}(L/8, r/4)) \subset S^F(C, q_2).$$

We then conclude that there exists $\hat{k}_i \in K^*$ such that

$$(18) \quad \phi_i(\hat{k}_i) = z, \quad \Phi_i(\hat{k}_i) + S^F(0, \kappa) \subset S^F(C, d_i),$$

where $\kappa \triangleq \frac{1}{8} \text{Min}(c_1^{-1}r, 2r)$ and $d_i \triangleq d[\Phi_i(k_0), C]$.

If $z \in S^F(\phi_i(k_0), c_1^{-1}r/2)$ then, for sufficiently large i , z is the endpoint of a polygon originating at $\phi_i(k_0)$ and of length $c_1^{-1}r$, and there exists \hat{k}_i satisfying (18). We can determine a sequence $J \subset (1, 2, \dots)$ such that $(\hat{k}_i)_{i \in J}$ converges to some $\hat{k} \in K^*$, and then it follows from (18) that

$$\phi(\hat{k}) = z, \quad \Phi(\hat{k}) + S^F(0, \kappa) \subset C. \quad \text{Q.E.D.}$$

Proof of Theorem 2.2. Step 1. Assume that there exists no (l_1, l_2, λ) as described in the theorem, and observe that the convex body C contains some ball $S^F(y_0, r)$ with $r > 0$. We shall first prove that there exists $\beta > 0$ such that, for every $(f, F) \in \Lambda^\beta(\phi, \Phi)(k_0)$, there exists $\hat{k} \in K$ satisfying

$$(1) \quad f(\hat{k} - k_0) = 0, \quad \Phi(k_0) + F(\hat{k} - k_0) + S^F(0, \beta) \subset C.$$

Indeed, assume the contrary. Then there exists a sequence $((f_i, F_i))_{i \geq i_0}$, with $i_0 > 2/r$ and $(f_i, F_i) \in \Lambda^{1/i}(\phi, \Phi)(k_0)$, such that for each $i \geq i_0$ the nonempty closed convex set

$$S_i \triangleq \{0\} \times \{y \in \mathcal{Y} | \Phi(k_0) + y + S^F(0, 1/i) \subset C\} \subset \mathbb{R}^m \times \mathcal{Y}$$

has no points in common with the nonempty compact convex set $W_i \triangleq (f_i, F_i)(K - k_0)$. Thus there exist $l^i = (l_1^i, l_2^i)$ such that

$$(2) \quad |l_1^i| + |l_2^i| = 1, \quad l^i w \geq l^i s \quad (w \in W_i, s \in S_i).$$

We may assume (choosing otherwise appropriate subsequences) that there exists $\tilde{l} = (\tilde{l}_1, \tilde{l}_2) \in \mathbb{R}^m \times \mathcal{Y}^*$ such that

$$\lim_i l_1^i = \tilde{l}_1, \quad \lim_i l_2^i y = \tilde{l}_2 y \quad (y \in \mathcal{Y}).$$

Since $0 \in W_i$, relation (2) yields

$$(3) \quad l^i s = l_2^i s_2 \leq 0 \quad (s = (0, s_2) \in S_i)$$

and, since $i_0 > 2/r$ and $S^F(y_0, r) \subset C$, this implies that

$$l_2^i (y_0 - \Phi(k_0) + z) \leq 0 \quad (z \in S^F(0, r/2)).$$

Thus

$$\frac{r}{2} |l_2^i| \leq l_2^i (\Phi(k_0) - y_0)$$

which together with the first relation of (2) yields

$$\tilde{l}_2 (\Phi(k_0) - y_0) \geq \frac{r}{2} (1 - |\tilde{l}_1|).$$

This shows that $\tilde{l} = (\tilde{l}_1, \tilde{l}_2) \neq 0$. Furthermore, if $c \in C^\circ$ then $(0, c - \Phi(k_0)) \in S_i$ for all sufficiently large i and therefore, by (3), $l_2^i (c - \Phi(k_0)) \leq 0$; hence $\tilde{l}_2 c \leq \tilde{l}_2 \Phi(k_0)$. This shows that

$$(4) \quad \tilde{l}_2 \Phi(k_0) = \max_{c \in C} \tilde{l}_2 c.$$

Since the collections $\{(f_i, F_i)|_K | i \geq i_0\}$ and $\{l^i | i \geq i_0\}$ are equicontinuous and bounded, there exist $J \subset (1, 2, \dots)$ and a linear functional $\tilde{\lambda}$ on the linear hull of K such that

$$(5) \quad \lim_{i \in J} l^i \circ (f_i, F_i)(k) = \tilde{\lambda}(k) \quad (k \in K)$$

and $\tilde{\lambda}|_K$ is continuous. We may arbitrarily extend $\tilde{\lambda}$ as a linear functional to the vector space \mathcal{X} and we may assume that the sequence $((f_i, F_i))$ was chosen so that $J = (1, 2, \dots)$. Furthermore, since $\Phi(k_0) \in C$ and $C^\circ \neq \emptyset$, we can select from each set S_i a point s_i so that $\lim_i |s_i| = 0$. It follows then from (2) and (5) that

$$\tilde{\lambda}(k - k_0) \geq 0 \quad (k \in K).$$

This relation, (3) and (4) show that $(\tilde{l}_1, \tilde{l}_2, \tilde{\lambda})$ has all the properties of (l_1, l_2, λ) , thus contradicting our first assumption.

Step 2. Let β be as defined in Step 1. If there exists no $\alpha \in (0, \beta]$ such that, for each $(f, F) \in \Lambda^\alpha(\phi, \Phi)(k_0)$,

$$(6) \quad S^F(0, \alpha) \subset f(K - k_0) \subset \mathbb{R}^m$$

then there exists a sequence $((f_i, F_i))$ such that $(f_i, F_i) \in \Lambda^{1/i}(\phi, \Phi)(k_0)$ and each convex and compact set $f_i(K - k_0)$ contains a boundary point w_i with $\lim_i w_i = 0$. There exist, therefore, $l_1^i \in \mathbb{R}^m$ such that $|l_1^i| = 1$ and

$$l_1^i f_i(k - k_0) \geq l_1^i w_i \quad (k \in K).$$

The same argument as in Step 1 shows that there exists $(\tilde{l}_1, 0, \tilde{\lambda}) \in \mathcal{L}\Lambda(\phi, \Phi, C)(k_0)$ such that

$$\tilde{\lambda}(k - k_0) \geq 0 \quad (k \in K).$$

Thus $(\tilde{l}_1, 0, \tilde{\lambda})$ has the properties of (l_1, l_2, λ) again contrary to assumption. Therefore there exists $\alpha > 0$ satisfying relation (6).

Step 3. Let α be defined as in Step 2, $(f, F) \in \Lambda^\alpha(\phi, \Phi)(k_0)$ and \hat{k} be defined as in (1). Since K is compact, there exists $c' > 0$ such that

$$|F(k' - k'')| \leq c' \quad (k', k'' \in K).$$

We set $\beta' \triangleq \frac{1}{2}(c' + \alpha)^{-1}\alpha$ and observe that, if $z \in S^F(0, \alpha) \subset \mathbb{R}^m$ then there exists $\tilde{k} \in K$ such that $f(\tilde{k} - k_0) = z$. We let $\bar{k} \triangleq \beta'\tilde{k} + (1 - \beta')\hat{k}$ and observe that

$$f(\bar{k} - k_0) = f(\beta'\tilde{k} + [1 - \beta']\hat{k} - k_0) = \beta'f(\tilde{k} - k_0) = \beta'z$$

and

$$\begin{aligned} \Phi(k_0) + F(\bar{k} - k_0) + S^F(0, \alpha/2) &= \Phi(k_0) + F(\hat{k} - k_0) + \beta'F(\tilde{k} - \hat{k}) + S^F(0, \alpha/2) \\ &\subset \Phi(k_0) + F(\hat{k} - k_0) + S^F(0, \beta) \subset C. \end{aligned}$$

This shows that

$$(7) \quad \begin{aligned} S^F(0, \beta'\alpha) &\subset \{f(k - k_0) | k \in K, F(k - k_0) + S^F(0, \alpha/2) \subset C - \Phi(k_0)\} \\ &\quad [(f, F) \in \Lambda^\alpha(\phi, \Phi)(k_0)]. \end{aligned}$$

Step 4. We can determine points $b_0, \dots, b_m \in S^F(0, \beta'\alpha)$ and $\varepsilon_1, \varepsilon_2 > 0$ such that

$$S^F(0, 2\varepsilon_2) \subset \text{co}\{b'_0, \dots, b'_m\}$$

whenever $|b'_j - b_j| \leq \varepsilon_1$ ($j = 0, \dots, m$). Furthermore, by assumption (3) in the definition of the directional derivate container, all the elements of $\Lambda^\alpha(\phi, \Phi)(k_0)$ are equicontinuous when restricted to K . It follows therefore from (7) that there exists a finite subset $\{k^1, \dots, k^N\}$ of the compact set K such that, for every $(f, F) \in \Lambda^\alpha(\phi, \Phi)(k_0)$ and every $j \in \{0, \dots, m\}$, there exists $\hat{k} \in \{k^1, \dots, k^N\}$ satisfying

$$|f(\hat{k} - k_0) - b_j| < \varepsilon_1, \quad F(\hat{k} - k_0) + S^F(0, \alpha/4) \subset C - \Phi(k_0).$$

This implies that

$$S^F(0, 2\varepsilon_2) \subset \text{co}\{f(k^j - k_0) | j \in \{1, \dots, N\}, F(k^j - k_0) + S^F(0, \alpha/4) \subset C - \Phi(k_0)\}$$

and therefore, setting $\hat{K} \triangleq \text{co}\{k_0, k^1, \dots, k^N\}$,

$$(8) \quad S^F(0, 2\varepsilon_2) \subset \{f(k - k_0) | k \in \hat{K}, F(k - k_0) + S^F(0, \alpha/4) \subset C - \Phi(k_0)\}.$$

Now let $\varepsilon = \alpha$ and let δ, i^* and K^* be correspondingly defined as in assumption (4) of Definition 2.1. Because of the equicontinuity of $(f, F) \in \Lambda^\alpha(\phi, \Phi)(k_0)$ when restricted to K , we may sufficiently reduce δ so that, for all such (f, F) and for all $p', p'' \in K^*$,

$$|f(p' - p'')| + |F(p' - p'')| \leq \min(\varepsilon_2, \alpha/8).$$

Then, for every $k' \in K^*$ and $i \geq i^*$, there exists $(\bar{f}, \bar{F}) \in \Lambda^\alpha(\phi, \Phi)(k_0)$ satisfying

$$D(\phi_i, \Phi_i)(k'; k - k') = (\bar{f}, \bar{F})(k - k') = (\bar{f}, \bar{F})(k - k_0) + (\bar{f}, \bar{F})(k_0 - k') \quad (k \in K)$$

and therefore, by (8),

$$(9) \quad S^F(0, \varepsilon_2) \subset \{D\phi_i(k'; k - k') | k \in \hat{K}, D\Phi_i(k'; k - k') + S^F(0, \alpha/8) \subset C - \Phi(k_0)\}.$$

If we set $\gamma \triangleq \min(\alpha/8, \varepsilon_2, \delta, 1)$ and replace the sequence $((\phi_i, \Phi_i))$ by $((\phi_i, \Phi_i))_{i \geq i^*}$, then it follows from (9) that the assumptions of Theorem 3.1 are satisfied. The second alternative of the present theorem then follows from Theorem 3.1, with γ renamed as δ . Q.E.D.

Proof of Theorem 2.3. Let

$$\begin{aligned}\tilde{\mathcal{X}} &\triangleq \mathcal{X} \times \mathbb{R}, \quad \tilde{K} \triangleq K \times [0, 1], \quad \tilde{k} \triangleq (k, a), \quad \tilde{k}_0 \triangleq (k_0, 0), \\ (\tilde{\phi}^1, \tilde{\Phi})(k) &\triangleq (\phi^1, \Phi)(k), \quad \tilde{\phi}^0(\tilde{k}) \triangleq \phi_0(k) + a.\end{aligned}$$

Then it is clear that we can define a directional derivative container for $(\tilde{\phi}^0, \tilde{\phi}^1, \tilde{\Phi})$ at \tilde{k}_0 by

$$\Lambda^e(\tilde{\phi}^0, \tilde{\phi}^1, \tilde{\Phi})(\tilde{k}_0) \triangleq \left\{ (h^0, h^1, H) \mid (h^1, H)(x, y) = (g^1, G)(x) \text{ and } h^0(x, y) = g^0(x) + y \text{ } (x \in \mathcal{X}, y \in \mathbb{R}), (g^0, g^1, G) \in \Lambda^e(\phi^0, \phi^1, \Phi)(k_0) \right\}.$$

We may apply Theorem 2.2 to the problem in which (ϕ, Φ) , K and k_0 are replaced by $((\tilde{\phi}^0, \tilde{\phi}^1), \tilde{\Phi})$, \tilde{K} and \tilde{k}_0 , respectively. Then the second alternative of Theorem 2.2 is invalid because \tilde{k}_0 minimizes $\tilde{\phi}^0$ on the set

$$\{\tilde{k} \mid \phi^1(\tilde{k}) = 0, \tilde{\Phi}(\tilde{k}) \in C\}.$$

Therefore there exists $((l_0, l_1), l_2, \lambda) \in \mathcal{L}\Lambda((\phi^0, \phi^1), \Phi, C)(k_0)$ such that

$$\lambda k_0 = \min_{k \in K} \lambda k, \quad l_2 \Phi(k_0) = \max_{c \in C} l_2 c$$

and

$$0 = \min_{a \in [0, 1]} l_0 a \leq l_0. \quad \text{Q.E.D.}$$

REFERENCES

- [1] F. H. CLARKE, *Necessary conditions for nonsmooth problems in optimal control and the calculus of variations*, Doctoral dissertation, Univ. of Washington, Seattle, 1973.
- [2] ———, *The maximum principle under minimal hypotheses*, this Journal, 14 (1976), pp. 1078–1091.
- [3] E. I. KUGUŠEV, *The maximum principle in problems of optimal control of systems with non-smooth right-hand side*, Vestnik Moskov. Univ. Ser. I Mat. Meh., 28 (1973), no. 3, pp. 103–117. (In Russian; English summary.)
- [4] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
- [5] ———, *Necessary conditions without differentiability assumptions in optimal control*, J. Differential Equations, 18 (1975), pp. 41–62.
- [6] ———, *Controllability and necessary conditions in unilateral problems without differentiability assumptions*, this Journal, 14 (1976), pp. 546–572.
- [7] ———, *Derivative containers, inverse functions, and controllability*, Calculus of Variations and Control Theory, D. L. Russell, ed., Academic Press, New York, 1976.
- [8] ———, Chapter XI, Appendix to the Russian translation of *Optimal Control of Differential and Functional Equations*, Nauka, Moscow, 1977.
- [9] ———, *Controllability of nondifferentiable hereditary processes*, this Journal, pp. 813–831.

CONTROLLABILITY OF NONDIFFERENTIABLE HEREDITARY PROCESSES*

J. WARGA†

Abstract. We derive a “controllability-multiplier rule” alternative for control problems defined by hereditary functional-integral equations of the form

$$y(t) = \int f(t, \tau, \xi(y)(\tau), u(\tau))\mu(d\tau) \quad (t \in T)$$

and by associated “isoperimetric” and unilateral constraints. Our assumptions are weaker than customary because $f(t, \tau, \cdot, u)$ and the functions of state variables that define the constraints and the cost functional are assumed to be Lipschitz-continuous but not necessarily differentiable. The arguments are based on a controllability model of a general nondifferentiable optimization problem developed in a companion paper.

1. Introduction. We shall derive a “controllability-multiplier rule” alternative for a class of optimal control problems defined by functional-integral equations of the form

$$(1) \quad y(t) = \int f(t, \tau, \xi(y)(\tau), u(\tau))\mu(d\tau) \quad (t \in T)$$

and by associated “isoperimetric” and unilateral constraints. The problems we consider are to some extent similar to those previously studied in [6, Chap. VII] but they differ in four respects: (a) We shall consider only “ p -hereditary equations” (which can be thought of as generalizations of Volterra-like functional-integral equations to higher-dimensional domains T), and (b) we shall only study relaxed controls u while both relaxed and ordinary controls were studied in [6, Chap. VII]. On the other hand, (c) we shall investigate the case where $f(t, \tau, \cdot, u)$ is Lipschitz continuous while in [6, Chap. VII] necessary conditions were derived only for functions $f(t, \tau, \cdot, u)$ in C^1 . Finally, (d) we shall consider the controllability aspects of the problem as well as the “alternative” necessary conditions for minimum.

Our approach to these nondifferentiable hereditary processes is based on a general model of nondifferentiable optimization problems investigated in a companion paper [13]. That paper owes much, in turn, to the more specialized studies, in [7]–[11], of the controllability and optimal control of ordinary differential equations without smoothness or convexity assumptions. The latter problem was also studied, using other methods, by Kugušev [5] and Clarke [1] (see also the references listed in [3]).

There is a fairly large and growing body of research on the optimal control and controllability of functional-integral equations, especially in the form of functional-differential equations, much of it apparently inspired by applications. However, aside from the applications, the present paper is also motivated by methodological considerations. It seems to us to demonstrate the usefulness of the general controllability model developed in [13] and of the concept of a directional derivative container which underlies it. We thus view the present paper and [13], as well as the previous more specialized studies in [7]–[11], as contributions to “Nonsmooth Analysis.” The latter subject is also being investigated, by different methods and often on different terrains, by F. H. Clarke [3], [4] (see also the references listed in [3]) and by H. Halkin [14].

* Received by the editors August 18, 1977.

† Department of Mathematics, Northeastern University, Boston, Massachusetts 02115. This work was supported in part by the National Science Foundation under Grant MCS 76-06756.

2. Assumptions and results. Let T and R be compact metric spaces, μ a positive nonatomic Radon measure on T , and $R^*(t)$ ($t \in T$) nonempty closed subsets of R such that the set $\{t \in T \mid R^*(t) \cap G \neq \emptyset\}$ is μ -measurable for every open $G \subset R$. We denote by $\text{frm}(R)$ the collection of the Radon measures on R , by $\text{frm}^+(R)$ resp. $\text{rpm}(R)$ the set of those that are nonnegative resp. probability measures, and by \mathcal{S}^* the set of *relaxed controls*, that is, of functions $\sigma: T \rightarrow \text{rpm}(R)$ such that $\sigma(t)(R^*(t)) = 1$ μ -a.e. and $t \mapsto \int \phi(r)\sigma(t)(dr)$ is μ -measurable for every continuous $\phi: R \rightarrow \mathbb{R}$. For any $s \in \text{frm}(R)$ and continuous $\phi: R \rightarrow \mathbb{R}^a$, we write $\phi(s) \triangleq \int \phi(r)s(dr)$, the symbol \triangleq meaning "equal by definition".

We assume given a point $t_1 \in T$, a convex body $A \subset \mathbb{R}^{m_2}$ (i.e. A is closed, convex and its interior A° is nonempty), and functions

$$f = (f^1, \dots, f^n): T \times T \times V \times R \rightarrow \mathbb{R}^n, \quad (h^0, h^1): W \rightarrow \mathbb{R} \times \mathbb{R}^m, \quad h^2: T^h \times W \rightarrow \mathbb{R}^{m_2},$$

$$\xi: C(T, W) \rightarrow L^\infty(\mu, V),$$

where $k, n, m, m_2 \in \mathbb{N} \triangleq \{1, 2, \dots\}$, T^h is a compact subset of T , V and W are open subsets of \mathbb{R}^k and \mathbb{R}^n , respectively, $C(T, W)$ is the subset of the Banach space $C(T, \mathbb{R}^n)$ with elements whose values are in W , and $L^\infty(\mu, V)$ is the subset of the Banach space $L^\infty(\mu, \mathbb{R}^k)$ whose elements are functions with values μ -a.e. in V .

We shall say that ξ is p -hereditary, where $p: T \rightarrow [0, 1]$ is a given function, if

$$\xi(y_1)(\tau) = \xi(y_2)(\tau) \quad \text{for } \mu\text{-a.a. } \tau \quad \text{with } p(\tau) \leq \alpha$$

whenever $y_1(\tau) = y_2(\tau)$ for $p(\tau) \leq \alpha$. Our arguments will be based on

ASSUMPTION 2.1. There exist $c_1 \in \mathbb{R}$, $\psi: T \times T \times T \rightarrow \mathbb{R}$ and a μ -measurable $p: T \rightarrow [0, 1]$ such that, for all $(\hat{t}, t, t', \tau, v, r) \in T^h \times T \times T \times T \times V \times R$,

- (1) the function $\alpha \mapsto \mu(p^{-1}([0, \alpha])): [0, 1] \rightarrow \mathbb{R}$ is continuous;
- (2) ξ has a continuous (Fréchet) derivative ξ' ,

$$|\xi'(y)| \leq c_1 \quad (y \in C(T, W)),$$

and ξ is p -hereditary;

- (3) $f(t, \tau, v, r) = 0$ if $p(\tau) > p(t)$;
- (4) the functions $f(t, \tau, \cdot, \cdot)$ and h^2 are continuous, $f(t, \tau, \cdot, r)$, h^0 , h^1 and $h^2(\hat{t}, \cdot)$ Lipschitz-continuous with a common bound and Lipschitz constant c_1 , and $f(\cdot, \cdot, v, r)$ Borel measurable;

$$(5) \quad \lim_{t' \rightarrow t} \int \psi(t, t', \tau) \mu(d\tau) = 0;$$

$$(6) \quad |f(t', \tau, v, r) - f(t, \tau, v, r)| \leq \psi(t, t', \tau),$$

$$|v_1 - v_2|^{-1} |[f(t', \tau, v_1, r) - f(t', \tau, v_2, r)] - [f(t, \tau, v_1, r) - f(t, \tau, v_2, r)]| \leq \psi(t, t', \tau)$$

$$(v_1, v_2 \in V, v_1 \neq v_2)$$

If E and F are real Banach spaces, $B(E, F)$ will represent the real Banach space of continuous linear operators on E to F , with the usual norm. In particular, with each \mathbb{R}^a being treated as a euclidean space, $B(\mathbb{R}^a, \mathbb{R}^b)$ will be identified with the space of real $b \times a$ matrices with the corresponding norm. We shall require the concept of a *derivate* container [7], [10], [11] for h^1 , (h^0, h^1) and h^2 as defined in [10].

DEFINITION 2.2. A bounded collection

$$\{\Lambda^\varepsilon h^2(t, v) \mid \varepsilon > 0, (t, v) \in T^h \times V\}$$

of nonempty closed subsets of $B(\mathbb{R}^n, \mathbb{R}^{m_2})$, also denoted by $\Lambda^\varepsilon h^2$, is a *derivate*

container for h^2 (with respect to v) if

$$\Lambda^\varepsilon h^2(t, v) \subset \Lambda^{\varepsilon'} h^2(t, v) \quad (\varepsilon' > \varepsilon)$$

and for every compact $V^* \subset V$ there exist a neighborhood \tilde{V} of V^* in V and a sequence of functions $h_i^2: T^h \times \tilde{V} \rightarrow \mathbb{R}^{m_2} (i \in \mathbb{N})$ such that each h_i^2 has a partial derivative $\mathcal{D}_2 h_i^2$ (with respect to the second argument of h^2), both h_i^2 and $\mathcal{D}_2 h_i^2$ are continuous, $\lim_i h_i^2 = h^2$ uniformly on $T^h \times V^*$, and for every $\varepsilon > 0$ there exist $i(\varepsilon, V^*)$ and $\delta(\varepsilon, V^*) > 0$ such that

$$\mathcal{D}_2 h_i^2(t, v) \in \Lambda^\varepsilon h^2(t, \tilde{v}) \quad [i \geq i(\varepsilon, V^*), (t, \tilde{v}) \in T^h \times V^*, |v - \tilde{v}| \leq \delta(\varepsilon, V^*)].$$

We write

$$\Lambda h^2(t, v) \triangleq \bigcap_{\varepsilon > 0} \Lambda^\varepsilon h^2(t, v),$$

and define derivate containers for (h^0, h^1) or h^1 similarly, dropping all references to the dependence on t . We shall also use a special kind of a derivate container for $f(t, \tau, \cdot, r)$, defined, as in [10, pp. 16, 19], by

$$\begin{aligned} \partial^\varepsilon f(t, \tau, v, r) &\triangleq \overline{\text{co}} \{ \mathcal{D}_3 f(t, \tau, \tilde{v}, r) | \tilde{v} \in V, |\tilde{v} - v| \leq \varepsilon, \mathcal{D}_3 f(t, \tau, \tilde{v}, r) \text{ exists} \}, \\ \partial f(t, \tau, v, r) &= \bigcap_{\varepsilon > 0} \partial^\varepsilon f(t, \tau, v, r), \end{aligned}$$

where $\overline{\text{co}}$ denotes the convex closure. (The set $\partial f(t, \tau, v, r)$ coincides with Clarke's generalized Jacobian of $f(t, \tau, \cdot, r)$ [2]).

For any $\sigma \in \mathcal{S}^*$, we shall denote by $y(\sigma)$ the unique solution of the equation

$$y(t) = \int f(t, \tau, \xi(y)(\tau), \sigma(\tau)) \mu(d\tau) \quad (t \in T)$$

if such a solution exists. Our results will be formulated in terms of the concept of a *relaxed extremal*.

DEFINITION 2.3. Let $\sigma_0 \in \mathcal{S}^*$ be such that $y(\sigma_0)$ exists, let $\Lambda^\varepsilon h^1$ and $\Lambda^\varepsilon h^2$ be derivate containers for h^1 and h^2 (with respect to v),

$$\begin{aligned} \Omega &\triangleq (f, \xi, h^1, \Lambda^\varepsilon h^1, h^2, \Lambda^\varepsilon h^2, A), \\ \Gamma(t, \tau) &\triangleq \partial f(t, \tau, \xi(y(\sigma_0))(\tau), \sigma_0(\tau)), \end{aligned}$$

and let \mathcal{G} denote the collection of all $g: T \times T \rightarrow B(\mathbb{R}^k, \mathbb{R}^n)$ such that $g(t, \cdot)$ is μ -measurable,

$$\begin{aligned} g(t, \tau) &\in \Gamma(t, \tau), \quad |g(t', \tau) - g(t, \tau)| \leq \psi(t, t', \tau) \\ &\quad (t, t' \in T, \mu\text{-a.a. } \tau \in T). \end{aligned}$$

We say that σ_0 is a *relaxed extremal relative to Ω* if

$$(1) \quad h^2(t, y(\sigma_0)(t)) \in A \quad (t \in T^h)$$

and there exist

$$\begin{aligned} g &\in \mathcal{G}, \quad H_1 \in \Lambda h^1(y(\sigma_0)(t_1)), \quad H_2: T^h \rightarrow B(\mathbb{R}^n, \mathbb{R}) \\ l_1 &\in \mathbb{R}^m, \quad \omega \in \text{frm}^+(T^h), \quad \zeta_1, \dots, \zeta_n \in \text{frm}(T) \end{aligned}$$

such that H_2 is bounded and Borel measurable,

$$(2) \quad H_2(t) \in \overline{\text{co}} \left\{ L_1 L_2 |L_1 h^2(t, y(\sigma_0)(t)) = \max_{a \in A} L_1 a, \right. \\ \left. |L_1| \geq 1, L_2 \in \bigcap_{\varepsilon > 0} \overline{\text{co}} \bigcup_{d(\tau, t) \leq \varepsilon} \Lambda^\varepsilon h^2(\tau, y(\sigma_0)(\tau)) \right\} \quad \omega\text{-a.e.},$$

$$(3) \quad |l_1| + \omega(T^h) > 0, \quad \omega(\{t \in T^h | h^2(t, y(\sigma_0)(t)) \in A^\circ\}) = 0, \\ \sum_{j=1}^n \Delta y^j(t) \zeta_j(dt) = l_1^T H_1 \mathcal{R}(\Delta y)(t_1) + \int H_2(t) \mathcal{R}(\Delta y)(t) \omega(dt) \\ [\Delta y = (\Delta y^1, \dots, \Delta y^n) \in C(T, \mathbb{R}^n)],$$

where $\mathcal{R}(\Delta y) \in C(T, \mathbb{R}^n)$ is the (unique) solution Δx of the equation

$$(4) \quad \Delta x(t) = \int g(t, \tau)(\xi'(y(\sigma_0))\Delta x)(\tau) \mu(d\tau) + \Delta y(t) \quad (t \in T^h), \\ \sum_{j=1}^n f^j(t, \tau, \xi(y(\sigma_0))(\tau), \sigma_0(\tau)) \zeta_j(dt) \\ = \min_{r \in R^*(\tau)} \sum_{j=1}^n f^j(t, \tau, \xi(y(\sigma_0))(\tau), r) \zeta_j(dt) \quad \text{for } \mu\text{-a.a. } \tau \in T.$$

We can now state our basic results. The symbol $S(x, \alpha)$ ($S^F(x, \alpha)$) denotes the open (closed) ball with center x and radius α in an appropriate metric space, while $S(B, \alpha)$, $S^F(B, \alpha)$ represent the open respectively closed α -neighborhoods of B .

THEOREM 2.4. *Let $\sigma_0 \in \mathcal{S}^*$ be such that $y(\sigma_0)$ exists and $h^2(t, y(\sigma_0)(t)) \in A$ ($t \in T^h$), and let Ω be as in Definition 2.3. Then either σ_0 is a relaxed extremal relative to Ω or there exist $\kappa, \delta > 0$, $N \in \mathbb{N}$, points $\sigma^1, \sigma^2, \dots, \sigma^N \in \mathcal{S}^*$, and a corresponding set*

$$\mathcal{S}^* \triangleq \left\{ \sigma_0 + \sum_{j=1}^N \omega^j(\sigma^j - \sigma_0) \mid \omega^j \geq 0, \sum_{j=1}^N \omega^j \leq \delta \right\}$$

such that

$$S^F(h^1(y(\sigma_0)(t_1)), \kappa) \subset \left\{ h^1(y(\sigma)(t_1)) \mid \sigma \in \mathcal{S}^*, h^2(t, y(\sigma)(t)) + S^F(0, \kappa) \subset A \quad (t \in T^h) \right\}.$$

THEOREM 2.5. *Let σ_0 yield the minimum of $h^0(y(\sigma)(t_1))$ on the set*

$$\{\sigma \in \mathcal{S}^* \mid h^1(y(\sigma)(t_1)) = 0, h^2(t, y(\sigma)(t)) \in A \quad (t \in T^h)\},$$

and let $\tilde{\Omega}$ be defined as Ω in Definition 2.3 except that $h^1, \Lambda^\varepsilon h^1$ are replaced by (h^0, h^1) , $\Lambda^\varepsilon(h^0, h^1)$, where the latter is a derivate container for (h^0, h^1) . Then σ_0 is a relaxed extremal relative to $\tilde{\Omega}$ with $(l_0, l_1) \in \mathbb{R} \times \mathbb{R}^m$ replacing l_1 and with $l_0 \geq 0$.

Remark 1. In the special case where $\mathcal{D}h^0, \mathcal{D}h^1, \mathcal{D}_2 h^2$ and $\mathcal{D}_3 f(t, \tau, \cdot, \cdot)$ exist and are continuous, Theorem 2.5 yields the same results as [6, VII.3.2, p. 421].

Remark 2. If T is an interval $[t_0, t_1] \subset \mathbb{R}$ then the choice of the cost functional $h^0(y(\sigma)(t_1))$ and of the “isoperimetric” constraint $h^1(y(\sigma)(t_1)) = 0$ appears quite natural. However, in other cases, one would expect a cost functional and an “isoperimetric” constraint of the form $h^0(z), h^1(z) = 0$, where

$$z = \int \gamma(\tau, \xi(y(\sigma))(\tau), \sigma(\tau)) \mu(d\tau) \in \mathbb{R}^a.$$

The problem considered in Theorems 2.4 and 2.5 will be of this new type if the function $p: T \rightarrow [0, 1]$ achieves its maximum at some point t_1 . We then set

$$\tilde{f}(t, \tau, v, r) = \begin{cases} \gamma(\tau, v, r) & [p(\tau) \leq p(t)] \\ 0 & [p(\tau) > p(t)], \end{cases}$$

$$\hat{f} = (f, \tilde{f}), \quad \hat{y} = (y, \tilde{y}), \quad \hat{h}^i(w, \tilde{w}) = h^i(\tilde{w}) \quad (i = 0, 1),$$

and consider our original problem, with f, y, h^0, h^1 replaced by $\hat{f}, \hat{y}, \hat{h}^0, \hat{h}^1$.

3. Auxiliary constructions and definitions.

3.1. The functions h_i^1, h_i^2 and f_i . Our main purpose is to prove Theorem 2.4, from which Theorem 2.5 will follow as a corollary. We shall assume henceforth (unless otherwise specified) that σ_0 and Ω are as described in Theorem 2.4. We choose fixed compact sets V^*, W^* such that $V^* \subset V, W^* \subset W$,

$$y(\sigma_0)(t) \in (W^*)^\circ \quad (t \in T), \quad \xi(y(\sigma_0))(t) \in (V^*)^\circ \quad \mu\text{-a.e.}$$

and select $\alpha_0 > 0$ such that $S^F(V^*, 3\alpha_0) \subset V$. As in our study of controlled ordinary differential equations [7]–[11], our present arguments are based on the construction of “nice” approximations to f, h^1, h^2 on $T \times T \times \tilde{V} \times R, \tilde{W}, T^h \times \tilde{W}$, respectively, where \tilde{V}, \tilde{W} are appropriate open neighborhoods of V^*, W^* in V, W . Such approximations (h_i^1) and (h_i^2) to h^1 and h^2 are inherent in the definition and choice of the derivate containers $\Lambda^\varepsilon h^1$ and $\Lambda^\varepsilon h^2$. The approximations (f_i) to f will be defined (as in [10, pp. 23, 30]) to fit our choice of $\partial f(t, \tau, v, r)$ as the derivate container for f (with respect to v).

Specifically, let $\tilde{V} \triangleq S(V^*, \alpha_0)$, let $q_i: \mathbb{R}^k \rightarrow \mathbb{R}$ ($i \in \mathbb{N}$) be a nonnegative C^1 function vanishing outside the ball $S_i \triangleq S^F(0, \alpha_0/i)$ and such that $\int_{S_i} q_i(v) \mathbf{m}(dv) = 1$, where \mathbf{m} denotes the Lebesgue measure on \mathbb{R}^k , and let

$$f_i(t, \tau, v, r) \triangleq \int_{S_i} f(t, \tau, v - w, r) q_i(w) \mathbf{m}(dw) \quad (t, \tau \in T, v \in \tilde{V}, r \in R).$$

By a known theorem of H. Rademacher, the Lipschitz continuous function $f(t, \tau, \cdot, r)$ is differentiable \mathbf{m} -a.e. We can verify (as in [10]) that, for each $i \in \mathbb{N}$ and $(t, \tau, v, r) \in T \times T \times \tilde{V} \times R$, $\mathcal{D}_3 f_i(t, \tau, v, r)$ exists, $f_i(\cdot, \cdot, v, r)$ is Borel measurable, $f_i(t, \tau, \cdot, \cdot)$ and $\mathcal{D}_3 f_i(t, \tau, \cdot, \cdot)$ are continuous,

$$\begin{aligned} \mathcal{D}_3 f_i(t, \tau, v, r) &= \int_{S_i} f(t, \tau, v - w, r) q_i'(w) \mathbf{m}(dw) \\ &= \int_{S_i} \mathcal{D}_3 f(t, \tau, v - w, r) q_i(w) \mathbf{m}(dw), \end{aligned}$$

$$|f_i(t, \tau, v, r)| \leq c_1, \quad |\mathcal{D}_3 f_i(t, \tau, v, r)| \leq c_1,$$

and for every $\varepsilon > 0$ there exist $i(\varepsilon)$ and $\delta(\varepsilon) > 0$ such that

$$(1) \quad \mathcal{D}_3 f_i(t, \tau, v, r) \in \partial^\varepsilon f(t, \tau, v, r) \quad [t, \tau \in T, \tilde{v} \in \tilde{V}, r \in R, i \geq i(\varepsilon), |v - \tilde{v}| \leq \delta(\varepsilon)].$$

Furthermore, by Assumption 2.1(6),

$$\begin{aligned} (2) \quad & \sup_{v, r} |f_i(t', \tau, v, r) - f_i(t, \tau, v, r)| \\ & \leq \sup_{v, r} |f(t', \tau, v, r) - f(t, \tau, v, r)| \leq \psi(t', \tau) \end{aligned}$$

and

$$\begin{aligned}
 & \sup_{v,r} |\mathcal{D}_3 f_i(t', \tau, v, r) - \mathcal{D}_3 f_i(t, \tau, v, r)| \\
 (3) \quad &= \sup_{v,r} \int_{S_i} |\mathcal{D}_3 f(t', \tau, v - w, r) - \mathcal{D}_3 f(t, \tau, v - w, r)| q_i(\omega) \mathbf{m}(d\omega) \\
 &\leq \psi(t, t', \tau).
 \end{aligned}$$

For a separable metric space S and a separable Banach space \mathcal{X} , we define $\mathcal{B}(\mu, S; \mathcal{X})$ as the real normed vector space of (μ -equivalence classes of) functions $\phi: T \times S \rightarrow \mathcal{X}$ such that $\phi(\cdot, s)$ is μ -measurable, $\phi(t, \cdot)$ continuous and $t \rightarrow |\phi(t, \cdot)|_{\sup}$ μ -integrable, with the norm $\phi \rightarrow \int |\phi(t, \cdot)|_{\sup} \mu(dt)$ [6, I.5.25, p. 135]. We have shown that the functions

$$t \rightarrow f_i(t, \cdot, \cdot, \cdot) \quad \text{and} \quad t \rightarrow \mathcal{D}_3 f_i(t, \cdot, \cdot, \cdot) \quad (i \in \mathbb{N})$$

are elements of

$$C(T, \mathcal{B}(\mu, V^* \times R; \mathbb{R}^n)) \quad \text{and} \quad C(T, \mathcal{B}(\mu, V^* \times R, B(\mathbb{R}^k, \mathbb{R}^n))),$$

respectively.

3.2. The normed vector space \mathcal{N} . The space \mathcal{S}^* of relaxed controls is defined [6, Chap. IV] as a convex and compact subset of a real normed vector space \mathcal{N} in the following manner: By a variant of the Dunford–Pettis theorem [6, IV.1.8, p. 268], the vector space $L^1(\mu, C(R))^*$ (where $C(R) \triangleq C(R, \mathbb{R})$ and $*$ denotes the topological dual) is isomorphic to the space \mathcal{N} of functions $\nu: T \rightarrow \text{frm}(R)$ such that $t \rightarrow \int \phi(r) \nu(t)(dr)$ is μ -measurable for every continuous $\phi: R \rightarrow \mathbb{R}$ and μ -ess sup $|\nu(t)|(R) < \infty$. This space \mathcal{N} is endowed with “weak” norm $|\cdot|_w$ [6, p. 272] such that any sequence (ν_j) in \mathcal{N} , with $|\nu_j(t)|(R) \leq 1$ μ -a.e., converges in $|\cdot|_w$ to ν if and only if

$$\lim_j \int \mu(dt) \int \phi(t, r) [\nu_j(t) - \nu(t)](dr) = 0$$

for every $\phi \in L^1(\mu, C(R))$.

4. Auxiliary lemmas.

LEMMA 4.1. For all $y, y_1, y_2 \in C(T, W)$, $\xi'(y)$ is p -hereditary and

$$|\xi(y_1)(t) - \xi(y_2)(t)| \leq nc_1 \sup_{p(\tau) \leq p(t)} |y_1(\tau) - y_2(\tau)| \quad \text{for } \mu\text{-a.a. } t \in T.$$

Proof. Let $y \in C(T, W)$, $\alpha \in [0, 1]$, $x \in C(T, \mathbb{R}^n)$ and $x(\tau) = 0$ ($p(\tau) \leq \alpha$). Since ξ is differentiable at y , we have

$$\xi'(y)x = \lim_{\beta \rightarrow 0} \beta^{-1} [\xi(y + \beta x) - \xi(y)]$$

and therefore

$$(\xi'(y)x)(\tau) = \lim_{\beta \rightarrow 0} \beta^{-1} [\xi(y + \beta x)(\tau) - \xi(y)(\tau)] = 0$$

for μ -a.a. $\tau \in T$ with $p(\tau) \leq \alpha$, thus showing that $\xi'(y)$ is p -hereditary.

Now let $y_1, y_2 \in C(T, W)$, $y_i \triangleq (y_i^1, \dots, y_i^n)$ and $\alpha \in [0, 1]$. It is easily seen that we can construct functions $z_1, z_2 \in C(T, W)$ such that

$$z_i(t) = y_i(t) \quad (i = 1, 2, p(t) \leq \alpha), \quad \sup_{t \in T} |z_1^j(t) - z_2^j(t)| = \sup_{p(t) \leq \alpha} |z_1^j(t) - z_2^j(t)|.$$

By 2.1(2), $\sup \{|\xi'(y)| \mid y \in C(T, W)\} \leq c_1$ and therefore, by the mean value theorem,

$$\begin{aligned} |\xi(y_1)(t) - \xi(y_2)(t)| &\leq |\xi(z_1) - \xi(z_2)| \leq c_1 |z_1 - z_2| \\ &\leq n \sup \{|y_1^j(\tau) - y_2^j(\tau)| \mid j = 1, \dots, n, p(\tau) \leq \alpha\} \end{aligned}$$

for μ -a.a. t with $p(t) \leq \alpha$. Q.E.D.

LEMMA 4.2. For each $\varepsilon > 0$, let

$$\Gamma^\varepsilon(t, \tau) \triangleq S^F(\partial^\varepsilon f(t, \tau, \xi(y(\sigma_0))(\tau), \sigma_0(\tau)), \varepsilon),$$

and let \mathcal{G}^ε be the collection of all $g: T \times T \rightarrow B(\mathbb{R}^k, \mathbb{R}^n)$ such that $g(t, \cdot)$ is μ -measurable,

$$\begin{aligned} g(t, \tau) &\in \Gamma^\varepsilon(t, \tau), \\ |g(t', \tau) - g(t, \tau)| &\leq \psi(t, t', \tau) \quad (t, t' \in T, \mu\text{-a.a. } \tau \in T). \end{aligned}$$

Then, for each $g \in \mathcal{G}^\varepsilon$, the relation

$$G_g(z)(t) = \int g(t, \tau) z(\tau) \mu(d\tau) \quad (z \in L^\infty(\mu, \mathbb{R}^k), t \in T)$$

defines an element $G_g \in B(L^\infty(\mu, \mathbb{R}^k), C(T, \mathbb{R}^n))$, and for every sequence (g_i) in \mathcal{G}^ε there exist $g_\infty \in \mathcal{G}^\varepsilon$ and $J \subset (1, 2, \dots)$ such that

$$\lim_i G_{g_i}(z) = G_{g_\infty}(z) \quad (z \in L^\infty(\mu, \mathbb{R}^k)).$$

Furthermore, there exists $\varepsilon_0 > 0$ such that, for each $g \in \mathcal{G}^{\varepsilon_0}$ and $y \in C(T, W)$ with $|y - y(\sigma_0)| \leq \varepsilon_0$, the mapping $I - G_g \xi'(y): C(T, \mathbb{R}^n) \rightarrow C(T, \mathbb{R}^n)$ is a homeomorphism and $\{[I - G_g \xi'(y)]^{-1} \mid g \in \mathcal{G}^{\varepsilon_0}, |y - y(\sigma_0)| \leq \varepsilon_0\}$ is a bounded subset of $B(C(T, \mathbb{R}^n), C(T, \mathbb{R}^n))$.

Proof. Step 1. Let $|\cdot|_1, |\cdot|_\infty$ denote the norms of $L^1(\mu, \mathbb{R}^k), L^\infty(\mu, \mathbb{R}^k)$, respectively. Let $L^{\infty(1)}$ denote the vector space $L^\infty(\mu, \mathbb{R}^k)$ endowed with the norm $|\cdot|_1$. We observe that, for every $z \in L^\infty(\mu, \mathbb{R}^k)$, we have $|z|_1 \leq \mu(T)|z|_\infty$.

For $t, t' \in T, z \in L^\infty(\mu, \mathbb{R}^k), \varepsilon > 0$ and $g \in \mathcal{G}^\varepsilon$, we have

$$\begin{aligned} |G_g(z)(t') - G_g(z)(t)| &\leq \int |g(t', \tau) - g(t, \tau)| |z(\tau)| \mu(d\tau) \\ &\leq |z|_\infty \int \psi(t, t', \tau) \mu(d\tau) \end{aligned}$$

and

$$\begin{aligned} \sup_{t \in T} |G_g(z)(t)| &\leq \sup_{t \in T} \int |g(t, \tau)| |z(\tau)| \mu(d\tau) \leq (c_1 + \varepsilon) |z|_1 \\ &\leq (c_1 + \varepsilon) |z|_\infty \mu(T). \end{aligned}$$

This shows that, for each $\varepsilon > 0$, the family

$$P_\varepsilon \triangleq \{G_g(z) \mid g \in \mathcal{G}^\varepsilon, z \in L^\infty(\mu, \mathbb{R}^k), |z|_\infty \leq 1\}$$

is an equicontinuous and bounded subset of $C(T, \mathbb{R}^n)$, and the family $\{G_g \mid g \in \mathcal{G}^\varepsilon\}$ is a bounded subset of both $B(L^{\infty(1)}, C(T, \mathbb{R}^n))$ and $B(L^\infty(\mu, \mathbb{R}^k), C(T, \mathbb{R}^n))$.

Now let $\varepsilon > 0, g_i \in \mathcal{G}^\varepsilon (i \in \mathbb{N})$ and $G_i \triangleq G_{g_i}$, and let S be the set $\{z \in L^\infty(\mu, \mathbb{R}^k) \mid |z|_\infty \leq 1\}$ with the relative topology of $L^1(\mu, \mathbb{R}^k)$. If we restrict each G_i to the separable set S , then these restrictions form a subset of $C(S, C(T, \mathbb{R}^n))$ whose elements have a common Lipschitz constant and whose values are contained in the

compact set \bar{P}_e . Thus, by a variant of the Arzèla–Ascoli theorem [6, I.2.18, p. 25], there exist $J \subset (1, 2, \dots)$ and $G_\infty \in C(S, C(T, \mathbb{R}^n))$ such that, for each $z \in S$,

$$(1) \quad \lim_{i \in J} G_i(z) = G_\infty(z).$$

If we extend G_∞ to all of $L^{\infty(1)}$ by setting

$$G_\infty(z) = |z|_\infty G_\infty(z/|z|_\infty) \quad \text{for } |z|_\infty > 1,$$

then it is clear that G_∞ is linear and bounded and relation (1) remains valid for all $z \in L^{\infty(1)}$

For each $t \in T$, the function $z \rightarrow G_\infty(z)(t): L^{\infty(1)} \rightarrow \mathbb{R}^n$ is continuous and linear and thus there exists $g_\infty(t, \cdot) \in L^\infty(\mu, B(\mathbb{R}^k, \mathbb{R}^n))$ such that

$$(2) \quad G_\infty(z)(t) = \int g_\infty(t, \tau) z(\tau) \mu(d\tau) \quad [z \in L^\infty(\mu, \mathbb{R}^k)].$$

Relations (1) and (2) imply that

$$\begin{aligned} & \left| \int [g_\infty(t', \tau) - g_\infty(t, \tau)] z(\tau) \mu(d\tau) \right| \\ &= \lim_i \left| \int [g_i(t', \tau) - g_i(t, \tau)] z(\tau) \mu(d\tau) \right| \leq \int \psi(t, t', \tau) |z(\tau)| \mu(d\tau) \\ & \quad [z \in L^\infty(\mu, \mathbb{R}^k)] \end{aligned}$$

from which we deduce that

$$(3) \quad |g_\infty(t', \tau) - g_\infty(t, \tau)| \leq \psi(t, t', \tau) \quad (t, t' \in T, \mu\text{-a.a. } \tau \in T).$$

Now let $\bar{d}(\cdot, \cdot)$ denote the distance in $B(\mathbb{R}^k, \mathbb{R}^n)$ when it is identified with the euclidean space \mathbb{R}^{kn} , and let \odot denote the corresponding scalar product and $\bar{d}[Q, q]$ the corresponding distance of a point q to a set Q . We shall complete the proof of the first part of the lemma by showing that $g_\infty(t, \tau) \in \Gamma^e(t, \tau)$ for all $t \in T$ and $\tau \in T'_i$, where $\mu(T \sim T'_i) = 0$. Indeed, otherwise there exist some $\bar{t} \in T$, $\beta > 0$ and $E \subset T$ such that $g_i(\bar{t}, \tau) \in \Gamma^e(\bar{t}, \tau)$ ($i \in \mathbb{N}$, $\tau \in E$),

$$\mu(E) > 0, \quad e(\tau) \triangleq \bar{d}[\Gamma^e(\bar{t}, \tau), g_\infty(\bar{t}, \tau)] > \beta \quad (\tau \in E).$$

For each $\tau \in T$, let $s(\tau)$ be the unique point in the compact convex set $\Gamma^e(\bar{t}, \tau)$ that minimizes the \bar{d} -distance to $g_\infty(\bar{t}, \tau)$. The function $\tau \rightarrow e(\tau): T \rightarrow \mathbb{R}$ is μ -measurable because $\Gamma^e(\bar{t}, \cdot)$ and $g_\infty(\bar{t}, \cdot)$ are μ -measurable and $(Q, q) \rightarrow \bar{d}[Q, q]$ continuous (with the topology in the space of nonempty compact subsets of \mathbb{R}^{kn} defined by the Hausdorff metric). Since the function $\tau \rightarrow s(\tau)$ satisfies the relation

$$(4) \quad \bar{d}(s(\tau), g_\infty(\bar{t}, \tau)) = e(\tau) \quad (\tau \in T),$$

it follows from the Filippov–Castaing theorem [6, I.7.10, p. 153] that there exists a μ -measurable $\tau \rightarrow \bar{s}(\tau)$ satisfying (4) with $s(\tau)$ replaced by $\bar{s}(\tau)$. This shows that $\tau \rightarrow s(\tau)$ is μ -measurable because, for every $\tau \in T$, $s(\tau)$ is the unique point satisfying relation (4).

Since $\Gamma^e(\bar{t}, \tau)$ is convex and

$$g_i(\bar{t}, \tau) \in \Gamma^e(\bar{t}, \tau) \quad (i \in \mathbb{N}, \tau \in E),$$

we have

$$\begin{aligned} (5) \quad & [g_i(\bar{t}, \tau) - g_\infty(\bar{t}, \tau)] \odot [s(\tau) - g_\infty(\bar{t}, \tau)] \\ & \geq [s(\tau) - g_\infty(\bar{t}, \tau)] \odot [s(\tau) - g_\infty(\bar{t}, \tau)] > \beta_2 \quad (i \in \mathbb{N}, \tau \in E). \end{aligned}$$

On the other hand, relations (1) and (2) imply that

$$\lim_{i \in J} \int [g_i(\bar{t}, \tau) - g_\infty(\bar{t}, \tau)] z(\tau) \mu(d\tau) = 0$$

for every $z \in L^\infty(\mu, \mathbb{R}^k)$, which contradicts (5).

Step 2. Let $\varepsilon > 0$, $g \in \mathcal{G}^\varepsilon$ and $y \in C(T, W)$. We have shown in Step 1 that G_g is a compact operator in $B(L^\infty(\mu, \mathbb{R}^k), C(T, \mathbb{R}^n))$ and it follows therefore that $G_g \xi'(y)$ is a compact operator in $B(C(T, \mathbb{R}^n), C(T, \mathbb{R}^n))$. Furthermore, by Lemma 4.1, $\xi'(y)$ is p -hereditary, and we have $g(t, \tau) = 0$ if $p(\tau) > p(t)$ because $f(t, \tau, v, r) = 0$ if $p(\tau) > p(t)$. It follows now from [6, II.5.6, p. 210] that $I - G_g \xi'(y)$ is one-to-one and therefore, by Riesz's theorem, it is a linear homeomorphism of $C(T, \mathbb{R}^n)$ onto itself. Thus $[I - G_g \xi'(y)]^{-1}$ exists and is bounded.

If the last assertion of the lemma is invalid then there exist sequences (ε_i) , (g_i) , (v_i) , (x_i) and (y_i) such that (ε_i) decreases to 0, $g_i \in \mathcal{G}^{\varepsilon_i}$, $v_i, x_i, y_i \in C(T, \mathbb{R}^n)$, $|y_i - y(\sigma_0)| \leq \varepsilon_i$, $\lim_i x_i = 0$, $|v_i| = 1$ and

$$(6) \quad v_i = G_{g_i} \xi'(y_i) v_i + x_i \quad (i \in \mathbb{N}).$$

The set $\{\xi'(y_i) v_i | i \in \mathbb{N}\}$ is a bounded subset of $L^\infty(\mu, \mathbb{R}^k)$ and therefore

$$\{G_{g_i} \xi'(y_i) v_i | i \in \mathbb{N}\} \subset \alpha P_{\varepsilon_1}$$

for some $\alpha > 0$, where P_{ε_1} is as defined in Step 1. Since P_{ε_1} is an equicontinuous and bounded subset of $C(T, \mathbb{R}^n)$ and we have proven the first part of the lemma, we may assume that the various sequences were chosen so that the sequence $(G_{g_i} \xi'(y_i) v_i)$ converges to some $w \in C(T, \mathbb{R}^n)$ and

$$\lim_i G_{g_i} z = G_{g_\infty} z \quad [z \in L^\infty(\mu, \mathbb{R}^k)]$$

for some $g_\infty \in \mathcal{G}^{\varepsilon_1}$. Furthermore, since all G_{g_i} belong to a bounded subset of $B(L^\infty(\mu, \mathbb{R}^k), C(T, \mathbb{R}^n))$, we also have

$$\lim_i G_{g_i} z_i = G_{g_\infty} z_0 \quad \text{whenever} \quad \lim_i z_i = z_0 \text{ in } L^\infty(\mu, \mathbb{R}^k).$$

It follows now from (6) that

$$w = \lim_i v_i = \lim_i G_{g_i} \xi'(y_i) v_i = G_{g_\infty} \xi'(y(\sigma_0)) w$$

and therefore $w = 0$ since $I - G_{g_\infty} \xi'(y(\sigma_0))$ is one-to-one. This contradicts the relations

$$\lim_i v_i = w, \quad |v_i| = 1. \quad \text{Q.E.D.}$$

In addition to ordinary (Fréchet) derivatives, we shall also use the more general concept of a derivative with respect to a set. For our present purposes, we can limit ourselves to the case of a function $\chi: B \rightarrow F$, where E and F are Banach spaces and B is a subset of E which is either open or convex with $B^\circ \neq \emptyset$. We say that $\chi'(b_0)$ is the derivative of χ at $b_0 \in B$ if $\chi'(b_0) \in B(E, F)$ and

$$\lim |b - b_0|^{-1} |\chi(b) - \chi(b_0) - \chi'(b_0)(b - b_0)| = 0 \quad \text{as } b \rightarrow b_0, b \in B \sim \{b_0\}.$$

Clearly, when $b_0 \in B^\circ$, this definition coincides with that of a Fréchet derivative.

If E and F are vector spaces, B a convex subset of E , $b_0 \in B$ and $\chi: B \rightarrow F$, we represent the directional derivative of χ at b_0 in the direction of $b \in B$ by

$$D\chi(b_0; b - b_0) \triangleq \lim_{\alpha \rightarrow 0+} \alpha^{-1} [\chi(b_0 + \alpha(b - b_0)) - \chi(b_0)].$$

LEMMA 4.3. Let $f_\infty \triangleq f$. There exist $i_0 \in \mathbb{N}$ and $\beta_0 > 0$ such that

(1) the equation

$$y(t) = \int f_i(t, \tau, \xi(y)(\tau), \sigma(\tau)) \mu(d\tau) \quad (t \in T)$$

has a unique solution $y_i(\sigma)$ for each $i = i_0, i_0 + 1, \dots, \infty$ and each

$$\sigma \in \mathcal{S}_0 \triangleq \{\sigma_0 + \beta(\sigma' - \sigma) \mid \sigma' \in \mathcal{S}^*, 0 \leq \beta \leq \beta_0\};$$

(2) each of the functions

$$\sigma \rightarrow y_i(\sigma): \mathcal{S}_0 \rightarrow C(T, \mathbb{R}^n) \quad (i = i_0, i_0 + 1, \dots, \infty)$$

is continuous;

(3) $\lim_i y_i(\sigma) = y_\infty(\sigma) = y(\sigma)$ uniformly for $\sigma \in \mathcal{S}_0$;

(4) for every choice of $N \in \mathbb{N}$, $\sigma^1, \dots, \sigma^N \in \mathcal{S}_0$, $i = i_0, i_0 + 1, \dots (i < \infty)$ and for

$$\mathcal{T}_N \triangleq \left\{ \omega = (\omega^1, \dots, \omega^N) \in \mathbb{R}^N \mid \omega^j \geq 0, \sum_{j=1}^N \omega^j \leq 1 \right\},$$

the functions

$$\omega \rightarrow y_i \left(\sigma_0 + \sum_{j=1}^N \omega^j (\sigma^j - \sigma_0) \right): \mathcal{T}_N \rightarrow C(T, \mathbb{R}^n)$$

are continuously differentiable and

$$Dy_i(\sigma'; \sigma - \sigma') = [I - G_\gamma \xi'(y_i(\sigma'))]^{-1} \mathcal{F}_i(\sigma')(\sigma - \sigma') \quad (\sigma, \sigma' \in \mathcal{S}_0),$$

where

$$\gamma(t, \tau) \triangleq \mathcal{D}_3 f_i(t, \tau, \xi(y_i(\sigma'))(\tau), \sigma'(\tau)),$$

$$\mathcal{F}_i(\sigma')(\nu)(t) \triangleq \int f_i(t, \tau, \xi(y_i(\sigma'))(\tau), \nu(\tau)) \mu(d\tau),$$

and G_γ is defined as in the statement of Lemma 4.2.

Proof. Step 1. Let $\bar{\mathbb{N}} \triangleq \mathbb{N} \cup \{\infty\}$, $\mathcal{Y} \triangleq C(T, \mathbb{R}^n)$, $\alpha' > 0$ be such that

$$S^F(y_0(\sigma_0), \alpha') \subset Y' \triangleq S(y_0(\sigma_0), 2\alpha') \subset \xi^{-1}(L^\infty(\mu, V)),$$

and set

$$F_i(y, \sigma)(t) \triangleq \int f_i(t, \tau, \xi(y)(\tau), \sigma(\tau)) \mu(d\tau) \quad (i \in \bar{\mathbb{N}}),$$

$$G_i(y) \triangleq F_i(y, \sigma_0) \quad (i \in \mathbb{N}),$$

$$Q \triangleq \{y \in \mathcal{Y} \mid |y| \leq c_1 \mu(T), |y(t) - y(t')| \leq \int \psi(t, t', \tau) \mu(d\tau) (t, t' \in T)\}.$$

By the Arzèla–Ascoli theorem, the set Q is compact. For each $(y, \sigma) \in Y' \times \mathcal{S}^*$, $i \in \bar{\mathbb{N}}$ and $t, t' \in T$, we have

$$\begin{aligned} |F_i(y, \sigma)(t) - F_i(y, \sigma)(t')| &\leq \int |f_i(t, \tau, \xi(y)(\tau), \sigma(\tau)) - f_i(t', \tau, \xi(y)(\tau), \sigma(\tau))| \mu(d\tau) \\ &\leq \int \psi(t, t', \tau) \mu(d\tau), \\ |F_i(y, \sigma)(\tau)| &\leq c_1 \mu(T). \end{aligned}$$

Thus, for each $i \in \bar{\mathbb{N}}$, we can define the function

$$(y, \sigma) \rightarrow F_i(y, \sigma): Y' \times \mathcal{S}^* \rightarrow Q$$

and, by [6, VII.2.1(5), p. 414], F_i is continuous and, for $i \neq \infty$, $G_i \triangleq F_i(\cdot, \sigma_0)$ is C^1 . By [6, II.5.8, p. 212], we have

$$(G'_i(y)\Delta y)(t) = \int \mathcal{D}_3 f_i(t, \tau, \xi(y)(\tau), \sigma_0(\tau))(\xi'(y)\Delta y)(\tau) \mu(d\tau) \\ (\Delta y \in \mathcal{Y}, t \in T, i \in \mathbb{N}).$$

Now let \mathcal{G}^ε and ε_0 be as defined in Lemma 4.2. Then there exists $c > 0$ such that $||[I - G_g \xi'(y)]^{-1}|| \leq c$ for all $g \in \mathcal{G}^{\varepsilon_0}$ and $y \in S^F(y_0(\sigma_0), \varepsilon_0)$. Since ξ is continuous, it follows from 3.1(1–3) that there exist $i_1 \in \mathbb{N}$ and $\alpha \in (0, \min[\varepsilon_0, \alpha'])$ such that $g \in \mathcal{G}^{\varepsilon_0}$ whenever

$$i_1 \leq i < \infty, \quad y \in S^F(y(\sigma_0), \alpha), \quad g(t, \tau) = \mathcal{D}_3 f_i(t, \tau, \xi(y)(\tau), \sigma_0(\tau)) \quad (t, \tau \in T).$$

This shows that $||[I - G'_i(y)]^{-1}|| \leq c$ for $i_1 \leq i < \infty$ and $y \in S^F(y_0(\sigma_0), \alpha)$.

We next observe that if $\beta_0 \triangleq \min\{[4cc_1\mu(T)]^{-1}\alpha, 1\}$, $\beta \in [0, \beta_0]$, $\sigma' \in \mathcal{S}^*$ and $y \in Y \triangleq S(y_0(\sigma_0), 2\alpha)$ then

$$|F_\infty(y, \sigma_0 + \beta(\sigma' - \sigma_0)) - F_\infty(y, \sigma_0)| \\ = \beta \sup_{t \in T} \left| \int f(t, \tau, \xi(y)(\tau), \sigma'(\tau) - \sigma_0(\tau)) \mu(d\tau) \right| \leq 2\beta c_1 \mu(T) \leq \frac{\alpha}{2c}$$

and that

$$\lim_i F_i(y, \sigma) = F_\infty(y, \sigma) \quad \text{uniformly for } y \in Y, \sigma \in \mathcal{S}_0.$$

Furthermore, by Lemma 4.1 and [6, II.5.4, p. 206], for each $\sigma \in \mathcal{S}_0$ and $i \in \mathbb{N}$, the equation

$$y(t) = \int f_i(t, \tau, \xi(y)(\tau), \sigma(\tau)) \mu(d\tau) \quad (t \in T)$$

can have at most one solution y in $C(T, W)$. Thus all the assumptions of [12, Thm. 1] are satisfied (with $X \triangleq \mathcal{S}_0$ and $K \triangleq Q$) and that theorem yields statements (1)–(3).

Step 2. Let $N \in \mathbb{N}$, $\sigma', \sigma^1, \dots, \sigma^N \in \mathcal{S}_0$ and $i = i_0, i_0 + 1, \dots$. For each $\omega \in \mathcal{T}_N$, $z(\omega) \triangleq y_i(\sigma' + \sum_{j=1}^N \omega^j(\sigma^j - \sigma'))$ is the unique solution of the equation

$$y = F_i(y, \sigma') + \sum_{j=1}^N \omega^j F_i(y, \sigma^j - \sigma')$$

and, by 3.1 and [6, II.5.8, p. 212],

$$(\mathcal{D}_1 F_i(y, \sigma)\Delta y)(t) = \int \mathcal{D}_3 f_i(t, \tau, \xi(y)(\tau), \sigma(\tau))[\xi'(y)\Delta y](\tau) \mu(d\tau) \\ [t \in T, \sigma \in \mathcal{S}^*, y \in C(T, W), \Delta y \in C(T, \mathbb{R}^n)].$$

We can show, by the same argument as in Step 2 of the proof of Lemma 4.2, that for every $y \in C(T, W)$ and $\sigma \in \mathcal{S}^*$, $I - \mathcal{D}_1 F_i(y, \sigma)$ is a linear homeomorphism of $C(T, \mathbb{R}^n)$ onto itself. It follows thus from a variant of the classical implicit function theorem [6, II.3.8, p. 176] that $\omega \rightarrow z(\omega)$ is continuously differentiable and that, for $N = 1$ and $\sigma = \sigma^1$, $z'(\omega) = Dy_i(\sigma'; \sigma - \sigma')$ has the form indicated in statement (4). Q.E.D.

LEMMA 4.4. Let $m_2 = 1$,

$$\omega, \omega^i \in \text{frm}^+(T^h), \quad \omega^i(T^h) \leq 1 \quad (i \in \mathbb{N}),$$

$$\lim_i \int z(t) \omega^i(dt) = \int z(t) \omega(dt) \quad [z \in C(T^h)],$$

$$\mathcal{H}^\varepsilon(t) \triangleq \bigcap_{\eta > \varepsilon} \overline{\text{co}} \bigcup_{d(\tau, t) \leq \eta} \Lambda^\varepsilon h^2(\tau, y_\infty(\sigma_0)(\tau)) \quad (t \in T^h, \varepsilon > 0),$$

$$H_2^i \in C(T^h, B(\mathbb{R}^n, \mathbb{R})), \quad H_2^i(t) \in \mathcal{H}^{1/i}(t) \quad (t \in T^h, i \in \mathbb{N}).$$

Then there exist a bounded Borel measurable $H_2: T^h \rightarrow B(\mathbb{R}^n, \mathbb{R})$ and $J \subset (1, 2, \dots)$ such that

$$H_2(t) \in \bigcap_{\varepsilon > 0} \mathcal{H}^\varepsilon(t) \quad \omega\text{-a.e.},$$

$$\lim_{i \in J} \int H_2^i(t) \Delta y(t) \omega^i(dt) = \int H_2(t) \Delta y(t) \omega(dt) \quad [\Delta y \in C(T^h, \mathbb{R}^n)].$$

Proof. We can identify $B(\mathbb{R}^n, \mathbb{R})$ with \mathbb{R}^n and choose some compact subset X of \mathbb{R}^n that contains $\Lambda^\varepsilon h^2(t, v)$ for all $t \in T^h$ and $v \in V$. We also verify that if

$$\varepsilon > 0, \quad A(t) \subset X \quad (t \in T^h), \quad B(t) \triangleq \bigcap_{\eta > \varepsilon} \text{closure} \bigcup_{d(\tau, t) \leq \eta} A(\tau)$$

then the set $\{(t, x) \mid t \in T^h, x \in B(t)\}$ is closed and, as a consequence, the set

$$\{(t, x) \mid t \in T^h, x \in \overline{\text{co}} B(t)\}$$

is also closed. It follows that the set

$$\{(t, x) \mid t \in T^h, x \in \mathcal{H}^\varepsilon(t)\}$$

is closed for every choice of $\varepsilon > 0$. Our conclusion then follows from [11, XI.3.5, p. 582] which remains valid, together with its proof, when T^h is an arbitrary compact metric space instead of being a compact interval as postulated in the reference.¹ Q.E.D.

5. Proofs of Theorems 2.4 and 2.5. It will be sufficient to prove Theorems 2.4 and 2.5 in the special case where $m_2 = 1$ and $A = (-\infty, 0]$. Indeed, if this is not the case, we can choose an arbitrary point $x_0 \in A^\circ$, set $B \triangleq A - x_0$,

$$\gamma(x) \triangleq \inf \left\{ \alpha > 0 \mid \frac{1}{\alpha} x \in B \right\} \quad (x \in \mathbb{R}^{m_2})$$

and

$$\tilde{h}^2(t, v) = \gamma(h^2(t, v) - x_0) - 1 \quad (t \in T^h, v \in V).$$

It is well known that γ (called the “gauge function of B ”) is Lipschitz continuous and that x belongs to the interior (boundary, exterior) of B if and only if $\gamma(x) < 1$ ($\gamma(x) = 1$, $\gamma(x) > 1$). Thus the statement

$$\text{“there exists } \kappa > 0 \text{ such that } h^2(t, v) + S^F(0, \kappa) \in A \text{”}$$

¹ The statement and proof of [11, XI.3.5, p. 582] contain a few typographical mistakes which can be corrected by replacing T by T^h everywhere. The arguments of [11, XI.3.5, p. 582] are similar to those in [8, Lemma 3.2, p. 30].

is equivalent to the statement

“there exists $\kappa' > 0$ such that $\tilde{h}^2(t, v) + \kappa' \in (-\infty, 0]$ ”.

This shows that we may replace h^2 and A by \tilde{h}^2 and $(-\infty, 0]$, respectively, without changing our problem. When we do this and establish the corresponding versions of Theorems 2.4 and 2.5, these results can then be translated into terms involving h^2 and $\Lambda^\varepsilon h^2$ directly, yielding the general form of Theorems 2.4 and 2.5. We shall not go through the details of this procedure because they are identical with the arguments in [10, 6.1, p. 31] or [11, XI.3.3, p. 578].

In view of the above remarks, we shall assume henceforth that $m_2 = 1$ and $A = (-\infty, 0]$. We shall also continue using the notation of the previous sections and shall refer, in particular, to ε_0 , G_g and \mathcal{G}^ε as defined in Lemma 4.2 and to \mathcal{T}_N , β_0 , \mathcal{S}_0 , $\sigma \rightarrow y_i(\sigma)$ ($i = i_0, i_0 + 1, \dots, \infty$) and $(\sigma', \nu) \rightarrow \mathcal{F}_i(\sigma')(\nu)$ as defined in Lemma 4.3. We shall assume that $i_0 = 1$ which can be done by replacing the sequence (f_i) by $(f_i)_{i \geq i_0}$ and renumbering the indices. We next set

$$\begin{aligned} (h_\infty^1, h_\infty^2) &\triangleq (h^1, h^2), \\ \phi_i(\sigma) &\triangleq h_i^1(y_i(\sigma)(t_1)), \quad \phi \triangleq \phi_\infty, \\ \Phi_i(\sigma)(t) &\triangleq h_i^2(t, y_i(\sigma)(t)), \quad \Phi \triangleq \Phi_\infty \quad (i \in \tilde{\mathbb{N}}, \sigma \in \mathcal{S}_0, t \in T^h). \end{aligned}$$

For each $\varepsilon \in (0, \varepsilon_0]$ we define the set $\Lambda^\varepsilon(\phi, \Phi)(\sigma_0)$ as the collection of all linear (but not necessarily continuous) operators $M = (M_1, M_2)$ on \mathcal{N} into $\mathbb{R}^m \times C(T^h)$ of the form

$$M_1(\nu) = H_1 x(\nu)(t_1), \quad M_2(\nu)(t) = H_2(t) x(\nu)(t) \quad (\nu \in \mathcal{N}, t \in T^h),$$

where

$$\begin{aligned} H_1 &\in \Lambda^\varepsilon h^1(y_\infty(\sigma_0)(t_1)), \quad H_2 \in C(T^h, B(\mathbb{R}^n, \mathbb{R})), \\ H_2(t) &\in \bigcap_{\varepsilon > 0} \overline{\bigcup_{d(\tau, t) \leq \varepsilon} \Lambda^\varepsilon h^2(\tau, y_\infty(\sigma_0)(t))} \quad (t \in T^h), \\ x(\nu) &= [I - G_g \xi'(y_i(\sigma'))]^{-1} \mathcal{F}_i(\sigma')(\nu), \\ g &\in \mathcal{G}^\varepsilon, \quad i \in \mathbb{N}, \quad i \geq 1/\varepsilon, \quad \sigma' \in \mathcal{S}_0, \quad |y_i(\sigma') - y_\infty(\sigma_0)| \leq \varepsilon. \end{aligned}$$

For $\varepsilon > \varepsilon_0$, we set $\Lambda^\varepsilon(\phi, \Phi)(\sigma_0) \triangleq \Lambda^{\varepsilon_0}(\phi, \Phi)(\sigma_0)$.

We next establish certain properties of (ϕ_i, Φ_i) and $\Lambda^\varepsilon(\phi, \Phi)(\sigma_0)$.

LEMMA 5.1.

- I. $(\phi_i, \Phi_i) \in C(\mathcal{S}_0, \mathbb{R}^m \times C(T^h))$ ($i \in \tilde{\mathbb{N}}$),
 $\lim_i (\phi_i, \Phi_i) = (\phi, \Phi)$ uniformly on \mathcal{S}_0 .
- II. $\Lambda^\varepsilon(\phi, \Phi)(\sigma_0) \subset \Lambda^{\varepsilon'}(\phi, \Phi)(\sigma_0)$ ($\varepsilon' > \varepsilon$).
- III. For every $\varepsilon > 0$, the set

$$\{M|_{\mathcal{S}_0} \mid M = (M_1, M_2) \in \Lambda^\varepsilon(\phi, \Phi)(\sigma_0)\}$$

is a bounded and equicontinuous subset of $C(\mathcal{S}_0, \mathbb{R}^m \times C(T^h))$. (Here $M|_{\mathcal{S}_0}$ represents the restriction of M to \mathcal{S}_0).

IV. For every choice of $N \in \mathbb{N}$, $\sigma^1, \dots, \sigma^N \in \mathcal{S}_0$ and $\varepsilon > 0$ there exist $\delta > 0$, $i^* \in \mathbb{N}$ and the corresponding set

$$\mathcal{S}_\delta^* \triangleq \left\{ \sigma_0 + \sum_{j=1}^N \omega^j (\sigma^j - \sigma_0) \mid (\omega^1, \dots, \omega^N) \in \delta \mathcal{T}_N \right\}$$

such that the functions

$$\omega \rightarrow (\phi_i, \Phi_i) \left(\sigma_0 + \sum_{j=1}^N \omega^j (\sigma^j - \sigma_0) \right) : \delta \mathcal{T}_N \rightarrow \mathbb{R}^m \times C(T^h) \quad (i \in \mathbb{N})$$

are continuously differentiable and for every $\sigma' \in \mathcal{S}_\delta^*$ and $i \in \mathbb{N}$, $i \geq i^*$, there exists $M \in \Lambda^\varepsilon(\phi, \phi)(\sigma_0)$ satisfying

$$D(\phi_i, \Phi_i)(\sigma'; \sigma - \sigma') = M(\sigma - \sigma') \quad (\sigma \in \mathcal{S}_0).$$

Proof. Step 1. Statement I follows from Lemma 4.3 and statement II from the definitions of $\Lambda^\varepsilon h^1$, $\Lambda^\varepsilon h^2$ and \mathcal{G}^ε . We next prove statement III.

Let $\tilde{\mathbb{N}}$ be endowed with the relative topology of $\bar{\mathbb{R}}$ (the compact metric space of extended real numbers). By Lemma 4.3, the functions $\sigma' \rightarrow y_i(\sigma') : \mathcal{S}_0 \rightarrow C(T, \mathbb{R}^n)$ ($i \in \tilde{\mathbb{N}}$) are continuous and $\lim_j y_j(\sigma') = y_\infty(\sigma')$ uniformly for all $\sigma' \in \mathcal{S}_0$; by Assumption 2.1 and by the construction in § 3.1, $\lim_j f_j = f_\infty$ uniformly, all $f_i(t, \tau, \cdot, r)$ have a common Lipschitz constant c_1 , all $\mathcal{F}_i(\sigma')(\sigma)(\cdot)$ are equicontinuous, and all functions $\sigma \rightarrow \mathcal{F}_i(\sigma')(\sigma) : \mathcal{S}^* \rightarrow C(T, \mathbb{R}^n)$ are continuous. Thus, if

$$i \in \tilde{\mathbb{N}}, \quad \lim_j \sigma'_j = \sigma' \text{ in } \mathcal{S}_0, \quad \lim_j \sigma_j = \sigma \text{ in } \mathcal{S}_0,$$

then

$$\lim_j \int \chi(\tau, \sigma_j(\tau) - \sigma(\tau)) \mu(d\tau) = 0$$

whenever $\chi(\cdot, r)$ is μ -measurable, $\chi(t, \cdot)$ continuous and χ bounded, and therefore

$$\begin{aligned} & \lim_j |\mathcal{F}_i(\sigma'_j)(\sigma_j) - \mathcal{F}_i(\sigma')(\sigma)| \\ & \leq \lim_j \sup_{t \in T} \left\{ \int |f_i(t, \tau, \xi(y_i(\sigma'_j))(\tau), \sigma_j(\tau)) - f_i(t, \tau, \xi(y_i(\sigma'))(\tau), \sigma_j(\tau))| \mu(d\tau) \right. \\ & \quad \left. + \left| \int f_i(t, \tau, \xi(y_i(\sigma'))(\tau), \sigma_j(\tau) - \sigma(\tau)) \mu(d\tau) \right| \right\} \\ & \leq \lim_j c_1 \int |\xi(y_i(\sigma'_j))(\tau) - \xi(y_i(\sigma'))(\tau)| \mu(d\tau) = 0 \end{aligned}$$

and

$$\begin{aligned} & \lim_j |\mathcal{F}_j(\sigma_j)(\sigma_j) - \mathcal{F}_\infty(\sigma')(\sigma)| \\ & \leq \lim_j \sup_{t \in T} \left\{ \int |f_j(t, \tau, \xi(y_j(\sigma'_j))(\tau), \sigma_j(\tau)) - f_j(t, \tau, \xi(y_\infty(\sigma'))(\tau), \sigma_j(\tau))| \mu(d\tau) \right. \\ & \quad + \int |f_j(t, \tau, \xi(y_\infty(\sigma'))(\tau), \sigma_j(\tau)) - f_\infty(t, \tau, \xi(y_\infty(\sigma'))(\tau), \sigma_j(\tau))| \mu(d\tau) \\ & \quad \left. + \left| \int f_\infty(t, \tau, \xi(y_\infty(\sigma'))(\tau), \sigma_j(\tau) - \sigma(\tau)) \mu(d\tau) \right| \right\} = 0. \end{aligned}$$

This shows that the function

$$(\sigma', \sigma, i) \rightarrow \mathcal{F}_i(\sigma')(\sigma) : \mathcal{S}_0 \times \mathcal{S}_0 \times \tilde{\mathbb{N}} \rightarrow C(T, \mathbb{R}^n)$$

is continuous. Since \mathcal{S}_0 and $\tilde{\mathbb{N}}$ are compact, this implies that the collection

$$\{\mathcal{F}_i(\sigma')(\cdot) \mid i \in \tilde{\mathbb{N}}, \sigma' \in \mathcal{S}_0\}$$

in $C(\mathcal{S}_0, C(T, \mathbb{R}^n))$ is equicontinuous. By § 3.1, $|f_i(t, \tau, v, r)| \leq c_1$ for all i, t, τ, v, r and thus $\mathcal{F}_i(\sigma')(\sigma)$ are uniformly bounded.

By Lemma 4.2, $[I - G_g \xi'(y)]^{-1}$ are uniformly bounded for $g \in \mathcal{G}^{\varepsilon_0}$ and $y \in C(T, W)$, with $|y - y_\infty(\sigma_0)| \leq \varepsilon_0$. It follows that, for each $\varepsilon \in (0, \varepsilon_0]$, the collection

$$\{[I - G_g \xi'(y_i(\sigma'))]^{-1} \mathcal{F}_i(\sigma')(\cdot) \mid g \in \mathcal{G}^\varepsilon, i \in \tilde{\mathbb{N}}, \sigma' \in \mathcal{S}_0, |y_i(\sigma') - y_\infty(\sigma_0)| \leq \varepsilon\}$$

is an equicontinuous and bounded subset of $C(\mathcal{S}_0, C(T, \mathbb{R}^n))$. Since $\Lambda^\varepsilon h^1(y_0(\sigma_0)(t_1))$ and $\Lambda^\varepsilon h^2(t, y_0(\sigma_0)(t))$ are uniformly bounded subsets of $B(\mathbb{R}^n, \mathbb{R}^m)$ and $B(\mathbb{R}^n, \mathbb{R})$, respectively, the conclusion of statement III follows.

Step 2. It remains to prove statement IV. By statement (4) of Lemma 4.3,

$$\omega \rightarrow (\phi_i, \Phi_i) \left(\sigma_0 + \sum_{j=1}^N \omega^j (\sigma^j - \sigma_0) \right): \mathcal{T}_N \rightarrow \mathbb{R}^m \times C(T^h)$$

are continuously differentiable for $i \in \mathbb{N}$ and we have

$$D(\phi_i, \Phi_i)(\sigma'; \sigma - \sigma') = (\tilde{M}_1(\sigma - \sigma'), \tilde{M}_2(\sigma - \sigma')) \quad (\sigma, \sigma' \in \mathcal{S}_0)$$

with

$$\tilde{M}_1(\nu) = H_1 x(\nu)(t_1), \quad \tilde{M}_2(\nu)(t) = H_2(t) x(\nu)(t) \quad (\nu \in \mathcal{N}),$$

$$H_1 = \mathcal{D}h^1(y_i(\sigma')(t_1)), \quad H_2(t) = \mathcal{D}_2 h^2(t, y_i(\sigma')(t))$$

$$x(\nu) = [I - G_g \xi'(y_i(\sigma'))]^{-1} \mathcal{F}_i(\sigma')(\nu),$$

$$g(t, \tau) = \mathcal{D}_3 f_i(t, \tau, \xi(y_i(\sigma'))(\tau), \sigma'(\tau)).$$

Since $\sigma' \rightarrow y_i(\sigma')$ are continuous, $\lim_i y_i(\sigma') = y_\infty(\sigma')$ uniformly on \mathcal{S}_0 and $|\mathcal{D}_3 f_i(t, \tau, v, r)| \leq c_1$, we may choose $\delta \in (0, \beta_0]$ and $i^* \in \mathbb{N}$ such that, for $\sigma' = \sigma_0 + \sum_{j=1}^N \omega^j (\sigma^j - \sigma_0) \in \mathcal{S}_\delta^*$ and $i^* \leq i < \infty$, we have

$$|y_i(\sigma') - y_\infty(\sigma_0)| \leq \varepsilon, \quad H_1 \in \Lambda^\varepsilon h^1(y_0(\sigma_0)(t_1)), \quad H_2(t) \in \Lambda^\varepsilon h^2(t, y_\infty(\sigma_0)(t)),$$

$$g(t, \tau) = \mathcal{D}_3 f_i(t, \tau, \xi(y_i(\sigma'))(\tau), \sigma'(\tau))$$

$$= \mathcal{D}_3 f_i(t, \tau, \xi(y_i(\sigma'))(\tau), \sigma_0(\tau)) + \sum_{j=1}^N \omega^j \mathcal{D}_3 f_i(t, \tau, \xi(y_i(\sigma'))(\tau), \sigma^j(\tau) - \sigma_0(\tau))$$

$$\in S(\partial^\varepsilon f(t, \tau, \xi(y_\infty(\sigma_0))(\tau), \sigma_0(\tau)), \varepsilon).$$

It follows then that $(\tilde{M}_1, \tilde{M}_2) \in \Lambda^\varepsilon(\phi, \Phi)(\sigma_0)$. Thus statement IV is valid. Q.E.D.

Properties I—IV of $\Lambda^\varepsilon(\phi, \Phi)(\sigma_0)$ as stated in Lemma 5.1 are precisely those that define a directional derivate container for (ϕ, Φ) at σ_0 [13, Def. 2.1]. We shall next find a representation for the elements of the corresponding *scalar directional derivate container* $\mathcal{L}\Lambda(\phi, \Phi, C)(\sigma_0)$ for (ϕ, Φ, C) at σ_0 , where

$$C \triangleq \{z \in C(T^h) \mid z(t) \leq 0 \ (t \in T^h)\}.$$

These elements are triplets (l_1, l_2, λ) such that $l = (l_1, l_2) \in \mathbb{R}^m \times C(T^h)^*$, $l \neq 0$, $\lambda \in \mathcal{L}(\mathcal{N}, \mathbb{R})$, $\lambda|_{\mathcal{S}_0}$ is continuous, and there exist sequences $(l^i) = ((l_1^i, l_2^i))$ and (M^i) with $l_2^i \in C(T^h)^*$ and $M^i \in \Lambda^{1/i}(\phi, \Phi)(\sigma_0)$, satisfying

$$|l_1^i| + |l_2^i| = 1, \quad \lim_i l_1^i = l_1, \quad \lim_i l_2^i z = l_2 z \quad (z \in C(T^h)),$$

$$\lim_i l^i M^i(\sigma) = \lambda(\sigma) \quad (\sigma \in \mathcal{S}_0), \quad l_2^i z \leq 0 \quad \text{if } z + \Phi(\sigma_0) + S^F(0, 1/i) \subset C.$$

LEMMA 5.2. Let $(l_1, l_2, \lambda) \in \mathcal{L}\Lambda(\phi, \Phi, C)(\sigma_0)$. Then there exist

$$g \in \mathcal{G}, \quad H_1 \in \Lambda h^1(y_\infty(\sigma_0)(t_1)), \quad H_2: T^h \rightarrow B(\mathbb{R}^n, \mathbb{R}), \quad \omega \in \text{frm}^+(T^h)$$

such that H_2 is bounded and Borel measurable,

$$(1) \quad |l_1| + \omega(T^h) > 0,$$

$$(2) \quad H_2(t) \in \bigcap_{\varepsilon > 0} \overline{\text{co}} \bigcup_{d(\tau, t) \leq \varepsilon} \Lambda^\varepsilon h^2(\tau, y_\infty(\sigma_0)(\tau)) \quad \omega\text{-a.e.}$$

$$(3) \quad l_2(z) = \int z(t) \omega(dt) \quad [z \in C(T^h)],$$

$$(4) \quad \lambda(\sigma) = l_1^T H_1 x(\sigma)(t_1) + \int H_2(t) x(\sigma)(t) \omega(dt) \quad (\sigma \in \mathcal{S}_0),$$

where

$$x(\nu) \triangleq [I - G_g \xi'(y_\infty(\sigma_0))]^{-1} \mathcal{F}_\infty(\sigma_0)(\nu) \quad (\nu \in \mathcal{N}).$$

Proof. Let $((l_1^i, l_2^i))$ and (M^i) be sequences corresponding to (l_1, l_2, λ) as in the definition of $\mathcal{L}\Lambda(\phi, \Phi, C)(\sigma_0)$. By the Riesz representation theorem, $l_2, l_2^i \in C(T^h)^*$ can be represented by some $\omega, \omega^i \in \text{frm}(T^h)$ such that

$$(5) \quad l_2^i z = \int z(t) \omega^i(dt) \xrightarrow{i} \int z(t) \omega(dt) = l_2 z \quad [z \in C(T^h)].$$

Since $l_2^i z \leq 0$ if $z + \Phi(\sigma_0) + S^F(0, 1/i) \subset C$, i.e. whenever

$$z(t) + h^2(t, y_\infty(\sigma_0)(t)) + 1/i \leq 0 \quad (t \in T^h),$$

it follows from (5) that ω and each ω^i are nonnegative measures. We also have $\omega^i(T^h) \leq 1$ because $|l_1^i| + |l_2^i| = 1$.

By the definition of

$$\Lambda^{1/i}(\phi, \Phi)(\sigma_0)$$

there exist $j(i), H_1^i, H_2^i, g_i, F_{j(i)}, \sigma_i', \eta_i$ and x_i such that

$$j(i) \in \{i, i+1, \dots\}, \quad H_1^i \in \Lambda^{1/i} h^1(y_\infty(\sigma_0)(t_1)), \quad H_2^i \in C(T^h, B(\mathbb{R}^n, \mathbb{R})),$$

$$H_2^i(t) \in \bigcap_{\eta > 1/i} \overline{\text{co}} \bigcup_{d(\tau, t) \leq \eta} \Lambda^{1/i} h^2(\tau, y_\infty(\sigma_0)(\tau)),$$

$$\sigma_i' \in \mathcal{S}_0, \quad g_i \in \mathcal{G}^{1/i}, \quad \eta_i = y_{j(i)}(\sigma_i'), \quad |\eta_i - y_\infty(\sigma_0)| \leq 1/i,$$

$$x_i(\nu) = [I - G_{g_i} \xi'(\eta_i)]^{-1} \mathcal{F}_{j(i)}(\sigma_i')(\nu) \quad (\nu \in \mathcal{N}),$$

$$M_1^i(\nu) = H_1^i x_i(\nu)(t_1), \quad M_2^i(\nu)(t) = H_2^i(t) x_i(\nu)(t) \quad (t \in T^h, \nu \in \mathcal{N}).$$

By Lemma 4.2, there exist $g \in \mathcal{G} (= \bigcap_{\varepsilon > 0} \mathcal{G}^\varepsilon)$ and $J \subset (1, 2, \dots)$ such that

$$\lim_{i \in J} G_{g_i} w = G_g w \quad \text{in } C(T, \mathbb{R}^n) \quad [w \in L^\infty(\mu, \mathbb{R}^k)].$$

Since $\lim_i \eta_i = y_\infty(\sigma_0)$ in $C(T, \mathbb{R}^n)$, ξ' is continuous, and $g_i(t, \tau)$ are uniformly bounded, we have

$$\begin{aligned} \lim_{i \in J} (G_{g_i} \xi'(\eta_i) \Delta y)(t) &= \lim_{i \in J} \int g_i(t, \tau) (\xi'(\eta_i) \Delta y)(\tau) \mu(d\tau) \\ &= \lim_{i \in J} \int g_i(t, \tau) (\xi'(y_\infty(\sigma_0)) \Delta y)(\tau) \mu(d\tau) \\ &= \int g(t, \tau) (\xi'(y_\infty(\sigma_0)) \Delta y)(\tau) \mu(d\tau) \end{aligned}$$

uniformly for all $t \in T$ and all Δy in the unit ball of $C(T, \mathbb{R}^n)$; hence

$$\lim_{i \in J} G_{g_i} \xi'(\eta_i) = G_g \xi'(y_\infty(\sigma_0)) \quad \text{in } B(C(T, \mathbb{R}^n), C(T, \mathbb{R}^n)).$$

We verify that

$$\lim_i \mathcal{F}_{j(i)}(\sigma_i')(\nu) = \mathcal{F}_\infty(\sigma_0)(\nu) \quad \text{in } C(T, \mathbb{R}^n) \quad (\nu \in \mathcal{N}),$$

and therefore

$$\lim_{i \in J} x_i(\nu) = x(\nu) \quad \text{in } C(T, \mathbb{R}^n) \quad (\nu \in \mathcal{N}).$$

We may choose J so that $(H_1^i)_{i \in J}$ converges to some $H_1 \in \Lambda h^1(y_\infty(\sigma_0)(t_1))$ and, by Lemma 4.4,

$$\lim_{i \in J} \int H_2^i(t) \Delta y(t) \omega^i(dt) = \int H_2(t) \Delta y(t) \omega(dt) \quad [\Delta y \in C(T^h, \mathbb{R}^n)]$$

for some bounded Borel measurable $H_2: T^h \rightarrow B(\mathbb{R}^n, \mathbb{R})$ satisfying relation (2). Because H_2^i are uniformly bounded, it follows that

$$\begin{aligned} \lambda(\sigma) &= \lim_i (l_1^i, l_2^i) M^i(\sigma) \\ &= \lim_{i \in J} \left\{ l_1^i T H_1^i x_i(\sigma)(t_1) + \int H_2^i(t) x_i(\sigma)(t) \omega^i(dt) \right\} \\ &= l_1^T H_1 x(\sigma)(t_1) + \int H_2(t) x(\sigma)(t) \omega(dt) \quad (\sigma \in \mathcal{S}_0) \end{aligned}$$

which proves relation (4). Relation (3) follows from (5), and relation (1) from (3) and $(l_1, l_2) \neq 0$. Q.E.D.

5.3. Proof of Theorem 2.4. Assume that the second alternative of the theorem is invalid. Then, a fortiori, a similar statement to that alternative, with $\sigma^1, \dots, \sigma^N$ restricted to \mathcal{S}_0 , also remains invalid. If we set $K \triangleq \mathcal{S}_0$, $K^* \triangleq \mathcal{S}^*$, $k_0 \triangleq \sigma_0$, then it follows from Lemma 5.1 and [13, Thm. 2.2] that there exists $(l_1, l_2, \lambda) \in \mathcal{L}\Lambda(\phi, \Phi, C)(\sigma_0)$ such that

$$(1) \quad \lambda(\sigma_0) = \text{Min}_{\sigma \in \mathcal{S}_0} \lambda(\sigma), \quad l_2 \Phi(\sigma_0) = \text{Max}_{c \in C} l_2 c.$$

By Lemma 4.3, $y(\sigma) = y_\infty(\sigma)$ ($\sigma \in \mathcal{S}_0$) and therefore, by Lemma 5.2, there exist

$$g \in \mathcal{G}, \quad H_1 \in \Lambda h^1(y_\infty(\sigma_0)(t_1)), \quad H_2: T^h \rightarrow B(\mathbb{R}^n, \mathbb{R}), \quad \omega \in \text{frm}^+(T^h)$$

such that H_2 is bounded and Borel measurable and relations 5.2(1)–(4) (that is, relations (1)–(4) of Lemma 5.2) are valid. The second relation of (1) and 5.2(3) yield

$$\begin{aligned} l_2 \Phi(\sigma_0) &= \int h^2(t, y(\sigma_0)(t)) \omega(dt) \\ &= \text{Max} \left\{ \int c(t) \omega(dt) \mid c(t) \leq 0 \ (t \in T), c \in C(T^h) \right\} \end{aligned}$$

which implies that

$$\omega(\{t \in T^h \mid h^2(t, y(\sigma_0)(t)) < 0\}) = 0, \quad h^2(t, y(\sigma_0)(t)) = 0 \quad \omega\text{-a.e.}$$

We can now deduce from 5.2(1)–(2) that relation 2.3(2) and the first two relations of 2.3(3) are satisfied (with $L_1 = 1$).

Since relation 2.3(1) is valid by assumption, it remains to prove the last relation of 2.3(3) and 2.3(4). To do so, we observe that, by (1) and 5.2(4),

$$(2) \quad \begin{aligned} \lambda(\sigma_0) &= l_1^T H_1 x(\sigma_0)(t_1) + \int H_2(t) x(\sigma_0)(t) \omega(dt) \\ &= \text{Min}_{\sigma \in \mathcal{S}_0} \left\{ l_1^T H_1 x(\sigma)(t_1) + \int H_2(t) x(\sigma)(t) \omega(dt) \right\} = \text{Min}_{\sigma \in \mathcal{S}_0} \lambda(\sigma), \end{aligned}$$

where

$$(3) \quad \begin{aligned} x(\sigma) &= [I - G_g \xi'(y(\sigma_0))]^{-1} \mathcal{F}_\infty(\sigma_0)(\sigma) \quad (\sigma \in \mathcal{S}_0), \\ \mathcal{F}_\infty(\sigma_0)(\sigma)(t) &= \int f(t, \tau, \xi(y(\sigma_0))(\tau)) \mu(d\tau) \quad (t \in T, \sigma \in \mathcal{S}_0). \end{aligned}$$

By Lemma 4.2, the function

$$\Delta y \rightarrow \mathcal{R}(\Delta y): C(T, \mathbb{R}^n) \rightarrow C(T, \mathbb{R}^n)$$

(as defined in 2.3(3)) is continuous and linear and therefore the function

$$\Delta y \rightarrow Z(\Delta y) \triangleq l_1^T H_1 \mathcal{R}(\Delta y)(t_1) + \int H_2(t) \mathcal{R}(\Delta y)(t) \omega(dt): C(T, \mathbb{R}^n) \rightarrow \mathbb{R}$$

is a continuous linear functional on $C(T, \mathbb{R}^n)$. Thus, by a variant of the Riesz representation theorem [6, I.5.9, p. 117], there exist $\zeta_1, \dots, \zeta_n \in \text{frm}(T)$, $\zeta_0 \in \text{frm}^+(T)$ and $\tilde{\zeta} \in L^1(\zeta_0, \mathbb{R}^n)$ such that

$$(4) \quad Z(\Delta y) = \int \tilde{\zeta}(t)^T \Delta y(t) \zeta_0(dt) = \sum_{i=1}^n \int \Delta y^i(t) \zeta_i(dt) \quad [\Delta y \in C(T, \mathbb{R}^n)],$$

which yields the last relation of 2.3(3).

By (3), $x(\sigma) = \mathcal{R}(\mathcal{F}_\infty(\sigma_0)(\sigma))$ and therefore (2), (3) and (4) yield

$$\begin{aligned} &\int \zeta_0(dt) \int \tilde{\zeta}(t)^T f(t, \tau, \xi(y(\sigma_0))(\tau), \sigma_0(\tau)) \mu(d\tau) \\ &= \text{Min}_{\sigma \in \mathcal{S}_0} \int \zeta_0(dt) \int \tilde{\zeta}(t)^T f(t, \tau, \xi(y(\sigma_0))(\tau), \sigma(\tau)) \mu(d\tau). \end{aligned}$$

Since $\sigma' = \sigma_0 + \beta_0(\sigma - \sigma_0) \in \mathcal{S}_0$ for every choice of $\sigma \in \mathcal{S}^*$, the above relation yields

$$\int \zeta_0(dt) \int \mu(d\tau) \int \tilde{\zeta}(t)^T \bar{f}(t, \tau, r)(\sigma - \sigma_0)(\tau)(dr) \geq 0 \quad (\sigma \in \mathcal{S}^*),$$

where $\bar{f}(t, \tau, r) \triangleq f(t, \tau, \xi(y(\sigma_0))(\tau), r)$. It follows, by [6, VII.3.1, p. 419], that

$$\int \zeta_0(dt) \int \tilde{\zeta}(t)^T \bar{f}(t, \tau, r) \sigma_0(\tau)(dr) = \text{Min}_{r \in R^*(\tau)} \int \tilde{\zeta}(t)^T \bar{f}(t, \tau, r) \zeta_0(dt)$$

for μ -a.a. $\tau \in T$

which, together with (4), implies the validity of 2.3(4). Q.E.D.

Proof of Theorem 2.5. Let

$$\phi^0(\sigma) \triangleq h^0(y(\sigma)(t_1)), \quad \phi^1(\sigma) \triangleq h^1(y(\sigma)(t_1)),$$

and let Φ be defined as before. By [13, Thm. 2.3], there exists $((l_0, l_1), l_2, \lambda) \in \mathcal{L}\Lambda((\phi^0, \phi^1), \Phi, C)(\sigma_0)$ such that

$$l_0 \geq 0, \quad \lambda(\sigma_0) = \text{Min}_{\sigma \in \mathcal{S}_0} \lambda(\sigma), \quad l_2 \Phi(\sigma_0) = \text{Max}_{c \in C} l_2 c.$$

Then the conclusion of Theorem 2.5 follows exactly as in the proof of Theorem 2.4. Q.E.D.

REFERENCES

- [1] F. H. CLARKE, *The maximum principle under minimal hypotheses*, this Journal, 14 (1976), pp. 1078–1091.
- [2] ———, *On the inverse function theorem*, Pacific J. Math., 64 (1976), pp. 97–102.
- [3] ———, *Generalized gradients of Lipschitz functionals*, Mathematics Research Center, Univ. of Wisconsin, Madison, Aug. 10, 1976.
- [4] ———, *Multiple integrals of Lipschitz functions in the calculus of variations*, preprint.
- [5] E. I. KUGUŠEV, *The maximum principle in problems of optimal control of systems with non-smooth right-hand side*, Vestnik Moskov. Univ. Ser. I Mat. Meh., 28 (1973), no. 3, pp. 103–117. (In Russian; English summary.)
- [6] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
- [7] ———, *Necessary conditions without differentiability assumptions in optimal control*, J. Differential Equations, 18 (1975), pp. 41–62.
- [8] ———, *Necessary conditions without differentiability assumptions in unilateral control problems*, Ibid., 21 (1976), pp. 25–38.
- [9] ———, *Controllability and necessary conditions in unilateral problems without differentiability assumptions*, this Journal, 14 (1976), pp. 546–572.
- [10] ———, *Derivate containers, inverse functions, and controllability*, Calculus of Variations and Control Theory, D. L. Russell, ed., Academic Press, New York, 1976.
- [11] ———, Chapter XI, Appendix to the Russian translation of *Optimal Control of Differential and Functional Equations*, Nauka, Moscow, 1977.
- [12] ———, *An implicit function theorem without differentiability*, Proc. Amer. Math. Soc., to appear.
- [13] ———, *Controllability and a multiplier rule for nondifferentiable optimization problems*, this Journal, pp. 803–812.
- [14] H. HALKIN, *Optimization without differentiability*, Proceedings of the Conference on Optimal Control Theory, Canberra (Australia, Aug. 1977), Springer-Verlag, to appear.

MULTI-AGENT LINEAR-QUADRATIC-GAUSSIAN SYSTEMS WITH "POOR" OBSERVATIONS*

J. F. RUDGE†

Abstract. Very few solutions are available for multi-agent control problems which have a nonclassical information structure. This is due to the fact that the agents can use their control actions not only for control itself, but also to pass information to other agents. We consider decentralized control situations in which information transfer is comparatively difficult because of large noise on some of the observation variables. By obtaining appropriate limits to this information transfer, linear control laws derived by disregarding the "poor" observations are shown to be "close" to optimal. It is then possible to derive other linear laws, from a similar situation again with static information structure, which are an order of magnitude closer, both in respect of the laws themselves and the optimized costs. It is shown however that iteration of this process in a natural way does not produce substantially better results.

1. Introduction. The extension of classical control theory to multi-agent situations is still only in its early stages despite the initial advances made by Radner [3] as long ago as 1962. The Linear-Quadratic-Gaussian (LQG) formulation, which is so successful for classical theory, does not have the same impact for decentralized control although no better framework has yet been found. Problems due to interaction between communication and control lie at the heart of multi-agent control whereas they are completely absent in single agent situations.

By introducing the important concept of partial nesting, Ho and Chu [2], classified many cases in which linearity of optimal control still held (as for classical theory). Witsenhausen [9], [10], [11] has considerably extended and generalized concepts of this nature in many papers which give a rigorous and very broadly based foundation to the subject. However, very few nonclassical problems have been solved. The Radner, and Ho and Chu papers, [3] and [2], mentioned above, together with Sandell and Athans [6], give examples classified by Witsenhausen [11] as quasiclassical. Radner's work was followed up in [4], which showed that the certainty equivalence principle has a natural extension to multi-agent situations in the form of repeated estimates or "second guesses"; a notion that had been accepted intuitively for some time; see, e.g. Chu and Ho [1]. Whittle and Rudge [7] considered a problem which was not quasiclassical but in which it was possible to separate the communication and control aspects. This enabled the two parts to be solved separately using communications theory and rate distortion theory.

Nonlinearity of optimal control occurs in general when it is possible for one agent to pass information to another by means of his control action. This information transfer will typically take place through an observation by the second agent of the first's action. The aim of this paper is to investigate the optimal control laws when this observation is badly distorted in a sense to be made precise later, so that intuitively, communication is difficult compared to control. One might expect the optimal control and the optimized cost both to approach those which hold when no such observations are allowed. By quantifying the amount of communication that can take place, we are able to state order of magnitude results for the differences between them in the two situations.

2. Problem definition. We suppose that the uncertainties of the external world or "states of nature" are represented by Gaussian random vectors ξ and η which are defined on some underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. There are N agents (or

* Received by the editors June 7, 1977, and in revised form January 20, 1978.

† Sheffield, S10 5PE England. This work was carried out at Cambridge University.

controllers or players) and each agent k ($1 \leq k \leq N$) has available to him information W_k before choosing his action u_k (an r_k column-vector which must be a measurable function of W_k). W_k is assumed to be separated into two parts Y_k and Z_k respectively, where Y_k is a linear function of ξ only and Z_k is linear in ξ , η , and some of the control actions of the other agents. Specifically,

$$(1) \quad Y_k = H_k \xi \quad Z_k = G_k \xi + \sum_{j=1}^N D_{kj} u_j + \eta_k$$

where $(\eta_1^T, \eta_2^T, \dots, \eta_N^T)$ is a partition of η ; ε is an n -vector, and the vectors η_i , Y_i , Z_i and matrices H_i , G_i , D_{ij} have consistent dimensions for all i and j . We assume that ξ , η_1, \dots, η_N are all mutually independent and that

$$\xi \sim N(0, \delta^2 \Sigma), \quad \eta_i \sim N(0, \Phi_i), \quad i \leq i \leq N,$$

where

$$\Sigma = \Sigma^T > 0, \quad \Phi_i = \Phi_i^T > 0.$$

We require causality of the system in that if the control action of player j affects the information of player k , i.e. $D_{kj} \neq 0$, then u_k cannot affect Z_j . Thus, in this case, there cannot exist l, m, \dots, n such that

$$D_{lk} \neq 0, \quad D_{ml} \neq 0, \dots, D_{jn} \neq 0.$$

If we denote:

$$u^T = (u_1^T; u_2^T; \dots; u_N^T)^T \quad \text{and} \quad r = \sum_{i=1}^N r_i$$

then the problem is to minimize the Bayes expected loss:

$$J(u) = \mathbb{E}(\frac{1}{2} u^T Q u + \xi S u)$$

where $Q = Q^T > 0$ is an $r \times r$ matrix and S is an $n \times r$ matrix. We partition Q into $r_i \times r_j$ blocks Q_{ij} and S into $n \times r_j$ blocks S_j . J is considered as a function of u merely to emphasize its dependence on the control law chosen.

Note that if $D_{kj} = 0 \forall k, j$, then the situation described is simply that of the static team. Our problem is to evaluate the nature of the solution to this more general problem as δ becomes small, although it should be noted that the term "solution" is used loosely here since it is not clear that an optimal control law does indeed exist. We obtain results relating given control laws to any others which may be candidates for optimal laws. The case of small δ corresponds to the situation in which the ξ dependent part of Z_k is poorly observed for each k . It is identical in principle to the case where the plant noise ξ is considered constant while the observation noise η become large, and we can transform the problem simply by considering ξ/δ and η/δ .

Before proving results for the above case, which we shall call "problem A", we need some notational conventions and essential lemmas.

3. Notation and background. It is convenient to work in the various Hilbert spaces \mathcal{H}_m say, of all m -vectors whose components are real valued, measurable, square integrable functions on Ω . The inner product on \mathcal{H}_m is defined by

$$(x, y)_m = \sum_{i=1}^m (x_i, y_i)_1 = \sum_{i=1}^m \int x_i(t) y_i(t) d\mathbb{P}(t) = \mathbb{E} x^T y$$

and the associated norm by $\|\cdot\|_m$. Subscripts will generally be omitted in cases where the space in question is clear.

Suppose that \mathcal{G} is a subsigma field of \mathcal{F} and that \mathcal{M} denotes the set of functions that are measurable with respect to \mathcal{G} . Then \mathcal{M} is a subspace in \mathcal{H}_1 and $\forall f \in \mathcal{H}_1$, $E(f|\mathcal{G})$ is the projection from f onto \mathcal{M} . In particular, for information denoted by W , say, we shall have constant use for a projection operator P defined by $Px = (E(x_k|W))$, for all $x = (x_k)$ in the appropriate space. We shall not in general distinguish between the vector W and the information subfield associated with it.

As stated above, we aim to quantify the amount of communication that can take place between agents. Each controller may use his control action u to pass information about ξ (and possibly η) to other agents in order that their controls are more effective. It is not surprising therefore that statistical communications theory can play a large part in solving some of the problems arising in this context. This theory includes the situation where a stationary sequence or stochastic process $\{u_i\}$ is input to a channel, which after distortion by noise gives an output sequence $\{y_i\}$. For any specified channel with input power restriction, a capacity may be defined by consideration of the number of input values that can be distinguished from the output, when the channel is operated over a long period of time. This capacity gives a limit to the information transfer that is possible between input and output.

Another branch of the theory, namely rate distortion theory deals with the case in which we wish to transmit information about some process $\{\xi_i\}$ through a channel of known capacity in such a way that the output $\{y_i\}$ may be "decoded" to give another sequence $\{w_i\}$ and some function L of $\{\xi_i\}$ and $\{w_i\}$ is thereby minimized. Typically, and in the present context, $\{w_i\}$ will be an estimate of $\{\xi_i\}$ and L will represent a function of the estimation error. The relevant parts of this theory are considerably expanded in Whittle and Rudge [7]. We require two lemmas which are essential for the main results of this paper. The proofs of these lemmas will not be given in full partly for reasons of brevity, and partly because from the outlines presented, the details may be inserted by reference to [7]. These details do appear in [5].

For fixed constants h, δ define

$$\Gamma_n = \{x: x \in \mathcal{H}_n, \|x\|_n^2 \leq h^2 \delta^2\}.$$

LEMMA 1. Let $\xi \sim N(0, \delta^2 \Sigma)$ where $\Sigma > 0$ and is $n \times n$. Let $\eta \sim N(0, V)$ with $V > 0$ and $m \times m$, and let ξ, η be independent with

$$y = f(\xi) + \eta$$

for some Borel measurable function f . Then $\exists k$ such that for all such $f(\xi) \in \Gamma_m$,

$$\|E(\xi|y)\|_n \leq k\delta^2.$$

Proof. In a communications theory setting of this problem, all the random variables would refer to single points in stationary processes.

However, any suitable function f gives an allowable input to the associated channel and so the capacity of the channel operated in this way (repeated inputs of $f(\xi)$) will not exceed the capacity using general stationary processes. Similarly, if we define a distortion measure L on the channel output by;

$$L = E\{(\xi - E(\xi|y))^T (\xi - E(\xi|y))\} = \|\xi - E(\xi|y)\|^2$$

then because we have an upper bound to the capacity, we shall obtain from rate distortion theory, a lower bound to L (since as the capacity increases more information can be transmitted and L decreases). For the special case $\Sigma = I_n$, $V = I_m$, the upper

bound to the capacity is given by

$$\frac{1}{2} \log_e (1 + h^2 \delta^2)^m$$

when the input power is restricted to $h^2 \delta^2$ (i.e. $f \in \Gamma_m$). For this capacity the minimal distortion is

$$L = \frac{n\delta^2}{(1 + h^2 \delta^2)^{m/n}}.$$

Since this is a lower bound we have

$$\begin{aligned} \|\mathbb{E}(\xi|y)\|^2 &= \|\xi\|^2 - \|\xi - \mathbb{E}(\xi|y)\|^2 \\ &\leq n\delta^2 - \frac{n\delta^2}{(1 + h^2 \delta^2)^{m/n}} \leq nmh^2 \delta^4. \end{aligned}$$

The general result is obtained from this by using the transformations

$$w = \Sigma^{-1/2} \xi, \quad \varepsilon = V^{-1/2} \eta, \quad g(\xi) = V^{-1/2} f(\xi), \quad h(w) = g(\Sigma^{1/2} w) = g(\xi),$$

so that

$$\mathbb{E}(\xi|y) = \Sigma^{1/2} \mathbb{E}(w|y) = \Sigma^{1/2} \mathbb{E}(w|h(w) + \varepsilon). \quad \square$$

LEMMA 2. Let ξ, η, f, y , be as above, with f having the form

$$f(\xi) = C\xi + h(\xi)$$

for some $m \times n$ matrix C and measurable function h . Let $\|h(\xi)\|_m = a\delta^j$ for some constants a, j with $j > 1$. Define

$$\Phi = \delta^2 \Sigma C^T (V + \delta^2 C \Sigma C^T)^{-1}.$$

Then

$$\|\mathbb{E}(\xi|y) - \Phi y\| = O(\delta^{(j+3)/2}) \quad \text{as } \delta \rightarrow 0.$$

Remark. Note that if $z = C\xi + \eta$, then $\mathbb{E}(\xi|z) = \Phi z$.

Proof. We shall give the proof only for the scalar case with $V = \Sigma = 1$ and $C = \phi$. In this case the capacity of the channel and the lower bound on L are actually attained when the input is linear in ξ . The vector version is more involved but essentially requires a transformation to simultaneously diagonal variance/covariance matrices and consideration of the channel as n parallel scalar channels.

Denote

$$g(y) = \mathbb{E}(\xi|y).$$

Let $\mathcal{O}y$ be the best linear estimate of ξ given y in that \mathcal{O} minimizes $\|\xi - \beta y\|^2$ over β . Then

$$\mathcal{O} = \frac{(\xi, f(\xi))}{\|y\|^2} = \frac{\phi\delta^2 + (\xi, h(\xi))}{1 + \|f\|^2}$$

where (\cdot, \cdot) denotes the Hilbert space inner product as before.

Now

$$\begin{aligned} \|g(y) - \mathcal{O}y\|^2 &= \|g(y)\|^2 + \mathcal{O}^2 \|y\|^2 - 2\mathcal{O}(y, g(y)) \\ &= \|g(y)\|^2 + \mathcal{O}^2 \|y\|^2 - 2\mathcal{O}(y, \xi) \end{aligned}$$

by the projection property of $g(y)$

$$\begin{aligned} &= \|g(y)\|^2 + \mathcal{O}^2 \|y\|^2 - 2\mathcal{O}(f(\xi), \xi) \\ &= \|g(y)\|^2 - \mathcal{O}^2 \|y\|^2. \end{aligned}$$

By definition, $\|f(\xi)\|^2 \leq \phi^2 \delta^2 + 2a\phi \delta^{j+1} + O(\delta^{2j})$, so that, using the linearity of optimal control for the scalar memoryless Gaussian channel [7], we have,

$$\begin{aligned} \|g(y)\|^2 &\leq \frac{(\phi^2 + 2a\phi \delta^{j-1})\delta^4}{1 + (\phi^2 + 2a\phi \delta^{j-1})\delta^2} + O(\delta^{2j+2}) \\ &= \frac{\phi^2 \delta^4}{1 + \phi^2 \delta^2} + O(\delta^{j+3}). \end{aligned}$$

Also

$$\begin{aligned} \mathcal{O}^2 \|y\|^2 &= \frac{(\phi \delta^2 + (\xi, h))^2}{1 + \|f\|^2} \geq \frac{\phi^2 \delta^4 - 2a\phi \delta^{j+3}}{1 + \phi^2 \delta^2 + 2a\phi \delta^{j+1}} + O(\delta^{2j+2}) \\ &= \frac{\phi^2 \delta^4}{1 + \phi^2 \delta^2} + O(\delta^{j+3}). \end{aligned}$$

Finally

$$\begin{aligned} \|g(y) - \Phi y\| &\leq \|g(y) - \mathcal{O}y\| + |\mathcal{O} - \Phi| \|y\| \\ &\leq |\mathcal{O} - \Phi| \|y\| + O(\delta^{(j+3)/2}) \end{aligned}$$

and

$$\begin{aligned} |\mathcal{O} - \Phi| &= \left| \frac{\phi \delta^2 + (\xi, h)}{1 + \|f\|^2} - \frac{\phi \delta^2}{1 + \phi^2 \delta^2} \right| \leq \left| \frac{\phi \delta^2 - 2a\phi \delta^{j+1}}{1 + \phi^2 \delta^2 + 2a\phi \delta^{j+1}} - \frac{\phi \delta^2}{1 + \phi^2 \delta^2} \right| \\ &= O(\delta^{j+1}) = o(\delta^{(j+3)/2}) \quad \text{for } j > 1, \end{aligned}$$

which completes the proof. \square

4. Main results. Returning to problem A, we show initially that any control law which gives a nonpositive value for J must be “small”. Precisely we have

LEMMA 3. *If a control law u gives a nonpositive cost then as $\delta \rightarrow 0$,*

$$\|u\| = O(\delta) \text{ and } |J(u)| = O(\delta^2).$$

Proof. If u gives a nonpositive cost, then

$$\begin{aligned} 0 \geq J(u) &= \frac{1}{2}(u, Qu) + (S^T \xi, u) \\ &\geq \frac{1}{2} \|Q^{1/2} u\|^2 - \|Q^{-1/2}\| \|Q^{1/2} u\| \|S^T \xi\| \\ &= \frac{1}{2} \|Q^{1/2} u\| \{ \|Q^{1/2} u\| - 2 \|Q^{-1/2}\| \|S^T \xi\| \}. \end{aligned}$$

Since $\xi \sim N(0, \delta^2 \Sigma)$, $\|S^T \xi\| = O(\delta)$ and thus

$$\|u\| \leq \|Q^{-1/2}\| \|Q^{1/2} u\| \leq 2 \|Q^{-1/2}\|^2 \|S^T \xi\| = O(\delta).$$

Also

$$|J(u)| \leq \|Q^{-1/2}\| \|Q^{1/2} u\| \|S^T \xi\| = O(\delta^2). \quad \square$$

Note that the control law $u \equiv 0$ trivially gives zero cost so that any candidates for an optimal law must obey the condition of this lemma.

We now consider two (suboptimal) linear control laws. Firstly suppose that the players disregard the η dependent part of their observations so that player k 's information is now effectively $W_k^0 = Y_k$. Since the information structure is now static in nature the optimal control u^0 , say, for this restricted problem which we call A^0 , is linear and is defined by (see [4]) the solution of

$$(2) \quad P^0 Q u^0 + P^0 w = 0, \quad P^0 u^0 = u^0,$$

where

$$w = (w_k) = (S_k^T \xi)$$

and

$$\forall x \in \mathcal{H}_r \quad P^0 x = (\mathbb{E}(x_k | W_k^0)) = (\mathbb{E}(x_k | Y_k)).$$

Note that (2) merely states the stationarity and measurability conditions, and that u^0 is a possible control law for problem A. The first result is

THEOREM 1. *For problem A defined above, all control laws u giving cost less than u^0 satisfy*

$$(3) \quad \|u - u^0\| = O(\delta^2).$$

Moreover,

$$J(u^0) - J(u) = O(\delta^4).$$

Remark. Control laws are defined as maps from the information subfield associated with the information structure, to an appropriate Euclidean space. However, since for any given control law, the information W_k , say, is a map defined on Ω , we shall use u_k to denote the function defined both on W_k and on Ω . For example in (3)

$$\|u - u^0\| \equiv \|u[W(\xi, \eta)] - u^0[Y(\xi)]\|.$$

Proof. For any u , let

$$u_k(W_k) = u_k^0(Y_k) + h_k(W_k). \quad 1 \leq k \leq N.$$

Then

$$\begin{aligned} J(u) &= \frac{1}{2}(u, Qu) + (w, u) \\ &= \frac{1}{2}(u^0, Qu^0) + (w, u^0) + \frac{1}{2}(h, Qh) + (Qu^0 + w, h) \\ &= J(u^0) + \frac{1}{2}(h, Qh) + (z^0, h), \end{aligned}$$

where

$$z^0 = Qu^0 + w.$$

Define the operator P , which depends on u , by

$$Px = (\mathbb{E}(x_k | W_k)) \quad \forall x \in \mathcal{H}_r.$$

Because of the measurability requirements on u , viz. $Pu = u$, we require also $Ph = h$. Hence, since P is Hermitian,

$$(z^0, h) = (z^0, Ph) = (Pz^0, h).$$

For each k , the distribution of ξ conditional on Y_k is Gaussian with mean $\hat{\xi}_k$, say, and var/cov matrix $\delta^2 \Gamma_k$ where

$$\hat{\xi}_k = \Sigma H_k^T (H_k \Sigma H_k^T)^{-1} Y_k, \quad \Gamma_k = \Sigma - \Sigma H_k^T (H_k \Sigma H_k^T)^{-1} H_k \Sigma.$$

Note that from Lemma 3 for each k ,

$$\left\| G_k \xi + \sum_j D_{kj} u_j \right\|_{r_k} = O(\delta).$$

Thus, using the causality condition, which implies that $\sum_j D_{kj} u_j$ is independent of η_k , and Lemma 1, we have

$$\|\mathbb{E}(\xi | W_k) - \hat{\xi}_k\|_{r_k} = O(\delta^2), \quad 1 \leq k \leq N.$$

Now z^0 is linear in ξ and $P^0 z^0 = 0$ by (2). Hence

$$\begin{aligned} \|Pz^0\|_r^2 &= \|(P - P^0)z^0\|_r^2 = \sum_k \|\mathbb{E}(z^0 | W_k) - \mathbb{E}(z^0 | Y_k)\|_{r_k}^2 \\ &\leq \text{const.} \cdot \sum_k \|\mathbb{E}(\xi | W_k) - \hat{\xi}_k\|_{r_k}^2 = O(\delta^4). \end{aligned}$$

As in lemma 3 we have

$$\begin{aligned} J(u) < J(u^0) &\Rightarrow 0 > \frac{1}{2} \|Q^{1/2} h\| \{ \|Q^{1/2} h\| - 2 \|Q^{-1/2}\| \|Pz^0\| \} \\ &\Rightarrow \|h\| \leq \text{const} \cdot \|Pz^0\| = O(\delta^2) \end{aligned}$$

i.e.

$$\|u - u^0\| = O(\delta^2).$$

Also

$$J(u^0) - J(u) \leq \|Q^{1/2} h\| \|Q^{-1/2}\| \|Pz^0\| = O(\delta^4),$$

which completes the proof. \square

Next we derive a linear control law which may be regarded as a first order approximation to the best linear law. Consider the static information structure W^1 say, defined by

$$\begin{aligned} W_k^1 &= \{Y_k, Z_k^1\}, \\ Z_k^1 &= G_k \xi + \sum_{j=1}^N D_{kj} u_j^0(Y_j) + \eta_k = J_k \xi + \eta_k \end{aligned}$$

for some matrix J_k . Let problem A^1 be that of minimizing J subject to this information structure. The optimal control law u^1 say, is again linear and is defined by the solution to

$$(4) \quad P^1 Q u^1 + P^1 w = 0, \quad P^1 u^1 = u^1,$$

where

$$\forall x \in \mathcal{X}_r \quad P^1 x = (\mathbb{E}(\dot{x}_k | W_k^1)).$$

LEMMA 4. *The control law u^1 has the form*

$$u_k^1(W_k^1) = u_k^0(Y_k) + \delta^2 N_k Z_k^1 + r_k(W_k^1), \quad 1 \leq k \leq N,$$

where the r_k are linear in W_k^1 for each k , $\|r_k(W_k^1)\| = O(\delta^3)$ and the matrices N_k are constant in δ .¹

Proof. Using the notation of Theorem 1 for the conditional distribution of ξ given Y_k , we have for each k ,

$$\mathbb{E}(\xi | W_k^1) = \hat{\xi}_k + \delta^2 \Gamma_k J_k^T (\Phi_k + \delta^2 J_k \Gamma_k J_k^T)^{-1} (Z_k^1 - J_k \hat{\xi}_k)$$

¹ The functions r_k defined in this lemma are not to be confused with the dimensions r_k of the controls u_k .

so that

$$(5) \quad \|\mathbb{E}(\xi | W_k^1) - \hat{\xi}_k - \delta^2 \Gamma_k J_k^T \Phi_k^{-1} Z_k^1\| = O(\delta^3).$$

Let

$$v_k(W_k^1) = u_k^1(W_k^1) - u_k^0(Y_k)$$

when from (2) and (4) it follows that

$$(6) \quad P^1 Q v + (P^1 - P^0) z^0 = 0.$$

But as stated above, z^0 is linear in ξ , and so using (5), we have

$$(7) \quad (P^1 - P^0) z^0 = \delta^2 x + y^1$$

where $x_k = M_k Z_k^1$ for some constant matrices M_k ($1 \leq k \leq N$), and $\|y^1\|_r = O(\delta^3)$.

Now let

$$(8) \quad v = v^1 + v^2$$

with

$$v_k^1 = -\delta^2 Q_{kk}^{-1} M_k Z_k^1, \quad 1 \leq k \leq N,$$

and let

$$(9) \quad \begin{aligned} y^2 &= P^1 Q v^1 + \delta^2 x \\ &= \left(\mathbb{E} \left(\sum_{j \neq k} \delta^2 Q_{kj} Q_{jj}^{-1} M_j Z_j^1 | W_k^1 \right) \right) = \delta^2 \left(\sum_{j \neq k} Q_{kj} Q_{jj}^{-1} M_j J_j \hat{\xi}_k \right) \end{aligned}$$

so that $\|y^2\| = O(\delta^3)$.

From (6)–(9) we have

$$P^1 Q v^2 + y^1 + y^2 = 0, \quad P^1 v^2 = v^2,$$

therefore

$$(P^1 Q v^2, v^2) = (Q v^2, v^2) = -(v^2, y^1 + y^2) \leq \|v^2\| \|y^1 + y^2\|$$

so that

$$\|v^2\|^2 \leq \|Q^{-1/2}\|^2 \|Q^{1/2} v^2\|^2 \leq \|Q^{-1/2}\|^2 \|v^2\| \|y^1 + y^2\|,$$

i.e.,

$$\|v^2\| \leq \|Q^{-1}\| \|y^1 + y^2\| = O(\delta^3).$$

Putting $N_k = -Q_{kk}^{-1} M_k$ and $r_k = v_k^2$ for $1 \leq k \leq N$ gives the result. The linearity of r_k is assured by that of u_k^1 . \square

Returning to the original problem A, consider the control law \bar{u} and associated information structure \bar{W} defined by

$$\bar{u}_k(\bar{W}_k) = u_k^0(Y_k) + \delta^2 N_k \bar{Z}_k, \quad 1 \leq k \leq N,$$

where

$$(10) \quad \begin{aligned} \bar{W}_k &= \{Y_k, \bar{Z}_k\}, \\ \bar{Z}_k &= G_k \xi + \sum_{j=1}^N D_{kj} \bar{u}_j(\bar{W}_j) + \eta_k = J_k \xi + s_k(\xi, \eta) + \eta_k \end{aligned}$$

$$(11) \quad \|s_k(\xi, \eta)\| = \left\| \sum_j D_{kj} \delta^2 N_j \bar{Z}_j \right\| = O(\delta^2).$$

THEOREM 2. For any control law u giving cost less than \bar{u} for problem A,

$$\|u(W) - \bar{u}(\bar{W})\| = O(\delta^3) \quad \text{and} \quad J(\bar{u}) - J(u) = O(\delta^6).$$

Proof. As before, let u be a control law, with associated information structure W , which gives cost less than \bar{u} . Denote by $\bar{u}(W)$ the vector of functions $\bar{u}_k(W_k)$ ($1 \leq k \leq N$) given by

$$(12) \quad \bar{u}_k(W_k) = u_k^0(Y_k) + \delta^2 N_k Z_k.$$

Note that whereas $\bar{u}(\bar{W})$ denotes a well defined control law, $\bar{u}(W)$ denotes only the functions given above with W defined by the specific control law u under consideration. Let

$$(13) \quad f_k(W_k) = u_k(W_k) - \bar{u}_k(W_k), \quad 1 \leq k \leq N,$$

and suppose that γ is such that

$$\max_k \|f_k(W_k)\|_{r_k} = O(\delta^\gamma) \quad \text{as } \delta \rightarrow 0.$$

We show initially that $\gamma \geq 2$. Thus

$$\begin{aligned} \|f(W)\| &= \|u(W) - \bar{u}(W)\| \\ &\leq \|u(W) - u^0(Y)\| + \|u^0(Y) - \bar{u}(W)\|. \end{aligned}$$

If we assume without loss of generality that $J(u) \leq J(u^0)$, the first term has order δ^2 . Since the second term has subvectors $\delta^2 N_k Z_k$, we have that $\gamma \geq 2$. Substituting for u from (13) in the cost function gives

$$(14) \quad \begin{aligned} J(u) &= \frac{1}{2}(\bar{u}(W), Q\bar{u}(W)) + (w, \bar{u}(W)) + \frac{1}{2}(f, Qf) + (Q\bar{u}(W) + w, f) \\ &= J(\bar{u}) + \frac{1}{2}(\Delta, Q\Delta) + (Q\bar{u}(\bar{W}) + w, \Delta) + \frac{1}{2}(f, Qf) + (Q\bar{u}(W) + w, f) \end{aligned}$$

where

$$\Delta = \bar{u}(W) - \bar{u}(\bar{W}) \in \mathcal{H}.$$

Corresponding to (10) but with \bar{u} replaced by u , we obtain using (12) and (13):

$$Z_k = J_k \xi + t_k(\xi, \eta) + \eta_k, \quad 1 \leq k \leq N,$$

where

$$(15) \quad \begin{aligned} t_k(\xi, \eta) &= \sum_j D_{kj}(\delta^2 N_j Z_j + f_j) \quad (1 \leq k \leq N) \\ &= O(\delta^{\min[2, \gamma]}) = O(\delta^2). \end{aligned}$$

Thus for each k ,

$$\begin{aligned} \|\Delta_k\|_{r_k} &= \delta^2 \|N_k(Z_k - \bar{Z}_k)\| \\ &= \delta^2 \|N_k(t_k - s_k)\| \\ &= \delta^2 \|N_k \sum_j D_{kj}[\delta^2 N_j(Z_j - \bar{Z}_j) + f_j]\| \\ &\leq \delta^4 \|N_k \sum_j D_{kj}(t_j - s_j)\| + \delta^2 \|N_k \sum_j D_{kj} f_j\| \\ &= O(\delta^\alpha) \quad \text{with the use of (11) and (15)} \end{aligned}$$

where

$$\alpha = \min [6, 2 + \gamma].$$

It follows that

$$(16) \quad \frac{1}{2}(\Delta, Q\Delta) + (Q\bar{u}(\bar{W}) + w, \Delta) \leq \frac{1}{2}\|Q^{1/2}\Delta\|^2 + \|Q\bar{u}(\bar{W}) + w\| \|\Delta\| = O(\delta^{\alpha+1}).$$

Also, from Lemma 4 we have

$$(17) \quad \begin{aligned} \|u_k^1(W_k^1) - \bar{u}_k(W_k)\| &= \|\delta^2 N_k(Z_k^1 - Z_k) + r_k(W_k^1)\| \\ &= \|\delta^2 N_k t_k + r_k(W_k^1)\| \quad (1 \leq k \leq N) \\ &= O(\delta^3). \end{aligned}$$

Thus

$$(18) \quad \begin{aligned} |(Q\bar{u}(W), f) - (Qu^1(W^1), f)| &\leq \|Q(\bar{u}(W) - u^1(W^1))\| \|f\| \\ &= O(\delta^{3+\gamma}). \end{aligned}$$

Now using (16) and (18) with (14), we deduce that

$$(19) \quad |J(u) - J(\bar{u}) - \frac{1}{2}(f, Qf) - (Qu^1(W^1) + w, f)| = O(\delta^\beta)$$

with

$$\beta = \min [7, 3 + \gamma].$$

Define

$$z^1 = Qu^1(W^1) + w$$

and note that $Pf = f$, where by definition

$$Px = (\mathbb{E}(x_k | W_k)) \quad \forall x \in \mathcal{H}_r.$$

Suppose that $\|Pz^1\| = O(\delta^\tau)$ for some τ as $\delta \rightarrow 0$. We need to establish a relationship between τ and γ . Lemma 4 gives

$$z_k^1 = (Qu^1 + w)_k = R_k \xi + \delta^2 \sum_j S_{kj} \eta_j, \quad 1 \leq k \leq N,$$

for some R_k, S_{kj} whose elements are $O(1)$ in δ . By repeated substitution for t_k from (15) we have

$$Z_k = J_k \xi + \bar{t}_k(\xi) + \bar{\eta}_k(\eta) + \eta_k, \quad 1 \leq k \leq N,$$

where

$$(20) \quad \bar{t}_k = \sum_j D_{kj} \left[\delta^2 N_j J_j \xi + f_j + \sum_l D_{jl} \left[\delta^2 N_l J_l \xi + f_l + \left[\dots \right] \right] \right]$$

and

$$\bar{\eta}_k = \delta^2 \sum_j D_{kj} \left[N_j \eta_j + \delta^2 \sum_l D_{jl} \left[N_l \eta_l + \dots \right] \right].$$

The causality condition implies that the summations involve only finitely many terms and that (using the independence of the η_i) $\bar{\eta}_k$ and η_k are independent. Clearly $\bar{\eta}_k$ is zero mean Gaussian. Suppose that its var/cov matrix is $\delta^2 \Xi_k$ where the elements of

Ξ_k are $O(1)$ in δ . Note that from (20);

$$\|\bar{t}_k\| = O(\delta^{\min\{3, \gamma\}}).$$

It follows from Lemma 2 that

$$\|\mathbb{E}(\xi | W_k) - \hat{\xi}_k - \delta^2 E_k(Z_k - J_k \hat{\xi}_k)\| = O(\delta^{(3+\gamma)/2})$$

where

$$E_k = \Gamma_k J_k^T (\Phi_k + \delta^2 \Xi_k + \delta^2 J_k \Gamma_k J_k^T)^{-1}.$$

Also if we use arguments similar to those in Lemma 2 (in that in the scalar case, posterior variance is maximized when the noise is Gaussian) it follows that

$$\|\mathbb{E}(\eta_k | W_k) - F_k(Z_k - J_k \hat{\xi}_k)\| = O(\delta^{(1+\gamma)/2})$$

where

$$F_k = \Phi_k (\delta^2 \Xi_k + \delta^2 J_k \Gamma_k J_k^T + \Phi_k)^{-1}.$$

Finally it is clear that $\forall j \neq k$,

$$\|\mathbb{E}(\eta_j | W_k)\| = O(\delta^2).$$

Hence

$$\begin{aligned} \|\mathbb{E}(z_k^1 | W_k)\| &\leq \|\mathbb{E}(R_k \xi + \delta^2 S_{kk} \eta_k | W_k)\| + O(\delta^4) \\ &\leq \|R_k [\mathbb{E}(\xi | W_k) - \hat{\xi}_k - \delta^2 E_k(Z_k - J_k \hat{\xi}_k)] + \delta^2 S_{kk} [\mathbb{E}(\eta_k | W_k) - F_k(Z_k - J_k \hat{\xi}_k)]\| \\ &\quad + \|R_k \hat{\xi}_k + \delta^2 (R_k E_k + S_{kk} F_k)(Z_k - J_k \hat{\xi}_k)\| + O(\delta^4) \\ &= O(\delta^{(3+\gamma)/2}) + \|R_k \hat{\xi}_k + \delta^2 (R_k E_k + S_{kk} F_k)(Z_k - J_k \hat{\xi}_k)\| + O(\delta^4). \end{aligned}$$

But

$$\begin{aligned} P^1 z^1 = 0 &\Rightarrow \mathbb{E} \left[R_k \xi + \delta^2 \sum_j S_{kj} \eta_j \middle| W_k^1 \right] = 0 \quad \text{for } 1 \leq k \leq N \\ &\Rightarrow R_k \hat{\xi}_k + \delta^2 [R_k \Gamma_k J_k^T + S_{kk} \Phi_k] (\phi_k + \delta^2 J_k \Gamma_k J_k^T)^{-1} (Z_k^1 - J_k \hat{\xi}_k) = 0 \\ &\Rightarrow \|R_k \hat{\xi}_k + \delta^2 (R_k E_k + S_{kk} F_k)(Z_k - J_k \hat{\xi}_k)\| = O(\delta^4). \end{aligned}$$

Thus

$$\|\mathbb{E}(z_k^1 | W_k)\| = O(\delta^{\min\{4, (3+\gamma)/2\}}),$$

i.e.

$$\tau \geq \min [4, (3 + \gamma)/2].$$

Considering the terms in (19) we have

$$\frac{1}{2} \langle f, Qf \rangle = O(\delta^{2\gamma})$$

and

$$(21) \quad |(z^1, f)| = |(z^1, Pf)| = |(Pz^1, f)| \leq \|Pz^1\| \|f\| = O(\delta^{\tau+\gamma}).$$

Hence

$$\begin{aligned} J(u) \leq J(\bar{u}) &\Rightarrow 2\gamma \geq \beta \text{ or } 2\gamma \geq \gamma + \min [4, (3 + \gamma)/2] \\ &\Rightarrow \gamma \geq \min [3\frac{1}{2}, (3 + \gamma)/2] \text{ or } \gamma \geq \min [4, (3 + \gamma)/2] \\ &\Rightarrow \gamma \geq 3. \end{aligned}$$

It follows that

$$\begin{aligned}\|u(W) - \bar{u}(\bar{W})\| &\leq \|u(W) - \bar{u}(W)\| + \|\Delta\| \\ &= \|f\| + O(\delta^5) = O(\delta^3)\end{aligned}$$

and from (19) and (21) that

$$J(\bar{u}) - J(u) = O(\delta^6),$$

which completes the proof. \square

The proof may be considered in a recursive fashion. Thus if $\gamma = 2$, the estimation of z^1 is such that $\|Pz^1\| = O(\delta^{5/2})$ implying that

$$0 < \frac{1}{2}(f, Qf) + (z^1, f) = O(\delta^4).$$

For this expression to be negative we require $\gamma \geq \frac{5}{2}$ but this gives $\|Pz^1\| = O(\delta^{11/4})$ and

$$0 < \frac{1}{2}(f, Qf) + (z^1, f) = O(\delta^5).$$

This argument clearly leads to the conclusion that $\gamma \geq 3$ as above.

At first sight it appears that a second approximation u^2 say to the optimal linear control law, found in the same way as u^1 , might achieve even better order of magnitude results. To see that this is in fact not so in general, define

$$W_k^2 = \{Y_k, Z_k^2\}, \quad 1 \leq k \leq N,$$

$$\begin{aligned}Z_k^2 &= G_k \xi + \sum_j D_{kj} u_j^1(W_j^1) + \eta_k \\ &= J_k \xi + \sum_j D_{kj} (\delta^2 N_j Z_j^1 + r_j(W_j^1)) + \eta_k.\end{aligned}$$

The r_j terms which have order δ^3 in norm remain an “unknown” limiting factor as shown by (17). These could be evaluated by further consideration of problem A^1 but the result of (16), which is essentially due to the f_j terms, still leads to the conclusion of (19). The results established do not mention the optimal linear control and it would indeed be interesting from a theoretical viewpoint to know how this compares with a) the linear sub-optimal laws considered and b) the “optimal” law. However, the optimal linear law would usually be extremely difficult to find since it involves the solution of nonlinear equations (a general feature of dynamic and nonnested information structures), whereas the controls considered above can in many cases be readily computed. It is important to note that application of the control rule \bar{u} does not require the solution to the static team problem A^1 . All that is required is the solution A^0 , which must be regarded as a starting point for any good sub-optimal law, and the set of matrices $\{N_k\}$, which may easily be found from (7) and (8).

5. Examples.

1. *Static Team.* This is the case where $D_{kj} = 0 \quad \forall k, j$. It is clear that u^1 is the optimal control and that $W^1 \equiv \bar{W}$. Suppose that

$$u_k^0 = \Lambda_k Y_k, \quad 1 \leq k \leq N,$$

so that

$$\begin{aligned}z_k^0 &= \sum_j Q_{kj} \Lambda_j Y_j + S_k^T \xi \quad (1 \leq k \leq N) \\ &= \Omega_k \xi, \quad \text{which defines } \Omega_k.\end{aligned}$$

From the proof of Lemma 4, noting that in this case $J_k \equiv G_k$, we have

$$N_k = -Q_{kk}^{-1} \Omega_k \Gamma_k G_k \Phi_k^{-1}, \quad 1 \leq k \leq N,$$

and

$$\bar{u}_k = \Lambda_k Y_k + \delta^2 N_k \bar{Z}_k.$$

Theorem 2 asserts that $J(\bar{u}) < J(u^1) + O(\delta^6)$. Hence by the calculation of u^0 alone we are immediately able to determine a control law \bar{u} which gives a cost within order δ^6 of the optimal. This becomes important when the dimension of the vector $(\xi^T : \eta^T)^T$ is appreciably greater than the dimension of ξ . In such cases the determination of u^1 may be unpractical or at least very time consuming, while the determination of u^0 could be straightforward.

2. *Witsenhausen's Counterexample.* This classic demonstration of nonlinearity of optimal control is described in [8], and in the present formulation is defined by

$$n = 1, \quad N = 2, \quad r_1 = r_2 = 1,$$

$$Q = \begin{pmatrix} 1+k^2 & -1 \\ -1 & 1 \end{pmatrix}, \quad S = (-k^2, 0),$$

$$Y_1 = \xi, \quad Z_1 = 0; \quad Y_2 = 0, \quad Z_2 = u_1 + \eta,$$

where $\xi \sim N(0, \delta^2)$, $\eta \sim N(0, 1)$.

Thus $w = (S_k^T \xi) = \begin{pmatrix} -k^2 \xi \\ 0 \end{pmatrix}$ and from (2) we obtain

$$u^0 = \begin{pmatrix} \mathcal{O} \xi \\ 0 \end{pmatrix}, \quad z^0 = \begin{pmatrix} 0 \\ -\mathcal{O} \xi \end{pmatrix}, \quad J(u^0) = \frac{1}{2} \mathcal{O} \delta^2, \quad \text{where } \mathcal{O} = k^2 / (1 + k^2).$$

As in the static team case we calculate using the proof of Lemma 4 that $N_1 = 0$ and $N_2 = \mathcal{O}^2$, i.e.

$$\bar{u} = \begin{pmatrix} \mathcal{O} \xi \\ \delta^2 \mathcal{O}^2 \bar{Z}_2 \end{pmatrix} = \begin{pmatrix} \mathcal{O} \xi \\ \delta^2 \mathcal{O}^2 (\mathcal{O} \xi + \eta) \end{pmatrix}.$$

It follows that $J(\bar{u}) = \frac{1}{2} (\mathcal{O} \delta^2 - \mathcal{O}^4 \delta^4 + \mathcal{O}^6 \delta^6)$. The optimal control law is unknown but we can find the optimal linear law u^L , say, which is given by

$$u^L = \begin{pmatrix} \lambda \xi \\ \mu (\lambda \xi + \eta) \end{pmatrix} \quad \text{where } \mu = \delta^2 \lambda^2 / (1 + \delta^2 \lambda^2)$$

and λ is a root of

$$(1 - \lambda)(1 + \lambda^2 \delta^2)^2 - \lambda / k^2 = 0.$$

For small δ , we have for some α ,

$$\lambda = \mathcal{O} + 2\mathcal{O}^3(1 - \mathcal{O})\delta^2 + \alpha\delta^4 + O(\delta^6)$$

and

$$J(u^L) = \frac{1}{2} [\mathcal{O} \delta^2 - \mathcal{O}^4 \delta^4 + \mathcal{O}^6 \delta^6 - 4(1 - \mathcal{O}) \mathcal{O}^6 \delta^6] + O(\delta^8),$$

verifying that up to order δ^6 , \bar{u} is as good as u^L . If we define successive approximations \bar{u}^j ($j = 1, 2, \dots$) to u^L by assuming the static information structures $Z_2^j = \bar{u}_1^{j-1} + \eta$, we obtain

$$\bar{u}^j = \begin{pmatrix} \gamma_j \xi \\ \beta_j (\gamma_j \xi + \eta) \end{pmatrix}, \quad j = 1, 2, \dots$$

where

$$\beta_j = \delta^2 \gamma_j \gamma_{j-1} / (1 + \delta^2 \gamma_{j-1}^2), \quad \gamma_{j+1} = \frac{k^2}{k^2 + 1 / (1 + \gamma_j^2 \delta^2)}, \quad \gamma_0 = 0.$$

Suppose that for some $j > 0$, $\gamma_j = \mathcal{O} + O(\delta^2)$. Then for some constants α_j we have

$$\gamma_{j+1} = \mathcal{O} + \mathcal{O}^3(1 - \mathcal{O})\delta^2 + \alpha_j \delta^4 + O(\delta^6).$$

Thus, since $\gamma_1 = \mathcal{O}$, this holds $\forall j > 0$, implying that

$$J(\bar{u}^j) = \frac{1}{2} [\mathcal{O}\delta^2 - \mathcal{O}^4\delta^4 + \mathcal{O}^6\delta^6 - 3(1 - \mathcal{O})\mathcal{O}^6\delta^6] + O(\delta^8).$$

We see that these approximations improve the term in δ^6 but do not achieve any better order of magnitude results than \bar{u} . The same results hold if we put

$$\bar{u}_2^j = \mathbb{E}(\bar{u}_1^j | \bar{u}_1^j + \eta) = \mathbb{E}(\gamma_j \xi | \gamma_j \xi + \eta) = \frac{\delta^2 \gamma_j^2}{1 + \gamma_j^2 \delta^2} (\gamma_j \xi + \eta).$$

6. Discussion. Many results for LQG systems hold when some of these conditions are relaxed. It is clear that in the present case, all three conditions are required unless substantially more work is done. The Linear-Gaussian character is essential to Lemmas 1 and 2 while the quadratic cost function lends itself naturally to the Hilbert space formulation as in Lemma 3. It is possible, however, to relax some of the independence conditions on ξ and η . Consider a subset of controllers who are in reality a single player effecting controls at different points in time. If this player has perfect memory as is often supposed, then the η_k corresponding to his various controls will be nonindependent subvectors of η . This nonindependence does not affect the results established because there is no communication problem in such situations and we merely need to take more specific account of some of the η_j in the section of Theorem 2 between (20) and (21). It is reasonable to suppose that the independence of the η_k can be relaxed in a more general fashion but this would require proof of the results of Lemmas 1 and 2 for such cases.

The generality of the above theory is limited by the specification of the information pattern described by (1) and it is of interest to consider to which other situations similar results might apply. Suppose that there are two controllers, the first of whom has information as in (1) while the second has the following information.

$$Y_2 = \begin{pmatrix} H_1 \\ H_2 \end{pmatrix} \xi, \quad Z_2 = G_2 \xi + D_{21} u_1 + \varepsilon$$

where

$$\varepsilon \sim N(0, \delta^2 \theta), \quad \theta \geq 0.$$

If player 1 neglects his “poor” observation Z_1 then his control can be deduced by player 2 and there is no need for communication—the optimal control law being linear as a result. One would expect $\|u_1(Y_1) - u_1(Y_1, Z_1)\|$ to be small, leading perhaps to results as before. However, this small change in control by player 1 may lead to a relatively large change in player 2’s information if, for example, we have the limiting case $G_2 = 0$, $\theta = 0$. A lower bound on the optimal cost can be obtained by including Z_1 in player 2’s information (leading to linear optimal controls) but while in this way a result such as Theorem 1 may be shown, it appears that iterative procedures along the lines considered above, are inappropriate here.

Acknowledgments. Thanks are due to a referee for helpful comments on an earlier draft.

REFERENCES

- [1] K. C. CHU AND Y.-C. HO, *On the generalized linear-quadratic-Gaussian problem*, Differential Games and Random Topics, H. Kuhn and G. Szego, eds., North-Holland, Amsterdam, 1971.
- [2] Y.-C. HO AND K. C. CHU, *Team decision theory and information structures in optimal control problems*, IEEE Trans. Automatic Control, AC-17 (1972), pp. 15–28.
- [3] R. RADNER, *Team decision problems* Ann. Math. Statist., 33 (1962), pp. 857–881.
- [4] J. F. RUDGE, *Series solutions to static team control problems*, Math. of Operations Res., 1 (1976), pp. 67–81.
- [5] ———, *Optimisation of multi-agent stochastic control*, Ph.D. thesis, University of Cambridge, England, 1976.
- [6] N. R. SANDELL AND M. ATHANS, *Solution of some nonclassical LQG stochastic control problems*, IEEE Trans. Automatic Control, AC-19 (1974), pp. 108–116.
- [7] P. WHITTLE AND J. F. RUDGE, *Simultaneous communication and control*, Advances in Appl. Probability, 8 (1976), pp. 365–384.
- [8] H. S. WITSENHAUSEN, *A counterexample in stochastic optimum control*, this Journal, 6 (1968), pp. 131–147.
- [9] ———, *On Information structures, feedback and causality*, this Journal, 9 (1971), pp. 149–160.
- [10] ———, *A standard form for sequential stochastic control*, Math. Systems Theory, 7 (1973), pp. 5–11.
- [11] ———, *The Intrinsic Model for Discrete Stochastic Control, Some Open Problems*, Lecture Notes in Economics and Mathematical Systems, vol. 107, Springer-Verlag, New York, 1975.

ON THE RELATIONSHIP BETWEEN FIRST AND SECOND ORDER CONTROLLABLE SYSTEMS IN BANACH SPACES*

ROBERTO TRIGGIANI†

Abstract. Approximate controllability properties of the 'abstract wave equation' $\mathcal{S}: \ddot{x} = Ax + Bu$ on $X \times X$ and of the 'abstract heat equation' $\mathcal{F}: \dot{x} = Ax + Bu$ on X are compared. It is shown that there exists a dense subspace X_0 of X , explicitly exhibited, such that, with $BU \subset X_0$, \mathcal{S} is approximately controllable on any finite time interval $[0, T]$ if and only if so is \mathcal{F} . Moreover, this is the case if and only if a suitable extension of the familiar finite dimensional rank condition holds.

1. Introduction and summary of results. Consider the abstract control systems

$$\mathcal{S}: \ddot{x} = Ax + Bu \quad (\mathcal{S}_m: \ddot{x} = Ax + \sum_{i=1}^m b_i u_i, b_i \in X, u_i = \text{scalar}),$$

$$\mathcal{F}: \dot{x} = Ax + Bu \quad (\mathcal{F}_m: \dot{x} = Ax + \sum_{i=1}^m b_i u_i, b_i \in X, u_i = \text{scalar}).$$

(\mathcal{S} and \mathcal{F} stand for second order and first order system, respectively), where both X and U are separable Banach spaces and B is a bounded linear operator from U into X . Unless otherwise stated, X is always infinite dimensional. \mathcal{S}_m and \mathcal{F}_m refer to the case when $\dim U = m$, or more generally, $\dim BU = m$, $BU = \text{range of } B$. When $m = 1$, we shall write b instead of b_1 . We shall also employ the notation $\mathcal{S}(A, B)$, $\mathcal{S}_m(A, (b_1, \dots, b_m))$, etc. to emphasize a particular pair $\langle A, B \rangle$ and $\langle A, (b_1, \dots, b_m) \rangle$ under consideration. The operator A is assumed throughout to satisfy the following assumption.

H1. A is (closed, linear, with domain $D(A)$ dense in X and range in X and) the infinitesimal generator of a strongly continuous cosine function $C(t)$ of bounded linear operators in X , $-\infty < t < \infty$.

Such assumption is necessary and sufficient for the homogeneous second order system to be uniformly well posed on the real line and of type $\leq \omega_0$ [5, I, Thm. 5.9, p. 91]. See, e.g., [5], [10]–[12], [16] and the references cited therein for the necessary background on the theory of abstract cosine functions.

A main theme of the present paper is a comparison between controllability properties of \mathcal{S} and \mathcal{F} (for precise definitions see § 2), which was already carried out in [4] by Fattorini. He showed the following results (using our terminology):

1) approximate controllability in finite time for \mathcal{S} is always stronger than approximate controllability in finite time for \mathcal{F} , while these two properties are equivalent if a certain assumption H2 (below) regarding the spectrum of A holds. The assumption in question is

H2. There exists a simple closed curve Γ entirely contained in $\rho_0(A)$ —the connected component of $\rho(A)$ that contains the half plane $\text{Re } \lambda > \omega_0^2$ —such that the origin is contained in the interior of Γ .

* Received by the editors February 25, 1977, and in revised form January 30, 1978.

† Mathematics Department, Iowa State University, Ames, Iowa 50011. The first draft of this paper was completed during a one-month stay (August 1975) at the Mathematics Research Center, University of Wisconsin, Madison, supported by United States Army under contract DAAG29-75-C0024. Preliminary versions of this paper were presented at, and have appeared in, the Proceedings of the 7th IFIP Working Conference held at the University of Rome, Italy, June 21–24, 1976 and at the 14th Allerton Conference (special session on Infinite Dimensional and Time Delay Systems) of the University of Illinois of Urbana-Champaign held on September 29, 30, 1976. See [23].

2) If H2 is not satisfied, the above equivalence may fail to hold. In fact, Fattorini provided the following example: Let $X = L_2[-\infty, \infty]$; A be the self-adjoint operator defined by $(Af)(\xi) = f''(\xi) + rf(\xi)$ with $D(A) = \{f \in X : f'' \in X\}$ (f'' understood in the sense of distributions), $r \geq 0$; $b_1(\xi) = e^{-|\xi|}$, and $b_2(\xi) = b_1(\xi + 1)$. Then, (i) assumption H2 is not satisfied for the spectrum $\sigma(A)$ of A is $\sigma(A) = (-\infty, r]$; (ii) $\mathcal{F}_2\langle A, (b_1, b_2) \rangle$ is approximately controllable in finite time, while (iii) $\mathcal{S}_2\langle A, (b_1, b_2) \rangle$ is not.

On the basis of the above results of Fattorini, therefore, one may be led to conjecture that assumption H2 is a necessary condition for the implication: " \mathcal{F} approximately controllable in finite time $\Rightarrow \mathcal{S}$ approximately controllable in finite time" to hold.

As the present paper will show, however, it turns out that Fattorini's result is crucially related to his approach, which is based on the intermediary action between \mathcal{F} and \mathcal{S} played by the resolvent operator $R(\lambda, A)$ of A , and it is expressed by

$$(1.1) \quad R(\lambda^2, A)x = \frac{1}{\lambda} \int_0^\infty e^{-\lambda t} C(t)x \, dt = \int_0^\infty e^{-\lambda^2 t} S(t)x \, dt, \quad \operatorname{Re} \lambda > \omega_0, \quad x \in X.$$

Here $S(t)$ is the semigroup associated with A (see § 2).

In the present paper we take a different approach, i.e., we work in the t domain rather than in the λ -domain, and we obtain new results which complement Fattorini's. In particular, we show (Theorem 2.3) that, *regardless of assumption H2*, there is a *dense subspace* X_0 of X , in fact $X_0 = \bigcup_{0 < t} S(t)X$, such that, with the range BU of B restricted to lie in X_0 , the system $\mathcal{S}\langle A, B \rangle$ is approximately controllable on an *arbitrary* interval $[0, T]$ if and only if the same property holds for $\mathcal{F}\langle A, B \rangle$. Moreover, this is the case if and only if a suitable extension of the familiar finite dimensional rank condition holds.

After the first draft of the present paper was completed (during a one-month stay in August 1975 at the Mathematics Research Center, University of Wisconsin, Madison), H. O. Fattorini kindly brought to our attention reference [19] by Tsujioka, which tackles a similar problem with a slightly different terminology from ours. While an instructive comparison between Tsujioka's results and those of the present paper is carried out in [23], it suffices to say here that: (1) in the present paper we treat the most general operator A for which the problem under consideration makes sense on an arbitrary Banach space (i.e. an operator A satisfying H1); in [19] instead, A is a self-adjoint operator on a Hilbert space; (2) our approach—which appears to us more natural—leaves the second order equation as such and employs the cosine operator theory; in [19] instead, the second order equation is rewritten as first order equation in a *suitable* cross product space, and the theory of group of operators is employed.

2. Comparison between approximate controllability of \mathcal{S} and \mathcal{F} . In what follows, we shall take for both \mathcal{F} and \mathcal{S} *zero* initial conditions without further mention.

The mild solution of the Cauchy problem associated with the system \mathcal{S} is, by definition,

$$(2.1) \quad x_s(t, u) = \int_0^t \left(\int_0^{t-\tau} C(s)Bu(\tau) \, ds \right) d\tau, \quad t \geq 0,$$

$$(2.2) \quad \dot{x}_s(t, u) = \int_0^t C(t-\tau)Bu(\tau) \, d\tau,$$

which make sense for any Bochner integrable (locally L_1) abstract function $u(\cdot)$. In particular, if $u(\cdot)$ is C^1 , the mild solution is indeed the strict solution (twice strongly continuously differentiable) of \mathcal{S} [4]. It is a fundamental fact—which will be exploited throughout in the sequel—that an operator A satisfying assumption H1 automatically

generates a strongly continuous (C_0) semigroup $S(t)$ of bounded linear operators on X , $t \geq 0$, which may be extended analytically into the half-plane $\operatorname{Re} t > 0$. Moreover, $S(t)$ is given explicitly for $t > 0$ by

$$(2.3) \quad S(t)x = \frac{1}{\sqrt{\pi t}} \int_0^\infty e^{-\tau^2/(4t)} C(\tau)x \, d\tau, \quad t > 0, \quad x \in X$$

[5, Eq. (5.17)], [11]. Hence the mild solution of the Cauchy problem associated with the system \mathcal{F} is by definition

$$(2.4) \quad x_f(t, u) = \int_0^t S(t-\tau)Bu(\tau) \, d\tau, \quad t \geq 0,$$

for any Bochner integrable $u(\cdot)$. In particular, if $u(\cdot)$ is Hölder continuous, the mild solution is indeed the strict solution (strongly continuously differentiable of \mathcal{F} , [9, p. 491]. As in Fattorini [4], we wish to compare the controllability properties of \mathcal{F} and \mathcal{S} according to the following definitions. Let $K_t(\mathcal{S})$ [resp. $K_t(\mathcal{F})$] be the set of attainability from the origin of \mathcal{S} [resp. \mathcal{F}] i.e., the linear subspace of $X \times X$ [resp. of X] consisting of mild solution pairs $(x_s(t, u), \dot{x}_s(t, u))$ [resp. mild solution point $x_f(t, u)$] when u runs over $L_1[[0, t], U]$. The space $X \times X$ of all pairs (x_1, x_2) of elements of X is endowed with pointwise operations and with norm $\|(x_1, x_2)\| = \{\|x_1\|^2 + \|x_2\|^2\}^{1/2}$, so that [9, p. 164] $(X \times X)^* = X^* \times X^*$ (as Banach spaces).

We then say that \mathcal{S} is approximately controllable on $[0, T]$, $0 < T < \infty$ (respectively, in finite time), in case $\bar{K}_T(\mathcal{S}) = X \times X$ (respectively, $\bigcup_{0 < t < T} K_t(\mathcal{S}) = X \times X$). Similarly, we say that \mathcal{F} is approximately controllable in $[0, T]$ in case $\bar{K}_T(\mathcal{F}) = X$. It is plain from (2.1) and (2.4) that the closure of the set of attainability $K_t(\cdot)$ does not change if, instead of taking $L_1[[0, t], U]$ -control functions, one restricts to any other linear class of control functions which are dense in $L_1[[0, t], U]$.

It is a standard fact (deducible from (2.5) below) that, since $S(t)$ is analytic for $t > 0$, we have $\bar{K}_T(\mathcal{F}) = \bigcup_{0 < t < T} K_t(\mathcal{F})$, with T arbitrary, $0 < T < \infty$, so that approximate controllability in $[0, T]$ and in finite time are the same concept for \mathcal{F} . We shall therefore often omit the time length specification for \mathcal{F} and talk only about its approximate controllability. This need not be true for \mathcal{S} , for, in general, we only have $\bar{K}_T(\mathcal{S}) \subset \bigcup_{0 \leq t \leq T} K_t(\mathcal{S})$.¹

Finally, we notice that exact controllability in finite time: $\bigcup_{0 \leq t \leq T} K_t(\mathcal{S}) = X \times X$ of the strict solution of \mathcal{S} is out of the question, when A is unbounded. This is so since the strict solution always lies in $D(A)$, which is never all of X , by the closed graph theorem. Similarly for \mathcal{F} . However, *even exact controllability in finite time of the mild solution of \mathcal{S} and \mathcal{F} can be excluded when the operator B is compact*. This in particular applies to the systems \mathcal{S}_m and \mathcal{F}_m for any finite m . The proof for \mathcal{S} follows along the same lines developed in [20, Remark 3.32] or [22] for \mathcal{F} , and is therefore omitted.

We now collect below a series of elementary but important results to be used in the sequel.

Claim (i). It follows easily, via (2.4) and a direct application of a standard consequence [8, p. 31] of the Hahn–Banach theorem, that \mathcal{F} (resp. \mathcal{F}_m) is approximately controllable in $[0, T]$ if and only if: $x^* \in X^*$

$$(2.5) \quad \left\{ \begin{array}{l} x^*(S(t)BU) = 0, \\ 0 \leq t \leq T, \end{array} \right\} \Rightarrow x^* = 0; \quad \left(\text{resp.} \left\{ \begin{array}{l} x^*(S(t)b_i) = 0, \\ 0 \leq t \leq T; i = 1, \dots, m, \end{array} \right\} \Rightarrow x^* = 0 \right).$$

For an equivalent formulation see [6, Prop. 2.1].

¹ E.g. use the proposition of [25] to select a vector b such that, in the notation of [25], $b \in X_m(T)$, but $b \notin X_m$. A specific example for $X = L_2(R)$, $A = d^2/d\xi^2$ and $T = 1$ is $\hat{b}(\omega) = 1/(1 + \omega^2)$.

Remark 2.1. It is plain from the characterization (2.5) that $\mathcal{F}\langle A, B \rangle$ is approximately controllable if and only if $\mathcal{F}\langle A, S(\bar{t})B \rangle$ is approximately controllable for any $\bar{t} \geq 0$.

Claim (ii). Similarly, via (2.1) and (2.2) \mathcal{S} (resp. \mathcal{S}_m) is approximately controllable on $[0, T]$, or else in finite time, if and only if: $x_1^*, x_2^* \in X^*$

$$(2.6) \quad \begin{cases} \int_0^t x_1^*(C(\tau)BU) d\tau + x_2^*(C(t)BU) \equiv 0, \\ 0 \leq t \leq T, \text{ or else } t \geq 0, \end{cases}$$

$$\left(\text{resp: } \begin{cases} \int_0^t x_1^*(C(\tau)b_i) d\tau + x_2^*(C(t)b_i) \equiv 0, \quad i = 1, \dots, m; \\ 0 \leq t \leq T, \text{ or else } t \geq 0, \end{cases} \right)$$

$$\text{imply } x_1^* = x_2^* = 0.$$

For an equivalent formulation see [4, Lemma 2.1].

Before illustrating the use of the above characterizations, we need a definition. Throughout this paper, a crucial role will be played by the set $X_0 \stackrel{\text{def}}{=} \bigcup_{0 < t} S(t)X$. X_0 is a subspace dense in X [8, p. 208]. That $b \in X_0$ means therefore that $b = S(\tau)\beta$, for some $\tau > 0$ and $\beta \in X$.

Two motivating illustrations on infinite spatial domains on the direct use of characterization (2.6). We wish to show that characterization (2.6) can be profitably used to derive in turn easy-to-check-tests in nontrivial, physically significant cases.

Example 2.1. Consider the same space X and the same operator A as in Fattorini's example reported in the Introduction. However, this time we choose the vectors b_i in a quite different way. Let us impose that the vectors b_i lie in the dense subspace X_0 defined above, i.e., let $b_i = S(\tau_i)\beta_i$ for some $\tau_i > 0$ and $\beta_i \in X$, $i = 1, \dots, m$. By using characterization (2.6) we shall now show that:

(a) The minimal number m of scalar controls which make the system $\mathcal{S}_m \langle A, (b_1, \dots, b_m) \rangle$ approximately controllable on an arbitrary interval $[0, T]$ is two; moreover,

(b) with $m = 2$, this is the case for $\mathcal{S}_2 \langle A, (b_1, b_2) \rangle$ if and only if

$$(A) \quad \hat{\beta}_1(\omega)\hat{\beta}_2(-\omega) - \hat{\beta}_1(-\omega)\hat{\beta}_2(\omega) \neq 0, \quad \text{a.e. in } \omega \geq 0,$$

which is equivalent to

$$\hat{b}_1(\omega)\hat{b}_2(-\omega) - \hat{b}_1(-\omega)\hat{b}_2(\omega) \neq 0, \quad \text{a.e. in } \omega \geq 0,$$

where

$$\hat{f}(\omega) = \text{l.i.m.}_{N \rightarrow \infty} (2\pi)^{-1/2} \int_{|\xi| \leq N} e^{i\omega\xi} f(\xi) d\xi$$

is the Fourier-Plancherel transform (isometric isomorphism of $L_2[-\infty, \infty]$ onto itself [26, Cor. V1, 2, p. 154]).

For $b_i = S(\tau_i)\beta_i$ as assumed, we have $\hat{b}_i(\omega) = e^{-\omega^2\tau_i}\hat{\beta}_i(\omega)$, so the above stated equivalence is checked directly.

Let us apply the Fourier-Plancherel transform with respect to the space coordinate to the homogeneous second order system with initial position equal to f and zero initial velocity, whose solution is therefore $x(t) = C(t)f$, when $f \in D(A)$. Since

$(\hat{A}f)(\omega) = -\omega^2 \hat{f}(\omega)$, standard computations yield $[\widehat{C(t)f}](\omega) = \cos(g(\omega)t) \hat{f}(\omega)$ where, for convenience, we set $\sqrt{r-\omega^2} = g(\omega)$. [Notice that for $r=0$, we get $[\widehat{C(t)f}](\omega) = \cos(\omega t) \hat{f}(\omega)$]. Then

$$\left[\int_0^t C(\tau) f d\tau \right] (\omega) = \int_0^t [\widehat{C(\tau)f}](\omega) d\tau = \frac{\sin h g(\omega)t}{g(\omega)} \hat{f}(\omega).$$

Since the Fourier–Plancherel transform defines an isometric isomorphism of X onto itself, the characterization (2.6) is equivalent to: $\hat{x}_1, \hat{x}_2 \in X$

$$\left(\hat{x}_1, \int_0^t C(\tau) b_i d\tau \right) + (\hat{x}_2, \widehat{C(t)b_i}) = 0, \quad i = 1 \cdots m, \quad 0 \leq t \leq T \Rightarrow \hat{x}_1 = \hat{x}_2 = 0,$$

i.e. to

$$\begin{aligned} \int_{-\infty}^{\infty} \left[\bar{x}_1(\omega) \frac{\sin h g(\omega)t}{g(\omega)} \hat{b}_i(\omega) + \bar{x}_2(\omega) \cos h g(\omega)t \hat{b}_i(\omega) \right] d\omega &\equiv 0 \\ \Rightarrow \hat{x}_1(\omega) &\equiv 0 \text{ and } \hat{x}_2(\omega) \equiv 0 \quad \text{a.e. in } -\infty < \omega < \infty, \\ i &= 1, \dots, m, \quad 0 \leq t \leq T, \end{aligned}$$

Split $\int_{-\infty}^{\infty} = \int_{-\infty}^0 + \int_0^{\infty}$ and change ω into $-\omega$ in $\int_{-\infty}^0$, so that the above identity can be written as

$$\begin{aligned} \int_0^{\infty} e^{-\omega^2 \tau_i} \left\{ \left[\frac{\sin h g(\omega)t}{g(\omega)} \bar{x}_1(\omega) + \cos h g(\omega)t \bar{x}_2(-\omega) \right] \hat{\beta}_i(-\omega) \right. \\ \left. + \left[\frac{\sin h g(\omega)t}{g(\omega)} \bar{x}_1(\omega) + \cos h g(\omega)t \bar{x}_2(\omega) \right] \hat{\beta}_i(\omega) \right\} d\omega &\equiv 0. \end{aligned}$$

Differentiating in t under the integral sign (which is legal by [24, p. 59]) and setting $t=0$ at each stage yields

$$\int_0^{\infty} (\sqrt{r-\omega^2})^{2n} e^{-\omega^2 \tau_i} [\bar{x}_1(-\omega) \hat{\beta}_i(-\omega) + \bar{x}_1(\omega) \hat{\beta}_i(\omega)] d\omega \equiv 0, \quad n = 0, 1, \dots,$$

as well as

$$\int_0^{\infty} (\sqrt{r-\omega^2})^{2n} e^{-\omega^2 \tau_i} [\bar{x}_2(-\omega) \hat{\beta}_i(-\omega) + \bar{x}_2(\omega) \hat{\beta}_i(\omega)] d\omega \equiv 0.$$

These identities, by virtue of the completeness of $\{\xi^n e^{-\xi}\}$ in $L_2[0, \infty]$ imply (modulo a change of variable), respectively,

$$(\neq) \quad \bar{x}_1(-\omega) \hat{\beta}_i(-\omega) + \bar{x}_1(\omega) \hat{\beta}_i(\omega) \equiv 0 \quad \text{and} \quad \bar{x}_2(-\omega) \hat{\beta}_i(-\omega) + \bar{x}_2(\omega) \hat{\beta}_i(\omega) \equiv 0, \\ \text{a.e. in } \omega \geq 0,$$

[18, p. 107] (the particular value of $\tau_i > 0$ is immaterial). Summing up and subtracting yields for $i = 1, \dots, m$:

$$\begin{aligned} \text{(B)} \quad \text{(i)} \quad & [\bar{x}_1(-\omega) + \bar{x}_2(-\omega)] \hat{\beta}_i(-\omega) + [\bar{x}_1(\omega) + \bar{x}_2(\omega)] \hat{\beta}_i(\omega) \equiv 0 \\ \text{(ii)} \quad & [\bar{x}_1(-\omega) - \bar{x}_2(-\omega)] \hat{\beta}_i(-\omega) + [\bar{x}_1(\omega) - \bar{x}_2(\omega)] \hat{\beta}_i(\omega) \equiv 0 \end{aligned} \quad \text{a.e. in } \omega \geq 0.$$

It is now readily seen that, for $m=1$, the above identities (B) do not imply $x_1(\omega) \equiv 0$ and $x_2(\omega) \equiv 0$ a.e. in $-\infty < \omega < \infty$. However for $m=2$, (B) (i) and (B) (ii) (each written as a system) imply, respectively

$$x_1(\omega) + x_2(\omega) \equiv 0, \quad x_2(\omega) - x_1(\omega) \equiv 0, \quad \text{a.e. in } -\infty < \omega < \infty$$

if and only if (A) holds. Therefore, the desired conclusion $x_1(\omega) \equiv 0$ and $x_2(\omega) \equiv 0$ a.e. in $-\infty < \omega < \infty$ is achieved if and only if (A) holds. By virtue of characterization (2.6), claims (a) and (b) are thus proved. Q.E.D.

Example 2.2. Let now $X = L_2[0, \infty]$, and A be the self-adjoint [3, p. 1384] operator defined by

$$(Af)(\xi) = f''(\xi) + rf(\xi); \quad D(A) = \{f \in X : f'' \in X, f(0) = 0\}$$

with $r \geq 0$. The spectrum $\sigma(A)$ of A is again $\sigma(A) = (-\infty, r]$, and so assumption H2 is again violated. We choose a vector b to again lie in the dense subspace X_0 defined after Claim (ii), i.e., we let $b = S(\tau)\beta$ for some $\tau > 0$ and $\beta \in X$.

We shall now show by means of the characterization (2.6) that: the system $\mathcal{S}\langle A, b \rangle$ is approximately controllable on an arbitrary interval $[0, T]$ if and only if:

$$\tilde{b}(\omega) \neq 0 \text{ a.e. in } \omega \geq 0 \text{ or, equivalently: } \tilde{\beta}(\omega) \neq 0 \text{ a.e. in } \omega \geq 0, \text{ where } \tilde{f}(\omega) = \text{l.i.m.}_{N \rightarrow \infty} (2/\pi)^{1/2} \int_0^N \sin \omega \xi f(\xi) d\xi$$

is the Fourier sine transform (isometric isomorphism of $L_2[0, \infty]$ onto itself [3, p. 1388]). A proof is through a parallel development of the previous Example 2.1. Apply the Fourier sine transform with respect to the space coordinate to the same homogeneous second order Cauchy problem; since $(\tilde{A}f)(\omega) = -\omega^2 f$ also for the Fourier sine transform [3, p. 1388] for $f \in D(A)$, one finds again $[C(t)f](\omega) = \cosh(g(\omega)t)\tilde{f}(\omega)$ with $g(\omega) = \sqrt{r - \omega^2}$ as in the previous example. Since the Fourier sine transform defines an isometric isomorphism of X onto itself, the characterization (2.6) is equivalent to: $\tilde{x}_1, \tilde{x}_2 \in X$

$$\left(\tilde{x}_1, \int_0^t \widetilde{C(\tau)b} d\tau \right) + (\tilde{x}_2, \widetilde{C(t)b}) = 0, \quad 0 \leq t \leq T, \Rightarrow \tilde{x}_1 = \tilde{x}_2 = 0.$$

For $b = S(\tau)\beta$ as assumed, we have $\tilde{b}(\omega) = e^{-\omega^2\tau}\tilde{\beta}(\omega)$ and the above identity becomes

$$\int_0^\infty e^{-\omega^2\tau} \left[\tilde{x}_1(\omega) \frac{\sinh(g(\omega)t)}{g(\omega)} + \tilde{x}_2(\omega) \cosh(g(\omega)t) \right] \tilde{\beta}(\omega) d\omega = 0, \quad 0 \leq t \leq T.$$

By proceeding exactly as in the previous example one arrives at

$$\tilde{x}_1(\omega)\tilde{\beta}(\omega) \equiv 0 \quad \text{and} \quad \tilde{x}_2(\omega)\tilde{\beta}(\omega) \equiv 0 \quad \text{a.e. in } \omega \geq 0,$$

which are the analogous counterpart of ($\#$) in Example 2.1. The above identities imply $\tilde{x}_1(\omega) \equiv 0$ and $\tilde{x}_2(\omega) \equiv 0$ a.e. in $\omega \geq 0$ if and only if $\tilde{\beta}(\omega) \neq 0$ a.e. in $\omega \geq 0$. Our claim is proved. Q.E.D.

Comments on Examples 2.1 and 2.2. These two examples, while indicating a general procedure on how to make direct use of characterization (2.6), show something more. In fact, the necessary and sufficient conditions for approximate controllability of these two second order systems on any $[0, T]$ (as well as the minimum number of scalar controls needed) turn out to be exactly the very same for approximate controllability of the corresponding first order systems. (Results for these first order systems were first derived in [6] by means of the ordered representation theory of a Hilbert space, and then rederived in [2.1, § 3.1] by using a suitable extension of the familiar rank condition for controllability of finite dimensional systems. The approach followed above in analyzing these second order systems is in the spirit of the derivation as in [21]). In other words: with the vectors b_i restricted to the subspace $X_0 = \bigcup_{0 < t} S(t)X$, dense in X , the wave equation $\mathcal{S}_m\langle A, (b_1, \dots, b_m) \rangle$ is approximately controllable on an arbitrary interval $[0, T]$ in both examples if and only if the corresponding heat equation $\mathcal{F}_m\langle A, (b_1, \dots, b_m) \rangle$ is also approximately controllable

on $[0, T]$, despite the fact that the operator A in both cases fails to satisfy Fattorini's assumption H2. That this behavior is not a coincidence pertaining only to a few examples will be shown in Theorem 2.3, part (b), below. Its proof will be the abstract analogous version of the procedure employed in the two special cases above. We begin with a lemma.

LEMMA 2.1. *Under assumption H1, the following holds for any $y \in X$ and T arbitrary and finite:*

$$\overline{\text{sp}} \{C(t)y, -\infty < t < \infty\} = \overline{\text{sp}} \{S(t)y, 0 \leq t < \infty\} = \overline{\text{sp}} \{S(t)y, 0 \leq t \leq T\},$$

where here and hereafter sp denotes 'span'.

Proof. $C(t)$ is an even function of t [5], [12]. By a consequence of the Hahn-Banach theorem, [8, p. 31], all we have to show to justify the first equality is that if $x^* \in X^*$ and $x^*(S(t)y) = 0, t \geq 0$, then $x^*(C(t)y) = 0, t \geq 0$, and conversely. Both implications follow from (1.1) (or (2.3)) by virtue of the uniqueness of the Laplace transform [3, p. 626]. The second equality is due to the analyticity of $S(t)$. Q.E.D.

For later reference we label another condition: $x^* \in X^*$,

$$(2.7) \quad x^*(C(t)BU) = 0, \quad 0 \leq t < \infty, \quad \Rightarrow x^* = 0.$$

Now, let \mathcal{F} fail to be approximately controllable in $[0, T]$, hence in finite time. Then, by claim (i) above, $\bar{x}^*(S(t)BU) = 0, t \geq 0$, for some nonzero $\bar{x}^* \in X^*$. By Lemma 2.1, $\bar{x}^*(C(t)BU) = 0, t \geq 0$ and therefore $\int_0^t \bar{x}^*(C(\tau)BU) d\tau = 0, t \geq 0$ and the left hand side of implication (2.6) is violated with $x_1^* = x_2^* = \bar{x}^* \neq 0$. We have thus established the following result, whose part a) was already proved in [4] in the λ -domain (i.e. using $R(\lambda, A)$).

THEOREM 2.2. *Let A satisfy H1. a) If \mathcal{S} is approximately controllable in finite time, then \mathcal{F} is approximately controllable in any $[0, T], 0 < T < \infty$. In symbols: $\bigcup_{0 < T < \infty} \bar{K}_T(\mathcal{F}) = X + X \Rightarrow \bar{K}_T(\mathcal{F}) = X$. b) The reverse implication also holds, if statement (2.7) implies statement (2.6).*

Remark 2.2. Statement (2.6) always implies statement (2.7) for any T , while the converse is not always true. However, since $C(t)$ is even and therefore $\int_0^t C(\tau)y d\tau$ is odd on $(-\infty, \infty)$, one also has that (2.7) implies (2.6) for $t \geq 0$, provided (2.6) valid for $t \geq 0$ can be extended to hold also for all negative values of t . This is the case, e.g., when A is bounded on X or in the physically significant case when A is normal and has compact resolvent (see § 3).

Before stating the next result, we define $D_\infty(A) = \bigcap_{n=1}^\infty D(A^n)$ and recall that, for an operator A generating a C_0 -semigroup, hence a fortiori for A satisfying H1, $D_\infty(A)$ is a subspace and is still dense in X [1, p. 12]. Also, with $X_0 = \bigcup_{0 < t} S(t)X$ (the dense subspace of X defined before), we have $X_0 \subset D_\infty(A)$ for any semigroup satisfying $S(t)X \subset D(A)$ for $t > 0$, i.e., for any differentiable semigroup, hence a fortiori for the analytic semigroup generated by A satisfying H1. The general result showing that condition H2 is not necessary for the approximate controllability of $\mathcal{F}\langle A, B \rangle$ to imply the same property for $\mathcal{S}\langle A, B \rangle$ is contained in Corollary 2.4 below. To this end, we shall focus our attention on smooth vectors of X contained in $D_\infty(A)$. The next result is in the spirit of the extension, as given by the author in [21, § 2], of the classical rank condition for controllability of finite dimensional systems, of which it is a generalization.

THEOREM 2.3. *Let A satisfy H1.*

a) *With $BU \subset D_\infty(A)$ (resp. with $b_i \in D_\infty(A), i = 1, \dots, m$), the condition*

$$\overline{\text{sp}} \{A^n BU\}_{n=0}^\infty = X \quad (\text{resp. } \overline{\text{sp}} \{A^n b_i, i = 1, i = 1, \dots, m\}_{n=0}^\infty = X)$$

is sufficient for approximate controllability on any $[0, T], 0 < T < \infty$, of $\mathcal{S}\langle A, B \rangle$ (resp. $\mathcal{S}_m\langle A, (b_1, \dots, b_m) \rangle$), hence of $\mathcal{F}\langle A, B \rangle$ (resp. $\mathcal{F}_m\langle A, (b_1, \dots, b_m) \rangle$). Such condition is

however not even necessary for approximate controllability of $\mathcal{F}_m\langle A, (b_1, \dots, b_m) \rangle$.

b) With $BU \subset X_0$ (resp. with $b_i \in X_0, i = 1, \dots, m$), the condition

$$\overline{\text{sp}} \{A^n BU\}_{n=0}^\infty = X \quad (\text{resp. } \overline{\text{sp}} \{A^n b_i, i = 1, \dots, m\}_{n=0}^\infty = X)$$

is necessary and sufficient for approximate controllability on any $[0, T], 0 < T < \infty$ of both $\mathcal{S}\langle A, B \rangle$ (resp. $\mathcal{S}_m\langle A, (b_1, \dots, b_m) \rangle$) as well as $\mathcal{F}\langle A, B \rangle$ (resp. $\mathcal{F}_m\langle A, (b_1, \dots, b_m) \rangle$).

COROLLARY 2.4. Let A satisfy H1. With $BU \subset X_0$ (resp. with $b_i \in X_0, i = 1, \dots, m$), approximate controllability on any $[0, T], 0 < T < \infty$, of $\mathcal{F}\langle A, B \rangle$ (resp. $\mathcal{F}_m\langle A, (b_1, \dots, b_m) \rangle$) implies the same property for $\mathcal{S}\langle A, B \rangle$ (resp. $\mathcal{S}_m\langle A, (b_1, \dots, b_m) \rangle$).

Remark 2.3. If the operator B of the original approximately controllable system $\mathcal{F}(A, B)$ does not satisfy $BU \subset X_0$ and perturbations are allowed, reference to Remark 2.1 is useful. This program is carried out in Remark 2.4. For the *computability* of the above tests to physically significant classes of systems, refer to [21] and also to the subsequent Remark 2.8 in the present paper.

Proof of Theorem 2.3. Part a). That the stated condition is not necessary for approximate controllability of $\mathcal{F}_m\langle A, (b_1, \dots, b_m) \rangle$ was already pointed out by the author in [21, Remark 2.4] by means of an example with $m = 2$ (Example 2.3 of the present paper, in fact). As for the sufficiency in part a) for \mathcal{S}_m , by contradiction let there exist $\bar{x}_1^*, \bar{x}_2^* \in X^*$, not both zero, such that (see Claim (ii))

$$(2.8) \quad \int_0^t \bar{x}_1^*(C(\tau)b_i) d\tau + \bar{x}_2^*(C(t)b_i) = 0, \quad 0 \leq t \leq T, \quad i = 1, \dots, m.$$

Recall now, e.g. [10], that for $n = 1, 2, \dots$,

$$(2.9) \quad \frac{d^{2n}C(t)y}{dt^{2n}} = A^n C(t)y = C(t)A^n y, \quad y \in D_\infty(A), \quad -\infty < t < \infty,$$

$$(2.10) \quad C(0)x = x, \quad x \in X \quad \text{and} \quad \left. \frac{dC(t)x}{dt} \right|_{t=0} = 0, \quad x \in D(A).$$

Setting $t = 0$ in (2.8) yields $\bar{x}_2^*(b_i) = 0$. Differentiate (2.8) successively using (2.9), and set $t = 0$ at each step using (2.10) to show, by induction, that

$$(2.11) \quad \bar{x}_1^*(A^n b_i) = 0, \quad \text{and} \quad \bar{x}_2^*(A^n b_i) = 0, \quad n = 0, 1, \dots,$$

But in view of the Hahn–Banach theorem, this contradicts the hypothesis, since \bar{x}_1^* and \bar{x}_2^* are not both zero. Similarly for \mathcal{S} . Since $X_0 \subset D_\infty(A)$, as was previously observed, it remains to show necessity for part b). To this end, let $b_i \in X_0$, i.e., $b_i = S(\tau_i)\beta_i$, for some $\tau_i > 0$ and $\beta_i \in X$ and let

$$\bar{x}^*(A^n b_i) = \bar{x}^*(A^n S(\tau_i)\beta_i) = 0, \quad i = 1, \dots, m, \quad n = 0, 1, \dots,$$

for some nonzero $\bar{x}^* \in X^*$. Then, by the analyticity of $S(t)$ which implies [1, pp. 15–16] $A^n S(t)$ bounded on X and

$$(2.12) \quad \frac{d^n S(t)}{dt^n} = A^n S(t), \quad t > 0,$$

it follows that $\bar{x}^*(S(t)\beta_i) = 0$ as well as $\bar{x}^*(S(t)b_i) = 0, 0 \leq t < \infty$. By Lemma 2.1, one also has $\bar{x}^*(C(t)b_i) = 0, -\infty < t < \infty$. Applying the previous claims (i) and (ii) yields easily a contradiction to the assumptions, since \bar{x}^* is nonzero. Q.E.D.

Necessity in part b) for \mathcal{S} can also be proved through a property of $C(t)X_0$ which is treated in appendix 1 of [23]; i.e., the map $t \rightarrow C(t)x, x \in X_0$ is entire with infinite radius

of convergence. See [25] for more details: These show in particular that X_0 is *not* the largest subspace of $D_\infty(A)$ containing BU for which $\overline{\text{sp}} \{A^n BU\} = X$ is a necessary condition for approximate controllability in finite time of $\mathcal{S}(A, B)$. The largest subspace with this property is denoted in [25] by X_m .

Remark 2.4. Recall Fattorini's example from § 1. Writing $A_r = \Delta + rI$, we have $S_{A_r}(t) = e^{rt} S_\Delta(t)$ with

$$(2.13) \quad (S_\Delta(t)f)(\xi) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} e^{-(\xi-\sigma)^2/(4t)} f(\sigma) d\sigma, \quad t > 0, \quad -\infty < \xi < \infty,$$

[1, p. 2]. Then $b_1(\xi) = e^{-|\xi|}$ and $b_2(\xi) = b_1(\xi + 1)$ do not belong to the subspace $X_0 = \bigcup_{0 < t} S_{A_r}(t)X$ in agreement with our Theorem 2.3. In fact, they do not even belong to $D_\infty(A)$ as one sees quickly by Plancherel–Fourier transform methods. However, for any $\varepsilon > 0$, there is $t_\varepsilon > 0$ such that the perturbations $\beta_i = S_{A_r}(t_\varepsilon)b_i$ satisfy $\|\beta_i - b_i\| < \varepsilon$, $i = 1, 2$ and make A_r approximately controllable on any $[0, T]$ for the first order system [see Remark 2.1] and also for the second order system [see Theorem 2.3b]. Finding a suitable t_ε is most easily done by Plancherel–Fourier methods: e.g.

$$\|\beta_1 - b_1\|^2 = \|\hat{\beta}_1 - \hat{b}_1\|^2 = \int_{-\infty}^{\infty} [e^{t(r-\omega^2)} - 1]^2 \left[\frac{2}{1+\omega^2} \right]^2 d\omega.$$

Remark 2.5. When $BU \subset D_\infty(A)$, the following inclusion is contained in the proof of Theorem 2.3a) and in Lemma 2.1:

$$(2.14) \quad \overline{\text{sp}} \{A^n BU\}_{n=0}^\infty = \overline{\text{sp}} \{S(t)BU, 0 \leq t \leq \infty\} = \overline{\text{sp}} \{C(t)BU, -\infty < t < \infty\}.$$

The next example with $m = 2$, illustrates the case when the inclusion (2.14) is proper, with the set on the right hand side being the entire space.

Example 2.3. Let $X = L_2[-\infty, \infty]$, $Af = d^2f/d\xi^2$ (in the sense of distributions with $D(A) = \{f: f \text{ and } f'' \in L_2[-\infty, \infty]\}$). Take $b_1(\xi)$ to be a C^∞ function with compact support. Say: $b_1(\xi) = \exp(-(\xi^2 - 1))^{-1}$, $-1 < \xi < 1$, and $b_1(\xi) = 0$ for $|\xi| \geq 1$. Define $b_2(\xi) = b_1(\xi - h)$, $h \neq 0$. Then $b_1(\cdot)$ and $b_2(\cdot)$ belong to $D_\infty(A)$, and, moreover, they vanish identically together with all their derivatives outside $[-1, 1+h]$ for $h > 0$ ($[-1+h, 1]$ for $h < 0$). Hence in this case we have $\overline{\text{sp}} \{A^n b_i, i = 1, 2\}_{n=0}^\infty \subsetneq X$ and there is a nonzero $\bar{x}^* \in X^*$ such that $\bar{x}^*(A^n b_i) \equiv 0$, $i = 1, 2$; $n = 0, 1, \dots$. Yet the identity $\bar{x}^*(S(t)b_i) \equiv 0$, $0 \leq t < \infty$, implies $\bar{x}^* = 0$ in view of the previous claim (i) since, as was shown in [5, Ex. 1], [21, Ex. 3.], the present system $\mathcal{F}_2(A, (b_1, b_2))$ is indeed approximately controllable. Notice that $b_i \notin X_0$, in agreement with Theorem 2.3b).

Remark 2.6. On the other hand, Theorem 2.3b) and Lemma 2.1 show that, if in particular $BU \subset X_0$, then

$$(2.15) \quad \overline{\text{sp}} \{A^n BU\}_{n=0}^\infty = \overline{\text{sp}} \{S(t)BU, 0 \leq t < \infty\} = \overline{\text{sp}} \{C(t)BU, -\infty < t < \infty\}.$$

COROLLARY 2.5. Let A be bounded on X . Then $\mathcal{S}(A, B)$ is approximately controllable in $[0, T]$ if and only if $\mathcal{F}(A, B)$ is approximately controllable, and this happens just in the case

$$(2.16) \quad \overline{\text{sp}} \{A^n BU\}_{n=0}^\infty = X.$$

Remark 2.7. When A is bounded on X (in which case assumption H2 is certainly satisfied), then

$$S(t) = e^{At} = \sum_{n=0}^{\infty} \frac{A^n t^n}{n!} \quad \text{and} \quad C(t) = \sum_{n=0}^{\infty} \frac{A^n t^{2n}}{(2n)!}, \quad -\infty < t < \infty,$$

and the above argument leading to Corollary 2.5 simplifies, by making direct use of the explicit power series expansions for $S(t)$ and $C(t)$. Also, the characterization (2.16) was

already given in [20, Th. 3.1.1] for \mathcal{F} , and one can deduce from it the corresponding statement for \mathcal{S} of Corollary 2.5 by rewriting \mathcal{S} as first order system on $X \times X$:

$$(2.17) \quad \dot{y} = \bar{A}y + \bar{B}v, \quad y \in Y = X \times X, \quad v \in V = U \times U, \quad \text{with}$$

$$\bar{A} = \begin{bmatrix} 0 & I \\ A & 0 \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} 0 \\ B \end{bmatrix}$$

bounded operators on Y and $V \rightarrow Y$, respectively. The powers $\bar{A}^n \bar{B}V$, $n = 0, 1, \dots$ are then

$$\begin{bmatrix} 0 \\ BU \end{bmatrix}, \begin{bmatrix} BU \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ ABU \end{bmatrix}, \begin{bmatrix} ABU \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ A^2BU \end{bmatrix}, \begin{bmatrix} A^2BU \\ 0 \end{bmatrix}, \text{ etc.,}$$

and they span $X \times X$ if and only if (2.16) holds. Q.E.D.

The next corollary is a comprehensive statement incorporating previous known results for first order systems ([6, Prop. 2.3] and [20, Th. 3.1.1]) together with Theorems 2.1 and 2.3 above as well as [4, Th. 2.3] for second order systems.

COROLLARY 2.6. *Let A satisfy H1 and consider the following four systems*

$$\begin{aligned} \mathcal{F}\langle A, B \rangle: \dot{x} &= Ax + Bu, & \mathcal{F}\langle R(\lambda_0, A), B \rangle: \dot{x} &= R(\lambda_0, A)x + Bu, \\ \mathcal{S}\langle A, B \rangle: \ddot{x} &= Ax + Bu, & \mathcal{S}\langle R(\lambda_0, A), B \rangle: \ddot{x} &= R(\lambda_0, A)x + Bu, \end{aligned}$$

where λ_0 is a fixed but otherwise arbitrary point in $\rho_0(A)$ (defined in the statement of H2 in § 1). Then

a) *either the systems $\mathcal{F}\langle A, B \rangle$, $\mathcal{F}\langle R(\lambda_0, A), B \rangle$ and $\mathcal{S}\langle R(\lambda_0, A), B \rangle$ are all simultaneously approximately controllable on any $[0, T]$, $0 < T < \infty$, or none of them is; the first alternative occurs if and only if*

$$(2.18) \quad \overline{\text{sp}} \{R^n(\lambda_0, A)BU\}_{n=0}^\infty = X.$$

b) *Equation (2.18) is a necessary condition for approximate controllability in finite time of $\mathcal{S}\langle A, B \rangle$, and it is also sufficient in the following two cases:*

- (i) *either A satisfies Fattorini's assumption H2*
- (ii) *or (with no assumption on A except H1) the range of B is contained in $\bigcup_{0 < t} S(t)X$.*

Remark 2.8. Equation (2.18) represents another version of the extension of the finite dimensional rank condition, as distinct from those in the statement of Theorem 2.3. We emphasize once more that all such versions are *computable*, and in fact they have been used by the author to derive in a systematic way easy-to-check-tests for large classes of physically significant dynamical systems [see 21]. See also [13] for an important case (delay differential equation) of an operator which generates a C_0 semigroup but *not* a cosine function.

Remark 2.9. We finally remark that Russell proved in [15] that a type of *exact* controllability in finite time of the wave equation, using *boundary* controls, implies the same property for the corresponding heat equation, on the same region and with controls of the same type. Although the problem in the present paper and the problem in Russell's are closely related (approximate controllability with distributed controls versus a type of exact controllability with boundary controls) the mathematical methods employed are—and appear to be bound to be—completely different.

3. The case when A is normal with compact resolvent. Throughout the present section, X will be specialized to be a Hilbert space and the operator A is assumed to satisfy, in addition to H1, the following assumption.

H3. A is normal and its resolvent $R(\mu_0, A)$ is (normal) and compact as an operator on X for some μ_0 (hence, for all μ in $\rho(A)$ [9, p. 187]).

The reason for singling out this special case lies in its importance in physical applications: the compactness of the resolvent is in fact automatically satisfied if the partial differential equation is defined on a bounded spatial domain [3 pp. 1739–40; p. 1330], [7, Remark 2.2], [14, Chap. 7], while the normality assumption covers all classical boundary conditions. Notice that assumption H3 certainly implies assumption H2 (see below) and hence the present special case is covered by Fattorini's result, reported in § 1: under H1 and H3, either \mathcal{F} and \mathcal{S} are both approximately controllable in finite time, or neither of them is. However, we feel that precisely because of the relevance of the present class of differential systems in mathematical physics, it is instructive to give a *direct*, ad hoc proof of the above results. Such proof, which is given below, will make use of the *particular* structures of the cosine function $C(t)$ and its corresponding semigroup $S(t)$, as implied by assumption H3.

In view of assumption H3, the following holds (for background see [9, p. 277], [14, p. 487], [3, p. 1330], etc.). The cosine operator $C(t)$ and the semigroup $S(t)$ are given by

$$(3.1) \quad C(t)x = \sum_{j=1}^{\infty} \cos \sqrt{-\lambda_j} t \sum_{k=1}^{r_j} (x, x_{jk}) x_{jk}, \quad -\infty < t < \infty, \quad x \in X,$$

$$(3.2) \quad S(t)x = \sum_{j=1}^{\infty} e^{\lambda_j t} \sum_{k=1}^{r_j} (x, x_{jk}) x_{jk}, \quad t \geq 0, \quad x \in X,$$

respectively. Here the $\{\lambda_j\}$ are the distinct isolated eigenvalues of A , $|\lambda_j| \rightarrow \infty$ as $j \rightarrow \infty$, with corresponding finite multiplicity r_j . The $\{x_{jk}\}$ are the complete orthonormal set of eigenvectors of A , $k = 1, \dots, r_j$. As a consequence of H1, such $\{\lambda_j\}$ are contained in a parabolic sector [5, Remark 5.6]

$$(3.3) \quad \{\lambda^2 : \operatorname{Re} \lambda \leq \omega_0\} = \left\{ \lambda : \operatorname{Re} \lambda \leq -\frac{(\operatorname{Im} \lambda)^2}{4\omega_0^2} + \omega_0^2 \right\}.$$

Now, let Claim (ii) in § 2 be violated, i.e., let

$$(3.4) \quad \sum_{j=1}^{\infty} \frac{\sin \sqrt{-\lambda_j} t}{\sqrt{-\lambda_j}} \sum_{k=1}^{r_j} (BU, x_{jk}) \bar{x}_1^*(x_{jk}) + \sum_{j=1}^{\infty} \cos \sqrt{-\lambda_j} t \sum_{k=1}^{r_j} (BU, x_{jk}) \bar{x}_2^*(x_{jk}) \equiv 0, \quad t \geq 0,$$

for $\bar{x}_1^*, \bar{x}_2^* \in X^*$ not both zero. Since the λ_j 's are in a parabolic sector as described in (3.3), it follows that $|\operatorname{Im} \sqrt{-\lambda_j}| \leq K$ uniformly in j , so that both $|\cos(\sqrt{-\lambda_j} t)|^2$ and $|\sin(\sqrt{-\lambda_j} t)|^2$ are uniformly bounded above by $\frac{1}{2}[\cosh 2Kt + 1]$. It follows that each series on the left hand side of (3.4) is uniformly convergent on any interval $[-T, T]$, since it is unconditionally convergent, i.e., independent of the order of the index j . Therefore, the left hand side of (3.4) is an almost periodic function on $(-\infty, \infty)$ [2, Thm. 1.6, p. 12]. Since it is identically zero for $t \geq 0$, by Bohr's definition [2, Property B, p. 14], (3.4) is identically zero also for $t < 0$. But the first series on the left hand side of (3.4) is an odd function of t , while the second series is an even function of t . Therefore, both series must be identically zero on $(-\infty, \infty)$ (refer to Remark 2.2), i.e.,

$$\bar{x}_2^*(C(t)BU) \equiv 0 \quad \text{and} \quad \int_0^t \bar{x}_1^*(C(\tau)BU) d\tau \equiv 0, \quad -\infty < t < \infty,$$

and hence also $\bar{x}_1^*(C(t)BU) \equiv 0$ in $(-\infty, \infty)$, the integrand being continuous. Since \bar{x}_1^* and \bar{x}_2^* are not both zero, we have therefore proved that *under assumptions H1 and H3, condition (2.7) implies condition (2.6)*. In view of Theorem 2.2b), this justifies the 'if' part of the following Theorem 3.1. Its 'only if' part is contained instead in Theorem 2.2a).

Moreover, the subsequent characterizations (3.5) and (3.6) are already known for the system \mathcal{F} [5, Example 4, for A self adjoint], [21, § 3.2]. In [21], they were derived using e.g., characterization (2.18) as applied to the present case.

THEOREM 3.1. *Let A satisfy H1 and H3. Then $\mathcal{S}\langle A, B \rangle$ is approximately controllable in finite time if and only if $\mathcal{F}\langle A, B \rangle$ is approximately controllable on an arbitrary $[0, T]$; this is the case if and only if*

$$(3.5) \quad P_j B U = X_j, \quad j = 1, 2, \dots, \quad B U = \text{range of } B.$$

For \mathcal{S}_m and \mathcal{F}_m the above characterization (3.6) becomes

$$(3.6) \quad \text{rank} \begin{vmatrix} (b_1, x_{j1}), & \dots & (b_m, x_{j1}) \\ (b_1, x_{j2}), & \dots & (b_m, x_{j2}) \\ \vdots & & \vdots \\ (b_1, x_{jr_j}), & \dots & (b_m, x_{jr_j}) \end{vmatrix} = r_j, \quad j = 1, 2, \dots,$$

which in turn implies: $\sup r_i \leq m$.

Moreover, $\mathcal{S}\langle A, B \rangle$ (resp. $(\mathcal{S})\langle A, (b_1, \dots, b_m) \rangle$) is approximately controllable in an arbitrary interval $[0, T]$, $0 < T < \infty$, if, in addition $B U \subset X_0$ (resp. $b_i \in X_0$: this is equivalent to having b_i of the form:

$$b_i = S(\tau_i) \beta_i = \sum_{j=1}^{\infty} e^{\lambda_j \tau_i} \sum_{k=1}^{r_j} (\beta_i, x_{jk}) x_{jk}$$

for some $\beta_i \in X$ and $\tau_i > 0$).

REFERENCES

- [1] P. L. BUTZER AND H. BERENS, *Semigroups of Operators and Approximations*, Springer-Verlag, Berlin, 1967.
- [2] C. CORDUNEANU, *Almost Periodic Functions*, Interscience, New York, 1968.
- [3] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, Parts 1 and 2, Interscience, New York, 1959 and 1963.
- [4] H. O. FATTORINI, *Controllability of higher order linear systems*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustad, eds., Academic Press, New York, 1967.
- [5] ———, *Ordinary differential equations in linear topological spaces*, I and II, J. Differential Equations, 5 (1968), pp. 72–105, and 6 (1969), pp. 50–70.
- [6] ———, *On complete controllability of linear systems*, Ibid., 3 (1967), pp. 391–402.
- [7] ———, *Some remarks on complete controllability*, this Journal 4 (1966), pp. 686–694.
- [8] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, American Mathematical Society, Providence, RI, 1958.
- [9] T. KATO, *Perturbation Theory of Linear Operators*, Springer-Verlag, Berlin, 1966.
- [10] J. KISINSKY, *On cosine operator functions and one parameter groups of operators*, Studia Math. T. XLIV (1972), pp. 93–105.
- [11] ———, *On the connection between cosine operator functions and one parameter semigroups and groups of operators*, Institute of Mathematics Report, University of Warsaw, Poland, 1972.
- [12] ———, *On operator-valued solutions of D'Alambert's functional equation*, II, Studia Math. T. XLII (1972), pp. 43–66.
- [13] A. MANITIUS AND R. TRIGGIANI, *Function space controllability of linear retarded systems: A derivation from abstract operator conditions*, this Journal, 16 (1978), pp. 599–646.
- [14] T. H. NAYLOR AND G. R. SELL, *Linear Operators in Engineering and Science*, Holt, Rinehart and Winston, New York, 1971.
- [15] D. L. RUSSELL, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Studies in Appl. Math., LII, 3 (September 1973), pp. 189–211.
- [16] M. SOVA, *Cosine operator functions*, Rozprawy Mat., XLIX (1966).
- [17] M. SLEMROD, *A note on complete controllability and stabilizability of linear control systems in Hilbert space*, this Journal, 12 (1974), pp. 500–508.

- [18] G. SZEGO, *Orthogonal Polynomials*, Colloquium Publications, American Mathematical Society, Providence RI, 1959.
- [19] K. TSUJIOKA, *Remarks on controllability of second order evolution equations in Hilbert spaces*, this Journal, 8, (1970), pp. 90–99.
- [20] R. TRIGGIANI, *Controllability and observability in Banach space with bounded operators*, this Journal, 13 (1975), pp. 462–491.
- [21] ———, *Extensions of rank conditions for controllability and observability to Banach spaces and unbounded operators*, this Journal, 14 (1976), pp. 313–338.
- [22] ———, *On the lack of exact controllability for mild solutions in Banach space*, J. Math. Anal. Appl., 50 (1975), 438–446.
- [23] ———, *Proceedings of 7th IFIP Conference (Rome, June 1976)*, Springer-Verlag.
- [24] E. C. TITCHMARSH, *The Theory of Functions*, Oxford University Press, 2nd ed., London, 1939.
- [25] R. TRIGGIANI AND S. NELSON, *Analytic properties of cosine operators*, submitted.
- [26] K. YOSIDA, *Functional Analysis*, Springer-Verlag, Berlin, 1965.

ERRATA: NONPARAMETRIC IDENTIFICATION FOR DIFFUSION PROCESSES*

G. BANON†

1. Assumption A_2 (p. 380) should read: " $m(\cdot)$ and $\sigma(\cdot)$ are real-valued functions on R satisfying, for $x, y \in R$, the Lipschitz condition . . ."
2. The last sentence on p. 385 should read: ". . . which converges pointwise, for all $a \in R$, . . ."
3. The statement of Lemma 3.2 (p. 388) should be altered to be: ". . . of the X_t process converges pointwise as $t \rightarrow \infty$. . ."
4. The text following the first display on p. 389 should be as follows: ". . . have successively: $r_1 \neq 0$, $r_2 = 0$, $\phi_1(x, 0) = 0$ for . . ."

* This Journal, 16 (1978), pp. 380–395. Received by the editors February 28, 1978.

† Laboratoire d'Automatique et d'Analyse des Systèmes du Centre National de la Recherche Scientifique, 31400 Toulouse, France. This research was supported by the National Science Foundation under Grant 01P75-04371 and the "Centre National de la Recherche Scientifique."

OPTIMAL CONTROL OF FUNCTIONAL DIFFERENTIAL SYSTEMS*

F. COLONIUS† AND D. HINRICHSSEN†

Abstract. This paper presents a unified approach to diverse optimal control problems for hereditary differential systems (HDS). An abstract local maximum principle is established via the Dubovitskii–Milyutin method. It yields necessary conditions for the optimal control of HDS towards surfaces in \mathbb{R}^n and towards target sets in function spaces. Nondegeneracy criteria are included. It is shown that the necessary conditions are sufficient in the case of linear HDS with convex cost functionals. Analogous results are obtained for systems described by Fredholm equations with general control action. For Fredholm systems with targets in function spaces, the attainability space \mathcal{A} is investigated, criteria for \mathcal{A} to be closed are established and full attainability is characterized.

Introduction. This paper is mainly concerned with the optimal control of hereditary differential systems (HDS) with finite memory $h \geq 0$ described by functional differential equations of the form:

$$(1) \quad \dot{x}(t) = f_1(x_t, u(t), t), \quad \text{a.e. } t \in [t_0, t_1],$$

where for every $t \in [t_0, t_1]$

$$(2) \quad x_t: s \mapsto x(t+s), \quad s \in [-h, 0],$$

is the past history of $x(\cdot)$ corresponding to the moving time-interval $[t-h, t]$. Throughout the text, $[t_0, t_1]$ is a fixed time interval. We suppose that an initial datum $x_{t_0} = \varphi$ is given which describes the motion of the system during the time interval $[t_0-h, t_0]$. Then, under suitable assumptions on f_1 , the trajectory $x(\cdot) = x_u(\cdot)$ of the system is completely determined by the control function $u: [t_0, t_1] \rightarrow \mathbb{R}^r$. Pertinent existence, uniqueness and continuity results are to be found in [19].

Mathematical models of this type play an important role in every field of science where causes do not produce their effects immediately but with some time delay (see [14] for a brief survey).

Our purpose is to establish a unifying framework in order to derive maximum principles for the following optimum control problems:

(P1) optimal control of HDS towards a given point $x_1 \in \mathbb{R}^n$:

$$x(t_1) = x_1,$$

(P2) optimal control towards a given target function $\varphi_1: [-h, 0] \rightarrow \mathbb{R}^n$:

$$x_{t_1} = \varphi_1.$$

For a discussion of these two problems and a survey of the results obtained as well as the abstract optimization methods employed up to 1973, we refer the reader to [2], [7], [8]. While the abstract variational theory of Neustadt has been efficiently used to derive necessary optimality conditions for problem (P1) (cf. [1], [8]), it encountered severe difficulties, when applied to problem (P2). More recently, some new ideas have been introduced into this context which seem to be quite promising [10], [28] (cf. § 3). The present paper, however, follows the alternative general theory of extremals, the theory of Dubovitskii–Milyutin (cf. [18]). More detailed bibliographical information will be given in the following sections.

* Received by the editors November 30, 1976, and in final revised form December 1, 1977.

† Fachbereich Mathematik, Forschungsschwerpunkt Dynamische Systeme, Universität Bremen, 28 Bremen 33, West Germany.

The problems considered jointly in this paper are usually studied separately in the literature. In order to give a comprehensive account we also include some known results or minor generalizations.

In § 1 we establish an abstract local maximum principle that covers the optimal control problems (P1) and (P2). It is shown that the *necessary* conditions in their nondegenerate form ($\lambda_0 = 1$) are also *sufficient* for optimality in the case of linear systems with convex cost functional. Conditions for nondegeneracy are included in order to clear up the relationship between necessary and sufficient optimality criteria.

A similar maximum principle has been obtained by Kurcyusz [22], as a necessary condition for the optimal control of systems with operator constraints. However, our assumptions are more readily verified. They easily admit the application to HDS with *distributed* lags and delays in the control, as well as to more general systems described by Fredholm equations.

In § 2 we briefly summarize the results which may be obtained by application of the abstract maximum principle to problem (P1).

In § 3 we apply the results of § 1 to problem (P2), confining ourselves to linear HDS and more generally to Fredholm systems with general control action. We obtain a maximum principle which states necessary and sufficient conditions for optimality. Possible generalizations to nonlinear systems are briefly indicated. Problem (P2) still has been treated by relatively few authors. To our knowledge, all attempts to establish a corresponding maximum principle for *general nonlinear* functional differential systems have failed up to now (cf. § 3).

Since the maximum principle obtained in § 3 depends essentially on the assumption that the attainability space \mathcal{A} is closed, we investigate this subspace in § 4. Our results generalize some of the theorems obtained in [6]. A conjecture of Banks–Jacobs–Langenhop [6, p. 619] concerning the necessity of their condition (H3) for \mathcal{A} to be closed is partially confirmed (Prop. 4.2). The known criteria for \mathcal{A} to be closed still remain rather unsatisfactory. Only full attainability is completely characterized.

Notation and terminology. Let X be a Banach space and let X^* denote its topological dual space. We define the symbol $\langle x^*, x \rangle_B$ by $\langle x^*, x \rangle_B := x^*(x)$, where the right-hand side is the value of the linear form x^* at the point x . Let $F: X_1 \times X_2 \rightarrow X$ be a map, X_1, X_2, X B-spaces. Then $DF(x) = D_1F(x_1, x_2) + D_2F(x_1, x_2)$ denotes the Fréchet derivative of F in $x = (x_1, x_2) \in X_1 \times X_2$. $\mathcal{L}(X_1, X_2)$ is the space of continuous linear operators, mapping X_1 into X_2 . A^* denotes the adjoint of a continuous linear operator A , while $\text{Im } A$ and $\text{Ker } A$ are its range and kernel, respectively; $\text{cl } Q$ is the topological closure of a set Q , $\text{int } Q$ its interior.

For any subset $Q \subset \mathbb{R}$, 1_Q denotes the characteristic function of Q on \mathbb{R} .

1. An abstract local maximum principle. We consider the following abstract control problem (ACP) which reflects the general structure of concrete control problems without presupposing the controls and trajectories to be functions of time:

(ACP) Let X, U, Z be Banach spaces, $Q \subset U$, $F_0: X \times U \rightarrow \mathbb{R}$, $F_1: X \times U \rightarrow X$, $F_2: X \rightarrow Z$.

Minimize

$$F_0(x, u),$$

subject to

$$(1.1) \quad x = F_1(x, u),$$

$$(1.2) \quad F_2(x) = 0,$$

$$(1.3) \quad u \in Q.$$

Later on, X will be interpreted as the space of trajectories, U as the space of control functions, Q as the set of admissible control functions and $F_2(x)=0$ as the end condition. $x = F_1(x, u)$ corresponds to the equation of motion of the dynamical system.

THEOREM 1.1 (Local maximum principle). *Let $(x^0, u^0) \in X \times U$ be a solution of (ACP) and suppose the following conditions:*

- (a) F_0 is Fréchet-differentiable in (x^0, u^0) , F_1 , F_2 are continuously Fréchet-differentiable in a neighborhood of (x^0, u^0) , resp. x^0 .
- (b) $D_1F_1(x^0, u^0)$ is a compact operator that satisfies

$$\text{Ker}(\text{Id}_X - D_1F_1(x^0, u^0)) = 0.$$

- (c) The “attainable subspace” \mathcal{A} of the linearized system

$$\mathcal{A} := \{DF_2(x^0)x \mid \exists u \in U: x = DF_1(x^0, u^0)(x, u)\}$$

is not a proper dense subspace of Z .

- (d) Q is convex and contains interior points.

Under these assumptions there exist $\lambda_0 \geq 0$, $x^ \in X^*$, $z^* \in Z^*$, not all zero, which satisfy the following two conditions:*

- (i) $x^* = \lambda_0 D_1F_0(x^0, u^0) + D_1F_1(x^0, u^0)^* x^* + DF_2(x^0)^* z^*$ (adjoint equation);
- (ii) $[\lambda_0 D_2F_0(x^0, u^0) + D_2F_1(x^0, u^0)^* x^*][u^0 - u] \leq 0$ for all $u \in Q$

(minimum condition).

Nondegeneracy condition: λ_0 is nonzero if the following additional assumptions are satisfied:

- (e) *There exist $\tilde{u} \in U$, $\tilde{x} \in X$ such that $u^0 + \tilde{u} \in \text{int } Q$ and*

$$\tilde{x} = DF_1(x^0, u^0)(\tilde{x}, \tilde{u}), \quad 0 = DF_2(x^0)(\tilde{x}).$$

- (f) *All the points in Z are attainable, that is, $\mathcal{A} = Z$.*

Proof. Construct local approximations of the objective function and the constraints by convex cones, and apply the theorem of Dubovitskii–Milyutin in order to establish the generalized equation of Euler–Lagrange for our problem [18]. Then, show that this equation is equivalent to (i) and (ii). For details see [12]. \square

Remark 1.1. In this remark we comment on some of the assumptions of the preceding theorem.

1) Assumption (b) is generally satisfied for ordinary and hereditary differential systems.

2) The most critical assumption of the theorem is (c). It is always satisfied, if we deal with pointwise end conditions (Z finite dimensional). However, it is difficult to verify, if Z is a function space.

3) If the closure of \mathcal{A} is a proper subspace of Z , λ_0 may be chosen to be zero and hence we get necessary conditions which are independent of the objective functional. In this case the theorem tells us something about the “system” (1.1) and its relation to the constraints (1.2), (1.3), but nothing about the optimal control problem: Every solution (x^0, u^0) of (1.1)–(1.3), for which the assumptions (a), (b) are met and $\text{cl}(\mathcal{A}) \neq Z$, satisfies conditions (i), (ii) with $\lambda_0 = 0$. The stronger assumption (f) excludes this possibility. It means that the system linearized at (x^0, u^0) is completely attainable. This implies that the differential of the function $X \times U \rightarrow X \times Z$ defined by the equality constraints (1.1), (1.2) is surjective. Hence the theorem of Lyusternik may be applied to compute the corresponding tangent cone [12].

4) If both (e) and (f) are satisfied it follows that $\lambda_0 \neq 0$. We have just seen that *only the nondegenerate version* of the local maximum principle yields pertinent information for the solution of the optimal control problem. This explains the

theoretical interest of sufficient conditions for nondegeneracy which should be included in maximum principles whenever possible. One of the strong points of the Dubovitskii–Milyutin method is that, in many cases, it yields conditions for nondegeneracy without much additional effort.

5) Halkin [20] has shown that for finite dimensional terminal conditions (i.e. Z finite dimensional) the assumptions concerning Q and the Fréchet-differentiability of F_0, F_1, F_2 can be relaxed.

We now consider the following linear version of the abstract control problem:

(LACP) Let X, U, Z, Q, F_0 be as before and $A \in \mathcal{L}(X, X)$, $B \in \mathcal{L}(U, X)$, $C \in \mathcal{L}(X, Z)$, $z \in Z$.

Minimize

$$F_0(x, u)$$

subject to

$$(1.4) \quad x = Ax + Bu,$$

$$(1.5) \quad Cx = z,$$

$$(1.6) \quad u \in Q.$$

We shall see that in the linear case the nondegenerate maximum principle is not only a necessary but also a sufficient condition for optimality.

Furthermore, the nondegeneracy assumptions may be weakened. It now suffices to require that the attainability space \mathcal{A} is closed in Z (instead of $\mathcal{A} = Z$). To see this, we simply regard the map $u \mapsto F_2(x_u)$, where x_u is the solution of $x = F_1(x, u)$, as a map from U onto \mathcal{A} .

THEOREM 1.2. *Let $(x^0, u^0) \in X \times U$ satisfy the constraints (1.4)–(1.6) and suppose that the following assumptions hold:*

- (a) F_0 is Fréchet-differentiable in (x^0, u^0) and convex.
- (b) A is a compact linear operator satisfying

$$\text{Ker}(\text{Id}_X - A) = 0.$$

- (c) $\mathcal{A} = \{Cx \mid \exists u \in U: x = Ax + Bu\}$ is closed in Z .
- (d) Q is convex and contains interior points.
- (e) There exists $(\tilde{x}, \tilde{u}) \in X \times U$ such that

$$\tilde{x} = A\tilde{x} + B\tilde{u}, \quad C\tilde{x} = z, \quad \tilde{u} \in \text{int } Q.$$

Then (x^0, u^0) is a solution of (LACP) iff the following conditions are satisfied:

There exist $x^* \in X^*$, $z^* \in Z^*$ such that

- (i) $x^* = D_1 F_0(x^0, u^0) + A^* x^* + C^* z^*$.
- (ii) $(D_2 F_0(x^0, u^0) + B^* x^*)(u^0 - u) \leq 0$ for all $u \in Q$.

Proof. Let $T: U \rightarrow X$ be the linear mapping which associates with every $u \in U$ the corresponding solution x_u of $x = Ax + Bu$. (LACP) may be reformulated as follows:

(LACP') Minimize $F_0(Tu, u)$

subject to $(C \circ T)u = z, \quad u \in Q.$

$C \circ T$ is a continuous, linear, surjective operator from U onto the Banach space \mathcal{A} (continuity of T follows from Banach's inverse theorem). It is now easy to derive the sufficiency and necessity of conditions (i) and (ii) from the corresponding theorem of Dubovitskii–Milyutin [18, p. 115]. \square

Remark 1.2. $G: u \mapsto F_0(x^0, u) + x^*(Ax^0 + Bu)$ is a convex functional on Q . Condition (ii) expresses that the Fréchet derivative of G in u^0 is a support functional of Q at u^0 . Since G is convex this means that u^0 is a minimum of G on Q . Hence the *local minimum condition* (ii) is equivalent to the following *global minimum condition*:

(ii') $F_0(x^0, u^0) + x^*(Ax^0 + Bu^0) \leq F_0(x^0, u) + x^*(Ax^0 + Bu)$ for all $u \in Q$.

Remark 1.3. With regard to the *necessity* of (i) and (ii), the central assumption of Theorem 1.2 is (c). However, here it is not needed for calculating the cone of tangent directions (because of linearity, this is a trivial problem). But it is needed in order to determine the dual cone. Kurcysz [22], [23] has shown the following interesting result: If \mathcal{A} is not closed in Z and $Q = U$, then an objective functional F_0 exists such that the unique optimal solution of the corresponding problem (LACP) does not satisfy the maximum principle ((i), (ii)).

Remark 1.4. According to Remark 1.1 only the *nondegenerate* maximum principle can be expected to yield a *sufficient* criterion for optimality. Assumptions (c) and (e) are used to establish the nondegenerate version of the maximum principle as a *necessary* criterion. These assumptions become redundant for the sufficiency part of the theorem which *presupposes* $\lambda_0 = 1$. This corresponds to the well-known fact that Slater's condition is only needed in order to prove the *necessity* of the saddle point condition for convex programs (see [18, p. 116], [25, p. 216]).

The following reasoning shows that the conditions (c), (e) and $\text{int } Q \neq \emptyset$ in (d) can be dispensed with in the proof of sufficiency. Suppose (i) and (ii). Then, by definition of T and the adjoint equation we have for $u \in Q$:

$$(B^*x^*)u = (x^* \circ (\text{Id}_X - A) \circ T)u = (D_1F_0(Tu^0, u^0) \circ T)u + (z^* \circ C \circ T)u.$$

This and the analogue equation for $(B^*x^*)u^0$ show that (ii) implies

$$(D_2F_0(Tu^0, u^0) + D_1F_0(Tu^0, u^0) \circ T)(u^0 - u) \leq 0$$

for all $u \in Q$ with $(C \circ T)u = z$.

Since $u \mapsto F_0(Tu, u)$ is convex on the convex set $\{u \in Q; (C \circ T)u = z\}$, we conclude as above that u^0 is an optimal solution of (LACP').

This remark again illustrates the theoretical significance of *conditions for nondegeneracy*: They specify assumptions under which the maximum principle can be established as a necessary condition in such a form that it becomes sufficient for optimality, if certain convexity conditions are satisfied.

Remark 1.5. Clearly, Theorems 1.1 and 1.2 may also be applied to optimal control problems without end condition. In this case $F_2 \equiv 0$ and $\mathcal{A} = Z = \{0\}$, assumption (c) is trivially satisfied; and the second Lagrange multiplier z^* is zero.

Remark 1.6. *Existence* of an optimal solution follows, if we assume, e.g., that U is reflexive and Q is closed and bounded. In this case, $Q \cap \{u \in U | (C \circ T)u = z\}$ is weakly compact and F_0 is weakly lower semicontinuous. *Unicity* of the optimal solution is guaranteed if F_0 is strictly convex.

2. Optimal control of hereditary differential systems towards target sets in \mathbb{R}^n .

Since a large number of papers has been published on this problem (cf. [8]), some of which expose similar optimality conditions, we only briefly indicate the kind of results which can be derived from § 1:

Consider a nonlinear HDS of type (1) which is to be guided from an initial datum φ towards a surface $\{x \in \mathbb{R}^n; g(x) = 0\}$, $g: \mathbb{R}^n \rightarrow \mathbb{R}^m$ within a fixed time interval $[t_0, t_1]$. Suppose that an integral cost criterion $F_0(x, u) = \int_{t_0}^{t_1} f_0(x(t), u(t), t) dt$ and a convex set $\Omega \subset \mathbb{R}^r$, $\text{int } \Omega \neq \emptyset$ of admissible control values are given. In order to apply the

theorems of § 1, choose the function spaces X and U of the trajectories and the control functions, respectively, as:

$$X := \mathcal{C}([t_0 - h, t_1], \mathbb{R}^n) \quad \text{and} \quad U := L_\infty([t_0, t_1], \mathbb{R}^r).$$

Then the following results are obtained by translation of Theorems 1.1, 1.2 into the present context:

RESULT 2.1. *Under the standard differentiability assumptions for f_0 , f_1 and g , a pointwise local maximum principle is derived which generalizes the local maximum principle for ordinary differential systems (cf. [18]).*

RESULT 2.2. *Nondegeneracy of the maximum principle is established under the following assumptions:*

- (a) *The set of admissible control functions $Q = \{u(\cdot) \in U; u(t) \in \Omega \text{ a.e.}\}$ contains in its interior a variation $u^0(\cdot) + \tilde{u}(\cdot)$ of the optimal control $u^0(\cdot)$ such that the trajectory of the linearized system corresponding to $\tilde{u}(\cdot)$:*

$$(2.1) \quad \begin{aligned} \dot{\tilde{x}}(t) &= D_1 f_1(x_t^0, u^0(t), t) \tilde{x}_t + D_2 f_1(x_t^0, u^0(t), t) \tilde{u}(t) \quad \text{on } [t_0, t_1], \\ \tilde{x}_{t_0} &\equiv 0 \end{aligned}$$

satisfies $Dg(x^0(t_1))\tilde{x}(t_1) = 0$.

- (b) *The linearized system (2.1) is output controllable at time t_1 in the following sense:*

$$\forall d \in \mathbb{R}^m \exists u \in U \exists x \in X: (x, u) \text{ solves (2.1) and}$$

$$Dg(x^0(t_1))(x(t_1)) = d.$$

RESULT 2.3. *For linear HDS with convex cost functional and affine end condition it is shown that the nondegenerate maximum principle is a necessary and sufficient optimality criterion, if condition (a) in Result 2.2 is assumed. Furthermore, the maximum principle may be extended to linear HDS with pointwise and distributed lags in the control as in [11].*

These results differ only in nuances (e.g. the conditions for nondegeneracy) from those in the literature (e.g., see [1], [3], [11], [27], [31]). The reader interested in explicit formulations and proofs is referred to [12].

However, it should be recognized that the optimal control problem with finite dimensional end condition cannot claim the same importance for HDS as for ordinary differential systems: Problem (P1) neglects the hereditary effects which may force the system to leave the desired value $x(t_1) = x_1$ after reaching it. Therefore, in the following section, we shall treat in more detail problem (P2) which does conform with the infinite dimensional character of HDS.

3. Optimal control of functional differential systems with function space terminal condition. In this section we consider optimal control problems with linear system equation and function space boundary condition. See Remark 3.2 (below) for the treatment of partially nonlinear systems.

We start with the following problem for HDS which will be generalized later to include systems governed by Fredholm equations.

Problem 3.1. Minimize $\int_{t_0}^1 f_0(x(t), u(t), t) dt$ subject to the constraints:

$$(3.1) \quad \begin{aligned} \dot{x}(t) &= A(t)x_t + B(t)u(t), & t \in [t_0, t_1], \\ x(t) &= 0, & t \in [t_0 - h, t_0], \end{aligned}$$

$$(3.2) \quad x(t_1 + s) = z(s), \quad s \in [-h, 0],$$

where $h \geq 0$, $f_0: \mathbb{R}^n \times \mathbb{R}^r \times [t_0, t_1] \rightarrow \mathbb{R}$, $A(t): \mathcal{C}([-h, 0], \mathbb{R}^n) \rightarrow \mathbb{R}^n$ is continuous linear, $B(t) \in \mathcal{L}_{nr}$ for $t \in [t_0, t_1]$, and $z(\cdot): [-h, 0] \rightarrow \mathbb{R}^n$ is a fixed target function.

Following the suggestion of Jacobs–Kao [21], we choose for Z a Sobolev space

$$Z = W^{1,p}([-h, 0], \mathbb{R}^n), \quad 1 \leq p \leq \infty,$$

where

$$W^{1,p}([\alpha, \beta], \mathbb{R}^n) = \{x \in AC([\alpha, \beta], \mathbb{R}^n) \mid \dot{x} \in L_p\}.$$

This is a Banach space with respect to the norm

$$\|x\| := |x(\alpha)| + \|\dot{x}\|_{L_p}.$$

The Lagrange multiplier corresponding to the terminal condition is an element of $(W^{1,p})^*$. While $(W^{1,p})^* \simeq W^{1,q}$ for $1 \leq p < \infty$, $p^{-1} + q^{-1} = 1$, the space $(W^{1,\infty})^*$ cannot be identified with a space of real valued functions (cf. [17, IV, 8.6]). Hence we exclude $p = \infty$.

Kurcyusz–Olbrod [24] analyzed linear systems with constant lag assuming that the coefficient matrices depend analytically on time. Supposing that the control functions are taken from $L_{\tilde{p}}$, they showed that the attainable subspace \mathcal{A} is closed in $W^{1,p}$ ($\tilde{p} \geq p \geq 1$) only if $p = \tilde{p}$. Hence it is not possible to choose $\tilde{p} = \infty$, i.e. $U = L_\infty([t_0, t_1], \mathbb{R}^r)$, since we have excluded $p = \infty$. Therefore we shall treat the control problem in the spaces

$$(3.3) \quad X := \underline{W}^{1,p}([t_0, t_1], \mathbb{R}^n), \quad U := L_p([t_0, t_1], \mathbb{R}^r), \quad Z := W^{1,p}([-h, 0], \mathbb{R}^n)$$

where¹ $2 \leq p < \infty$ and

$$\underline{W}^{1,p}([t_0, t_1], \mathbb{R}^n) := \{x \in W^{1,p}([t_0 - h, t_1], \mathbb{R}^n) \mid x|_{[t_0 - h, t_0]} = 0\}.$$

The choice of L_p instead of L_∞ entails that pointwise constraints for the controls be excluded since the abstract maximum principle requires that the set of admissible control functions has nonempty interior. $\{u \in L_p \mid u(t) \in \Omega \text{ a.e.}\}$ may not have any interior point although $\text{int } \Omega \neq \emptyset$ in \mathbb{R}^r . Therefore we are only allowed to consider control restrictions referring to the whole function, for example energy restrictions of the form $\int_{t_0}^{t_1} |u(t)|^2 dt \leq \alpha$ ($\alpha > 0$).

Let us analyze (3.1) in some more detail. We regard $A(s)$ as a continuous linear map of $X = \underline{W}^{1,p}([t_0, t_1], \mathbb{R}^n)$ into \mathbb{R}^n defining

$$(3.4) \quad A(s)x := A(s)x_s, \quad x \in X, \quad s \in [t_0, t_1].$$

Identifying $(\underline{W}^{1,p})^*$ with $\underline{W}^{1,q}$, $p^{-1} + q^{-1} = 1$, we may assume $A(s) \in \underline{W}^{1,q}([t_0, t_1], \mathcal{L}_{nn})$ for all $s \in [t_0, t_1]$.

LEMMA 3.1. *For every $A \in L_q([t_0, t_1], \underline{W}^{1,q}([t_0, t_1], \mathcal{L}_{nn}))$ there is a unique $n \times n$ matrix function $\eta \in L_q([t_0, t_1] \times [t_0, t_1], \mathcal{L}_{nn})$ such that*

$$(3.5) \quad A(s)x = \int_{t_0}^{t_1} \eta(s, t) \dot{x}(t) dt, \quad x \in X, \quad s \in [t_0, t_1].$$

If A is defined by (3.4), η satisfies

$$(3.6) \quad A(s)x = \int_{t_0}^s \eta(s, t) \dot{x}(t) dt.$$

¹ The condition $p \geq 2$ is needed to guarantee Fréchet-differentiability of the cost functional.

Proof. The derivation operator $D: W^{1,q} \rightarrow L_q$, $Dx = \dot{x}$, is a linear isometry of $W^{1,q}$ onto L_q . Let $A_i(s) \in W^{1,q}([t_0, t_1], (\mathbb{R}^n)^*)$ be the i th row-vector function of the matrix function $A(s)$ ($i = 1, \dots, n$). Then $D \circ A_i \in L_q([t_0, t_1], L_q([t_0, t_1], (\mathbb{R}^n)^*))$. Reasoning as Dunford–Schwartz [17, III. 11, Lemma 16 and Thm. 17], we obtain uniquely determined functions $\eta_i \in L_q([t_0, t_1] \times [t_0, t_1], (\mathbb{R}^n)^*)$ ($i = 1, \dots, n$), satisfying $\eta_i(s, \cdot) = D \circ A_i(s)$ for a.e. $s \in [t_0, t_1]$, i.e.

$$(3.7) \quad \begin{aligned} A_i(s)x &= \langle A_i(s), x \rangle_{W^p} = \langle D \circ A_i(s), \dot{x} \rangle_{L_p} = \langle \eta_i(s, \cdot), \dot{x} \rangle_{L_p} \\ &= \int_{t_0}^{t_1} \langle \eta_i(s, t), \dot{x}(t) \rangle_{\mathbb{R}^n} dt. \end{aligned}$$

If we define $\eta(s, t)$ to be the matrix composed of the row vectors $\eta_i(s, t)$ ($s, t \in [t_0, t_1]$), η evidently satisfies (3.5). If $A(s)x$ depends only on the values of x on $[s-h, s]$ according to (3.4) it follows from (3.7) that $\eta_i(s, t)$ is zero for $t > s$. \square

The following conclusion is immediate: $x \in X$ is a solution of (3.1) iff \dot{x} is the solution of the following Volterra equation of the second kind [39]:

$$(3.8) \quad v(s) = \int_{t_0}^s \eta(s, t)v(t) dt + B(s)u(s), \quad s \in [t_0, t_1].$$

We see that the HDS (3.1) may be described by a *Volterra equation*. It seems natural to generalize our analysis to systems described in the same way by *Fredholm equations*. These Fredholm systems are noncausal (anticipatory), if we continue to interpret t as time. Dynamical systems involving both retarded and advanced effects are employed as models in classical relativistic mechanics [35] and electro-dynamics [16]. Better known is the use of Fredholm equations for the study of two point boundary problems in mathematical physics [13] (e.g. deformation of an elastic rod). In this case t has to be interpreted as a spatial variable and the control is some stationary external force.

In the following, we shall deal with Fredholm systems described by an equation of the form

$$(3.9) \quad \dot{x}(s) = A(s)x + y(s), \quad s \in [t_0, t_1],$$

where $y \in L_p([t_0, t_1], \mathbb{R}^n)$ and

$$(3.10) \quad A \in L_\infty([t_0, t_1], W^{1,q}([t_0, t_1], \mathcal{L}_{nn}));$$

or equivalently, by the Fredholm equation

$$(3.11) \quad \dot{x}(s) = \int_{t_0}^{t_1} \eta(s, t)\dot{x}(t) dt + y(s), \quad s \in [t_0, t_1],$$

where η is uniquely determined by A according to Lemma 3.1. While (3.8) always has a unique solution $v \in L_p([t_0, t_1], \mathbb{R}^n)$ for every $u \in L_p([t_0, t_1], \mathbb{R}^r)$, we need an additional assumption to assure unique solubility of (3.11).

The linear operator $\hat{A}: L_p([t_0, t_1], \mathbb{R}^n) \rightarrow L_p([t_0, t_1], \mathbb{R}^n)$, defined by

$$(3.12) \quad (\hat{A}v)(s) = \int_{t_0}^{t_1} \eta(s, t)v(t) dt, \quad s \in [t_0, t_1], \quad v \in L_p,$$

is compact (cf. [17, VI.9.53]).

According to Fredholm's alternative, equation (3.11) has a unique solution $\dot{x}(\cdot) \in L_p$ for every $y(\cdot) \in L_p$ iff the following condition is satisfied:

(3.13) The homogeneous Fredholm equation

$$v(s) = \int_{t_0}^{t_1} \eta(s, t)v(t) dt, \quad s \in [t_0, t_1]$$

has only the trivial solution $v \equiv 0$ in $L_p([t_0, t_1], \mathbb{R}^n)$.

If we assume (3.13), the Fredholm operator $\text{Id}_{L_p} - \hat{A}$ is injective and of index 0, hence bijective. Therefore the solution operator

$$(3.14) \quad (D - \hat{A} \circ D)^{-1}: L_p([t_0, t_1], \mathbb{R}^n) \rightarrow W^{1,p}([t_0, t_1], \mathbb{R}^n)$$

of the equation $\dot{x}(s) = A(s)x + y(s)$, $x_{t_0} \equiv 0$ is well defined; it is continuous by the open mapping theorem.

We now generalize Problem 3.1 in order to include Fredholm systems with retarded controls:

Problem 3.2. Minimize $\int_{t_0}^{t_1} f_0(x(t), u(t), t) dt$ subject to the constraints

$$(3.15) \quad \dot{x}(t) = A(t)x + \sum_{i=1}^k B_i(t)u(t-h_i) + \int_{t_0}^{t_1} B_0(t, s)u(s) ds, \quad t \in [t_0, t_1],$$

$$x|_{[a, t_0]} = 0,$$

$$(3.16) \quad C(x|_{[b, t_1]}) = z,$$

$$(3.17) \quad u \in Q,$$

where $a \leq t_0 \leq b \leq t_1$, $f_0: \mathbb{R}^n \times \mathbb{R}^r \times [t_0, t_1] \rightarrow \mathbb{R}$,
 $A \in L_\infty([t_0, t_1], W^{1,q}([t_0, t_1], \mathcal{L}_{nn}))$,
 $B_i \in L_\infty(\mathbb{R}, \mathcal{L}_{nr})$, $B_i = 0$ outside $[t_0, t_1]$ for $i = 1, \dots, k$,
 $B_0 \in L_\infty([t_0, t_1] \times [t_0, t_1], \mathcal{L}_{nr})$,
 $C \in \mathcal{L}(W^{1,p}([b, t_1], \mathbb{R}^n), W^{1,p}([b, t_1], \mathbb{R}^n))$,
 $z \in W^{1,p}([b, t_1], \mathbb{R}^n)$,
 $Q \subset L_p([t_0, t_1], \mathbb{R}^r)$ are fixed.

We assume $u = 0$ outside $[t_0, t_1]$.

The following theorem is obtained by application of Theorem 1.2 to Problem 3.2.

THEOREM 3.1. Let $(x^0, u^0) \in W^{1,p}([a, t_1], \mathbb{R}^n) \times L_p([t_0, t_1], \mathbb{R}^r)$ satisfy the constraints of Problem 3.2 and assume the following conditions:

- (a) $f_0(x, u, t)$ is continuously Fréchet-differentiable and convex with respect to (x, u) , measurable in t and for every $K > 0$ there exist $m_1(\cdot) \in L_1([t_0, t_1], \mathbb{R})$, $m_2 \in \mathbb{R}_+$, $m_3(\cdot) \in L_q([t_0, t_1], \mathbb{R})$ such that

$$|f_0(x, u, t)| + |D_1 f_0(x, u, t)| \leq m_1(t) + m_2|u|^p,$$

$$|D_2 f_0(x, u, t)| \leq m_3(t) + m_2|u|^{p-1}$$

for all $x \in \mathbb{R}^n$, $|x| \leq K$, all $u \in \mathbb{R}^r$ and a.e. $t \in [t_0, t_1]$.

- (b) Condition (3.13) is satisfied with η defined as in Lemma 3.1.

- (c) $\mathcal{A} := \{C(x|_{[b, t_1]}) | \exists u \in L_p([t_0, t_1], \mathbb{R}^r): (x, u) \text{ satisfies (3.15)}\}$ is a closed linear subspace of $W^{1,p}([b, t_1], \mathbb{R}^n)$.

- (d) $Q \subset L_p([t_0, t_1], \mathbb{R}^r)$ is convex and has nonempty interior.

- (e) There is $(\tilde{x}, \tilde{u}) \in W^{1,p}([a, t_1], \mathbb{R}^n) \times L_p([t_0, t_1], \mathbb{R}^r)$ satisfying (3.15), (3.16) and $\tilde{u} \in \text{int } Q$.

- (f) $\text{Im } C$ is closed.

Then (x^0, u^0) is optimal iff there exist $\psi, \rho \in L_q([t_0, t_1], \mathbb{R}^n)$, $p^{-1} + q^{-1} = 1$, $\rho|_{[t_0, b]}$ constant such that

$$(3.18) \quad \psi(t) = \int_t^{t_1} D_1 f_0(x^0(s), u^0(s), s) ds + \int_{t_0}^{t_1} \eta(s, t)^* \psi(s) ds + \rho(t) \quad \text{a.e. on } [t_0, t_1]$$

$$(3.19) \quad \int_{t_0}^{t_1} \left\langle D_2 f_0(x^0(t), u^0(t), t) + \sum_{i=1}^k B_i(t + h_i)^* \psi(t + h_i) + \int_{t_0}^{t_1} B_0(s, t)^* \psi(s) ds, u^0(t) - u(t) \right\rangle_{\mathbb{R}^r} dt \leq 0 \quad \text{for all } u \in Q.$$

(3.20) The function $\rho(b) + \int_b^t \rho(s) ds$, $t \in [b, t_1]$, is orthogonal to $\text{Ker } C$.

Proof. Let $X := \underline{W}^{1,p}([t_0, t_1], \mathbb{R}^n)$, $U := L_p([t_0, t_1], \mathbb{R}^r)$, $Z := W^{1,p}([b, t_1], \mathbb{R}^n)$, and define

$$\begin{aligned} F_0(x, u) &= \int_{t_0}^{t_1} f_0(x(t), u(t), t) dt, \\ (\tilde{A}x)(t) &:= \int_{t_0}^t A(s)x ds, \quad t \in [t_0, t_1], \\ (\tilde{B}u)(t) &:= \int_{t_0}^t \left[\sum_{i=1}^k B_i(s)u(s - h_i) + \int_{t_0}^{t_1} B_0(s, \tau)u(\tau) d\tau \right] ds, \quad t \in [t_0, t_1], \\ \tilde{C}x &:= C(x|_{[b, t_1]}) \end{aligned}$$

for $x \in X$, $u \in U$.

Then Problem 3.2 is equivalent to

$$\begin{aligned} &\text{Minimize} && F_0(x, u) \text{ on } X \times U \\ &\text{subject to} && x = \tilde{A}x + \tilde{B}u, \\ & && \tilde{C}x = z, \\ & && u \in Q. \end{aligned}$$

Using (a) we have to prove that F_0 is well-defined and Fréchet-differentiable in (x, u) . By use of Lebesgue's theorem on dominated convergence, it may be shown that F_0 is continuously differentiable in x . It remains to prove continuous differentiability in u . This is much more difficult (see [34, Thm. 21.1]) and requires application of a theorem of Gavurin. The operators $\tilde{A}: X \rightarrow X$, $\tilde{B}: U \rightarrow X$ and $\tilde{C}: X \rightarrow Z$ are well-defined, continuous and linear. $\tilde{A} = D^{-1} \circ \hat{A} \circ D$ is a compact linear operator on $\underline{W}^{1,p}$ (cf. (3.12)) satisfying $\text{Ker}(\text{Id}_{\underline{W}^{1,p}} - \tilde{A}) = 0$ by (b).

The other conditions of Theorem 1.2 follow immediately from (c)–(e).

Thus (x^0, u^0) is a solution of Problem 3.2 iff there are $x^* \in X^*$, $z^* \in Z^*$ with

$$(3.21) \quad x^* = D_1 F_0(x^0, u^0) + \tilde{A}^* x^* + \tilde{C}^* z^*,$$

$$(3.22) \quad [D_2 F_0(x^0, u^0) + \tilde{B}^* x^*](u^0 - u) \leq 0 \quad \text{for all } u \in Q.$$

$(\underline{W}^{1,p})^*$ and $(L^p)^*$ are identified with $\underline{W}^{1,q}$ and L^q , respectively ($p^{-1} + q^{-1} = 1$).

Computation of the operators in (3.21) and (3.22) yields

$$\begin{aligned} D_1 F_0(x^0, u^0)(t) &= \int_{t_0}^t \int_{\tau}^{t_1} D_1 f_0(x^0(s), u^0(s), s) ds d\tau, \quad t \in [t_0, t_1]; \\ (\tilde{A}^* x^*)(t) &= \int_{t_0}^t \int_{t_0}^{t_1} \eta(s, \tau)^* x^*(s) ds d\tau, \quad t \in [t_0, t_1]; \\ (\tilde{C}^* z^*)(t) &= \begin{cases} (t - t_0)(C^* z^*)(b), & t \in [t_0, b], \\ (b - t_0 - 1)(C^* z^*)(b) + (C^* z^*)(t), & t \in [b, t_1]; \end{cases} \\ D_2 F_0(x^0, u^0)(t) &= D_2 f_0(x^0(t), u^0(t), t), \quad t \in [t_0, t_1]; \\ (\tilde{B}^* x^*)(t) &= \sum_{i=1}^k B_i(t + h_i)^* \frac{dx^*}{dt}(t + h_i) + \int_{t_0}^{t_1} B_0(s, t)^* \frac{dx^*}{ds}(s) ds, \quad t \in [t_0, t_1]. \end{aligned}$$

Define

$$\begin{aligned} \psi(t) &:= \frac{d}{dt} x^*(t), \quad t \in [t_0, t_1]; \\ \rho(t) &:= \frac{d}{dt} (\tilde{C}^* z^*)(t) = \begin{cases} (C^* z^*)(b), & t \in [t_0, b], \\ \frac{d}{dt} (C^* z^*)(t), & t \in (b, t_1]. \end{cases} \end{aligned}$$

Then (3.21) and (3.22) yield (3.18) and (3.19). The transversality condition (3.20) is a consequence of the definition of \tilde{C} and ρ .

Conversely (3.18)–(3.20) imply the existence of x^* , z^* satisfying (3.21) and (3.22), because by (f): $\text{Im } C^* = (\text{Ker } C)^\perp$. \square

Remark 3.1. Convexity of f_0 and condition (f) are needed only for sufficiency, while conditions (c), (e) and $\text{int } Q \neq \emptyset$ are needed only for necessity.

Remark 3.2. It is possible to generalize Theorem 3.1 to nonlinear systems by application of the abstract maximum principle (Theorem 1.1); but the control u must appear linearly in the system's equation to ensure Fréchet-differentiability of F_1 (see [22], [33]). In fact, suppose the system's equation is given as in (1) by

$$\dot{x}(t) = f_1(x, u(t), t), \quad t \in [t_0, t_1].$$

If $F_1: W^{1,p} \times L_p \rightarrow W^{1,p}$, defined by $F_1(x, u)(t) = \int_{t_0}^t f_1(x_\tau, u(\tau), \tau) d\tau$, is Fréchet-differentiable, then for any fixed $x(\cdot) \in W^{1,p}$ the map

$$\begin{aligned} u &\mapsto \left(t \mapsto \int_{t_0}^t f_1(x_\tau, u(\tau), \tau) d\tau \right) \\ L_p([t_0, t_1], \mathbb{R}^r) &\rightarrow W^{1,p}([t_0, t_1], \mathbb{R}^n) \end{aligned}$$

and hence the map

$$\begin{aligned} u &\mapsto (t \mapsto f_1(x_t, u(t), t)) \\ L_p([t_0, t_1], \mathbb{R}^r) &\rightarrow L_p([t_0, t_1], \mathbb{R}^n) \end{aligned}$$

must be Fréchet-differentiable.

Now the latter mapping is a superposition operator for which Vainberg [33, pp. 90–91] has proved (for $p = 2$, $r = n = 1$) the following:

LEMMA. Suppose $G: L_p([t_0, t_1], \mathbb{R}^r) \rightarrow L_p([t_0, t_1], \mathbb{R}^n)$ is given by

$$G(u)(t) = g(u(t), t), \quad \text{a.e. } t \in [t_0, t_1]$$

where $g: \mathbb{R}^r \times [t_0, t_1] \rightarrow \mathbb{R}^n$. Then G is Fréchet-differentiable iff $g(\cdot, t)$ is affine for a.e. $t \in [t_0, t_1]$.

Thus F_1 may be expected to be Fréchet-differentiable only if it is linear in u . Therefore, necessary conditions analogous to Theorem 3.1 may be derived from Theorem 1.1 only for Fredholm systems described by equations of the following form (neglecting delays in the control):

$$\dot{x}(t) = f(x, t) + g(x, t)u(t)$$

where $f: W^{1,p} \times [t_0, t_1] \rightarrow \mathbb{R}^n$, $g: W^{1,p} \times [t_0, t_1] \rightarrow \mathcal{L}_{nr}$.

Conditions on f, g which are sufficient for F_1 to be continuously Fréchet-differentiable can be found in [22], [40].

Remark 3.3. Comparison of Theorem 3.1 with the maximum principle known for \mathbb{R}^n -targets (cf. § 2) shows the following differences:

- in Theorem 3.1 the minimum condition can be established only in integral form as distinguished from the usual pointwise form;
- in the case of \mathbb{R}^n -targets, the solution ψ of the adjoint equation is known to be of bounded variation on $[t_0, t_1]$, left continuous on $(t_0, t_1]$, continuous at t_0 . The corresponding function ψ in Theorem 3.1 is only in L^q . But suppose that for $s \in [t_0, t_1]$, $A(s): W^{1,p}([t_0, t_1], \mathbb{R}^n) \rightarrow \mathbb{R}^n$ can be extended to a continuous linear function on $\mathcal{C}([t_0, t_1], \mathbb{R}^n)$. Then $\eta(s, \cdot)$ is of bounded variation and thus, since $\rho|_{[t_0, b]}$ is constant, ψ is of bounded variation on $[t_0, b]$.

Remark 3.4. For HDS the adjoint equation (3.18) has the form

$$\psi(t) = \int_t^{t_1} D_1 f_0(x^0(s), u^0(s), s) ds + \int_t^{t_1} \eta(s, t)^* \psi(s) ds + \rho(t)$$

because $\eta(s, t) = 0$ for $s < t$.

Problem 3.2 includes also problems with lagged controls, where $h_i \geq 0$ for $i = 1, \dots, k$ and $B_0 \equiv 0$.

Remark 3.5. The terminal condition $C(x|_{[b, t_1]}) = z$ is rather flexible and includes the following cases:

- 1) $C = 0$ and $z = 0$: no terminal condition.
- 2) $b = t_1$: finite dimensional terminal condition as in § 2.
- 3) suppose, it is required that x vanishes in l fixed time points $t^{(i)} \in [b, t_1]$, $i = 1, \dots, l$. Let $p = 2$, so $Z = W^{1,2}$ is a Hilbert space, and define

$$V = \{y \in W^{1,2}([b, t_1], \mathbb{R}^n) : y(t^{(i)}) = 0, i = 1, \dots, l\}.$$

V is a closed subspace of $W^{1,2}$ with finite dimensional orthogonal complement V^\perp . If we take C as the projection of Z onto V^\perp , the desired terminal condition is described by

$$C(x|_{[b, t_1]}) = 0.$$

- 4) $C = \text{Id}_{W^{1,p}([b, t_1], \mathbb{R}^n)}$: fixed target function as in Problem 3.1.

In the first three cases \mathcal{A} is clearly a closed subspace of $W^{1,p}$. In the fourth case \mathcal{A} is closed only under additional assumptions which will be studied later.

The work we know which has been done on control problems with function space terminal conditions is restricted to HDS, which are sometimes allowed to be of neutral type, i.e. have delays in \dot{x} , too.

It is easy to show that the adjoint equation (3.18) coincides with those in [3], [4], [7], [10], [21], [22] for the corresponding classes of HDS (cf. [8]). We remark that the maximum principle in [21] is true only under the additional assumption that the control u appears linearly in the system's equation (cf. Remark 3.2 above).

The necessary conditions derived by Banks–Kent [7] for quite general neutral HDS are not totally satisfactory, because nontriviality is not guaranteed; however, assuming nondegeneracy, they established sufficiency. Theorem 3.1 gives conditions, which imply nondegeneracy. Bien [10] has tried to solve Problem 3.1 by transforming the functional end condition into a mixed control-phase variable equality constraint plus a finite dimensional end condition. He then applied the Neustadt–Makowski theory [26]. But in order to get Lagrange multipliers, which can be identified with functions on $[t_0, t_1]$, a rather strong regularity condition is needed.

For a discussion of results on optimal control problems involving integral equations of Volterra type compare [8], [27]. See also [36], [37], resp. [38], for a treatment of dynamical systems described by functional integral equations and Fredholm integral equations. In [38], it is shown that under additional assumptions, it is not necessary to assume unique solubility (3.13). However, all these papers on integral equations only apply to problems with finite dimensional end conditions.

The crucial assumption of our Theorem 3.1 is (c): \mathcal{A} has to be closed in the infinite-dimensional B-space $Z = W^{1,p}([b, t_1], \mathbb{R}^n)$. The counterexample (cf. § 1) of Kurcyusz [22] shows that this assumption may not be weakened. If \mathcal{A} is not closed (resp. dense and not closed), a quadratic real function f_0 exists, so that the unique solution of the corresponding optimal control problem, Problem 3.2, does not satisfy the nondegenerate maximum principle (3.18)–(3.20) (resp. only satisfies the trivial form of the maximum principle with $\lambda_0 = 0$, $\psi = 0$, $\rho = 0$).

The next section will be concerned with the problem of finding necessary and/or sufficient conditions for \mathcal{A} to be closed or to equal Z . The equality $\mathcal{A} = Z$ which represents a very restrictive attainability condition (Proposition 4.3 below), is necessary in order to derive a nondegenerated maximum principle for nonlinear HDS (see Theorem 1.1).

4. Analysis of the attainable subspace. We exclude time-delays of the control function and consider the system

$$(4.1) \quad \begin{aligned} \dot{x}(t) &= A(t)x + B(t)u(t), & t \in [t_0, t_1], \\ x|_{[a, t_0]} &= 0 \end{aligned}$$

where $u \in U = L^p([t_0, t_1], \mathbb{R}^r)$, $1 < p < \infty$, $a \leq t_0 < b \leq t_1$.

In order to simplify the analysis, we assume

$$\mathcal{A} = \{x|_{[b, t_1]} \mid \exists u \in U, (x, u) \text{ satisfies (4.1)}\},$$

that is, we regard the problem of control towards a fixed target function in Z . We presuppose (3.10), (3.13) and $B \in L_\infty([t_0, t_1], \mathcal{L}_{nr})$. Hence (4.1) has a unique solution for each $u \in U$.

We decompose the operator $A(s)$, $s \in [t_0, t_1]$, into two additive components (cf. Lemma 3.1)

$$A^1(s): W^{1,p}([t_0, b], \mathbb{R}^n) \rightarrow \mathbb{R}^n, \quad A^1(s)x = \int_{t_0}^b \eta(s, t)\dot{x}(t) dt,$$

$$A^2(s): W^{1,p}([b, t_1], \mathbb{R}^n) \rightarrow \mathbb{R}^n, \quad A^2(s)x = \int_b^{t_1} \eta(s, t)\dot{x}(t) dt.$$

Then $A(s)x = A^1(s)(x|_{[t_0, b]}) + A^2(s)(x|_{[b, t_1]})$, for $x \in X = W^{1,p}([t_0, t_1], \mathbb{R}^n)$.

Throughout the rest of this section we assume that A^1, B satisfy the following two conditions:

$$(4.2) \quad (t \rightarrow \|B(t)^+ A^1(t)\|) \in L_p([b, t_1], \mathbb{R}),$$

where $B(t)^+$ is the generalized inverse of the matrix $B(t)$ (Penrose [30]) and

$$\|B(t)^+ A^1(t)\| := \sup \{\|B(t)^+ A^1(t)x\| \mid x \in W^{1,p}, \|x\| \leq 1\}.$$

(4.3) The equation

$$\dot{x}(s) = A^1(s)x \mid [t_0, b] + w(s), \quad s \in [t_0, b], \quad x \mid [a, t_0] = 0$$

has a unique solution x for each $w \in L_p([t_0, b], \mathbb{R}^n)$.

The following hypotheses will be referred to in the sequel:

(H1) For all $d \in \mathbb{R}^n$ there is $v \in L_p([t_0, b], \mathbb{R}^r)$ such that the solution y of

$$y \mid [a, t_0] = 0, \quad \dot{y}(t) = A^1(t)y + B(t)v(t), \quad t \in [t_0, b],$$

satisfies

$$y(b) = d.$$

(H2) $\text{Im } A^1(t) \subset \text{Im } B(t)$ for a.e. $t \in [b, t_1]$.

(H3) $B(t)^+$ is bounded a.e. on $[b, t_1]$.

(H4) $\text{Rank } B(t) = n$ for a.e. $t \in [b, t_1]$.

Some comments on these hypotheses are appropriate: (H1) is equivalent to complete pointwise attainability at time b for HDS. (H2) is equivalent to

$$(4.4) \quad A^1(t)x = B(t)B(t)^+ A^1(t)x \quad \text{for all } x \in W^{1,p}([t_0, b], \mathbb{R}^n), \quad \text{a.e. } t \in [b, t_1],$$

i.e. $A^1(t)$ factors through $B(t)$ for a.e. t .

Condition (4.2) implies that

$$(t \mapsto B(t)^+ A^1(t)x) \in L_p([b, t_1], \mathbb{R}^r).$$

Thus (H2) admits the following intuitive interpretation: The hereditary effects on $\dot{x} \mid [b, t_1]$, produced by the values $x(s)$, $s \in [t_0, b]$ via A^1 , can be compensated by suitable control functions in L_p .

If $B(\cdot)$ is continuous, Kurcyusz-Olbrot [24] have shown that (H3) is satisfied iff $\text{Rank } B(t)$ is constant on $[b, t_1]$.

(H4) is a very strong condition, requiring in particular that the dimension r of the control space is not less than the dimension n of the phase space. Evidently (H4) implies (H2).

For HDS, (H1)–(H3) correspond to the hypotheses (H1)–(H3) in [6], letting $b = t_1 - h$.

PROPOSITION 4.1. *If the hypotheses (H1), (H4) are valid, \mathcal{A} is dense in $W^{1,p}([b, t_1], \mathbb{R}^n)$. Conversely, if \mathcal{A} is dense in $W^{1,p}([b, t_1], \mathbb{R}^n)$, (H1) must be valid.*

Proof. If \mathcal{A} is dense in $W^{1,p}([b, t_1], \mathbb{R}^n)$ and $d \in \mathbb{R}^n$, a sequence (u_k) exists in $L_p([t_0, t_1], \mathbb{R}^r)$ such that the corresponding solutions (x_k) of (4.1) converge to the constant function $z(t) \equiv d$ in $W^{1,p}([b, t_1], \mathbb{R}^n)$. Hence by (3.10)

$$(4.5) \quad x_k(b) \rightarrow d \quad \text{and} \quad A^2(t)x_k \mid [b, t_1] = \int_b^{t_1} \eta(t, s)\dot{x}_k(s) ds \rightarrow 0$$

uniformly in $t \in [b, t_1]$, if $k \rightarrow \infty$. Let T be the solution operator of the equation in (4.3). Then

$$x_k|_{[t_0, b]} = T(A^2(\cdot)x_k|_{[b, t_1]} + B(\cdot)u_k(\cdot)).$$

Since T is continuous (cf. (3.14)), we obtain for $z_k := T(B(\cdot)u_k(\cdot))$ by (4.5)

$$\|z_k - x_k|_{[t_0, b]}\| = \|T(A^2(\cdot)x_k|_{[b, t_1]})\| \rightarrow 0$$

where the norm is taken in $W^{1,p}([t_0, b], \mathbb{R}^n)$.

Hence $\lim_{k \rightarrow \infty} z_k(b) = d$. (H1) follows, because the only dense linear subspace of \mathbb{R}^n is the whole space.

Now assume (H1) and (H4) and let $z \in W^{1,p}([b, t_1], \mathbb{R}^n)$. We shall construct a sequence $(x_k, u_k) \in W^{1,p}([t_0, t_1], \mathbb{R}^n) \times L_p([t_0, t_1], \mathbb{R}^r)$, satisfying the system's equation such that $(x_k|_{[b, t_1]})$ converges to z .

Condition (4.3) implies the existence of \tilde{y} with

$$\tilde{y}|_{[a, t_0]} = 0, \quad \dot{\tilde{y}}(s) = A^1(s)\tilde{y} + A^2(s)z, \quad s \in [t_0, b].$$

By (H1) there are \hat{y}, v with

$$\hat{y}|_{[a, t_0]} = 0, \quad \dot{\hat{y}}(s) = A^1(s)\hat{y} + B(s)v(s), \quad s \in [t_0, b],$$

and

$$\hat{y}(b) = z(b) - \tilde{y}(b).$$

Define $y = \hat{y} + \tilde{y}$. Then

$$y|_{[a, t_0]} = 0, \quad \dot{y}(s) = A^1(s)y + B(s)v(s) + A^2(s)z, \quad s \in [t_0, b]$$

and $y(b) = z(b)$.

Define $x \in X := W^{1,p}([t_0, t_1], \mathbb{R}^n)$ by

$$x(s) = \begin{cases} y(s) & \text{for } s \in [a, b], \\ z(s) & \text{for } s \in [b, t_1], \end{cases}$$

and a measurable, not necessarily integrable function w by

$$w(s) = \begin{cases} v(s), & s \in [t_0, b], \\ B(s)^*[B(s)B(s)^*]^{-1}[\dot{x}(s) - A(s)x], & s \in [b, t_1]. \end{cases}$$

w is a.e. defined by condition (H4). By definition we get $B(\cdot)w(\cdot) \in L_p([t_0, t_1], \mathbb{R}^n)$ and

$$\dot{x}(s) = A(s)x + B(s)w(s), \quad s \in [t_0, t_1].$$

Let $M_k := \{s \in [t_0, t_1] : |w(s)| < k\}$ and

$$w_k(s) := \begin{cases} w(s) & \text{for } s \in M_k, \\ 0 & \text{for } s \in [t_0, t_1] \setminus M_k. \end{cases}$$

Then $w_k \in L_p([t_0, t_1], \mathbb{R}^r)$ and $B(\cdot)w_k(\cdot)$ converges to $B(\cdot)w(\cdot)$ in $L_p([t_0, t_1], \mathbb{R}^n)$ if $k \rightarrow \infty$.

The solutions $x_k \in X$ of

$$x_k|_{[a, t_0]} = 0, \quad \dot{x}_k(s) = A(s)x_k + B(s)w_k(s), \quad s \in [t_0, t_1],$$

converge to x in X , since the solution operator of this equation is continuous by (3.14). Hence the sequence $(x_k|_{[b, t_1]})$ converges to z in $W^{1,p}([b, t_1], \mathbb{R}^n)$. \square

PROPOSITION 4.2. *Let (H2) be satisfied. Then \mathcal{A} is closed in $W^{1,p}([b, t_1], \mathbb{R}^n)$ iff (H3) is valid.*

Proof. We define the continuous multiplication operator \tilde{B} from $L_p([b, t_1], \mathbb{R}^r)$ to $L_p([b, t_1], \mathbb{R}^n)$ by

$$(\tilde{B}u)(t) := B(t)u(t), \quad t \in [b, t_1].$$

Kurcyusz–Olbrod [24] have shown that (H3) is satisfied (i.e. $B(\cdot)^+ \in L_\infty([b, t_1], \mathcal{L}_m)$), iff $\text{Im } \tilde{B}$ is closed. Suppose that (H3) is not valid. Then there exists a function $w \in \text{cl}(\text{Im } \tilde{B}) \setminus \text{Im } \tilde{B}$. Define x as solution of

$$x|[a, t_0] = 0, \quad \dot{x}(t) = A(t)x + \begin{cases} 0, & t \in [t_0, b], \\ w(t), & t \in [b, t_1]. \end{cases}$$

Since $w \in \text{cl}(\text{Im } \tilde{B})$ and the solution operator is continuous by (3.14), we get $x|[b, t_1] \in \text{cl } \mathcal{A}$. Let us suppose for a moment that $x|[b, t_1] \in \mathcal{A}$. In this case there is a pair $(y, u) \in X \times U$ with

$$y|[a, t_0] = 0, \quad \dot{y}(t) = A(t)y + B(t)u(t), \quad t \in [t_0, t_1],$$

and $y|[b, t_1] = x|[b, t_1]$. Then a.e. on $[b, t_1]$,

$$\begin{aligned} A^1(t)x|[t_0, b] + A^2(t)x|[b, t_1] + w(t) &= \dot{x}(t) = \dot{y}(t) \\ &= A^1(t)y|[t_0, b] + A^2(t)y|[b, t_1] + B(t)u(t). \end{aligned}$$

By (4.4), we obtain for a.e. $t \in [b, t_1]$:

$$\begin{aligned} w(t) &= A^1(t)(y - x)|[t_0, b] + B(t)u(t) \\ &= B(t)[B(t)^+ A^1(t)(y - x)|[t_0, b] + u(t)]. \end{aligned}$$

Condition (4.2) implies that

$$B(\cdot)^+ A^1(\cdot)(y - x)|[t_0, b] \in L_p([b, t_1], \mathbb{R}^r),$$

and thus $w \in \text{Im } \tilde{B}$. This contradiction shows that $x|[b, t_1] \in \text{cl}(\mathcal{A}) \setminus \mathcal{A}$ and hence \mathcal{A} is not closed.

Suppose now, conversely, that (H3) is satisfied and define for $\rho > 0$

$$\mathcal{A}^\rho := \{x|[b, t_1] \mid \exists u \in L_p: \|u\|_p \leq \rho \text{ and } (x, u) \text{ satisfies (4.1)}\}.$$

Let $(x^k|[b, t_1])$ be a sequence in $\mathcal{A}^\rho \subset Z$ corresponding to a sequence (u^k) in $\{u \in U: \|u\|_p \leq \rho\}$ and converging to $z \in Z$. Since the spheres in L_p are weakly compact, there is a subsequence (u^{k_i}) of (u^k) converging weakly to an element u of U . By the continuity of the solution operator (3.14) $x(u^{k_i})$ converges weakly to $x(u)$. Hence $z = x(u)|[b, t_1] \in \mathcal{A}^\rho$, because the limit is unique.

In the general case a sequence $(x^k|[b, t_1])$ in \mathcal{A} converging to $z \in W^{1,p}$ does not necessarily correspond to a bounded sequence of control functions. But, by appropriately generalizing Lemma 3.3 in [6] it is possible to derive, from (H2) and (H3), that there exists a bounded sequence of control functions yielding the same end functions $x^k|[b, t_1]$. Thus the general case may be reduced to the one just analyzed and the proposition is proved. \square

PROPOSITION 4.3. $\mathcal{A} = W^{1,p}([b, t_1], \mathbb{R}^n)$ iff (H1), (H3), (H4) are valid.

Proof. Suppose (H1), (H3) and (H4) are satisfied. Then \mathcal{A} is dense by Proposition 4.1 and closed by Proposition 4.2 since (H4) implies (H2).

Conversely, suppose $\mathcal{A} = W^{1,p}$. Then (H1) is satisfied. If (H4) is valid, (H3) follows by Proposition 4.2. It remains to prove the rank condition (H4).

For each $z \in W^{1,p}$ there is $(x, u) \in X \times U$ satisfying

$$(4.6) \quad \dot{z}(s) = \dot{x}(s) = A(s)x + B(s)u(s), \quad s \in [b, t_1].$$

We assume now that there is a compact subset $M \subset [b, t_1]$ of positive measure, such that $\text{Rank } B(s) < n$ for $s \in M$. Then there exists a measurable map $s \mapsto e(s)$ from M into the unit sphere of \mathbb{R}^n such that $B(s)B(s)^*e(s) = 0$ and hence $B(s)^*e(s) = 0$ a.e. on M (cf. [32]). The map $A: [t_0, t_1] \rightarrow W^{1,q}([t_0, t_1], \mathcal{L}_{nn})$ is measurable by assumption (3.10). Hence by Lusin's theorem (cf. [9]), there is a subset $N \subset M$ of positive measure such that $e|_N$ and $A|_N$ are continuous.

Let α be any function in $L_p(N, \mathbb{R})$. Define $f \in L_p([b, t_1], \mathbb{R}^n)$ by

$$f(s) = \begin{cases} 0, & \text{for } s \in [b, t_1] \setminus N, \\ \alpha(s)e(s) & \text{for } s \in N \end{cases}$$

and $z \in W^{1,p}([b, t_1], \mathbb{R}^n)$ by

$$z(s) = \int_b^s f(\tau) d\tau.$$

The scalar product of both sides of (4.6) with $e(s)$ yields, for a.e. $s \in N$,

$$\langle \alpha(s)e(s), e(s) \rangle_{\mathbb{R}^n} = \langle A(s)x, e(s) \rangle_{\mathbb{R}^n} + \langle B(s)u(s), e(s) \rangle_{\mathbb{R}^n}$$

i.e.,

$$\begin{aligned} \alpha(s) &= \langle A(s)x, e(s) \rangle_{\mathbb{R}^n} + \langle u(s), B(s)^*e(s) \rangle_{\mathbb{R}^r} \\ &= \langle A(s)x, e(s) \rangle_{\mathbb{R}^n}. \end{aligned}$$

This means that α is a.e. equal to a continuous function. This is a contradiction, because it may be shown that there is a function in $L_p(N, \mathbb{R})$ which is not almost everywhere equal to a continuous function. \square

COROLLARY. \mathcal{A} is a dense proper subspace of $W^{1,p}([b, t_1], \mathbb{R}^n)$, if (H1), (H4) are satisfied and (H3) is not.

Proposition 4.3 yields a full attainability criterion for HDS. In particular, we learn from it that it is impossible to steer hereditary differential systems from 0 to arbitrary target functions in $Z = W^{1,p}([t_1 - h, t_1], \mathbb{R}^n)$ if the number of input components r is smaller than n , the number of state components. Since the condition $r \geq n$ is rarely satisfied in practice, the concept of full attainability does not seem to be very practical. Evidently, the demand to hit *exactly* any target function in $W^{1,p}([t_1 - h, t_1], \mathbb{R}^n)$, by choosing appropriate control functions in $L_p([t_0, t_1], \mathbb{R}^r)$, is too strong. Alternatively, the concept of approximate attainability (controllability) may be pursued (cf. e.g. [15], [29]). However, this concept means that \mathcal{A} is dense in Z , and it is just this case which is difficult to handle by the maximum principle (see Theorem 1.1). Therefore Olbrot replaced the equality end constraint by the condition that the final state lies in a ball in a function space. Unfortunately it seems that his approach in [28] is not generalizable to systems with time-varying lag.

While Proposition 4.3 completely characterizes full attainability, there remains some distance between the necessary and the sufficient conditions for the closedness of \mathcal{A} . In particular, the rather strong assumption (H2) is not necessary, as is shown by the following result of Kurcyusz–Olbrot [24].

PROPOSITION. *Consider*

$$\dot{x}(t) = A_1 x(t-1) + A_2 x(t) + Bu(t), \quad t \in [t_0, t_1],$$

$$x|_{[t_0-1, t_0]} = 0$$

where $A_1, A_2 \in \mathcal{L}_{nn}$, $B \in \mathcal{L}_{nr}$, $t_1 - 1 > t_0$, $A_1 \neq 0$, and $u \in L_p([t_0, t_1], \mathbb{R}^r)$.

Then the attainable set \mathcal{A} is closed in $W^{1,p}([t_1-1, t_1], \mathbb{R}^n)$ iff $\text{Im } A_1 A_2^i B \subset \text{Im } B$, $i = 0, 1, \dots, n-1$.

Propositions 4.1–4.3 improve the results which have already been presented in the literature, especially by Banks–Jacobs–Langenhop [6]. We generalize established results from HDS to Fredholm systems. Furthermore, Proposition 4.2 shows that (H3) is not only a sufficient [6, Thm. 3.3] but also a necessary condition for \mathcal{A} to be closed (if (H2) and (4.2) are assumed). This enables us to substitute (H3) by the weaker assumption (4.2) as premise in the characterization of full attainability (compare Proposition 4.3 and [6, Thm. 3.1]).

REFERENCES

- [1] H. T. BANKS, *Variational problems involving functional differential equations*, this Journal, 7 (1969), pp. 1–17.
- [2] ———, *Control of functional differential equations with function space boundary conditions*, Delay and Functional Differential Equations and their Application, K. Schmitt, ed., New York, 1972.
- [3] H. T. BANKS AND M. Q. JACOBS, *The optimization of trajectories of linear functional differential equations*, this Journal, 8 (1970), pp. 461–488.
- [4] ———, *An attainable sets approach to optimal control of functional differential equations with function space boundary conditions*, J. Differential Equations, 13 (1973), pp. 127–149.
- [5] 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.
- [6] ———, *Characterization of the controlled states in $W_2^{(1)}$ of linear hereditary systems*, this Journal, 13 (1975), pp. 611–649.
- [7] H. T. BANKS AND G. A. KENT, *Control of functional differential equations of retarded and neutral type with target sets in function space*, this Journal, 10 (1972), pp. 567–593.
- [8] H. T. BANKS AND A. MANITIUS, *Application of abstract variational theory to hereditary systems—A survey*, IEEE Trans. Automatic Control, AC-19 (1974), pp. 524–533.
- [9] H. BAUER, *Wahrscheinlichkeitstheorie und Grundzüge der Maßtheorie*, 2. Auflage, De Gruyter, Berlin, 1974.
- [10] Z. BIEN, *Optimal control of delay systems*, Ph.D. thesis, University of Iowa, Ames, 1975.
- [11] D. H. CHYUNG AND E. B. LEE, *Delayed action control systems*, Automatica, 6 (1970), pp. 395–400.
- [12] F. COLONIUS AND D. HINRICHSSEN, *Optimal control of hereditary differential systems, Part I*, Arbeitspapiere Mathematik Nr. 2, Univ. Bremen, Bremen, W. Germany, 1976.
- [13] R. COURANT AND D. HILBERT, *Methoden der Mathematischen Physik*, Springer-Verlag, Berlin, 1968.
- [14] M. C. DELFOUR AND A. MANITIUS, *Control systems with delays: Areas of applications and present status of the linear theory*, Rep. CRM-658, Univ. Montréal, 1976.
- [15] M. C. DELFOUR AND S. K. MITTER, *Controllability, observability and optimal feedback control of affine hereditary differential systems*, this Journal, 10 (1972), pp. 298–328.
- [16] R. D. DRIVER, *A “backwards” two-body problem of classical relativistic electrodynamics*, Phys. Rev., 178 (1969), pp. 2051–2057.
- [17] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part I: General Theory*, Wiley-Interscience, New York, 1967.
- [18] I. V. GIRSANOV, *Lectures on Mathematical Theory of Extremum Problems*, Springer-Verlag, Berlin, 1972.
- [19] J. HALE, *Theory of Functional Differential Equations*, 2nd ed., Springer-Verlag, New York, 1977.
- [20] H. HALKIN, *A satisfactory treatment of equality and inequality constraints in the Dubovitskii–Milyutin optimization formalism*, J. Optimization Theory Appl., 6 (1970), pp. 138–149.

- [21] M. Q. JACOBS AND T. J. KAO, *An optimum settling problem for time-lag systems*, J. Math. Anal. Appl., 40 (1972), pp. 687–707.
- [22] S. KURCYUSZ, *A local maximum principle for operator constraints and its application to systems with time lag*, Control and Cybernetics, 2 (1973), pp. 99–125.
- [23] ———, *On the existence and nonexistence of Lagrange multipliers in Banach spaces*, J. Optimization Theory Appl., 20 (1976), pp. 81–110.
- [24] 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.
- [25] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1968.
- [26] K. MAKOWSKI AND L. W. NEUSTADT, *Optimal control problems with mixed control phase variable equality and inequality constraints*, this Journal, 12 (1974), pp. 184–228.
- [27] L. W. NEUSTADT, *Optimization—A Theory of Necessary Conditions*, Princeton University Press, Princeton, NJ, 1976.
- [28] A. W. OLBROT, *Control of retarded systems with function space constraints; Necessary optimality conditions*, Control and Cybernetics, 5 (1976), pp. 5–31.
- [29] ———, *Control of retarded systems with function space constraints, Part 2: Approximate controllability*, Ibid., 6 (1977), pp. 17–71.
- [30] R. PENROSE, *A generalized inverse for matrices*, Proc. Cambridge Philos. Soc., 51 (1955), pp. 406–413.
- [31] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISCHENKO, *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962.
- [32] W. T. REID, *Some elementary properties of proper values and proper vectors of matrix functions*, SIAM J. Appl. Math., 18 (1970), pp. 259–266.
- [33] M. M. VAINBERG, *Some problems in the differential calculus in linear spaces*, Uspehi Mat. Nauk, 7 (1952), No. 4, pp. 55–102. (In Russian.)
- [34] ———, *Variational Methods for the Study of Nonlinear Operators*, Holden-Day, San Francisco, 1964.
- [35] H. VAN DAM AND E. P. WIGNER, *Classical relativistic mechanics of interacting point particles*, Phys. Rev. B, 138 (1965), pp. 1576–1582.
- [36] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York-London, 1972.
- [37] ———, *Optimal controls with pseudo-delays*, this Journal, 12 (1974), pp. 286–299.
- [38] L. v. WOLFFSDORF, *Optimale Steuerungsprobleme bei linearen Integralgleichungen, I. Fredholmsche Integralgleichungen*, Beiträge Anal., 7 (1975), pp. 113–130.
- [39] A. C. ZAAANEN, *Linear Analysis*, North-Holland, Amsterdam, 1964.
- [40] P. P. ZABREYKO ET AL., *Integral Equations—A Reference Text*, Noordhoff, Leyden, 1975.

DEADBEAT FUNCTION OBSERVERS FOR DISCRETE-TIME LINEAR SYSTEMS*

HIDENORI KIMURA†

Abstract. A deadbeat function observer is an estimator that estimates exactly a linear function of the state of a deterministic discrete-time linear system via an incomplete state observation. In this paper structural properties of deadbeat function observers are studied by a geometric approach. The main result is the derivation of the minimal order of deadbeat function observers, which has been a longstanding open problem. The result reveals a fundamental relation between the observer order and the number of exogenous data required for exact estimation. An upper bound for the minimal order, which is at the same time the minimal order for generic cases, is derived in terms of the observability indices of the plant. A simple design algorithm of a minimal-time minimal-order function observer is also derived. Some numerical examples are also discussed.

1. Introduction. The discrete-time version of the Luenberger observer [6] has an advantage, compared with its continuous-time counterpart, of being capable of estimating exactly the current state vector of the plant within a finite length of time. Such observers were called *deadbeat observers* by Tse et al. [8]. They provided an algorithm for constructing minimal order deadbeat observers for time-varying systems from the viewpoint of statistically optimal construction [8], which was further improved by Yoshikawa et al. [11] and Leondes et al. [5].

Sometimes, we only need to estimate a linear function of the state vector rather than the state vector itself. This is usually the case where we are to implement a linear feedback control law via an incomplete state observation. Such observers are called *function observers*. Function observers for continuous-time systems have been studied by many authors (e.g., [1], [2]). However, the most fundamental problem of finding its minimal order still remains unsolved. As for the deadbeat observer, there is another facet of the problem, that is, a question of time-optimality. A desirable deadbeat observer is the one that estimates the function as fast as possible. Such an estimator is called a *minimal-time function observer* [7]. Nagata et al. [7] gave a design algorithm of minimal-time function observer. Their algorithm, however, does not lead to a minimal-order observer. The question may arise: How is the order-minimality related to the time-optimality? Putting it differently, is it possible to reduce the order at the cost of increasing the estimation time? The present paper gives the complete answer to this question. The results are summarized as follows:

(i) An explicit form for the minimal order of deadbeat function observer is derived.

(ii) The minimal order is shown to be independent on the estimation time. In other words, the minimal order of the minimal-time observer is the minimal order of any deadbeat observers.

(iii) An upper bound for the minimal order is derived in terms of the observability indices of the plant, which provides the *generic* minimal order.

(iv) A simple sequential algorithm for designing a minimal-time minimal-order function observer is derived.

In § 2 a fundamental preliminary result is derived. Based on a property of nilpotent matrices, we introduce a concept of *L-chain* which is an alternative

* Received by the editors June 17, 1977, and in final revised form December 29, 1977.

† Department of Control Engineering, Faculty of Engineering Science, Osaka University, Toyonaka, Osaka, Japan.

representation of deadbeat function observers (§ 3). In § 4 an explicit form of the minimal order of deadbeat observers is derived by using properties of L -chain. Some consequences of this result are discussed. The final section is devoted to the derivation of a design algorithm and the discussions of examples.

Mathematics is mainly developed on the dual space. A matrix W is always associated with its row space (the subspace spanned by its row vectors) that is denoted by the corresponding script letter \mathcal{W} and vice versa. We use a notation $\mathcal{W}A^{-1} = \{z: zA \in \mathcal{W}\}$. We denote $\langle \mathcal{C}|A \rangle = \mathcal{C} + \mathcal{C}A + \cdots + \mathcal{C}A^{n-1}$. The notation $\mathbf{i} = (1, 2, \dots, i)$ is used throughout.

2. Problem statement and preliminaries. Consider a linear system described by a difference equation

$$(1) \quad x(i+1) = Ax(i) + Bu(i), \quad y(i) = Cx(i),$$

where $x \in \mathcal{R}^n$, $u \in \mathcal{R}^r$ and $y \in \mathcal{R}^m$ denote the state, input and output vectors, respectively, and A , B , and C are constant matrices of appropriate dimensions. We assume that a system (1) is controllable.

Now a linear function of the state

$$(2) \quad \xi = Lx, \quad L \in \mathcal{R}^{p \times n}, \quad \xi \in \mathcal{R}^p,$$

is given. Our problem is to construct an estimator

$$(3) \quad z(i+1) = Fz(i) + Gy(i) + Tu(i), \quad z \in \mathcal{R}^p,$$

$$(4) \quad \hat{\xi}(i) = Hz(i) + Jy(i)$$

such that the output $\hat{\xi}(i)$ estimates $\xi(i) = Lx(i)$ exactly after a finite number of steps, i.e.,

$$(5) \quad \hat{\xi}(i) = \xi(i), \quad i = \pi, \pi+1, \dots$$

for some integer π starting at $i=0$. The estimator (3)(4) evaluates $\xi(\pi)$ based on the $2\pi+1$ exogenous data ($y(0), y(1), \dots, y(\pi), u(0), u(1), \dots, u(\pi-1)$). We call this estimator an (L, π) -observer. The integer π represents the estimation time. Of course it is desirable to design an (L, π) -observer with π as small as possible. This leads to a minimal-time observer [7]. On the other hand, the dimension p of z , that is, the order of the observer, represents the complexity of the estimation scheme. From various practical reasons, the observer with the simplest scheme, the minimal order, is preferable among those exhibiting the same performance. Thus the problem has the two main aspects: To attain the quickest response (the minimal time) and to give the simplest estimation scheme (the minimal order). In what follows we shall show that the simplest scheme is always realized in the observer of the quickest response.

To begin with, we derive some fundamental matrix relations. Their continuous-time versions were, with a slight additional restriction, found in [1].

LEMMA 1. *An estimator (3)(4) is an (L, π) -observer for (1)(2) if and only if there exists a matrix $W \in \mathcal{R}^{p \times n}$ such that the quintuple (W, F, G, H, J) satisfies*

$$(6) \quad HF^k(WA - FW - GC) = 0,$$

$$(7) \quad HF^{\pi+k} = 0 \quad \text{for each } k \geq 0,$$

$$(8) \quad HF^k(WB - T) = 0,$$

$$(9) \quad L = HW + JC.$$

Proof. Assume that (6)–(9) are satisfied for some W . From (1)–(3), we have

$$\hat{\xi}(k + \pi) - \xi(k + \pi) = HF^{\pi+k}(z(0) - Wx(0)) = 0$$

for each $k \geq 0$. Thus the sufficiency part has been proved.

Assume that a system (3)(4) is an (L, π) -observer. Due to (5),

$$Hz(\pi + k) = (L - JC)x(\pi + k)$$

for each $k \geq 0$. A straightforward algebraic manipulation yields

$$HF^i z(\pi) = M_i x(\pi) + \sum_{j=1}^i (M_{j-1}B - HF^{j-1}T)u(\pi + i - j)$$

where M_i is a matrix defined sequentially by

$$(10) \quad M_0 = L - JC, \quad M_{i+1} = M_i A - HF^i GC.$$

Since $u(i)$ and $x(\pi)$ are arbitrary (note the controllability assumption on (1)), we conclude that

$$M_k B = HF^k T$$

and there exists a matrix W such that

$$(11) \quad HF^k W = M_k$$

for each k . These relations obviously imply (8). From (11), we have

$$HF^k WA = M_{k+1} + HF^k GC = HF^{k+1} W + HF^k GC.$$

This implies (6). Letting $k = 0$ in (11) shows (9). On the basis of (6)(8)(9), we have

$$\hat{\xi}(\pi + k) = \xi(\pi + k) + HF^{\pi+k}(z(0) - Wx(0)).$$

Therefore $HF^{\pi+k}(z(0) - Wx(0)) = 0$, for each k , $z(0)$ and $x(0)$. This implies (7). Thus the proof has been completed.

In the above lemma, we can assume, without loss of generality, that the pair (H, F) is observable, since otherwise, we can find another (L, π) -observer of lower order by picking up the observable mode of the system (3)(4). In this respect the relations (6)(7) and (9) are written as

$$(12) \quad WA = FW + GC, \quad L = HW + JC, \quad F^\pi = 0$$

and (8) as

$$(13) \quad T = WB.$$

The relations (12) represent the fundamental relations that should be satisfied by an (L, π) -observer. The relation (13) is concerned with the input and is of secondary importance because the matrix T is readily obtained once W is found. The matrix relation (12) is simply denoted by $(W, F, G, H, J) \xrightarrow{(L, \pi)} (A, C)$, implying that a quintuple (W, F, G, H, J) represents an (L, π) -observer.

3. L -chain. In this section we shall introduce a concept of L -chain that is an alternative characterization of (L, π) -observers and is much easier to work with than the matrix relations (12). We begin with an important property of a nilpotent matrix.

LEMMA 2. Let $F \in \mathbb{R}^{\rho \times \rho}$ be a nilpotent matrix with nilpotency index π . If a pair (H, F) is observable, then there exist subspaces \mathcal{H}_i ($i \in \pi$) of \mathcal{H} such that

$$(14) \quad \mathcal{R}^\rho = \mathcal{H}_1 + \mathcal{H}_2 F + \cdots + \mathcal{H}_\pi F^{\pi-1},$$

$$(15) \quad \mathcal{H} F^i \subset \mathcal{H}_{i+1} F^i + \cdots + \mathcal{H}_\pi F^{\pi-1}, \quad i = 0, 1, \dots, \pi-1,$$

Proof. Let $\lambda^{\pi_1}, \lambda^{\pi_2}, \dots, \lambda^{\pi_k}$ ($\pi = \pi_1 \geq \pi_2 \geq \cdots \geq \pi_k$, $\pi_1 + \pi_2 + \cdots + \pi_k = \rho$) be invariant factors of F , where k is the cyclic index of F . The observability assumption on (H, F) implies [10, p. 16] that $k \leq \text{rank } H$ and there exist vectors $h_i \in \mathcal{H}$ ($i \in \mathbf{k}$) such that

$$\mathcal{R}^\rho = \text{span} \{h_1, h_1 F, \dots, h_1 F^{\pi_1-1}, h_2, \dots, h_2 F^{\pi_2-1}, \dots, h_k F^{\pi_k-1}\}.$$

Let $\sigma_i = \{\text{number of } j \text{ for which } \pi_j \geq i\}$. Define

$$\mathcal{H}_i = \text{span} \{h_j, j \in \sigma_i\}.$$

Then,

$$\mathcal{R}^\rho = \mathcal{H}_1 + \mathcal{H}_2 F + \cdots + \mathcal{H}_\pi F^{\pi-1}.$$

It is well known that there exists vectors z_i such that

$$(16) \quad \mathcal{R}^\rho = \text{span} \{z_1, z_1 F, \dots, z_1 F^{\pi_1-1}, \dots, z_k, \dots, z_k F^{\pi_k-1}\},$$

$$z_i F^{\pi_i} = 0, \quad i \in \mathbf{k}.$$

Let $\mathcal{Z} = \text{span}\{z_i, i \in \mathbf{k}\}$. From (16),

$$\mathcal{H} \subset \mathcal{Z} + \mathcal{Z} F + \cdots + \mathcal{Z} F^{\pi-1}.$$

Since $F^\pi = 0$, we have, for each i

$$(17) \quad \mathcal{H} F^i \subset \mathcal{Z} F^i + \mathcal{Z} F^{i+1} + \cdots + \mathcal{Z} F^{\pi-1}.$$

Therefore,

$$(18) \quad \mathcal{H}_{i+1} F^i + \mathcal{H}_{i+2} F^{i+1} + \cdots + \mathcal{H}_\pi F^{\pi-1} \subset \mathcal{Z} F^i + \cdots + \mathcal{Z} F^{\pi-1}.$$

Since $\dim(\mathcal{H}_{i+1} F^i + \cdots + \mathcal{H}_\pi F^{\pi-1}) = \dim(\mathcal{Z} F^i + \cdots + \mathcal{Z} F^{\pi-1}) = \sigma_{i+1} + \sigma_{i+2} + \cdots + \sigma_\pi$, we conclude that “ \subset ” in (18) can be replaced by “ $=$ ”. This, together with (17), shows the validity of (15).

Now we introduce another characterization of (L, π) -observer.

DEFINITION 1. A sequence of subspaces $(\mathcal{W}_1, \mathcal{W}_2, \dots, \mathcal{W}_\pi)$ is called an L -chain of length π if the following conditions hold:

$$(19) \quad \mathcal{W}_\pi \subset \mathcal{W}_{\pi-1} \subset \cdots \subset \mathcal{W}_1, \quad \mathcal{L} \subset \mathcal{W}_1 + \mathcal{C}.$$

$$(20) \quad \mathcal{W}_\pi A \subset \mathcal{C}, \quad \mathcal{W}_i A \subset \mathcal{W}_{i+1} + \mathcal{C}, \quad i \in \pi - \mathbf{1}.$$

The integer $\rho = \dim \mathcal{W}_1$ is called the chain dimension.

Note that L -chain is a restricted type of (A, C) -invariant subspaces [10, p. 91]. The significance of L -chain is demonstrated in the following result.

THEOREM 1. There exists an (L, π) -observer of order ρ if and only if there exists an L -chain of length π with dimension ρ .

Proof. Let $(W, F, G, H, J) \xrightarrow{(L, \pi)} (A, C)$. Due to Lemma 2, there exist subspaces $\mathcal{H}_i \subset \mathcal{H}$ satisfying (15). Define $\mathcal{W}_i = (\mathcal{H}_i F^{i-1} + \cdots + \mathcal{H}_\pi F^{\pi-1})W$. We have, from (12),

$$\mathcal{W}_\pi A \subset \mathcal{H}_\pi F^{\pi-1} W A \subset \mathcal{H}_\pi F^{\pi-1} (F W + G C) \subset \mathcal{C},$$

and due to (15),

$$\begin{aligned}\mathcal{W}_i A &\subset (\mathcal{H}_i F^{i-1} + \cdots + \mathcal{H}_\pi F^{\pi-1})(FW + GC) \\ &\subset (\mathcal{H}_{i+1} F^i + \cdots + \mathcal{H}_\pi F^{\pi-1})W + \mathcal{C} \\ &\subset \mathcal{W}_{i+1} + \mathcal{C},\end{aligned}$$

which establishes (20). The relation (19) follows immediately from (12) and the observability of (H, F) . Thus the “only if” part has been established.

Let $(\mathcal{W}_1, \mathcal{W}_2, \dots, \mathcal{W}_\pi)$ be an L -chain of length π . Define subspaces $\tilde{\mathcal{W}}_i$ ($i \in \pi$) as satisfying $\mathcal{W}_i = \tilde{\mathcal{W}}_i \oplus \mathcal{W}_{i+1}$, $i \in \pi - 1$, and $\tilde{\mathcal{W}}_\pi = \mathcal{W}_\pi$. We have, from (19) and (20).

$$\begin{aligned}\tilde{\mathcal{W}}_i A &\subset (\tilde{\mathcal{W}}_{i+1} \oplus \tilde{\mathcal{W}}_{i+2} \oplus \cdots \oplus \tilde{\mathcal{W}}_\pi) + \mathcal{C}, \quad \tilde{\mathcal{W}}_\pi A \subset \mathcal{C}, \\ \mathcal{L} &\subset (\tilde{\mathcal{W}}_1 \oplus \cdots \oplus \tilde{\mathcal{W}}_\pi) + \mathcal{C},\end{aligned}$$

or representing in the matrix form,

$$\tilde{\mathcal{W}}_i A = \sum_{j=i+1}^{\pi} \Phi_{ij} \tilde{\mathcal{W}}_j + \Xi_i C, \quad L = \sum_{j=1}^{\pi} \Psi_j \tilde{\mathcal{W}}_j + \Xi_0 C.$$

It is easily seen that $(W, F, G, H, J) \xrightarrow{(L, \pi)} (A, C)$ with

$$\begin{aligned}W &= \begin{bmatrix} \tilde{\mathcal{W}}_\pi \\ \tilde{\mathcal{W}}_{\pi-1} \\ \vdots \\ \tilde{\mathcal{W}}_1 \end{bmatrix}, \quad F = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \Phi_{\pi-1\pi} & 0 & \cdots & 0 \\ \vdots & & \cdots & \\ \Phi_{1\pi} & \Phi_{2\pi} & \cdots & 0 \end{bmatrix}, \quad G = \begin{bmatrix} \Xi_\pi \\ \Xi_{\pi-1} \\ \vdots \\ \Xi_1 \end{bmatrix}, \\ H &= [\Psi_1, \dots, \Psi_\pi], \quad J = \Xi_0.\end{aligned}$$

Thus the proof has been completed.

The above lemma implies that each (L, π) -observer is associated with an L -chain of length π and vice versa. Thus the problem is reduced to finding the minimal dimension and/or the minimal length of L -chain.

Now we make the following assumptions:

(A₁) A is nonsingular,

(A₂) $\mathcal{L} \subset \langle \mathcal{C} | A^{-1} \rangle$.

If the system (1) is a discretization of a continuous-time system, then the matrix A is the exponential function of a matrix and the assumption (A₁) is satisfied. Hence, (A₁) is by no means a strong restriction on discrete-time system. The assumption (A₂) is always satisfied if (C, A) is observable. As will be shown later, it is a necessary condition for the existence of (L, π) -observer.

Now we choose subspaces $\mathcal{C}_i \subset \mathcal{C}$ ($i = 0, 1, \dots, n-1$) as satisfying

$$\begin{aligned}\langle \mathcal{C} | A^{-1} \rangle &= \mathcal{C}_0 \oplus \mathcal{C}_1 A^{-1} \oplus \cdots \oplus \mathcal{C}_{n-1} A^{-(n-1)}, \\ \mathcal{C}_i &\subset \mathcal{C}_{i-1}, \quad \mathcal{C}_0 = \mathcal{C}.\end{aligned}$$

The existence of such \mathcal{C}_i is exhibited by the following procedure: Let $\{c_1, \dots, c_m\}$ be an arbitrary basis of \mathcal{C} . Consider the following sequence of vectors:

$$c_1, c_2, \dots, c_m, c_1 A^{-1}, \dots, c_m A^{-1}, c_1 A^{-2}, \dots, c_i A^{-j}, \dots$$

Let k_i be the smallest integer j such that $c_i A^j$ is contained in the subspace spanned by the preceding vectors in the above sequence. The integers (k_1, \dots, k_m) are the so-called observability indices for (C, A^{-1}) . Obviously,

$$\langle \mathcal{C} | A^{-1} \rangle = \text{span} \{c_i A^{-j+1}, i \in \mathbf{m}, j \in \mathbf{k}_i\}$$

and the subspaces \mathcal{C}_i given by

$$\mathcal{C}_i = \text{span} \{c_j; \text{ for all } j \text{ satisfying } k_j \geq i+1\}$$

satisfy the above conditions.

Define the subspaces \mathcal{Y}_i and \mathcal{Z}_i as

$$\mathcal{Y}_i = \mathcal{C}_0 \oplus \mathcal{C}_1 A^{-1} \oplus \dots \oplus \mathcal{C}_i A^{-i} = \mathcal{C} + \mathcal{C} A^{-1} + \dots + \mathcal{C} A^{-i},$$

$$\mathcal{Z}_i = \mathcal{C}_{i+1} A^{-(i+1)} \oplus \dots \oplus \mathcal{C}_{n-1} A^{-(n-1)}.$$

Then we have the decomposition

$$(21) \quad \langle \mathcal{C} | A^{-1} \rangle = \mathcal{Y}_i \oplus \mathcal{Z}_i.$$

We denote by P_i the projection operator on \mathcal{Z}_i along \mathcal{Y}_i with the domain $\langle \mathcal{C} | A^{-1} \rangle$. Since the subspaces \mathcal{Y}_i and \mathcal{Z}_i depend on the choice of \mathcal{C} , the projection operators P_i also depend on the choice of \mathcal{C}_i . For later use, we define

$$\tilde{\mathcal{Y}}_i = \mathcal{C} A^{-1} + \mathcal{C} A^{-2} + \dots + \mathcal{C} A^{-i}.$$

The following lemma states some useful properties of $P_i[\cdot]$.

LEMMA 3. For each i and j ,

$$(22) \quad P_{j-1}[\cdot] A^j \in P_j[\cdot] A^j + \mathcal{C},$$

$$(23) \quad P_j[P_i[\cdot] A^i] \in P_{i+j}[\cdot] A^i.$$

Proof. The relation (22) is an immediate consequence of an obvious fact $P_{j-1}[\cdot] \in P_j[\cdot] + \mathcal{C} A^{-j}$. Note also that $P_i[\cdot] \in P_{i+j}[\cdot] + \mathcal{C} A^{-(i+1)} + \dots + \mathcal{C} A^{-(i+j)}$. This implies that $P_i[\cdot] A^i \in P_{i+j}[\cdot] A^i + \mathcal{Y}_j$. On the other hand, $P_{i+j}[\cdot] A^i \in \mathcal{Z}_{i+j} A^i \subset \mathcal{Z}_j$. These two relations establish (23).

Now we single out a particular L -chain of central importance in what follows.

LEMMA 4. Let us define, for each $i \in \mathbf{n}-1$,

$$(24) \quad \mathcal{W}_i^* = P_{i-1}[\mathcal{L}] A^{i-1} + \dots + P_{n-2}[\mathcal{L}] A^{n-2}.$$

Then, $(\mathcal{W}_1^*, \mathcal{W}_2^*, \dots, \mathcal{W}_{n-1}^*)$ is an L -chain.

Proof. Due to (A_2) , $P_{n-1}[\mathcal{L}] = \{0\}$. Therefore, due to (22),

$$\mathcal{W}_{n-1}^* A = P_{n-2}[\mathcal{L}] A^{n-1} \subset P_{n-1}[\mathcal{L}] A^{n-1} + \mathcal{C} \subset \mathcal{C}.$$

Analogously,

$$\begin{aligned} \mathcal{W}_i^* A &= P_{i-1}[\mathcal{L}] A^i + \dots + P_{n-2}[\mathcal{L}] A^{n-1} \\ &\subset P_i[\mathcal{L}] A^i + \dots + P_{n-2}[\mathcal{L}] A^{n-2} + \mathcal{C} \\ &= \mathcal{W}_{i+1}^* + \mathcal{C}. \end{aligned}$$

Thus we have established (20). From the definition (24) and the obvious fact that $\mathcal{L} \subset P_0[\mathcal{L}] + \mathcal{C}$, we can easily show the validity of (19) for $(\mathcal{W}_1^*, \mathcal{W}_2^*, \dots, \mathcal{W}_{n-1}^*)$. Thus the proof has been completed.

In view of (20), each L -chain $(\mathcal{W}_1, \dots, \mathcal{W}_\pi)$ satisfies $\mathcal{W}_i \subset \mathcal{Y}_{\pi-i+1}$, $i \in \pi$. Therefore, the second relation of (19) implies $\mathcal{L} \subset \mathcal{Y}_\pi$. On the other hand, if $\mathcal{L} \subset \mathcal{Y}_\pi$,

$P_i[\mathcal{L}] = \{0\}$ for each $i \geq \pi$. This implies that $\mathcal{W}_i^* = \{0\}$ for each $i \geq \pi + 1$. This observation immediately leads to the following result.

LEMMA 5. *There exists an L -chain of length π , if and only if $\mathcal{L} \subset \mathcal{Y}_\pi$.*

A direct consequence of this lemma is the following characterization of the minimal estimation time, which was first established by Nagata et al. [7] by a different approach.

COROLLARY 1. *The minimal estimation time π^* (the minimal number of steps for exact estimation of ξ in (2)) is given by*

$$(25) \quad \pi^* = \min \{j: \mathcal{L} \subset \mathcal{Y}_j(\mathcal{L}A^j \subset \mathcal{C} + \mathcal{C}A + \cdots + \mathcal{C}A^j)\}.$$

4. Minimal-dimensional L -chain. In Lemma 4 we showed that the subspaces \mathcal{W}_i^* given by (24) constitute an L -chain. Using π^* defined by (25), we can write (24) as

$$(24') \quad \mathcal{W}_i^* = P_{i-1}[\mathcal{L}]A^{i-1} + \cdots + P_{\pi^*-1}[\mathcal{L}]A^{\pi^*-1}$$

because $P_j[\mathcal{L}] = \{0\}$ for $j \geq \pi^*$. In this section, we shall show that this L -chain is of minimal dimension by proving that the subspace (24') is contained in any L -chain. Towards this end, we need some preliminary result.

LEMMA 6. *Let $(\mathcal{W}_1, \mathcal{W}_2, \dots, \mathcal{W}_\pi)$ be an L -chain. Then, for each i and j ,*

$$P_i[\mathcal{W}_i]A \subset P_{j-1}[\mathcal{W}_{i+1}].$$

Proof. Obviously, $P_i[\mathcal{W}_i] \subset \mathcal{W}_i + \tilde{\mathcal{Y}}_i$. Therefore, $P_i[\mathcal{W}_i]A \subset \mathcal{W}_iA + \tilde{\mathcal{Y}}_iA \subset \mathcal{W}_{i+1} + \mathcal{Y}_{j-1}$. On the other hand, since $P_j[\mathcal{W}_i] \subset \mathcal{L}_j$, $P_i[\mathcal{W}_i]A \subset \mathcal{L}_jA \subset \mathcal{L}_{j-1}$. Therefore,

$$P_i[\mathcal{W}_i]A \subset P_{j-1}[P_i[\mathcal{W}_i]A] \subset P_{j-1}[\mathcal{W}_{i+1}],$$

which was to be proved.

LEMMA 7. *For any L -chain $(\mathcal{W}_1, \mathcal{W}_2, \dots, \mathcal{W}_\pi)$,*

$$(26) \quad \mathcal{W}_i^* \subset \mathcal{W}_i$$

for each $i \in \pi$, where \mathcal{W}_i^* is given in (24').

Proof. From the second relation of (19), we have $P_j[\mathcal{L}] \subset P_j[\mathcal{W}_1]$ for each $j = 0, 1, \dots, \pi - 1$. The successive application of Lemma 6 yields

$$P_j[\mathcal{L}]A^j \subset P_j[\mathcal{W}_1]A^j \subset P_{j-1}[\mathcal{W}_2]A^{j-1} \subset \cdots \subset P_0[\mathcal{W}_{j+1}] = \mathcal{W}_{j+1}.$$

In view of the first relation of (19), the above relation obviously implies (26).

Let us call the L -chain $(\mathcal{W}_1^*, \mathcal{W}_2^*, \dots, \mathcal{W}_{\pi^*}^*)$ of length π^* defined by (24') a *basic L -chain*. The above lemma demonstrates the significances of the basic L -chain. Any L -chain of arbitrary length contains the basic L -chain, or (if we define a *subchain* in a usual way) any L -chain contains a basic L -chain as a subchain. Recall that the projection operators P_i depend on the choice of \mathcal{C}_i in (21). However, Lemma 7 implies that the basic L -chain is unique. Indeed, if $(\mathcal{W}_1^*, \mathcal{W}_2^*, \dots, \mathcal{W}_{\pi^*}^*)$ and $(\tilde{\mathcal{W}}_1^*, \tilde{\mathcal{W}}_2^*, \dots, \tilde{\mathcal{W}}_{\pi^*}^*)$ are two basic L -chains, the relation (26) implies both $\mathcal{W}_i^* \subset \tilde{\mathcal{W}}_i^*$ and $\tilde{\mathcal{W}}_i^* \subset \mathcal{W}_i^*$, which are equivalent to $\mathcal{W}_i^* = \tilde{\mathcal{W}}_i^*$. Thus we have established the main result of this paper.

THEOREM 2. *Under the assumptions (A₁) and (A₂), the minimal order ρ^* of deadbeat function observer is given by*

$$(27) \quad \rho^* = \dim \mathcal{W}_1^* = \dim (P_0[\mathcal{L}] + P_1[\mathcal{L}]A + \cdots + P_{\pi^*-1}[\mathcal{L}]A^{\pi^*-1}),$$

where π^* is the integer in (25). There exists a minimal-time observer $((L, \pi^*)$ -observer) with order ρ^* .

Remark. The estimation time π (the length of the L -chain) is at our disposal under the condition $\pi \geq \pi^*$. The above result implies that the minimal order is independent on the choice of π . The minimal order of the minimal-time observer is also the minimal order of any deadbeat observer.

5. Upper bound for the minimal order. In this section we discuss the result further and derive an upper bound for the minimal order (27) in terms of the observability indices of the plant. This upper bound turns out to be the minimal order for generic L .

LEMMA 8. Let $\mu_1 \geq \mu_2 \geq \dots \geq \mu_m$ be the observability indices of the pair (C, A) . If $k = \mu_s - 1$ for some $0 < s \leq m$, then

$$\dim(\langle \mathcal{C}|A \rangle) - \dim(\mathcal{C} + \mathcal{C}A + \dots + \mathcal{C}A^k) = \sum_{i=1}^s (\mu_i - k - 1).$$

The proof is straightforward from the existence of $c_i \in \mathcal{C}$, $i \in \mathbf{m}$, satisfying

$$\langle \mathcal{C}|A \rangle = \text{span} \{c_1, c_1A, \dots, c_1A^{\mu_1-1}, \dots, c_m, \dots, c_mA^{\mu_m-1}\}.$$

THEOREM 3. Assume that A is nonsingular and let $\mu_1 \geq \mu_2 \geq \dots \geq \mu_m$ be the observability indices of the pair (C, A) . Then, for each $L \in \mathcal{R}^{p \times n}$, the minimal order ρ^* given in (27) satisfies

$$(28) \quad \rho^* \leq \sum_{i=1}^s (\mu_i - 1), \quad s = \min(p, m).$$

Moreover, the equality in (28) holds for generic L .

Remark 1. The nonsingularity assumption on A is always satisfied for usual discrete-time systems, because A is always an exponential function of a matrix.

Remark 2. Although the word “generic” is now popular in the control literature (e.g. [10, p. 26]), we state its precise meaning in the present context. The theorem asserts that the equality in (28) holds for all matrices $L \in \mathcal{R}^{p \times n}$ except those whose entries l_{ij} satisfy a finite number of nontrivial algebraic equations, the coefficients of which are some algebraic function of A and C .

Proof. Let $p \geq m$. In the decomposition (21), we have

$$\dim(\mathcal{X}_0) = \dim(\langle \mathcal{C}|A^{-1} \rangle) - \dim(\mathcal{Y}_0) = \dim(\langle \mathcal{C}|A \rangle) - \dim(\mathcal{C}) = \sum_{i=1}^m \mu_i - m.$$

Here we assumed that $\text{rank } C = m$ in order to avoid the triviality. Since we can easily show that $\mathcal{W}_1^* \subset \mathcal{X}_0$, we obtain $\rho^* = \dim(\mathcal{W}_1^*) \leq \sum_{i=1}^m (\mu_i - 1)$, which implies (28). Now we assume $p < m$. Write $k = \mu_p - 1$. Due to Lemma 8,

$$\begin{aligned} \dim(\mathcal{X}_k A^k) &= \dim(\langle \mathcal{C}|A^{-1} \rangle A^k) - \dim(\mathcal{Y}_k A^k) \\ &= \sum_{i=1}^p (\mu_i - k - 1). \end{aligned}$$

Note that $\mathcal{W}_1^* = P_0[\mathcal{L}] + P_1[\mathcal{L}]A + \dots + P_{k-1}[\mathcal{L}]A^{k-1} + (P_k[\mathcal{L}] + \dots + P_{\pi^*-1} \times [\mathcal{L}]A^{\pi^*-1})A^k \subset P_0[\mathcal{L}] + \dots + P_{k-1}[\mathcal{L}]A^{k-1} + \mathcal{X}_k A^k$. Since $\dim(P_i[\mathcal{L}]) \leq p$ for each i ,

$$\begin{aligned} \dim(\mathcal{W}_1^*) &\leq \dim(P_0[\mathcal{L}] + \dots + P_{k-1}[\mathcal{L}]A^{k-1}) + \sum_{i=1}^p (\mu_i - k - 1) \\ &\leq kp + \sum_{i=1}^p (\mu_i - k - 1) = \sum_{i=1}^p (\mu_i - 1). \end{aligned}$$

Thus the first part of the theorem has been proved.

If we fix a basis of $\langle \mathcal{C}|A^{-1} \rangle$, the basis of $P_i[\mathcal{L}]$ constitutes a matrix whose entries are algebraic functions of L . Hence ρ^* in (27) is represented by a rank condition of a matrix of this type. Since the rank condition of a matrix is reduced to determining

whether its minors vanish or not, the second assertion will be proved if we can show the existence of at least one $L = L_0$ for which the equality in (27) holds. We construct such an L_0 as follows. Since the observability indices of (C, A^{-i}) are equal to those of (C, A) , we can find $c_i \in \mathcal{C}(i \in \mathbf{s})$ such that, for each $j < \mu_i$,

$$c_i A^{-1} \notin \mathcal{Y}_{j-1}.$$

Define

$$l_i = c_i A^{-(\mu_i-1)}, \quad i \in \mathbf{s}, \quad \mathcal{L}_0 = \text{span} \{l_i, i \in \mathbf{s}\}.$$

From the selection of c_i , we can easily show that the vectors $\{v_{11}, \dots, v_{1\mu_1-1}, v_{21}, \dots, v_{2\mu_2-1}, \dots, v_{s1}, \dots, v_{s\mu_s-1}\}$ are linearly independent, where $v_{ij} = P_{j-1}[l_i]A^{j-1} = c_i A^{-(\mu_i-i)}$. Obviously, $v_{ij} \in \text{span } P_0[\mathcal{L}_0] + P_1[\mathcal{L}_0]A + \dots + P_{\mu_i-1}[\mathcal{L}_0]A^{\mu_i-1}$. If $s = p$ ($p \leq m$), we set $\mathcal{L}_0 = \mathcal{L}_0$ and if $s = m$ ($p > m$) we take an arbitrary subspace $\mathcal{L}_0 \supset \mathcal{L}_0$ of dimension p . From what has been shown above, the equality in (27) holds for \mathcal{L}_0 . Thus the proof has been completed.

We write

$$(29) \quad \rho_g^* = \sum_{i=1}^s (\mu_i - 1),$$

indicating the generic minimal order. If $p \geq m$ and the pair (C, A) is observable, then $\rho_g^* = \sum_{i=1}^m \mu_i - m = n - m$. This coincides with the well-known result on the minimal order of observers for estimating the whole state ($p = n$) [6]. If $p = 1$ (the observer for estimating single linear function), $\rho_g^* = \mu_1 - 1$. This again coincides with the well-known minimal order of functional observers [6]. These observations provide a good reason to bring forward a conjecture that the integer ρ_g^* in (29) is the generic minimal order of function observers having an arbitrary set of poles.

If we are interested only in the generic L , the expression (29) is much simpler than (27). However, this does not reduce the value of the exact expression (27) of the minimal order. The word "generic" only means that the minimal order is given by (29) with probability one, if the entries of the matrix L in (2) are chosen at random. From the standpoint of control system design, however, nongeneric cases are of great importance because a function observer of lower dimension is desirable. Furthermore, the expression (27) gives an insight into the complicated relation among A , C and L .

6. Design algorithm. The geometric arguments developed in § 4 gives in principle a design algorithm of a minimal-time minimal-order function observer. However, their straightforward translation into the language of matrices may lead to an observer containing redundancy or unnecessarily complicated estimation scheme. In this section we shall derive a simple sequential design algorithm of a minimal-time minimal-order function observer which results in the simplest estimation scheme in the sense that the number of the nonzero entries of the coefficient matrices F and H in (3)(4) is minimal. Moreover, it does not require the inversion of A .

ALGORITHM.

Step 1 (Construction of a matrix V_1 —Initialization.) Find the integer π_1 such that

$$(30) \quad \pi_1 = \min \{j : \mathcal{L}A^j \subset \mathcal{C} + \mathcal{C}A + \dots + \mathcal{C}A^j = \mathcal{Y}_j A^j\}.$$

Note that the integer π_1 is equal to the minimal estimation time π^* in (25). Choose $l_1 \in \mathcal{L}$ for which

$$(31) \quad l_1 A^j \notin \mathcal{Y}_j A^j,$$

for each $j < \pi_1$. From the definition of π_1 , there exist vectors $c_{10}, c_{11}, \dots, c_{1\pi_1} \in \mathcal{C}$ such that

$$(32) \quad l_1 A^{\pi_1} = c_{10} + c_{11}A + \dots + c_{1\pi_1}A^{\pi_1}.$$

Define vectors v_{1j} , $j \in \pi_1$ as

$$(33) \quad v_{1j} = l_1 A^{\pi_1-j} - c_{1j} - c_{1j+1}A - \cdots - c_{1\pi_1}A^{\pi_1-j} \\ = (c_{10} + c_{11}A + \cdots + c_{1j-1}A^{j-1})A^{-j}.$$

Obviously,

$$(34) \quad v_{11}A = c_{10}, \quad v_{1j+1}A = v_{1j} + c_{1j}, \quad j \in \pi_1 - \mathbf{1}, \quad l_1 = v_{1\pi_1} + c_{1\pi_1}.$$

The vectors v_{1j} in (33) are obtained by multiplying A^{-j} in the right hand side of (32) and discarding all the terms with nonnegative powers of A . The actual calculation of v_{1j} does not require the inversion of A . In geometric terms, we can write

$$(35) \quad v_{1j} = P_{\pi_1-j}[l_1]A^{\pi_1-j}.$$

Set $V_1 = [v_{11}^T \ v_{12}^T \ \cdots \ v_{1\pi_1}^T]^T$.

Step i (Construction of a matrix V_i .) Assume that matrices V_1, V_2, \dots, V_{i-1} have already been constructed. Find the integer π_i such that

$$(36) \quad \pi_i = \min \{j : \mathcal{L}A^j \subset (\mathcal{Y}_j + \mathcal{V}_1 + \cdots + \mathcal{V}_{i-1})A^j\}.$$

Choose $l_i \in \mathcal{L}$ such that, for each $j < \pi_i$,

$$(37) \quad l_i A^j \notin (\mathcal{Y}_j + \mathcal{V}_1 + \cdots + \mathcal{V}_{i-1})A^j.$$

From the definition of π_i , we have

$$(38) \quad l_i A^{\pi_i} = z_i A^{\pi_i} + w_i A^{\pi_i}, \quad w_i \in \mathcal{V}_{i-1}, \quad w_i \notin \mathcal{Y}_j, \quad j < \pi_i. \\ z_i A^{\pi_i} = c_{10} + c_{11}A + \cdots + c_{i\pi_i}A^{\pi_i} \in \mathcal{Y}_{\pi_i}A^{\pi_i}, \quad c_{ij} \in \mathcal{C}.$$

Analogously to Step 1, we define vectors v_{ij} , $j \in \pi_i$, as

$$(39) \quad v_{ij} = (l_i - w_i)A^{\pi_i-j} - c_{ij} - c_{ij+1}A - \cdots - c_{i\pi_i}A^{\pi_i-j}.$$

We have

$$(40) \quad v_{i1}A = c_{i0}, \quad v_{ij+1}A = v_{ij} + c_{ij}, \quad j \in \pi_i - \mathbf{1},$$

$$(41) \quad l_i = v_{i\pi_i} + w_i + c_{i\pi_i}.$$

In geometric terms, we can write

$$(42) \quad v_{ij} = P_{\pi_i-j}[z_i]A^{\pi_i-j}.$$

Set $V_i = [v_{i1}^T \ v_{i2}^T \ \cdots \ v_{i\pi_i}^T]^T$. If $i = p$, stop. Otherwise, let $i = i + 1$ and repeat the step.

After concluding the above algorithms, we have a matrix $W = [V_1^T \ V_2^T \ \cdots \ V_p^T]^T \in \mathcal{R}^{(\pi_1 + \pi_2 + \cdots + \pi_p) \times n}$. The relations (34)(40) are summarized as

$$(43) \quad WA = FW + GC,$$

where G is the matrix representing c_{ij} and F is of the form

$$F = \text{blockdiag}[\Lambda_{\pi_1} \ \Lambda_{\pi_2} \ \cdots \ \Lambda_{\pi_p}], \\ \Lambda_j = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \in \mathcal{R}^{j \times j}.$$

An inspection of (40) proves that $v_{ij} \in \mathcal{Y}_j$. This shows that the vector w_i in (38) is contained in $\text{span} \{v_{jk}, j \leq i-1, k > \pi_i\}$. Therefore, the relations (41) are summarized as

$$(44) \quad L = HW + JC,$$

where J is the matrix representing $c_{i\pi_1}$ and

$$(45) \quad H = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1p} \\ \cdots & \cdots & \cdots & \cdots \\ h_{p1} & h_{p2} & \cdots & h_{pp} \end{bmatrix}, \quad \begin{aligned} h_{ii} &= (0 \cdots 0 \quad 1) \in \mathcal{R}^{\pi_i} \\ h_{ij} &= (0 \cdots 0 \quad 0) \in \mathcal{R}^{\pi_j}, \quad i < j, \\ h_{ij} &= (\underbrace{0 \cdots 0}_{\pi_i} \quad \underbrace{* \cdots *}_{\pi_j - \pi_i}) \in \mathcal{R}^{\pi_j}, \quad i > j. \end{aligned}$$

Since $\pi_1 \geq \pi_2 \geq \cdots \geq \pi_p$, we easily see that $F^{\pi_1} = 0$. Therefore the relations (43)(44) implies that $(W, F, G, H, J) \xrightarrow{(L, \pi_1)} (A, C)$ (recall the fundamental relations (12)). Since $\pi_1 = \pi^*$, the above algorithm generates a minimal-time deadbeat observer. The structured forms of F and H obtained above are schematically illustrated in Fig. 1, in which exogenous data passes are omitted in order to emphasize the internal dynamics. The longest string of the delay elements specifies the minimal estimation time. Other shorter strings accommodates the information from longer strings, instead of storing the incoming data via delay elements.

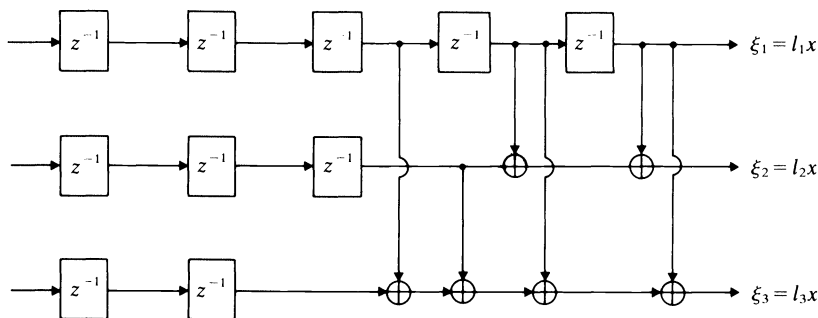


FIG. 1. An example of dynamic structure of a deadbeat function observer with $p=3$ and $(\pi_1, \pi_2, \pi_3) = (5, 3, 2)$.

It remains to show that the above algorithm generates a minimal-order observer. Towards this end we investigate the algorithm from the geometric view point.

LEMMA 9. Let $V_i, i \in \mathbf{p}$ be matrices constructed in the algorithm. Then,

$$(46) \quad \mathcal{V}_i \subset \mathcal{W}_1^* = P_0[\mathcal{L}] + P_1[\mathcal{L}]A + \cdots + P_{\pi_i-1}[\mathcal{L}]A^{\pi_i-1}.$$

Proof. From (35), (46) is evident for $i=1$. Assume that (46) is valid up to $i-1$. Due to (38),

$$P_k[z_i]A^k = P_k[l_i]A^k + P_k[w_i]A^k \in P_k[\mathcal{L}]A^k + P_k[\mathcal{W}_1^*]A^k.$$

From the definition of \mathcal{W}_1^* and (23),

$$P_k[z_i]A^k \in P_k[\mathcal{L}]A^k + P_{k+1}[\mathcal{L}]A^{k+1} + \cdots + P_{\pi_i-1}[\mathcal{L}]A^{\pi_i-1}.$$

In view of (42), the above relation implies (46) for i . Thus the proof has been completed.

LEMMA 10. Let $V_i, i \in \mathbf{p}$, be matrices constructed in the above algorithm. Then,

$$\text{rank } W = \text{rank } [V_1^T \cdots V_p^T]^T = \pi_1 + \pi_2 + \cdots + \pi_p.$$

Proof. If $\text{rank } W < \pi_1 + \pi_2 + \cdots + \pi_p$, there exist integers i and k such that

$$v_{ik} = t_1 + t_2, \quad t_1 \in \text{span} \{v_{i1}, \dots, v_{ik-1}\}, \quad t_2 \in \mathcal{V}_1 + \cdots + \mathcal{V}_{i-1}.$$

Since $v_{ij} \in \mathcal{Y}_j$, $P_{k-1}[t_1] = 0$. Due to (23) and (42), $P_{k-1}[v_{ik}]A^{k-1} = P_{\pi_1-1}[z_i]A^{\pi_1-1} \in \mathcal{Y}_1$. Since $P_{k-1}[t_2]A^{k-1} \in \mathcal{V}_1 + \cdots + \mathcal{V}_{i-1}$, we conclude that $P_{\pi_1-1}[z_i]A^{\pi_1-1} \in \mathcal{Y}_1 \cap (\mathcal{V}_1 + \cdots + \mathcal{V}_{i-1}) = \text{span} \{v_{11}, v_{21}, \dots, v_{i-1,1}\} \subset (\mathcal{V}_1 + \cdots + \mathcal{V}_{i-1} + \mathcal{Y}_{\pi_1-1})A^{\pi_1-1}$. Therefore, from (38), $P_{\pi_1-1}[l_i] \in \mathcal{V}_1 + \cdots + \mathcal{V}_{i-1}$ or equivalently, $l_i \in \mathcal{Y}_{\pi_1-1} + \mathcal{V}_1 + \cdots + \mathcal{V}_{i-1}$. This obviously contradicts the selection of l_i indicated in (37). Thus the proof has been completed.

Since $(W, F, G, H, J) \xrightarrow{(L, \pi_i)} (A, C)$, Lemma 7 implies that $\mathcal{W} = \mathcal{V}_1 + \cdots + \mathcal{V}_p \supset \mathcal{W}^*$. Therefore, from Lemma 9, $\mathcal{W} = \mathcal{W}^*$. This, together with Lemma 10, establishes the following result, which not only verifies that the algorithm generates a minimal order deadbeat observer but also provides another characterization of the minimal order.

THEOREM 4. Let the integer $\pi_i, i \in \mathbf{p}$, be defined via (30) and (36). Then

$$\rho^* = \pi_1 + \pi_2 + \cdots + \pi_p.$$

Remark 1. This result implies that the integers π_i (obviously, $\pi_1 \geq \pi_2 \geq \cdots \geq \pi_p$) do not depend on the choice of l_1, \dots, l_{i-1} . It is not difficult to see that $\pi_i \leq \mu_i - 1$ and the equality holds for generic L .

Example 1. Consider a linear system (1) with

$$A = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 2 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix},$$

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad L = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

(Step 1.) We have $\pi_1 = 2$. Take $l_1 = (0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0)$. Simple calculation yields a relation corresponding to (32),

$$l_1 A^2 = \frac{1}{2}c_2 - \frac{1}{2}c_3 - \frac{1}{2}c_1 A + c_3 A + \frac{1}{2}c_1 A^2.$$

The relations (33) define

$$w_{11} = l_1 A + \frac{1}{2}c_1 - c_3 - \frac{1}{2}c_1 A = \begin{pmatrix} \frac{1}{2} & 0 & 0 & -\frac{1}{2} & 0 & 0 & \frac{1}{2} \end{pmatrix},$$

$$w_{12} = l_1 - \frac{1}{2}c_1 = \begin{pmatrix} -\frac{1}{2} & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

We have

$$w_{11}A = \frac{1}{2}(c_2 - c_3), \quad w_{12}A - w_{11} - \frac{1}{2}c_1 + c_3, \quad l_1 = w_{12} + \frac{1}{2}c_1.$$

(Step 2.) We have $\pi_2 = 1$. The relation corresponding (38) is calculated to be

$$l_2A = -c_1 + 2c_2 + 2c_3 + \frac{1}{2}c_1A + c_2A - w_{12}A.$$

We set

$$w_{21} = l_2 - \frac{1}{2}c_1 - c_2 + w_{12} = (-1 \quad -1 \quad 0 \quad 1 \quad 1 \quad 0 \quad 0),$$

which gives rise to

$$w_{21}A = -c_1 + 2c_2 + 2c_3, \quad l_2 = w_{21} + \frac{1}{2}c_1 + c_2 - w_{12}.$$

Since $p = 2$, the algorithm has been carried out.

We have thus constructed a quintuple $(W, F, G, H, J) \xrightarrow{(L, 2)} (A, C)$ as

$$W = \begin{bmatrix} \frac{1}{2} & 0 & 0 & -\frac{1}{2} & 0 & 0 & \frac{1}{2} \\ -\frac{1}{2} & 0 & 0 & 1 & 0 & 0 & 0 \\ -1 & -1 & 0 & 1 & 1 & 0 & 0 \end{bmatrix}, \quad F = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$G = \begin{bmatrix} 0 & \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 0 & 1 \\ -1 & 2 & 2 \end{bmatrix}, \quad H = \begin{bmatrix} 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}, \quad J = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 1 & 0 \end{bmatrix}.$$

From this we have a minimal-time minimal-order observer

$$\begin{aligned} z_1(i+1) &= \frac{1}{2}y_2(i) - \frac{1}{2}y_3(i) + \frac{1}{2}u(i), \\ z_2(i+1) &= z_1(i) - \frac{1}{2}y_1(i) + y_3(i), \\ z_3(i+1) &= -y_1(i) + 2y_2(i) + 2y_3(i), \\ \hat{\xi}_1(i) &= z_2(i) - \frac{1}{2}y_1(i), \\ \hat{\xi}_2(i) &= -z_2(i) + z_3(i) + \frac{1}{2}y_1(i) - y_2(i). \end{aligned}$$

Now that $(\mu_1, \mu_2, \mu_3) = (3, 2, 2)$ and $\mu_1 + \mu_2 - 2 = 3$. Therefore, the above observer is of generic case.

Example 2. We construct a minimal-time minimal-order observer for a system treated by Nagata et al. [7]. The coefficient matrices are given by

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix},$$

$$L = \begin{bmatrix} 1 & 2 & 1 & 2 & 1 & 0 & 1 \\ 0 & 1 & 2 & 1 & 2 & 0 & 1 \end{bmatrix}.$$

(Step 1.) We have $\pi_1 = 1$, and taking $l_1 = (1 \quad 2 \quad 1 \quad 2 \quad 1 \quad 0 \quad 1)$ yields

$$\begin{aligned} l_1A &= c_1 + c_2 + c_3 + (c_1 + c_2)A, \\ w_{11} &= l_1 - c_1 - c_2 = (1 \quad 1 \quad 0 \quad 2 \quad 0 \quad 0 \quad 1), \\ w_{11}A &= c_1 + c_2 + c_3. \end{aligned}$$

(Step 2.) We have $\pi_2 = 1$, and taking $l_2 = (0 \ 1 \ 2 \ 1 \ 2 \ 0 \ 1)$ yields

$$l_2 A = c_3 + 2c_1 A + c_2 A,$$

$$w_{21} = l_2 - 2c_1 - c_2 = (0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 1), \quad w_{21} A = c_3.$$

The resulting quintuple $(W, F, G, H, J) \xrightarrow{(L, 1)} (A, C)$ is given as

$$W = \begin{bmatrix} 1 & 1 & 0 & 2 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}, \quad F = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$

$$G = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \quad H = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad J = \begin{bmatrix} 1 & 1 & 0 \\ 2 & 1 & 0 \end{bmatrix}.$$

Thus we have constructed a minimal-time minimal-order observer

$$z_1(i+1) = y_1(i) + y_2(i) + y_3(i) + u(i),$$

$$z_2(i+1) = y_3(i) + u(i);$$

$$\hat{\xi}_1(i) = z_1(i) + y_1(i) + y_2(i), \quad \hat{\xi}_2(i) = z_2(i) + y_3(i).$$

In [7] a fourth order observer was constructed. The order of the above observer is $2 < 4$. Thus the present algorithm provides a much simpler configuration.

7. Conclusion. The structural properties of deadbeat function observers have been studied by a geometric approach. An explicit form of the minimal order of deadbeat function observers has been derived which turns out to be independent on the choice of the estimation time. The minimal order of the minimal-time observer is shown to be the minimal order of any deadbeat function observers. An upper bound of the minimal order, which provides the generic minimal order, has been found in terms of the observability indices of the plant. A simple sequential algorithm for designing a minimal-time minimal-order observer has been devised accompanied by some numerical examples. Extension of these results to time-varying systems is straightforward by the approach in [9].

The results resolve completely the question of the minimal dimension of function observers for a special class. The present approach seems to be promising for attacking the same problem for other types of function observers. Moreover, the results throw a new light on the dual version of the problem, the problem of deadbeat control [3], [4].

Acknowledgment. The author wishes to express his grateful thanks to Dr. Hara of Iwate University and Mr. Adachi of Kyoto University for valuable comments on the early draft of this paper.

REFERENCES

- [1] T. E. FORTMAN AND D. WILLIAMSON, *Design of low order observers for linear feedback control laws*, IEEE Trans. Automatic Control, AC-17 (1972), pp. 301–308.
- [2] H. KIMURA, *Geometric structure of observers for linear feedback control laws*, Ibid., AC-22 (1977), pp. 845–855.
- [3] V. KUČERA, *The structure properties of time-optimal discrete linear systems*, Ibid., AC-16 (1971), pp. 375–377.
- [4] B. LEDEN, *Multivariable dead-beat control*, Automatica, 13 (1977), pp. 185–188.
- [5] C. T. LEONDES AND L. M. NOVAK, *Reduced-order observers for linear discrete-time systems*, IEEE Trans. Automatic Control, AC-19 (1974), pp. 42–46.

- [6] D. G. LUENBERGER, *Observers for multivariable systems*, Ibid., AC-11 (1966), pp. 190–197.
- [7] A. NAGATA, T. NISHIMURA AND M. IKEDA, *Linear function observer for linear discrete-time systems*, Ibid., AC-20 (1975), pp. 401–407.
- [8] E. TSE AND M. ATHANS, *Optimal minimal-order observer-estimators for discrete linear time-varying systems*, Ibid., AC-15 (1970), pp. 416–426.
- [9] L. WEISS, *Controllability, realization and stability of discrete-time systems*, this Journal, 10 (1972), pp. 230–251.
- [10] W. M. WONHAM, *Linear Multivariable Control—A Geometric Approach*, Springer-Verlag, New York, 1974.
- [11] T. YOSHIKAWA AND H. KOBAYASHI, *Comments on “minimal-order observer-estimators for discrete linear time-varying systems”*, IEEE Trans. Automatic Control, AC-17 (1972), pp. 272–273.

CONVEX CONTROL PROBLEMS AND HAMILTONIAN SYSTEMS ON INFINITE INTERVALS*

VIOREL BARBU†

Abstract. This paper presents results on optimal control problems over $[0, +\infty[$ associated with a linear process in a Hilbert space and with convex cost criterion. Results on synthesis of optimal control and necessary conditions for optimality in Hamiltonian form are obtained.

Corollaries to the main result include the behavior of Hamiltonian systems of the form: $x' - Ax \in \partial_p H(x, p)$, $p' + A^*p \in -\partial_x H(x, p)$, in a neighborhood of a saddle point.

Introduction. This paper considers a convex control problem over $R^+ = [0, +\infty[$ associated with time invariant systems whose input-state dynamics are described by linear evolution equations in a Hilbert space. The optimal control problem will be defined in terms of an integral convex cost criterion.

The main results are Theorem 1, in which we prove that the optimal control can be synthesized via a nonlinear feedback law of subgradient type, and Theorem 2, in which the necessity (and in a certain sense the sufficiency) of Euler–Lagrange conditions for optimality are established. The basic hypotheses for the development include a controllability assumption and some boundedness assumptions for the Hamiltonian function associated with the convex cost criterion. Under somewhat stronger assumptions, Theorem 1 has already been established in [1]. For linear control system with quadratic criterion a great deal of work has been done in this direction. We cite the works of Lukes and Russell [11], R. Datko [7], [8], Curtain and Pritchard [6], M. C. Delfour, C. McCalla and S. K. Mitter [9] (for linear hereditary systems) and in particular the book of J. L. Lions [10]. Our results extend those of the above-mentioned authors, in that we allow for a more general cost criterion which in general does not require any differentiability assumption.

The main results are proved in §§ 3 and 4. These results are used in § 5 to study the “saddle point” behavior of a general class of Hamiltonian systems. In this context Theorem 3 partially extends the results of Rockafellar [16] and those of [2] in that the space is infinite-dimensional and the Hamiltonian is unbounded. Two examples are considered in § 6.

For the basic facts in convex analysis relevant to this paper we refer the reader to the books [3], [5], [14] and the survey by R. T. Rockafellar [15].

1. Problem statement. The basic assumptions which will be in effect throughout our work will be set forth in this section. Later, in order to prove the main results, we shall make further hypotheses.

Let E and U be real Hilbert spaces whose norms are denoted $|\cdot|$ and $\|\cdot\|$. The inner product in E and U will be denoted by (\cdot, \cdot) and $\langle \cdot, \cdot \rangle$ respectively. We assume that A is a (possibly) unbounded closed and densely defined linear operator in E which generates a continuous semigroup $S(t): R^+ \rightarrow L(E, E)$. ($L(E, E)$ denotes the algebra of linear continuous operators on E .) Then the adjoint operator A^* generates the dual semigroup $S^*(t)$.

Our control system is

$$(1.1) \quad x'(t) = Ax(t) + Bu(t); \quad x(0) = x_0,$$

* Received by the editors July 20, 1977.

† Faculty of Mathematics, University of Iași, Iași, Romania.

where B is a linear continuous operator from U to E , $u: R^+ \rightarrow U$ is a given locally summable function and x_0 is a given element of E . In the sequel we shall be concerned with "mild" solutions of the Cauchy problem (1.1), i.e., with continuous functions $x: R^+ \rightarrow E$ satisfying

$$(1.2) \quad x(t) = S(t)x_0 + \int_0^t S(t-s)Bu(s) ds, \quad \text{for } t \geq 0,$$

where $u \in L^2_{\text{loc}}(R^+; U)$. By $L^2_{\text{loc}}(R^+; U)$ we denote the space of all strongly measurable functions $u: R^+ \rightarrow U$ which are "square integrable" on every finite interval $[0, T]$. It is well known that every (strong) solution of (1.1) can be written as (1.2); the differential equation (1.1) itself may not be satisfied by the "mild" solution (1.2), however.

In the sequel we shall denote by $x(t; x_0, u)$ the "mild" solution (1.2) of equation (1.1) corresponding to a given control $u(t)$ and with initial value x_0 .

Let x_0 be a given element of E and let $u \in L^2_{\text{loc}}(R^+; U)$ be a given control. On the trajectories of system (1.1) we define the cost functional

$$(1.3) \quad G(u, x_0) = \int_0^\infty L(x(t, x_0, u), u(t)) dt.$$

The extended real valued function $L: E \times U \rightarrow [0, +\infty]$ is assumed to be convex, lower-semicontinuous and nonidentically $+\infty$. We shall also assume that there exist real constants C_i , $i = 1, 2, 3$; $C_1 > 0$ such that

$$(1.4) \quad L(x, u) \geq C_1 \|u\|^2 + C_2 |x|^2 + C_3 \quad \text{for all } (x, u) \in E \times U.$$

It should be recalled that the conditions imposed on L imply that $L(x(t), u(t))$ is a Lebesgue measurable function of t whenever x and u are Lebesgue measurable. Since $L \geq 0$ we may infer that $G(u, x_0)$ is well-defined (unambiguously either a real number or $+\infty$).

Let x_0 be arbitrary but fixed in E . The optimal control problem we shall to work with is the following:

$$(P) \quad \min \{G(u, x_0); u \in L^2_{\text{loc}}(R^+; U)\}.$$

A function u where the infimum in (P) is attained will be referred to as *optimal control* of problem (P) while the corresponding function x (given by (1.2)) will be called the *optimal arc*. The pair (x, u) will be called *optimal pair* of problem (P).

One of our main purposes here is to synthesize the optimal controls u and to characterize the optimal pairs (x, u) in terms of the operator A and function L . To this aim we introduce the function $\phi: E \rightarrow [0, +\infty[$

$$(1.5) \quad \phi(x_0) = \inf \{G(u, x_0); u \in L^2_{\text{loc}}(R^+; U)\}.$$

The next lemma may be viewed as an existence result for problem (P).

LEMMA 1. *For every $x_0 \in E$ the infimum defining $\phi(x_0)$ is attained. Moreover, the function ϕ is convex and lower-semicontinuous on E .*

Proof. Since the argument is standard, the proof will be only outlined. Let $x_0 \in D(\phi) = \{y \in E, \phi(y) < +\infty\}$ be fixed and let $\{u_n\} \subset L^2_{\text{loc}}(R^+; U)$ be such that

$$(1.6) \quad \phi(x_0) \leq G(u_n, x_0) \leq \phi(x_0) + 1/n, \quad n = 1, 2, \dots$$

Denote by x_n the corresponding solution of (1.2). We fix any $T > 0$. By a simple calculation involving (1.2), condition (1.4) and Gronwall's lemma, it follows that the sequence $\{u_n\}$ is bounded in $L^2(0, T; U)$. Thus extracting a subsequence, if necessary,

we may assume that for each $T > 0$,

$$(1.7) \quad \begin{aligned} u_n &\rightarrow u \quad \text{weakly in } L^2(0, T; U), \\ x_n &\rightarrow x \quad \text{weakly in } L^2(0, T; E) \end{aligned}$$

where $u \in L^2_{\text{loc}}(R^+; U)$ and $x(t) = S(t)x_0 + \int_0^t S(t-s)Bu(s) ds$, for $t \geq 0$. On the other hand the lower-semicontinuity of L implies by a standard argument that for every $T > 0$ the convex function

$$(y, v) \rightarrow \int_0^T L(y(t), v(t)) dt$$

is lower-semicontinuous (equivalently, weakly lower-semicontinuous) on $L^2(0, T; E) \times L^2(0, T; U)$. The latter combined with (1.6) shows that $\int_0^T L(x(t), u(t)) dt \leq \phi(x_0)$ for each $T > 0$. Hence $G(u, x_0) = \phi(x_0)$ as claimed. A similar argument involving Fatou's lemma may be used to verify that the function ϕ is lower-semicontinuous. The convexity of ϕ is an immediate consequence of the corresponding property of L .

2. Synthesis of optimal controls. To begin with, we recall for easy references some concepts and notations.

Given the convex and lower-semicontinuous function $L: E \times U \rightarrow [0, +\infty]$ we denote by $H: E \times U \rightarrow [-\infty, +\infty]$ the corresponding Hamiltonian function, i.e.,

$$(2.1) \quad H(y; q) = \sup \{ \langle q, v \rangle - L(y, v); v \in U \}.$$

By $\partial H = \{-\partial_y H, \partial_q H\}$ we shall denote the subdifferential of H [14], [15], i.e.,

$$(2.2) \quad \partial_y H(y, q) = \{y^* \in E; H(y, q) \geq H(y_1, q) + \langle y^*, y - y_1 \rangle \text{ for all } y_1 \in E\},$$

$$(2.3) \quad \partial_q H(y, q) = \{v^* \in U; H(y, q) \leq H(y, q_1) + \langle v^*, q - q_1 \rangle \text{ for all } q_1 \in U\}.$$

A continuous arc $y: R^+ \rightarrow E$ is said to be *feasible* in problem (P) if there exists a control $u \in L^2_{\text{loc}}(R^+; U)$ such that

$$y(t) = y(t; y(0); v) = S(t)y(0) + \int_0^t S(t-s)Bv(s) ds$$

for $t \geq 0$ and $L(y, v) \in L^1(R^+)$.

We shall need the following assumptions:

Assumption A. The Hamiltonian function H is everywhere finite on $E \times U$. Moreover, one has

$$(2.4) \quad H(x, 0) \leq 0; \quad H(x, q) \leq C_1(|x|^2 + \|q\|^2) + C_2 \quad \text{for all } (x, q) \in E \times U$$

where C_i , $i = 1, 2$, are real constants.

Assumption B. There exists $R > 0$ such that:

(a) For each $x_0 \in S(0; R)$ there is at least one feasible arc y in problem (P) such that $y(0) = x_0$.

(b) For each $x_0 \in S(0; R)$ there exists a sequence $T_n \rightarrow \infty$ and the controls $\{u_n\} \subset L^2(0, T_n; U)$ such that $x(T_n; x_0, u_n) \in S(0; R)$.

Here $S(0; R)$ denotes the open ball $\{x_0 \in E; |x_0| < R\}$.

Next we shall discuss the meaning and some circumstances in which the above assumptions are satisfied.

First of all it should be noted that the condition that $-\infty < H < +\infty$ on $E \times U$ implies that H is separately continuous on $E \times U$ (see, e.g., [15]).

The condition that $H > -\infty$ on $E \times U$ in Assumption A requires that for each $y \in E$ the function $L(y, \cdot)$ is nonidentically $+\infty$. The rest of the assumption (i.e., condition (2.4)) implies that $L \geq 0$ on $E \times U$ and condition (1.4). This follows by a standard argument involving the conjugacy formula (see (2.1))

$$(2.5) \quad L(y, v) = \sup \{ \langle p, v \rangle - H(y, p); p \in U \}.$$

So Assumption A precludes the existence of state constraints into problem (P) but allows implicit control constraint of the form

$$u(t) \in U_0 \quad \text{a.e. } t > 0$$

where U_0 is a closed convex subset of U .

Assumption B may be viewed as a controllability condition relating system (1.1) and the function L . Part (a) simply requires that $S(0; R) \subset D(\phi)$ for some R . We shall present in Lemma 2 below some important cases in which it is satisfied.

LEMMA 2. *Assume that the controlled system (1.1) is stabilizable and that L satisfies the following conditions*

$$(2.6) \quad L(0, 0) = 0,$$

$$(2.7) \quad (0, 0) \in \text{int } D(L).$$

Then Assumption B holds. If either $D(L) = E \times U$ or the uncontrolled system (1.1) is asymptotically stable and

$$(2.8) \quad L(0, 0) = 0; \quad L(y, 0) < +\infty \quad \text{for all } y \in E,$$

then Assumption B is satisfied with $R = +\infty$.

Here $D(L)$ denotes the effective domain of L i.e.,

$$D(L) = \{(y, v) \in E \times U; L(y, v) < +\infty\}.$$

Proof of Lemma 2. By condition (2.7) there exists $r > 0$ such that

$$(2.9) \quad L(y, v) < +\infty \quad \text{for } |y| \leq r \text{ and } \|v\| \leq r$$

while the stabilizability hypothesis means that there exists a bounded linear operator $K: E \rightarrow U$ such that every "mild" solution of the closed loop system

$$(2.10) \quad y'(t) = (A + BK)y(t), \quad t \in \mathbb{R}^+,$$

is exponentially stable, i.e., there exists $\gamma > 0$ and $M > 0$ such that

$$(2.11) \quad |y(t)| \leq M \exp(-\gamma t) |y(0)| \quad \text{for all } t \geq 0.$$

We take $R = \inf \{r/M, \|K\|r/M\}$ where $\|K\|$ denotes the operatorial norm of K . Then estimate (2.11) together with relation (2.9) imply that

$$(2.12) \quad (y(t), Ky(t)) \in \text{int } D(L) \quad \text{for all } t \geq 0$$

for each solution y to system (2.10) with initial value $x_0 = y(0)$ in $S(0; R)$. Since the subdifferential ∂L of L is locally bounded on $\text{int } D(L)$ we deduce from (2.11) and (2.12) that

$$(2.13) \quad |z_1(t)| + \|z_2(t)\| \leq C \quad \text{for } t \geq 0$$

where $(z_1(t), z_2(t))$ is any section of $\partial L(y(t), Ky(t))$. Estimates (2.11) and (2.13)

combined with condition (2.6) and definition of ∂L , yield

$$\begin{aligned} L(y(t), Ky(t)) &\leq (z_1(t), y(t)) + (z_2(t), Ky(t)) \\ &\leq C_1 \exp(-\gamma t) \quad \text{for all } t \geq 0 \end{aligned}$$

and therefore $L(y, Ky) \in L^1(R^+)$. Hence Assumption B, part (a) holds with R defined above. As for part (b), it follows from estimate (2.11) that for all $T \geq T_0$, with T_0 sufficiently large, $y(T) \in S(0, R)$.

If in particular $D(L)$ is all of $E \times U$ then $r = +\infty$ and therefore Assumption B holds with $R = +\infty$. This also happens if the uncontrolled system (1.1) is asymptotically stable and condition (2.8) holds because in this case each $y(t) = S(t)x_0$ is a feasible arc in problem (P) corresponding to null control. The proof of Lemma 2 is thereby complete.

Remark 2.1. It is apparent from the preceding proof that part (b) of Assumption B automatically holds if the system (1.1) is stabilizable. Sufficient conditions for stabilizability in the framework developed here may be found in the paper [13] of Pritchard and Triggiani.

We are now ready to formulate the first theorem.

THEOREM 1. *Let Assumptions A and B be satisfied. Let (x, u) be any optimal pair for problem (P) with $x(0) = x_0 \in S(0; R)$. Then the optimal control u is expressed as a function of x by the feedback law*

$$(2.14) \quad u(t) \in \partial_q H(x(t), -B^* \partial \phi(x(t))), \quad \text{a.e. } t > 0.$$

Here $\partial \phi: E \rightarrow E$ denotes the subdifferential of ϕ .

It is noted that in particular Theorem 1 implies that each optimal arc x of problem (P) is a "mild" solution to the closed loop system

$$(2.15) \quad x' \in Ax + B \partial_q H(x, -B^* \partial \phi(x)), \quad t \in R^+.$$

Remark 2.2. If the function L is quadratic then $\partial \phi$ is a linear, self-adjoint continuous and positive operator from E into itself and so Theorem 1 reduces to some earlier results by Lukes and Russell [11] and Datko [8]. Furthermore, in this case $\partial \phi$ is the solution of a Riccati operational equation (see, e.g., [6], [10]).

Let us assume that each "mild" solution $x(\cdot)$ to closed loop equation (2.15) with $x(0) \in D(A)$ is a strong solution. This happens, for example, if A generates an analytic semigroup or if the function $x \rightarrow B \partial_q H(x, -B^* \partial \phi(x))$ is Fréchet differentiable (in particular, if L is quadratic).

For the sake of simplicity we shall also assume that $\partial \phi$ is single valued and for every $x_0 \in E$, problem (P) has a unique solution (x, u) . We have (see (3.3) below)

$$(\phi(x(t)))' + L(x(t), u(t)) = 0, \quad \text{a.e. } t > 0$$

while conjugacy formula (2.1) and relation (2.14) yield

$$L(x(t), u(t)) + H(x(t), -B^* \partial \phi(x(t))) = -(u(t), B^* \partial \phi(x(t))), \quad \text{a.e. } t > 0.$$

Using the well-known formula

$$(\phi(x(t)))' = (\partial \phi(x(t)), x'(t)), \quad \text{a.e. } t > 0,$$

and remembering that $x(t)$ is a strong solution to (1.1), we get

$$H(x(t), -B^* \partial \phi(x(t))) - (Ax(t), \partial \phi(x(t))) = 0 \quad \text{for } t \geq 0.$$

Inasmuch as $x(0)$ can be taken arbitrary in E we find that $K = \partial \phi$ satisfies the

operational equation

$$(2.16) \quad H(x, -B^*Kx) - (Ax, Kx) = 0 \quad \text{for all } x \in D(A).$$

Parenthetically, we note that in the special case in which L is quadratic, equation (2.16) can be brought into the form of a stationary Riccati operator equation associated with the given system. Equation (2.16) is hard to solve even in simple cases. In this sense, (2.16) is of theoretical rather than of practical interest. However, it can be used to gain some insight into the nature of $\partial\phi$ as well as to construct appropriate suboptimal feedback law for given problem.

3. Proof of Theorem 1. We fix x_0 in $S(0; R)$ and consider an optimal pair (x, u) for problem (P) corresponding to x_0 (the existence of a such pair was proved in Lemma 1).

Let ϕ be the function defined by (1.5). It is noted that for every $T > 0$, (x, u) is also a solution of the following control problem on the interval $[0, T]$

$$(3.1) \quad \min \left\{ \int_0^T L(y(t), v(t)) dt + \phi(y(T)); \right. \\ \left. y(t) = S(t)x_0 + \int_0^t S(t-s)Bv(s) ds, v \in L^2(0, T; U) \right\}.$$

Here is the argument. Let $v \in L^2(0, T; U)$ and let $y(t) = S(t)x_0 + \int_0^t S(t-s)Bv(s) ds$, $0 \leq t \leq T$. As mentioned earlier the infimum defining ϕ is attained. Thus there exists $w \in L^2_{\text{loc}}(0, \infty; U)$ such that $\phi(y(T)) = \int_0^\infty L(z(t), w(t)) dt$ and $z(t) = S(t)y(T) + \int_0^t S(t-s)Bw(s) ds$ for $t \geq 0$. We observe that the pair (\tilde{y}, \tilde{v}) defined by

$$\tilde{y}(t) = \begin{cases} y(t), & 0 \leq t \leq T, \\ z(t-T), & T < t < \infty, \end{cases} \quad \tilde{v}(t) = \begin{cases} v(t), & 0 \leq t \leq T, \\ w(t-T), & T < t < \infty, \end{cases}$$

satisfies (1.2). Hence

$$(3.2) \quad \phi(x_0) = \int_0^\infty L(x(t), u(t)) dt \leq \int_0^\infty L(\tilde{y}(t), \tilde{v}(t)) dt \\ = \int_0^T L(y(t), v(t)) dt + \phi(y(T)).$$

In particular, it follows that

$$\phi(x(T)) \geq \int_T^\infty L(x(s), u(s)) ds.$$

Since the converse inequality is obvious, we may infer that

$$(3.3) \quad \phi(x(T)) = \int_T^\infty L(x(t), u(t)) dt.$$

Comparison of relations (3.2) and (3.3) shows that (x, u) is an optimal pair of problem (3.1), as claimed. Incidentally we have proved also that the value of (3.1) is $\phi(x_0)$.

Let $\{T_n\}$ be the sequence arising in Assumption B.

By Assumption A, $-\infty < H(x, q) < +\infty$ for all $(x, q) \in E \times U$ while Assumption B implies that for n sufficiently large there exists an admissible control $u(t)$ on $[0, T_n]$ such that $x_T = x(T_n, x_0, u) \in \text{int } D(\phi)$. Thus we may apply Theorem 1 in [4] concerning necessary and sufficient condition for optimality in problem (3.1) to deduce the

existence of a continuous function $p_{T_n}: [0, T_n] \rightarrow E$ (the dual extremal arc) which satisfies the equations

$$(3.4) \quad p_{T_n}(t) = S^*(T_n - t)p_{T_n}(T_n) - \int_t^{T_n} S^*(s - t)q_1(s) ds, \quad 0 \leq t \leq T_n,$$

$$(3.5) \quad B^*p_{T_n}(t) = q_2(t), \quad 0 \leq t \leq T_n$$

and transversality conditions

$$(3.6) \quad p_{T_n}(T_n) \in -\partial\phi(x(T_n)).$$

Here $(q_1, q_2) \in L^2(0, T_n; E) \times L^2(0, T_n; U)$ are defined by

$$(3.7) \quad (q_1(t), q_2(t)) \in \partial L(x(t), u(t)), \quad \text{a.e. } t \in]0, T_n[.$$

As pointed out earlier $\partial L: E \times U \rightarrow E \times U$ denotes the subdifferential of function L . Since the Hamiltonian H is the partial conjugate of L , we deduce from (3.7) that

$$(3.8) \quad q_1(t) \in -\partial_x H(x(t), B^*p_{T_n}(t)), \quad \text{a.e. } t \in]0, T_n[,$$

and

$$(3.9) \quad u(t) \in \partial_q H(x(t), B^*p_{T_n}(t)), \quad \text{a.e. } t \in]0, T_n[.$$

Let h be arbitrary in E and let $v \in L^2_{\text{loc}}(R^+; U)$ and $y(t) = S(t)h + \int_0^t S(t-s)Bv(s) ds$, $t \geq 0$, be such that $\phi(h) = \int_0^\infty L(y(s), v(s)) ds$. Let t be arbitrary but fixed on the interval $[0, T_n]$ and let $y_t(s) = y(s-t)$, $v_t(s) = v(s-t)$, $t < s < +\infty$. It follows from (3.7) that

$$L(x(s), u(s)) \leq L(y_t(s), v_t(s)) + (x(s) - y_t(s), q_1(s)) + \langle u(s) - v_t(s), q_2(s) \rangle, \quad \text{a.e. } s > 0.$$

We integrate the latter on the interval $[t, T_n]$ to obtain after some calculation involving (3.4) and (3.5) that

$$\begin{aligned} & -(p_{T_n}(t), x(t) - h) + (p_{T_n}(T_n), x(T_n) - y(T_n - t)) \\ & \quad \geq \int_t^{T_n} L(x(s), u(s)) ds - \int_0^{T_n-t} L(y(s), v(s)) ds \\ & \quad = \phi(x(t)) - \phi(x(T_n)) - \phi(h) + \phi(y(T_n - t)). \end{aligned}$$

(Here we have used, in particular, (3.3).) Combining the latter with (3.6) we get

$$-(p_{T_n}(t), x(t) - h) \geq \phi(x(t)) - \phi(h).$$

As h is arbitrary we may infer that

$$(3.10) \quad -p_{T_n}(t) \in \partial\phi(x(t)) \quad \text{for every } t \in [0, T_n]$$

and therefore (see (3.9)),

$$(3.11) \quad u(t) \in \partial_p H(x(t), -B^*\partial\phi(x(t))), \quad \text{a.e. } t > 0,$$

which completes the proof.

4. The Hamiltonian condition for optimality. We shall say that given continuous arcs $x: R^+ \rightarrow E$ and $p: R^+ \rightarrow E$ satisfy in the “mild” sense the Hamiltonian system

$$(4.1) \quad \begin{aligned} x'(t) - Ax(t) & \in B \partial_q H(x(t), B^*p(t)), \\ p'(t) + A^*p(t) & \in -\partial_x H(x(t), B^*p(t)) \end{aligned}$$

over R^+ if there exist two functions $q_i \in L^2_{\text{loc}}(R^+; E)$, $i = 1, 2$, such that

$$(4.2) \quad \begin{aligned} x(t) &= S(t)x(0) + \int_0^t S(t-s)Bq_1(s) \, ds, \quad \text{for all } t \geq 0, \\ p(t) &= S^*(T-t)p(T) - \int_t^T S^*(s-t)q_2(s) \, ds, \quad \text{for all } 0 < t \leq T < +\infty \end{aligned}$$

and

$$(4.3) \quad q_1(t) \in \partial_q H(x(t), B^*p(t)), \quad q_2(t) \in -\partial_x H(x(t), B^*p(t)) \quad \text{a.e. } t > 0.$$

We note that system (4.3) can be brought into the following form:

$$(4.4) \quad (q_2(t), B^*p(t)) \in \partial L(x(t), q_1(t)), \quad \text{a.e. } t > 0.$$

In the sequel, we shall require at times the following assumptions:

- (i) $H(x, q)$ is strictly concave in x for every $q \in U$.
- (ii) $H(x, q)$ is strictly convex in q for every $x \in E$.
- (iii) $H(x, q)$ is Gâteaux differentiable in x and the pair (A, B) is “controllable”, i.e., if $B^*S^*(t)p_0 = 0$, over some interval $[0, T]$ then $p_0 = 0$.
- (iv) $(0, 0)$ is a saddle point of H and $H(0, 0) = 0$.
- (v) For each $r > 0$, $\inf \{-H(x, 0); |x| = r\} > 0$.
- (vi) For each $r > 0$ there exists a real positive function ω such that $\lim_{t \rightarrow 0} \omega(t)/t = 0$ and

$$(4.5) \quad H(x, q) \leq \omega(\|q\|) \quad \text{for all } |x| \leq r \text{ and } q \in U.$$

THEOREM 2. *Let $x(t)$ be an optimal arc of problem (P) with $|x(0)| < R$. Then under Assumptions A, B and (ii), (iii) there exists an unique function p satisfying along with x the Hamiltonian system (4.1) (in the “mild” sense). Furthermore, one has*

$$(4.6) \quad p(t) \in -\partial\phi(x(t)) \quad \text{for every } t \geq 0.$$

Further, suppose that assumptions (iv) up to (vi) hold. Then

$$(4.7) \quad \lim_{t \rightarrow \infty} x(t) = 0 \quad \text{and} \quad |p(t)| \text{ is bounded over } R^+.$$

If in addition, assumption (i) is satisfied then the function ϕ is Gâteaux differentiable on $S(0, R)$ and

$$(4.8) \quad \lim_{t \rightarrow \infty} p(t) = 0 \quad \text{weakly in } E.$$

Furthermore, for each $x_0 \in E$ there is at most one solution (x, p) to (4.1) satisfying the initial value condition $x(0) = x_0$ and either (4.6) or (4.7).

In particular, Theorem 2 shows that under assumptions (ii)–(vi) the feedback control law (2.14) stabilizes system (1.1). Moreover, $p + \partial\phi(x) = 0$ is an invariant manifold for the Hamiltonian system (1.1) in which the origin is asymptotically stable. Theorem 2 also implies that the motion in the manifold projects as the optimal closed loop motion x .

Remark 3.1. It should be noted that under assumptions A and (v), part (b) of Assumption B is implied by part (a). Here is the argument. Let $x_0 \in S(0; R)$ and let y be a feasible arc in problem (P) satisfying $y(0) = x_0$. Then the conjugacy formula (2.1) implies that $-H(y, 0) \in L^1(R^+)$, which in conjugation with assumption (v), implies that $y(T_n) \rightarrow 0$ for a sequence $T_n \rightarrow +\infty$.

Proof of Theorem 2. Let (x, u) be any optimal pair of problem (P) with $x(0) = x_0 \in S(0; R)$. Let $T_n \rightarrow +\infty$ be the sequence which arises in Assumption B, part (b).

As seen in the proof of Theorem 1 for each n there exists a continuous function $p_n = p_{T_n}: [0, T_n] \rightarrow E$ satisfying together with x equations (3.4), (3.5), (3.8) and (3.9) over $[0, T_n]$. We shall prove that for each n there exists at most one such function p . Assume the contrary, and let p_n^1 be another solution to this system. We set $q_n(t) = p_n(t) - p_n^1(t)$ and notice that (3.8) implies that $B^*q_n(t) \equiv 0$ over $[0, T_n]$, because by virtue of the strict convexity of $H(x, \cdot)$ $\partial_q H(x, q_1) \cap \partial_q H(x, q_2) \neq \emptyset$ unless $q_1 = q_2$. Then (3.9) and first part of assumption (iii) yield

$$q_n(t) = S^*(T_n - t)q(T_n) \quad \text{for all } t \in [0, T_n].$$

Thus by virtue of controllability of the pair (A, B) we conclude that $q_n \equiv 0$ over $[0, T_n]$, as desired.

Hence the continuous function $p(t): R^+ \rightarrow E$ defined by

$$p(t) = p_{T_n}(t) \quad \text{for } t \in [0, T_n]$$

satisfies together with $x(t)$ (in the "mild" sense) the system (4.1) and relation (4.6). The uniqueness of $p(t)$ follows by the same argument. It remains to prove under assumptions (iv)–(vi) relations (4.7) and (4.8). To this end we need the following lemma.

LEMMA 3. *There exist real numbers $\mu_0 > 0$ and μ_1 such that*

$$(4.9) \quad L(y, v) \geq \mu_0(|y| + \|v\|) + \mu_1 \quad \text{for all } (y, v) \in E \times U.$$

Proof. Denote by $M: U \times E \rightarrow]-\infty, +\infty]$ the function

$$(4.10) \quad \begin{aligned} M(p, q) &= \sup \{ \langle p, v \rangle + \langle q, y \rangle - L(y, v); (y, v) \in E \times U \} \\ &= \sup \{ \langle q, y \rangle + H(y, p); y \in E \}. \end{aligned}$$

Since $H(\cdot, p)$ is concave and by virtue of assumption (iv),

$$(4.11) \quad H(y, 0) \leq 0 \leq H(0, 0) \quad \text{for all } (y, p) \in E \times H,$$

we get

$$H(y, p) \leq |y|H(y/|y|, p) \quad \text{for } |y| \geq 1.$$

We set $w = y/|y|$ and use the definition of $\partial_p H$ to obtain

$$(4.12) \quad H(y, p) \leq |y|H(w, 0) + |y|\|p\|\|z_w\| \quad \text{for all } |y| \geq 1, p \in U,$$

where $z_w \in \partial_p H(w, p)$. By assumption (v), $H(w, 0) \leq -\gamma < 0$ for all $|w| = 1$. Furthermore, since ∂H is locally bounded on a neighborhood of the origin (as mentioned earlier, this is a consequence of condition (1.5) and of Assumption A, we may also suppose that $\|z_w\| \leq C$ for all $|w| = 1$, and $\|p\|$ sufficiently small. Applying this to (4.12) we obtain

$$H(y, p) \leq -\gamma|y|/2$$

for all $|y| \geq 1$ and $\|p\| \leq \delta$, where δ is sufficiently small. We deduce from this and formula (4.10) that

$$(4.13) \quad M(p, q) \leq \sup \{ \langle q, y \rangle - \gamma|y|/2; y \in E \} < +\infty$$

if $|q| \leq \gamma/2$ and $\|p\| \leq \delta$. By the theory of conjugate convex function, we have

$$(4.14) \quad L(y, v) = \sup \{ \langle p, v \rangle + \langle q, y \rangle - M(p, q); (q, p) \in E \times U \}$$

and therefore

$$L(y, v) \geq \rho(|y| + \|v\|) - M(\rho w, \rho w_1), \quad \|w\| = 1, \quad |w_1| = 1$$

for all $(y, v) \in E \times U$ and $\rho > 0$. Applying (4.13), we obtain (4.9), as claimed.

Proof of Theorem 2 (continued). From inequality (4.9) we see that

$$\int_t^{t+h} (|x(s)| + \|u(s)\|) ds \leq C(h+1) \quad \text{for all } t, h \geq 0,$$

while formula (1.2) yields

$$(4.15) \quad |x(t+h)| \leq C \left(|x(t)| + \int_t^{t+h} \|u(s)\| ds \right) \quad \text{for } t \geq 0, \quad h \in [0, 1],$$

and therefore

$$(4.16) \quad |x(t+h)| \leq C(|x(t)| + 1) \quad \text{for all } t \geq 0, \quad h \in [0, 1].$$

(We shall denote by C several positive constants independent of t and h .)

On the other hand, as remarked earlier, the conjugacy formula (2.1) yields

$$L(x(t), u(t)) \geq -H(x(t), 0) \quad \text{a.e. } t > 0,$$

which implies that $-H(x, 0) \in L^1(R^+)$. Hence

$$-H(x(t), 0) \leq \delta(T) \quad \text{for } t \in [T, +\infty[\setminus E_T$$

where $\delta(T)$ and the Lebesgue measure $m(E_T)$ of E_T tends to zero as $T \rightarrow \infty$. From this and assumption (v) we conclude that

$$(4.17) \quad |x(t)| \leq \eta(T) \quad \text{for } t \in [T, +\infty[\setminus E_T$$

where $\lim_{T \rightarrow \infty} \eta(T) = 0$. Combining (4.16) and (4.17), we deduce first that $|x(t)|$ is bounded over R^+ . Next, we use the inequality

$$L(x, u) \geq \langle q, u \rangle - H(x, q) \quad \text{for all } q \in U$$

to get

$$\rho \|u\| \leq L(x, u) + H(x, \rho w)$$

where $\rho > 0$ and $w = q/\|q\|$. Thus assumption (vi) yields

$$\int_T^{T+h} \|u(s)\| ds \leq \frac{1}{\rho} \int_T^{T+h} L(x(s), u(s)) ds + \omega(\rho)h \quad \text{for all } T, h > 0$$

where $\lim_{\rho \rightarrow 0} \omega(\rho) = 0$. Inasmuch as $L(x, u) \in L^1(R^+)$, we obtain

$$(4.18) \quad \int_T^{T+h} \|u(s)\| ds \leq \chi(T) \quad \text{for all } T > 0 \text{ and } h \in [0, 1]$$

where $\lim_{T \rightarrow \infty} \chi(T) = 0$. We deduce further from (4.15) and (4.18) that

$$|x(t+h)| \leq C(|x(t)| + \chi(t)) \quad \text{for } t \geq 0 \text{ and } h \in [0, 1]$$

where C is independent of t and h . Then, making use once again of inequality (4.17), we deduce that $\lim_{t \rightarrow \infty} x(t) = 0$. On the other hand $\partial\phi$ is locally bounded at 0 because by Assumption B, $D(\phi)$ (and therefore $D(\partial\phi)$) contains the ball $S(0; R)$. Then we see from (4.7) that $|p(t)|$ is bounded over R^+ . Next we shall prove the uniqueness of the pair (x, p) satisfying the system (4.1), the condition $x(0) = x_0$ and (4.7). Let (\tilde{x}, \tilde{p}) be

another solution to (4.1) satisfying (4.7) and let \tilde{q}_1, \tilde{q}_2 be the corresponding functions arising in formula (4.3). Invoking formula (4.4), we get, after some calculations involving equations (4.2),

$$\int_0^T L(\tilde{x}, \tilde{q}_1) dt \leq \int_0^T L(y, v) dt - (\tilde{p}(T), y(T) - \tilde{x}(T))$$

for all $T \geq 0$, $v \in L^2_{\text{loc}}(R^+; U)$ and $y(t) = S(t)x_0 + \int_0^t S(t-s)Bv(s) ds$. Assume further that $\lim_{t \rightarrow \infty} y(t) = 0$. Then the latter inequality yields

$$\int_0^\infty L(\tilde{x}, \tilde{q}_1) dt \leq \int_0^\infty L(y, v) dt.$$

Since assumption (i) implies the strict convexity of L , taking $y = (x + \tilde{x})/2$ and $v = (u + \tilde{q}_1)/2$ in the preceding inequality we see that $x = \tilde{x}$ and $u = \tilde{q}_1$ as claimed. The uniqueness of solution (x, p) to (4.1) satisfying (4.6) and the initial condition $x(0) = x_0$ follows by a parallel argument by observing that if the arc $x(t)$ satisfies these conditions then for every $T > 0$ it is optimal in the control problem,

$$(4.19) \quad \min \left\{ \int_0^T L(y(t, x_0, v), v(t)) dt + \phi(y(T)); v \in L^2(0, T; U) \right\}.$$

Let T be a positive number with the property that there exists a control $u \in L^2(0, T; U)$ such that $y(T; x_0; u) \in S(0, R)$. (Since part (b) of Assumption B holds, such numbers always exist.)

Denote by $\Lambda \subset E \times E$ the set of all pairs (x_0, p_0) with the property that there is a solution ("mild") (\tilde{x}, \tilde{p}) to system (4.1) satisfying

$$(4.20) \quad \tilde{x}(0) = x_0, \quad \tilde{p}(0) = p_0,$$

$$(4.21) \quad \tilde{p}(T) \in -\partial\phi(\tilde{x}(T)).$$

By virtue of equivalence between problem (4.19) with $|x_0| < R$ and the system (4.1) and transversality condition (4.21) and $\tilde{x}(0) = x_0$ (here we use again Theorem 1 in [4]) we deduce that $\Lambda x_0 \neq \emptyset$ for each $x_0 \in S(0; R)$. Furthermore, by uniqueness of such solution (\tilde{x}, \tilde{p}) we see that $\tilde{x} = x$ and $\tilde{p} = p$ over $[0, T]$ where (x, p) denotes as above the unique solution to system (4.1) satisfying (4.7) and initial condition $x(0) = x_0$. Thus Λ can be equivalently defined as

$$\Lambda x_0 = p(0).$$

In particular we deduce that Λ is single valued and $-\Lambda x_0 \in \partial\phi(x_0)$ for all $x_0 \in S(0; R)$. We shall prove that $-\Lambda = \partial\phi$ on $S(0; R)$. To this end, first we shall show that $-\Lambda$ is maximal monotone on $E \times E$, i.e. $R(I - \Lambda) = E$. Let y_0 be any fixed element of E . Consider the equation

$$(4.22) \quad x_0 - \Lambda x_0 = y_0$$

and the differential system over $[0, T]$

$$(4.23) \quad \begin{aligned} \tilde{x}' - A\tilde{x} &\in B \partial_q H(\tilde{x}, B^* \tilde{p}), \\ \tilde{p}' + A^* \tilde{p} &\in -\partial_x H(\tilde{x}, B^* \tilde{p}) \end{aligned}$$

with boundary value conditions

$$(4.24) \quad \tilde{p}(0) - \tilde{x}(0) = y_0, \quad \tilde{p}(T) \in -\partial\phi(\tilde{x}(T)).$$

Here $[0, T]$ is an arbitrary finite interval. Again invoking Theorem 1 in [4] we deduce by virtue of Assumptions A and B (part (a)) that system (4.23), (4.24) represent the Hamiltonian conditions for optimality in control problem

$$(4.25) \quad \min \left\{ \int_0^T L(y(t); y(0); v), v(t) dt + \phi(y(T)) + \frac{1}{2}|y(0) - y_0|^2; v \in L^2(0, T; U) \right\},$$

which by virtue of condition (2.4) admits a solution (\tilde{x}, \tilde{u}) . We conclude therefore that system (4.23) and (4.24) has a solution (\tilde{x}, \tilde{p}) . Recalling the definition of Λ we finally deduce that $x_0 = \tilde{x}(0)$ satisfies equation (4.22) as claimed. Hence $-\Lambda$ is a maximal monotone single valued operator from E into itself which agrees with $\partial\phi$ on $S(0, R)$. We have therefore shown that for each $x_0 \in S(0, R)$, $\partial\phi(x_0)$ is singleton. This fact implies that ϕ is Gâteaux differentiable at every point $x_0 \in S(0, R)$ and $\partial\phi(x_0) = \nabla\phi(x_0)$ (the gradient of ϕ at x_0). On the other hand, $-\Lambda$ is demicontinuous on $S(0, R)$ (i.e., strongly-weakly continuous) because it is maximal monotone single valued and everywhere defined on $S(0, R)$. Since $\Lambda = -\partial\phi$ on $S(0, R)$ we see from relations (4.6) and (4.7) that

$$p(t) \rightarrow 0 \text{ weakly in } E \quad \text{as } t \rightarrow +\infty$$

thereby completing the proof of Theorem 2.

COROLLARY 1. *The conclusions of Theorem 2 remain valid if instead of condition (iii) we assume that $N(B^*) = \{0\}$.*

Proof. A glance to the preceding proof reveals that assumption (iii) was used only to prove the uniqueness of a function p which satisfies together with a given function x the Hamiltonian system (4.1) over a given interval. If the kernel $N(B^*)$ of B^* consists only of the zero element, then as seen earlier this fact is implied by the strict convexity of $H(x, q)$ as a function of q .

5. The “saddle point” behavior of Hamiltonian systems. In this section the conclusions of Theorem 2 are restated in a form which resembles the classical behavior of ordinary differential systems in a neighborhood of a saddle point.

The differential system studied here is

$$(5.1) \quad \begin{aligned} x'(t) - Ax(t) &\in \partial_p H(x(t), p(t)), \\ p'(t) + A^*p(t) &\in -\partial_x H(x(t), p(t)) \end{aligned}$$

where as before, A is the infinitesimal generator of a continuous semigroup $S(t): R^+ \rightarrow L(E, E)$ and H is a concave-convex function on $E \times E$. Of course it will be assumed that the function H satisfies Assumptions A and (i) up to (vi) (by virtue of Corollary 1, assumption (iii) will be superfluous) and, in addition:

(v)' For each $r > 0$, $\inf \{H(0, p); |p| = r\} > 0$.

(vi)' For each $r > 0$ there exists a real positive function such that $\lim_{t \rightarrow 0} \vartheta(t)/t = 0$ and

$$(5.2) \quad -H(x, q) \leq \vartheta(|x|) \quad \text{for } |q| \leq r \text{ and all } x \in E.$$

Let Λ denote the set of all pairs $(x_0, p_0) \in E \times E$ such that the Hamiltonian system (5.1) has a “mild” solution $(x(t), p(t))$ over R^+ satisfying

$$(5.3) \quad x(0) = x_0, \quad p(0) = p_0,$$

$$(5.4) \quad \lim_{t \rightarrow \infty} x(t) = 0, \quad \lim_{t \rightarrow \infty} p(t) = 0$$

where the above limits are taken in the strong topology of E . Λ is the graph of an operator (possibly multivalued) which will be identified with Λ by setting $p_0 \in \Lambda x_0$ iff $(x_0, p_0) \in \Lambda$. We shall denote with $D(\Lambda)$ the domain of Λ , i.e. $D(\Lambda) = \{x_0 \in E; \Lambda x_0 \neq \emptyset\}$. In general Λ is not everywhere defined on E . However, one has

THEOREM 3. *Let H satisfy assumptions (i), (ii), (iv)–(vi) and (v)', (vi)'. Then:*

(a) *For each $(x_0, p_0) \in \Lambda$ the solution to (5.1) satisfying (5.3) and (5.4) is unique and remains in Λ .*

(b) *There is an open neighborhood V of 0 such that $V \subset D(\Lambda)$ and Λ is single valued, injective and demicontinuous on V .*

(c) *There exists a convex lower-semicontinuous function ϕ on E which is Gâteaux differentiable on V and whose Gâteaux differential $\nabla \phi$ agrees with $-\Lambda$ on V .*

Proof. Let L be the Lagrangian function associated with H , i.e.,

$$(5.5) \quad L(y, v) = \sup \{(p, v) - H(y, p); p \in E\}, \quad (y, v) \in E \times E.$$

By the theory of concave-convex functions (see, e.g., [15]) L is convex and lower-semicontinuous on $E \times E$. Moreover, the minimax inequality (4.11) in the proof of Theorem 2 yields

$$(5.6) \quad L(0, 0) = 0; \quad L(y, v) \geq 0 \quad \text{for all } (y, v) \in E \times E.$$

We have in mind applying Theorem 2 in the special case $U = E$, $B = I$ and L defined by formula (5.5). To this purpose we need to verify Assumption B in this particular case. Since the system (1.1) with $B = I$ is obviously stabilizable, by Lemma 2 it suffices to show that $(0, 0) \in \text{int } D(L)$. This is carried out in the following lemma.

LEMMA 4. *The function L is finite on a neighborhood of the origin.*

Proof. The argument of the proof is based on conjugacy formula (5.5) and it closely resembles that used in the proof of Lemma 3 for proving inequality (4.13). Note, however, that in this case assumption (v)' is used instead of (v).

We are now in a position to conclude that Assumption B is satisfied for some R suitably chosen. We set $V = S(0; R)$.

Thus, we can place ourselves in the context of Theorem 2, according to which for each $x_0 \in V$ there exists a solution $(x(t), p(t))$ to (5.1) over R^+ satisfying $x(0) = x_0$, $\lim_{t \rightarrow \infty} x(t) = 0$, $\lim_{t \rightarrow \infty} p(t) = 0$ weakly in E . Since the assertions (a) and (c) of the theorem about the uniqueness and the Gâteaux differentiability of ϕ are covered by Theorem 2, we confine ourselves to prove that

$$(5.7) \quad \lim_{t \rightarrow \infty} p(t) = 0 \quad \text{strongly in } E.$$

Keeping in mind the definition of Λ we could then conclude that $V \subset D(\Lambda)$ thereby completing the proof of (b). The proof of convergence of $p(t)$ closely parallels the proof of convergence of $x(t)$ but with some simplifications.

To this end we consider again the conjugate function M (see (4.10)) and note that the Hamiltonian system (5.1) can be equivalently written as

$$L(x(t), q_1(t)) + M(p(t), q_2(t)) = (p(t), q_1(t)) + (x(t), q_2(t)), \quad \text{a.e. } t > 0$$

where q_1 and q_2 are the functions arising in (4.2) and (4.3). Integrating the latter over $[0, T]$ we find that

$$\int_0^T L(x(t), q_1(t)) dt + \int_0^T M(p(t), q_2(t)) dt = (x(T), p(T)) - (x_0, p_0).$$

Since

$$M \geq 0 \quad \text{on } E \times E \quad \text{and} \quad \lim_{T \rightarrow \infty} (x(T), p(T)) = 0,$$

we infer that $M(p, q_2) \in L^1(R^+)$ and therefore the inequality

$$M(p(t), q_2(t)) \geq (q_2(t), y) + H(y, p(t)) \quad \text{for all } y \in E$$

implies that $H(0, p) \in L^1(R^+)$. Then assumption (vi)' yields

$$(5.8) \quad |p(t)| \leq \delta(T) \quad \text{for } t \in [T, +\infty[\setminus E_T$$

where the Lebesgue measure of E_T and $\delta(T)$ tends to zero as $T \rightarrow +\infty$.

On the other hand, by definition of "mild" solution one has

$$p(t) = S^*(h)p(t+h) - \int_t^{t+h} S^*(s-t)q_2(s) \, ds, \quad \text{for all } t, h \geq 0$$

and therefore

$$(5.9) \quad |p(t)| \leq C \left(|p(t+h)| + \int_t^{t+h} |q_2(s)| \, ds \right) \quad \text{for all } t \geq 0 \text{ and } h \in [0, 1]$$

where C is independent of t and h .

Next we see from formula (4.10) that

$$M(p, q_2) \geq \rho |q_2| + H(\rho w, p) \quad \text{for all } \rho > 0$$

where $|w| = 1$. Invoking assumption (vi)' and recalling that $M(p, q_2) \in L^1(R^+)$ we get that

$$\int_t^{t+1} |q_2(s)| \, ds \leq \chi(t) \quad \text{for all } t \geq 0$$

where $\lim_{t \rightarrow \infty} \chi(t) = 0$. Substituting the latter in (5.9) we see that

$$|p(t)| \leq C(|p(t+h)| + \chi(t)) \quad \text{for all } t \geq 0, \quad h \in [0, 1],$$

which in conjunction with (5.8) implies that $p(t)$ approaches zero as t tends to $+\infty$. The proof of Theorem 3 is, therefore, complete.

Remark 5.1. One might ask whether Λ is a homeomorphism of V onto another neighborhood V_1 of the origin. This happens, for example, if A is bounded (see [2]) and it is easy to see that this fact requires that $\lim_{|x_0| \rightarrow \infty} \phi(x_0)/|x_0| \geq \beta > 0$. However, the results established by A. Pazy [12] show that this does not hold unless A is the infinitesimal generator of a continuous group of linear bounded operators.

6. Examples. In this section we shall illustrate the general theory with two specific examples.

Example 1. Consider the distributed control problem; Minimize

$$(6.1) \quad \int_0^\infty \int_\Omega g(x, y(x, t), u(x, t)) \, dx \, dt$$

in $y(x, t): \Omega \times R^+ \rightarrow R$ and $u \in L^2_{\text{loc}}(R^+; L^2(\Omega))$ subject to the constraints

$$(6.2) \quad \partial y / \partial t - \Delta y = Bu \quad \text{for } x \in \Omega, \quad t > 0,$$

$$(6.3) \quad y(x, t) = 0 \quad \text{for } x \in \Gamma, \quad t > 0,$$

$$(6.4) \quad y(x, 0) = y_0(x) \quad \text{for } x \in \Omega,$$

$$(6.5) \quad 0 \leq u(x, t) \leq 1 \quad \text{for } x \in \Omega, \quad t > 0,$$

where Ω is a bounded and open subset of R^n with a sufficiently smooth boundary Γ , B is a linear continuous operator from $L^2(\Omega)$ into $L^2(\Omega)$ and $y_0 \in L^2(\Omega)$ is a given function. As for the function $g: R^n \times R \times R \rightarrow R$, it is assumed measurable in x , convex in (y, u) and $g(x, 0, 0) = 0$, a.e. $x \in \Omega$.

Further suppose that there exist $\alpha \in R$ and $\beta \in L^1(\Omega)$ such that

$$(6.6) \quad g(x, y, 0) \leq \alpha|y|^2 + \beta(x) \quad \text{a.e. } x \in \Omega \quad \text{for all } y \in R.$$

To formulate (6.1) as a problem of type (P) we set $E = U = L^2(\Omega)$, $A = \Delta$ (the Laplace operator) with $D(A) = H_0^1(\Omega) \cap H^2(\Omega)$ and $L(\cdot, \cdot)$ defined by

$$(6.7) \quad L(y, v) = \int_{\Omega} g_0(x, y(x), v(x)) \, dx \quad \text{for } (y, v) \in E \times E$$

where

$$g_0(x, y, v) = \begin{cases} g(x, y, v) & \text{if } 0 \leq v \leq 1, \\ +\infty & \text{otherwise.} \end{cases}$$

Obviously Assumption A of Theorem 1 is implied by condition (6.6) and the fact that for every $y \in L^2(\Omega)$ the effective domain of $L(y, \cdot)$ reduces to $\{v \in L^2(\Omega); 0 \leq v(x) \leq 1 \text{ a.e. } x \in \Omega\}$. As for Assumption B, it is implied by the condition that $L(y, 0) = 0$ and the fact that the semigroup generated by A is exponentially stable (see Remark 2.1). Thus, invoking Theorem 1, we see that any optimal control $u(x, t)$ of problem (6.1) is expressed as a function of optimal state $y(x, t)$ as

$$\begin{aligned} u(x, t) &= 0 \quad \text{if } (B^* \partial \phi(y))(x, t) - \partial_v g(x, y(x, t), 0) \leq 0, \\ u(x, t) &= 1 \quad \text{if } (B^* \partial \phi(y))(x, t) - \partial_v g(x, y(x, t), 1) \geq 1, \\ u(x, t) &= (\partial_v g(x, y(x, t), \cdot))^{-1}(B^* \partial \phi(y)(x, t)) \\ &\quad \text{if } 0 < (\partial_v g(x, y(x, t), \cdot))^{-1}(B^* \partial \phi(y)(x, t)) < 1, \end{aligned}$$

where $\partial_v g$ stands for the subdifferential of $g(x, y, v)$ as a function of v and $\phi: L^2(\Omega) \rightarrow R^+$ is the function defined by (1.5). Under obvious additional assumptions on the function g , Theorem 2 is also applicable but the details are left to the reader.

Remark 6.1. According to (2.16) and comments preceding it, we may say that the operator $K = \partial \phi$ satisfies the equation

$$(6.8) \quad \int_{\Omega} (h_0(x, y(x), -(B^* K)(y)(x)) - \Delta y(x) K(y)(x)) \, dx = 0$$

for all $y \in H_0^1(\Omega) \cap H^2(\Omega)$ where

$$h_0(x, y, p) = \sup \{pv - g(x, y, v); 0 \leq v \leq 1\}$$

for all $x \in \Omega$ and $y, p \in R$.

Example 2. We present an example illustrating Theorem 3 on a mixed boundary value problem involving hyperbolic systems in two variables. Specifically, we are concerned with a system of the form

$$(6.9) \quad \begin{aligned} \frac{\partial y}{\partial t} - A_1(x) \frac{\partial y}{\partial x} - A_2(x) y &\in \partial_z G(x, y, z), \\ \frac{\partial z}{\partial t} - \frac{\partial}{\partial x} (A_1^*(x) z) + A_2^*(x) z &\in -\partial_y G(x, y, z) \end{aligned}$$

where the unknown n -dimensional functions y and z are to be found in $\{(x, t); 0 \leq x \leq 1, t \geq 0\}$ and $A_1(x), A_2(x)$ are $n \times n$ matrices which satisfy

$$(6.10) \quad A_1(x) \text{ is symmetric for each } x \in [0, 1]$$

$$(6.11) \quad A_1 \in C^1([0, 1]; R^n \times R^n), \quad A_2 \in C([0, 1]; R^n \times R^n).$$

Further, suppose that for each $x \in [0, 1]$, $A_1(x)$ has eigenvalues $\mu_1(x), \dots, \mu_n(x)$ satisfying

$$(6.12) \quad \mu_1(x) \leq \dots \leq \mu_p(x) < 0 < \mu_{p+1}(x) \leq \dots \leq \mu_n(x), \quad \text{for } x \in [0, 1]$$

and

$$\mu_j = \mu_{j+1} \quad \text{if } \mu_j(x_0) = \mu_{j+1}(x_0) \quad \text{for some } x_0 \in [0, 1].$$

It will be convenient to write the vectorial functions y and z in the form

$$y = \begin{pmatrix} y^- \\ y^+ \end{pmatrix}, \quad z = \begin{pmatrix} z^- \\ z^+ \end{pmatrix}$$

where

$$y^-, z^- \in R^p \quad \text{and} \quad y^+, z^+ \in R^q; \quad p + q = n.$$

We impose the boundary conditions

$$(6.13) \quad y^-(0, t) - D_0 y^+(0, t) = 0, \quad y^+(1, t) - D_1 y^-(1, t) = 0,$$

$$(6.14) \quad \begin{aligned} D_0^* (A_1(0)z(0, t))^- + (A_1(0)z(0, t))^+ &= 0, \\ D_1^* (A_1(1)z(1, t))^+ + (A_1(1)z(1, t))^- &= 0 \end{aligned}$$

where the matrices D_0 and D_1 have the dimensions $q \times p$ and $p \times q$, respectively. Making an orthogonal transformation (see, e.g., [17]) we may reduce the problem to the case in which the matrix A has the form

$$A_1(x) = \begin{pmatrix} A_1^-(x) & 0 \\ 0 & A_1^+(x) \end{pmatrix}$$

where $A_1^-(x) = \text{diag}(\mu_1(x), \dots, \mu_p(x))$ and $A_1^+(x) = \text{diag}(\mu_{p+1}(x), \dots, \mu_n(x))$. Then the boundary conditions (6.14) may be rewritten as

$$(6.15) \quad \begin{aligned} D_0^* A_1^-(0) z^-(0, t) + A_1^+(0) z^+(0, t) &= 0, \\ D_1^* A_1^+(1) z^+(1, t) + A_1^-(1) z^-(1, t) &= 0. \end{aligned}$$

As for the function G , it is assumed to be measurable in x , continuous in (y, z) , strictly concave in y and strictly convex in z for each $x \in [0, 1]$. Further, assume that

$$(6.16) \quad G(x, y, 0) \leq G(x, 0, 0) = 0 \leq G(x, 0, z),$$

$$(6.17) \quad -\beta(x)|y|^q \leq G(x, y, z) \leq \alpha(x)|z|^p$$

for all $x \in [0, 1]$ and y, z in R . Here α, β are positive L^∞ functions on $[0, 1]$ and $p, q \in [1, 2]$.

It is easily seen that under these conditions the concave-convex function

$$H: L^2(0, 1; R^n) \times L^2(0, 1; R^n) \rightarrow R$$

defined by

$$H(y, z) = \int_{\Omega} G(x, y(x), z(x)) \, dx$$

satisfies assumptions A, (i), (ii), (iv)–(vi), (v)' and (vi)'.

Furthermore, if we assume that

$$(6.18) \quad A_2(x) + A_2^*(x) - A_1'(x) \leq 0 \quad \text{for } x \in [0, 1]$$

and

$$(6.19) \quad A_1^+(0) + D_0^* A_1^-(0) D_0 \geq 0, \quad A_1^-(1) + D_1^* A_1^+(1) D_1 \geq 0,$$

it follows that the operator $A: L^2(0, 1; R^n) \rightarrow L^2(0, 1; R^n)$ defined by $(Ay)(x) = A_1 y'(x) + A_2 y(x)$ a.e. $x \in]0, 1[$ with domain $D(A) = \{y \in L^2(0, 1); y' \in L^2(0, 1), y \text{ satisfies (6.13)}\}$ is maximal dissipative and therefore generates a strongly continuous semigroup on $L^2(0, 1; R^n)$. Thus, in terms of H and A , the system (6.9) with boundary value conditions (6.13) and (6.14) can be expressed in form (5.1). Then, invoking Theorem 3, we may infer: *Under assumptions (6.10)–(6.12), (6.16)–(6.19), there exists a neighborhood $V = \{y_0 \in L^2(0, 1; R^n); \|y_0\|_{L^2(0, 1; R^n)} \leq R\}$ of the origin and a demicontinuous dissipative operator $\Lambda: V \rightarrow L^2(0, 1; R^n)$ such that for each $y_0 \in V$ the system (6.9) with boundary value conditions (6.13), (6.14) has a unique solution $(y(x, t), z(x, t))$ satisfying*

$$(6.20) \quad y(x, 0) = y_0(x), \quad z(x, 0) = \Lambda y_0(x) \quad \text{a.e. } x \in]0, 1[,$$

$$(6.21) \quad \lim_{t \rightarrow \infty} y(\cdot, t) = 0, \quad \lim_{t \rightarrow \infty} z(\cdot, t) = 0 \quad \text{in } L^2(0, 1; R^n).$$

REFERENCES

- [1] V. BARBU, *Convex control problems on infinite intervals*, J. Math. Anal. Appl., to appear.
- [2] ———, *Hamiltonian systems in a neighborhood of a saddle point*, Trans. Amer. Math. Soc., to appear.
- [3] ———, *Nonlinear Semigroups and Evolution Equations in Banach Spaces*, Noordhoff, Leyden (Netherlands)-Bucharest (Romania), 1976.
- [4] ———, *Constrained control problems with convex cost in Hilbert spaces*, J. Math. Anal. Appl., 57 (1976), no. 3, pp. 502–528.
- [5] H. BRÉZIS, *Opérateurs Maximaux Monotones et Semigroupes de Contractions dans les Espaces de Hilbert*, Math. Studies, 5, North-Holland, Amsterdam, 1973.
- [6] R. CURTAIN AND A. J. PRITCHARD, *The infinite-dimensional Riccati equations*, J. Math. Anal. Appl., 47 (1974), pp. 43–58.
- [7] R. DATKO, *A linear control problem in an abstract Hilbert spaces*, J. Differential Equations, 9 (1971), pp. 346–359.
- [8] ———, *Unconstrained control problems with quadratic cost*, this Journal, 11 (1973), pp. 32–52.
- [9] M. C. DELFOUR, C. MCCALLA AND S. K. MITTER, *Stability and the infinite quadratic cost problem for linear hereditary differential systems*, this Journal, 13 (1975), pp. 48–88.
- [10] J. L. LIONS, *Sur le Contrôle Optimal de Systèmes Governés par des Equations aux Dérivées Partielles*, Dunod, Gauthier-Villars, Paris, 1968.
- [11] D. L. LUKES AND D. L. RUSSELL, *The quadratic criterion for distributed systems*, this Journal, 7 (1969), pp. 101–121.
- [12] A. PAZY, *On the applicability of Lyapunov's theorem in Hilbert space*, SIAM J. Math. Anal., 2 (1972), pp. 291–294.
- [13] A. J. PRITCHARD AND R. TRIGGIANI, *Stabilizability in Banach space*, Rep. 35, Control Theory Centre, University of Warwick, England.
- [14] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1969.
- [15] ———, *Monotone operators associated with saddle functions and minimax problems*, Nonlinear Functional Analysis, F. Browder, ed., Proceedings Symposia in Pure Mathematics, vol. XVIII, Part I, American Mathematical Society, Providence, RI, 1970, pp. 241–251.
- [16] ———, *Saddle points of Hamiltonian systems in convex problems of Lagrange*, J. Optimization Theory Appl., 12 (1973), pp. 367–390.
- [17] D. L. RUSSELL, *Quadratic performance criteria in boundary control of linear symmetric hyperbolic systems*, SIAM J. Control, 11 (1973), pp. 475–509.

MARTINGALE PROJECTION AND LINEAR FILTERING IN HILBERT SPACES. I: THE THEORY*

JEAN-YVES OUVRARD†

Abstract. Using a theorem on the projection of Hilbertian martingales and a Hilbertian theorem of the Girsanov type, previously established by the author, to study the filtering problem for systems governed by linear evolution equations in Hilbert spaces, we show that if the observation is a function of the state through an unbounded operator and with values in a Hilbert space (under some hypotheses on the observation noise), the filtered state is obtained as the solution of a differential stochastic equation, the coefficients of which are given by a solution of an operator differential Riccati equation. We show that the linear filtering problem can be solved only if the solution of this equation is unique.

If the observation is made through a bounded operator, we find again, with the innovation method, results from A. Bensoussan (1971) in the partial differential equation context, and from R. F. Curtain (1975) and S. K. Mitter–R. Vinter (1974) in the linear hereditary equation context; the application to the linear hereditary equation will be presented in another paper.

Introduction. We study a linear filtering problem in Hilbert spaces. The state $(x_t)_{t \in [0, T]}$ is a process with values in a real separable Hilbert space H , the solution of a linear stochastic differential equation; the linear operators $(A_t)_{t \in [0, T]}$ which appear in this equation are unbounded on H . Though the formulation adopted here is different, we recall that an investigation of such filtering problems has already been made fairly extensively (cf. e.g. [2], [3], [5], [11]). In some papers, even more general evolution systems are considered, since we consider here only systems having a strong solution. But, what is new, is that we consider an observation process $(y_t)_{t \in [0, T]}$ related with the state through a family of unbounded operators $(C_t)_{t \in [0, T]}$ on H , and also that the observation takes its values in an infinite dimensional space K (which allows us to consider observations on the boundary of a domain).

The method used is quite original for studying infinite dimensional systems: we have used the innovation approach, well known in the finite dimensional case (cf. e.g. [1]). The principle consists in exhibiting two martingales Z_0 and Z_s related with the state and observation processes respectively. But, contrary to the finite dimensional case, we do not know how to solve the famous problem of the equality of the σ -fields generated by the observation and innovation processes. So we have used a measure transformation method for using a representation theorem for an Hilbertian martingale adapted to the family of the σ -fields generated by a Gaussian Hilbertian martingale. This idea proved to account successfully for the one dimensional case (M. Fujisaki, G. Kallianpur and H. Kunita [7]). This idea is equally effective in the infinite dimensional case, but in the linear context. This way, we obtain the filtering equations of the Kalman–Bucy type. To have a complete solution for this filtering problem, we have to make sure that the solution of the Riccati equation is unique—its existence being proved by the method used for solving the problem. We have given a sufficient condition for the uniqueness—it seems difficult to obtain a general answer and we have approached the uniqueness problem by a duality technique: we introduce a “dual” linear deterministic control problem, the solution of which is unique and is a function of a solution of the Riccati equation. Thus, we can prove the unicity of the solution of this Riccati equation.

* Received by the editors July 13, 1977.

† I.R.M.A. Université Scientifique et Médicale de Grenoble, B.P.53, F-38041 Grenoble-Cedex-France.

1. Description of the system. D, H, K, L are Hilbert spaces such that D is a dense subspace of H . The canonical injection i from D to H is continuous and dense. $(A_t)_{t \in [0, T]}$ is a family of linear operators in H , the state space, with domain containing D , such that $A_t|_D \in \mathcal{L}(D, H)$ and, $\forall h \in D$, $A \cdot h$ is continuous from $[0, T]$ to H . Elsewhere, let us have a family $(C_t)_{t \in [0, T]}$ of linear operators from H to K , the state of the observations, with domain containing D and such that $C_t|_D \in \mathcal{L}(D, K)$ and that, $\forall h \in D$, the application $C \cdot h$ is measurable and bounded from $[0, T]$ to K .

We consider a Wiener process W on L relative to the process basis $(\Omega, \mathcal{F}(\mathcal{F}_t)_{t \in [0, T]}, P)$ that satisfies the customary conditions, (Ω, \mathcal{F}) being moreover a Blackwell space; the covariance operator of W is the nuclear operator $\mathcal{W} \in \mathcal{L}_1(L)$.

Let B and D be two operator valued functions satisfying:

$B \in \mathcal{L}^\infty([0, T]; \mathcal{L}(L, H))$ and $D \in \mathcal{L}^\infty([0, T]; \mathcal{L}(L, K))$.

We make the following hypothesis about the observation noise: if $Q_t = D_t \mathcal{W} D_t^*$, $\forall t \in [0, T]$, we suppose that:

- (i) $\text{rg } C_t \subset \text{rg } Q_t, \forall t \in [0, T]$;
- (ii) $\forall h \in D$ the set $\{\|Q_t^+ C_t h\| \mid t \in [0, T]\}$ is bounded in K where Q_t^+ is the pseudo-inverse (unbounded, with domain containing $\text{rg } Q_t$)—for a definition see [13].
- (iii) $\text{rg } D_t \mathcal{W} B_t^* \subset \mathcal{D}(Q_t^+), \forall t \in [0, T]$.

Remarks. 1) The hypothesis about the independence of noises on the state and the observation is equivalent to the following condition: $D_t \mathcal{W} B_t^* = 0, \forall t \in [0, T]$; condition (iii) is then obviously fulfilled.

2) Hypothesis (iii) is automatically satisfied in the finite dimensional case; indeed, in this case, we have $\text{rg } D_t \mathcal{W} B_t^* \subset \text{rg } Q_t$. In the infinite dimensional case, we can only show that we have $\text{rg } D_t \mathcal{W} D_t^* \subset \text{rg } \overline{Q_t}$.

Now we make the following hypothesis on the family $(A_t)_{t \in [0, T]}$. We suppose that a Green kernel is associated with this family, i.e. a family of operators $\{G(t, s) \in \mathcal{L}(H) \mid 0 \leq s \leq t \leq T\}$ such that:

- 1) $G(t, t) = 1_H, \forall t \in [0, T]$,
- 2) $G(u, s) = G(u, t) \circ G(t, s)$ if $0 \leq s \leq t \leq u \leq T$,
- 3) $\forall h \in H, G(\cdot, s)h$ is continuous from $[s, T]$ in H ,
- 4) $\forall h \in D$ the Cauchy problem

$$\begin{aligned} \frac{dx_t}{dt} &= A_t x_t, & s \leq t \leq T \\ x_s &= h \end{aligned}$$

has a unique solution in $L_D^2([0, T]) \cap C_H([0, T])$ given by $x_t = G(t, s)h$ (i.e. $x_t = h + \int_s^t A_u x_u du$).

The Green kernel G and B are supposed to have extra properties assuring the existence and unicity of the stochastic differential equation:

$$\begin{aligned} (1.1) \quad dx_t &= A_t x_t dt + B_t dW_t \\ x(0) &= x_0 \end{aligned}$$

(see, for example, [4] and [10] for different theorems about existence and unicity). A solution of (1.1) is a process with values in H , such that any trajectory is almost everywhere (for the Lebesgue measure on $[0, T]$) in D and which satisfies (1.1). Moreover, the family $(A_t)_{t \in [0, T]}$ is supposed to be such that the solution of (1.1) is in $L_D^2([0, T] \times \Omega, dt \otimes P)$ (cf. [10] for such a condition).

The state x_t is subjected to the observation y_t , a process with values in K , related with the state through the equation:

$$(1.2) \quad \begin{aligned} dy_t &= C_t x_t dt + D_t dW_t \\ y(0) &= 0. \end{aligned}$$

The two equations (1.1) and (1.2) constitute the filtering system.

Let $\mathcal{B}_t = \sigma(y_s | s \leq t)$ be the σ -field generated by the observation up to time t .

We intend to establish equations giving the filtered state $E^{\mathcal{B}_t} x_t$ recursively, that is to establish the equations of the Kalman filter and this, by the innovation approach (see [1] in the finite dimensional case).

2. Martingales related to the state and observation—innovation process.

LEMMA 2.1. *There is a $(\hat{x}_t)_{t \in [0, T]}$ predictable process \hat{x} with values in H , the trajectories of which are almost everywhere in D so that:*

$$P(\hat{x}_t = E^{\mathcal{B}_t} x_t) = 1 \quad \forall t \in [0, T]$$

where \hat{x} satisfies:

$$(2.1.1) \quad E \int_0^T \|\hat{x}_t\|_D^2 dt < +\infty.$$

Proof. This property results from the corresponding property in the finite dimensional case (cf. e.g. [8, proposition 2.2, p. 55]) since D is separable and dense in the separable space H and since we have $x \in L_D^2([0, T] \times \Omega, dt \otimes dP)$. Then, (2.1.1) follows from the fact that we have, when applying the Banach–Steinhaus theorem to the family of linear bounded operators from D to H , $(A_t|_D)_{t \in [0, T]}$:

$$\sup \{ \|A_t|_D\|_{\mathcal{L}(D, H)} | t \in [0, T] \} < +\infty.$$

DEFINITION 2.2. The two processes Z_s and Z_0 with values respectively in H and K are defined by:

$$\begin{aligned} Z_s(t) &= \hat{x}_t - \int_0^t A_u \hat{x}_u du \\ \forall t \in [0, T] \\ Z_0(t) &= y_t - \int_0^t C_u \hat{x}_u du. \end{aligned}$$

Z_0 is called an *innovation process*.

Remark. These two processes are well defined by virtue of (2.1.1) and of the boundedness of the operator families $(A_t|_D)_{t \in [0, T]}$ and $(C_t|_D)_{t \in [0, T]}$. Moreover, we have:

$$(2.2.1) \quad E \int_0^T \|A_u \hat{x}_u\|_H^2 du < +\infty$$

$$(2.2.2) \quad E \int_0^T \|C_u \hat{x}_u\|_K^2 du < +\infty.$$

DEFINITION 2.3. We call *error* the process e with values in H defined by:

$$e_t = x_t - \hat{x}_t \quad \forall t \in [0, T].$$

Remark. The trajectories of e are almost everywhere in D ; the process e satisfies

$$(2.3.1) \quad E^{\mathcal{B}_u} e_t = 0 \quad \text{if } 0 \leq u \leq t \leq T$$

$$(2.3.2) \quad E \int_0^T \|e_t\|_D^2 dt < +\infty.$$

PROPOSITION 2.4. Z_s and Z_0 are square integrable martingales with values in H and K respectively, with respect to the process basis $(\Omega, \mathcal{F}, (\mathcal{B}_t)_{t \in [0, T]}, P)$.

Proof. It will be given only for Z_s , since it is the same for Z_0 . First, it is obvious that these processes are adapted to the family $(\mathcal{B}_t)_{t \in [0, T]}$. Further, by virtue of (2.1.1), (2.2.1) and (2.2.2), we have:

$$E\|Z_s(t)\|_H^2 < +\infty \quad \text{and} \quad E\|Z_0(t)\|_K^2 < +\infty \quad \forall t \in [0, T].$$

Let $u \leq t$. We have:

$$E^{\mathcal{B}_u}(Z_s(t) - Z_s(u)) = E^{\mathcal{B}_u}(\hat{x}_t - \hat{x}_u) - E^{\mathcal{B}_u} \int_u^t A_v \hat{x}_v dv.$$

Now

$$\begin{aligned} E^{\mathcal{B}_u}(\hat{x}_t - \hat{x}_u) &= E^{\mathcal{B}_u}(E^{\mathcal{B}_t} x_t - E^{\mathcal{B}_u} x_u) \\ &= E^{\mathcal{B}_u}(x_t - x_u) \\ &= E^{\mathcal{B}_u} \left[\int_u^t A_v x_v dv + \int_u^t B_v dW_v \right]. \end{aligned}$$

Since we have

$$E^{\mathcal{B}_u} \int_u^t B_v dW_v = E^{\mathcal{B}_u} \left[E^{\mathcal{F}_u} \int_u^t B_v dW_v \right] = 0$$

we have

$$E^{\mathcal{B}_u}(Z_s(t) - Z_s(u)) = E^{\mathcal{B}_u} \int_u^t A_v (x_v - \hat{x}_v) dv.$$

Since Ax and $A\hat{x}$ have $dt \otimes P$ -square integrable norms, we can apply the Fubini theorem; so

$$E^{\mathcal{B}_u} \int_u^t A_v (x_v - \hat{x}_v) dv = \int_u^t E^{\mathcal{B}_u} A_v (x_v - \hat{x}_v) dv.$$

But x and \hat{x} belong to $L_D^2([0, T] \times \Omega, dt \otimes P)$ and $A_{u|D}$ belongs to $\mathcal{L}(D, H)$, so we have

$$\begin{aligned} E^{\mathcal{B}_u} A_v (x_v - \hat{x}_v) &= E^{\mathcal{B}_u} [A_v (E^{\mathcal{B}_v} (x_v - \hat{x}_v))] \quad \text{if } u \leq v \\ &= 0. \end{aligned}$$

Finally

$$E^{\mathcal{B}_u}(Z_s(t) - Z_s(u)) = 0.$$

PROPOSITION 2.5. Z_0 is a martingale a.s. with continuous trajectories in K ; its natural process $\langle Z_0 \rangle$ is given by

$$(2.5.1) \quad \langle Z_0 \rangle_t = \int_0^t \widehat{D_u \mathcal{W} D_u^*} du \quad \forall t \in [0, T].$$

Remark. If $A \in \mathcal{L}_1(K)$, we write \hat{A} the element of the projective tensor product $K \otimes_1 K$ corresponding to it by the canonical isometry from $\mathcal{L}_1(K)$ to $K \otimes_1 K$. Conversely, if $u \in K \otimes_1 K$, we write \hat{u} the operator of $\mathcal{L}_1(K)$ associated with it.

Proof. The continuity of the trajectories of Z_0 results from those of y and from (2.2.2). Notice that Z_0 can be written

$$(2.5.2) \quad Z_0(t) = \int_0^t C_u e_u du + \int_0^t D_u dW_u.$$

So, by means of Ito's formula [9], we can write

$$(2.5.3) \quad \begin{aligned} Z_0(t) \otimes_K Z_0(t) &= \int_0^t Z_0(u) \otimes_K (C_u e_u) du + \int_0^t (C_u e_u) \otimes_K Z_0(u) du \\ &+ \int_0^t Z_0(u) \otimes_K D_u dW_u + \int_0^t (D_u dW_u) \otimes_K Z_0(u) \\ &+ \int_0^t D_u \otimes D_u (\hat{\mathcal{W}}) du. \end{aligned}$$

But, by virtue of (2.3.2) and of the uniform boundedness in $\mathcal{L}(D, K)$ of the family $\mathcal{C}_{|D}$, we may apply Fubini's theorem; so, for $u \leq t$

$$(2.5.4) \quad \begin{aligned} E^{\mathcal{B}_u} \int_0^t Z_0(v) \otimes_K (C_v e_v) dv &= \int_u^t E^{\mathcal{B}_u} [Z_0(v) \otimes_K (C_v E^{\mathcal{B}_v} e_v)] dv \\ &= 0. \end{aligned}$$

The latter equality holds by virtue of (2.3.1). Moreover, the stochastic integral $\int_0^t Z_0(u) \otimes_K D_u dW_u$ is well defined by application of the existence theorem by M. Metivier ([10, § 3, th. 2]). More precisely, let X be a process with values in $\mathcal{L}(L, K \otimes_2 K)$ defined by:

$$\forall (t, \omega) \in [0, T] \times \Omega \quad X(t, \omega) = Z_0(t, \omega) \otimes_K (D_{t,\cdot}).$$

Then, $X \in L^2_{\mathcal{L}(L, K \otimes_2 K)}([0, T] \times \Omega, \mathcal{P}, \text{tr } \mathcal{W} dt \otimes P)$ where \mathcal{P} is the predictable σ -field on $[0, T] \times \Omega$. In fact, we have

$$\|X(t, \cdot)\|_{\mathcal{L}(L, K \otimes_2 K)}^2 \leq \|Z_0(t)\|_K^2 \|D_t\|_{\mathcal{L}(L, K)}^2.$$

D , being bounded on $[0, T]$ by $\alpha > 0$, we have:

$$E \int_0^T \|X_t\|_{\mathcal{L}(L, K \otimes_2 K)}^2 \text{tr } \mathcal{W} dt \leq \text{tr } \mathcal{W} \alpha^2 \int_0^T E \|Z_0(t)\|_K^2 dt.$$

But as $\|Z_0\|_K^2$ is a sub-martingale, we obtain

$$E \int_0^T \|X_t\|_{\mathcal{L}(L, K \otimes_2 K)}^2 \text{tr } \mathcal{W} dt \leq \text{tr } \mathcal{W} \alpha^2 T E \|Z_0(T)\|_K^2 < +\infty.$$

Then the stochastic integral $\int_0^T X dW$ is well defined; it is written $\int_0^T Z_0(v) \otimes (D_v dW_v)$. Moreover, if $0 \leq s \leq t \leq T$, we have

$$(2.5.5) \quad E^{\mathcal{B}_u} \int_u^t Z_0(v) \otimes_K (D_v dW_v) = E^{\mathcal{B}_u} E^{\mathcal{F}_u} \int_u^t Z_0(v) \otimes_K (D_v dW_v).$$

Then it results from (2.5.3), (2.5.4) and (2.5.5) that

$$(2.5.6) \quad E^{\mathcal{B}_u}(Z_0(t) \otimes_K Z_0(t) - Z_0(u) \otimes_K Z_0(u)) = \int_u^t \widehat{D_v W D_v^*} dv.$$

This proves (2.5.1) since the right member of (2.5.6) is predictable. \square

We do not know—and it may perhaps be wrong (as in the finite dimensional case; see e.g. [1])—the equality of the σ -fields \mathcal{B}_t and the one generated by the innovation process Z_0 up to time t . However, by using an idea of Fujisaki [7] in the case of *nonlinear* filtering of one-dimensional systems, we establish a representation theorem for a square-integrable martingale as a stochastic integral with respect to Z_0 . The basic tool is a Girsanov theorem [12].

We shall prove the following fundamental result:

REPRESENTATION THEOREM 2.6. *Any separable centered square integrable martingale M on the process basis $(\Omega, \mathcal{F}, (\mathcal{B}_t)_{t \in [0, T]}, P)$ can be written $\int_0^\cdot \Phi dZ_0$, where $\Phi \in \bar{\Lambda}_T^2(Z_0; K, H)$.*

First, we must define $\bar{\Lambda}_T^2(Z_0; K, H)$.

DEFINITION 2.7. Let $\Lambda_T^2(Z_0; K, H)$ be the set of predictable processes X on the process basis $(\Omega, \mathcal{F}, (\mathcal{B}_t)_{t \in [0, T]}, P)$, with values in $\mathcal{L}(K, H)$ satisfying:

1) $\forall k \in K, Xk$ is strongly predictable with values in H .

2) $\|X\|_{\Lambda_T^2}^2 \equiv E \int_0^T \text{tr } X_t Q_t X_t^* dt < +\infty$.

$\Lambda_T^2(Z_0; K, H)$ is a pre-Hilbertian space for the semi-norm $\|\cdot\|_{\Lambda_T^2}$. Let Λ_T^{2*} be the Hilbert space obtained by completeness of $\Lambda_T^2(Z_0; K, H)$ with respect to this semi-norm. The space $\bar{\Lambda}_T^2(Z_0; K, H)$ is the closure in Λ_T^{2*} of the space \mathcal{E} of the uniformly predictable step-processes with values in $\mathcal{L}(K, H)$. The stochastic integral with respect to Z_0 of any $X \in \bar{\Lambda}_T^2(Z_0; K, H)$ is then defined by isometry [10].

Remark. Remember that in $\bar{\Lambda}_T^2(Z_0; K, H)$ there are processes with values unbounded linear operators from K to H [10].

Proof of Representation Theorem 2.6. The idea is to use a theorem of the Girsanov type [12], then a representation theorem for a martingale adapted to the family of σ -fields generated by a Hilbertian Gaussian martingale [13]. Here, the hypotheses (i) and (ii) on the observation noise are used in a crucial way.

Let φ be the K -valued process defined by:

$$\varphi_u = -Q_u^+ C_u \hat{x}_u \quad \forall u \in [0, T].$$

This process is well defined since \hat{x} takes its values in $D \subset \mathcal{D}(C_u) dt \otimes P$ —almost everywhere and as we have $\text{rg } (C_u) \subset \text{rg } Q_u \subset \mathcal{D}(Q_u^+)$. It is a (\mathcal{B}_t) -predictable process; indeed, $(C_u \hat{x}_u)_{u \in [0, T]}$ is (\mathcal{B}_t) -predictable; now, we have [13]

$$\forall h \in \text{rg } Q_u \quad Q_u^+ h = \lim_{n \rightarrow \infty} \int_0^n \exp[(s-n)Q_u^2] ds Q_u h$$

and, for $n \in \mathbb{N}$, the process $(\int_0^n \exp[(s-n)Q_u^2] ds Q_u C_u \hat{x}_u)_{u \in [0, T]}$ is (\mathcal{B}_t) -predictable; therefore φ is also predictable.

Elsewhere, for $u \in [0, T]$, $Q_u^+ C_u|_D \in \mathcal{L}(D, K)$. Indeed, by virtue of hypothesis (i), $\mathcal{D}(Q_u^+ C_u) \supset D$; so, it is sufficient to show that $Q_u^+ C_u|_D$ is a closed operator. To do so, let us consider a sequence $(x_n)_{n \in \mathbb{N}} \subset D$ such that

$$(2.6.1) \quad \lim_n x_n \stackrel{D}{=} x$$

$$\lim_n Q_u^+ C_u x_n = y.$$

By using the continuity of Q_u and properties of the pseudo-inverse, we may write

$$\lim_n Q_u Q_u^+ C_u x_n = \lim_n C_u x_n = Q_u y.$$

So, as C belongs to $\mathcal{L}(D, K)$ we have

$$C_u x = Q_u y.$$

But $C_u x$ belongs to $\mathcal{D}(Q_u^+)$, so

$$(2.6.2) \quad Q_u^+ C_u x = Q_u^+ Q_u y.$$

However, $Q_u^+ Q_u = \Pi_{\overline{\text{rg } Q_u}}$, the projection operator on $\overline{\text{rg } Q_u}$; as we have $Q_u^+ C_u x_n \in \overline{\text{rg } Q_u}$, y belongs to $\overline{\text{rg } Q_u}$; then from (2.6.2) we get

$$Q_u^+ C_u x = y.$$

Thus, we have proved that $Q_u^+ C_u|_D \in \mathcal{L}(D, K)$.

Elsewhere, by using the Banach–Steinhaus theorem and hypothesis (ii), we see that the set $\{\|Q_u^+ C_u\|_{\mathcal{L}(D, K)} | u \in [0, T]\}$ is bounded in K . Then it results from (2.1.1) that

$$(2.6.3) \quad E \int_0^T \|Q_u^{1/2} \varphi_u\|_K^2 du < +\infty.$$

Since the natural process of Z_0 is absolutely continuous with respect to the Lebesgue measure on $[0, T]$, we can show, as in the finite dimensional case, that $\Lambda_T^2(Z_0; K, \mathbb{R})$ is identical with the set of progressive processes Ψ with values in K verifying (2.6.3) (where Ψ_u takes the place of φ_u). Thus, the stochastic integral $\int_0^t \langle -Q_u^+ C_u \hat{x}_u, dZ_0(u) \rangle_K$ is well defined.

Let α be now the process with positive values defined by

$$(2.6.4) \quad \forall t \in [0, T] \quad \alpha_t = \exp \left[- \int_0^t \langle Q^+ C \hat{x}, dZ_0 \rangle_K - \frac{1}{2} \int_0^t \|Q_u^{1/2} Q_u^+ C_u \hat{x}_u\|_K^2 du \right].$$

For any $n \in \mathbb{N}$, we define the (\mathcal{B}_t) -stopping-time T_n by

$$(2.6.5) \quad T_n = \inf \left\{ t \in [0, T] \mid \int_0^t \|Q_u^{1/2} Q_u^+ C_u \hat{x}_u\|_K^2 du > n \right\}$$

where we set $\inf \emptyset = T$. By virtue of (2.6.3), the sequence $(T_n)_{n \in \mathbb{N}}$ converges almost surely towards T when n goes to infinity.

For any $n \in \mathbb{N}$ let us consider $\tilde{P}_{T_n} = \alpha_{T_n} P$, where α_{T_n} is the process α stopped in T_n . In view of [12], \tilde{P}_n is a probability measure on (Ω, \mathcal{F}) , and the process U^n , defined by

$$(2.6.6) \quad U_t^n = Z_0(t) + \int_0^{t \wedge T_n} Q_u Q_u^+ C_u \hat{x}_u du \quad \forall t \in [0, T]$$

is a continuous, centred, square integrable martingale with values in K with respect to the process basis $(\Omega, \mathcal{F}, (\mathcal{B}_t)_{t \in [0, T]}; \tilde{P}_n)$; its natural process is $\int_0^t Q_u du$. Moreover,

$$(2.6.7) \quad U_t^n = Z_0(t) + \int_0^{t \wedge T_n} C_u \hat{x}_u du \quad \forall t \in [0, T].$$

Let us consider now a martingale M verifying the properties mentioned in Theorem 2.6 and let \tilde{M} be the process defined by:

$$(2.6.8) \quad \tilde{M}_t = \alpha_t^{-1} M_t \quad \forall t \in [0, T].$$

It is a classical property that $\tilde{M}_{\cdot \wedge T_n}$ is a square integrable martingale with respect to the process basis $(\Omega, \mathcal{F}, (\mathcal{B}_{t \wedge T_n})_{t \in [0, T]}, \tilde{P}_n)$. Elsewhere

$$\begin{aligned} U_{t \wedge T_n}^n &= Z_0(t \wedge T_n) + \int_0^{t \wedge T_n} C_u \hat{x}_u du \\ &= y(t \wedge T_n) - \int_0^{t \wedge T_n} C_u \hat{x}_u du + \int_0^{t \wedge T_n} C_u \hat{x}_u du \end{aligned}$$

whence

$$(2.6.9) \quad U_{t \wedge T_n}^n = y(t \wedge T_n).$$

Since (Ω, \mathcal{F}) is a Blackwell space, we may thus write the following equalities of σ -fields:

$$(2.6.10) \quad \mathcal{B}_{t \wedge T_n} = \sigma(y(s \wedge T_n) | s \leq t) = \sigma(U_{s \wedge T_n}^n | s \leq t) = \mathcal{C}_{t \wedge T_n}^n$$

where the σ -field $\sigma(U_s^n | s \leq t)$ will be designate by \mathcal{C}_t^n . So, $\tilde{M}_{\cdot \wedge T_n}$ is a \tilde{P}_n -martingale adapted to the family of σ -fields $(\mathcal{C}_{t \wedge T_n}^n)_{t \in [0, T]}$ and consequently adapted to the family $(\mathcal{C}_t^n)_{t \in [0, T]}$. Then, the representation theorem for Hilbertian Gaussian martingales [13] assures the existence of a unique process Φ^n belonging to $\tilde{\Lambda}_T^2(U^n; K, \mathbb{R})$ such that

$$(2.6.11) \quad \tilde{M}_{t \wedge T_n} = \int_0^t \Phi^n dU^n \quad \forall t \in [0, T].$$

Then we can write, by using the localization properties of the stochastic integral and the formula (2.6.9),

$$(2.6.12) \quad \tilde{M}_{t \wedge T_n} = \int_0^{t \wedge T_n} \Phi^n dy$$

or

$$(2.6.13) \quad \tilde{M}_{t \wedge T_n} = \int_0^{t \wedge T_n} \Phi^n dZ_0 + \int_0^{t \wedge T_n} \Phi_u^n C_u \hat{x}_u du.$$

Elsewhere, by using Ito formula, we can prove the classical exponential formula

$$(2.6.14) \quad \alpha_{t \wedge T_n} = 1 + \int_0^{t \wedge T_n} \langle \alpha_u Q_u^+ C_u \hat{x}_u, dZ_0(u) \rangle_K,$$

and, from the equality (2.6.8), we may write

$$(2.6.15) \quad M_{t \wedge T_n} = \alpha_{t \wedge T_n} \tilde{M}_{t \wedge T_n}.$$

The equalities (2.6.13), (2.6.14) and (2.6.15) allow us to write, by use of the Ito's formula applied to the process $(L_t + V_t, \psi_t)_{t \in [0, T]}$ with values in $H \times \mathbb{R}$ and to the function f from $H \times \mathbb{R}$ to H defined by $f(h, a) = ah$ for any $(h, a) \in H \times \mathbb{R}$,

$$\begin{aligned} (2.6.16) \quad M_{t \wedge T_n} &= \psi_t(L_t + V_t) \\ &= \int_0^t (L_s + V_s) d\psi_s + \int_0^t \psi_s dV_s + \int_0^t \psi_s dL_s + \frac{1}{2} \int_0^t d\langle L, \psi \rangle_s \end{aligned}$$

where we have set

$$\begin{aligned} L_t &= \int_0^{t \wedge T_n} \Phi^n dZ_0 \\ V_t &= \int_0^{t \wedge T_n} \Phi_u^n C_u \hat{x}_u du \\ \psi_t &= \alpha_{t \wedge T_n}. \end{aligned}$$

From (2.6.14) and (2.6.16) it results that

$$(2.6.17) \quad M_{t \wedge T_n} = \int_0^t \Psi^n dZ_0 + \int_0^t \Theta_u^n du$$

where $\int_0 \Theta_u^n du$ is a process with bounded variation with values in H and $\Psi^n \in \bar{\Lambda}_T^2(Z_0; K, H)$. It results from (2.6.17) that $\int_0 \Theta_u^n du$ is a continuous martingale; by use of the separability of H , a classical result on the one-dimensional continuous martingale shows that $\int_0 \Theta_u^n du$ is a vanishing process; so

$$(2.6.18) \quad M_{t \wedge T_n} = \int_0^{t \wedge T_n} \Psi_u^n dZ_0(u).$$

Then, for $n \leq m$, we have

$$(2.6.19) \quad M_{t \wedge T_n} = M_{(t \wedge T_m) \wedge T_n} = \int_0^{(t \wedge T_m) \wedge T_n} \Psi^m dZ_0.$$

Combining (2.6.18) and (2.6.19), we obtain

$$(2.6.20) \quad E \int_0^{t \wedge T_n} \|\Psi_u^n - \Psi_u^m\| Q_u^{1/2} \|^2_2 du = 0.$$

Then, we define the process Ψ by

$$\Psi = \sum_{p=0}^{\infty} \Psi^p 1_{\|T_p, T_{p+1}\|}.$$

In view of (2.6.20), we have for any $n \leq m$

$$1_{\|0, T_n\|} \Psi^n Q^{1/2} = 1_{\|0, T_n\|} \Psi^m Q^{1/2}.$$

So,

$$\begin{aligned} 1_{\|0, T_n\|} \Psi Q^{1/2} &= \sum_{p=0}^{n-1} 1_{\|T_p, T_{p+1}\|} \Psi^p Q^{1/2} \\ &= \sum_{p=0}^{n-1} 1_{\|T_p, T_{p+1}\|} \Psi^n Q^{1/2}. \end{aligned}$$

Consequently,

$$(2.6.21) \quad 1_{\|0, T_n\|} \Psi Q^{1/2} = 1_{\|0, T_n\|} \Psi^n Q^{1/2}.$$

Then in view of (2.6.18) we have

$$E \|M_{t \wedge T_n}\|_H^2 = E \int_0^{t \wedge T_n} \|\Psi_u^n Q_u^{1/2}\|_2^2 du$$

¹If $S \leq T$ are two stopping times, we designate by $\|S, T\|$ the stochastic interval $\{(s, \omega) \in [0, T] \times \Omega \mid S(\omega) < s \leq T(\omega)\}$.

and by using (2.6.21)

$$(2.6.22) \quad E\|M_{t \wedge T_n}\|_H^2 = E \int_0^{t \wedge T_n} \|\Psi_u Q_u^{1/2}\|_2^2 du.$$

$\|M_{\cdot \wedge T_n}\|_H^2$ is a sub-martingale, so we have

$$E \int_0^{T \wedge T_n} \|\Psi_u Q_u^{1/2}\|_2^2 du \leq E\|M_T\|_H^2 < +\infty.$$

Then, by using the Beppo-Levi lemma, it follows that

$$E \int_0^T \|\Psi_u Q_u^{1/2}\|_2^2 du < +\infty.$$

By virtue of (2.6.18) and (2.6.21) we have

$$M_{t \wedge T_n} = \int_0^{t \wedge T_n} \Psi dZ_0 \quad \forall n \in \mathbb{N}.$$

The proof of Theorem 2.6 is complete.

COROLLARY 2.8. Z_s is a (\mathcal{B}_t) -martingale almost surely with continuous trajectories and there is a process $\Psi \in \bar{\Lambda}_T^2(Z_0; K, H)$ such that

$$(2.8.1) \quad Z_s(t) = \int_0^t \Psi dZ_0 \quad \forall t \in [0, T].$$

Proof. It follows directly from Proposition 2.4 and Theorem 2.6. \square

We shall show now that the covariance of the error e is deterministic; then we shall compute the process with bounded variation $\langle Z_0, Z_s \rangle$ associated with the two martingales Z_0 and Z_s . To do so, we establish two preliminary lemmas.

LEMMA 2.9. For any $t \in]0, T]$, let us call $\mathcal{B}_t^{(n)}$ the σ -field generated by the random variables

$$y_{t/2^n}, y_{2t/2^n}, \dots, y_{(2^n-1)t/2^n}, y_t.$$

Let us set

$$\hat{Q}_n(t) = E^{\mathcal{B}_t^{(n)}}[(x_t - E^{\mathcal{B}_t^{(n)}} x_t) \otimes_H (x_t - E^{\mathcal{B}_t^{(n)}} x_t)].$$

Then, for any $t \in]0, T]$, $\hat{Q}_n(t)$ is an almost surely constant random variable with values in $H \otimes_1 H$.

Proof. Since the process x may be written as

$$x_t = x_0 + \int_0^t G(t, s) B_s dW_s$$

it results from (1.2) that the random vector

$$(x_t, y_{t/2^n}, \dots, y_{(2^n-1)t/2^n}, y_t)$$

is a Gaussian vector with values in $H \times (K)^2$. Then, the property results from the same property for the finite dimensional Gaussian vectors (take a dense sequence in the dual space of $H \otimes_1 H$).

LEMMA 2.10. (a) The σ -field $\tilde{\mathcal{B}}_t$ generated by the increasing sequence of σ -fields $\mathcal{B}_t^{(n)}$ is equal to \mathcal{B}_t .

(b) For any $t \in [0, T]$ we have

$$\lim_n E^{\mathcal{B}_t^{(n)}}[x_t - E^{\mathcal{B}_t^{(n)}} x_t]^{\otimes_2 H} = E^{\mathcal{B}_t}[x_t - E^{\mathcal{B}_t} x_t]^{\otimes_2 H}.$$

(c) *There is a unique predictable process (with respect to the family $(\mathcal{B}_t)_{t \in [0, T]}$) with values in $\mathcal{L}_1(D, H)$ such that*

$$(2.10.1) \quad P[\hat{P}_t = E^{\mathcal{B}} e_t \otimes_{D, H} e_t] = 1 \quad \forall t \in [0, T].$$

This process is undistinguishable from a deterministic process.

Proof. (a) The announced property results from the almost-sure continuity of the trajectories of the process y .

(b) Since x takes its values in the separable Hilbertian space H , it follows from (a) and convergence theorems of strong Hilbertian martingales that

$$(2.10.2) \quad \lim_n E^{\mathcal{B}_t^{(n)}} x_t = E^{\mathcal{B}} x_t$$

the limit being an almost-sure strong limit in H . In the same way, $H \otimes_1 H$ being the separable dual space of a Banach space, we may write

$$(2.10.3) \quad \lim_n E^{\mathcal{B}_t^{(n)}} x_t^{\otimes_2 H} = E^{\mathcal{B}} x_t^{\otimes_2 H}$$

the limit being an almost-sure limit in $H \otimes_1 H$. Now, for any $n \in \mathbb{N}$

$$(2.10.4) \quad E^{\mathcal{B}_t^{(n)}} [x_t - E^{\mathcal{B}_t^{(n)}} x_t]^{\otimes_2 H} = E^{\mathcal{B}_t^{(n)}} x_t^{\otimes_2 H} - (E^{\mathcal{B}_t^{(n)}} x_t)^{\otimes_2 H}.$$

It follows that

$$\lim_n E^{\mathcal{B}_t^{(n)}} [x_t - E^{\mathcal{B}_t^{(n)}} x_t]^{\otimes_2 H} = E^{\mathcal{B}} x_t^{\otimes_2 H} - (E^{\mathcal{B}} x_t)^{\otimes_2 H} = E^{\mathcal{B}} [x_t - E^{\mathcal{B}} x_t]^{\otimes_2 H}$$

where the limits are taken in $H \otimes_1 H$.

(c) First, notice that we have

$$(2.10.5) \quad E \int_0^T \|e_t \otimes_{D, H} e_t\|_1 dt < +\infty.$$

Then, by using analogous results in the finite dimensional case ([8, Proposition 2.2, p. 55]) and the fact that $D \otimes_1 H$ is the separable dual space of a Banach space, we show that there exists a predictable process \hat{P} with values in $D \otimes_1 H$ such that

$$P[\hat{P}_t = E^{\mathcal{B}} e_t \otimes_{D, H} e_t] = 1 \quad \text{f.a.e. } t \in [0, T].$$

The process mentioned in Lemma 2.10 is then the process with values in $\mathcal{L}_1(D, H)$ associated with \hat{P} by the canonical isometry from $D \otimes_1 H$ onto $\mathcal{L}_1(D, H)$. It follows from Lemma 2.9 and from the assertion (b) of this lemma that the process P is a predictable version of the deterministic process $t \rightarrow E e_t \otimes_{D, H} e_t$. Since this process is measurable, Fubini's theorem makes sure that P is in fact undistinguishable from this process.

Remark. Condition (2.3.2) shows that, for almost every $t \in [0, T]$, we have $\text{rg } P_t \subset i(D)$, P -almost surely.

PROPOSITION 2.11. *The predictable process with bounded variation associated with the two martingales Z_0 and Z_s , with values in $K \otimes_1 H$ is written as*

$$(2.11.1) \quad \langle Z_0, Z_s \rangle_t = \int_0^t \widehat{(P_u C_u^* + B_u W D_u^*)} du$$

where P is the process introduced in Lemma 2.10(c) and where $C_u^* \in \mathcal{L}(K, D)$ is the adjoint operator of $C \in \mathcal{L}(D, K)$.

Proof. For $u < t$, we have

$$(2.11.2) \quad Z_s(t) = \int_0^t A_v e_v dv + \int_0^t B_v dW_v - e_t.$$

Now, from (2.3.1), we have

$$(2.11.3) \quad \begin{aligned} E^{\mathfrak{B}_u}[Z_0(t) \otimes_{K,H} Z_s(t) - Z_0(u) \otimes_{K,H} Z_s(u)] \\ = E^{\mathfrak{B}_u}[Z_0(t) \otimes_{K,H} \beta(t) - Z_0(u) \otimes_{K,H} \beta_u] \end{aligned}$$

where we have set

$$\beta_t = Z_s(t) + e_t.$$

Let us apply the Ito formula to the process (Z_0, β) with values in $K \times H$ and to the function F from $K \times H$ to $K \otimes_1 H$ defined by $F(k, h) = k \otimes h$ for any $(k, h) \in K \times H$; from (2.5.2) and (2.11.2) we get

$$(2.11.4) \quad \begin{aligned} Z_0(t) \otimes_{K,H} \beta_t &= \int_0^t Z_0(v) \otimes (A_v e_v) dv + \int_0^t Z_0(v) \otimes (B_v dW_v) \\ &+ \int_0^t (C_v e_v) \otimes \beta_v dv + \int_0^t (D_v dW_v) \otimes \beta_v \\ &+ \int_0^t D_v \otimes B_v d\langle W \rangle_v. \end{aligned}$$

Taking into account (2.3.1), (2.2.1) and (2.2.2), the linearity of A_v and B_v and the fact that Z_0 and Z_s are adapted to the family $(\mathfrak{B}_t)_{t \in [0, T]}$, we obtain, by using Fubini's theorem, that

$$(2.11.5) \quad \begin{aligned} E^{\mathfrak{B}_u}[Z_0(t) \otimes Z_s(t) - Z_0(u) \otimes Z_s(u)] \\ = E^{\mathfrak{B}_u} \left[\int_u^t (C_v e_v) \otimes \beta_v dv + \int_u^t D_v \otimes B_v (\hat{W}) dv \right]. \end{aligned}$$

Now, by virtue of the Fubini theorem

$$\begin{aligned} E^{\mathfrak{B}_u} \left[\int_u^t (C_v e_v) \otimes \beta_v dv \right] &= \int_u^t E^{\mathfrak{B}_u} E^{\mathfrak{B}_v} [(C_v e_v) \otimes \beta_v] dv \\ &= \int_u^t E^{\mathfrak{B}_u} [(C_v E^{\mathfrak{B}_v} e_v) \otimes Z_s(v) + E^{\mathfrak{B}_v} \{(C_v e_v) \otimes e_v\}] dv. \end{aligned}$$

So, we have from (2.3.1)

$$E^{\mathfrak{B}_u} \int_u^t (C_v e_v) \otimes \beta_v dv = \int_u^t E^{\mathfrak{B}_u} E^{\mathfrak{B}_v} [(C_v e_v) \otimes e_v] dv;$$

and, using the usual identifications of tensor products and linear operators, we have

$$E^{\mathfrak{B}_u} \int_u^t (C_v e_v) \otimes \beta_v dv = E^{\mathfrak{B}_u} \int_u^t \widehat{P_v C_v^*} dv.$$

In view of (2.11.5), we see that the process

$$(Z_0(t) \otimes Z_s(t) - \int_0^t \widehat{(P_v C_v^* + B_v \mathcal{W} D_v^*)} dv)_{t \in [0, T]}$$

is a (\mathcal{B}_t) -martingale; since the process

$$\left(\int_0^t \widehat{(P_v C_v^* + B_v \mathcal{W} D_v^*)} dv \right)_{t \in [0, T]}$$

is adapted to the family $(\mathcal{B}_t)_{t \in [0, T]}$ and continuous, then predictable, the proof is complete.

Remark. The martingales Z_0 and Z_s having almost all their trajectories continuous are still martingales with respect to the family $(\mathcal{B}_t^+)_{t \in [0, T]}$.

THEOREM 2.12. *Let us consider now the two martingales Z_0 and Z_s with respect to the family of σ -fields $(\mathcal{B}_t^+)_{t \in [0, T]}$; the process Ψ appearing in (2.8.1) for the representation of Z_s as a stochastic integral with respect to Z_0 is unique in $\bar{\Lambda}_T^2(Z_0; K, H)$; moreover, for almost every $u \in [0, T]$, Ψ_u is an extension to $\text{rg}(D_u \mathcal{W} D_u^*)^{1/2}$ of the operator $(P_u C_u^* + B_u \mathcal{W} C_u^*)(D_u \mathcal{W} D_u^*)^+$. We shall write*

$$(2.12.1) \quad Z_s(t) = \int_0^t (P_u C_u^* + B_u \mathcal{W} D_u^*)(D_u \mathcal{W} D_u^*)^+ dZ_0(u).$$

Proof. It follows immediately from the martingale representation [13], from Corollary 2.8 and Propositions 2.5 and 2.11.

COROLLARY 2.13. *There exists a continuous version of the process \hat{x} with values in H ; it is the unique solution of the stochastic differential equation*

$$(2.13.1) \quad \hat{x}_t = \int_0^t A_u \hat{x}_u du + \int_0^t (P_u C_u^* + B_u \mathcal{W} C_u^*)(D_u \mathcal{W} D_u^*)^+ dZ_0(u).$$

Remark. We give the same meaning to the word “solution” as for equation (1.1).

Proof. It is a direct consequence of the Z_0 definition and of Theorem 2.12. Notice that the method used before assured the existence of a solution of (2.13.1), that is \hat{x} ; elsewhere, the unicity of the solution is guaranteed by the unicity of the deterministic Cauchy problem connected with the family $(A_t)_{t \in [0, T]}$ (cf. § 1).

PROPOSITION 2.14. *Let Π be the covariance function of the error e , with values in $\mathcal{L}_1(D, H)$, defined for almost every $t \in [0, T]$ by*

$$\hat{\Pi}_t = E e_t \otimes_{D, H} e_t.$$

Almost all trajectories of the process P coincide almost everywhere with the function Π ; they satisfy the Ricatti equation

$$(2.14.1) \quad \begin{aligned} \frac{dP_t}{dt} &= A_t P_t^* + P_t A_t^* + B_t \mathcal{W} B_t^* \\ &\quad - (P_t C_t^* + B_t \mathcal{W} D_t^*)(D_t \mathcal{W} D_t^*)^+ (P_t C_t^* + B_t \mathcal{W} D_t^*)^* \\ P_0 &= E \widehat{x_0 \otimes_{D, H} x_0} \in \mathcal{L}_1(D, H). \end{aligned}$$

Proof. By virtue of Lemma 2.10, almost all trajectories of P coincide with the function Π . So we shall show that Π satisfies equation (2.14.1). Let us first introduce the function \hat{R} with values in $H \otimes_1 H$ defined by

$$\begin{aligned} \hat{R}_t &= E(x_t - \hat{x}_t)^{\otimes_2 H} \\ &= E E^{\mathcal{B}_t} [x_t^{\otimes_2 H} - x_t \otimes_H \hat{x}_t - \hat{x}_t \otimes_H x_t + \hat{x}_t^{\otimes_2 H}] \\ &= E[E^{\mathcal{B}_t} x_t^{\otimes_2 H} - \hat{x}_t^{\otimes_2 H}]. \end{aligned}$$

So,

$$(2.14.2) \quad \hat{R}_t = E[x_t^{\otimes_2 H} - \hat{x}_t^{\otimes_2 H}].$$

Note that we have, in view of the classical identification of tensor products

$$(2.14.3) \quad \hat{R}_t = \widehat{\Pi_t i^*} = \widehat{i \Pi_t^*}.$$

Then, applying the Ito formula to the process x and to the function F from H to $H \otimes_1 H$ defined by

$$\forall h \in H \quad F(h) = h \otimes_H h \quad (\text{cf. [6 bis]}),$$

we have, if $s < t$

$$(2.14.4) \quad \begin{aligned} x_t^{\otimes_2 H} - x_s^{\otimes_2 H} &= \int_s^t (A_u x_u) \otimes_H x_u \, du + \int_s^t x_u \otimes_H (A_u x_u) \, du \\ &\quad + \int_s^t (A_u x_u) \otimes_H (B_u \, dW_u) \\ &\quad + \int_s^t (B_u \, dW_u) \otimes_H (A_u x_u) + \int_s^t B_u^{\otimes 2}(\mathcal{W}) \, du. \end{aligned}$$

We easily verify that the process $(Ax) \otimes (B \cdot)$ with values in $\mathcal{L}(H, H \otimes_2 H)$ belongs to $L^2_{\mathcal{X}(H, H \otimes_2 H)}([0, T] \times \Omega, \mathcal{P}, \text{tr } \mathcal{W} \, dt \otimes P)$; then the process $(\int_0^\cdot (A_u x_u) \otimes_H (B_u \, dW_u))$ is a martingale with values in $H \otimes_2 H$. So

$$(2.14.5) \quad E(x_t^{\otimes_2 H} - x_s^{\otimes_2 H}) = E \left[\int_s^t (A_u x_u) \otimes_H x_u \, du + \int_s^t x_u \otimes_H (A_u x_u) \, du \right] + \int_s^t \widehat{B_u \mathcal{W} B_u^*} \, du.$$

Elsewhere, let us recall that we have, by virtue of the definition of the process Z_s

$$(2.14.6) \quad \hat{x}_t = \int_0^t A_u \hat{x}_u \, du + Z_s(t)$$

and that Z_s is a (\mathcal{B}_t^+) -square integrable martingale. Then applying the Ito formula [6 bis] to the process \hat{x} and to the same function F , we obtain

$$(2.14.7) \quad \begin{aligned} \hat{x}_t^{\otimes_2 H} - \hat{x}_s^{\otimes_2 H} &= \int_s^t (A_u \hat{x}_u) \otimes_H \hat{x}_u \, du + \int_s^t \hat{x}_u \otimes_H (A_u \hat{x}_u) \, du \\ &\quad + \int_s^t (A_u \hat{x}_u) \otimes_H dZ_s(u) + \int_s^t dZ_s(u) \otimes_H (A_u \hat{x}_u) + \int_s^t d\langle Z_s \rangle_u. \end{aligned}$$

So,

$$(2.14.8) \quad E[\hat{x}_t^{\otimes_2 H} - \hat{x}_s^{\otimes_2 H}] = E \left[\int_s^t (A_u \hat{x}_u) \otimes_H \hat{x}_u \, du + \int_s^t \hat{x}_u \otimes_H (A_u \hat{x}_u) \, du \right] + E \int_s^t d\langle Z_s \rangle_u.$$

Using the classical identifications of tensor products, then remembering that x and \hat{x} are $dt \otimes P$ -almost everywhere with values in D and finally using the Fubini theorem—

applicable by virtue of (2.1.1) and (2.2.1)—we obtain

$$\begin{aligned}
 & E \left[\int_s^t (A_u x_u) \otimes_H x_u du - \int_s^t (A_u \hat{x}_u) \otimes_H \hat{x}_u du \right] \\
 &= E \left[\int_s^t \widehat{x_u \otimes_{D,H} x_u A_u^*} du - \int_s^t \widehat{\hat{x}_u \otimes_{D,H} \hat{x}_u A_u^*} du \right] \\
 (2.14.9) \quad &= \int_s^t E \left(\widehat{x_u \otimes_{D,H} x_u - \hat{x}_u \otimes_{D,H} \hat{x}_u} \right) A_u^* du \\
 &= \int_s^t \widehat{\Pi_u A_u^*} du.
 \end{aligned}$$

$$(2.14.10) \quad \langle Z_s \rangle_t = \int_0^t \widehat{(P_u C_u^* + B_u \mathcal{W} D_u^*)(D_u \mathcal{W} D_u^*)^+ (P_u C_u^* + B_u \mathcal{W} D_u^*)^*} du.$$

Combining the relations (2.14.2), (2.14.3), (2.14.5), (2.14.8), (2.14.9) and (2.14.10), and then using the fact that P is undistinguishable from the deterministic process Π , we obtain

$$\begin{aligned}
 (2.14.11) \quad (\Pi_t - \Pi_s) i^* &= \int_s^t [\Pi_u A_u^* + A_u \Pi_u^* + B_u \mathcal{W} B_u^* \\
 &\quad + (\Pi_u C_u^* + B_u \mathcal{W} D_u^*)(D_u \mathcal{W} D_u^*)^+ (\Pi_u C_u^* + B_u \mathcal{W} D_u^*)^*] du.
 \end{aligned}$$

By virtue of the hypotheses and the relation (2.10.4), the integrand appearing in (2.14.11) is integrable on $[0, T]$; this completes the proof.

LEMMA 2.15. *The natural process $\langle Z_s \rangle$ associated with the martingale Z_s is given by*

$$(2.15.1) \quad \langle Z_s \rangle = \int_0^t \widehat{(P_u C_u^* + B_u \mathcal{W} D_u^*)(D_u \mathcal{W} D_u^*)^+ (P_u C_u^* + B_u \mathcal{W} D_u^*)^*} du.$$

Proof. Remember that there exists a process (which is in fact deterministic) $\Psi \in \bar{\Lambda}_T^2(Z_0; K, H)$ such that the following relation holds:

$$(2.8.1) \quad Z_s(t) = \int_0^t \Psi dZ_0 \quad \forall t \in [0, T].$$

Elsewhere, from [13], there exists a sequence $(Q_n)_{n \in \mathbb{N}} \subset L_{\mathcal{L}_1(K)}^2([0, T])$ such that

$$(2.15.2) \quad \Phi Q_n \xrightarrow[n \rightarrow \infty]{\bar{\Lambda}_T^2(Z_0; K, H)} \Psi$$

where

$$(2.15.3) \quad \Phi_t = P_t C_t^* + B_t \mathcal{W} D_t^* \quad \forall t \in [0, T].$$

This may be written (which makes sense in view of the properties of the process Ψ proved in Theorem 2.12)

$$(2.15.4) \quad \Phi Q_n Q^{1/2} \xrightarrow[n \rightarrow \infty]{L_{\mathcal{L}_2(K, H)}^2([0, T])} \Psi Q^{1/2}.$$

Using the usual properties of nuclear and Hilbert–Schmidt operators, we obtain

$$(2.15.5) \quad (\Phi Q_n Q^{1/2})(\Phi Q_n Q^{1/2})^* \xrightarrow[n \rightarrow +\infty]{L^1_{\mathcal{L}_1(H)}([0, T])} (\Psi Q^{1/2})(\Psi Q^{1/2})^*.$$

We shall prove that, under hypotheses (ii) and (iii), we have

$$(2.15.6) \quad (\Psi Q^{1/2})(\Psi Q^{1/2})^* = \Phi Q^+ \Phi^* \quad \lambda\text{-a.e.}$$

where λ is the Lebesgue measure on $[0, T]$.

For some time we shall be working with a fixed $t \in [0, T]$, and so shall omit the index t (e.g. Q_t will be written Q) until stated otherwise. First notice that, for any k and k' belonging to $\mathcal{D}(Q^+)$, we have, since the operators Q_n and Q^+ are symmetrical,

$$(2.15.7) \quad |((Q_n - Q^+)Q(Q_n - Q^+)k, k')_K| \leq \|Q(Q_n - Q^+)k\|_K \|(Q_n - Q^+)k'\|_K.$$

But, by virtue of the properties of pseudo-inverses and of the approximating sequence $(Q_n)_{n \in \mathbb{N}}$ (cf. [13] for memory) we have

$$(2.15.8) \quad \lim_n (Q_n - Q^+)k = 0 \quad \forall k \in \mathcal{D}(Q^+).$$

Since $\mathcal{D}(Q^+)$ is dense in K , we have

$$(2.15.9) \quad \lim_n (Q_n - Q^+)Q(Q_n - Q^+)k = 0 \quad \text{weakly } \forall k \in \mathcal{D}(Q^+).$$

Taking into account hypotheses (ii) and (iii) we obtain

$$(2.15.10) \quad \lim_n (Q_n - Q^+)Q(Q_n - Q^+)\Phi^*h = 0 \quad \text{weakly } \forall h \in H.$$

In particular

$$(2.15.11) \quad \lim_n \langle (Q_n - Q^+)Q(Q_n - Q^+)\Phi^*h, \Phi^*h' \rangle_K = 0 \quad \forall h \in H, \quad \forall h' \in H.$$

This may be written

$$(2.15.12) \quad \lim_n \langle \Phi(Q_n - Q^+)Q(Q_n - Q^+)\Phi^*h, h' \rangle_H = 0 \quad \forall h \in H, \quad \forall h' \in H.$$

Henceforth we shall no longer work with a fixed t . In view of (2.15.5), at least for a sub-sequence, we have

$$(2.15.13) \quad \lim_n \Phi Q_n Q Q_n \Phi^* = (\Psi Q^{1/2})(\Psi Q^{1/2})^* \quad \lambda\text{-a.e.}$$

Since it follows from (2.15.8) that, for any $t \in [0, T]$,

$$(2.15.14) \quad \lim_n \langle \Phi_t(Q_n(t) - Q_t^+)Q_t Q_t^+ \Phi_t^*h, h' \rangle_H = 0 \quad \forall h \in H, \quad \forall h' \in H,$$

it follows from (2.15.12)

$$(2.15.15) \quad \lim_n \langle \Phi_t Q_n(t) Q_t Q_n(t) \Phi_t^*h, h' \rangle_H = \langle \Phi_t Q_t^+ Q_t Q_t^+ \Phi_t^*h, h' \rangle_H \quad \forall h \in H, \quad \forall h' \in H.$$

Taking into account (2.15.13), we obtain (2.15.6).

Then let Z_n be the $(\mathcal{B}_t)_{t \in [0, T]}$ -martingale with square-integrable norm with values in H defined by

$$(2.15.16) \quad Z_n(t) = \int_0^t \Phi Q_n dZ_0 \quad \forall t \in [0, T].$$

By virtue of [9], we have

$$(2.15.17) \quad \langle Z_n \rangle_t = \int_0^t (\Phi Q_n)^{\otimes 2} d\langle Z_0 \rangle.$$

So, by use of (2.5.1)

$$(2.15.18) \quad \langle Z_n \rangle_t = \int_0^t \widehat{\Phi Q_n Q Q_n \Phi^*} d\lambda.$$

Then, let $0 \leq u < t \leq T$; we have

$$(2.15.19) \quad E^{\mathcal{B}_u}(Z_n^{\otimes 2}(t) - Z_n^{\otimes 2}(u)) = \int_u^t \widehat{\Phi Q_n Q Q_n \Phi^*} d\lambda.$$

But the definition of the stochastic integral and the relations (2.15.2) and (2.15.16) allow us to write

$$(2.15.20) \quad \lim_n E^{\mathcal{B}_u}(Z_n^{\otimes 2}(t) - Z_n^{\otimes 2}(u)) = E^{\mathcal{B}_u}(Z_s^{\otimes 2}(t) - Z_s^{\otimes 2}(u)).$$

By virtue of (2.15.5), (2.15.6) and (2.15.19), we obtain

$$(2.15.21) \quad E^{\mathcal{B}_u}(Z_s^{\otimes 2}(t) - Z_s^{\otimes 2}(u)) = \int_u^t \Phi Q^+ \Phi^* d\lambda.$$

The proof is complete.

Remark. Proposition 2.14 shows the existence of a solution of equation (2.14.1); it is evident that the filtering problem will be solved completely if we make sure of the unicity of this solution. The following section will be devoted to investigating this unicity.

3. About the uniqueness of the solution of the Ricatti equation (2.14.1). Let us first determine what we mean exactly by a solution of (2.14.1); the announced uniqueness will then be a uniqueness within the class of these solutions.

DEFINITION 3.1. By a solution of (2.14.1) we understand any absolutely continuous and λ -integrable function from $[0, T]$ to $\mathcal{L}_1(D, H)$ satisfying the equation (2.14.1) for almost every $t \in [0, T]$ and such that

$$(3.1.1) \quad \begin{aligned} P_t i^* &= i P_t^* \\ P_t(D) &\subset i(D) \end{aligned} \quad \text{for almost every } t \in [0, T].$$

We shall examine the uniqueness of this solution by introducing a “dual” deterministic control problem associated with the studied filtering problem. By this duality, well known in the finite dimensional case, and studied by certain authors also in the infinite dimensional case (cf. e.g. [2] and [10]), we want to reduce the uniqueness problem for (2.14.1) to the uniqueness problem of the dual control problem. Sufficient conditions for the uniqueness of the control problem will yield sufficient conditions for the uniqueness of the solution of (2.14.1).

Later on, we make the following hypothesis:

$$(3.1.2) \quad Q_t > 0 \quad \forall t \in [0, T].$$

Notice that the pseudo-inverse Q_t^+ is then the unbounded inverse Q_t^{-1} of Q_t .

An easy computation shows that the equation (2.14.1) is equivalent to the

equation

$$(3.1.3) \quad \frac{dP_t i^*}{dt} = \mathcal{A}_t P_t^* + P_t \mathcal{A}_t^* + E_t - P_t C_t^* Q_t^{-1} C_t P_t^*$$

$$P(0) = P_0 \in \mathcal{L}_1(D, H)$$

where we have set

$$(3.1.4) \quad \begin{aligned} R_t &= B_t \mathcal{W} D_t^* \\ \mathcal{A}_t &= A_t - R_t Q_t^{-1} C_t \\ E_t &= B_t \mathcal{W} B_t^* - R_t Q_t^{-1} R_t^*. \end{aligned}$$

In the second section, we have seen that (3.1.3) has at least one solution.

Then, let us introduce the deterministic control problem defined by the two relations (3.1.5) and (3.1.6):

$$(3.1.5) \quad \frac{di' iz(T)}{dt} = -\mathcal{A}_t' iz(t) - C_t' u_t$$

$$iz(T) = a \in H;$$

$$(3.1.6) \quad J(u) = \langle iz(0), P_0 i^* iz(0) \rangle_H + \int_0^T [\langle iz(t), E_t iz(t) \rangle_H + \langle u_t, Q_t u_t \rangle_K] dt$$

$$\inf \{J(u) | u \in L_K^2([0, T])\}.$$

Remark. The “prime” sign denotes the dual operator whereas the “star” sign denotes the adjoint operator of the same operator.

We suppose that (3.1.5) has a unique solution in

$$W(0, T) = \left\{ f \in L_{D^*}^2([0, T]) \mid \frac{di' if}{dt} \in L_{D^*}^2([0, T]) \right\} \quad \text{for any } u \in L_K^2([0, T]).$$

We recall that, with a modification on a zero-measure set, it belongs to $C_H^0([0, T])$, the set of continuous functions from $[0, T]$ to H . So $iz(T) \in H$ is well defined.

Otherwise remark that, under hypothesis (3.1.2), the cost function J is strictly convex on $L_K^2([0, T])$. So, if there exists a solution of the control problem (3.1.5)–(3.1.6), it is unique.

PROPOSITION 3.2. *For any $u \in L_K^2([0, T])$, we have*

$$(3.2.1) \quad \begin{aligned} J(u) &= \langle iz(T), P_T i^* iz(T) \rangle_H \\ &+ \int_0^T \langle Q_t [Q_t^{-1} C_t P_t^* iz(t) + u_t], Q_t^{-1} C_t P_t^* iz(t) + u(t) \rangle_K dt \end{aligned}$$

where $z \in W(0, T)$ is the solution of (3.1.5) associated with the control $u \in L_K^2([0, T])$ and P is a solution of (3.1.3).

Proof. Let us define the function F on $[0, T]$ by

$$(3.2.2) \quad F(t) = \langle iz(t), i P_t^* iz(t) \rangle_H \quad \forall t \in [0, T].$$

F is differentiable almost everywhere, with its derivative being given, for almost every t , by

$$(3.2.3) \quad F'(t) = \left\langle \frac{di' iz(t)}{dt}, P_t^* iz(t) \right\rangle_{D^*, D} + \left\langle iz(t), \frac{dP_t}{dt} i^* iz(t) \right\rangle_H + \left\langle iz(t), P_t \frac{di^* iz(t)}{dt} \right\rangle_H.$$

In fact, for any $h \in \mathbb{R}^*$ such that $t+h \in [0, T]$, we have

$$(3.2.4) \quad \begin{aligned} \frac{F(t+h)-F(t)}{h} = & \left\langle \frac{i(z(t+h)-z(t))}{h}, iP_{t+h}^* iz(t+h) \right\rangle_H \\ & + \left\langle iz(t), \frac{(iP_{t+h}^* - iP_t^*)}{h} iz(t+h) \right\rangle_H \\ & + \left\langle iz(t), iP_t^* \left(\frac{iz(t+h)-iz(t)}{h} \right) \right\rangle_H. \end{aligned}$$

Let us study separately the three terms of the right member of (3.2.4). We have

$$(3.2.5) \quad \left\langle i \frac{(z(t+h)-z(t))}{h}, iP_{t+h}^* iz(t+h) \right\rangle_H = \left\langle i' i \frac{(z(t+h)-z(t))}{h}, P_{t+h}^* iz(t+h) \right\rangle_{D', D}.$$

Now

$$(3.2.6a) \quad \lim_{h \rightarrow 0} i' i \frac{(z(t+h)-z(t))}{h} = \frac{di' iz(t)}{dt}$$

and

$$(3.2.6b) \quad \lim_{h \rightarrow 0} P_{t+h}^* (iz(t+h)) = P_t^* iz(t)$$

since P^* is a solution of (3.1.3) and so is absolutely continuous and integrable on $[0, T]$, and since iz is continuous on H . Thus

$$(3.2.7) \quad \lim_{h \rightarrow 0} \left\langle i \frac{(z(t+h)-z(t))}{h}, iP_{t+h}^* iz(t+h) \right\rangle_H = \left\langle \frac{di' iz(t)}{dt}, P_t^* iz(t) \right\rangle_{D', D}.$$

Let us study now the second term; by virtue of (3.1.1), we have

$$(3.2.8) \quad \lim_{h \rightarrow 0} \frac{iP_{t+h}^* - iP_t^*}{h} = \frac{dP_t^*}{dt} = \frac{dP_t^{i*}}{dt}.$$

Since iz belongs to $C_H^0([0, T])$, we get

$$(3.2.9) \quad \lim_{h \rightarrow 0} \left\langle iz(t), \frac{iP_{t+h}^* - iP_t^*}{h} iz(t+h) \right\rangle_H = \left\langle iz(t), \frac{dP_t^{i*}}{dt} iz(t) \right\rangle_H.$$

As for the third term, we have by virtue of (3.1.1)

$$(3.2.10) \quad \begin{aligned} \lim_{h \rightarrow 0} \left\langle iz(t), iP_t^* \left(\frac{iz(t+h)-iz(t)}{h} \right) \right\rangle_H &= \lim_{h \rightarrow 0} \left\langle iz(t), P_t \frac{i^* iz(t+h) - i^* iz(t)}{h} \right\rangle_H \\ &= \left\langle iz(t), P_t \frac{di^* iz(t)}{dt} \right\rangle_H. \end{aligned}$$

Combining (3.2.7), (3.2.9) and (3.2.10), we get (3.2.3). Then, in view of (3.1.3) and (3.1.5), (3.2.3) allows us to write

$$(3.2.11) \quad \begin{aligned} F(0) = F(T) - \int_0^T [& \langle -\mathcal{A}_t' iz(t) - C_t' u_t, P_t^* iz(t) \rangle_{D', D} \\ & + \langle iz(t), \mathcal{A}_t P_t^* iz(t) + P_t \mathcal{A}_t^* iz(t) \rangle_H + \langle iz(t), E_t iz(t) \rangle_H \\ & - \langle iz(t), P_t C_t^* Q_t^{-1} C_t P_t^* iz(t) \rangle_H + \langle iz(t), -P_t \mathcal{A}_t^* iz(t) - P_t C_t^* u_t \rangle_H] dt. \end{aligned}$$

Therefrom we obtain

$$(3.2.12) \quad F(0) = F(T) - \int_0^T [\langle iz(t), E_t iz(t) \rangle_H - 2\langle u_t, C_t P_t^* iz(t) \rangle_K - \langle iz(t), P_t C_t^* Q_t^{-1} C_t P_t^* iz(t) \rangle_H] dt.$$

Putting (3.2.12) into (3.2.6), we get

$$(3.2.13) \quad J(u) = \langle iz(T), P_T^* iz(T) \rangle_H + \int_0^T [\langle u_t, Q_t u_t \rangle_K + 2\langle u_t, C_t P_t^* iz(t) \rangle_K + \langle iz(t), P_t C_t^* Q_t^{-1} C_t P_t^* iz(t) \rangle_H] dt.$$

Taking into account hypothesis (i), the relation

$$(3.2.14) \quad Q_t Q_t^{-1} y = y \quad \forall y \in \mathcal{D}(Q_t^{-1})$$

and the fact that Q_t is self-adjoint, we see that the relation (3.2.13) yields immediately (3.2.1).

PROPOSITION 3.3. *Let P be a solution of (3.1.3). If there exists a unique solution $z_{\text{op}}^{P,a}$ belonging to $W(0, T)$ of the homogeneous equation*

$$(3.3.1) \quad \frac{d}{dt} i z(t) = (-\mathcal{A}_t' + C_t' Q_t^{-1} C_t P_t^*) i z(t) \\ i z(T) = a \in i(D)$$

and if we define the control $u_{\text{op}}^{P,a}$ by

$$(3.3.2) \quad u_{\text{op}}^{P,a}(t) = -Q_t^{-1} C_t P_t^* i z_{\text{op}}^{P,a}(t) \quad \forall t \in [0, T]$$

then the pair $(u_{\text{op}}^{P,a}, z_{\text{op}}^{P,a})$ is an optimal solution of the control problem (3.1.5)–(3.1.6). Moreover, if S is another solution of (3.1.3) such that there exists a unique solution $z_{\text{op}}^{S,a}$ of the equation (3.3.1') obtained by changing P into S in (3.3.1), we have

$$(3.3.3) \quad Q_t^{-1} C_t P_t^* i z_{\text{op}}^{P,a}(t) = Q_t^{-1} C_t S_t^* i z_{\text{op}}^{S,a}(t)$$

for almost every $t \in [0, T]$.

Proof. The first part of the proposition is a consequence of Proposition 3.2. The second part results directly from the uniqueness of the solution of the control problem (3.1.5)–(3.1.6).

PROPOSITION 3.4. *The hypotheses of Proposition 3.3 being satisfied, we suppose moreover that, for almost every $t \in [0, T]$, the set $\{i z_{\text{op}}^{P,a}(t) | a \in i(D)\}$ is dense in H ; then the solutions P and S of (3.1.3) considered in Proposition 3.3 satisfy*

$$(3.4.1) \quad C_t P_t^* = C_t S_t^* \quad \text{for almost every } t \in [0, T].$$

Proof. By virtue of (3.3.1), (3.3.1') and (3.3.3), we may write for any $a \in i(D)$

$$(3.4.2) \quad \frac{d}{dt} i(z_{\text{op}}^{P,a} - z_{\text{op}}^{S,a}(t)) = -\mathcal{A}_t' i(z_{\text{op}}^{P,a}(t) - z_{\text{op}}^{S,a}(t)) \\ i(z_{\text{op}}^{P,a}(T) - z_{\text{op}}^{S,a}(T)) = 0.$$

Taking into account the uniqueness of the solution of the equation (3.1.5)—with $u = 0$ —we have

$$(3.4.3) \quad z_{\text{op}}^{P,a}(t) = z_{\text{op}}^{S,a}(t) \quad \forall t \in [0, T].$$

So, using the relation (3.3.3), we get

$$(3.4.4) \quad Q_t^{-1} C_t (P_t^* - S_t^*) i z_{\text{op}}^{P,a}(t) = 0 \quad \text{for almost every } t \in [0, T].$$

Then, the relation (3.4.1) results from the density hypothesis on the set $\{i z_{\text{op}}^{P,a}(t) | a \in i(D)\}$. \square

We give now a sufficient condition bearing only on the operator families A, B, C, D , assuring the uniqueness of the solution of (2.14.1). In fact, this condition is a perturbation hypothesis and was suggested by [5].

PROPOSITION 3.5. *We suppose that a Green kernel $\{\mathcal{G}(t, s) \in \mathcal{L}(H) | 0 \leq s \leq t \leq T\}$ is associated with the operator family $(\mathcal{A}_t)_{t \in [0, T]}$ satisfying, besides its definition properties (see § 1) the following hypotheses:*

- (i) $\forall h \in H$, the mapping $(s, t) \mapsto \mathcal{G}(t, s)h$ is continuous from $\{(s, t) | 0 \leq s \leq t \leq T\}$ to H .
- (ii) $\forall s < t$, $\mathcal{G}(t, s)(H) \subset D$ and $\mathcal{G}(t, s) \in \mathcal{L}(H, D)$.
- (iii) $\sup_{0 \leq t \leq T} \int_0^t \|\mathcal{G}(t, s)\|_{\mathcal{L}(H, D)}^2 ds < +\infty$.
- (iv) For any $t \in [0, T]$ and $\alpha < t$:

$$\sup_{s \leq \alpha} \|\mathcal{A}_t \mathcal{G}(t, s)\|_{\mathcal{L}(H)} < +\infty.$$

Then, for any solution P of (3.1.3), the equation

$$(3.5.1) \quad \begin{aligned} \frac{diz_t}{dt} &= \mathcal{A}_t z_t - P_t C_t^* Q_t^{-1} C_t z_t \\ iz(0) &= b \in H \end{aligned}$$

has a unique solution $z^{P,b}$ belonging to $W(0, T)$.

Proof. For any $t \in [0, T]$, let us define

$$U_t = -P_t C_t^* Q_t^{-1} C_t \in \mathcal{L}(D, H)$$

and let z be the unique solution in $L_D^2([s, T]) \cap C_H^0([s, T])$ of the integral equation

$$(3.5.2) \quad z_t = \mathcal{G}(t, s)z_s + \int_s^t \mathcal{G}(t, \alpha) U_\alpha z_\alpha d\alpha.$$

In fact, this equation has a unique solution in $L_D^2([s, T])$ since, by virtue of (iii) and of the boundedness of the family $(U_t)_{t \in [0, T]}$, we have

$$\sup_{s \leq t \leq T} \int_s^t \|\mathcal{G}(t, \alpha) U_\alpha\|_{\mathcal{L}(D)}^2 d\alpha \leq \left(\sup_{s \leq t \leq T} \int_s^t \|\mathcal{G}(t, \alpha)\|_{\mathcal{L}(H, D)}^2 d\alpha \right) \left(\sup_\alpha \|U_\alpha\|_{\mathcal{L}(D, H)} \right)^2.$$

Moreover, its solution belongs to $C_H^0([s, T])$; in fact we have for any $u > 0$

$$(3.5.3) \quad \begin{aligned} & \left\| \int_s^{t+u} \mathcal{G}(t+u, \alpha) U_\alpha z_\alpha d\alpha - \int_s^t \mathcal{G}(t, \alpha) U_\alpha z_\alpha d\alpha \right\|_H \\ & \leq \int_s^t \|\mathcal{G}(t+u, \alpha) - \mathcal{G}(t, \alpha)\|_H U_\alpha z_\alpha d\alpha + \int_t^{t+u} \|\mathcal{G}(t+u, \alpha) U_\alpha z_\alpha\|_H d\alpha. \end{aligned}$$

But, for almost every α , it results from (3.1.1) that $U_\alpha z_\alpha \in D$, so for almost every α

$$(3.5.4) \quad \lim_{u \rightarrow 0} (\mathcal{G}(t+u, \alpha) - \mathcal{G}(t, \alpha)) U_\alpha z_\alpha \stackrel{H}{=} 0.$$

Moreover, we have

$$(3.5.5) \quad \|(G(t+u, \alpha) - \mathcal{G}(t, \alpha))U_\alpha z_\alpha\|_H \leq 2 \left(\sup_{0 \leq \alpha \leq t \leq T} \|\mathcal{G}(t, \alpha)\|_{\mathcal{L}(H)} \right) \|U_\alpha z_\alpha\|_H.$$

By virtue of the integrability of Uz , we may apply the Lebesgue theorem and show that the first term of the right member of (3.5.3) tends to zero with u . This is also true for the second term, since we have

$$\|\mathcal{G}(t+u, \alpha)U_\alpha z_\alpha\|_H \leq \sup_{0 \leq \alpha \leq t \leq T} \|\mathcal{G}(t, \alpha)\|_{\mathcal{L}(H)} \|U_\alpha z_\alpha\|_H.$$

So, (3.5.2) has a unique solution in $L_D^2([s, T]) \cap C_H^0([s, T])$. Then we shall show that this solution is the solution of (3.5.1). In fact, by virtue of the hypotheses made on $\mathcal{G}(\cdot, \cdot)$, we have for $\alpha < t$

$$(3.5.6) \quad \frac{d}{dt} \mathcal{G}(t, \alpha)h = \mathcal{A}_t \mathcal{G}(t, \alpha)h \quad \forall h \in D.$$

Taking into account hypothesis (iv), the solution z of (3.5.1) satisfies, for almost every $t \in [0, T]$

$$\begin{aligned} \frac{dz_t}{dt} &= \frac{d}{dt} \mathcal{G}(t, s)z_s + \mathcal{G}(t, u)U_t z_t + \int_s^t \frac{d}{dt} \mathcal{G}(t, \alpha)U_\alpha z_\alpha d\alpha \\ &= \mathcal{A}_t \mathcal{G}(t, s)z_s + U_t z_t + \int_s^t \mathcal{A}_t \mathcal{G}(t, \alpha)U_\alpha z_\alpha d\alpha \\ &= \mathcal{A}_t [\mathcal{G}(t, s)z_s + \int_s^t \mathcal{G}(t, \alpha)U_\alpha z_\alpha d\alpha] + U_t z_t \\ &= (\mathcal{A}_t + U_t)z_t. \end{aligned}$$

So, z is a solution of (3.5.1).

Conversely, if z is a solution of (3.5.1) it is the solution of the integral equation (3.5.2). In fact, from the relation

$$(3.5.7) \quad \frac{\partial}{\partial s} \mathcal{G}(t, s)d = -\mathcal{G}(t, s)\mathcal{A}_s d \quad \forall d \in D, \quad 0 \leq s < t \leq T$$

which is true in view of the hypotheses made on $\mathcal{G}(\cdot, \cdot)$, we get

$$\begin{aligned} \frac{\partial}{\partial s} [\mathcal{G}(t, s)z_s] &= -\mathcal{G}(t, s)\mathcal{A}_s z_s + \mathcal{G}(t, s)(\mathcal{A}_s + U_s)z_s \\ &= \mathcal{G}(t, s)U_s z_s \end{aligned}$$

which yields the announced result, since the application $s \mapsto \mathcal{G}(t, s)U_s z_s$ is square integrable in D . The existence and uniqueness of the solution of (3.5.1) result from this.

PROPOSITION 3.6. *We suppose that, for any $a \in i(D)$, the equation (3.1.5) has a unique solution z^a in $W(0, T)$ and that the mapping $a \mapsto z^a$ is continuous from H to $W(0, T)$. Then a Green kernel $\{\mathcal{U}(t, s) \in \mathcal{L}(H) | 0 \leq s \leq t \leq T\}$ is associated with the operator family $(-\mathcal{A}'_t)_{t \in [0, T]}$, the solution of (3.1.5) for $u = 0$ satisfying*

$$(3.6.1) \quad iz(t) = \mathcal{U}(t, s)iz_s \quad \text{if } 0 \leq s \leq t \leq T.$$

We make the following hypotheses: There exists a real separable Hilbert space V such

that

$$D \subset V \subset H \subset V' \subset D'$$

with continuous and dense injections, and such that

(i) $\forall t \in [0, T]$, $C_t|_V \in \mathcal{L}(V, K)$ and the set $\{\|C_t\|_{\mathcal{L}(V, K)} | t \in [0, T]\}$ is bounded by $c > 0$.

(ii) For any $0 \leq s < t \leq T$, there exists an operator $\bar{\mathcal{U}}(t, s) \in \mathcal{L}(V', D)$ such that $\bar{\mathcal{U}}(t, s)|_H = \mathcal{U}(t, s)$, the family $\{\bar{\mathcal{U}}(t, s) | 0 \leq s < t \leq T\}$ satisfying

(a) for any $\alpha \in [0, T]$ and $k \in K$, the mapping $t \mapsto i' \bar{\mathcal{U}}(t, \alpha) C'_\alpha k$ is a differentiable mapping from $[\alpha, T]$ to D' satisfying

$$(3.6.2) \quad \frac{\partial}{\partial t} i' \bar{\mathcal{U}}(t, \alpha) C'_\alpha k = -\mathcal{A}'_t \bar{\mathcal{U}}(t, \alpha) C'_\alpha k,$$

$$(b) \quad \sup_{0 \leq t \leq T} \int_0^t \|\bar{\mathcal{U}}(t, \alpha)\|_{\mathcal{L}(V', D)}^2 d\alpha < +\infty.$$

Then, for any solution P of (3.1.3), the equation (3.3.1) has a unique solution $\varphi^{P,a}$ belonging to $W(0, T)$.

Proof. Let z be the function of $L_D^2([0, T])$ verifying for $0 \leq s \leq t \leq T$

$$(3.6.3) \quad iz(t) = \mathcal{U}(t, s)iz(s) + \int_s^t \bar{\mathcal{U}}(t, \alpha) C'_\alpha u_\alpha d\alpha.$$

The last integral makes sense, since we have tried by virtue of hypotheses (i) and (ii) (b)

$$(3.6.4) \quad \left(\int_s^t \|\bar{\mathcal{U}}(t, \alpha) C'_\alpha u_\alpha\|_D d\alpha \right)^2 \leq C \left(\int_0^t \|\bar{\mathcal{U}}(t, \alpha)\|_{\mathcal{L}(V', D)}^2 d\alpha \right) \left(\int_0^t \|u_\alpha\|_K^2 d\alpha \right).$$

Then, we have almost everywhere

$$\begin{aligned} \frac{di'iz_t}{dt} &= \frac{\partial}{\partial t} i' \mathcal{U}(t, s) iz_s + \int_s^t \frac{\partial}{\partial t} i' \bar{\mathcal{U}}(t, \alpha) C'_\alpha u_\alpha d\alpha + C'_t u_t \\ &= -\mathcal{A}'_t \mathcal{U}(t, s) iz_s + \int_s^t -\mathcal{A}'_t \bar{\mathcal{U}}(t, \alpha) C'_\alpha u_\alpha d\alpha + C'_t u_t \\ &= -\mathcal{A}'_t iz_t + C'_t u_t. \end{aligned}$$

The first equality is true by virtue of hypothesis (ii). This shows that z is the solution of (3.1.5) in $W(0, T)$. Then, if P is a solution of (3.1.3), for any $\Psi \in L_H^2([0, T])$, the equation

$$(3.6.5) \quad \frac{di'iz(t)}{dt} = -\mathcal{A}'_t iz_t + C'_t Q_t^{-1} C_t P_t^* \Psi_t$$

$$iz(T) = a \in H$$

has a unique solution $z \in W(0, T)$ satisfying, for $0 \leq s \leq t \leq T$

$$(3.6.6) \quad iz(t) = \mathcal{U}(t, s)iz(s) + \int_s^t \bar{\mathcal{U}}(t, \alpha) F_\alpha \Psi_\alpha d\alpha$$

where we have set

$$(3.6.7) \quad F_t = C'_t Q_t^{-1} C_t P_t^* \in \mathcal{L}(H, V').$$

Elsewhere, from hypotheses (i) and (ii)(b), the integral equation

$$(3.6.8) \quad \mathcal{Y}(t, s) = \tilde{\mathcal{U}}(t, s) + \int_s^t \tilde{\mathcal{U}}(t, \alpha) F_\alpha \mathcal{Y}(\alpha, s) d\alpha$$

has a unique solution in $L^2_{\mathcal{L}(V', D)}([s, T])$ for any $s \in [0, T]$. Then, since we have $\tilde{\mathcal{U}}(t, s)_H = \mathcal{U}(t, s)$, it results from (3.6.6) and (3.6.8) that we have for any $d \in D$, $\mathcal{Y}(\cdot, s)d \in L^2_D([0, T])$ and that this application is the unique solution of the equation (3.6.5) where Ψ is taken as the application defined by $\Psi_t = i\mathcal{Y}(t, s)d$. This shows that (3.3.1) admits $\mathcal{Y}(\cdot, s)d \in W(0, T)$ as a solution.

This solution is unique; in fact, if $z^j (j = 1, 2)$ are two solutions of (3.3.1), the mapping $u^j = -Q^{-1}.C.P^*.iz^j$ belongs to $L^2_K([0, T])$ and so, by the hypothesis, the equation

$$(3.6.9) \quad \begin{aligned} \frac{d i' i z(t)}{dt} &= -\mathcal{A}' i z(t) - C'_t u^j_t \\ i z(T) &= a \in H \end{aligned}$$

has a unique solution for $j = 1, 2$. But $z^j (j = 1, 2)$ satisfies (3.6.9), so we have $z^1 = z^2$.

PROPOSITION 3.7. *Under the hypotheses of Propositions 3.5 and 3.6, the conclusion of Proposition 3.4 is true.*

Proof. It is sufficient to show that the hypothesis appearing in Proposition 3.4 is satisfied. So, let P be a solution of the Riccati equation (3.1.3); the equations (3.5.1) and (3.3.1) evolving in the time interval $[s, T]$ may be written respectively:

$$(3.7.1) \quad \frac{d i z_t}{dt} = \mathcal{B}_t z_t \quad t \in [s, T],$$

$$i z(s) = b \in H$$

$$(3.7.2) \quad \frac{d i' i \varphi_t}{dt} = -\mathcal{B}'_t i \varphi_t \quad t \in [s, T],$$

$$i \varphi(T) = a \in H$$

where we have set

$$\mathcal{B}_t = \mathcal{A}_t - P_t C_t^* Q_t^{-1} C_t.$$

These equations have a unique solution z^b and φ^a respectively, for any a and b belonging to $i(D) \subset H$. Then, we have

$$(3.7.3) \quad \frac{d}{dt} \langle i \varphi^a(t), i z^b(t) \rangle_H = \left\langle \frac{d}{dt} i' i \varphi^a(t), z^b(t) \right\rangle_{D', D} + \left\langle i \varphi^a(t), \frac{d}{dt} i z^b(t) \right\rangle_H.$$

So

$$(3.7.4) \quad \begin{aligned} &\langle i \varphi^a(T), i z^b(T) \rangle_H - \langle i \varphi^a(s), i z^b(s) \rangle_H \\ &= \int_s^T \langle -\mathcal{B}'_t i \varphi^a(t), i z^b(t) \rangle_{D', D} dt + \int_s^T \langle i \varphi^a(t), \mathcal{B}_t z_t \rangle_H dt. \end{aligned}$$

We obtain

$$(3.7.5) \quad \langle b, i \varphi^a(s) \rangle_H = \langle a, i z^b(T) \rangle_H.$$

Then, let b be orthogonal to the set $\{i\varphi^a(s)|a \in i(D)\}$. It results from (3.7.5) that

$$\langle a, iz^b(T) \rangle_H = 0 \quad \forall a \in i(D),$$

that is, $i(D)$ being dense in H

$$iz^b(T) = 0.$$

It results from this that $b = 0$, which shows that the set $\{i\varphi^a(s)|a \in i(D)\}$ is dense in H ; the proof is complete.

Remark. If, for any $t \in [0, T]$, C_t is a bounded operator on H , we may take the space V equal to H . In this case, we obtain the uniqueness of the solution of the Ricatti equation in the sense of Proposition 3.4. We must remark that the introduced space V serves to measure the “unboundedness degree” of the operator C_t and that the hypotheses made in Proposition 3.6 will a priori be more difficult to realize as V will be “nearer” D .

Conclusion. We have shown how we can approach linear filtering problems in Hilbert spaces with unbounded observation and with a possibly infinite dimensional observation space by a Hilbertian martingales projection theory. An example of application to the smoothing problem for hereditary systems will be treated in another paper. Although we have given a sufficient uniqueness condition for the solution of the Ricatti equation, it seems nevertheless difficult to give a total answer to the uniqueness question. In the case of nonuniqueness, it would be interesting to find a characterization of the solutions of the Ricatti equation which, when substituted in the stochastic differential equation giving the filtered state, would give the same filter—could it be connected with the notion of minimal solution of a Ricatti equation introduced by L. Tartar?

Acknowledgment. It is a pleasure for me to thank Professor M. Metivier for the discussions we have had on the subject of this paper.

REFERENCES

- [1] A. V. BALAKRISHNAN, *A martingale approach to linear recursive state estimation*, this Journal, 10 (1972), pp. 754–766.
- [2] A. BENSOUSSAN, *Filtrage Optimal des Systèmes Linéaires*, Dunod, Paris, 1971.
- [3] R. F. CURTAIN, *Infinite dimensional filtering*, this Journal, 13 (1975), pp. 89–104.
- [4] R. F. CURTAIN AND P. FALB, *Stochastic differential equations in Hilbert space*, J. Differential Equations, 10 (1971), pp. 412–430.
- [5] R. F. CURTAIN AND A. J. PRITCHARD, *The infinite dimensional Ricatti equation*, J. Math. Anal. Appl., 47 (1974), pp. 43–57.
- [6] M. C. DELFOUR AND S. K. MITTER, *Controllability, observability and optimal feedback control of hereditary differential systems*, this Journal, 10 (1972), pp. 298–328.
- [6 bis] B. GRAVERAUX AND J. PELLAUMAIL, *Formule d'Ito pour les processus non continus à valeurs dans un espace de Banach*, Ann. Inst. H. Poincaré Sect. B, 10 (1974), no. 4, pp. 399–422.
- [7] M. FUJISAKI, G. KALLIANPUR AND H. KUNITA, *Stochastic differential equations for the non-linear filtering problem*, Osaka J. Math., 9 (1972), pp. 19–40.
- [8] J. MEMIN, *Sur quelques problèmes fondamentaux de la théorie du filtrage*, Thèse 3ème Cycle 1974, Université de Rennes, France.
- [9] M. METIVIER, *Intégrale stochastique par rapport à des processus à valeurs dans un espace de Banach réflexif*, Theor. Probability Appl., 19 (1974), pp. 577–606.
- [10] M. METIVIER AND G. PISTONE, *Une formule d'isométrie pour l'intégrale stochastique hilbertienne-et intégration de processus non uniformément prévisible*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 33 (1975), pp. 1–18.

- [11] S. K. MITTER AND R. VINTER, *Filtering for linear stochastic hereditary differential systems*, Internat. Symp. Control Theory Numerical Methods and Computer Systems Modelling, Institute de Recherche d'Informatique et d'Automatique, Rocquencourt, France, June 1974.
- [12] J. Y. OUVRARD, *Martingales locales et théorèmes de Girsanov dans les espaces de Hilbert réels séparables*, Ann. Inst. H. Poincaré Sect. B, 9 (1973), no. 4, pp. 351–368.
- [13] J. Y. OUVRARD, *Représentation de martingales vectorielles de carré intégrable à valeurs dans des espaces de Hilbert réels séparables*, Z. Wahrscheinlichkeitstheorie Verw. Gebiete, 33 (1975), pp. 195–208.
- [14] K. YOSIDA, *Functional Analysis*, Springer-Verlag, Berlin, 1965.

LINEAR FILTERING IN HILBERT SPACES. II: AN APPLICATION TO THE SMOOTHING THEORY FOR HEREDITARY SYSTEMS WITH OBSERVATION DELAYS*

JEAN-YVES OUVRARD†

Abstract. An example is given where the linear filtering theory in Hilbert spaces [8] is applied to solving a smoothing problem for hereditary systems with observation delays. As a particular case, we find again the new results from A. Bagchi (1976). This problem is transposed into a filtering problem in Hilbert spaces, using the formalism introduced by M. C. Delfour and S. K. Mitter (1975) in their investigations of hereditary systems. The filter equations then yield a system of equations—stochastic linear equations and Riccati integro-differential equations—which allows us to find the smoothed state.

Introduction. Recently A. Bagchi [1] has solved the smoothing problem for delay systems with observation delays by the innovation method. We shall show how we can find the equations yielding the smoothed state for hereditary systems with hereditary observation as a direct application of the filtering theory in Hilbert spaces [8]—we thus find again the results of [1] as a particular case. To do so, we transpose the initial problem into a filtering problem in Hilbert spaces by using the formalism introduced by M. C. Delfour and S. K. Mitter [5] in their investigations of hereditary systems. The results of the abstract theory [8] only need to be expressed in terms of the initial problem. Note that this functional approach of investigating the filtering problem of hereditary systems has already been used by R. F. Curtain [2] and S. K. Mitter and R. Vinter [7]; however these latter two papers exclude the delays in the observation.

1. Description of the system. We intend to establish the equations giving the smoothed state of the following hereditary system evolving during the time interval $[0, T]$:

$$(1.1) \quad dX(t) = A_0(t)X(t) dt + \sum_{i=1}^k A_i(t) \begin{cases} X(t-\theta_i) & \text{if } t-\theta_i \geq 0 \\ \alpha(t-\theta_i) & \text{if } t-\theta_i < 0 \end{cases} dt \\ + \int_{-b}^0 M(t, \theta) \begin{cases} X(t+\theta) & \text{if } t+\theta \geq 0 \\ \alpha(t+\theta) & \text{if } t+\theta < 0 \end{cases} d\theta dt + B_t dW_t$$

$$X(0) = \alpha(0);$$

$$(1.2) \quad dY(t) = C_0(t)X(t) dt + \sum_{i=1}^k C_i(t) \begin{cases} X(t-\theta_i) & \text{if } t-\theta_i \geq 0 \\ \alpha(t-\theta_i) & \text{if } t-\theta_i < 0 \end{cases} dt \\ + \int_{-b}^0 N(t, \theta) \begin{cases} X(t+\theta) & \text{if } t+\theta \geq 0 \\ \alpha(t+\theta) & \text{if } t+\theta < 0 \end{cases} d\theta dt + D_t dW_t$$

$$Y(0) = 0$$

where X is the state of the system with values in the finite-dimensional Euclidean space E , and Y is the observation with values in the finite-dimensional Euclidean space K . We suppose that the coefficients satisfy the following hypotheses:

$$A_i \in \mathcal{C}_{\mathcal{L}(E)}([0, T]) \quad \forall i = 0, 1, \dots, k$$

$$C_i \in \mathcal{C}_{\mathcal{L}(E, K)}([0, T]) \quad \forall i = 0, 1, \dots, k$$

$$0 = \theta_0 < \theta_1 < \dots < \theta_k = b.$$

* Received by the editors July 13, 1977.

† I.R.M.A. Université Scientifique et Médicale de Grenoble, B.P. 53, F-38041, Grenoble-Cedex-France.

The kernels M and N verify:

$M \in \mathcal{L}_{\mathcal{L}(E)}^\infty([0, T] \times [-b, 0])$ and $M(\cdot, \theta)$ is continuous with values in $\mathcal{L}(E)$ for any $\theta \in [-b, 0]$.

$N \in \mathcal{L}_{\mathcal{L}(E, K)}^\infty([0, T] \times [-b, 0])$ and $N(\cdot, \theta)$ is continuous with values in $\mathcal{L}(E, K)$ for any $\theta \in [-b, 0]$.

Otherwise, W is a Wiener process defined on a standard process basis $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in [0, T]}, P)$ with values in a separable real Hilbert space L . Its covariance operator is the nuclear operator $\mathcal{W} \in \mathcal{L}_1(L)$. We suppose that:

$$B \in \mathcal{C}_{\mathcal{L}(L, E)}([0, T])$$

and

$$D \in \mathcal{C}_{\mathcal{L}(L, K)}([0, T]).$$

We write $Q_t = D_t \mathcal{W} D_t^*$ for any $t \in [0, T]$. We make the following hypotheses on the noises (they are automatically satisfied if Q_t is invertible for any $t \in [0, T]$):

- $$\begin{aligned} \text{(i)} \quad & \left. \begin{array}{l} \text{rg } C_i(t) \subset \text{rg } Q_t \quad \forall i = 0, 1, \dots, k \\ \text{rg } N(t, \theta) \subset \text{rg } Q_t \quad \forall \theta \in [-b, 0] \end{array} \right\} \quad \forall t \in [0, T] \\ \text{(ii)} \quad & \text{rg } D_t \mathcal{W} B_t^* \subset \text{rg } Q_t \quad \forall t \in [0, T]. \end{aligned}$$

(This latter hypothesis is obviously satisfied if we suppose the independence of the noises on the state and on the observation).

(iii) $\forall e \in E, \quad \forall i = 0, 1, \dots, k$ the sets $\{Q_i^+ C_i(t)e | t \in [0, T]\}$ and $\{Q_i^+ N(t, \alpha)e | (t, \alpha) \in [0, T] \times [-b, 0]\}$ are bounded in K , where Q_i^+ designates the pseudo-inverse of the positive operator $Q_i \in \mathcal{L}(K)$.

Finally, the "thick" initial condition α is a Gaussian centered random variable with covariance P_0 , independent from the Wiener process W , with values in the separable Hilbert space D (cf. [5]) which we shall define again below. The differential stochastic equation (1.1) giving the state X has a unique solution (cf. [2], [3], [7]).

Following the theory developed in [2] and [7], we introduce an abstract evolution system the solution of which allows us to find the solution of (1.1). Thus the smoothing problem is naturally transformed into a filtering problem in this functional context.

DEFINITION 1.1. Let \mathcal{H} be the set of the applications from $[-b, 0]$ to E with square integrable norm. It is provided with the pre-Hilbertian semi-norm defined by

$$\forall h \in \mathcal{H} \quad \|h\|_{\mathcal{H}}^2 = \|h(0)\|_E^2 + \int_{-b}^0 \|h(\theta)\|_E^2 d\theta.$$

The quotient space H of \mathcal{H} by this semi-norm is a Hilbert space which is isomorphic to $E \times L_E^2([-b, 0])$ (cf. [5]).

DEFINITION 1.2. Let D be the set of the elements h of \mathcal{H} being absolutely continuous on $[-b, 0]$, with a derivative belonging to $\mathcal{L}_E^2([-b, 0])$. We provide D with the Hilbertian norm defined by

$$\forall h \in D \quad \|h\|_D^2 = \sum_{i=0}^k \|h(-\theta_i)\|_E^2 + \int_{-b}^0 \left[\|h(\theta)\|_E^2 + \left\| \frac{d}{d\theta} h(\theta) \right\|_E^2 \right] d\theta.$$

(Remark that this norm is equivalent to the one introduced in [2] and [7], in the case where b is finite.)

The space D provided with this norm is then isomorphic to the space $E \times H_E^{1,2}([-b, 0])$, where $H_E^{1,2}([-b, 0])$ is the Sobolev space of the functions belonging to $L_E^2([-b, 0])$ with a first derivative in $L_E^2([-b, 0])$.

Let i be the mapping from D to H defined by

$$(h(0), h(\cdot)) \in D \xrightarrow{i} (h(0), h(\cdot)) \in H.$$

Any $h \in D$ such that $i(h) = 0$ is identically zero (use the absolute continuity of $h(\cdot)$). So the mapping i is an injection from D to H . Moreover, this injection is continuous; in view of properties of Sobolev spaces, $i(D)$ is dense and i is a compact injection.

Now, we shall define operator families A_\cdot , B_\cdot , C_\cdot .

DEFINITION 1.3 (operator families A_\cdot , B_\cdot , C_\cdot). For any $t \in [0, T]$, we define the linear operators A_t (from D to H), C_t (from D to K) and \check{B}_t (from L to H) by

$$(1.3.1) \quad \forall h \in D \quad A_t(h)(\theta) = \begin{cases} \sum_{i=0}^k A_i(t)h(-\theta_i) + \int_{-b}^0 M(t, \alpha)h(\alpha) d\alpha & \text{if } \theta = 0 \\ \frac{dh(\theta)}{d\theta} & \text{if } \theta \in [-b, 0[, \end{cases}$$

$$(1.3.2) \quad \forall h \in D \quad C_t(h) = \sum_{i=0}^k C_i(t)h(-\theta_i) + \int_{-b}^0 N(t, \alpha)h(\alpha) d\alpha,$$

$$(1.3.3) \quad \forall l \in L \quad \check{B}_t(l)(\theta) = \begin{cases} B_t(l) & \text{if } \theta = 0 \\ 0 & \text{if } \theta \in [-b, 0[. \end{cases}$$

LEMMA 1.4. (a) $\forall t \in [0, T]$, $A_t \in \mathcal{L}(D, H)$.

(b) $\forall h \in D$, $A_\cdot h$ is continuous from $[0, T]$ to H .

Proof. (a) For any $h \in D$, we have

$$\begin{aligned} \|A_\cdot h\|_H^2 &= \|A_t(h)(0)\|_E^2 + \int_{-b}^0 \|A_t(h)(\theta)\|_E^2 d\theta \\ &\leq 2 \left\| \sum_{i=0}^k A_i(t)h(-\theta_i) \right\|_E^2 \\ &\quad + 2 \int_{-b}^0 \|M(t, \alpha)h(\alpha) d\alpha\|_E^2 + \int_{-b}^0 \left\| \frac{dh}{d\theta}(\theta) \right\|_E^2 d\theta \end{aligned}$$

and in view of the fact that the operators A_i and the kernel M are uniformly bounded in t , there is a constant $K > 0$ such that

$$\begin{aligned} \|A_\cdot h\|_H^2 &\leq K \left[\sum_{i=0}^k \|h(-\theta_i)\|_E^2 + \int_{-b}^0 \left(\|h(\alpha)\|_E^2 + \left\| \frac{dh}{d\alpha}(\alpha) \right\|_E^2 \right) d\alpha \right] \\ &= K \|h\|_D^2. \end{aligned}$$

(b) For any $h \in D$ and $t, t' \in \mathbb{R}$, we have

$$\begin{aligned} \|A_\cdot h - A_{\cdot'} h\|_H^2 &= \left\| \sum_{i=0}^k (A_i(t) - A_i(t'))h(-\theta_i) \right\|_E^2 \\ &\quad + \int_{-b}^0 \|(M(t, \alpha) - M(t', \alpha))h(\alpha)\|_E^2 d\alpha \\ &\leq \sum_{i=0}^k \|A_i(t) - A_i(t')\|^2 \|h(-\theta_i)\|_E^2 \\ &\quad + \int_{-b}^0 \|M(t, \alpha) - M(t', \alpha)\|^2 \|h(\alpha)\|_E^2 d\alpha. \end{aligned}$$

If t' tends towards t , the first term of the right member of this inequality tends towards zero by virtue of the continuity of the operators $A_i(\cdot)$; this is true, too, for the second term by virtue of the Lebesgue theorem which we may use by virtue of the hypothesis made on M .

LEMMA 1.5. $\forall t \in [0, T]$, $C_t \in \mathcal{L}(D, K)$ and $\forall h \in D$, $C_t h$ is continuous from $[0, T]$ to K .

Proof. It is analogous to the preceding one.

LEMMA 1.6. We have for all $t \in [0, T]$:

$$(1.6.1) \quad \text{rg } C_t \subset \text{rg } Q_t$$

$$(1.6.2) \quad \text{rg } D_t \check{W} \check{B}_t^* \subset \text{rg } Q_t$$

$$(1.6.3) \quad \{\|Q_t^+ C_t h\| \mid t \in [0, T]\} \text{ is bounded in } K \text{ for any } h \in D.$$

Proof. (a) The relation (1.6.1) is a direct consequence of hypothesis (i) and of the definition of C_t .

(b) First, let us compute $\check{B}_t^* \in \mathcal{L}(H, L)$: for any $h \in H$ and $l \in L$ we have from the definition of \check{B}_t

$$\begin{aligned} \langle \check{B}_t l, h \rangle_H &= \langle B_t l, h(0) \rangle_E \\ &= \langle l, B_t^* h(0) \rangle_L, \end{aligned}$$

which shows that

$$(1.6.4) \quad \forall h \in H \quad \check{B}_t^* h = B_t^* h(0).$$

So, for any $h \in H$

$$(1.6.5) \quad D_t \check{W} \check{B}_t^* h = D_t \check{W} B_t^* h(0).$$

Then (1.6.2) results from hypothesis (ii).

(c) The property (1.6.3) is a direct consequence of hypothesis (iii). \square

Next we consider the following filtering problem in H :

$$(1.3) \quad \begin{aligned} dx_t &= A_t x_t dt + \check{B}_t dW_t \\ x_0 &= \alpha \end{aligned}$$

$$(1.4) \quad \begin{aligned} dy_t &= C_t x_t dt + D_t dW_t \\ y_0 &= 0. \end{aligned}$$

The equation (1.3) has a unique solution x almost all trajectories of which are continuous in H and with values almost everywhere in D ; this solution satisfies

$$(1.5) \quad E \int_0^T \|x_t\|_D^2 dt < +\infty$$

(cf. e.g. [6] for an existence and unicity theorem of the solution of (1.3)). Moreover (cf. e.g. [7]), $X = x_0$ is the unique solution of (1.1) and there is a version of x , still written x , such that

$$(1.6) \quad \forall t \in [0, T] \quad x_t(\theta) = \begin{cases} X(t+\theta) & \text{if } t+\theta \geq 0 \\ a(t+\theta) & \text{if } t+\theta < 0 \end{cases} \text{ almost surely.}$$

If \mathcal{B}_t is the σ -field generated by the observation y up to time t , the filtered state $\hat{x}_t = E^{\mathcal{B}_t} x_t$ will give the smoothed state

$$\hat{X}(t, \theta) = E^{\mathcal{B}_t} X(t+\theta) \quad \text{for any } \theta \in [-b, 0].$$

The filtering theory in Hilbert spaces ([8]–[9]) shows that \hat{x} is the solution of the differential stochastic equation

$$(1.7) \quad \hat{x}_t = \int_0^t A_u \hat{x}_u du + \int_0^t (P_u C_u^* + \check{B}_u \mathcal{W} D_u^*) (D_u \mathcal{W} D_u^*)^+ dZ_0(u)$$

where Z_0 is the innovation process defined by

$$(1.8) \quad Z_0(t) = y_t - \int_0^t C_u \hat{x}_u du.$$

The mapping $t \mapsto P_t \in \mathcal{L}_1(D, H)$ is a solution of the Ricatti equation

$$(1.9) \quad \begin{aligned} \frac{dP_t}{dt} &= A_t P_t^* + P_t A_t^* + \check{B}_t \mathcal{W} \check{B}_t^* - (P_t C_t^* + \check{B}_t \mathcal{W} D_t^*) (D_t \mathcal{W} D_t^*)^+ (P_t C_t^* + \check{B}_t \mathcal{W} D_t^*)^* \\ P_0 &= \widehat{E\alpha \otimes_{D,H} \alpha}^1 \end{aligned}$$

2. Solving the smoothing problem by using the filtering problem in Hilbert spaces.

LEMMA 2.1. *The innovation process Z_0 satisfies*

$$(2.1.1) \quad \begin{aligned} Z_0(t) = y_t - \sum_{i=0}^k \int_0^t C_i(u) &\begin{pmatrix} \hat{X}(u, -\theta_i) & \text{if } u - \theta_i \geq 0 \\ \hat{\alpha}(u, -\theta_i) & \text{if } u - \theta_i < 0 \end{pmatrix} du \\ &- \int_0^t \left(\int_{-b}^0 N(u, \theta) \begin{pmatrix} \hat{X}(u, \theta) & \text{if } u + \theta \geq 0 \\ \hat{\alpha}(u, \theta) & \text{if } u + \theta < 0 \end{pmatrix} d\theta \right) du \end{aligned}$$

where we write

$$\hat{\alpha}(t, \theta) = E^{\otimes} \alpha(t + \theta) \quad \text{if } -b < t + \theta < 0.$$

Proof. It is a direct consequence of the relation between x and X and of the definition of C_i . Note that the observations y and Y coincide.

Then it is sufficient to transcribe the filtering equations (1.7), (1.9) in order to establish the smoothing equations. To do so, we shall compute all the operators appearing in these equations.

LEMMA 2.2. *For any $t \in [0, T]$, the dual operator $A'_t \in \mathcal{L}(H, D')$ of A_t is given by:*

$$(2.2.1) \quad \forall h \in D \text{ such that } h(-b) = 0: \quad (A'_t h)(\theta) = \begin{cases} A_0^* h(0) + h(0) & \text{if } \theta = 0 \\ A_i^* h(0) & \text{if } \theta = -\theta_i, i = 1, \dots, k \\ M^*(t, \theta) h(0) - \frac{dh}{d\theta}(\theta) & \text{if } \theta \in [-b, 0[, \theta \neq -\theta_i. \end{cases}$$

Proof. First notice that the space $V = E^{k+1} \times L_E^2([-b, 0])$ is included in D' and that for any $d \in D$ and $v \in V$ we have

$$(2.2.2) \quad \langle d, v \rangle_{D, D'} = \sum_{i=0}^k \langle d(-\theta_i), v(-\theta_i) \rangle_E + \int_{-b}^0 \langle d(\theta), v(\theta) \rangle_E d\theta.$$

¹ If $u \in D \otimes_1 H$, the projective tensor product of D and H , we designate by \tilde{u} the nuclear operator isometrically associated with u .

Then let $h \in D$ such that $h(-b) = 0$ and $d \in D$; we have

$$\begin{aligned}
 \langle A_t d, h \rangle_H &= \left\langle \sum_{i=0}^k A_i(t) d(-\theta_i) + \int_{-b}^0 M(t, u) d(u) du, h(0) \right\rangle_E + \int_{-b}^0 \left\langle \frac{dd(u)}{du}, h(u) \right\rangle_E du \\
 (2.2.3) \quad &= \sum_{i=0}^k \langle d(-\theta_i), A_i^*(t) h(0) \rangle_E + \int_{-b}^0 \langle d(u), M^*(t, u) h(0) \rangle_E du \\
 &\quad + \int_{-b}^0 \left\langle d(u), -\frac{d}{du} h(u) \right\rangle_E du + \langle d(0), h(0) \rangle_E.
 \end{aligned}$$

So, if A'_t is defined by (2.2.1), the relations (2.2.2) and (2.2.3) show that $A'_t h$ belongs to D' and that $A'_{t|D}$, the restriction of A'_t to D provided with the topology of H , is continuous, since we have

$$\begin{aligned}
 \|A'_t h\|_{D'} &= \sup_{\|d\| \leq 1} |\langle d, A'_t h \rangle_{D, D'}| \\
 &= \sup_{\|d\| \leq 1} |\langle A_t d, h \rangle_H| \\
 &\leq \|A_t\|_{\mathcal{L}(D, H)} \|h\|_H.
 \end{aligned}$$

Then it is obvious that A'_t is the dual operator of A_t .

LEMMA 2.3. For any $t \in [0, T]$, the dual operator $C'_t \in \mathcal{L}(K, D')$ of C_t is given by:

$$\begin{aligned}
 &\forall k \in K \\
 (2.3.1) \quad (C'_t k)(\theta) &= \begin{cases} C_i^*(t)k & \text{if } \theta = -\theta_i, i = 0, 1, \dots, k, \\ N^*(t, \theta)k & \text{if } \theta \in [-b, 0], \theta \neq -\theta_i. \end{cases}
 \end{aligned}$$

Proof. For any $k \in K$ and $d \in D$, we have

$$\begin{aligned}
 \langle C_t d, k \rangle_K &= \sum_{i=0}^k \langle C_i(t) d(-\theta_i), k \rangle_K + \int_{-b}^0 \langle N(t, \theta) d(\theta), k \rangle_K d\theta \\
 &= \sum_{i=0}^k \langle d(-\theta_i), C_i^*(t) k \rangle_E + \int_{-b}^0 \langle d(\theta), N^*(t, \theta) k \rangle_E d\theta \\
 &= \langle d, C'_t k \rangle_{D, D'}
 \end{aligned}$$

where $C'_t k$ belongs to $V \subset D'$ and is defined by (2.3.1).

LEMMA 2.4. For any $t \in [0, T]$, the operator $\check{B}_t \mathcal{W} \check{B}_t^* \in \mathcal{L}_1(H)$ is given by:

$$(2.4.1) \quad \forall h \in H \quad [\check{B}_t \mathcal{W} \check{B}_t^* h](\theta) = \begin{cases} B_t \mathcal{W} B_t^* h(0) & \text{if } \theta = 0 \\ 0 & \text{if } \theta \in [-b, 0[. \end{cases}$$

The operator $\check{B}_t \mathcal{W} D_t^* \in \mathcal{L}(K, H)$ is given by:

$$(2.4.2) \quad \forall k \in K \quad [\check{B}_t \mathcal{W} D_t^* k](\theta) = \begin{cases} B_t \mathcal{W} D_t^* k & \text{if } \theta = 0 \\ 0 & \text{if } \theta \in [-b, 0[. \end{cases}$$

The operator $D_t \mathcal{W} \check{B}_t^* \in \mathcal{L}(H, K)$ is given by:

$$(2.4.3) \quad \forall h \in H \quad D_t \mathcal{W} \check{B}_t^* h = D_t \mathcal{W} B_t^* h(0).$$

Proof. It is a direct consequence of the definition of the operator \tilde{B}_t and of the relation (1.6.4) giving the expression of the operator \tilde{B}_t^* . \square

Now we introduce a kernel and study its properties; then we represent the nuclear operator P_t by means of this kernel.

DEFINITION 2.5. For any $t \in [0, T]$, we define the kernel $p(t, \cdot, \cdot)$ on $[-b, 0]^2$ with values in $\mathcal{L}(E)$ by

$$(2.5.1) \quad p(t, \theta_1, \theta_2) = E(x_t(\theta_1) - \hat{x}_t(\theta_1))(x_t(\theta_2) - \hat{x}_t(\theta_2))^*.$$

PROPOSITION 2.6. For almost every $t \in [0, T]$, we have for any $h \in H$ and any $\theta_1 \in [-b, 0]$

$$(2.6.1) \quad [P_t^* h](\theta_1) = p(t, \theta_1, 0)h(0) + \int_{-b}^0 p(t, \theta_1, \theta_2)h(\theta_2) d\theta_2.$$

Moreover, for almost every θ_2 and for $\theta_2 = 0$, $p(t, \cdot, \theta_2)$ is an absolutely continuous function with its derivative belonging to $\mathcal{L}^2([-b, 0])$. Likewise, for almost every θ_1 and for $\theta_1 = 0$, $p(t, \theta_1, \cdot)$ is an absolutely continuous function with its derivative belonging to $\mathcal{L}^2([-b, 0])$.

Remark. From the definition of p , we have: $p(t, \theta_2, \theta_1) = p^*(t, \theta_1, \theta_2)$. So, the properties of p will be the same in each variable θ_1 and θ_2 .

Proof. It is not difficult to prove that if X is an integrable random variable with values in D , then the application $\theta \rightarrow E[X(\theta)]$ belongs to D and that, for any $\theta \in [-b, 0]$, we have

$$[EX](\theta) \stackrel{E}{=} E[X(\theta)].$$

We recall (cf. [9]) that, for almost every $t \in [0, T]$, $P_t^* \in \mathcal{L}_1(H, D)$ is defined for any $h \in H$ by

$$(2.6.2) \quad P_t^* \stackrel{D}{=} E[(x_t - \hat{x}_t, h)_H (x_t - \hat{x}_t)].$$

So we may write for any $\theta_1 \in [-b, 0]$

$$(2.6.3) \quad \begin{aligned} [P_t^* h](\theta_1) &= E[(x_t - \hat{x}_t, h)_H (x_t(\theta_1) - \hat{x}_t(\theta_1))] \\ &= E[(x_t(\theta_1) - \hat{x}_t(\theta_1))(x_t(0) - \hat{x}_t(0))^* h(0)] \\ &\quad + E \int_{-b}^0 (x_t(\theta_1) - \hat{x}_t(\theta_1))(x_t(\theta_2) - \hat{x}_t(\theta_2))^* h(\theta_2) d\theta_2. \end{aligned}$$

Remembering that we have for almost every $t \in [0, T]$

$$(2.6.4) \quad E\|x_t - \hat{x}_t\|_D^2 < +\infty$$

we obtain in particular

$$(2.6.5) \quad E \int_{-b}^0 \|x_t(\theta) - \hat{x}_t(\theta)\|_E^2 d\theta < +\infty.$$

So, when applying the Fubini theorem in (2.6.3), we obtain (2.6.1). Otherwise, since P_t^* belongs to $\mathcal{L}_1^+(H, D)$, there are orthonormal bases $(h_n)_{n \in \mathbb{N}}$ and $(d_n)_{n \in \mathbb{N}}$ of H and D respectively and a sequence of positive numbers $(\lambda_n)_{n \in \mathbb{N}}$ verifying $\sum_{n \in \mathbb{N}} \lambda_n < +\infty$, such that

$$(2.6.6) \quad \forall h \in H \quad P_t^* h \stackrel{D}{=} \sum_{n=0}^{\infty} \lambda_n \langle h_n, h \rangle_H d_n.$$

Particularly, for any $h \in H$ and $\theta_1 \in [-b, 0]$, we have

$$(2.6.7) \quad [P_t^* h](\theta_1) \stackrel{E}{=} \sum_{n=0}^{\infty} \lambda_n \left[\langle h_n(0), h(0) \rangle_E + \int_{-b}^0 \langle h_n(\theta), h(\theta) \rangle_E d\theta \right] d_n(\theta_1).$$

But as the sequence $(\sum_{n=0}^P \lambda_n h_n)_{P \in \mathbb{N}}$ converges in H , there is a subsequence $(\sum_{n=0}^{P_k} \lambda_n h_n)_{k \in \mathbb{N}}$ which converges absolutely almost everywhere in $[-b, 0]$. Moreover, for this sub-sequence, taking into account the fact that d_n is absolutely continuous, we obtain when applying the Schwarz inequality

$$(2.6.8) \quad \left| \sum_{n=0}^{P_k} \lambda_n \langle h_n(\theta), h(\theta) \rangle_E d_n(\theta_1) \right| \leq \left(\sum_{n=0}^{\infty} \lambda_n \|h_n(\theta)\|_E \|h(\theta)\|_E \right) \sqrt{b}.$$

In view of the integrability of the right member and of the almost-everywhere convergence of the sequence $(\sum_{n=0}^{P_k} \lambda_n \langle h_n(\theta), h(\theta) \rangle_E d_n(\theta_1))_{k \in \mathbb{N}}$, we obtain by using the Lebesgue theorem

$$(2.6.9) \quad \begin{aligned} & \sum_{n=0}^{\infty} \lambda_n \int_{-b}^0 \langle h_n(\theta), h(\theta) \rangle_E d_n(\theta_1) d\theta \\ &= \int_{-b}^0 \lim_k \sum_{n=0}^{P_k} \lambda_n [d_n(\theta_1) h_n(\theta)^*] h(\theta) d\theta. \end{aligned}$$

Then it results from (2.6.1), (2.6.7) and (2.6.9) that

$$(2.6.10) \quad p(t, \theta_1, \theta_2) = \lim_k \sum_{n=0}^{P_k} \lambda_n [d_n(\theta_1) h_n(\theta_2)^*] \quad \text{for almost every } \theta_2 \in [-b, 0[$$

and

$$(2.6.11) \quad p(t, \theta_1, 0) = \sum_{n=0}^{\infty} \lambda_n [d_n(\theta_1) h_n(0)^*].$$

That proves the second part of the lemma.

PROPOSITION 2.7. *For any $t \in [0, T]$ and $h \in H$, we have for almost every θ_1 and θ_2 belonging to $]-b, 0[$*

$$(2.7.1) \quad \frac{\partial}{\partial \theta_1} \int_{-b}^0 p(t, \theta_1, \theta_2) h(\theta_2) d\theta_2 = \int_{-b}^0 \frac{\partial}{\partial \theta_1} p(t, \theta_1, \theta_2) h(\theta_2) d\theta_2$$

$$(2.7.2) \quad \frac{\partial}{\partial \theta_2} \int_{-b}^0 p(t, \theta_1, \theta_2) h(\theta_1) d\theta_1 = \int_{-b}^0 \frac{\partial}{\partial \theta_2} p(t, \theta_1, \theta_2) h(\theta_1) d\theta_1.$$

Proof. Note first that, in view of (2.6.10), $p(t, \cdot, \cdot)$ is a measurable mapping. Moreover, we have for almost every $\theta_2 \in]-b, 0[$ and for all $\theta_1 \in]-b, 0[$

$$(2.7.3) \quad |p(t, \theta_1, \theta_2)| \leq \sqrt{b} \sum_{n=0}^{\infty} \lambda_n |h_n(\theta_2)|.$$

Since the right member of (2.7.3) is integrable (see the proof of Proposition 2.6), $p(t, \cdot, \cdot)$ is likewise integrable. Otherwise, for all $e \in E$, the sequence $\sum_{n=0}^{P_k} \lambda_n d_n(\cdot) \langle h_n(\theta_2), e \rangle_E$ is a sequence of functions of D which converges in D , since we have, for any $n \in \mathbb{N}$

$$(2.7.4) \quad \|\lambda_n d_n(\cdot) \langle h_n(\theta_2), e \rangle_E\|_D \leq \lambda_n \|\langle h_n(\theta_2), e \rangle_E\|$$

since the right member of (2.7.4) is the general term of a convergent series for almost

every θ_2 . The limit of the studied function sequence is thus an absolutely continuous function and

$$(2.7.5) \quad \frac{\partial}{\partial \theta_1} \lim_k \sum_{n=0}^{P_k} \lambda_n d_n(\theta_1) \langle h_n(\theta_2), e \rangle_E = \lim_k \sum_{n=0}^{P_k} \lambda_n \frac{d}{d\theta_1} d_n(\theta_1) \langle h_n(\theta_2), e \rangle_E.$$

Then, (2.6.10) shows that $(\partial/\partial \theta_1)p(t, \cdot, \cdot)$ is a measurable and integrable mapping.

Then, let φ be a function belonging to $\mathcal{D}(-b, 0]$, the space of infinitely differentiable functions with their support in $]-b, 0]$. The definition of distributional derivatives and the double application of the Fubini theorem allows us to write

$$\begin{aligned} & \int_{-b}^0 \frac{\partial}{\partial \theta_1} \left(\int_{-b}^0 p(t, \theta_1, \theta_2) h(\theta_2) d\theta_2 \right) \varphi(\theta_1) d\theta_1 \\ &= - \int_{-b}^0 \left[\int_{-b}^0 p(t, \theta_1, \theta_2) h(\theta_2) d\theta_2 \right] \varphi'(\theta_1) d\theta_1 \\ &= - \int_{-b}^0 \left[\int_{-b}^0 p(t, \theta_1, \theta_2) \varphi'(\theta_1) d\theta_1 \right] h(\theta_2) d\theta_2 \\ &= \int_{-b}^0 \left[\int_{-b}^0 \frac{\partial}{\partial \theta_1} p(t, \theta_1, \theta_2) \varphi(\theta_1) d\theta_1 \right] h(\theta_2) d\theta_2 \\ &= \int_{-b}^0 \left[\int_{-b}^0 \frac{\partial}{\partial \theta_1} p(t, \theta_1, \theta_2) h(\theta_2) d\theta_2 \right] \varphi(\theta_1) d\theta_1. \end{aligned}$$

Since, in view of Proposition 2.6, the mapping $\theta_1 \mapsto \int_{-b}^0 p(t, \theta_1, \theta_2) h(\theta_2) d\theta_2$ is absolutely continuous, Proposition 2.7 is completely proved. \square

We shall now express the operators appearing in the Ricatti equation in terms of the kernel p .

LEMMA 2.8. *For almost every $t \in [0, T]$, we have for any $h \in H$ and $\theta \in [-b, 0]$*

$$(2.8.1) \quad [A_t P_t^* h](\theta) = \begin{cases} \sum_{i=0}^k A_i(t) p(t, -\theta_i, 0) h(0) + \sum_{i=0}^k A_i(t) \int_{-b}^0 p(t, -\theta_i, v) h(v) dv \\ \quad + \int_{-b}^0 M(t, u) p(t, u, 0) h(0) du \\ \quad + \int_{-b}^0 M(t, u) \left(\int_{-b}^0 p(t, u, v) h(v) dv \right) du & \text{if } \theta = 0 \\ \frac{\partial}{\partial \theta} p(t, \theta, 0) h(0) \\ \quad + \int_{-b}^0 \frac{\partial}{\partial \theta} p(t, \theta, v) h(v) dv & \text{if } \theta \in [-b, 0]. \end{cases}$$

Proof. In view of (1.3.1) we have

$$(2.8.2) \quad [A_t P_t^* h](\theta) = \begin{cases} \sum_{i=0}^k A_i(t) (P_t^* h)(-\theta_i) + \int_{-b}^0 M(t, v) (P_t^* h)(v) dv & \text{if } \theta = 0 \\ \frac{\partial}{\partial \theta} (P_t^* h)(\theta) & \text{if } \theta \in [-b, 0]. \end{cases}$$

The formulae (2.6.1) and (2.7.1) yield the announced result.

LEMMA 2.9. For almost every $t \in [0, T]$, we have for any $h \in D$ such that $h(-b) = 0$ and for any $\theta \in [-b, 0]$

$$(2.9.1) \quad \begin{aligned} [P_t A_t^* h](\theta) &= \sum_{i=0}^k p(t, \theta, -\theta_i) A_i^*(t) h(0) + \int_{-b}^0 p(t, \theta, v) M^*(t, v) h(0) dv \\ &\quad + \int_{-b}^0 \frac{\partial}{\partial v} p(t, \theta, v) h(v) dv. \end{aligned}$$

Proof. We recall that we have for any $d \in D$

$$(2.9.2) \quad P_t d \stackrel{H}{=} E[\langle x_t - \hat{x}_t, d \rangle_D (x_t - \hat{x}_t)].$$

Then we may write by virtue of (2.2.1) and (2.2.2)

$$(2.9.3) \quad \begin{aligned} P_t A_t^* h &\stackrel{H}{=} E[\langle x_t - \hat{x}_t, A_t' h \rangle_{D, D'} (x_t - \hat{x}_t)] \\ &= E \left\{ \left[\langle x_t(0) - \hat{x}_t(0), A_0^* h(0) + h(0) \rangle + \sum_{i=1}^k \langle x_t(-\theta_i) - \hat{x}_t(-\theta_i), A_i^* h(0) \rangle \right. \right. \\ &\quad \left. \left. + \int_{-b}^0 \left\langle x_t(\theta) - \hat{x}_t(\theta), M^*(t, \theta) h(0) - \frac{dh(\theta)}{d\theta} \right\rangle d\theta \right] (x_t - \hat{x}_t) \right\}. \end{aligned}$$

Then the formula (2.9.1) results from (2.9.3) by use of the definition and the properties of the kernel $p(t, \cdot, \cdot)$ and by the fact that we have chosen h such that

$$(2.9.4) \quad \int_{-b}^0 p(t, \theta, v) \frac{dh}{dv}(v) dv = p(t, \theta, 0) h(0) - \int_{-b}^0 \frac{\partial}{\partial v} p(t, \theta, v) h(v) dv.$$

LEMMA 2.10. For almost every $t \in [0, T]$, we have for any $k \in K$ and any $\theta \in [-b, 0]$

$$(2.10.1) \quad [P_t C_t^* k](\theta) = \sum_{i=0}^k p(t, \theta, -\theta_i) C_i^*(t) k + \int_{-b}^0 p(t, \theta, v) N^*(t, v) k dv.$$

Proof. By virtue of (2.2.2), (2.3.1) and (2.9.2) we have

$$(2.10.2) \quad \begin{aligned} P_t C_t^* k &\stackrel{H}{=} E[\langle x_t - \hat{x}_t, C_t' k \rangle_{D, D'} (x_t - \hat{x}_t)] \\ &= E \left\{ \left[\sum_{i=0}^k \langle x_t(-\theta_i) - \hat{x}_t(-\theta_i), C_i^*(t) k \rangle \right. \right. \\ &\quad \left. \left. + \int_{-b}^0 \langle x_t(\theta) - \hat{x}_t(\theta), N^*(t, \theta) k \rangle d\theta \right] (x_t - \hat{x}_t) \right\}. \end{aligned}$$

By using the definition and the properties of the kernel $p(t, \cdot, \cdot)$, we obtain the formula (2.10.1).

LEMMA 2.11. For almost every $t \in [0, T]$, we have for any $h \in H$

$$(2.11.1) \quad \begin{aligned} C_t P_t^* h &= \sum_{i=0}^k C_i(t) p(t, -\theta_i, 0) h(0) + \sum_{i=0}^k C_i(t) \int_{-b}^0 p(t, -\theta_i, v) h(v) dv \\ &\quad + \int_{-b}^0 N(t, u) p(t, u, 0) h(0) du + \int_{-b}^0 N(t, u) \left(\int_{-b}^0 p(t, u, v) h(v) dv \right) du. \end{aligned}$$

Proof. It is a direct consequence of (1.3.2) and (2.6.1).

LEMMA 2.12. *For almost every $t \in [0, T]$, we have for any $h \in H$ and any $\theta \in [-b, 0]$*

$$(2.12.1) \quad [iP_t^*h](\theta) = [Pi^*h](\theta) = p(t, \theta, 0)h(0) + \int_{-b}^0 p(t, \theta, v)h(v) dv.$$

Proof. It is a direct consequence of the definition and the properties of the kernel $p(t, \cdot, \cdot)$. \square

Before exploiting all these results, we shall compute the quadratic term of the Ricatti equation (1.9).

DEFINITION 2.13. For any $t \in [0, T]$ and $h \in H$, we define the vectors of H , $F_i(t)h$ ($i = 0, 1, \dots, 4$), by

$$(2.13.1) \quad F_0(t)h = (P_t C_t^* + \check{B}_t \mathcal{W} D_t^*) Q_t^+ (P_t C_t^* + \check{B}_t \mathcal{W} D_t^*)^* h$$

$$(2.13.2) \quad F_1(t)h = (P_t C_t^*) Q_t^+ (C_t P_t^*) h$$

$$(2.13.3) \quad F_2(t)h = (P_t C_t^*) Q_t^+ (D_t \mathcal{W} \check{B}_t^*) h$$

$$(2.13.4) \quad F_3(t)h = (\check{B}_t \mathcal{W} D_t^*) Q_t^+ (C_t P_t^*) h$$

$$(2.13.5) \quad F_4(t)h = (\check{B}_t \mathcal{W} D_t^*) Q_t^+ (D_t \mathcal{W} \check{B}_t^*) h.$$

Remark. We have

$$(2.13.6) \quad F_0(t)h = \sum_{i=1}^4 F_i(t)h.$$

LEMMA 2.14. *For almost every $t \in [0, T]$, we have for any $h \in H$ and $\theta \in [-b, 0]$*

$$(2.14.1) \quad \begin{aligned} [F_1(t)h] = & \int_{-b}^0 \left[\sum_{i,j=0}^k p(t, \theta, -\theta_j) C_j^*(t) Q_t^+ C_i(t) p(t, -\theta_i, u) \right] h(u) du \\ & + \int_{-b}^0 \left[\sum_{j=0}^k p(t, \theta, -\theta_j) C_j^*(t) Q_t^+ \int_{-b}^0 N(t, u) p(t, u, v) du \right] h(v) dv \\ & + \int_{-b}^0 \left[\left\{ \int_{-b}^0 p(t, \theta, v) N^*(t, v) dv \right\} Q_t^+ \sum_{i=0}^k C_i(t) p(t, -\theta_i, \alpha) \right] h(\alpha) d\alpha \\ & + \int_{-b}^0 \left[\int_{-b}^0 \int_{-b}^0 p(t, \theta, v) N^*(t, v) Q_t^+ N(t, u) p(t, u, \alpha) du dv \right] h(\alpha) d\alpha. \end{aligned}$$

Proof. It is a direct application of Lemmas 2.10 and 2.11.

LEMMA 2.15. *For almost any $t \in [0, T]$, we have for any $h \in H$ and $\theta \in [-b, 0]$*

$$(2.15.1) \quad \begin{aligned} [F_2(t)h](\theta) = & \sum_{i=0}^k p(t, \theta, -\theta_i) C_i^*(t) D_t \mathcal{W} B_t^* h(0) \\ & + \int_{-b}^0 p(t, \theta, v) N^*(t, v) Q_t^+ D_t \mathcal{W} B_t^* h(0) dv. \end{aligned}$$

Proof. It is an immediate consequence of (2.10.1) and (2.4.3).

LEMMA 2.16. *For almost every $t \in [0, T]$, we have for any $h \in H$*

$$(2.16.1) \quad [F_3(t)h](\theta) = 0 \quad \text{if } \theta \in [-b, 0]$$

$$\begin{aligned}
 [F_3(t)h](\theta) = & (B_t \mathcal{W} D_t^*) Q_t^+ \left[\sum_{i=0}^k C_i(t) p(t, -\theta_i, 0) h(0) + \int_{-b}^0 N(t, u) p(t, u, 0) h(0) du \right] \\
 (2.16.2) \quad & + (B_t \mathcal{W} D_t^*) Q_t^+ \left[\sum_{i=0}^k C_i(t) \int_{-b}^0 p(t, -\theta_i, v) h(v) dv \right. \\
 & \left. + \int_{-b}^0 N(t, u) \left(\int_{-b}^0 p(t, u, v) h(v) dv \right) du \right].
 \end{aligned}$$

Proof. It is an immediate consequence of (2.4.2) and (2.11.1).

LEMMA 2.17. For almost every $t \in [0, T]$, we have for any $h \in H$

$$(2.17.1) \quad [F_4(t)h](\theta) = \begin{cases} (B_t \mathcal{W} D_t^*) Q_t^+ (D_t \mathcal{W} B_t^*) h(0) & \text{if } \theta = 0 \\ 0 & \text{if } \theta \in [-b, 0[. \end{cases}$$

Proof. This results from (2.4.3) and (2.4.2).

3. Smoothing equations. We shall transcribe the filtering equations (1.7), (1.9) in H into a functional form and in this way obtain the smoothing equations for the state X , the solution of (1.1).

PROPOSITION 3.1. The smoothed state $\hat{X}(t, \theta)$, where $\theta \in [-b, 0]$ and $t \in [0, T]$, is the unique solution of the stochastic integro-partial differential equation (3.1.1) subject to the boundary condition that $\hat{X}(t, 0)$ is the unique solution of the equation (3.1.2) with the initial condition $\hat{X}(0, \theta) = 0$ for any $\theta \in [-b, 0]$.

$$\begin{aligned}
 (3.1.1) \quad \begin{pmatrix} \hat{X}(t, \theta) & \text{if } t + \theta \geq 0 \\ \hat{\alpha}(t, \theta) & \text{if } t + \theta < 0 \end{pmatrix} = & \int_0^t \frac{\partial}{\partial \theta} \begin{pmatrix} \hat{X}(u, \theta) & \text{if } u + \theta \geq 0 \\ \hat{\alpha}(u, \theta) & \text{if } u + \theta < 0 \end{pmatrix} du \\
 & + \int_0^t \sum_{i=0}^k p(u, \theta, -\theta_i) C_i^*(u) Q_u^+ dZ_0(u) \\
 & + \int_0^t \left(\int_{-b}^0 p(u, \theta, v) N^*(u, v) dv \right) Q_u^+ dZ_0(u) \\
 & \text{if } \theta \in [-b, 0[,
 \end{aligned}$$

$$\begin{aligned}
 (3.1.2) \quad \hat{X}(t, 0) = & \int_0^t \sum_{i=0}^k A_i(u) \begin{pmatrix} \hat{X}(u, -\theta_i) & \text{if } u + \theta_i \geq 0 \\ \hat{\alpha}(u, -\theta_i) & \text{if } u + \theta_i < 0 \end{pmatrix} du \\
 & + \int_0^t \left(\int_{-b}^0 M(u, \alpha) \begin{pmatrix} \hat{X}(u, \alpha) & \text{if } u + \alpha \geq 0 \\ \hat{\alpha}(u, \alpha) & \text{if } u + \alpha < 0 \end{pmatrix} d\alpha \right) du \\
 & + \int_0^t \sum_{i=0}^k p(u, 0, -\theta_i) C_i^*(u) Q_u^+ dZ_0(u) \\
 & + \int_0^t \left(\int_{-b}^0 p(u, 0, v) N^*(u, v) dv \right) Q_u^+ dZ_0(u) + \int_0^t B_u \mathcal{W} D_u^* Q_u^+ dZ_0(u).
 \end{aligned}$$

Remarks. 1) $\hat{X}(t, 0)$ is the filtered state of the system (1.1)–(1.2).

2) For solving the smoothing problem, we must first compute (for $t < -\theta$) an estimation $\hat{\alpha}$ of the thick initial condition (cf. (3.1.1)).

3) We recall that the innovation process Z_0 is given by (2.1.1); so, the stochastic integrals appearing in (3.1.1) and (3.1.2) can be in fact computed by means of Stieljes' integrals as we can see by using a stochastic formula of integration by parts.

Proof. We recall that

$$(3.1.3) \quad \hat{x}_t(\theta) = E^{\mathcal{B}_t}[x_t(\theta)] = \begin{cases} \hat{X}(t, \theta) & \text{if } t + \theta \geq 0 \\ \hat{\alpha}(t, \theta) & \text{if } t + \theta < 0. \end{cases}$$

The equations (3.1.1) and (3.1.2) are thus obtained directly from the equation (1.7) by use of Lemmas 2.4 and 2.10 and the formula (1.3.1).

PROPOSITION 3.2 (Ricatti equations). *The kernel $p(\cdot, \cdot, \cdot)$ is the solution of the following Ricatti system (3.2.1), (3.2.2), (3.2.3), (3.2.4):*

$$(3.2.1) \quad \begin{aligned} \frac{\partial}{\partial t} p(t, u, v) &= \frac{\partial}{\partial u} p(t, u, v) + \frac{\partial}{\partial v} p(t, u, v) \\ &- \sum_{i,j=0}^k p(t, u, -\theta_j) C_j^*(t) Q_i^+ C_i(t) p(t, -\theta_i, v) \\ &- \sum_{j=0}^k p(t, u, -\theta_j) C_j^*(t) Q_i^+ \left(\int_{-b}^0 N(t, \alpha) p(t, \alpha, v) d\alpha \right) \\ &- \sum_{i=0}^k \left(\int_{-b}^0 p(t, u, \alpha) N^*(t, \alpha) d\alpha \right) Q_i^+ C_i(t) p(t, -\theta_i, v) \\ &- \left(\int_{-b}^0 p(t, u, \alpha) N^*(t, \alpha) d\alpha \right) Q_i^+ \left(\int_{-b}^0 N(t, \alpha) p(t, \alpha, v) d\alpha \right) \end{aligned}$$

in the domain $[0, T] \times]0, b]^2$,

$$(3.2.2) \quad \begin{aligned} \frac{\partial}{\partial t} p(t, u, 0) &= \frac{\partial}{\partial u} p(t, u, 0) + \sum_{i=0}^k p(t, u, -\theta_i) A_i^*(t) + \int_{-b}^0 p(t, u, \alpha) M^*(t, \alpha) d\alpha \\ &- \sum_{i,j=0}^k p(t, u, -\theta_j) C_j^*(t) Q_i^+ C_i(t) p(t, -\theta_i, 0) \\ &- \sum_{i=0}^k p(t, u, -\theta_i) C_i^*(t) Q_i^+ (D_i \mathcal{W} B_i^*) \\ &- \sum_{i,j=0}^k p(t, u, -\theta_j) C_j^*(t) Q_i^+ \int_{-b}^0 N(t, \alpha) p(t, \alpha, 0) d\alpha \\ &- \left(\int_{-b}^0 p(t, u, \alpha) N^*(t, \alpha) d\alpha \right) Q_i^+ \left(\sum_{i=0}^k C_i(t) p(t, -\theta_i, 0) \right) \\ &- \left(\int_{-b}^0 p(t, u, \alpha) N^*(t, \alpha) d\alpha \right) Q_i^+ \left(\int_{-b}^0 N(t, \beta) p(t, \beta, 0) d\beta \right) \\ &- \left(\int_{-b}^0 p(t, u, \alpha) N^*(t, \alpha) d\alpha \right) Q_i^+ (D_i \mathcal{W} B_i^*) \end{aligned}$$

in the domain $[0, T] \times]0, b]$,

$$\begin{aligned}
 \frac{\partial p}{\partial t}(t, 0, v) = & \frac{\partial}{\partial v} p(t, 0, v) + \sum_{i=0}^k A_i(t) p(t, -\theta_i, v) + \int_{-b}^0 M(t, \alpha) p(t, \alpha, v) d\alpha \\
 & - \sum_{i,j=0}^k p(t, 0, -\theta_j) C_j^*(t) Q_t^+ C_i(t) p(t, -\theta_i, v) \\
 & - \left(\sum_{j=0}^k p(t, 0, -\theta_j) C_j^*(t) \right) Q_t^+ \left(\int_{-b}^0 N(t, \alpha) p(t, \alpha, v) d\alpha \right) \\
 (3.2.3) \quad & - \left(\int_{-b}^0 p(t, 0, \alpha) N^*(t, \alpha) d\alpha \right) Q_t^+ \left(\sum_{i=0}^k C_i(t) p(t, -\theta_i, v) \right) \\
 & - \left(\int_{-b}^0 p(t, 0, \alpha) N^*(t, \alpha) d\alpha \right) Q_t^+ \left(\int_{-b}^0 N(t, \alpha) p(t, \alpha, v) d\alpha \right) \\
 & - \sum_{i=0}^k (B_i \mathcal{W} D_i^*) Q_t^+ C_i(t) p(t, -\theta_i, v) \\
 & - (B_i \mathcal{W} D_i^*) Q_t^+ \left(\int_{-b}^0 N(t, \alpha) p(t, \alpha, v) d\alpha \right)
 \end{aligned}$$

in the domain $[0, T] \times]0, b]$,

$$\begin{aligned}
 \frac{\partial p}{\partial t}(t, 0, 0) = & \sum_{i=0}^k A_i(t) p(t, -\theta_i, 0) + \int_{-b}^0 M(t, \alpha) p(t, \alpha, 0) d\alpha \\
 & + \sum_{i=0}^k p(t, 0, -\theta_i) A_i^*(t) + \int_{-b}^0 p(t, 0, \alpha) M^*(t, \alpha) d\alpha + B_i \mathcal{W} B_i^* \\
 & - \sum_{i,j=0}^k p(t, 0, -\theta_j) C_j^*(t) Q_t^+ C_i(t) p(t, -\theta_i, 0) \\
 & - \sum_{j=0}^k p(t, 0, -\theta_j) C_j^*(t) Q_t^+ \left(\int_{-b}^0 N(t, \alpha) p(t, \alpha, 0) d\alpha \right) \\
 (3.2.4) \quad & - \left(\int_{-b}^0 p(t, 0, \alpha) N^*(t, \alpha) d\alpha \right) Q_t^+ \sum_{i=0}^k C_i(t) p(t, -\theta_i, 0) \\
 & - \left(\int_{-b}^0 p(t, 0, \alpha) N^*(t, \alpha) d\alpha \right) Q_t^+ \left(\int_{-b}^0 N(t, \alpha) p(t, \alpha, 0) d\alpha \right) \\
 & - \sum_{i=0}^k p(t, 0, -\theta_i) C_i^*(t) Q_t^+ (D_i \mathcal{W} B_i^*) \\
 & - \left(\int_{-b}^0 p(t, 0, \alpha) N^*(t, \alpha) d\alpha \right) Q_t^+ (D_i \mathcal{W} B_i^*) \\
 & - (B_i \mathcal{W} D_i^*) Q_t^+ (D_i \mathcal{W} B_i^*).
 \end{aligned}$$

The initial condition is given by

$$(3.2.5) \quad p(0, u, v) = E\alpha(u)\alpha(v^*) \quad \forall (u, v) \in [-b, 0]^2.$$

Remark. The kernel p is thus a solution of a very complicated system of integro-partial-differential equations; the equations (3.2.2), (3.2.3), (3.2.4) are boundary conditions.

Proof. These equations are only the translation of the abstract Ricatti equation (1.9) by use of Lemmas 2.8, 2.9, 2.12, 2.14, 2.15, 2.16, 2.17 and the formula (2.4.1).

To obtain the equations (3.2.1) and (3.2.3) it suffices that the abstract Ricatti equation (1.9) be satisfied for any $h \in H$ such that $h(0)=0$; to obtain the equations (3.2.2) and (3.2.4), we write that (1.9) is satisfied for any $h \in H$ such that $h(\theta)=0 \forall \theta \in [-b, 0[$. The initial condition (3.2.5) is obtained by expressing the initial condition P_0 of the abstract Ricatti equation by means of the kernel $p(0, \cdot, \cdot)$.

Conclusion. We find again the results of [1] as a particular case: lack of integrated delays ($N=0$ and $M=0$), nonrandom and null thick initial condition ($\alpha=0$). Moreover, contrary to [1], we do not suppose the independence of the noises on the observation and the state. Finally, we allow the noise on the observation to be degenerate; however in this case, we are not sure of the unicity of the solution of the Ricatti system—when this noise is nondegenerate we do have the unicity (cf. [9]). The method used here has the advantage over [1] of really obtaining the properties of absolute continuity with respect to each variable of the kernel $p(t, \cdot, \cdot)$; that is not demonstrated in [1]. Otherwise, remark that we find again the classical results of the filtering and smoothing theory by taking the coefficients $A_i(t) \equiv 0$ and $C_i(t) \equiv 0$ for any $t \in [0, T]$ and $i = 1, \dots, k$, and the initial condition $\alpha \in D$ such that $\alpha(0)$ is a Gaussian random variable in E and $\alpha(\theta)=0$ for any $\theta \in [-b, 0[$. Finally, let us recall that the filtering theory presented in [2] and [7] corresponds to the case $C_i(t) \equiv 0$ for any $t \in [0, T]$ and $i = 1, \dots, k$ and that the results are given only in the Hilbert context without translating them into terms of initial data, as we have done by establishing a kernel theorem (Proposition 2.6).

REFERENCES

- [1] A. BAGCHI, *A martingale approach to state estimation in delay-differential equations*, J. Math. Anal. Appl., 56 (1976), pp. 195–210.
- [2] R. F. CURTAIN AND A. KALMAN, *Filtering theory for affine hereditary differential equations*, Internat. Symp. Control theory, Numerical methods and computer systems modelling, Institut de Recherche d'Informatique et d'Automatique, Rocquencourt, France, June 1974.
- [3] R. F. CURTAIN, *Infinite dimensional estimation theory for linear systems*, Rep. 38, Control Theory Centre, Warwick, England, Aug. 1975.
- [4] ———, *The infinite dimensional Ricatti equation with applications to affine hereditary differential systems*, this Journal, 13 (1975), pp. 1130–1143.
- [5] M. C. DELFOUR AND S. K. MITTER, *Hereditary differential systems with constant delays*, J. Differential Equations, I. General case, 12 (1972), pp. 213–235; II. A class of affine systems and the adjoint problem, 18 (1975), pp. 18–28.
- [6] M. METIVIER AND G. PISTONE, *Une formule d'isométrie pour l'intégrale stochastique hilbertienne et équations d'évolution linéaires stochastiques*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 33 (1975), pp. 1–18.
- [7] S. K. MITTER AND R. VINTER, *Filtering for linear stochastic hereditary differential systems*, Internat. Symp. Control Theory, Numerical Methods and Computer systems modelling, Institut de Recherche d'Informatique et d'Automatique, Rocquencourt, France, June 1974.
- [8] J. Y. OUVRARD, *Projection of martingales and linear filtering in Hilbert spaces*, Proceedings of the European Congress of Statisticians, (Grenoble, Sept. 1976); Recent Developments in Statistics, North-Holland, Amsterdam, (1977), pp. 551–558.
- [9] ———, *Martingale projection and linear filtering in Hilbert spaces. I: The theory*, this Journal, 16 (1978), pp. 912–937.

ALTERNATIVE THEORETICAL FRAMEWORKS FOR FINITE HORIZON DISCRETE-TIME STOCHASTIC OPTIMAL CONTROL*

STEVEN E. SHREVE† AND DIMITRI P. BERTSEKAS‡

Abstract. Stochastic optimal control problems are usually analyzed under one of three types of assumptions: a) Countability assumptions on the underlying probability space—this eliminates all difficulties of measure theoretic nature; b) Semicontinuity assumptions under which the existence of optimal Borel measurable policies can be guaranteed; and c) Borel measurability assumptions under which the existence of p -optimal or p - ϵ -optimal Borel measurable policies can be guaranteed (Blackwell [3], Strauch [31]). In this paper we introduce a general theoretical framework based on outer integration which contains these three models as special cases. Within this framework all known results for finite horizon problems together with some new ones are proved and subsequently specialized. An important new feature of our specialization to the Borel measurable model is the introduction of universally measurable policies. We show that everywhere optimal or nearly optimal policies exist within this class and this enables us to dispense with the notion of p -optimality.

1. Introduction. Consider a stochastic optimal control problem with cost function

$$(1) \quad J = \sum_{k=0}^{N-1} g(x_k, u_k, w_k), \quad N: \text{positive integer or } +\infty,$$

subject to the system equation

$$(2) \quad x_{k+1} = f(x_k, u_k, w_k), \quad k = 0, 1, \dots,$$

where x_k , u_k are the state and control of the system and w_k is a random object with probability distribution parameterized by x_k and u_k . We wish to choose a policy, that is, a sequence of functions $\{\mu_k\}$ from the state space S to the control space C so that when $u_k = \mu_k(x_k)$ the expected value of J is minimized. (A precise definition of the problem will be postponed for later.)

The equation

$$(3) \quad J_{k+1}(x) = \inf_u E\{g(x, u, w) + J_k[f(x, u, w)] | x, u\}$$

with $J_0(x) \equiv 0$, and its limiting form

$$(4) \quad J^*(x) = \inf_u E\{g(x, u, w) + J^*[f(x, u, w)] | x, u\}$$

are the Dynamic Programming (DP for short) equations related to the problem above. In the case where w takes a single value and the problem is deterministic (more generally, where w can take a countable number of values), the functions in these equations exist in a well-defined mathematical sense and the theory of DP is well developed (see e.g. [1]). When w can take uncountably many values, acute difficulties arise from the need to impose a proper measure theoretic structure on the problem so that the expected value of the cost J of (1) and the expected values in (3) and (4) are well defined. A related difficulty stems from the need to balance the measurability restrictions on policies (necessary so that the expected cost corresponding to a policy is

* Received by the editors April 14, 1977, and in revised form December 23, 1977. This work was supported by the National Science Foundation under Grant No. ENG 74-19332.

† Department of Statistics, University of California, Berkeley, California 94720.

‡ Department of Electrical Engineering and Coordinated Science Lab., University of Illinois, Urbana, Illinois 61801.

well defined) against a desire to admit enough policies to consideration so as to be able to find one which selects at or near the infimum in (3).

The measurability questions described above have been dealt with by a number of authors under the assumption that all spaces underlying the problem are Borel spaces (Borel subsets of complete separable metric spaces). There have been two main approaches. In the first approach, semicontinuity and compactness assumptions are imposed on the data of the problem (see § 5). Under these assumptions, the functions J_k in (3) can be shown to be semicontinuous and there exists a Borel measurable function μ_k such that $u_k = \mu_k(x)$ selects at or near the infimum in (3) for *every* x . For results in this direction see Maitra [15], Schäl [24]–[25], and Freedman [9]. Much of the work in stochastic programming (see Olsen [18]–[20], Rockafellar and Wets [22], [23] and the references quoted therein) also utilizes assumptions of this type. Some of this work employs additional convexity assumptions and is geared toward convex programming type results, i.e., duality and Kuhn–Tucker conditions for optimality, rather than resolution of the measure theoretic questions.

The second approach was introduced by Blackwell [3] and further refined by Strauch [31], Dynkin and Juskevic [8], Hinderer [12] and others. No assumptions other than Borel measurability of the data of the problem are made, and admissible policies are required to be Borel measurable. Under these conditions it is possible to prove the universal measurability of the optimal cost function and the existence for every $\varepsilon > 0$ and probability measure p on S of a p - ε -optimal policy (Strauch [31, Thms 7.1 and 8.1]). A p - ε -optimal policy is one which leads to a cost which differs from the optimal cost by less than ε for p -almost every initial state. Even over a finite horizon the optimal cost function need not be Borel measurable, and there need not exist an everywhere ε -optimal policy (Blackwell [3, Example 2]). The difficulty arises from the inability to choose a Borel measurable function $\mu_k: S \rightarrow C$ which nearly achieves the infimum in (3) uniformly in x . The nonexistence of such a function interferes with the construction of optimal policies via the DP algorithm (3), since one must first determine, at each stage k , a measure p_k with respect to which it is satisfactory to nearly achieve the infimum in (3) for p_k -almost every x . The difficulties in constructing nearly optimal policies over an infinite horizon are more acute. Furthermore, from an applications point of view, a p - ε -optimal policy, even if it can be constructed, is a much less appealing object than an everywhere ε -optimal policy, since in many situations the distribution p is unknown or may change when the system is operated repetitively, in which case a new p - ε -optimal policy must be computed.

In view of the undesirable features of p -optimality, Blackwell, Freedman and Orkin [4] have considered analytically measurable policies—a class that properly contains Borel measurable policies (see § 6). Their work deals with a special type of problem, that of minimization when the cost per stage is nonpositive. They show that a history remembering policy which is *everywhere* ε -optimal exists, and if the optimal cost functions J_k^* , $k = 1, \dots, N$, are everywhere finite, this policy can be taken to be Markov. We relax the assumption of a nonpositive cost per stage and show the existence in Corollary 5.1 of an analytically measurable ε -optimal Markov policy under the assumption that the functions J_k^* , $k = 1, \dots, N$, are everywhere finite. However we have been unable to show the strongest possible existence results for finite horizon problems within an analytically measurable policies framework (compare Corollaries 5.1 and 5.2). For this reason we have extended the class of admissible policies to include all universally measurable policies (a class properly containing the analytically measurable policies). A key fact here is that the composition of two universally measurable functions is universally measurable, while the

composition of two analytically measurable functions need not be analytically measurable [2]. In this paper we admit only Markov nonrandomized policies and, within this framework, cannot prove the result of Blackwell, Freedman and Orkin mentioned earlier for the case where $J_k^*(x)$ can be $-\infty$ for some x and k . It is shown elsewhere [2], [29], however, that under the assumption of Corollary 5.2(b), there exists for every $\varepsilon > 0$ a nonrandomized semi-Markov and a randomized Markov universally measurable ε -optimal policy. We do not know whether such a policy can be taken to be analytically measurable rather than universally measurable. The fact mentioned earlier relating to composition of two analytically measurable functions interferes with the constructions involved in the proofs of [29].

The present paper has two main objectives. The first is to provide a general framework for finite horizon stochastic optimal control that includes as special cases the formulations described earlier. The second is to demonstrate that when universally measurable policies are admitted in the Borel space framework of Blackwell, then all basic results for stochastic problems can be shown to hold in a form that is as strong as for problems where measurability questions are of no essential concern. In particular, the existence of *everywhere* ε -optimal policies is assured as opposed to policies which are ε -optimal p -almost everywhere. Thus the notion of p -optimality can be dispensed with.

The paper is organized as follows. Section 2 formulates a general stochastic optimal control problem without any topological assumptions. The formulation is based on a notion of outer integration developed in Appendix A. The main results regarding the validity of the DP algorithm and the existence of optimal and nearly optimal policies are provided in § 3. These include all results known for special cases together with a new result [Proposition 1(b)] relating to the existence of a sequence of policies exhibiting what is referred to as $\{\varepsilon_n\}$ dominated convergence to optimality. The results of § 3 are applied to special cases in § 4 (model without topological assumptions) and in § 5 (Borel space models with semicontinuity assumptions). Slight extensions of results by Freedman [9] are given in Corollaries 4.1 and 4.2. Section 6 is devoted to general Borel space models. We consider both analytically and universally measurable policies and prove an extended version of a measurable selection theorem by Brown and Purves [5]. Using this theorem we show that all the results of § 3 carry over to the Borel space model when universally measurable policies are allowed.

We note that some of the ideas and analysis in this paper (particularly the employment of universally measurable policies) have infinite horizon and imperfect state information counterparts described elsewhere [2], [28], [29]. Also, this paper considers exclusively nonrandomized Markov policies. Existence results relating to randomized and semi-Markov policies may be found in [2], [27], [29].

2. Problem formulation. Our notation will be as follows. For a set X we denote by F_X the set of all functions $J: X \rightarrow [-\infty, +\infty]$. For $J_1, J_2 \in F_X$ we write $J_1 = J_2$ if $J_1(x) = J_2(x) \forall x \in X$, and $J_1 \leq J_2$ if $J_1(x) \leq J_2(x) \forall x \in X$. If $J(x) \geq \varepsilon (J(x) \leq \varepsilon) \forall x \in X$, where ε is a scalar, we write $J \geq \varepsilon (J \leq \varepsilon)$. If a sequence $\{J_k\} \subset F_X$ increases (decreases) monotonically to $J \in F_X$, we write $J_k \uparrow J (J_k \downarrow J)$. If $\{J_k\}$ converges pointwise to J we write $J_k \rightarrow J$. If $J_1, J_2 \in F_X$, ε is a scalar, and $J_1(x) \leq J_2(x) + \varepsilon \forall x \in X$, we write $J_1 \leq J_2 + \varepsilon$. We adopt the usual conventions regarding ordering and arithmetic in the set of extended real numbers $[-\infty, +\infty]$, except that we take

$$-\infty + \infty = +\infty - \infty = +\infty.$$

The Cartesian product of sets A_1, A_2, \dots, A_n is denoted by $A_1 A_2 \dots A_n$. If X and Y

are sets, then proj_X is the projection mapping from XY to X . If E is a subset of some space X , we denote by χ_E the indicator function of E [$\chi_E(x) = 1$ if $x \in E$, $\chi_E(x) = 0$ if $x \notin E$]. For any function $f: X \rightarrow [-\infty, +\infty]$, where X is some space, we use the notation $f^+(x) = \max\{0, f(x)\}$, $f^-(x) = \max\{0, -f(x)\}$. The infimum over the empty set is taken by convention to be $+\infty$ ($\inf \emptyset = +\infty$).

The stochastic optimal control model we consider consists of the eight elements listed below:

S —*State space*. A nonempty set.

C —*Control space*. A nonempty set.

(W, \mathcal{F}) —*Disturbance space*. A measurable space.

$p(dw|x, u)$ —*Disturbance kernel*. For fixed $(x, u) \in SC$, $p(\cdot|x, u)$ is a probability measure on (W, \mathcal{F}) .

$f(x, u, w)$ —*System function*. A mapping from SCW to S .

$g(x, u, w)$ —*One-stage cost function*. A mapping from SCW to $[-\infty, +\infty]$.

M —*Control function space*. A nonempty set of mappings from S to C .

N —*Horizon*. A positive integer.

The model is stationary in that the data does not change from one stage to the next. There is no essential loss of generality in this assumption, since a nonstationary model can be reduced to a stationary one by state augmentation ([24, § 8], [1, § 6.7]). We impose no assumptions for the time being on the set of control functions M . However, specific results will assume explicitly or implicitly various conditions on M , and in fact our line of analysis is geared toward demonstrating the type of properties of M that are essential for specific results to hold. In particular special cases the set M could be as large as the set of all functions $\mu: S \rightarrow C$ or as restricted as the set of all linear functions $\mu: S \rightarrow C$ (S, C assumed to be linear spaces). We shall use the letter x to represent an element of S , and the letter u to represent an element of C . Denote by Π_N the Cartesian product of N copies of M and define

$$(7) \quad \Gamma = \{(x, u): x \in S, u = \mu(x) \text{ for some } \mu \in M\}.$$

We denote by Γ_x the cross-section $\{u: (x, u) \in \Gamma\}$. We refer to an element of Π_N as a *policy*.

We have in mind a system operating as follows. A policy $\pi = (\mu_0, \dots, \mu_{N-1}) \in \Pi_N$ is chosen. The system begins in some initial state x_0 and subsequent states are specified by the system equation

$$(6) \quad x_{k+1} = f(x_k, u_k, w_k), \quad k = 0, \dots, N-2,$$

where

$$(7) \quad u_k = \mu_k(x_k), \quad k = 0, \dots, N-1,$$

and w_k is random with distribution $p(dw_k|x_k, u_k)$. The cost incurred at each stage of the operation is $g(x_k, u_k, w_k)$, so the total cost is

$$\sum_{k=0}^{N-1} g(x_k, u_k, w_k).$$

The expected total cost corresponding to the policy π is obtained by taking the expectation of the total cost with respect to the appropriate probability measure. If the integrals can be defined, this can be represented by

$$(8) \quad \int_W \cdots \int_W \sum_{k=0}^{N-1} g(x_k, u_k, w_k) p(dw_{N-1}|x_{N-1}, u_{N-1}) \cdots p(dw_0|x_0, u_0),$$

where (6) and (7) hold. But we have not yet imposed sufficient structure for the integrals to be defined, so we specify the cost corresponding to each policy and initial state by means of outer integration.

Given a probability space (X, \mathcal{B}, p) and a function $f \in F_X$ with $f \geq 0$, the outer integral of f with respect to p is defined in Appendix A as

$$\int^* f dp = \inf \left\{ \int g dp : f \leq g, g \text{ is } \mathcal{B}\text{-measurable} \right\}.$$

Given an arbitrary $f \in F_X$, we define its *outer integral with respect to p* by

$$(9) \quad \int^* f dp = \int^* f^+ dp - \int^* f^- dp.$$

Since we take $\infty - \infty$ to be $+\infty$, the outer integral of f is defined for every $f \in F_X$. If f is measurable with respect to the σ -algebra \mathcal{B} , we write $\int f dp$ in place of $\int^* f dp$. Note that for such an f and $\alpha \in (-\infty, +\infty]$, we have

$$(10) \quad \alpha + \int f dp = \int (\alpha + f) dp.$$

For each $\mu \in M$, we define the mapping $T_\mu: F_S \rightarrow F_S$ by

$$(11) \quad T_\mu(J)(x) = \int^* \{g[x, \mu(x), w] + J[f(x, \mu(x), w)]\} p(dw | x, \mu(x)) \quad \forall J \in F_S, \quad x \in S.$$

We also define the mapping $T: F_S \rightarrow F_S$ by

$$(12) \quad T(J)(x) = \inf_{u \in \Gamma_x} \int^* \{g(x, u, w) + J[f(x, u, w)]\} p(dw | x, u) \quad \forall J \in F_S, \quad x \in S.$$

Note that since we have by definition $\Gamma_x = \{u \in C : u = \mu(x) \text{ for some } \mu \in M\}$, it follows that (12) can be replaced by

$$(13) \quad T(J)(x) = \inf_{\mu \in M} T_\mu(J)(x) \quad \forall x \in S.$$

For any $\mu_1, \dots, \mu_k \in M$, we denote by $(T_{\mu_1} \cdots T_{\mu_k})$ the composition of the mappings $T_{\mu_1}, \dots, T_{\mu_k}$. Similarly, we denote by T^k the composition of T with itself k times. For convenience we also use T^0 to denote the identity mapping on F_S .

The *cost function corresponding to the policy π* is defined by

$$(14) \quad J_{N,\pi} = (T_{\mu_0} \cdots T_{\mu_{N-1}})(J_0),$$

where $J_0(x) = 0$ for every $x \in S$. The *optimal cost function* is given by

$$(15) \quad J_N^*(x) = \inf_{\pi \in \Pi_N} J_{N,\pi}(x) \quad \forall x \in S.$$

If measurability assumptions are made so that reference to the outer integral is unnecessary and finiteness assumptions are imposed to allow the interchange of summation and integration, then (14) reduces to the more traditional definition of expected cost corresponding to a policy given by (8). One type of measurability assumption is to assume W is countable and \mathcal{F} is the power set of W , so that integration reduces to summation. No measure structure need be imposed on S and C . A less trivial set of assumptions is obtained by letting S have a σ -algebra \mathcal{S} , C have a σ -algebra \mathcal{C} , and assuming f is $(\mathcal{S} \otimes \mathcal{F}, \mathcal{S})$ measurable and g is $(\mathcal{S} \otimes \mathcal{F}, \mathcal{B})$ measurable,

where \mathcal{SCF} is the product σ -algebra and \mathcal{B} is the Borel σ -algebra in $[-\infty, +\infty]$. One must also assume $p(B|x, u)$ is \mathcal{SC} measurable for fixed $B \in \mathcal{F}$. If all mappings in M are $(\mathcal{S}, \mathcal{C})$ measurable, then whenever $\mu \in M$ and J is $(\mathcal{S}, \mathcal{B})$ measurable, $T_\mu(J)$ is also. Despite these assumptions, it may still occur that (8) and (14) do not agree, but if the possibility of $+\infty - \infty$ occurring is limited, agreement can be guaranteed. This can be accomplished by requiring that $g(x, u, w) > -\infty \forall x \in S, u \in C, w \in W$ (see (10)), or by requiring that for each $\pi = (\mu_0, \dots, \mu_{N-1}) \in \Pi_N$ and each $x_0 \in S$,

$$(16) \quad \int_W \cdots \int_W \sum_{k=0}^{N-1} g^+(x_k, u_k, w_k) p(dw_{N-1}|x_{N-1}, u_{N-1}) \cdots p(dw_0|x_0, u_0) < +\infty,$$

where (6) and (7) hold. If (16) holds for each $\pi \in \Pi_N$ and $x_0 \in S$, one can in fact show by Fubini's theorem that

$$J_{N,\pi}(x) = E_{(\pi,x)} \left\{ \sum_{k=0}^{N-1} g(x_k, u_k, w_k) \right\},$$

where (6) and (7) hold and the expectation is with respect to the product measure on $W \cdots W$ generated by π from $x_0 = x$. This is also the case if for each $\pi \in \Pi_N$ and $x_0 \in S$, (16) holds with g^+ replaced by g^- .

We now introduce various notions of optimality. Let $x \in S$ and $\varepsilon > 0$ be given. A policy $\pi \in \Pi_N$ is ε -optimal at x if

$$J_{N,\pi}(x) \leq \begin{cases} J_N^*(x) + \varepsilon & \text{if } J_N^*(x) > -\infty, \\ -1/\varepsilon & \text{if } J_N^*(x) = -\infty. \end{cases}$$

If $\pi \in \Pi_N$ is ε -optimal at every $x \in S$, we say π is ε -optimal. A policy $\pi \in \Pi_N$ is optimal at x if

$$J_{N,\pi}(x) = J_N^*(x).$$

If $\pi \in \Pi_N$ is optimal at every $x \in S$, we say π is optimal.

Let $\{\varepsilon_n\}$ be a sequence of positive numbers with $\varepsilon_n \downarrow 0$. A sequence of policies $\{\pi_n\} \subset \Pi_N$ is said to exhibit $\{\varepsilon_n\}$ dominated convergence to optimality if

$$J_{N,\pi_n} \rightarrow J_N^*, \\ J_{N,\pi_n}(x) \leq J_N^*(x) + \varepsilon_n \quad \text{if } J_N^*(x) > -\infty,$$

and

$$J_{N,\pi_n}(x) \leq J_{N,\pi_{n-1}}(x) + \varepsilon_n \quad \text{if } J_N^*(x) = -\infty.$$

If $\{\pi_n\}$ exhibits $\{\varepsilon_n\}$ dominated convergence to optimality and $J_N^*(x) > -\infty$ for every $x \in S$, then by definition π_n is ε_n -optimal.

3. Main results. For our results we shall need some regularity assumptions on the model. We list them here for convenience and shall refer to them explicitly when we wish to include them in the hypotheses of a proposition.

Assumption A: There is a subset F of F_S such that $J_0 \in F$ and whenever $J \in F$, then $T(J) \in F$.

Assumption B: If $J \in F$ as given in Assumption A and $\varepsilon > 0$, then there exists $\mu_\varepsilon \in M$ such that

$$T_{\mu_\varepsilon}(J)(x) \leq \begin{cases} T(J)(x) + \varepsilon & \text{if } T(J)(x) > -\infty, \\ -1/\varepsilon & \text{if } T(J)(x) = -\infty. \end{cases}$$

Assumption C: If $J \in F$ as given in Assumption A and the infimum in (12) is achieved for every $x \in S$, then there exists $\mu \in M$ such that

$$T_\mu(J)(x) = T(J)(x) \quad \forall x \in S.$$

Assumption D: For $J \in F$ as given in Assumption A, define

$$A(J) = \{(x, u) \in \Gamma: p^*(\{w: J[f(x, u, w)] = -\infty\} | x, u) > 0\},$$

where $p^*(\cdot | x, u)$ represents outer measure. For each $J \in F$, there is a $\mu_J \in M$ such that $(x, \mu_J(x)) \in A(J)$ whenever $x \in \text{proj}_S A(J)$. Furthermore if $\mu \in M$ and μ_J is as above, then $\hat{\mu}$ defined by

$$\hat{\mu}(x) = \begin{cases} \mu_J(x) & \text{if } x \in \text{proj}_S A(J), \\ \mu(x) & \text{otherwise,} \end{cases}$$

is in M . Also if $J \in F$ and $\mu_1, \mu_2 \in M$, then $\bar{\mu}$ defined by

$$\bar{\mu}(x) = \begin{cases} \mu_1(x) & \text{if } T_{\mu_1}(J)(x) \leq T_{\mu_2}(J)(x), \\ \mu_2(x) & \text{otherwise,} \end{cases}$$

is in M .

Assumption A will be used to show properties of $T^N(J_0)$, which is often identical to J_N^* . By choosing F to be the set of functions having measurability or continuity properties and showing that Assumption A holds, we can immediately deduce properties of $T^N(J_0)$. We will find it very important to be able to choose a control function which nearly achieves the infimum in the definition of $T(J)$ for $J \in F$. This is the condition given in Assumption B. Assumption C states that M is rich enough to allow exact selection of this infimum if it is achieved. This is necessary in order to construct an optimal policy. Assumption D states that M contains enough functions to allow certain constructions necessary for the proof of Proposition 1 below.

The following lemma provides some properties of the mappings T_μ and T that we shall need.

LEMMA 1. (a) If $J_1, J_2 \in F_S$ and $J_1 \leq J_2$, then $T(J_1) \leq T(J_2)$, and for all $\mu \in M$ $T_\mu(J_1) \leq T_\mu(J_2)$.

(b) If $J_1, J_2 \in F_S$ and $J_2 \leq J_1 + \varepsilon$ for some $\varepsilon > 0$, then $T_\mu(J_2) \leq T_\mu(J_1) + 2\varepsilon$ for all $\mu \in M$.

(c) If $J_1, J_2 \in F_S$ and for some $\varepsilon > 0$ we have

$$(17) \quad J_2(x) \leq J_1(x) + \varepsilon \quad \text{if } J_1(x) > -\infty,$$

then for all $\mu \in M$

$$(18) \quad T_\mu(J_2)(x) \leq T_\mu(J_1)(x) + 2\varepsilon \quad \text{if } T_\mu(J_1)(x) > -\infty.$$

Proof. (a) and (b) follow directly from Lemma A.3(a), (b), so we concentrate on proving (c). Let $x \in S$ be such that $T_\mu(J_1)(x) > -\infty$. Then either $T_\mu(J_1)(x) = +\infty$, in which case (18) is trivial or else from Lemma A.3(g) we obtain

$$(19) \quad p^*(A | x, \mu(x)) = 0,$$

where $A = \{w | J_1[f(x, \mu(x), w)] = -\infty\}$. From (17) we obtain for all $w \notin A$

$$g[x, \mu(x), w] + J_2[f(x, \mu(x), w)] \leq g[x, \mu(x), w] + J_1[f(x, \mu(x), w)] + \varepsilon$$

and (18) follows from Lemma A.3(b), (e) and (19). Q.E.D.

We now arrive at our first main result:

PROPOSITION 1. *Let Assumptions A and B hold.*

(a) *If $J_k^*(x) > -\infty$ for all $x \in S$ and $k = 1, 2, \dots, N$, then*

$$J_N^* = T^N(J_0),$$

and for every $\varepsilon > 0$ there exists an ε -optimal policy.

(b) *If Assumption D holds and $J_{k,\pi}(x) < +\infty$ for all $x \in S$, $\pi \in \Pi_k$ and $k = 1, 2, \dots, N$, then*

$$J_N^* = T^N(J_0),$$

and for every sequence $\{\varepsilon_n\}$ with $\varepsilon_n > 0$, $n = 1, 2, \dots$, $\varepsilon_n \downarrow 0$, there exists a sequence of policies exhibiting $\{\varepsilon_n\}$ dominated convergence to optimality.

Proof. (a) For any $\pi = (\mu_0, \dots, \mu_{k-1}) \in \Pi_k$,

$$\begin{aligned} J_{k,\pi} &= (T_{\mu_0} \cdots T_{\mu_{k-2}} T_{\mu_{k-1}})(J_0) \\ &\geq (T_{\mu_0} \cdots T_{\mu_{k-2}} T)(J_0) && \text{(by Lemma 1(a))} \\ &\geq \cdots \\ &\geq T^k(J_0). \end{aligned}$$

Hence

$$(20) \quad J_k^* \geq T^k(J_0), \quad k = 1, 2, \dots.$$

We conclude the proof by induction. Assumption B and (20) guarantee that when $N = 1$, (a) holds. Suppose (a) holds for $N - 1$. Then for $\varepsilon > 0$ there exists $\pi \in \Pi_{N-1}$ such that

$$J_{N-1,\pi} \leq J_{N-1}^* + \varepsilon/4.$$

From the induction assumption and Lemma 1(a), (b),

$$\begin{aligned} T^N(J_0) &= T(J_{N-1}^*) \\ &\geq T(J_{N-1,\pi} - \varepsilon/4) \\ &\geq T(J_{N-1,\pi}) - \varepsilon/2 \\ &= \inf_{\mu} T_{\mu}(J_{N-1,\pi}) - \varepsilon/2 \\ &\geq J_N^* - \varepsilon/2. \end{aligned}$$

Combining this with (20) we obtain

$$J_N^* = T^N(J_0).$$

Use Assumptions A and B to find $\mu \in M$ for which

$$T_{\mu}(J_{N-1}^*) \leq T(J_{N-1}^*) + \varepsilon/2.$$

With π as above, we have

$$\begin{aligned} J_{N,(\mu,\pi)} &= T_{\mu}(J_{N-1,\pi}) \\ &\leq T_{\mu}(J_{N-1}^*) + \varepsilon/2 \\ &\leq T(J_{N-1}^*) + \varepsilon \\ &= J_N^* + \varepsilon, \end{aligned}$$

so (μ, π) is an ε -optimal N -stage policy.

(b) The proof proceeds by induction. Let $\{\varepsilon_n\}$ be a sequence with $\varepsilon_n > 0$, $\varepsilon_n \downarrow 0$. For $N = 1$, Assumptions A and B imply the existence of a sequence of policies $\pi_n = (\mu_0^n) \in \Pi_1$ for which

$$(21) \quad T_{\mu_0^n}(J_0)(x) \leq \begin{cases} T(J_0)(x) + \varepsilon_n & \text{if } T(J_0)(x) > -\infty, \\ -1/\varepsilon_n & \text{if } T(J_0)(x) = -\infty. \end{cases}$$

By the last part of Assumption D we can assume without loss of generality that

$$(22) \quad T_{\mu_0^n}(J_0)(x) \downarrow -\infty \quad \text{if } T(J_0)(x) = -\infty.$$

This implies $J_1^* \leq T(J_0)$, which together with (20) establishes

$$(23) \quad J_1^* = T(J_0).$$

From (21)–(23) we see that $\{\pi_n\}$ exhibits $\{\varepsilon_n\}$ dominated convergence to optimality.

Suppose the result holds for $N-1$. Let $\pi_n = (\mu_1^n, \dots, \mu_{N-1}^n)$ be a sequence of $(N-1)$ -stage policies exhibiting $\{\varepsilon_n/4\}$ dominated convergence to optimality, i.e.,

$$(24) \quad J_{N-1, \pi_n} \rightarrow J_{N-1}^*,$$

$$(25) \quad J_{N-1, \pi_n}(x) \leq J_{N-1}^*(x) + \varepsilon_n/4 \quad \text{if } J_{N-1}^*(x) > -\infty,$$

$$(26) \quad J_{N-1, \pi_n}(x) \leq J_{N-1, \pi_{n-1}}(x) + \varepsilon_n/4 \quad \text{if } J_{N-1}^*(x) = -\infty.$$

We assume without loss of generality that $\sum_{n=1}^{\infty} \varepsilon_n < \infty$. By the induction hypothesis and Assumption A

$$(27) \quad J_{N-1}^* = T^{N-1}(J_0) \in F,$$

so by Assumption B there is a sequence $\{\mu^n\} \subset M$ such that

$$(28) \quad T_{\mu^n}(J_{N-1}^*)(x) \leq \begin{cases} T^N(J_0)(x) + \varepsilon_n/2 & \text{if } T^N(J_0)(x) > -\infty, \\ -2/\varepsilon_n & \text{if } T^N(J_0)(x) = -\infty. \end{cases}$$

By the last part of Assumption D we can assume without loss of generality that

$$(29) \quad T_{\mu^n}(J_{N-1}^*) \leq T_{\mu^{n-1}}(J_{N-1}^*), \quad n = 2, 3, \dots$$

By Assumption D there is a $\mu \in M$ such that $(x, \mu(x)) \in A(J_{N-1}^*)$ whenever $x \in \text{proj}_S A(J_{N-1}^*)$, i.e.,

$$(30) \quad p^*({w: J_{N-1}^* [f(x, \mu(x), w)] = -\infty} | x, \mu(x)) > 0$$

whenever for some $u \in \Gamma_x$

$$p^*({w: J_{N-1}^* [f(x, u, w)] = -\infty} | x, u) > 0.$$

Define

$$\hat{\mu}^n(x) = \begin{cases} \mu(x) & \text{if } x \in \text{proj}_S A(J_{N-1}^*), \\ \mu^n(x) & \text{otherwise.} \end{cases}$$

Then $\{\hat{\mu}^n\} \subset M$ by Assumption D and

$$\hat{\pi}_n = (\hat{\mu}^n, \pi_n) \in \Pi_N.$$

For $x \in \text{proj}_S A(J_{N-1}^*)$, we have

$$\begin{aligned}
 & \limsup_{n \rightarrow \infty} J_{N, \hat{\pi}_n}(x) \\
 &= \limsup_{n \rightarrow \infty} T_\mu(J_{N-1, \pi_n})(x) \\
 &= \limsup_{n \rightarrow \infty} \int^* \{g[x, \mu(x), w] + J_{N-1, \pi_n}[f(x, \mu(x), w)]\} p(dw | x, \mu(x)) \\
 &= \int^* \{g[x, \mu(x), w] + J_{N-1}^*[f(x, \mu(x), w)]\} p(dw | x, \mu(x))
 \end{aligned}$$

by (24)–(26), the fact that $J_{N, \hat{\pi}_1}(x) < +\infty$ for every $x \in S$, and Corollary A.1.1. Relation (30) and Lemma A.3(g) imply that $T_\mu(J_{N-1}^*)(x) = \pm\infty$. But $T_\mu(J_{N-1}^*)(x) \leq J_{N, \hat{\pi}_1}(x) < +\infty$, so

$$\begin{aligned}
 & \limsup_{n \rightarrow \infty} J_{N, \hat{\pi}_n}(x) = T_\mu(J_{N-1}^*)(x) \\
 (31) \quad & = -\infty \\
 & \leq T^N(J_0)(x).
 \end{aligned}$$

For $x \notin \text{proj}_S A(J_{N-1}^*)$, we have for each $u \in \Gamma_x$

$$(32) \quad p^*({w : J_{N-1}^*[f(x, u, w)] = -\infty} | x, u) = 0.$$

If $u = \mu^n(x)$ satisfies (32), then denoting $E = \{w : J_{N-1}^*[f(x, u, w)] = -\infty\}$, we have from Lemma A.3(e)

$$\begin{aligned}
 & J_{N, \hat{\pi}_n}(x) = T_{\mu^n}(J_{N-1, \pi_n})(x) \\
 &= \int^* \chi_{W-E}(w) \{g[x, \mu^n(x), w] + J_{N-1, \pi_n}[f(x, \mu^n(x), w)]\} p(dw | x, \mu^n(x)) \\
 (33) \quad & \cong \int^* \chi_{W-E}(w) \{g[x, \mu^n(x), w] + J_{N-1}^*[f(x, \mu^n(x), w)]\} p(dw | x, \mu^n(x)) + \varepsilon_n/2 \\
 & \quad \text{(by (25) and Lemma A.3(b))} \\
 &= T_{\mu^n}(J_{N-1}^*)(x) + \varepsilon_n/2.
 \end{aligned}$$

Inequality (33) implies for $x \notin \text{proj}_S A(J_{N-1}^*)$

$$\begin{aligned}
 & \limsup_{n \rightarrow \infty} J_{N, \hat{\pi}_n}(x) \leq \limsup_{n \rightarrow \infty} T_{\mu^n}(J_{N-1}^*)(x) \\
 (34) \quad & \leq T^N(J_0)(x) \quad \text{by (28).}
 \end{aligned}$$

Combining (31) and (34) we have

$$(35) \quad \limsup_{n \rightarrow \infty} J_{N, \hat{\pi}_n}(x) \leq T^N(J_0)(x)$$

for every $x \in S$, and this proves

$$(36) \quad J_N^* \leq T^N(J_0).$$

Combining (36) and (20), we obtain $J_N^* = T^N(J_0)$ and (35) can be replaced by

$$(37) \quad \lim_{n \rightarrow \infty} J_{N, \hat{\pi}_n}(x) = J_N^*(x)$$

for every $x \in S$.

To see that the convergence to optimality given in (37) is $\{\varepsilon_n\}$ dominated, note that if $T^N(J_0)(x) > -\infty$, then $T_{\hat{\mu}^n}(J_{N-1}^*)(x) > -\infty$ for every n . By Lemma 1(c) and (25)

$$(38) \quad T_{\hat{\mu}^n}(J_{N-1, \pi_n})(x) \leq T_{\hat{\mu}^n}(J_{N-1}^*)(x) + \varepsilon_n/2 \quad \text{if } T^N(J_0)(x) > -\infty.$$

If $x \notin \text{proj}_S A(J_{N-1}^*)$, then

$$(39) \quad \begin{aligned} T_{\hat{\mu}^n}(J_{N-1}^*)(x) &= T_{\mu^n}(J_{N-1}^*)(x) \\ &\leq T^N(J_0)(x) + \varepsilon_n/2 \quad \text{if } T^N(J_0)(x) > -\infty \quad \text{by (28)} \\ &= J_N^*(x) + \varepsilon_n/2 \quad \text{if } J_N^*(x) > -\infty \quad \text{by the fact } T^N(J_0) = J_N^*. \end{aligned}$$

Combining (38) and (39), we have for $x \notin \text{proj}_S A(J_{N-1}^*)$

$$(40) \quad J_{N, \hat{\pi}_n}(x) \leq J_N^*(x) + \varepsilon_n \quad \text{if } J_N^*(x) > -\infty.$$

If $x \in \text{proj}_S A(J_{N-1}^*)$, then it is clear from (31) that

$$J_N^*(x) = T(J_{N-1}^*)(x) = -\infty,$$

and so (40) is true for all $x \in S$. If $x \notin \text{proj}_S A(J_{N-1}^*)$, (33) and (29) can be used to show for $n \geq 2$

$$\begin{aligned} J_{N, \hat{\pi}_n}(x) &\leq T_{\mu^n}(J_{N-1}^*)(x) + \varepsilon_n/2 \\ &\leq T_{\mu^{n-1}}(J_{N-1}^*)(x) + \varepsilon_n/2 \\ &\leq T_{\mu^{n-1}}(J_{N-1, \pi_{n-1}})(x) + \varepsilon_n/2 \\ &= J_{N, \hat{\pi}_{n-1}}(x) + \varepsilon_n/2. \end{aligned}$$

If $x \in \text{proj}_S A(J_{N-1}^*)$,

$$\begin{aligned} J_{N, \hat{\pi}_n}(x) &= \int^* \{g[x, \mu(x), w] + J_{N-1, \pi_n}[f(x, \mu(x), w)]\} p(dw | x, \mu(x)) \\ &\leq \int^* \{g[x, \mu(x), w] + J_{N-1, \pi_{n-1}}[f(x, \mu(x), w)]\} p(dw | x, \mu(x)) + \varepsilon_n/2 \\ &\quad \text{(by (25), (26) and Lemma A.3(b))} \\ &= J_{N, \hat{\pi}_{n-1}}(x) + \varepsilon_n/2, \end{aligned}$$

so for every $x \in S$

$$J_{N, \hat{\pi}_n}(x) \leq J_{N, \hat{\pi}_{n-1}}(x) + \varepsilon_n/2. \quad \text{Q.E.D.}$$

To see that an assumption such as $J_k^*(x) > -\infty$ or $J_{k, \pi}(x) < +\infty$ for all x, π and k is necessary in order for $J_N^* = T^N(J_0)$ to hold, consider the following example:

Example. Let $N = 2$, $S = \{\alpha, \beta\}$, $C = (-\infty, +\infty)$, $W = \{1, 2, \dots\}$, $M = \{\mu: \mu(\alpha), \mu(\beta) \in (-\infty, +\infty)\}$, $p(w = k | x, u) = 1/(k^2 \sum_{j=1}^{\infty} 1/j^2)$, $k = 1, 2, \dots$, $f(\alpha, u, w) = f(\beta, u, w) = \beta \quad \forall u \in C, w \in W$, $g(\alpha, u, w) = w$, $g(\beta, u, w) = u$, $\forall u \in C, w \in W$. Here there are two states, α and β , and we always have $x_1 = \beta$ so that the cost of the second stage is $\mu_1(\beta)$ and can be made arbitrarily small. On the other hand,

$\int w dp = +\infty$, so that $J_{2,\pi}(\alpha) = \int \{g[\alpha, \mu_0(\alpha), w] + \mu_1(\beta)\} dp = \int \{w + \mu_1(\beta)\} dp = +\infty$ for all $\pi \in \Pi_2$. We have by a straightforward calculation $J_2^*(\alpha) = +\infty$, $J_2^*(\beta) = -\infty$, while $T^2(J_0)(\alpha) = -\infty$, $T^2(J_0)(\beta) = -\infty$.

Despite the need for various assumptions in order to show the equality $J_N^* = T^N(J_0)$, the following result, which establishes the validity of the DP algorithm as a means for obtaining optimal policies, requires none of the assumptions of Proposition 1. We say that a policy $\pi = (\mu_0, \dots, \mu_{N-1}) \in \Pi_N$ is *uniformly N -stage optimal* if for $k = 1, \dots, N$, $\pi^k = (\mu_{N-k}, \dots, \mu_{N-1})$ is optimal in the k -stage problem of minimizing $J_{k,\pi}$ over Π_k .

PROPOSITION 2. (a) *If there exists a uniformly N -stage optimal policy then*

$$(41) \quad J_k^* = T^k(J_0), \quad \forall k = 1, \dots, N.$$

(b) *A policy $\pi = (\mu_0, \dots, \mu_{N-1}) \in \Pi_N$ is uniformly N -stage optimal if and only if*

$$(42) \quad (T_{\mu_{N-k}} T^{k-1})(J_0) = T^k(J_0), \quad \forall k = 1, \dots, N.$$

Proof. (a) Let $\pi = (\mu_0, \dots, \mu_{N-1})$ be uniformly N -stage optimal. Then

$$(43) \quad T(J_0) = J_1^* = T_{\mu_{N-1}}(J_0)$$

by definition. For every $\mu \in M$,

$$(T_\mu T)(J_0) = (T_\mu T_{\mu_{N-1}})(J_0),$$

which implies

$$T^2(J_0) = \inf_{\mu \in M} (T_\mu T)(J_0) = \inf_{\mu \in M} (T_\mu T_{\mu_{N-1}})(J_0) \geq J_2^* = (T_{\mu_{N-2}} T_{\mu_{N-1}})(J_0) \geq T^2(J_0).$$

Therefore

$$(44) \quad T^2(J_0) = J_2^* = (T_{\mu_{N-2}} T)(J_0) = (T_{\mu_{N-2}} T_{\mu_{N-1}})(J_0).$$

Replace (43) by (44) and continue. This proves (41).

(b) This follows from (a) and (20). Q.E.D.

It follows from Proposition 2(b) that when M is rich enough so that Assumption C holds, then existence of a uniformly N -stage optimal policy is equivalent to attainment of the infimum in the DP algorithm. We state this as a separate proposition.

PROPOSITION 3. *Let Assumption C hold. A uniformly N -stage optimal policy exists if and only if the infimum in*

$$(45) \quad \inf_{u \in \Gamma_x} \int^* \{g(x, u, w) + T^{k-1}(J_0)[f(x, u, w)]\} p(dw | x, u)$$

is achieved for every $x \in S$, $k = 1, \dots, N$.

4. The model without topological assumptions. The simplest special case of our model is when we take $F = F_S$ in Assumption A and place restrictions on M only as follows. A subset Γ of SC is given with the property $\text{proj}_S \Gamma = S$ and M is taken to be the set of all mappings from S to C whose graphs lie in Γ . Note that Γ and M correspond as in (5). It is easy to see that Assumptions A, B, C, and D hold when F and M have been so chosen. Hence the results of Propositions 1–3 apply.

It is not customary to use outer integration in connection with Dynamic Programming, so the model outlined here is somewhat unusual. A special case of this model often considered is the case of a countable disturbance space [1]. As mentioned

earlier, integration reduces to summation in such a model. The countable disturbance space model is more restrictive than the Borel space model of § 6 in that a countable disturbance space is a special case of a Borel space. It is more general in that S and C are not required to have a Borel structure. The model described in the first paragraph of this section is, of course, more general than both. The main advantage that it offers is simplicity—there is no need to introduce elaborate topological assumptions in order to ascertain the validity of the DP algorithm. There are, however, inherent limitations centering around the pathologies of outer integration (Appendix A) in the nontopological model. Topological assumptions also play an important role in the treatment of problems with imperfect state information (see [2], [29]).

5. Borel space models with semicontinuity assumptions. We first introduce some notation and definitions. For any topological space Y we denote by \mathcal{B}_Y the Borel σ -algebra generated by the open sets. A *Borel space* X is a topological space such that there exists a complete separable metric space Y and a homeomorphism φ of X into Y with $\varphi(X) \in \mathcal{B}_Y$.

It follows that a Borel space X is metrizable and separable. Note that if X and Y are Borel spaces, then the product space XY equipped with the product topology is also a Borel space and \mathcal{B}_{XY} equals the product σ -algebra $\mathcal{B}_X \mathcal{B}_Y$ on XY [21, Chap. 1]. Also, every Borel subset of a Borel space becomes a Borel space when endowed with the relative topology. The extended real line $[-\infty, +\infty]$ with the topology generated by the open real intervals together with the sets $[-\infty, \alpha)$, and $(\alpha, +\infty]$, α real, is a Borel space. In what follows we implicitly assume that every Borel subset of $[-\infty, +\infty]$ is endowed with the corresponding relative topology and is thus a Borel space. If X and Y are Borel spaces and $f: X \rightarrow Y$ is such that $f^{-1}(B) \in \mathcal{B}_X$ for each $B \in \mathcal{B}_Y$, then we say that f is *Borel measurable*.

If X is a Borel space, we denote by $P(X)$ the set of probability measures on \mathcal{B}_X . We take the topology on $P(X)$ to be the weakest with respect to which all mappings of the form $p \rightarrow \int f dp$ are continuous, as f ranges over the set of bounded continuous real-valued functions on X . With this topology $P(X)$ becomes a Borel space [21, Chap. 2]. Let X and Y be Borel spaces and for each $x \in X$, let $q(dy|x)$ be a probability measure on \mathcal{B}_Y . If the mapping $x \rightarrow q(dy|x)$ is continuous from X to $P(Y)$, we will say that $q(dy|x)$ is a *continuous stochastic kernel on Y given X* .

We now define two special cases of the model of § 2. In both cases, S , C and W are Borel spaces, $\mathcal{F} = \mathcal{B}_W$, $p(dw|x, u)$ is a continuous stochastic kernel on W given SC , and $f(x, u, w)$ is continuous from SCW to S .

Lower semicontinuous model. Here C is compact, and g is lower semi-continuous and bounded below. A subset Γ of SC is given and is assumed to be of the form

$$\Gamma = \bigcup_{j=1}^{\infty} \Gamma_j,$$

where for all j , $\Gamma_j \subset \Gamma_{j+1}$, Γ_j is a closed subset of SC , and for all $(x, u) \in SC$

$$(46) \quad \lim_{j \rightarrow \infty} \inf_{(x, u) \in \Gamma_j - \Gamma_{j-1}} \int g(x, u, w) p(dw|x, u) = +\infty.$$

(By convention the infimum over the empty set is $+\infty$. Thus we allow the possibility that for some \bar{j} , $\Gamma_j = \Gamma_{\bar{j}}$ for all $j \geq \bar{j}$.) It is also assumed that $\text{proj}_S \Gamma = S$. The set M of admissible control functions is taken to be the set of all Borel measurable functions from S to C whose graphs lie in Γ . (Notice that if the sets Γ_j , $j = 1, 2, \dots$, are compact then there is no loss of generality in assuming that C is compact. This is true because if

C is not compact, it can be homeomorphically embedded in a compact Borel space \tilde{C} [7, Chap. 9, Cor. 9.2] and the images of Γ_j are compact and hence closed in $S\tilde{C}$. There is no need to extend f and g to $S\tilde{C}W$ nor $p(dw|x, u)$ to $S\tilde{C}$ for the proof we give of Proposition 4.)

Upper semicontinuous model. Here g is upper semicontinuous and bounded above. An open subset Γ of SC is given and it is assumed that $\text{proj}_S \Gamma = S$. The set M is the set of all Borel measurable functions from S to C whose graphs lie in Γ .

By selecting an appropriate subset $F \subset F_S$ for each model, we show now that some of the Assumptions A, B, C and D are satisfied. This in turn will allow application of some of the results of § 3.

PROPOSITION 4. (a) *In the lower semicontinuous model Assumptions A, B and C are satisfied with F being the set of lower semicontinuous functions $J: S \rightarrow (-\infty, +\infty]$ which are bounded below.*

(b) *In the upper semicontinuous model Assumptions A and B are satisfied with F being the set of upper semicontinuous functions $J: S \rightarrow [-\infty, +\infty)$ which are bounded above.*

Proof. (a) If $J: S \rightarrow (-\infty, +\infty]$ is lower semicontinuous and bounded below, then

$$H(x, u) = \int \{g(x, u, w) + J[f(x, u, w)]\} p(dw|x, u)$$

is also [25, Lemma 3.4]. Condition (46) guarantees that $\{(x, u) \in \Gamma: H(x, u) \leq \alpha\}$ is closed for each real α . The result follows from a simple modification of Lemma 3.4 and the Selection Theorem of [15].

(b) This follows from (17) of [9] and the fact that proj_S is an open mapping. Q.E.D.

By combining Proposition 4 with the results of § 2, we obtain the following:

COROLLARY 4.1. *In the lower semicontinuous model, we have $J_N^* = T^N(J_0)$ and there exists a uniformly N -stage optimal policy.*

COROLLARY 4.2. *In the upper semicontinuous model, if $J_k^*(x) > -\infty$ for all $x \in S$ and $k = 1, 2, \dots, N$, then $J_N^* = T^N(J_0)$, and for every $\varepsilon > 0$ there exists an ε -optimal policy.*

David Freedman [9] has proved results quite similar to Corollaries 4.1 and 4.2 by placing control constraints, not directly on the control u as we have done by requiring $(x, u) \in \Gamma$, but rather on the pair (x, P) , where P is the distribution of the subsequent state. Since the mapping $(x, u) \rightarrow (x, P)$ is continuous in the semicontinuous models, requiring (x, u) to be in an open set (our upper semicontinuous model) is slightly more general than requiring (x, P) to be in an open set (Freedman's model), while requiring (x, u) to be in the union of an increasing sequence of closed sets (our lower semicontinuous model) is significantly more general than requiring (x, P) to be in a closed set (Freedman's model). Our lower semicontinuous model does not require a compact state space. For example, we can take $S = R^n$, C to be the one point compactification of R^m ,

$$\Gamma_j = \{(x, u): u'u \leq j\}, \quad g(x, u, w) = x'Qx + u'Ru,$$

where Q is a positive semidefinite and R is a positive definite matrix of appropriate dimension.

6. General Borel space models with perfect state information. For the models of this section we shall need the notions of analytic sets, universally measurable sets and related facts. For more detailed treatments we refer the reader to [2], [6], [11], [14], [16], [21].

Let \mathcal{N} be the cross product of countably many copies of the positive integers. Let the set of positive integers have the discrete topology and \mathcal{N} the product topology. A separable metric space A is *analytic* if there is a continuous function f mapping \mathcal{N} onto A . In what follows the empty set will also be considered analytic.

We list some properties of analytic sets that we shall be using:

(a) Every Borel space is analytic but in every uncountable Borel space there exist analytic subsets which are not Borel spaces [14], § 38VI.

(b) The countable union, intersection, and cross product of analytic sets is analytic [21, Chap. 1, Thms. 3.1 and 3.2].

(c) If X and Y are Borel spaces, $A \subset X$ and $B \subset Y$ are analytic sets, and f is a Borel measurable function from X to Y , then $f(A)$ and $f^{-1}(B)$ are analytic [21, Chap. 1, Thm. 3.5]. As a consequence, if D is an analytic subset of XY , then $\text{proj}_X D$ is analytic.

In addition to the Borel σ -algebra, we are interested in two more σ -algebras that arise naturally in a Borel space X . The *analytic σ -algebra*, denoted \mathcal{A}_X , is the σ -algebra generated by the analytic subsets of X . The *universal σ -algebra*, denoted \mathcal{U}_X , is the intersection of all completions with respect to finite measures of the Borel σ -algebra \mathcal{B}_X . We have

$$\mathcal{B}_X \subset \mathcal{A}_X \subset \mathcal{U}_X,$$

and if X is uncountable both inclusions are strict. In fact, it is possible to prove that if X is uncountable, then under the continuum hypothesis \mathcal{U}_X has a larger cardinality than both \mathcal{B}_X and \mathcal{A}_X . (We are indebted to Professor J. Doob for pointing out this fact to us.)

Let X and Y be Borel spaces, D be a subset of X , and $f: D \rightarrow Y$. If $D \in \mathcal{A}_X$ and $f^{-1}(B) \in \mathcal{A}_X$ for all $B \in \mathcal{B}_Y$, we say that f is *analytically measurable*. If $D \in \mathcal{U}_X$ and $f^{-1}(B) \in \mathcal{U}_X$ for all $B \in \mathcal{B}_Y$, we say that f is *universally measurable*. If D is analytic, $Y = [-\infty, +\infty]$ and the set $\{x \in D: f(x) < \alpha\}$ is analytic for every real α , we say that f is *lower semianalytic*. For a lower semianalytic f , the sets $\{x \in D: f(x) \leq \alpha\}$ are also analytic for every $\alpha \in [-\infty, +\infty]$. Note that a lower semianalytic function is analytically measurable and hence also universally measurable, the sum of two lower semianalytic functions is lower semianalytic, and a Borel measurable function from X to $[-\infty, +\infty]$ is lower semianalytic.

If X is a Borel space and $p \in P(X)$, then p has a unique extension to a probability measure on \mathcal{U}_X . We denote this extension by p also, and we write $p(E)$ instead of $p^*(E)$ when $E \in \mathcal{U}_X$. Likewise, if $f: X \rightarrow [-\infty, +\infty]$ is a universally measurable function we will write $\int f dp$ in place of $\int^* f dp$. Under these circumstances $\int f dp$ obeys the rules of classical integration, provided we take care in handling the expression $+\infty - \infty$.

If X and Y are Borel spaces, $q(dy|x)$ is a probability measure on \mathcal{B}_Y for each $x \in X$, and the function $q(B|\cdot)$ is Borel measurable from X to $[0, 1]$ for all $B \in \mathcal{B}_Y$, we say that $q(dy|x)$ is a *Borel measurable stochastic kernel on Y given X* . The stochastic kernel $q(dy|x)$ is Borel measurable if and only if the mapping $x \rightarrow q(\cdot|x)$ is Borel measurable from X to $P(Y)$.

We now specify the two special cases of the problem of § 2 to be considered in this section. In both cases, S , C and W are Borel spaces, $\mathcal{F} = \mathcal{B}_W$, $p(dw|x, u)$ is a Borel measurable stochastic kernel on W given SC , f is Borel measurable, and g is lower semianalytic. In both models an analytic subset Γ of SC is given with $\text{proj}_S \Gamma = S$. The models differ only in the specification of M .

Borel model with analytically measurable policies (BAP for short): Here M consists of all analytically measurable functions from S to C whose graphs lie in Γ .

Borel model with universally measurable policies (BUP for short): Here M consists of all universally measurable functions from S to C whose graphs lie in Γ .

Our main result of this section is the following:

PROPOSITION 5. (a) *In BAP Assumptions A and B are satisfied with F being the class of lower semianalytic functions $J: S \rightarrow [-\infty, +\infty]$.*

(b) *In BUP Assumptions A, B, C and D are satisfied with F being the class of lower semianalytic functions $J: S \rightarrow [-\infty, +\infty]$.*

We postpone the proof of Proposition 5 until we develop some further machinery. By combining Proposition 5 with the results of § 3 we obtain the following:

COROLLARY 5.1. *In BAP if $J_k^*(x) > -\infty$ for all $x \in S$ and $k = 1, \dots, N$, then $J_N^* = T^N(J_0)$ and for every $\varepsilon > 0$ there exists an ε -optimal policy.*

COROLLARY 5.2. *Consider BUP.*

(a) *If $J_k^*(x) > -\infty$ for all $x \in S$ and $k = 1, \dots, N$, then $J_N^* = T^N(J_0)$ and for every $\varepsilon > 0$ there exists an ε -optimal policy.*

(b) *If $J_{k,\pi}(x) < +\infty$ for all $x \in S$, $\pi \in \Pi_k$ and $k = 1, \dots, N$, then $J_N^* = T^N(J_0)$ and for every sequence $\{\varepsilon_n\}$, $\varepsilon_n > 0$, $n = 1, 2, \dots$, $\varepsilon_n \downarrow 0$, there exists a sequence of policies exhibiting $\{\varepsilon_n\}$ dominated convergence to optimality. If in addition $J_N^*(x) > -\infty$ for all $x \in S$, then for every $\varepsilon > 0$ there exists an ε -optimal policy.*

(c) *If the infimum in*

$$\inf_{u \in \Gamma_x} \int^* \{g(x, u, w) + T^{k-1}(J_0)[f(x, u, w)]\} p(dw | x, u)$$

is attained for every $x \in S$ and $k = 1, \dots, N$, then $J_k^ = T^k(J_0)$, $k = 1, \dots, N$, and there exists a uniformly N -stage optimal policy.*

We now provide two results that are crucial in our development. The first is often attributed to von Neuman [17], but was also proved by Jankov [13]. A proof of the version given here may be found in Blackwell, Freedman and Orkin [4]. Part (a) of the second result is contained in a proof given by Blackwell, Freedman and Orkin [4, Thm. (43)]. Part (b) is an extension of a selection theorem of Brown and Purves [5, Thm. 2] in that f is allowed to be lower semianalytic rather than Borel measurable. Our proof parallels the proofs of [4] and [5].

JANKOV-VON NEUMANN LEMMA. *Let X and Y be Borel spaces and $A \subset XY$ be an analytic set. Then there exists an analytically measurable function $\varphi: \text{proj}_X A \rightarrow Y$ such that $(x, \varphi(x)) \in A$ for every $x \in \text{proj}_X A$.*

SELECTION THEOREM. *Let X and Y be Borel spaces, $D \subset XY$ be an analytic set, and $f: D \rightarrow [-\infty, +\infty]$ be a lower semianalytic function. Define $g: \text{proj}_X D \rightarrow [-\infty, +\infty]$ by*

$$(47) \quad g(x) = \inf_{y \in D_x} f(x, y),$$

where $D_x = \{y: (x, y) \in D\}$. Then g is lower semianalytic. Furthermore:

(a) *For every $\varepsilon > 0$ there exists an analytically measurable function $\varphi: \text{proj}_X D \rightarrow Y$ such that for all $x \in \text{proj}_X D$*

$$f[x, \varphi(x)] \leq g(x) + \varepsilon \quad \text{if } g(x) > -\infty,$$

$$f[x, \varphi(x)] \leq -\frac{1}{\varepsilon} \quad \text{if } g(x) = -\infty.$$

(b) *The set*

$$I = \{x \in \text{proj}_X D : \text{for some } y_0 \in D_x, f(x, y_0) = g(x)\}$$

is universally measurable, and for every $\varepsilon > 0$ there exists a universally measurable function $\varphi: \text{proj}_X D \rightarrow Y$ such that for all $x \in \text{proj}_X D$

$$\begin{aligned} f[x, \varphi(x)] &= g(x) & \text{if } x \in I, \\ f[x, \varphi(x)] &\leq g(x) + \varepsilon & \text{if } x \notin I, g(x) > -\infty, \\ f[x, \varphi(x)] &\leq -\frac{1}{\varepsilon} & \text{if } x \notin I, g(x) = -\infty. \end{aligned}$$

Proof. (a) Since $\{x: g(x) < \alpha\} = \text{proj}_X \{(x, y) \in D: f(x, y) < \alpha\}$, g is lower semi-analytic. For $k = 0, \pm 1, \pm 2, \dots$ define

$$\begin{aligned} A(k) &= \{(x, y) \in D: f(x, y) < k\varepsilon\}, \\ B(k) &= \{x \in \text{proj}_X D: (k-1)\varepsilon \leq g(x) < k\varepsilon\}, \\ B(-\infty) &= \{x \in \text{proj}_X D: g(x) = -\infty\}, \\ B(+\infty) &= \{x \in \text{proj}_X D: g(x) = +\infty\}. \end{aligned}$$

The sets $A(k)$, $k = 0, \pm 1, \pm 2, \dots$ and $B(-\infty)$ are analytic, while the sets $B(k)$, $k = 0, \pm 1, \pm 2, \dots$ and $B(+\infty)$ are analytically measurable. By the Jankov-Von Neumann Lemma there exists, for each $k = 0, \pm 1, \pm 2, \dots$, an analytically measurable $\varphi_k: \text{proj}_X A(k) \rightarrow C$ with $(x, \varphi_k(x)) \in A(k)$ for all $x \in \text{proj}_X A(k)$, and an analytically measurable $\bar{\varphi}: \text{proj}_X D \rightarrow C$ such that $(x, \bar{\varphi}(x)) \in D$ for all $x \in \text{proj}_X D$. Let k^* be an integer such that $k^* \leq -1/\varepsilon^2$. Define $\varphi: \text{proj}_X D \rightarrow C$ by

$$\varphi(x) = \begin{cases} \varphi_k(x) & \text{if } x \in B(k), \quad k = 0, \pm 1, \pm 2, \dots, \\ \bar{\varphi}(x) & \text{if } x \in B(+\infty), \\ \varphi_{k^*}(x) & \text{if } x \in B(-\infty). \end{cases}$$

Since $B(k) \subset \text{proj}_X A(k)$ and $B(-\infty) \subset \text{proj}_X A(k)$ for all k this definition is possible. Then φ has the required properties.

(b) Denote by Q the set of rationals and let $Q^* = Q \cup \{-\infty, +\infty\}$. Denote also by R^* the extended reals. Consider the set $E \subset XYR^*$ defined by

$$E = \{(x, y, b): (x, y) \in D, f(x, y) \leq b\}.$$

Since

$$E = \bigcap_{k=1}^{\infty} \bigcup_{r \in Q^*} \left\{ (x, y, b): (x, y) \in D, f(x, y) \leq r, r \leq b + \frac{1}{k} \right\},$$

it follows that E is analytic in XYR^* and hence the set

$$A = \text{proj}_{XR^*}(E)$$

is analytic in XR^* . The mapping $T: \text{proj}_X D \rightarrow XR^*$ defined by

$$T(x) = (x, g(x))$$

is analytically measurable and

$$I = \{x: (x, g(x)) \in A\} = T^{-1}(A).$$

Since the inverse image under a universally measurable function of a universally measurable set is universally measurable, I is universally measurable.

Since E is analytic, by the Jankov–Von Neumann Lemma there is an analytically measurable $\rho: A \rightarrow Y$ such that $(x, \rho(x, b), b) \in E$ for every $(x, b) \in A$. Define $\Psi: I \rightarrow Y$ by

$$\Psi(x) = \rho(x, g(x)) = (\rho \circ T)(x) \quad \forall x \in I.$$

Then Ψ is universally measurable and by construction

$$(48) \quad f[x, \Psi(x)] = g(x) \quad \forall x \in I.$$

By part (a) there exists an analytically measurable $\Psi_\varepsilon: \text{proj}_X D \rightarrow Y$ such that

$$(49) \quad f[x, \Psi_\varepsilon(x)] \leq g(x) + \varepsilon \quad \text{if } g(x) > -\infty,$$

$$(50) \quad f[x, \Psi_\varepsilon(x)] \leq -1/\varepsilon \quad \text{if } g(x) = -\infty.$$

Define $\varphi: \text{proj}_X D \rightarrow Y$ by

$$\varphi(x) = \begin{cases} \Psi(x) & \text{if } x \in I, \\ \Psi_\varepsilon(x) & \text{if } x \in \text{proj}_X D - I. \end{cases}$$

Then φ is universally measurable and, by (48)–(50), it has the required properties. Q.E.D.

Suppose X and Y are Borel spaces, $f: XY \rightarrow [-\infty, +\infty]$ is universally measurable (i.e., measurable with respect to \mathcal{U}_{XY}), and $q(dy|x)$ is a Borel measurable stochastic kernel. Then it can be shown that $g(x) = \int f(x, y)q(dy|x)$ is universally measurable. If f is actually Borel measurable, so is g . If f is lower semianalytic, then g is also lower semianalytic. This last fact can be obtained by modifying Lemma (29) of [4] (see [29]).

We are now ready to prove Proposition 5.

Proof of Proposition 5. Let $J: S \rightarrow [-\infty, +\infty]$ be lower semianalytic. Then the function $H: SC \rightarrow [-\infty, +\infty]$ defined by

$$H(x, u) = \int \{g(x, u, w) + J[f(x, u, w)]\}p(dw|x, u)$$

is lower semianalytic. It follows that

$$T(J)(x) = \inf_{u \in \Gamma_x} H(x, u),$$

is lower semianalytic. Since J_0 is lower semianalytic, Assumption A is satisfied for both BAP and BUP. The Selection Theorem guarantees that Assumption B is satisfied for BAP, while Assumptions B and C are satisfied for BUP. It remains to verify Assumption D for BUP. We first show that the function

$$(51) \quad (x, u) \rightarrow -p(\{w: J[f(x, u, w)] = -\infty\}|x, u)$$

is lower semianalytic whenever J is. Define $t(dx'|x, u)$ by

$$t(E|x, u) = p(\{w: f(x, u, w) \in E\}|x, u) \quad \forall E \in \mathcal{B}_S.$$

We will show that $t(dx'|x, u)$ is a Borel measurable stochastic kernel. Clearly for fixed (x, u) , $t(\cdot|x, u)$ is a probability measure on S . We need to show that $p(B_{(x,u)}|x, u)$ is Borel measurable for each Borel subset B of SCW [$B_{(x,u)}$ is the cross section

$\{w: (x, u, w) \in B\}$. It is easy to show that the sets $B \in \mathcal{B}_{SCW}$ for which $p(B_{(x,u)}|x, u)$ is measurable form a Dynkin system, so that by the Dynkin system theorem, we need only verify that $p((B_1 B_2 B_3)_{(x,u)}|x, u)$ is measurable for all $B_1 \in \mathcal{B}_S$, $B_2 \in \mathcal{B}_C$, $B_3 \in \mathcal{B}_W$. But

$$p((B_1 B_2 B_3)_{(x,u)}|x, u) = \begin{cases} p(B_3|x, u) & \text{if } (x, u) \in B_1 B_2, \\ 0 & \text{otherwise,} \end{cases}$$

and this is measurable since $p(B_3|x, u)$ is measurable. Hence $t(dx'|x, u)$ is a Borel measurable stochastic kernel. We have for all $(x, u) \in SC$

$$(52) \quad \begin{aligned} -p(\{w: J[f(x, u, w)] = -\infty\}|x, u) &= -t(\{x': J(x') = -\infty\}|x, u) \\ &= \int -\chi_{\{J = -\infty\}} t(dx'|x, u). \end{aligned}$$

The function $-\chi_{\{J = -\infty\}}$ can be easily seen to be lower semianalytic. It follows from the remark preceeding the proof that the function of (51) is lower semianalytic. Hence the set

$$A(J) = \{(x, u) \in \Gamma: p(\{w: J[f(x, u, w)] = -\infty\}|x, u) > 0\}$$

is analytic, and by the Jankov–Von Neumann Lemma there is a $\mu_J \in M$ such that $(x, \mu_J(x)) \in A(J)$ whenever $x \in \text{proj}_S A(J)$. If $\mu \in M$ and $\hat{\mu}$ is as in Assumption D, it follows from the fact that $\text{proj}_S A(J)$ is analytic and hence universally measurable that $\hat{\mu} \in M$. If $\mu_1, \mu_2 \in M$, then $T_{\mu_1}(J)$ and $T_{\mu_2}(J)$ can be easily shown to be universally measurable. Hence if μ is as in Assumption D it follows that $\bar{\mu} \in M$. Q.E.D.

Remark. In the models in which C is equipped with a σ -algebra, one can speak of randomized policies $\pi = (\mu_0, \dots, \mu_{N-1})$, where $\mu_k(du_k|x_k)$ is an appropriately measurable stochastic kernel on C given S . Control constraints can be introduced by requiring that $\mu_k(\Gamma_x|x) = 1$ for every $x \in S$, $k = 0, \dots, N-1$, where $\Gamma = \{(x, u): u \in \Gamma_x\}$ is some prescribed subset of SC . The cost corresponding to such a policy is

$$J_{N,\pi} = (T_{\mu_0} \cdots T_{\mu_{N-1}})(J_0),$$

where

$$T_{\mu_k}(J)(x) = \int_C \int_W \{g(x, u, w) + J[f(x, u, w)]\} p(dw|x, u) \mu_k(du|x).$$

It is clear that $J_{N,\pi}$ is bounded below by $T^N(J_0)$, so if $J_N^* = T^N(J_0)$, the admission of randomized policies to the models considered does not alter the optimal cost function. Note however that in the example of § 3 if randomized policies are admitted, then the optimal cost function becomes $J_2^*(\alpha) = J_2^*(\beta) = -\infty$ and is different from the one corresponding to nonrandomized policies. Furthermore an optimal randomized policy exists.

Remark. There is the σ -algebra of “ C -sets” studied by Selivanovskij [26]. This σ -algebra, which we call the limit σ -algebra, is contained strictly between the analytic and universal σ -algebras in Borel spaces and has the property that all the results of this section remain valid if the words “universally measurable” are replaced by “limit measurable”. This σ -algebra is the minimal acceptable σ -algebra for DP in the sense that the composition of limit measurable functions is limit measurable and every analytically measurable function is limit measurable, but no smaller σ -algebra has these two properties. The limit σ -algebra is discussed more fully in [2], [30].

Appendix A: The outer integral. Throughout this appendix, (X, \mathcal{B}, p) is a probability space. Unless otherwise specified f , g , and h are functions from X to $[-\infty, +\infty]$.

DEFINITION A.1. If $f \geq 0$, the *outer integral* of f with respect to p is defined by

$$(A.1) \quad \int^* f dp = \inf \left\{ \int g dp : f \leq g, g \text{ is } \mathcal{B}\text{-measurable} \right\}.$$

If f is arbitrary, define

$$(A.2) \quad \int^* f dp = \int^* f^+ dp - \int^* f^- dp.$$

If $\int^* f^+ dp < +\infty$, we say f is *outer summable above*. If $\int^* f^- dp < +\infty$, we say f is *outer summable below*. If f is outer summable above or outer summable below, we say f is *outer summable*. In the following discussion, simple proofs are omitted.

LEMMA A.1. If $f \geq 0$, then there exists a \mathcal{B} -measurable g with $g \geq f$, such that

$$(A.3) \quad \int^* f dp = \int g dp.$$

LEMMA A.2. If $f \geq 0$, $h \geq 0$, then

$$(A.4) \quad \int^* (f+h) dp \leq \int^* f dp + \int^* h dp.$$

If either f or h is \mathcal{B} -measurable, then equality holds in (A.4).

We provide an example to show that strict inequality can occur in (A.4), even if $f+h$ is \mathcal{B} -measurable. For this and subsequent examples we will need the following observation: For any $E \subset X$,

$$(A.5) \quad \int^* \chi_E dp = p^*(E),$$

where $p^*(E)$ is the p -outer measure defined by

$$(A.6) \quad p^*(E) = \inf \{ p(B) : E \subset B, B \in \mathcal{B} \}.$$

This follows from the fact that for any set E there exists a set $A \in \mathcal{B}$ such that $E \subset A$ and $p(A) = p^*(E)$.

Example A.1. Let $X = [0, 1]$, \mathcal{B} be the Borel σ -algebra, and p be Lebesgue measure restricted to \mathcal{B} . Let $E \subset X$ be a set for which $p^*(X-E) = 1$ [10, § 16, Thm. E]. Then

$$\int (\chi_E + \chi_{X-E}) dp = \int 1 dp = 1, \quad \int^* \chi_E dp + \int^* \chi_{X-E} dp = 2,$$

and strict inequality holds in (A.4).

Lemma A.2 cannot be extended to (possibly negative) bounded functions even if h is \mathcal{B} -measurable, as the following example demonstrates.

Example A.2. Let (X, \mathcal{B}, p) and E be as before. Let $f = \chi_E - \chi_{X-E}$, $h = 1$. Then

$$\begin{aligned} \int^* (f+h) dp &= \int^* 2\chi_E dp = 2, \\ \int^* f dp + \int h dp &= \int^* \chi_E dp - \int^* \chi_{X-E} dp + 1 = 1. \end{aligned}$$

LEMMA A.3. (a) If $f \leq g$ then $\int^* f dp \leq \int^* g dp$.

(b) If $\varepsilon > 0$ and $f \leq g \leq f + \varepsilon$, then

$$(A.7) \quad \int^* f dp \leq \int^* g dp \leq \int^* f dp + 2\varepsilon.$$

(c) If f is outer summable, then

$$(A.8) \quad \int^* (-f) dp = - \int^* f dp.$$

(d) If $A, B \in \mathcal{B}$ are disjoint, then for any f

$$(A.9) \quad \int^* \chi_{A \cup B} f dp = \int^* \chi_A f dp + \int^* \chi_B f dp.$$

(e) If $E \subset X$ satisfies $p^*(E) = 0$, then for every f

$$\int^* f dp = \int^* \chi_{X-E} f dp.$$

(f) If $p^*({x: f(x) = +\infty}) > 0$ then, for every g , $\int^* (g + f) dp = +\infty$.

(g) If $p^*({x: f(x) = -\infty}) > 0$ then, for every g , either $\int^* (g + f) dp = +\infty$ or $\int^* (g + f) dp = -\infty$.

Proof. (b) In light of (a), it remains only to show that

$$(A.10) \quad \int^* (f + \varepsilon) dp \leq \int^* f dp + 2\varepsilon.$$

For $g_1 \geq f^+$, g_1 \mathcal{B} -measurable and

$$\int^* f^+ dp = \int g_1 dp,$$

we have

$$(f + \varepsilon)^+ \leq g_1 + \varepsilon$$

so

$$(A.11) \quad \int^* (f + \varepsilon)^+ dp \leq \int g_1 dp + \varepsilon = \int^* f^+ dp + \varepsilon.$$

For $g_2 \geq (f + \varepsilon)^-$, g_2 \mathcal{B} -measurable and

$$(A.12) \quad \int^* (f + \varepsilon)^- dp = \int g_2 dp,$$

we have

$$\begin{aligned} g_2 + \varepsilon &\geq (f + \varepsilon)^- + \varepsilon \\ &= \max \{f^- - \varepsilon, 0\} + \varepsilon \\ &\geq f^-, \end{aligned}$$

so

$$\begin{aligned}
 \varepsilon + \int^* (f + \varepsilon)^- dp &= \varepsilon + \int g_2 dp \\
 (A.13) \qquad \qquad \qquad &= \int (g_2 + \varepsilon) dp \\
 &\cong \int^* f^- dp \quad \text{by (a).}
 \end{aligned}$$

Combine (A.11) and (A.13) to conclude (A.10).

(f) We have $(g+f)^+(x) = +\infty$ if $f(x) = +\infty$, so that $p^*({x: (g+f)^+(x) = +\infty}) > 0$. Hence $\int^* (g+f)^+ dp = +\infty$ and it follows that $\int^* (g+f) dp = +\infty$.

(g) Consider the sets $E = {x: f(x) = -\infty}$ and $E_g = {x: f(x) = -\infty, g(x) < +\infty}$. If $p^*(E_g) = 0$ then we have

$$(A.14) \qquad p^*(E - E_g) = p^*(E - E_g) + p^*(E_g) \geq p^*(E) > 0.$$

Since we have $f(x) + g(x) = +\infty$ for $x \in E - E_g$, it follows from (f) that $\int^* (g+f) dp = +\infty$. If $p^*(E_g) > 0$, then $p^*({x: (g+f)^-(x) = +\infty}) \geq p^*(E_g) > 0$ and hence, by (f), $\int^* (g+f)^- dp = +\infty$. Hence if $\int^* (g+f)^+ dp = +\infty$ then $\int^* (g+f) dp = +\infty$, while if $\int^* (g+f)^+ dp < +\infty$ then $\int^* (g+f) dp = -\infty$. Q.E.D.

The bound given in (A.7) is the sharpest possible. To see this, let f be as defined in Example A.2, $g = f + 1$ and $\varepsilon = 1$. Despite these pathologies of outer integration, there is a monotone convergence theorem, which we now prove.

THEOREM A.1. *If $\{f_n\}$ is a sequence of nonnegative functions and $f_n \uparrow f$, then*

$$(A.15) \qquad \int^* f_n dp \uparrow \int^* f dp.$$

If $\{f_n\}$ is a sequence of nonpositive functions and $f_n \downarrow f$, then

$$\int^* f_n dp \downarrow \int^* f dp.$$

Proof. We prove the first statement of the theorem. The second follows from the first by Lemma A.3(c). Assume $f_n \geq 0$ and $f_n \uparrow f$. Let $\{g_n\}$ be a sequence of \mathcal{B} -measurable functions such that $g_n \geq f_n$ and

$$(A.16) \qquad \int^* f_n dp = \int g_n dp.$$

If for some n , $\int g_n dp = \int^* f_n dp = +\infty$, then (A.15) is assured. If not, then for every n

$$(A.17) \qquad \int g_n dp < \infty.$$

Suppose (A.17) holds for every n and for some n ,

$$p({x: g_n(x) > g_{n+1}(x)}) > 0.$$

Then since $g_{n+1} \geq f_{n+1} \geq f_n$, we have that \bar{g} defined by

$$\bar{g}(x) = \begin{cases} g_n(x) & \text{if } g_n(x) \leq g_{n+1}(x), \\ g_{n+1}(x) & \text{if } g_n(x) > g_{n+1}(x), \end{cases}$$

satisfies $g_n \geq \bar{g} \geq f_n$ everywhere and $\bar{g} < g_n$ on a set of positive measure. This contradicts (A.16). We may therefore assume without loss of generality that $g_1 \leq g_2 \leq \dots$. Let $g = \lim_{n \rightarrow \infty} g_n$. Then $g \geq f$ and

$$\lim_{n \rightarrow \infty} \int^* f_n dp = \lim_{n \rightarrow \infty} \int g_n dp = \int g dp \geq \int^* f dp.$$

But $f_n \leq f$ for every n , so the reverse inequality holds as well. Q.E.D.

One might hope that if $\{f_n\}$ is a sequence of functions which are bounded below and $f_n \uparrow f$, then (A.15) remains valid. This is not the case, as the following example shows.

Example A.3. Let $X = [0, 1]$, \mathcal{B} be the Borel σ -algebra, and p be Lebesgue measure restricted to \mathcal{B} . Define an equivalence relation \sim on X by

$$x \sim y \Leftrightarrow x - y \text{ is rational.}$$

Let F_0 be constructed by choosing one representative from each equivalence class. Let $Q = \{q_0, q_1, \dots\}$ be an enumeration of the rationals in $[0, 1]$ with $q_0 = 0$ and define

$$F_k = F_0 + q_k \pmod{1}, \quad k = 0, 1, \dots$$

Then F_0, F_1, \dots is a sequence of disjoint sets with

$$(A.18) \quad \bigcup_{k=0}^{\infty} F_k = [0, 1).$$

If for some $n < \infty$, we have $p^*(\bigcup_{k=n}^{\infty} F_k) < 1$, then $E = \bigcup_{k=0}^{n-1} F_k$ contains a \mathcal{B} -measurable set with measure $\delta > 0$. For $k = 1, \dots, n-1$, let $q_k = r_k/s_k$, where r_k and s_k are integers and r_k/s_k is reduced to lowest terms. Let $\{p_1, p_2, \dots\}$ be a sequence of prime numbers such that

$$\max_{1 \leq k \leq n-1} s_k < p_1 < p_2 < \dots$$

Then the sets $E, E + p_1^{-1} \pmod{1}, E + p_2^{-1} \pmod{1}, \dots$ are disjoint, and by the translation invariance of p , each contains a \mathcal{B} -measurable set with measure $\delta > 0$. It follows that $[0, 1)$ must contain a \mathcal{B} -measurable set of infinite measure. This contradiction implies

$$(A.19) \quad p^*\left(\bigcup_{k=n}^{\infty} F_k\right) = 1$$

for every n . Define

$$f_n = -\chi_{\bigcup_{k=n}^{\infty} F_k}, \quad n = 0, 1, \dots$$

Then $f_n \uparrow 0$, but (A.5) and (A.19) imply that for every n

$$\int^* f_n dp = -1.$$

By a change of sign in Example A.3, we see that the second part of Theorem A.1 cannot be extended to functions which are bounded above unless additional conditions are imposed. We impose such conditions in order to prove a corollary.

COROLLARY A.1.1. Let $\{\varepsilon_n\}$ be a sequence of positive numbers with $\sum_{n=1}^{\infty} \varepsilon_n < \infty$.

Let $\{f_n\}$ be a sequence with

$$(A.20) \quad \lim_{n \rightarrow \infty} f_n = f,$$

$$(A.21) \quad f \leq f_n, \quad n = 1, 2, \dots,$$

$$(A.22) \quad f_n(x) \leq f(x) + \varepsilon_n \quad \text{if } f(x) > -\infty,$$

$$(A.23) \quad f_n(x) \leq f_{n-1}(x) + \varepsilon_n \quad \text{if } f(x) = -\infty, \quad n = 2, 3, \dots,$$

$$(A.24) \quad \int^* f_1 dp < +\infty.$$

Then

$$(A.25) \quad \lim_{n \rightarrow \infty} \int^* f_n dp = \int^* f dp.$$

Proof. From (A.20) we have $\lim_{n \rightarrow \infty} f_n^+ = f^+$ and $\lim_{n \rightarrow \infty} f_n^- = f^-$. Now

$$\inf_{k \geq n} f_k^- \leq f_n^- \leq f^-$$

and

$$\inf_{k \geq n} f_k^- \uparrow f^-$$

as $n \rightarrow \infty$. By the theorem

$$\int^* f^- dp = \lim_{n \rightarrow \infty} \int^* \inf_{k \geq n} f_k^- dp \leq \lim_{n \rightarrow \infty} \int^* f_n^- dp \leq \int^* f^- dp,$$

so

$$(A.26) \quad \lim_{n \rightarrow \infty} \int^* f_n^- dp = \int^* f^- dp.$$

Let $A = \{x : f(x) = -\infty\}$. If $p^*(A) = 0$, then (A.21), (A.22), (A.24) and Lemma A.3(b) and (d) imply

$$\int^* f^+ dp \leq \int^* f_n^+ dp \leq 2\varepsilon_n + \int^* f^+ dp < +\infty,$$

so

$$(A.27) \quad \lim_{n \rightarrow \infty} \int^* f_n^+ dp = \int^* f^+ dp < +\infty.$$

Combine (A.26) and (A.27) to conclude (A.25). If $p^*(A) > 0$, then $\int^* f^- dp = -\infty$ and (A.26) will imply (A.25) provided that

$$(A.28) \quad \int^* f^+ dp < +\infty$$

and

$$(A.29) \quad \limsup_{n \rightarrow \infty} \int^* f_n^+ dp < +\infty.$$

Conditions (A.21) and (A.24) imply (A.28). Conditions (A.21), (A.22) and (A.23) imply for every $x \in X$,

$$f_n(x) \leq f_{n-1}(x) + \varepsilon_n, \quad n = 2, 3, \dots,$$

so

$$\int^* f_n^+ dp \leq 2\varepsilon_n + \int^* f_{n-1}^+ dp$$

and

$$\int^* f_n^+ dp \leq 2 \sum_{k=2}^n \varepsilon_k + \int^* f_1^+ dp.$$

The finiteness of $\sum_{k=2}^{\infty} \varepsilon_k$ and (A.24) imply (A.29). Q.E.D.

REFERENCES

- [1] D. P. BERTSEKAS, *Dynamic Programming and Stochastic Control*, Academic Press, New York, 1976.
- [2] D. P. BERTSEKAS AND S. SHREVE, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.
- [3] D. BLACKWELL, *Discounted dynamic programming*, Ann. Math. Statist., 36 (1965), pp. 226–235.
- [4] D. BLACKWELL, D. FREEDMAN AND M. ORKIN, *The optimal reward operator in dynamic programming*, Ann. Probability, 2 (1974), pp. 926–941.
- [5] L. D. BROWN AND R. PURVES, *Measurable selections of extrema*, Ann. Statist., 1 (1973), pp. 902–912.
- [6] C. DELLACHERIE, *Ensembles Analytiques, Capacites, Mesures des Hausdorff*, Springer-Verlag, New York, 1972.
- [7] J. DUNGUNDJI, *Topology*, Allyn and Bacon, Boston, 1966.
- [8] E. B. DYNKIN AND A. A. JUSKEVIC, *Controlled Markov Processes and their Applications*, Moscow, 1975. English translation to be published by Springer-Verlag.
- [9] D. FREEDMAN, *The optimal reward operator in special classes of dynamic programming problems*, Ann. Probability, 2 (1974), pp. 992–994.
- [10] P. R. HALMOS, *Measure Theory*, Van Nostrand, New York, 1950.
- [11] F. HAUSDORFF, *Set Theory*, Chelsea, New York, 1957.
- [12] K. HINDERER, *Foundations of Nonstationary Dynamic Programming with Discrete Time Parameter*, Springer-Verlag, New York, 1970.
- [13] B. JANKOV, *On the uniformisation of A-sets*, Doklady Acad. Nauk. SSSR, 30 (1941), pp. 591–592 (in Russian).
- [14] K. KURATOWSKI, *Topology I*, Academic Press, New York, 1966.
- [15] A. MAITRA, *Discounted dynamic programming on compact metric spaces*, Sankhyā Ser. A, 30 (1968), pp. 211–216.
- [16] P. A. MEYER, *Probability and Potentials*, Blaisdell, Waltham, MA, 1966.
- [17] J. VON NEUMANN, *On rings of operators. Reduction theory*, Ann. Math., (1949), pp. 401–485.
- [18] P. OLSEN, *Multistage stochastic programming with recourse: The equivalent deterministic problem*, this Journal, 14 (1976), pp. 495–517.
- [19] ———, *When is a multistage stochastic programming problem well-defined?* this Journal, 14 (1976), pp. 518–527.
- [20] ———, *Multistage stochastic programming with recourse as mathematical programming in an L_p space*, this Journal, 14 (1976), pp. 528–537.
- [21] K. PARTHASARATHY, *Probability Measures on Metric Spaces*, Academic Press, New York, 1967.
- [22] R. T. ROCKAFELLAR AND R. WETS, *Stochastic convex programming: Relatively complete recourse and induced feasibility*, this Journal, 14 (1976), pp. 574–589.
- [23] ———, *Stochastic convex programming: Basic duality*, Pacific J. Math., 62 (1976), pp. 173–195.
- [24] M. SCHÄL, *Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 32 (1975), pp. 179–196.
- [25] ———, *On dynamic programming: Compactness of the space of policies*, Stochastic Processes Appl., 3 (1975), pp. 345–364.

- [26] E. SELIVANOVSKIĬ, *Ob odnom klasse effektivnykh mnozhestv (mnozhestva C)*, Mat. Sbornik, 35 (1928), pp. 379–413.
- [27] S. SHREVE AND D. P. BERTSEKAS, *A new theoretical framework for finite horizon stochastic control*, Proc. 14th Annual Allerton Conf. on Circuit and System Theory (Univ. of Ill., Urbana, 1976), pp. 336–343.
- [28] ———, *Equivalent stochastic and deterministic optimal control problems*, Proc. 1976 IEEE Conf. on Decision and Control, Institute of Electrical and Electronics Engineers, Inc., New York, 1976, pp. 705–709.
- [29] S. SHREVE, *Dynamic programming in complete separable spaces*, Ph.D. thesis, Dept. of Math., Univ. of Ill., Urbana, IL, 1977.
- [30] ———, *Probability measures and the C -sets of Selivanovskij*, Pacific J. Math., to appear.
- [31] R. E. STRAUCH, *Negative dynamic programming*, Ann. Math. Statist., 37 (1966), pp. 871–890.

EXACT BOUNDARY CONTROL FOR SOME EVOLUTION EQUATIONS*

THOMAS I. SEIDMAN†

Abstract. It is verified that Russell's results [19], [20] on exact boundary control of the wave and heat equations ($\ddot{u} = \Delta \dot{u}$, $\dot{u} = \Delta u$, respectively) can be extended to a class of second order hyperbolic and parabolic equations with spatially variable coefficients. For the diffusion processes a characterization is given of the 'optimal control manifold' $\mathcal{M} \subset L_2([0, T] \times \partial \mathcal{R})$ and a continuity result is obtained for the dependence of the optimal exact control on the coefficients of the equation and the boundary operator.

1. Introduction. After some preliminary abstract material, the structure of this paper falls naturally into two parts: exact null-controllability for certain classes of wave and diffusion processes and, assuming such controllability, a demonstration for diffusion processes of the continuous dependence of the optimal controls. For the first of these parts, consisting of §§ 3 and 4, the principal ideas are largely taken from work of Russell's [19], [20] in the specific context of the wave and heat equations: $\ddot{u} = \Delta u$, $\dot{u} = \Delta u$, although there are some differences of exposition. Essentially, we need only verify, here, that no new considerations are introduced by the generalization and then refer to the literature in scattering theory (e.g., [28]) for relevant results on decay rates. The second part, consisting of §§ 5 and 6, generalizes the discussion in [24], [25] of the dependence of the optimal null-control (for $\dot{u} = \Delta u$) on the length of the available time interval.

In some sense the principal justification for developing the material of the earlier sections, as is done here, is to ensure that the results of the second part are not vacuous, i.e., it is possible to speak of varying the coefficients of the governing diffusion equation and boundary conditions within a reasonably general class for which one has exact controllability. The one new idea which arises is that of "(locally) *uniform* controllability/observability" for a family of systems.

The significance of the continuity result Theorem 6.3 which concludes the paper is perhaps most pointed in the context of (numerical) computation of the optimal exact boundary control (see [29]): no approximate procedure could be considered appropriate without some hold on the variation of the desired result with variation of the parameters of the problem. Indeed, it would seem desirable to investigate the sensitivity of the control of perturbations of the systems by considering the *derivative* of this control with respect to the coefficients σ . The characterization of the optimal control manifold \mathcal{M} given by Theorem 5.3 is used in the proof of Theorem 6.3 but is of independent interest on two grounds. It provides a suggestive result toward a possible theory of vector-valued Dirichlet series generalizing results of [21]. More concretely, it provides a means of obtaining regularity for the optimal control, again a desideratum for computation especially with the use of finite element methods.

2. Abstract preliminaries. We begin by presenting, in rather abstract forms, some results which will subsequently be applied in the more concrete context of the title.

DEFINITION. An *abstract linear control system* (briefly, a *system*) \mathbf{S} denotes specification of

$$(2.1) \quad \mathbf{E} := (\mathbf{E}_x, \mathbf{E}_y): \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{Z}$$

* Received by the editors April 29, 1977, and in revised form February 14, 1978. This research was partly supported by the U.S. Army Research Office under ARO grant DAAG-29-77-G-0061.

† Department of Mathematics, University of Maryland, Baltimore County, Baltimore, Maryland 21228.

where $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ are reflexive Banach spaces and \mathbf{E} is a continuous linear map. Let \mathcal{N} be the nullspace of \mathbf{E} in $\mathcal{X} \times \mathcal{Y}$, let $\mathcal{N}_{\mathcal{Y}}$ be the nullspace of $\mathbf{E}_{\mathcal{Y}}$ in \mathcal{Y} and let \mathcal{D}_C be the projection on \mathcal{X} of \mathcal{N} , i.e.,

$$\mathcal{D}_C := \{\mathbf{x} \in \mathcal{X} : (\mathbf{x}, \mathbf{y}) \in \mathcal{N} \text{ for some } \mathbf{y} \in \mathcal{Y}\} = \mathbf{E}_{\mathcal{X}}^{-1} \mathcal{R}(\mathbf{E}_{\mathcal{Y}}).$$

Let \mathcal{M} in \mathcal{Y}^* be the closure of the range \mathcal{M}_0 of $\mathbf{E}_{\mathcal{Y}}^*: \mathcal{Z}^* \rightarrow \mathcal{Y}^*$

We now give the basic duality theorem; see [4], but also the earlier [22] for the more special context of parabolic boundary control.

THEOREM 2.1 *Assume that the range of $\mathbf{E}_{\mathcal{Y}}$ is dense in \mathcal{Z} . Then the range of \mathbf{E}^* is the graph of a linear map $\mathbf{P}_0: \mathcal{M}_0 \rightarrow \mathcal{X}^*$. The following are equivalent:*

- (a) $\mathcal{D}_C = \mathcal{X}$ (i.e., $\mathcal{R}(\mathbf{E}_{\mathcal{X}}) \subset \mathcal{R}(\mathbf{E}_{\mathcal{Y}})$), for each $\mathbf{x} \in \mathcal{X}$ there is a $\mathbf{y} \in \mathcal{Y}$ such that $\mathbf{E}(\mathbf{x}, \mathbf{y}) = 0$,
- (b) there is a (unique) bounded map $\mathbf{C}: \mathcal{X} \rightarrow \mathcal{Y}_1 := \mathcal{Y}/\mathcal{N}_{\mathcal{Y}}$ such that $\mathbf{E}(\mathbf{x}, \mathbf{y}) = 0$ for $\mathbf{y} \in \mathbf{C}\mathbf{x}$,
- (c) the map \mathbf{P}_0 extends to a bounded map $\mathbf{P}: \mathcal{M} \rightarrow \mathcal{X}^*$.

If (a), (b), (c) hold, then, on identifying \mathcal{M}^* with $\mathcal{Y}_1 := \mathcal{Y}/\mathcal{M}^{\perp}$ one has $\mathbf{C} = -\mathbf{P}^*$. If \mathcal{Y} is a Hilbert space, then \mathcal{Y}_1 is identified with \mathcal{M} and in this case—even if (a), (b), (c) do not hold—whenever \mathbf{z} in \mathcal{Z} is reachable from \mathbf{x} in \mathcal{X} (i.e., when $(\mathbf{z} - \mathbf{E}_{\mathcal{X}}\mathbf{x})$ is in $\mathcal{R}(\mathbf{E}_{\mathcal{Y}})$) the optimal (minimum \mathcal{Y} -norm) control \mathbf{y} is in \mathcal{M} and is uniquely determined by that.

Proof. If $\mathcal{R}(\mathbf{E}_{\mathcal{Y}})$ is dense, then $\mathcal{N}(\mathbf{E}_{\mathcal{Y}}^*) = \{0\}$. Thus, $(\mathbf{x}_1^*, \mathbf{y}_1^*) = \mathbf{E}_{\mathcal{Y}}^* \mathbf{z}_1^*$ and $(\mathbf{x}_2^*, \mathbf{y}_2^*) := \mathbf{E}^* \mathbf{z}_2^*$ with $\mathbf{y}_1^* = \mathbf{y}_2^*$ imply $\mathbf{z}_1^* = \mathbf{z}_2^*$ so $\mathbf{x}_1^* = \mathbf{x}_2^*$ and $\mathbf{P}_0: \mathbf{y}^* \mapsto \mathbf{x}^*$ is well-defined. The linearity of \mathbf{P}_0 follows from that of \mathbf{E} . Now \mathcal{D}_C is the set of $\mathbf{x} \in \mathcal{X}$ for which $\mathbf{E}(\mathbf{x}, \mathbf{y}) = 0$ has a “solution” $\mathbf{y} \in \mathcal{Y}$ so there is always a unique map $\mathbf{C}: \mathcal{D}_C \rightarrow \mathcal{Y}_1$ with $[\mathbf{E}_{\mathcal{X}} + \mathbf{E}_{\mathcal{Y}}^* \mathbf{C}] = 0: \mathcal{X} \rightarrow \mathcal{Z}$. Here $\mathbf{E}_{\mathcal{Y}}^*$ is the factorization of $\mathbf{E}_{\mathcal{Y}}$ through the quotient map: $\mathcal{Y} \rightarrow \mathcal{Y}_1 := \mathcal{Y}/\mathcal{N}(\mathbf{E}_{\mathcal{Y}})$. Note that $\mathbf{E}_{\mathcal{Y}}^*$ is well-defined and bounded since $\mathbf{E}_{\mathcal{Y}}$ is continuous. The graph of \mathbf{C} is closed, being the nullspace of the bounded operator, $\mathbf{E}' := (\mathbf{E}_{\mathcal{X}}, \mathbf{E}_{\mathcal{Y}}^*)$. Thus (a) implies (b) by the closed graph theorem. That (b) implies (a) is trivial. If $\mathbf{x}^* = \mathbf{P}_0 \mathbf{y}^*$ with $\mathbf{y}^* \in \mathcal{M}_0$ (i.e., if $(\mathbf{x}^*, \mathbf{y}^*) := \mathbf{E}^* \mathbf{z}^*$ for some $\mathbf{z}^* \in \mathcal{Z}^*$), then

$$(2.2) \quad \mathbf{x}^* \mathbf{x} + \mathbf{y}^* \mathbf{y} = (\mathbf{E}^* \mathbf{z}^*)(\mathbf{x}, \mathbf{y}) = \mathbf{z}^*(\mathbf{E}_{\mathcal{X}} \mathbf{x} + \mathbf{E}_{\mathcal{Y}} \mathbf{y}).$$

Thus, if \mathbf{x} is in \mathcal{D}_C and \mathbf{y} is in the coset $\mathbf{C}\mathbf{x}$ so $\mathbf{E}_{\mathcal{Y}} \mathbf{y} = -\mathbf{E}_{\mathcal{X}} \mathbf{x}$, one has

$$(\mathbf{P}_0 \mathbf{y}^*) \mathbf{x} + \mathbf{y}^* (\mathbf{C}\mathbf{x}) = \mathbf{z}^* \cdot 0 = 0.$$

Formally, then, $-\mathbf{P}_0$ and \mathbf{C} are adjoint. It follows that if either $-\mathbf{P}_0$ or \mathbf{C} is continuous then so is the other and $\mathbf{C} = -\mathbf{P}^*$ with $\|\mathbf{C}\| = \|\mathbf{P}\|$. The set $\mathcal{G}_{\mathbf{x}, \mathbf{z}}$ of controls \mathbf{y} taking \mathbf{x} to \mathbf{z} , if nonempty, is a translate of the closed subspace $\mathcal{N}(\mathbf{E}_{\mathcal{Y}})$. If \mathcal{Y} is a Hilbert space then the norm is minimized in $\mathcal{G}_{\mathbf{x}, \mathbf{z}}$ for \mathbf{y} orthogonal to $\mathcal{N}(\mathbf{E}_{\mathcal{Y}})$, i.e., for \mathbf{y} in \mathcal{M} since $\mathcal{N}(\mathbf{E}_{\mathcal{Y}}) = \mathcal{R}(\mathbf{E}_{\mathcal{Y}}^*)^{\perp} = \mathcal{M}^{\perp}$. Clearly $\mathcal{G}_{\mathbf{x}, \mathbf{z}} \cap \mathcal{M}$ must then be singleton: the optimal control. \square

DEFINITION. We call \mathbf{S} *nullcontrollable* if (a)—and so also (b), (c)—hold.

DEFINITION. A system \mathbf{S}' will be called an *extension* of a system \mathbf{S} if there is a surjection $f: \mathcal{X}' \rightarrow \mathcal{X}$, a map $h: \mathcal{Z}' \rightarrow \mathcal{Z}$ with $h(0) = 0$ and a map $g: \mathcal{X}' \times \mathcal{Y}' \rightarrow \mathcal{Y}$ such that the diagram:

$$(2.3) \quad \begin{array}{ccc} \mathcal{X}' \times \mathcal{Y}' & \xrightarrow{\mathbf{E}'} & \mathcal{Z}' \\ f \downarrow g & & \downarrow h \\ \mathcal{X} \times \mathcal{Y} & \xrightarrow{\mathbf{E}} & \mathcal{Z} \end{array}$$

commutes, where $\mathcal{X}' \times \mathcal{Y}' \rightarrow \mathcal{X} \times \mathcal{Y}: (\mathbf{x}', \mathbf{y}') \mapsto (f(\mathbf{x}'), g(\mathbf{x}', \mathbf{y}'))$. Note that no requirements of linearity or continuity are imposed on f, g, h , although those would typically obtain in practice.

THEOREM 2.2. If \mathbf{S}' is an extension of \mathbf{S} then $f(\mathfrak{D}'_C) \subset \mathfrak{D}_C$. Thus, since f is surjective, \mathbf{S} is nullcontrollable if it has any nullcontrollable extension.

Proof. Let $\mathbf{x}' \in \mathfrak{D}'_C$. Then there exists $\mathbf{y}' \in Y'$ for which $\mathbf{E}'(\mathbf{x}', \mathbf{y}') = \mathbf{0}$. Let $\mathbf{x} := f(\mathbf{x}')$, $\mathbf{y} := g(\mathbf{x}', \mathbf{y}')$ and note that $\mathbf{E}(\mathbf{x}, \mathbf{y}) \in \mathfrak{Z}$ is $h(\mathbf{E}'(\mathbf{x}', \mathbf{y}')) = h(\mathbf{0}) = \mathbf{0}$ so $\mathbf{x} \in \mathfrak{D}_C$. \square

THEOREM 2.3. Let \mathbf{S} be a system for which there exist maps $f: \mathfrak{X} \rightarrow \mathfrak{X}$ and $g: \mathfrak{X} \rightarrow \mathfrak{Y}$ such that:

$$(2.4) \quad \|g(\mathbf{x})\| \leq K \|\mathbf{x}\|, \quad \|f(\mathbf{x})\| \leq \theta \|\mathbf{x}\| \quad (\theta < 1),$$

$$(2.5) \quad \mathbf{E}_{\mathfrak{X}}[f(\mathbf{x}) - \mathbf{x}] = \mathbf{E}_{\mathfrak{Y}}g(\mathbf{x}).$$

Then \mathbf{S} is nullcontrollable and

$$(2.6) \quad \|\mathbf{C}\| = \|\mathbf{P}\| \leq K/(1 - \theta).$$

Proof. For any $\mathbf{x} = \mathbf{x}_0 \in \mathfrak{X}$, recursively define the sequences: $\mathbf{x}_{k+1} := f(\mathbf{x}_k)$, $\mathbf{y}_k := g(\mathbf{x}_k)$. By induction.

$$\|\mathbf{x}_k\| \leq \theta^k \|\mathbf{x}_0\|, \quad \|\mathbf{y}_k\| \leq K\theta^k \|\mathbf{x}_0\| \quad (k = 0, 1, 2, \dots).$$

Set $\mathbf{y} := \sum \mathbf{y}_k$; the series converges absolutely with $\|\mathbf{y}\| \leq K \|\mathbf{x}_0\|/(1 - \theta)$. Observe that

$$\begin{aligned} \mathbf{E}_{\mathfrak{Y}}\mathbf{y} &= \sum_0^\infty \mathbf{E}_{\mathfrak{Y}}\mathbf{y}_k = \sum \mathbf{E}_{\mathfrak{Y}}g(\mathbf{x}_k) = \sum \mathbf{E}_{\mathfrak{X}}[f(\mathbf{x}_k) - \mathbf{x}_k] \\ &= \sum_0^\infty [\mathbf{E}_{\mathfrak{X}}\mathbf{x}_{k+1} - \mathbf{E}_{\mathfrak{X}}\mathbf{x}_k] = -\mathbf{E}_{\mathfrak{X}}\mathbf{x}_0. \end{aligned}$$

Thus $\mathbf{E}(\mathbf{x}_0, \mathbf{y}) = \mathbf{0}$ so $\mathbf{x}_0 \in \mathfrak{D}_C$. \square

Let \mathfrak{W} be a reflexive complex Banach space and suppose a specified sequence $\Lambda = \{(\lambda_j, \mathbf{w}_j): j = 1, 2, \dots\}$ is given in $\mathbb{R}^+ \times \mathfrak{W}^*$. We will consider \mathfrak{W}^* valued Dirichlet series, i.e., functions given on intervals $[0, T]$ by (finite) sums of the forms:

$$(2.7) \quad \mathbf{f}(t) = \sum_j c_j e^{-\lambda_j t} \mathbf{w}_j$$

and, with $0 < \omega_j := \sqrt{\lambda_j}$,

$$(2.8) \quad \mathbf{g}(t) = \sum_j (c_j^+ e^{i\omega_j t} + c_j^- e^{-i\omega_j t}) \mathbf{w}_j.$$

We introduce the L_q type norm

$$(2.9) \quad \|\mathbf{f}\|_{q,T} = \left(\int_0^T \|\mathbf{f}(t)\|_{\mathfrak{W}^*}^q dt \right)^{1/q}$$

for (strongly measurable) \mathfrak{W}^* -valued functions. It is known (cf., e.g., [5]) that, with \mathfrak{W} reflexive and $1 < q < \infty$, the indicated Banach space $L_q([Q, T] \rightarrow \mathfrak{W}^*)$ is reflexive with dual space $L_p([0, T] \rightarrow \mathfrak{W})$ where $1/p + 1/q = 1$.

THEOREM 2.4. Given $T > 0$, $p > 1$ and k , suppose there is a constant μ such that

$$(2.10) \quad |c_k^\pm| \leq \mu \|\mathbf{g}\|_{q,T} \quad (\text{all } \mathbf{g} \text{ as in (2.8)}).$$

Then, for any $T' > 0$, there is a function Φ in $C_0^\infty(\mathbb{R} \rightarrow \mathfrak{W})$ with support in $[0, T']$ such that

$$(2.11) \quad c_k = \int_0^{T'} \langle \Phi(t), \mathbf{f}(t) \rangle_{\mathfrak{W}} dt \quad (\mathbf{f} \text{ as in (2.7)}).$$

Further, for $j = 0, 1, \dots$ and any $p' \geq 1$, there is a constant K_j —depending on p, T, p', T'

and j but not on \mathfrak{B} , Λ , μ or k —such that

$$(2.12) \quad \|\Phi^{(j)}\|_{p', T'} \leq K_j \mu \exp \left[\frac{1}{2} T' \lambda_k \right].$$

In particular, (2.11) and (2.12) with $j = 0$ give

$$(2.13) \quad |c_k| \leq K_0 \mu \exp \left[\frac{1}{2} T' \lambda_k \right] \|\mathbf{f}\|_{q', T'} \quad (\mathbf{f} \text{ as in (2.7)}).$$

This asserts that continuity of the linear functionals $\gamma_k^\pm: \mathbf{g} \mapsto c_k^\pm$ (for (2.8)) in *some* $L_q(0, T)$ norm implies continuity of $\gamma_k: \mathbf{f} \mapsto c_k$ (for (2.7)) in *every* $L_q(0, T')$ norm (indeed, in every $W^{-j, q'}(0, T')$ norm) with the estimate (2.12). For the proof of the theorem it is convenient to note the following lemma.

LEMMA 2.5. For any σ, τ there is an entire function $R = R_{\sigma, \tau}$ such that

- (a) R is even and of exponential type with $|R(z)| \leq e^{\tau|z|}$,
- (b) on the imaginary axis R is real and $R \geq 1$,
- (c) on the real axis, $R(x) \exp[\sigma|x|^{1/2}]$ is bounded.

Proof. The construction is due to Redheffer [18]: choosing $\alpha > e\sigma^2$ and n so large that $\sum_{k>n} k^{-2} < \tau/\alpha$, set

$$(2.14) \quad R(z) := \prod_n \psi\left(\frac{\alpha z}{k^2}\right) \quad \left[\psi(\zeta) := \frac{\sin \zeta}{\zeta} \right].$$

Note that ψ is even with $\psi \geq 1$ on the imaginary axis and $|\psi(\zeta)| \leq \psi(i|\zeta|) \leq e^{|\zeta|}$ from which (a), (b) follow. Property (c) is obtained in the proof of Lemma 6.2 (with $\beta = 2$) of [6]. \square

Proof. (of Theorem 2.4). We have, from (2.10), that $\|\gamma^+\| \leq \mu$ where γ^+ is here viewed as extended to be a functional on $L_q(\mathbb{R} \rightarrow \mathfrak{B})$ with (2.8) taken to define functions vanishing outside $[0, T]$. Consider the \mathfrak{B} -valued function $\gamma(\cdot)$ on \mathbb{C} defined by

$$\gamma(z): \mathbf{w} \mapsto \langle \gamma^+, (e(z) - e(-z))\mathbf{w} \rangle_{L_q(\mathbb{R} \rightarrow \mathfrak{B})} \quad \mathbf{w} \in \mathfrak{B}^*$$

where $e(z)$ is e^{izt} for t in $[0, T]$ and vanishes outside. Note that $\gamma(\omega_j)\mathbf{w}_j = \delta_{j,k}$ and that analyticity of $\gamma(\cdot)$ follows easily from the continuity of γ^+ (2.10). We also have

$$(2.15) \quad \|\gamma(z)\| \leq 2\mu \left(\frac{1 - e^{-qyT}}{qy} \right)^{1/p} \leq 2\mu T^{1/q} e^{T|z|} \quad (z := x + iy).$$

Choosing $\tau < \frac{1}{2}T'$ and $\sigma > T$, let

$$(2.16) \quad \hat{\Phi}(z) := R_{\sigma, \tau}(z) \gamma(i[z]^{1/2}).$$

Note that $\hat{\Phi}$ is entire (as $\gamma(\cdot)$ is even) and of exponential type with

$$(2.17) \quad \hat{\Phi}(z) = \mathcal{O}(\exp[\tau|z| + T|z|^{1/2}]) = \mathcal{O}(e^{1/2 T'|z|}),$$

by (2.15) and Lemma 2.5(a). On the real axis

$$(2.18) \quad |\hat{\Phi}(x)| \leq 2\mu T^{1/q} |R(x) \exp[\sigma|x|^{1/2}]| \exp[(\sigma - T)|x|^{1/2}]$$

from (2.15). By Lemma 2.5(c), this and (2.17) show that the inverse Fourier transform of $\hat{\Phi}$ is a \mathfrak{B} -valued function Φ_0 with $\Phi_0(t)$ definable, e.g., as a Bochner integral (convergent by (2.13)).

By (2.17) and the Paley–Wiener Theorem [14], Φ_0 vanishes outside $[-\frac{1}{2}T', \frac{1}{2}T']$. It is infinitely smooth with $\|\Phi_0^{(j)}\|_{p'}$ estimable from (2.18)—indeed, the constant K_j of (2.12) can be taken to be $2T^{1/q} \sup\{|R(x)| \exp[\sigma|x|^{1/2}]\}$ times an estimate obtained from $\exp[-\varepsilon|x|^{1/2}]$ with $\varepsilon := (\sigma - T)$ and otherwise depending only on j and p' . We

finally take

$$(2.19) \quad \Phi(t) := \frac{\exp[\frac{1}{2}T'\lambda_k]}{R(i\lambda_k)} \Phi_0(t - \frac{1}{2}T')$$

which has support in $[0, T']$ and satisfies (2.12), using (2.18) as noted and Lemma 2.5(b). Now

$$\begin{aligned} \int_0^{T'} \langle \Phi(t), e^{-\lambda_j t} \mathbf{w}_j \rangle_{\mathfrak{W}} dt &= \frac{\exp[\frac{1}{2}T'\lambda_k] \exp[-\frac{1}{2}T'\lambda_j]}{R(i\lambda_k)} \langle \Phi(i\lambda_j), \mathbf{w}_j \rangle_{\mathfrak{W}} \\ &= \frac{R(i\lambda_j)}{R(i\lambda_k)} \exp[\frac{1}{2}T'(\lambda_k - \lambda_j)] \langle \gamma(\omega_j), \mathbf{w}_j \rangle_{\mathfrak{W}} \\ &= \delta_{j,k}, \end{aligned}$$

which gives (2.11). \square

3. Concrete consequences. We now consider systems in which the maps \mathbf{E} of (2.1) are constructed in terms of partial differential equations of wave or diffusion type. Let \mathcal{R} be a (smoothly) bounded region in \mathbb{R}^n ; for $T > 0$ let $\mathcal{Q} := [0, T] \times \mathcal{R}$ and $\mathcal{S} := [0, T] \times \partial\mathcal{R}$. In $\bar{\mathcal{R}}$ (but extending smoothly to \mathbb{R}^n) specify coefficients $a > 0$, $q \geq 0$ and $A := ((a_{ij}))$ positive definite; on $\partial\mathcal{R}$ specify functions α, β with $\alpha^2 + \beta^2 = 1$. Set

$$(3.1) \quad \mathbf{L}u := \nabla \cdot \mathbf{A}u - qu \quad \text{in } \mathcal{Q},$$

$$(3.2) \quad \mathbf{B}u := \alpha u + \beta u_\nu \quad \text{on } \mathcal{S}$$

where $u_\nu := \partial u / \partial \nu := \mathbf{A}\mathbf{n} \cdot \nabla u$ (\mathbf{n} is the unit outward normal to $\partial\mathcal{R}$).

First consider the diffusion equation

$$(3.3) \quad au_t = \mathbf{L}u$$

with initial/boundary conditions

$$(3.4) \quad u(0, \cdot) = \mathbf{x}, \quad \mathbf{B}u = \mathbf{y}.$$

Here the *initial state* \mathbf{x} is to be taken in a Banach space \mathfrak{X} of functions on \mathcal{R} and the *boundary data* \mathbf{y} in a Banach space \mathfrak{Y} of functions on \mathcal{S} . Quite generally, (3.4) determines a unique solution u of (3.3) and so determines the *terminal state* $\mathbf{z} := u(T, \cdot)$ —i.e., the trace on the “face” $\{t = T\}$ of \mathcal{Q} —in an appropriately chosen space \mathfrak{Z} . We require that $\mathfrak{X}, \mathfrak{Y}, \mathfrak{Z}$ be such that the linear map $\mathbf{E}: (\mathbf{x}, \mathbf{y}) \mapsto \mathbf{z}$ defined in this way by (3.3), (3.4) is continuous from $\mathfrak{X} \times \mathfrak{Y}$ to \mathfrak{Z} . To be able to apply Theorem 2.1 we compute $\mathbf{E}_{\mathfrak{Y}}^*: \mathfrak{Z}^* \rightarrow \mathfrak{Y}^*$. It is convenient to take the $\mathfrak{X} - \mathfrak{X}^*$ and $\mathfrak{Z} - \mathfrak{Z}^*$ dualities in the forms

$$(3.5) \quad \begin{aligned} \langle \xi, \mathbf{x} \rangle_{\mathfrak{X}} &:= \int_{\mathcal{R}} a(0, \cdot) \mathbf{x}(\cdot) \bar{\xi}(\cdot) \quad (\mathbf{x} \in \mathfrak{X}, \xi \in \mathfrak{X}^*), \\ \langle \zeta, \mathbf{z} \rangle_{\mathfrak{Z}} &:= \int_{\mathcal{R}} a(T, \cdot) \mathbf{z}(\cdot) \bar{\zeta}(\cdot) \quad (\mathbf{z} \in \mathfrak{Z}, \zeta \in \mathfrak{Z}^*). \end{aligned}$$

If v is the solution of the *adjoint equation*

$$(3.6) \quad -(av)^{\cdot} = \mathbf{L}v,$$

$$(3.7) \quad v(T, \cdot) = \zeta, \quad \mathbf{B}v = \mathbf{0}$$

with $\zeta \in \mathfrak{Z}^*$, we consider the *complementary data* $\eta := \mathbf{B}'v$ where

$$(3.8) \quad \mathbf{B}'v := \beta v - \alpha v_\nu \quad \text{on } \mathcal{S}.$$

By the divergence theorem, we then have the identity:

$$u_\nu v - uv_\nu \equiv (\mathbf{B}u)(\mathbf{B}'v) - (\mathbf{B}'u)(\mathbf{B}v).$$

Using this identity and (3.3)–(3.7), one then has

$$(3.9) \quad \langle \zeta, \mathbf{E}(\mathbf{x}, \mathbf{y}) \rangle_{\mathfrak{Z}} = \langle \xi, \mathbf{x} \rangle_{\mathfrak{X}} + \int_{\mathcal{S}} \mathbf{y} \bar{\eta}.$$

This gives $\mathbf{E}_{\mathfrak{Y}}^* \zeta \mapsto \eta$ in \mathfrak{Y}^* and, similarly, $\xi := v(0, \cdot) = \mathbf{E}_{\mathfrak{X}}^* \zeta$ in \mathfrak{X}^* . This shows that \mathfrak{M} is the closure in \mathfrak{Y}^* of

$$(3.10) \quad \mathfrak{M}_0 = \{\eta := \mathbf{B}'v : (3.6), (3.7) \text{ with } \zeta \text{ in } \mathfrak{Z}^*\} \subset \mathfrak{Y}^*.$$

Note that the map \mathbf{P} of Theorem 2.1 just determines ξ in \mathfrak{X}^* from ‘observation’ of $\eta := \mathbf{B}'v$ (given (3.6) and $\mathbf{B}v = \mathbf{0}$).

Next, consider the ‘corresponding’ wave equation

$$(3.11) \quad (a\dot{u}) = \mathbf{L}u$$

with the initial/boundary conditions

$$(3.12) \quad u(0, \cdot) = \mathbf{x}_0, \quad \dot{u}(0, \cdot) = \mathbf{x}_1, \quad \mathbf{B}u = \mathbf{y}.$$

Here, the ‘initial state’ $\mathbf{x} := [\mathbf{x}_0, \mathbf{x}_1]$ is to be taken in a Banach space $\hat{\mathfrak{X}} := \hat{\mathfrak{X}}_0 \times \hat{\mathfrak{X}}_1$ of \mathbb{C}^2 -valued functions on \mathcal{R} with, as before, $\mathbf{y} \in \mathfrak{Y}$. Again require that $\hat{\mathfrak{X}}, \mathfrak{Y}, \hat{\mathfrak{Z}}$ be such that the linear map

$$\mathbf{E}: ([\mathbf{x}_0, \mathbf{x}_1], \mathbf{y}) \mapsto \mathbf{z} = [u(T, \cdot), \dot{u}(T, \cdot)] := [\mathbf{z}_0, \mathbf{z}_1] \in \hat{\mathfrak{Z}} := \hat{\mathfrak{Z}}_0 \times \hat{\mathfrak{Z}}_1$$

is continuous from $\hat{\mathfrak{X}} \times \mathfrak{Y}$ to $\hat{\mathfrak{Z}}$. The dualities are now taken in the forms

$$(3.13) \quad \begin{aligned} \langle \xi, \mathbf{x} \rangle_{\hat{\mathfrak{X}}} &:= \int_{\mathcal{R}} a(0, \cdot) [\mathbf{x}_1 \bar{\xi}_0 - \mathbf{x}_0 \bar{\xi}_1] & (\mathbf{x} \in \hat{\mathfrak{X}}, \xi \in \hat{\mathfrak{X}}^*), \\ \langle \zeta, \mathbf{z} \rangle_{\hat{\mathfrak{Z}}} &:= \int_{\mathcal{R}} a(T, \cdot) [\mathbf{z}_1 \bar{\zeta}_0 - \mathbf{z}_0 \bar{\zeta}_1] & (\mathbf{z} \in \hat{\mathfrak{Z}}, \zeta \in \hat{\mathfrak{Z}}^*). \end{aligned}$$

If \hat{v} is a solution of (3.11) with

$$(3.14) \quad \hat{v}(T, \cdot) = \zeta_0, \quad \dot{\hat{v}}(T, \cdot) = \zeta_1, \quad \mathbf{B}\hat{v} = \mathbf{0}$$

and $\eta := \mathbf{B}'\hat{v}$, one again has the same identity (3.1)—with, of course, a new interpretation. Again $\mathbf{E}_{\mathfrak{Y}} \zeta = \eta$, $\mathbf{E}_{\hat{\mathfrak{X}}}^* \zeta = \xi = [\hat{v}(0, \cdot), \dot{\hat{v}}(0, \cdot)]$ and $\mathbf{P}: \eta \mapsto \xi$ in terms of (3.11), (3.14).

We shall be considering control spaces which restrict the support of the control \mathbf{y} . Let $\partial_a \mathcal{R}$ be a specified *active* portion of the boundary ($\partial_a \mathcal{R} \subset \partial \mathcal{R}$), and let the *passive* portion of the boundary be the complement, $\partial_\# \mathcal{R} := \partial \mathcal{R} \setminus \partial_a \mathcal{R}$. We impose constraints by taking \mathfrak{Y} to be the space of (generalized) functions on $\mathcal{S}_a := [0, T] \times \partial_a \mathcal{R}$, extended to be zero on the passive portion $\mathcal{S}_\# := \mathcal{S} \setminus \mathcal{S}_a = [0, T] \times \partial_\# \mathcal{R}$. In this case the dual space \mathfrak{Y}^* will also consist of functions on \mathcal{S}_a . Thus \mathbf{B}' is given by the restriction of (3.8) to \mathcal{S}_a and the integral on the right of (3.9) is taken over \mathcal{S}_a .

In the present context Theorem 2.1 takes the form:

THEOREM 3.1. *Let \mathbf{S} be a system given by (3.3), (3.4) using $\mathfrak{X}, \mathfrak{Y}, \mathfrak{Z}$ (respectively, $\hat{\mathfrak{S}}$ a system given by (3.11), (3.12) using $\hat{\mathfrak{X}}, \mathfrak{Y}, \hat{\mathfrak{Z}}$). Then a nullcontrol $\mathbf{y} \in \mathfrak{Y}$ exists for*

each $\mathbf{x} \in \mathfrak{X}$ (i.e., for each initial state \mathbf{x} there is a \mathbf{y} such that the solution u of (3.3), (3.4) has final state $u(T, \cdot) = \mathbf{0}$) if and only if there is a constant $C (= \|\mathbf{P}\|)$ such that

$$(3.15a) \quad \|v(0, \cdot)\|_{\mathfrak{X}^*} \leq C \|\mathbf{B}'v\|_{\mathfrak{Y}^*}$$

for every solution v of (3.6), (3.7). (Similarly, $\hat{\mathbf{S}}$ is nullcontrollable if and only if

$$(3.15b) \quad \|[\hat{v}(0, \cdot), \hat{v}'(0, \cdot)]\|_{\hat{\mathfrak{X}}^*} \leq C \|\mathbf{B}'\hat{v}\|_{\hat{\mathfrak{Y}}^*}$$

for some $C (= \|\hat{\mathbf{P}}\|)$ and every solution \hat{v} of (3.11), (3.14)). In this case the nullcontrol \mathbf{y} may be chosen such that $\|\mathbf{y}\| \leq C \|\mathbf{x}\|$. \square

We now focus attention on the autonomous case—all the coefficients a, A, q, α, β independent of t —for which, following Russell [19], [20] (but see, also, Triggiani [27]), we wish to show that nullcontrollability for (3.11) implies that for (3.3). The falsity of the converse is known already for the heat equation [25].

Introduce the eigenpairs $\{(-\lambda_j, e_j(\cdot))\}$ and the corresponding complementary boundary data $\{w_j(\cdot)\}$ given by

$$(3.16) \quad \mathbf{L}e_j = -\lambda_j a e_j, \quad \mathbf{B}e_j = \mathbf{0}; \quad w_j := \mathbf{B}'e_j.$$

We have $\lambda_j \geq 0$ and $\{e_j\}$ orthonormal in $L_{2,a}(\mathcal{R})$ (i.e., using an a -weighted L_2 -norm for functions on \mathcal{R}). The $\{w_j\}$ are (generalized) functions on $\partial_a \mathcal{R}$, extended as zero on $\partial_{\neq} \mathcal{R}$. Note that e_j, w_j will be smooth under suitable smoothness assumptions on the coefficients and on $\partial \mathcal{R}$. The (formal) general solutions v, \hat{v} of (3.4), (3.11), respectively, are now given by the series

$$(3.17) \quad v(t, \cdot) = \sum_j c_j e^{-\lambda_j(T'-t)} e_j(\cdot),$$

$$(3.18) \quad \hat{v}(t, \cdot) = \sum_j [c_j^+ e^{i\omega_j t} + c_j^- e^{-i\omega_j t}] e_j(\cdot)$$

with $\omega_j^2 = \lambda_j$. We set

$$(3.19) \quad \mathbf{f}(t) := \mathbf{B}'v(T'-t, \cdot), \quad \mathbf{g}(t) := \mathbf{B}'\hat{v}(t, \cdot)$$

for solutions v, \hat{v} of (3.4), (3.11), respectively. Note that, from (3.17), (3.18), these are given by (2.7), (2.8). In each case the functions $w_j(\cdot)$ are viewed as elements \mathbf{w}_j of a reflexive complex Banach space \mathfrak{W}^* . Thus \mathbf{f} and \mathbf{g} are to be viewed as elements of $\mathfrak{Y}^* := L_q([0, T] \rightarrow \mathfrak{W}^*)$ and $\hat{\mathfrak{Y}}^* := L_q([0, T] \rightarrow \mathfrak{W}^*)$, and we take

$$(3.20) \quad \mathfrak{Y} := L_p([0, T'] \rightarrow \mathfrak{W}), \quad \hat{\mathfrak{Y}} := L_p([0, T] \rightarrow \mathfrak{W}).$$

THEOREM 3.2. *Let $\hat{\mathbf{S}}$ be a hyperbolic system given by (3.11), (3.12) using $\hat{\mathfrak{X}}, \hat{\mathfrak{Y}}, \hat{\mathfrak{Z}}$. Suppose $\hat{\mathfrak{X}}_0^*, \hat{\mathfrak{X}}_1^*$ each contain $\{e_k\}$ and the coefficients of (3.18) satisfy*

$$(3.21) \quad |c_k^\pm| \leq \hat{C} \|\xi\|_{\hat{\mathfrak{X}}^*} \quad (\xi = [\hat{v}(0, \cdot), \hat{v}'(0, \cdot)] \in \hat{\mathfrak{X}}^*)$$

for each k . Suppose $\hat{\mathfrak{Y}}$ has the form (3.20) for some $p > 1$ and some reflexive complex Banach space \mathfrak{W} for which \mathfrak{W}^* contains $\{\mathbf{w}_j := \mathbf{B}'e_j\}$. Let \mathbf{S} be a parabolic system given by (3.3), (3.4) using $\mathfrak{X}, \mathfrak{Y}, \mathfrak{Z}$. Suppose \mathfrak{X}^* contains $\{e_j\}$ and \mathfrak{Y} has the form (3.20) for some $p' > 1$ and with the same space \mathfrak{W} as for $\hat{\mathfrak{Y}}$. Finally, suppose there exists $T' > 0$ such that

$$(3.22) \quad \sum_k e^{-(1/2)T'\lambda_k} \|e_k\|_{\mathfrak{X}^*} =: C' < \infty.$$

Then if $\hat{\mathbf{S}}$ is nullcontrollable, so is \mathbf{S} .

Proof. By Theorem 3.1, nullcontrollability of $\hat{\mathbf{S}}$ implies (3.15) $^\wedge$; i.e., $\|\xi\|_{\hat{\mathcal{X}}^*} \leq \|\hat{\mathbf{P}}\|_g \|g\|_{\mathcal{Y}^*}$. Combining this with (3.21) gives (2.10) with $\mu := \hat{C} \|\hat{\mathbf{P}}\|$ and we may apply Theorem 2.4 to obtain (2.13). Since $\mathbf{B}'v(t, \cdot) = \mathbf{f}(T' - t)$ we have $\|\mathbf{B}'v\|_{\mathcal{Y}^*} = \|\mathbf{f}\|_{q', T'}$. Combining this with (2.13) and (3.17) gives

$$\begin{aligned} \|v(0, \cdot)\|_{\mathcal{X}^*} &\leq \sum_k |c_k| e^{-T'\lambda_k} \|e_k\|_{\mathcal{X}^*} \\ &\leq \sum_k (K_0 \hat{C} \|\hat{\mathbf{P}}\| e^{(1/2)T'\lambda_k} \|\mathbf{B}'v\|_{\mathcal{Y}^*}) e^{-T'\lambda_k} \|e_k\|_{\mathcal{X}^*} \\ &= (K_0 \hat{C} \|\hat{\mathbf{P}}\| C') \|\mathbf{B}'v\|_{\mathcal{Y}^*} \end{aligned}$$

which has the form of (3.15). Another application of Theorem 3.1, now in the reverse direction, completes the proof that \mathbf{S} is nullcontrollable. \square

Remark 3.3. Note that the proof above shows that

$$(3.23) \quad \|\mathbf{P}\| \leq K_0 \hat{C} C' \|\hat{\mathbf{P}}\|$$

where \hat{C}, C' are given by the conditions (3.21), (3.22), and K_0 depends only on the choices of p, T (such that $\hat{\mathbf{S}}$ is nullcontrollable) and p', T' . Thus, suppose $\{\hat{\mathbf{S}}\}$ is a *family* of such settings (e.g., with varying conditions: region \mathcal{R} , coefficients of \mathbf{L} or of \mathbf{B} , specification of $\partial_\omega \mathcal{R}$, etc.) with the same p, T used in defining the $\{\mathcal{Y}\}$ and the spaces $\{\hat{\mathcal{X}}\}$ and suppose that the same bound \hat{C} may be used in (3.21) for each. Suppose, also, that the corresponding settings $\{\mathbf{S}\}$ use the same p', T' throughout in defining the $\{Y\}$ and that a fixed C' can be used in (3.22). Suppose the family $\{\hat{\mathbf{S}}\}$ is *uniformly nullcontrollable* (i.e., a uniform bound on $\{\|\hat{\mathbf{P}}\|\}$, corresponding to a requirement that the nullcontrol \mathbf{y} for each initial state \mathbf{x} and each $\hat{\mathbf{S}}$ in the family can be so chosen as to have $\{\|\mathbf{y}\|/\|\mathbf{x}\|\}$ uniformly bounded). Then we conclude from (3.23) that the corresponding family $\{\mathbf{S}\}$ is also uniformly nullcontrollable. \square

Remark 3.4. The elements $\{e_k\}$ are orthonormal in $L_{2,a}(\mathcal{R})$ so $\{\|e_k\|_{L_2(\mathcal{R})}\}$ is bounded. For \mathbf{L} given by (3.1) with $\mathbf{B}v = \mathbf{0}$ (using (3.2)) one has (cf., e.g., [3]):

$$(3.24) \quad \lambda_k \sim Ck^{2/n}.$$

With the choice $\mathcal{X} := L_2(\mathcal{R})$ (or any ‘stronger’ space so \mathcal{X}^* norms would be smaller), it follows that (3.22) holds for *any* $T' > 0$. While the constant C in (3.24) depends on $\mathcal{R}, \mathbf{L}, \mathbf{B}$, one has local uniformity in these for (3.22) under quite mild conditions.

From (3.18) we have $\xi := [\xi_0, \xi_1]$ given by

$$\xi_0 = \sum_i (c_i^+ + c_i^-) \mathbf{e}_i, \quad \xi_1 = \sum_i i\omega_i (c_i^+ - c_i^-) \mathbf{e}_i.$$

Setting $\mathbf{q}_k^\pm := \frac{1}{2}[(\pm i/\omega_k)\mathbf{e}_k, \mathbf{e}_k]$ and recalling (3.13), one has $c_k^\pm = \langle \xi, \mathbf{q}_k^\pm \rangle_{\mathcal{X}^*}$ and \hat{C} in (3.21) is any bound for $\{\|\mathbf{q}_k^\pm\|_{\mathcal{X}^*}\}$. Suppose one were to use, for $\hat{\mathcal{X}}_0$ and $\hat{\mathcal{X}}_1$, the norms

$$(3.25) \quad \begin{aligned} \|x_0\|_{\hat{\mathcal{X}}_0}^2 &:= \int_{\mathcal{R}} [\nabla x_0 \cdot A \nabla x_0 + q|x_0|^2], \\ \|x_1\|_{\hat{\mathcal{X}}_1}^2 &:= \int_{\mathcal{R}} a|x_1|^2. \end{aligned}$$

Then

$$\|\mathbf{q}_k^\pm\|_{\hat{\mathcal{X}}} = \frac{1}{2}(\|\mathbf{e}_k\|_{\hat{\mathcal{X}}_0}^2/\lambda_k + \|\mathbf{e}_k\|_{\hat{\mathcal{X}}_1}^2)^{1/2} = \frac{1}{2}\sqrt{2},$$

and we may take $\hat{C} = 1/\sqrt{2}$. Note that (3.25) just gives the *energy norm* for $\hat{\mathcal{X}}: \|\xi\|_{\hat{\mathcal{X}}}^2 =$

$\mathcal{E}(v; 0, \mathcal{R})$ where

$$(3.26) \quad \mathcal{E}(v, t, \mathcal{R}) = \int_{\mathcal{R}} [\nabla v \cdot A \nabla v + q|v|^2 + a|\dot{v}|^2]_t.$$

It is easily seen that if $q \neq 0$ the norm above for $\hat{\mathcal{X}}_0$ is equivalent to that of $H^1(\mathcal{R}) = W^{1,2}(\mathcal{R})$; if $q \equiv 0$, the ‘norm’ given for $\hat{\mathcal{X}}_0$ in (3.25) vanishes for constants and must, in any case, be modified (e.g., in a fashion relevant to the boundary conditions, especially if $\alpha \equiv 0$). One may take $\hat{\mathcal{X}} := H^1(\mathcal{R}) \times H^0(\mathcal{R})$. The norm equivalence is such that one still has local uniformity of (3.21) under quite mild conditions. \square

Remark 3.5. No conditions are imposed on $\mathfrak{J}, \hat{\mathfrak{J}}$ in either Theorem 3.1 or 3.2—except the implicit requirement that the terminal state depend continuously on the initial state and boundary control. The specific choice is essentially irrelevant to the observation problem ($\eta \mapsto \xi$) and, for the control problem, may be taken as indicating the sense in which the terminal constraint is imposed. \square

4. Scattering theory and control of wave equations. The aim of this section is the application of Theorem 2.3 to obtain controllability results for certain hyperbolic equations (3.11), (3.12) in energy norm (3.26). (Later, by Theorem 3.2, this will imply controllability results for corresponding parabolic equations (3.3), (3.4).) An essentially similar argument was developed by Russell [19], [20] for the wave equation $\ddot{u} = \Delta u$. The principal innovation here, other than in exposition, is that the reference to [12] is replaced by reference to Zachmanoglou’s 1966 paper [28]. This enables us to obtain a class of equations with spatially variable coefficients for which one has controllability. Indeed, the controllability is (locally) uniform in its dependence on the coefficients—this uniformity being relevant to the arguments of Theorems 5.3 and 6.4 below.

Suppose we are given a region $\mathcal{R} \subset \mathbb{R}^n$ and coefficients as in the preceding section but *autonomous*. (As the available results (e.g., [28]) are primarily for Dirichlet conditions, we will work only with that case but note that suitable generalizations of [28]—cf. Remark 4.6—would lead to corresponding generalizations of our arguments here.) Now introduce an ‘obstacle’ Ω in the complement of \mathcal{R} , though possibly sharing some boundary, and assume that the coefficients a, A, q are appropriately extended to all of $\hat{\mathcal{R}} := \mathbb{R}^n \setminus \Omega$. We consider *finite energy solutions* u of

$$(4.1) \quad a\ddot{u} = \mathbf{L}u := \nabla \cdot A \nabla u - qu \quad \text{in } \hat{\mathcal{Q}} := \mathbb{R}_+ \times \hat{\mathcal{R}},$$

$$(4.2) \quad u|_{\hat{\mathcal{S}}} = 0 \quad \text{where } \hat{\mathcal{S}} := \mathbb{R}_+ \times \partial\hat{\mathcal{R}}.$$

These are solutions for which

$$(4.3) \quad \mathcal{E}(u) := \mathcal{E}(u, t, \hat{\mathcal{R}}) := \int_{\hat{\mathcal{R}}} [\nabla u \cdot A \nabla u + q|u|^2 + a|\dot{u}|^2]$$

is finite. Note that \mathcal{E} is independent of t for solutions u of (4.1), (4.2).

DEFINITION. For $\mathcal{R} \subset \hat{\mathcal{R}}$, a *decay rate* for \mathcal{R} is a function $\zeta: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ such that

$$(4.4) \quad \mathcal{E}(u, t, \mathcal{R}) \leq \zeta(t) \mathcal{E}(u) = \zeta(t) \mathcal{E}(u, 0, \mathcal{R}_1)$$

for every solution u of (4.1), (4.2) such that $u(0, \cdot), \dot{u}(0, \cdot)$ vanish outside some arbitrary but fixed bounded set \mathcal{R}_1 with $\mathcal{R} \subset \mathcal{R}_1 \subset \hat{\mathcal{R}}$.

In connection with the choice of \mathcal{R}_1 in this definition we introduce the space $\mathcal{X}_{0*} := \{x \in H^1(\hat{\mathcal{R}}) : x|_{\hat{\mathcal{R}} \setminus \mathcal{R}_1} = 0, x|_{\partial\Omega} = 0\}$ with the norm given as in (3.25), by $[\int_{\mathcal{R}_1} \nabla x \cdot A \nabla x + q|x|^2]^{1/2}$. We henceforth assume the following technical condition:

(G₁) There is a bounded embedding (extension map) $\epsilon_0: \mathfrak{X}_0 \rightarrow \mathfrak{X}_{0*}$, i.e., each function x_0 in $\mathfrak{X}_0 := \{x \in H^1(\mathcal{R}): x|_{\partial\mathcal{R} \cap \partial\Omega} = 0\}$ (the norm of \mathfrak{X}_0 is given by (3.25)) is the restriction to \mathcal{R} of a function in \mathfrak{X}_{0*} .

(We remark that a sufficient condition on \mathcal{R}, Ω for the existence of \mathcal{R}_1 satisfying (G₁) is that $\partial\mathcal{R}$ and $\partial\Omega$ be piecewise C^1 with $\overline{\partial\mathcal{R} \cup \Omega}$ satisfying a cone condition. Note also that, since the norm of \mathfrak{X}_{0*} is equivalent to that of $H^1(\hat{\mathcal{R}})$ for functions with fixed bounded support, the existence (and choices) of \mathcal{R}_1 and ϵ_0 as in (G₁) are of purely geometric significance, independent of the particular coefficients, although the number $\|\epsilon_0\|$ will, of course depend on the particular norm.)

DEFINITION. We say that *waves are not trapped in \mathcal{R}* if there is *some* choice of Ω, \mathcal{R}_1 and the extensions to $\hat{\mathcal{R}}$ of the coefficients a, A, q for which there is a decay rate with $\zeta(t) \rightarrow 0$ as $t \rightarrow \infty$.

THEOREM 4.1. *Let (3.11) be autonomous and suppose waves are not trapped in \mathcal{R} . Let $\partial_{\mathcal{R}}\mathcal{R} = \partial\mathcal{R} \cap \partial\Omega$ so $\partial_a\mathcal{R} = \partial\mathcal{R} \setminus \partial\hat{\mathcal{R}}$. Then there are positive T, K_1 (depending only on $\mathcal{R}, \mathcal{R}_1, \|\epsilon_0\|$, the decay rate ζ , and the norm equivalence of \mathfrak{X}_0 with $H^1(\mathcal{R})$) such that \mathbf{S} is nullcontrollable with $\|\mathbf{P}\| \leq K_1$. Here \mathbf{S} is the system defined by (3.11), (3.12) with $\mathfrak{X} = \mathfrak{X}_0 \times \mathfrak{X}_1$, \mathfrak{X}_0 as in (G₁), $\mathfrak{X}_1 := L_{2,a}(\mathcal{R})$, $\mathfrak{Z} = \mathfrak{X}$, and $\mathfrak{Y} = L_2([0, T] \times \partial_a\mathcal{R})$.*

Proof. Let ϵ_1 be the obvious embedding of $L_{2,a}(\mathcal{R})$ in $L_{2,a}(\hat{\mathcal{R}})$: $\epsilon_1\mathbf{x}_1$ vanishes on $\hat{\mathcal{R}} \setminus \mathcal{R}$ and coincides with \mathbf{x}_1 on \mathcal{R} . Let $\epsilon := \epsilon_0 \times \epsilon_1: \mathfrak{X} \rightarrow \mathfrak{X}_{0*} \times L_2(\hat{\mathcal{R}})$ and choose $T > 0$ such that

$$(4.5) \quad \zeta(T)\|\epsilon\|^2 =: \theta < 1.$$

For $\mathbf{x} = [\mathbf{x}_0, \mathbf{x}_1]$ in \mathfrak{X} , let u be the solution in $\hat{\mathcal{Q}}_T := [0, T] \times \hat{\mathcal{R}}$ of (4.1), (4.2) with the initial conditions

$$(4.6) \quad u(0, \cdot) = \epsilon_0\mathbf{x}_0, \quad \dot{u}(0, \cdot) = \epsilon_1\mathbf{x}_1.$$

Next, let \mathbf{x}^* be the restriction to \mathcal{R} of $[u(T, \cdot), \dot{u}(T, \cdot)]$ and let \mathbf{y}^* be the trace of u on $\mathcal{S} := [0, T] \times \partial\mathcal{R}$, i.e., $\mathbf{y}^*(t, \cdot) = u(t, \cdot)|_{\partial\mathcal{R}}$. Note that \mathbf{x}^* will be in $\mathfrak{Z} = \mathfrak{X}$ and also that the map $\mathbf{x} \mapsto u \mapsto \mathbf{y}^* \in L_2(\mathcal{S})$ is continuous (indeed, \mathbf{y}^* will be in $L_2([0, T] \rightarrow H^{1/2}(\partial\mathcal{R}))$) with \mathbf{y}^* vanishing on $\mathcal{S}_{\mathcal{R}} := [0, T] \times \partial_{\mathcal{R}}\mathcal{R}$ so \mathbf{y}^* is in \mathfrak{Y} . Similarly, with “reversed time” obtain a new solution u^* in $\hat{\mathcal{Q}}_T$ of (4.1), (4.2) with data $\epsilon\mathbf{x}^*$ at $t = T$. Then let \mathbf{x}_* be the restriction to \mathcal{R} of $[u^*(0, \cdot), \dot{u}^*(0, \cdot)]$ and let \mathbf{y}_* be the trace of u^* on \mathcal{S} .

For application of Theorem 2.3, set $f(\mathbf{x}) := \mathbf{x}_*$ and $g(\mathbf{x}) = \mathbf{y}^* - \mathbf{y}_*$. Note that time-reversibility of the autonomous equations (4.1), (4.2) gives

$$(4.7) \quad \mathcal{E}(u^*, t, \mathcal{R}) \leq \zeta(T-t)\hat{\mathcal{E}}(u^*) = \zeta(T-t)\hat{\mathcal{E}}(\epsilon\mathbf{x}^*).$$

Thus,

$$\begin{aligned} \|\mathbf{x}^*\|^2 &= \mathcal{E}(u, T, \mathcal{R}) \leq \zeta(T)\hat{\mathcal{E}}(u) = \zeta(T)\hat{\mathcal{E}}(\epsilon\mathbf{x}) \leq \theta\|\mathbf{x}\|^2, \\ \|\mathbf{x}_*\|^2 &= \mathcal{E}(u^*, 0, \mathcal{R}) \leq \zeta(T)\hat{\mathcal{E}}(\epsilon\mathbf{x}^*) \leq \theta^2\|\mathbf{x}\|^2, \\ \|\mathbf{y}^*\|^2 &\leq \|\tau\| \left[\int_0^T \zeta \right]^{1/2} \|\epsilon\| \|\mathbf{x}\|, \quad \|\mathbf{y}_*\| \leq \|\tau\| \left[\int_0^T \zeta \right]^{1/2} \|\epsilon\| \|\mathbf{x}^*\| \end{aligned}$$

where τ is the trace operator— $\tau u := u|_{\mathcal{S}}$. This gives (2.4). By uniqueness for (3.11), (3.12) one has

$$\mathbf{E}_{\mathfrak{X}}\mathbf{x} + \mathbf{E}_{\mathfrak{Y}}\mathbf{y} =: \mathbf{x}^* = \mathbf{E}_{\mathfrak{X}}\mathbf{x}_* + \mathbf{E}_{\mathfrak{Y}}\mathbf{y}_*$$

which gives (2.5). Nullcontrollability of \mathbf{S} is then given by Theorem 2.3 with

$$(4.8) \quad \|\mathbf{P}\| \leq \|\tau\| \|\epsilon\| \left[\int_0^T \zeta \right]^{1/2} \frac{1 + \theta^{1/2}}{1 - \theta}.$$

Observe that $\|\tau\|$ and $\|\epsilon\|$ depend on the coefficients through the use of the norms given by (3.25). For any class of problems for which the norm-equivalences of \mathfrak{X}_0 with $H^1(\mathcal{R})$, \mathfrak{X}_1 with $L_2(\mathcal{R})$, etc., is uniform and for which a decay rate can be chosen uniformly, one may choose T uniformly (so, e.g., $\theta = \frac{1}{4}$) and then there will be a uniform bound K_1 on the right hand side of (4.8). \square

Remark 4.2. Note that the control \mathbf{y} constructed by Theorem 2.3 (*not*, in general, the minimum norm control) is given by the series $\sum \mathbf{y}_k$ with $\mathbf{y}_k := g(\mathbf{x}_k)$, $\mathbf{x}_{k+1} := f(\mathbf{x}_k)$, $\mathbf{x}_0 := \mathbf{x}$. Thus, each step is associated with a solution \hat{u}_k of (3.11) and (3.12) using $\mathbf{y}_k (= \mathbf{y}_k^* - \mathbf{y}_{*k})$. Thus $\hat{u}_k := u_k - u_k^*$ where u_k is the restriction to \mathcal{R} of the solution of (4.1), (4.2) with data $\epsilon \mathbf{x}_k$ at $t=0$ and, similarly, u_k^* is the restriction of the solution with data $\epsilon \mathbf{x}_k^*$ at $t=T$. From (4.4), (4.7) we have ($0 \leq t \leq T$)

$$\mathcal{E}(\hat{u}_k, t, \mathcal{R}) \leq \zeta(t) \|\epsilon \mathbf{x}_k\|^2 + \zeta(T-t) \|\epsilon \mathbf{x}_k^*\|^2.$$

Hence, as $\|\mathbf{x}_k\| \leq \theta^k \|\mathbf{x}\|$, $\|\mathbf{x}_k^*\| \leq \theta^{1/2} \|\mathbf{x}_k\|$, one has

$$\mathcal{E}(u, t, \mathcal{R}) \leq \|\epsilon\| [\zeta(t) + \theta^{1/2} \zeta(T-t)] \|\mathbf{x}\| / (1 - \theta),$$

where u is the solution of the controlled problem: (3.11), (3.12) using \mathbf{y} . This choice of control gives a ‘state trajectory’ with finite energy at the intermediate times: $0 < t < T$; cf., the comment following Theorem 1.1 of [8] \square

Remark 4.3. The use of $\mathcal{Y} = L_2(\mathcal{S}_a)$ for $\mathfrak{X} = \mathfrak{J}$ as here depends on the use of Dirichlet boundary conditions so that the relevant trace τ is continuous where needed. If the boundary conditions involved u_ν one would have to take, e.g., $\mathcal{Y} := L_2([0, T] \rightarrow \mathfrak{B})$ with $\mathfrak{B} := H^{-1/2}(\partial_a \mathcal{R})$ to employ essentially the same argument as here. Russell, [19], [20], treated such conditions ($u = 0$ on \mathcal{S}_a , u_ν controlled on \mathcal{S}_a) with $\mathcal{Y} := L_2(\mathcal{S}_a)$ but was then forced to work with smoother (initial) states: $u(0, \cdot)$ in $H^2(\mathcal{R})$, $\dot{u}(0, \cdot)$ in $H^1(\mathcal{R})$.

We also remark that the argument presented here remains valid in the *time-dependent* case provided one has *both* a forward *and* a backward decay rate on $[0, T]$, i.e., ζ for which $\zeta(T)\|\epsilon\| =: \theta < 1$ and *both* (4.1) and (4.7). As an example of such results, see [2]. \square

We now wish to identify, following [28], a class of problems for which the ‘non-trapping’ hypothesis of Theorem 4.1 is known to be satisfied. To this end, we introduce besides (G_1) , a number of conditions on the geometry and on the coefficients, as extended to \mathcal{R} .

- (*) There exists a constant γ such that any finite energy solution u of (4.1), (4.2) with initial data vanishing outside \mathcal{R}_1 satisfies

$$\left| \int_{\mathcal{R}} a u \dot{u} \right| \leq \gamma \mathcal{E}(u) \quad (t > 0).$$

It is observed in [28] that the condition (*) is known to follow from the conditions (G) , (C_a) , (C_g) below—at least for $n \geq 3$ and moderately smooth coefficients: A in $C^2(\mathcal{R})$ and q in $C^1(\mathcal{R})$. The argument given there rests on the existence, for each \mathbf{x}_1 in $L_2(\mathcal{R})$ vanishing outside \mathcal{R}_1 , of a solution h with $h|_{\partial\Omega} = 0$ and $\int_{\mathcal{R}} [\nabla h \cdot A \nabla h + q|h|^2] < \infty$ of the elliptic equation: $\nabla \cdot A \nabla h - qh = a x_1$.

The significant conditions which we now state, are (G) on the geometry and (C_a) , (C_g) on the coefficients: (C_g) represents *global* conditions on a, A, q while (C_a) represents *asymptotic* conditions ($r := |r| \rightarrow \infty$).

- (G) The pair $\mathcal{R}, \partial_a \mathcal{R}$ is *star-complemented*: i.e., there exists an obstacle Ω as above such that (G_1) holds and such that (shifting the origin appropriately into Ω) $\mathbf{r} \cdot \mathbf{n} \leq 0$ on $\partial\Omega = \partial \hat{\mathcal{R}}$ (note that \mathbf{n} is the normal out of $\hat{\mathcal{R}}$, into Ω).

Having chosen Ω , \mathcal{R}_1 as in the geometric conditions (G), (G_1) , we now assume the coefficients a , A , q —initially only defined in \mathcal{R} —can be extended to all of $\hat{\mathcal{R}} := \mathbb{R}^n \setminus \bar{\Omega}$ so as to satisfy $a > 0$, $A > 0$, $q \geq 0$ and also:

(C_g) There exists $\alpha = \alpha(r)$ with $0 < \alpha(r) < \min\{\alpha_0, r^{-1-\delta}\} < 1$ for which

$$(a) \quad -\mathbf{r} \cdot \nabla a \leq \alpha a,$$

$$(b) \quad \mathbf{r} \cdot \nabla A \leq \alpha A,$$

$$(c) \quad (2 - \alpha)q \leq -\mathbf{r} \cdot \nabla q.$$

THEOREM 4.4. *Let the conditions (G), (C_a), (C_g) and (*) be satisfied. Then for $0 < \varepsilon < 1 - \alpha_0$ there is a constant \hat{c} , depending only on ε , n , r_0 , α_0 , δ , γ , $a_1 := \max\{a\}$ and $a_0 := \max\{\|A^{-1}\|\}$, such that $\zeta(t) \leq \min\{1, \hat{c}t^{-\varepsilon}\}$ in (4.4). \square*

Remark 4.5. The theorem is stated in [28] only for *smooth* solutions u of (4.1), (4.2) but, obviously, extends to all finite energy solutions by continuity in $L_\infty([0, T] \rightarrow \mathcal{X})$.

The *asymptotic* condition (C_a) is only mildly restrictive for our purposes as rather general coefficients in \mathcal{R} can be so extended (retaining (C_g)) to $\hat{\mathcal{R}}$ provided they are ‘fairly flat’ with A/a ‘not far’ from 1 in \mathcal{R} . One must be careful to preserve the (local) uniformity of the norm-equivalence $(\mathcal{X}_0(\hat{\mathcal{R}}))$ with $H^1 \times H^0$, of a_1 and a_0 , and of \mathcal{R}_1 and γ .

The most restrictive conditions are (G) and (C_g). (Note that (G) may, on occasion, be evaded through the use of a spatial transformation, such as inversion with respect to a point outside $\bar{\mathcal{R}}$, but this is likely to lead to compensating “infringement” of (C_g).) While these conditions may be weakened somewhat (see, e.g., [1], [26], [13] or, for a time-dependent result, [21]), it is known [16], [13] that *some* such restrictions are necessary if waves are not to be trapped (to avoid, e.g., the possibility of infinitely long [recirculating] ray paths in \mathcal{R}). This seems an essential restriction as it is also known [15] that Theorem 4.1 is, to a considerable extent, reversible; see, also, the comments in § 4 of [24]. For application to (3.3), (3.4) by Theorem 3.2, the conditions (G), (C_g) seem much too strong (indeed, see the discussion in [25]) but, while some additional flexibility is available (e.g., through altering q by a constant, replacing u by $e^{-ct}u$ in (3.3)), this seems not to be useful [17]. It seems possible that a quite different choice of \mathcal{X} involving “microlocal” considerations might lead to a different decay rate result for (4.1), (4.2), thus permitting more general application of Theorem 3.2, but at present this is conjectural. \square

5. Control of diffusion equations. In this section we apply the results above (Theorems 2.1, 2.2, 3.2, 4.2, 4.4) to obtain controllability results for a class of diffusion equations of the form

$$(5.1) \quad \begin{aligned} a\dot{u} &= \nabla \cdot A \nabla u - qu =: \mathbf{L}u & \text{in } \mathcal{Q} := (0, T] \times \mathcal{R}, \\ \mathbf{y} &= \alpha u + \beta A \mathbf{n} \cdot \nabla u =: \mathbf{B}u & \text{on } \mathcal{I} := (0, T] \times \partial\mathcal{R}, \end{aligned}$$

and then proceed to characterize more conveniently the *optimal control space* $\mathcal{M} := \mathcal{M}_0$ for autonomous settings.

It will be convenient to denote by $\boldsymbol{\sigma} := (a, A, q, \alpha, \beta)$ the set of coefficients for (5.1). We assume, of course, that $a > 0$, $A > 0$, $q \geq 0$ (this last can be achieved with no loss of generality by substituting $e^{ct}u$ for u), that $\alpha^2 + \beta^2 \equiv 1$ and that the usual initial/boundary value problem is well-posed. With no confusion we alternatively view $\boldsymbol{\sigma}$ as defined on \mathcal{R} , $\partial\mathcal{R}$ in the autonomous case with which we are here principally concerned. We introduce, as a condition on \mathcal{R} , $\boldsymbol{\sigma}$, the ‘smoothing’ property:

(\mathcal{S}_ν) For every interval $[s, t] \subset [0, T]$ there is a constant κ such that for every

solution v of

$$(5.2) \quad -(av) = \mathbf{L}v, \quad \mathbf{B}v = \mathbf{0}$$

in $[s, t] \times \mathcal{R}$ one has

$$(5.3) \quad \|v(s, \cdot)\|_{H^{2\nu}(\mathcal{R})} \leq \kappa \|v(t, \cdot)\|_{L_2(\mathcal{R})}.$$

LEMMA 5.1. *Let $\partial\mathcal{R}, \sigma$ be sufficiently smooth (for simplicity we also assume β non-vanishing) and autonomous. Then (\mathcal{S}_ν) holds. The smoothness required of $\partial\mathcal{R}, \sigma$ depends on ν and, topologizing $\{\sigma\}$ in $C^{2\nu-2}(\mathcal{R}), C^{2\nu-1}(\partial\mathcal{R})$, the bound κ depends locally uniformly on σ .*

Proof. Let $\mathbf{A}: w \mapsto [-(1/a)\mathbf{L}w; \mathbf{B}w]$ for functions w defined on \mathcal{R} and let $\{(e_j, \lambda_j)\}$ be the eigenpairs: $\mathbf{L}e_j = -a\lambda_j e_j, \mathbf{B}e_j = \mathbf{0}$ with $\lambda_j > 0$ and $\{e_j\}$ orthonormal in $\mathfrak{X} := L_{2,a}(\mathcal{R})$, using positivity and selfadjointness of $-\mathbf{L}$ in $L_2(\mathcal{R})$ with the given boundary conditions). For a solution v of (5.2) we have the expansions:

$$v(t, \cdot) = \sum_j c_j e_j(\cdot), \quad v(s, \cdot) = \sum_j c_j e^{-\lambda_j \tau} e_j(\cdot),$$

$$\mathbf{A}_*^k v(s, \cdot) = \sum_j c_j [e^{-\tau \lambda_j} \lambda_j^k] e_j(\cdot),$$

where $\tau := t - s, \mathbf{A}_* := -(1/a)\mathbf{L}$. Thus,

$$(5.4) \quad \begin{aligned} \|\mathbf{A}_*^k v(s, \cdot)\|_{\mathfrak{X}}^2 &= \sum_j |c_j|^2 \lambda_j^{2k} e^{-2\tau \lambda_j} \leq \left(\frac{k}{e\tau}\right)^{2k} \|v(t, \cdot)\|_{\mathfrak{X}}^2, \\ \|\mathbf{A}_*^k v(s, \cdot)\|_{L_2(\mathcal{R})}^2 &\leq \frac{\max\{a\}}{\min\{a\}} \left(\frac{k}{e\tau}\right)^{2k} \|v(t, \cdot)\|_{L_2(\mathcal{R})}^2. \end{aligned}$$

With suitable smoothness, the formal operator \mathbf{A} induces bounded and (see, e.g., [10, Th. 5.2]) boundedly invertible operators

$$\mathbf{A}_k: H^{2k}(\mathcal{R}) \rightarrow H^{2k-2}(\mathcal{R}) \times H^{2k-3/2}(\partial\mathcal{R}), \quad k = 1, \dots, \nu.$$

Thus, for each k we may interpret \mathbf{A}_*^{-1} as the solution operator for the elliptic problem: $-\mathbf{L}w = af, \mathbf{B}w = 0$ with f in $H^{2k-2}(\mathcal{R})$ giving w in $H^{2k}(\mathcal{R})$; we have $\mathbf{A}_k w = [f, 0]$ so $\|w\| \leq \|\mathbf{A}_k^{-1}\| \|f\|$. Combining this with (5.4) gives the desired estimate (5.3) with

$$(5.5) \quad \kappa := \left[\frac{\max\{a\}}{\min\{a\}} \right]^{1/2} \nu^\nu e^{-\nu} (t-s)^{-\nu} \|\mathbf{A}_\nu^{-1}\| \cdots \|\mathbf{A}_1^{-1}\|.$$

It is easily verified that each \mathbf{A}_k depends continuously in operator norm on a, A, q and their derivatives of order up to $2k-2$ topologized in $L_\infty(\mathcal{R})$ and on $\alpha, \beta \mathbf{A} \mathbf{n}$ and their derivatives of order up to, say, $2k-1$ in $L_\infty(\partial\mathcal{R})$. Thus, (5.5) shows that κ can be taken continuously dependent on the coefficients σ topologized in $C^{2\nu-2}(\mathcal{R}), C^{2\nu-1}(\partial\mathcal{R})$ as appropriate. Note that if one were to specify Dirichlet conditions ($\beta \equiv 0$) instead, then one would take $\mathbf{A}_k: H^{2k}(\mathcal{R}) \rightarrow H^{2k-2}(\mathcal{R}) \times H^{2k-1/2}(\partial\mathcal{R})$ and proceed accordingly with κ continuously dependent on a, A, q in $C^{2\nu-2}(\mathcal{R})$. \square

With this lemma in hand we proceed to obtain the desired controllability/observability result.

THEOREM 5.2. *Let $\mathcal{R}, \partial_a \mathcal{R}$ satisfy (G); let a, A, q be autonomous with $\beta \equiv 0$ on $\partial_a \mathcal{R}$ and satisfying the conditions $(C_a), (C_g)$ and $(*)$. Then for any $T > 0$ the system $\mathbf{S} = \mathbf{S}_T(\sigma)$ with $\mathfrak{X} = \mathfrak{J} := L_2(\mathcal{R}), \mathfrak{Y} := L_2([0, T] \times \partial_a \mathcal{R})$ is exactly nullcontrollable.*

Further, as σ varies (topologized by, e.g., $C^1(\mathcal{R})$, $C^1(\partial_a \mathcal{R})$ subject to the conditions imposed) the controllability is locally uniform.

Proof. By the conditions assumed on the geometry and σ , we can apply Theorem 4.4 to assert existence of a decay rate $\zeta = \zeta(\cdot, \sigma)$ for the corresponding wave equation. By Theorem 4.1, then, one has exact nullcontrollability for that wave equation for large enough T . Observe that as the coefficients a, A, q vary C^1 -continuously the parameters $\|\varepsilon_0\|, r_0, \alpha_0, \delta, \gamma$ of the conditions will vary continuously. Thus, the bound on $\zeta(\cdot, \sigma)$ depends continuously on σ and so, therefore, do T, K_1 of Theorem 4.1.

Select a bounded region \mathcal{R}' such that $\mathcal{R}_1 \subset \mathcal{R}' \subset \hat{\mathcal{R}}$ (recall that \mathcal{R}_1 may be chosen for (G_1) independently of σ) and such that $\partial \mathcal{R}$ lies in the interior of the closure of \mathcal{R}' (relative to the closure of $\hat{\mathcal{R}}$); let $\mathbf{S}' = \mathbf{S}'_T(\sigma)$ be the system defined for $0 \leq t \leq T$ by the original parabolic equation extended to \mathcal{R}' with Dirichlet conditions. Observe that \mathbf{S}' is an extension of \mathbf{S} in the sense of (2.3): (i) take f, h to be given by restriction from \mathcal{R}' to \mathcal{R} , (ii) given \mathbf{x}' in $\mathcal{X}' := L_{2,a}(\mathcal{R}')$ and \mathbf{y}' in \mathcal{Y}' , let w be the solution of: $a\dot{w} = \mathbf{L}w$ in $[0, T] \times \mathcal{R}'$ with $w = \mathbf{y}'$ on $[0, T] \times \partial \mathcal{R}'$ and $w(0, \cdot) = \mathbf{x}'$, (iii) let $g(\mathbf{x}', \mathbf{y}')$ be the trace of w on $[0, T] \times \partial \mathcal{R}$. Note that commutativity of the diagram (2.3) follows by uniqueness of the initial/boundary value problem for (5.1) in $[0, T] \times \mathcal{R}$. The trace in step (iii) is in $\mathcal{Y} := L_2([0, T] \times \partial_a \mathcal{R})$ since (cf., [11]) w is sufficiently regular to give an L_2 trace (for other than Dirichlet boundary conditions on $\partial_a \mathcal{R}$ one might employ here a trick used earlier, for the same purpose, in [23] to obtain suitable smoothness) and since \mathcal{Y}' is such that \mathbf{y}' vanishes on $[0, T] \times \partial_{\neq} \mathcal{R}' \supset [0, T] \times \partial_{\neq} \mathcal{R}$.

Applying Theorem 3.2 now gives (locally uniform in σ) nullcontrollability of $\mathbf{S}'_T(\sigma)$ —and so of $\mathbf{S} = \mathbf{S}_T(\sigma)$ —for every $T > 0$. \square

As noted in § 2, the optimal control for a given initial state $u(0, \cdot) = \mathbf{x}$ must lie in the subspace $\mathcal{M} := \mathcal{M}_0 \subset \mathcal{Y}$ given by (3.10): \mathcal{M}_0 consists of the complementary boundary data $\boldsymbol{\eta} := \mathbf{B}'v$ for solutions v of the adjoint equation (5.2) for which $\boldsymbol{\eta}$ is in $L_2(\mathcal{S}_a)$ and $v(T, \cdot)$ is in $\mathcal{Z}^* := L_{2,a}(\mathcal{R})$. We wish to characterize the closure $\bar{\mathcal{M}} := \bar{\mathcal{M}}_0$ directly as given in the same fashion but omitting the requirement that $v(T, \cdot)$ be in \mathcal{Z}^* . I.e., letting

$$(5.6) \quad \bar{\mathcal{M}} := \{\boldsymbol{\eta} := \mathbf{B}'v : \boldsymbol{\eta} \in L_2(\mathcal{S}_a), v \text{ satisfies (5.2) on } [0, T] \times \mathcal{R}\},$$

we wish to show that $\mathcal{M} = \bar{\mathcal{M}}$. This alternate characterization of \mathcal{M} will be applied in the final section. Note that we no longer restrict ourselves to Dirichlet boundary conditions.

THEOREM 5.3. *Let σ be autonomous and sufficiently smooth; let $\sigma, \mathcal{R}, \partial_a \mathcal{R}$ be such that for arbitrarily short intervals (as $[0, \tau]$ for any $\tau > 0$) the system $\mathbf{S} = \mathbf{S}_\tau$ given by (5.1) with $\mathcal{X} = \mathcal{Z} := L_{2,a}(\mathcal{R})$, $\mathcal{Y}_\tau := L_2([0, \tau] \times \partial_a \mathcal{R})$ is null controllable. Then, with $\mathcal{M}, \bar{\mathcal{M}}$ given as above in $\mathcal{Y} := L_2([0, T] \times \partial_a \mathcal{R})$, one has $\mathcal{M} = \bar{\mathcal{M}}$, i.e., the optimal control space \mathcal{M} consists precisely of the complementary boundary data $\mathbf{B}'v$ for solutions v of (5.2). Further, for every $\boldsymbol{\eta} = \mathbf{B}'v$ in $\mathcal{M} = \bar{\mathcal{M}}$ one has κ_1, κ_2 (locally uniformly bounded in their dependence on σ —as in Lemma 5.1) such that*

$$(5.7) \quad \|v(t, \cdot)\|_{H^s(\mathcal{R})} \leq \kappa_1 \|\boldsymbol{\eta}\|_{L_2([t, T] \times \partial_a \mathcal{R})}, \quad (0 \leq t < T)$$

and

$$(5.8) \quad \|\boldsymbol{\eta}\|_{H^{r,s}([0, t] \times \partial_a \mathcal{R})} \leq \kappa_2 \|\boldsymbol{\eta}\|_{L_2([t, T] \times \partial_a \mathcal{R})}.$$

Proof. We proceed to construct a function v on $[0, T] \times \mathcal{R}$ for given $\boldsymbol{\eta}$ in \mathcal{M} , next obtain (5.7), and then show that this v satisfies (5.2) and $\boldsymbol{\eta} = \mathbf{B}'v$ whence (5.8); this shows $\mathcal{M} \subset \bar{\mathcal{M}}$. To show $\bar{\mathcal{M}} \subset \mathcal{M}$ we need only show, finally, that \mathcal{M}_0 is dense in $\bar{\mathcal{M}}$.

Let $\boldsymbol{\eta}$ be in \mathcal{M}_0 , so $\boldsymbol{\eta} = \mathbf{B}'v$ for a unique solution v of (5.2) with $v(T, \cdot)$ in $\mathcal{J}^* = \mathcal{J} = \mathcal{X} := L_{2,a}(\mathcal{R})$. Define

$$(5.9) \quad \mathbf{P}_t \boldsymbol{\eta} := v(t, \cdot) \quad (0 \leq t < T, \quad \boldsymbol{\eta} \in \mathcal{M}_0).$$

The nullcontrollability assumption of the Theorem (with $\tau := T - t$) implies, by the Duality Theorem 2.1 and by translating the autonomous system \mathbf{S}_τ from $[0, \tau]$ to $[t, T]$, that each such \mathbf{P}_t is bounded:

$$(5.10) \quad \begin{aligned} \|v(t, \cdot)\|_{\mathcal{X}} &\leq \|\mathbf{P}_t\| \|\boldsymbol{\eta}\|_{L_2([t, T] \times \partial_a \mathcal{R})} \\ &\leq \|\mathbf{P}_t\| \|\boldsymbol{\eta}\|_{\mathcal{Y}}. \end{aligned}$$

Thus, each \mathbf{P}_t ($0 \leq t < T$) extends by continuity to all of \mathcal{M} . For any $\boldsymbol{\eta}$ in \mathcal{M} we now define a function v on $[0, T] \times \mathcal{R}$ (actually, $v: (0, T) \rightarrow \mathcal{X}$) by setting

$$(5.11) \quad v(t, \cdot) := [\mathbf{P}\boldsymbol{\eta}](t) := \mathbf{P}_t \boldsymbol{\eta} \quad (0 \leq t < T; \quad \boldsymbol{\eta} \in \mathcal{M}),$$

reversing (5.9).

Now define $\mathbf{P}_{s,t}$ for $0 \leq s < t < T$ by letting

$$(5.12) \quad \mathbf{P}_{s,t} \boldsymbol{\eta} := v_t(s, \cdot) \quad (0 \leq s < t < T, \quad \boldsymbol{\eta} \in \mathcal{M})$$

where v_t is defined as the solution of (5.2) on $[0, t] \times \mathcal{R}$ with

$$(5.13) \quad v_t(t, \cdot) = \mathbf{P}_t \boldsymbol{\eta} \quad (0 < t < T, \quad \boldsymbol{\eta} \in \mathcal{M}).$$

Using (5.4) with $k = 0$ and (5.10) gives

$$\|\mathbf{P}_{s,t} \boldsymbol{\eta}\|_{\mathcal{X}} = \|v_t(s, \cdot)\|_{\mathcal{X}} \leq \|v_t(t, \cdot)\|_{\mathcal{X}} \leq \|\mathbf{P}_t\| \|\boldsymbol{\eta}\|_{\mathcal{Y}}$$

so each $\mathbf{P}_{s,t}$ is bounded. On the other hand, for $\boldsymbol{\eta} = \mathbf{B}'v$ in \mathcal{M}_0 and $0 \leq s < t < T$ we have $\mathbf{P}_{s,t} \boldsymbol{\eta} = \mathbf{P}_s \boldsymbol{\eta}$ —each equals $v(s, \cdot)$ by the semigroup property of the solution operator for (5.2). By continuity we have $\mathbf{P}_s \boldsymbol{\eta} = \mathbf{P}_{s,t} \boldsymbol{\eta}$ for any t in (s, T) and any $\boldsymbol{\eta}$ in \mathcal{M} . It follows that $v(s, \cdot) := [\mathbf{P}\boldsymbol{\eta}](s)$ coincides with $v_t(s, \cdot)$ for $0 \leq s < t < T$ and so satisfies (5.2) there. As t can be arbitrary in $(0, T)$, this shows that $v = \mathbf{P}\boldsymbol{\eta}$ satisfies (5.2) on $[0, T] \times \mathcal{R}$. Further, by (5.10) and Lemma 5.1,

$$(5.14) \quad \|v(t, \cdot)\|_{H^{2\nu}(\mathcal{R})} \leq \kappa \|\mathbf{P}_s\| \|\boldsymbol{\eta}\|_{L_2([s, T] \times \partial_a \mathcal{R})}$$

for any convenient choice of s in (t, T) . This proves (5.7).

Next, for $0 < r < T$ and $\boldsymbol{\eta}$ in \mathcal{M} define $\mathbf{R}_r \boldsymbol{\eta}$ by

$$[\mathbf{R}_r \boldsymbol{\eta}](t) := \mathbf{B}' \mathbf{P}_t \boldsymbol{\eta} = \mathbf{B}' v_s(t) \quad (0 \leq t \leq r < s < T).$$

We have \mathbf{B}' continuous from $H^2(\mathcal{R})$ to $L_2(\partial_a \mathcal{R})$ —indeed, as $\beta = 0$, to $H^{3/2}(\partial_a \mathcal{R})$ —and, using that norm for \mathbf{B}' , (5.14) gives

$$(5.15) \quad \begin{aligned} \|\mathbf{R}_r \boldsymbol{\eta}\|_{L_2([0, r] \times \partial_a \mathcal{R})} &\leq \|\mathbf{B}'\| \left[\int_0^r \|v_s(t, \cdot)\|_{H^2(\mathcal{R})}^2 dt \right]^{1/2} \\ &\leq \frac{\|\mathbf{B}'\| \|\mathbf{A}_1^{-1}\|}{e} \left[\frac{\log(s/(s-r))}{\min_{\mathcal{R}} \{a\}} \right]^{1/2} \|\mathbf{P}_s\| \|\boldsymbol{\eta}\|_{L_2([s, r] \times \partial_a \mathcal{R})} \end{aligned}$$

This shows $\mathbf{R}_r: L_2([r, T] \times \partial_a \mathcal{R}) \rightarrow L_2([0, r] \times \partial_a \mathcal{R})$ is continuous. Since $\mathbf{R}_r \boldsymbol{\eta}$ just gives the restriction of $\boldsymbol{\eta}$ to $[0, r] \times \partial_a \mathcal{R}$ for $\boldsymbol{\eta}$ in \mathcal{M}_0 , this must also be true, by continuity, for all $\boldsymbol{\eta}$ in \mathcal{M} . Thus, $\mathbf{B}'v$ ($v := \mathbf{P}\boldsymbol{\eta}$) coincides with $\boldsymbol{\eta}$ on $[0, r] \times \partial_a \mathcal{R}$ and, since r is arbitrary in $(0, \tau)$, this shows $\boldsymbol{\eta} = \mathbf{B}'v$ in \mathcal{Y} . We have shown that $\mathcal{M} \subset \mathcal{M}$.

The inequality (5.8) is obtained from (5.7) much as was the special case (5.15). We consider $H^{r,s}([0, t] \times \partial_a \mathcal{R})$ as $H^r([0, t] \rightarrow H^s(\partial_a \mathcal{R}))$ and observe that (5.2) gives $\partial^k \boldsymbol{\eta} / \partial t^k = \mathbf{B}' \mathbf{A}_*^k v$. We only consider r, ν integral ($s = 2\nu - \frac{3}{2}$) for simplicity. For $0 \leq \tau \leq t < \ell < T$ one has

$$\begin{aligned}
 \|\boldsymbol{\eta}^{(k)}(t, \cdot)\|_{H^s(\partial_a \mathcal{R})} &\leq \|\mathbf{B}'\| \|\mathbf{A}_*^k v(\tau, \cdot)\|_{H^{2\nu}(\mathcal{R})} \\
 (5.16) \quad &\leq \|\mathbf{B}'\| \|\mathbf{A}_\nu^{-1}\| \cdots \|\mathbf{A}_1^{-1}\| \|\mathbf{A}_*^{k+\nu} v(\tau, \cdot)\|_{L_2(\mathcal{R})} \\
 &\leq \|\mathbf{B}'\| \|\mathbf{A}_\nu^{-1}\| \cdots \|\mathbf{A}_1^{-1}\| \left(\frac{\nu + k}{e(\ell - \tau)} \right)^{\nu+k} [\min\{a\}]^{-1/2} \|\mathbf{P}_\ell\| \|\boldsymbol{\eta}\|_{L_2([\ell, T] \times \partial_a \mathcal{R})}.
 \end{aligned}$$

Now (5.8) follows by integrating over $[0, t]$ with respect to t . (Note that $\|\mathbf{B}'\| < \infty$ is here the norm of $\mathbf{B}' : H^{2\nu}(\mathcal{R}) \rightarrow H^s(\partial_a \mathcal{R})$). The locally uniform dependence on σ is shown as in the proof of Lemma 5.1.

Finally, we wish to prove that \mathfrak{M}_0 is dense in \mathfrak{M} , i.e., that for $\boldsymbol{\eta} = \mathbf{B}'v$ in \mathfrak{M} one can find $\boldsymbol{\eta}_j = \mathbf{B}'v_j$ in \mathfrak{M}_0 for which $\boldsymbol{\eta}_j \rightarrow \boldsymbol{\eta}$ in \mathfrak{Y} . Let $T_j \rightarrow T -$ (so $\varepsilon_j := T - T_j \rightarrow 0^+$) and let v_j be the solution of (5.2) in $[0, T] \times \mathcal{R}$ with $v_j(T, \cdot) := v(T_j, \cdot)$ in \mathcal{X} . Clearly each $\boldsymbol{\eta}_j$ is in \mathfrak{M}_0 . Extending v and $\boldsymbol{\eta} = \mathbf{B}'v$ to $-\varepsilon_0 \leq t < 0$ by (5.2) and using the time-independence of (5.2), we have

$$v_j(t, \cdot) = v(t - \varepsilon_j, \cdot), \quad \boldsymbol{\eta}_j(t, \cdot) = \boldsymbol{\eta}(t - \varepsilon_j, \cdot) \quad (0 \leq t \leq T)$$

so

$$\begin{aligned}
 \|\boldsymbol{\eta} - \boldsymbol{\eta}_j\|_{\mathfrak{Y}}^2 &= \int_0^T \|\boldsymbol{\eta}(t, \cdot) - \boldsymbol{\eta}(t - \varepsilon_j, \cdot)\|_{L_2(\partial_a \mathcal{R})}^2 dt \\
 (5.17) \quad &\leq \int_{-\infty}^{\infty} \|\hat{\boldsymbol{\eta}}(t) - \hat{\boldsymbol{\eta}}(t - \varepsilon_j)\|_{L_2(\partial_a \mathcal{R})}^2 dt \\
 &= \|\hat{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}(\cdot - \varepsilon_j)\|_{L_2(\mathbb{R} \rightarrow L_2(\partial_a \mathcal{R}))}^2.
 \end{aligned}$$

Here $\boldsymbol{\eta}$ is just $\boldsymbol{\eta}(t, \cdot)$ for $-\varepsilon \leq t < T$ and vanishes elsewhere on \mathbb{R} . A standard result (see, e.g., [9 Th. 13.24] for the case of scalar-valued functions; the result is obvious for simple functions in $L_2(\mathbb{R} \rightarrow L_2(\partial_a \mathcal{R}))$ and those are dense) now gives: $\|\hat{\boldsymbol{\eta}}(\cdot) - \hat{\boldsymbol{\eta}}(\cdot - \varepsilon_j)\| \rightarrow 0$ as $\varepsilon_j \rightarrow 0$ and this completes the proof. \square

Remark 5.4. Note that the choice of r in (5.8) does not affect the required spatial smoothness of the coefficients but, rather, follows from the assumed time-independence. Indeed, (see, e.g., [7]), solution semigroups for such problems are actually analytic whence $v(t, \cdot)$ (and so also $\boldsymbol{\eta}(t, \cdot) := \mathbf{B}'v(t, \cdot)$) are analytic in t for $\operatorname{Re} t < T$. Although not done in quite that way above, it is easily seen that the spatial smoothness of the estimates (5.7), (5.8) can be reduced to finding comparable boundary estimates for the eigenfunctions $\{e_j(\cdot)\}$. Note, also, that the Sobolev embedding theorem can be used to obtain *pointwise* estimates for $\boldsymbol{\eta}$ (and its derivatives) away from T —provided r, s can be taken large enough. \square

Remark 5.5. A careful look at the proof of Theorem 5.3 above shows that the major portion of the proof—that $\mathfrak{M} \subset \mathfrak{M}$ together with the estimates (5.7), (5.8)—goes through with only minor alterations if the coefficients are time-dependent. In that case we must have (\mathcal{S}_ν) and must now require nullcontrollability for arbitrarily short intervals $[r, T]$ so that each \mathbf{P}_r exists as a bounded operator: $L_2([r, T] \times \partial_a \mathcal{R}) \rightarrow L_{2,a(t, \cdot)}(\mathcal{R})$. The particular argument used above to show $\mathfrak{M} \subset \mathfrak{M}$, however, seems to depend intrinsically on the autonomy of the equation. \square

6. Continuous dependence. In this section we investigate the dependence of the optimal control on the set of coefficients σ . We begin by noting that Theorem 5.3 resolves a question raised in Remark 11 of [25].

COROLLARY 6.1. *Let the hypotheses of Theorem 5.3 hold. For $T > 0$ and $0 \neq \mathbf{x}$ in \mathcal{X} , let $\mathbf{y}(T)$ in $\mathcal{Y}_T := L_2([0, T] \times \partial_a \mathcal{R})$ be the minimum norm nullcontrol ($u(0, \cdot) = \mathbf{x}$, $\mathbf{B}u = \mathbf{y}(T)$, $u(T, \cdot) = \mathbf{0}$). Then $[T \mapsto \|\mathbf{y}(T)\|]$ is strictly decreasing.*

Proof. Following Remark 11 of [25], we note that if $\|\mathbf{y}(T')\| = \|\mathbf{y}(T)\|$ for some $0 < T' < T$ one must have $\mathbf{y}(T') = \mathbf{y}(T)$, i.e., as a function: $[0, T] \rightarrow L_2(\partial_a \mathcal{R})$, one would have $\mathbf{y}(T)$ vanishing for $T' < t < T$. Now, $\mathbf{y}(T)$ in $\mathcal{M} = \mathcal{M}(T)$ has the form $\mathbf{y}(T) = \mathbf{B}'v$ for some solution v of (5.2). The analyticity in t of the solution semigroup for (5.2) and so of this v implies similar analyticity of $\mathbf{y}(T)$. Thus, $\mathbf{y}(T)$ vanishing on the interval (T', T) would imply $\mathbf{y}(T) = \mathbf{0}$. However, this is impossible for $\mathbf{x} \neq \mathbf{0}$ by the backward uniqueness property of the (autonomous) equation (5.2). \square

For the discussion of the dependence on σ of optimal controls we introduce a new topology for \mathcal{Y} . First make a choice of ν determined by the smoothness of the coefficients σ to be considered as in the proof/discussion of the estimate (5.8) and noting Remark 5.4. We assume $\nu \geq 1$. Introduce the seminorms

$$N(\mathbf{y}) := N(\mathbf{y}; \tau, j, s) := \|\mathbf{y}^{(j)}\|_{L_2([0, \tau] \rightarrow H^s(\partial_a \mathcal{R}))}$$

for $0 < \tau < T$, $0 \leq s \leq 2\nu - \frac{3}{2}$ (if $\beta = 0$, then $s < 2\nu - \frac{1}{2}$) and $j = 0, 1, \dots$. Finally, let \mathcal{Y} be the subset of $\mathcal{Y} := L_2([0, T] \times \partial_a \mathcal{R})$ for which these are all finite, topologized by these together with the weak topology of \mathcal{Y} . We now fix \mathcal{R} , $\partial_a \mathcal{R}$ (with $\partial \mathcal{R}$ suitably smooth) and let Σ be the set of suitably smooth σ for which the hypotheses of Theorem 5.3 hold, topologized as in Lemma 5.1—i.e., in $C^{2\nu+2}(\mathcal{R})$ for a, A, q and in $C^{2\nu-1}(\partial \mathcal{R})$ for $\alpha, \beta A$. We write, as necessary, $\mathbf{S}(\sigma)$ for the system defined using (5.1) with the coefficients σ in Σ . Before proceeding to the continuity theorem itself, we need a lemma regarding certain continuous dependence on σ for solutions of the usual initial-boundary value problem which we take in the adjoint form (5.2).

LEMMA 6.2. *Let $\mathcal{R} \subset \mathbb{R}^n$ with $\partial \mathcal{R}$, σ , σ_j sufficiently smooth and $\sigma_j \rightarrow \sigma$ in the sense: $a_j \rightarrow a$, $q_j \rightarrow q$ in $C^2(\mathcal{R})$, $A_j \rightarrow A$ in $C^3(\mathcal{R})$, $\alpha_j \rightarrow \alpha$, $\beta_j A_j \rightarrow \beta A$ in $C^2(\partial \mathcal{R})$. For fixed ξ in $C_0^\infty(\mathcal{R})$ let z, z_j be the solutions, respectively, of*

$$(6.1) \quad -a z = \mathbf{L}z, \quad \mathbf{B}z = \mathbf{0}, \quad z(T, \cdot) = \xi,$$

$$(6.2) \quad -a_j z_j = \mathbf{L}_j z_j, \quad \mathbf{B}_j z_j = \mathbf{0}, \quad z_j(T, \cdot) = \xi$$

(of course, $\mathbf{L}_j x := \nabla \cdot A_j \nabla x - q_j x$ in \mathcal{Q} , $\mathbf{B}_j x := \alpha_j x + \beta_j A_j \mathbf{n} \cdot \nabla x$ on \mathcal{S}). Then $z_j(0, \cdot) \rightarrow z(0, \cdot)$ in $L_{2,a}(\mathcal{R})$ and $\mathbf{B}'_j z_j \rightarrow \mathbf{B}'z$ in $L_2(\mathcal{S}_a)$.

Proof. Let $\mathbf{A}_j := (1/a_j)\mathbf{L}_j$ and let $\mathcal{X}_j := \{\mathbf{x} \in H^4(\mathcal{R}): \mathbf{B}_j \mathbf{x} = \mathbf{0}, \mathbf{B}_j \mathbf{A}_j \mathbf{x} = \mathbf{0}\}$ with the norm

$$\|\mathbf{x}\|_j := \left[\int_{\mathcal{R}} |[\mathbf{A}_j^2 \mathbf{x}](\cdot)|^2 a_j \right]^{1/2} = \|\mathbf{A}_j^2 \mathbf{x}\|_{L_{2,a_j}(\mathcal{R})}.$$

Since $\mathbf{A}_j \rightarrow \mathbf{A} := -(1/a)\mathbf{L}$ and $\mathbf{B}_j \rightarrow \mathbf{B}$ suitably, we have, as in the proof of Lemma 5.1, a bound C (uniform in j —i.e., locally uniform in its dependence on σ) such that

$$(6.3) \quad C^{-1} \|\mathbf{x}\|_{H^4(\mathcal{R})} \leq \|\mathbf{x}\|_j \leq C \|\mathbf{x}\|_{H^4(\mathcal{R})} \quad (\mathbf{x} \in \mathcal{X}_j).$$

Without confusion we also use \mathbf{A}_j to denote the “same” operator as densely defined: $\mathcal{D}_j \subset \mathcal{X}_j \rightarrow \mathcal{X}_j$ with domain $\mathcal{D}_j := \{\mathbf{x} \in \mathcal{X}_j: \mathbf{A}_j \mathbf{x} \in \mathcal{X}_j\}$. It is easily verified that the inner product of \mathcal{X}_j has been taken just so as to make \mathbf{A}_j selfadjoint (using the selfadjointness of \mathbf{L}_j in $L_2(\mathcal{R})$). Further, \mathbf{A}_j is negative semidefinite (as $a_j, A_j > 0, q_j \geq 0$). It

follows that $\|\lambda(\lambda - \mathbf{A}_j)^{-1}\| \leq 1$ for $\lambda > 0$ so \mathbf{A}_j is the infinitesimal generator of a C_0 contraction semigroup on \mathfrak{X}_j . Thus,

$$(6.4) \quad \|z_j(t, \cdot)\|_{H^4(\mathcal{R})} \leq C \|z_j(t, \cdot)\|_j \leq C \|\zeta\|_j \leq C^2 \|\zeta\|_{H^4(\mathcal{R})},$$

uniformly in j and $t \leq T$, noting that ζ is in each \mathfrak{X}_j . Using (6.2) we also obtain uniform bounds on $\{\dot{z}_j\}, \{\ddot{z}_j\}$ in $H^2(\mathcal{R}), L_2(\mathcal{R})$ respectively. Indeed we have boundedness (uniformly in j and for $t \leq T$) for $\{z_j\}, \{\nabla z_j\}, \{\ddot{z}_j\}, \{\nabla \ddot{z}_j\}$, respectively, in the spaces $H^4(\mathcal{R}), h^3(\mathcal{R}), H^2(\mathcal{R}), H^1(\mathcal{R})$ and, taking traces, in $H^{7/2}(\partial\mathcal{R}), H^{5/2}(\partial\mathcal{R}), H^{3/2}(\partial\mathcal{R}), H^{1/2}(\partial\mathcal{R})$. We now introduce

$$(6.5) \quad \begin{aligned} f_j &:= (a_j - a)z_j + (\mathbf{L} - \mathbf{L}_j)z_j = (a_j - a)\dot{z}_j + \nabla \cdot (\mathbf{A}_j - \mathbf{A})\nabla z_j - (q_j - q)z_j, \\ g_j &:= (\mathbf{B}_j - \mathbf{B})z_j = (\alpha_j - \alpha)z_j + (\beta_j \mathbf{A}_j - \beta \mathbf{A})\mathbf{n} \cdot \nabla z_j. \end{aligned}$$

Observe that $\dot{f}_j \rightarrow 0, \nabla f_j \rightarrow 0$ in $L_2(\mathcal{R})$ uniformly in t (so $f_j \rightarrow 0$ in $H^{2,1}(\mathcal{Q})$, in the notation of [11]) provided $a_j \rightarrow a$ in $C^2(\mathcal{R})$, $\mathbf{A}_j \rightarrow \mathbf{A}$ in $C^3(\mathcal{R})$, $q_j \rightarrow q$ in $C^2(\mathcal{R})$. Similarly, $g_j \rightarrow 0$ in $H^{2,1}(\mathcal{S})$ provided $\alpha_j \rightarrow \alpha$ and $\beta_j \mathbf{A}_j \mathbf{n} \rightarrow \beta \mathbf{A} \mathbf{n}$ in $C^2(\partial\mathcal{R})$. Each $w_j := z - z_j$ satisfies

$$(6.6) \quad -a\dot{w}_j = \mathbf{L}w_j + f_j, \quad \mathbf{B}w_j = g_j, \quad w_j(T, \cdot) = \mathbf{0}.$$

We may apply, e.g., Theorem 6.2 of [11] with r just less than $\frac{3}{4}$ (to avoid the anomalous half-integral case for $2r$) to show that $w_j \rightarrow 0$ in $H^{7/2-, 7/4-}(\mathcal{Q})$ and $\nabla w_j \rightarrow 0$ in $H^{5/2-, 5/4-}(\mathcal{Q})$. Taking traces, we have $w_j(0, \cdot) \rightarrow 0$ and $\mathbf{B}'w_j := (\beta w_j - \alpha \mathbf{A} \mathbf{n} \cdot \nabla w_j) \rightarrow 0$ in $L_{2,a}(\mathcal{R})$ and $L_2(\mathcal{S}_a)$, respectively, with smoothness to spare. Note that $\{z_j(t, \cdot)\}$ is bounded in $H^4(\mathcal{R})$, uniformly in j and $t(0 \leq t \leq T)$ and that we have $\mathbf{B}'_j \rightarrow \mathbf{B}$ in operator norm: $H^4(\mathcal{R}) \rightarrow L_2(\partial_a \mathcal{R})$ as $\sigma_j \rightarrow \sigma$. Since $\mathbf{B}'z - \mathbf{B}'_j z_j = \mathbf{B}'w_j + (\mathbf{B}' - \mathbf{B}'_j)z_j$, we have the desired result. \square

If we had topologized the coefficients as multipliers on appropriate Sobolev spaces, the smoothness required in the sense of $\sigma_j \rightarrow \sigma$ could be reduced, say, replacing $\mathfrak{X}_j \approx H^4, r = \frac{3}{4}-$ by $\mathfrak{X}_j \approx H^3, r = \frac{1}{5}$. Independently, the required smoothness can be reduced if it happens that (pointwise) the direction of $\{\beta_j \mathbf{A}_j \mathbf{n}\}$ is fixed: $\beta_j \mathbf{A}_j \mathbf{n} = \lambda_j \beta \mathbf{A} \mathbf{n}$ with $\alpha_j/\lambda_j \rightarrow \alpha$ —e.g., if \mathbf{A}, \mathbf{A}_j are scalar.

We come now to the desired result on the continuous dependence of the optimal nullcontrol on the coefficients involved.

THEOREM 6.3. *Let $\{\sigma_j\}$ be in Σ with $\sigma_j \rightarrow \sigma$ in Σ . (Assume ν large enough for Lemma 6.2 to apply.) Suppose for each $T'(0 \leq T' < T)$ the systems $\mathbf{S}_{[T', T]}(\sigma_j)$ are uniformly (in j) nullcontrollable. Let \mathbf{y}_j in $\mathcal{Y} := L_2(\mathcal{S}_a)$ be the optimal control associated in $\mathbf{S}_j := \mathbf{S}(\sigma_j)$ with the fixed initial state \mathbf{x} in $\mathfrak{X} := L_2(\mathcal{R})$ and terminal states $\mathbf{z}_j \rightarrow \mathbf{z}$ in $\mathcal{Z} := L_2(\mathcal{R})$. Assume $\{\mathbf{y}_j\}$ is bounded in \mathcal{Y} (this is automatic if these are nullcontrols: $\mathbf{z}_j = \mathbf{0} = \mathbf{z}$ noting the uniform nullcontrollability assumption above). Then \mathbf{x} is controllable to \mathbf{z} in $\mathbf{S} := \mathbf{S}(\sigma)$ with an optimal control \mathbf{y} and $\mathbf{y}_j \rightarrow \mathbf{y}$ in \mathcal{Y} .*

Proof. The proof both uses and resembles those of Theorems 5.3 and 6.2. We show, first, that a subsequence of $\{\mathbf{y}_j\}$ converges in \mathcal{Y} to \mathbf{y}_∞ . Next that \mathbf{y}_∞ is a control from \mathbf{x} to \mathbf{z} in \mathbf{S} (so there does exist an optimal control \mathbf{y}) and, finally, that \mathbf{y}_∞ is in $\mathcal{M} = \mathcal{M}(\sigma)$ so that it must be the (unique) optimal control \mathbf{y} and $\mathbf{y}_j \rightarrow \mathbf{y}$ in \mathcal{Y} .

By Theorem 2.1, each \mathbf{y}_j is in the appropriate optimal control space $\mathcal{M}_j := \mathcal{M}(\sigma_j) \subset \mathcal{Y}$ so, by Theorem 5.3, there is a (unique) solution v_j of

$$(6.7) \quad -a_j \dot{v}_j = \mathbf{L}_j v_j, \quad \mathbf{B}_j v_j = \mathbf{0}$$

for which $\mathbf{y}_j = \mathbf{B}'_j v_j$. There is no difficulty in extending (6.7) to $\mathcal{Q}^\delta := [-\delta, T] \times \mathcal{R}$ and so extending each \mathbf{y}_j to \mathbf{y}_j^δ in $\mathcal{S}_a^\delta := [-\delta, T] \times \partial_a \mathcal{R}$. The arguments for (5.7), (5.9) continue to be valid for $-\delta \leq t \leq 0$ as well as on $[0, T]$. Thus we may conclude from (5.8) that

the *extended* data \mathbf{y}_j^δ is in $\mathcal{Y}^\delta := L_2(\mathcal{S}_a^\delta)$ for each j and, indeed, that $\{\mathbf{y}_j^\delta\}$ is bounded in \mathcal{Y}^δ .

For $0 \leq \tau \leq T$ and suitable k, s (depending on ν) there will be a constant κ such that, uniformly in j .

$$(6.8) \quad N^\delta(\mathbf{y}_j^\delta; \tau, k, s) \leq \kappa \|\mathbf{y}_j\|_{L_2([\tau, T] \times \partial_a \mathcal{R})} \leq \kappa C$$

where N^δ is defined on \mathcal{S}_a^δ as N was on \mathcal{S}_a and $C := \sup_j \{\|\mathbf{y}_j\|_{\mathcal{Y}}\}$. To see this we choose t conveniently in (τ, T) and use (5.8), noting the local uniformity of κ_2 over Σ and the assumed uniform nullcontrollability of $\mathbf{S}_{[t, T]}(\sigma_j)$.

As $\{\mathbf{y}_j^\delta\}$ is bounded in \mathcal{Y}^δ we may extract a weakly convergent sequence $\mathbf{y}_{j(i)}^\delta \rightarrow \mathbf{y}_\infty^\delta$. Clearly, $\mathbf{y}_{j(i)} \rightarrow \mathbf{y}_\infty$ (:= the restriction of \mathbf{y}_∞^δ to \mathcal{S}_a). Since we will show $\mathbf{y}_\infty = \mathbf{y}(\mathbf{y}_\infty^\delta = \mathbf{y}^\delta)$ for *any* such sequence, it will follow that the *full* sequence $\{\mathbf{y}_j^\delta\}$ converges to \mathbf{y}^δ so there is no loss of generality in taking $\{\mathbf{y}_{j(i)}^\delta\} = \{\mathbf{y}_j^\delta\}$, i.e., in writing for convenience $\mathbf{y}_j^\delta \rightarrow \mathbf{y}_\infty^\delta$ in \mathcal{Y}^δ . To obtain convergence in \mathcal{Y}^δ we now use (6.8) and show $N^\delta(\mathbf{y}_j^\delta - \mathbf{y}_\infty^\delta; \tau, k, s) \rightarrow 0$ for any specified τ, k, s . Take $k' > k, s' > s$ and observe that $\{N^\delta(\mathbf{y}_j^\delta; \tau, k', s')\}$ is bounded by (6.8). By compactness of the map: $H^{k'}([-\delta, \tau] \rightarrow H^s(\partial_a \mathcal{R})) \rightarrow H^k([-\delta, \tau] \rightarrow H^s(\partial_a \mathcal{R}))$, a subsequence converges strongly in $H^k([-\delta, T] \rightarrow H^s(\partial_a \mathcal{R}))$ to say, \mathbf{y}_* . As we already have $\mathbf{y}_j^\delta \rightarrow \mathbf{y}_\infty^\delta$, this \mathbf{y}_* must be the restriction of \mathbf{y}_∞^δ to $[-\delta, \tau] \times \partial_a \mathcal{R}$. As this is true for each subsequence, we have $\mathbf{y}_j^\delta \rightarrow \mathbf{y}_\infty^\delta$ in the sense of $N^\delta(\cdot; \tau, k, s)$. Varying τ, k, s , this gives convergence of $\{\mathbf{y}_j^\delta\}$ to \mathbf{y}_∞^δ in \mathcal{Y}^δ .

Next we show that \mathbf{y}_∞ is actually a control in \mathbf{S} , steering the fixed initial state \mathbf{x} to the terminal state \mathbf{z} . Let u, u_j be the “trajectories,” i.e., the solutions, respectively of

$$(6.9) \quad a\dot{u} = \mathbf{L}u, \quad \mathbf{B}u = \mathbf{y}_\infty, \quad u(0, \cdot) = \mathbf{x},$$

$$(6.10) \quad a_j \dot{u}_j = \mathbf{L}_j u_j, \quad \mathbf{B}_j u_j = \mathbf{y}_j, \quad u_j(0, \cdot) = \mathbf{x}.$$

For ζ in $\mathcal{Z}^* = L_2(\mathcal{R})$, one has, by (3.9) and noting (3.5), etc., that

$$(6.11) \quad \int_{\mathcal{R}} a u(T, \cdot) = \int_{\mathcal{R}} a \mathbf{x} z(0, \cdot) + \int_{\mathcal{S}_a} \mathbf{y}_\infty (\mathbf{B}' z),$$

$$(6.12) \quad \int_{\mathcal{R}} a_j z_j \zeta = \int_{\mathcal{R}} a_j \mathbf{x} z_j(0, \cdot) + \int_{\mathcal{S}_a} \mathbf{y}_j (\mathbf{B}'_j z_j),$$

where z, z_j are the solutions of (6.1), (6.2), respectively. Thus,

$$(6.13) \quad \begin{aligned} \int_{\mathcal{R}} a[u(T, \cdot) - \mathbf{z}] \zeta = & - \int_{\mathcal{R}} a(\mathbf{z} - \mathbf{z}_j) \zeta - \int_{\mathcal{R}} (a - a_j) \mathbf{z}_j \zeta + \int_{\mathcal{R}} a[z(0, \cdot) - z_j(0, \cdot)] \mathbf{x} \\ & + \int_{\mathcal{R}} (a - a_j) z_j(0, \cdot) \mathbf{x} + \int_{\mathcal{S}_a} (\mathbf{y}_\infty - \mathbf{y}_j) (\mathbf{B}' z) \\ & + \int_{\mathcal{S}_a} \mathbf{y}_j (\mathbf{B}' z - \mathbf{B}'_j z_j). \end{aligned}$$

Numbering the integrals on the right side of (6.13) from [i] to [vi], we consider this as $j \rightarrow \infty$ for ζ in $C_0^\infty(\mathcal{R})$. We have [i] $\rightarrow 0$ as $\mathbf{z}_j \rightarrow \mathbf{z}$ in $L_2(\mathcal{R})$; [ii], [iv] $\rightarrow 0$ since $\sigma_j \rightarrow \sigma$ and $\{z_j\}, \{z_j(0, \cdot)\}$ are bounded in $L_2(\mathcal{R})$; [v] $\rightarrow 0$ since $\mathbf{y}_j \rightarrow \mathbf{y}_\infty$ in \mathcal{Y} ; [iii], [vi] $\rightarrow 0$ by Lemma 6.2, as we note that $\{\mathbf{y}_j\}$ is bounded in \mathcal{Y} . Thus, $\int [u(T, \cdot) - \mathbf{z}] \zeta = 0$ for all ζ in $C_0^\infty(\mathcal{R})$ whence $u(T, \cdot) = \mathbf{z}$, so that \mathbf{y}_∞ is, indeed, a suitable control from \mathbf{x} to \mathbf{z} in \mathbf{S} .

Finally, we wish to show that \mathbf{y}_∞ is in $\mathcal{M} = \mathcal{M}(\sigma)$. For any ε ($0 < \varepsilon < \delta, T$), let $\hat{\mathbf{y}}^\varepsilon$ be the translation by ε of \mathbf{y}_∞^δ viewed as an element of \mathcal{Y} :

$$(6.14) \quad \hat{\mathbf{y}}^\varepsilon(t, \cdot) := \mathbf{y}_\infty^\delta(t - \varepsilon, \cdot) \quad \text{for } 0 < t < T.$$

We observe, much as in the argument following (5.17), that $\hat{\mathbf{y}}^\varepsilon \rightarrow \mathbf{y}_\infty$ in \mathcal{Y} as $\varepsilon \rightarrow 0^+$. Since \mathcal{M} is closed, we have \mathbf{y}_∞ in \mathcal{M} once it is shown that $\hat{\mathbf{y}}^\varepsilon$ is in \mathcal{M} for (small) $\varepsilon > 0$. Fixing ε , we define $\hat{\mathbf{y}}_j^\varepsilon$ from \mathbf{y}_j^ε as in (6.14). As noted earlier, the constant κ_1 of (5.7) may be taken uniform in j and $\{\mathbf{y}_j^\varepsilon\}$ is bounded in $L_2([T - \varepsilon, T] \times \partial_a \mathcal{R})$. By (5.7), then, there is a constant C_ε such that the solutions v_j of (6.7) with $\mathbf{B}'v_j = \mathbf{y}_j^\varepsilon$ satisfy

$$(6.15) \quad \|v_j(T - \varepsilon, \cdot)\|_{H^4(\mathcal{R})} \leq C_\varepsilon.$$

We can then select a subsequence (again denoted by $\{\hat{\mathbf{y}}_j^\varepsilon, v_j, \cdot \cdot \cdot\}$) for which $\{v_j(T - \varepsilon, \cdot)\}$ converges weakly in $H^4(\mathcal{R})$, i.e., $v_j(T - \varepsilon, \cdot) \rightharpoonup v_*$. Let v_* be the solution in $[-\varepsilon, T - \varepsilon] \times \mathcal{R}$ of

$$(6.16) \quad -a\dot{v}_* = \mathbf{L}v_*, \quad \mathbf{B}v_* = \mathbf{0}, \quad v_*(T - \varepsilon, \cdot) = v_*,$$

then define $\hat{v}_j^\varepsilon, \hat{v}_*^\varepsilon$ by translation, as in (6.14), and let $w_j := \hat{v}_*^\varepsilon - \hat{v}_j^\varepsilon$. Then w_j satisfies (much as with (6.6))

$$(6.17) \quad -a\dot{w}_j = \mathbf{L}w_j + f_j, \quad \mathbf{B}w_j = g_j, \quad w_j(T, \cdot) = \mathbf{w}_j$$

where f_j, g_j are defined as in (6.5):

$$\begin{aligned} f_j &:= (a_j - a)\hat{v}_j^\varepsilon + \nabla \cdot (A_j - A)\nabla \hat{v}_j^\varepsilon - (q_j - q)\hat{v}_j^\varepsilon && \text{in } \mathcal{Q}, \\ g_j &:= (\alpha_j - \alpha)\hat{v}_j^\varepsilon + (\beta_j A_j - \beta A)\mathbf{n} \cdot \nabla \hat{v}_j^\varepsilon && \text{on } \mathcal{S}. \end{aligned}$$

and

$$\mathbf{w}_j := v_* - v_j(T - \varepsilon, \cdot).$$

As in the proof of Lemma 6.2, we have $\hat{v}_j^\varepsilon(t, \cdot)$ bounded, uniformly in j and t ($0 \leq t \leq T$). Thus, as $\sigma_j \rightarrow \sigma$ in Σ we have $f_j \rightarrow 0$ in $H^{2,1}(\mathcal{Q})$ and $g_j \rightarrow 0$ in $H^{2,1}(\mathcal{S})$. Since $\mathbf{w}_j \rightarrow 0$ in $H^4(\mathcal{R})$, we have, by the compactness of the embedding, $\mathbf{w}_j \rightarrow 0$ in, e.g., $H^3(\mathcal{R})$. Once again we apply Theorem 6.2 of [11] with $r = \frac{3}{4}$ — and conclude, as there, that $\mathbf{B}'_j \hat{v}_j^\varepsilon \rightarrow \mathbf{B}' \hat{v}_*^\varepsilon$ in \mathcal{Y} as $j \rightarrow \infty$. Since $\mathbf{B}'_j \hat{v}_j^\varepsilon = \hat{\mathbf{y}}_j^\varepsilon \rightarrow \hat{\mathbf{y}}^\varepsilon$, this shows that $\hat{\mathbf{y}}^\varepsilon = \mathbf{B}' \hat{v}_*^\varepsilon$ and so is in $\mathcal{M}_0(\sigma)$. Thus \mathbf{y}_∞ , being $\lim_{\varepsilon \rightarrow 0^+} \hat{\mathbf{y}}^\varepsilon$, is in the closure $\mathcal{M} = \mathcal{M}(\sigma)$ and so must be the unique *optimal* (minimum \mathcal{Y} -norm) control \mathbf{y} in \mathbf{S} steering \mathbf{x} to \mathbf{z} . \square

Remark 6.4. If Theorem 5.3 could be modified to cover time-dependent equations (note Remark 5.5), then a somewhat more cumbersome construction (along the lines of the proof that $\mathcal{M} \subset \mathcal{M}$ for Theorem 5.3) could have been used in the last portion of the proof above to show \mathbf{y}_∞ is in \mathcal{M} and so, therefore, is in \mathcal{M} for time-dependent diffusion equations as well. \square

REFERENCES

- [1] C. O. BLOOM AND N. D. KAZARINOFF, *Energy decay in inhomogeneous media*, Proc. Symp. on Analysis (Rio de Janeiro, August, 1972), Hermann, Paris.
- [2] ———, *Energy decays locally even if total energy grows algebraically with time*, Bull. Amer. Math. Soc., 79 (1973), pp. 969–972.
- [3] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, vol. 1, Interscience, New York, 1953.
- [4] S. DOLECKI AND D. L. RUSSELL, *A general theory of observation and control*, this Journal, 15 (1977), pp. 185–220.
- [5] R. EDWARDS, *Functional Analysis*, Holt, Rhinehart and Winston, New York, 1965.
- [6] H. O. FATTORINI AND D. L. RUSSELL, *Uniform bounds on biorthogonal functions for real exponentials with applications to the control theory of parabolic equations*, Quart. Appl. Math., (1974), pp. 45–69.
- [7] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rhinehart and Winston, New York, 1969.

- [8] K. GRAHAM AND D. L. RUSSELL, *Boundary value control of the wave equation in a spherical region*, this Journal, 13 (1975), pp. 174–196.
- [9] E. HEWITT AND K. STROMBERG, *Real and Abstract Analysis*, Springer-Verlag, New York, 1965.
- [10] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, vol. I, Springer-Verlag, New York, 1972.
- [11] ———, vol. II, Springer-Verlag, New York, 1972.
- [12] C. S. MORAWETZ, *Exponential decay of solutions of the wave equation*, Comm. Pure Appl. Math., 19 (1966), pp. 439–444.
- [13] C. S. MORAWETZ, J. V. RALSTON AND W. A. STRAUSS, *Decay of solutions of the wave equation outside non-trapping obstacles*, (preprint: 1976), to appear.
- [14] R. E. A. C. PALEY AND N. WIENER, *The Fourier Transform in the Complex Domain*, Colloquium Publications, vol. 19, American Mathematical Society, New York, 1934.
- [15] J. QUINN AND D. L. RUSSELL, *Asymptotic stability and energy decay rates for solutions of hyperbolic equations with boundary damping*, U. Wisc. MRC Report 1575.
- [16] J. V. RALSTON, *Solutions of the wave equation with localized energy*, Comm. Pure Appl. Math. 22 (1969), pp. 807–823.
- [17] ———, private communication.
- [18] R. REDHEFFER, *Elementary remarks on completeness*, Duke Math. J., 35 (1968), pp. 103–116.
- [19] D. L. RUSSELL, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Stud. Appl. Math., LII (1973), pp. 189–211.
- [20] ———, *Exact boundary value controllability theorems for wave and heat processes in star-complemented regions*, Differential Games and Control Theory, Marcel Dekker, New York, 1974.
- [21] L. SCHWARTZ, *Étude des Sommes d'Exponentielles*, 2^{me} ed., Hermann, Paris, 1959.
- [22] T. I. SEIDMAN, *Problems of boundary control and observation for diffusion processes*, Math. Res. Rep: 73–10, UMBC, 1973.
- [23] ———, *Observation and prediction for the heat equation, III*, J. Differential Equations, 20 (1976), pp. 18–27.
- [24] ———, *Boundary observation and control for the heat equation*, Calculus of Variations and Control Theory, Academic Press, New York, 1976.
- [25] ———, *Observation and prediction for the heat equation, IV: patch observability and controllability*, this Journal, 15 (1977), pp. 412–427.
- [26] N. SHENK AND D. THOE, *Outgoing solutions of $(-\Delta + q - k^2)u = f$ in an exterior domain*, J. Math. Anal. Appl., 31 (1970), pp. 81–116.
- [27] R. TRIGGIANI, *On the relationship between first and second order controllable systems in Banach spaces*, to appear.
- [28] E. C. ZACHMANOGLOU, *The decay of solutions of the initial-boundary value problem for hyperbolic equations*, J. Math. Anal. Appl., 13 (1966), pp. 504–515.
- [29] W. C. CHEWNING AND T. I. SEIDMAN, *A convergent scheme for boundary control of the heat equation*, this Journal 15 (1977), pp. 64–72.

EXACT BOUNDARY VALUE CONTROLLABILITY OF A CLASS OF HYPERBOLIC EQUATIONS*

JOHN LAGNESE†

Abstract. Let $c(t)$ be a real-valued function which is analytic for $t \geq 0$ and Ω be a bounded, open set in R^n with smooth boundary. Sufficient conditions are given which insure that control processes modeled by partial differential equations of the form

$$\frac{\partial^2 u}{\partial t^2} - \sum_{i=1}^n \frac{\partial^2 u}{\partial x_i^2} + c(t)u = 0$$

in the cylinder $\Omega \times [0, \infty)$ are exactly controllable in any finite time T which exceeds the diameter of Ω by control forces applied on the wall of the cylinder.

1. Introduction. This paper is concerned with exact boundary value control of solutions of equations of the form

$$(1.1) \quad \begin{aligned} L_n u &= \frac{\partial^2 u}{\partial t^2} - \Delta_n u + c(t)u = 0, \\ \Delta_n &= \sum_{i=1}^n \frac{\partial^2}{\partial x_i^2}, \quad n \geq 2, \end{aligned}$$

where $c(t)$ is a real-valued analytic function for $t \geq 0$. Let Ω be a bounded, open set in R^n with smooth boundary. Initial conditions

$$(1.2) \quad u(0, x) = f_0(x), \quad \frac{\partial u}{\partial t}(0, x) = g_0(x), \quad x \in \Omega,$$

are given, and a solution of (1.1), (1.2) is allowed to evolve in a cylinder $[0, T] \times \Omega$ subject to a boundary control

$$(1.3) \quad F = \alpha u + \beta \frac{\partial u}{\partial \nu} \quad \text{on } [0, T] \times \partial\Omega,$$

($\alpha^2 + \beta^2 > 0$, ν = outward unit normal). The object is to select the control function F such that, after a specified time T and for a specified state (f_1, g_1) , it will be the case that

$$(1.4) \quad u(T, x) = f_1(x), \quad \frac{\partial u}{\partial t}(T, x) = g_1(x), \quad x \in \Omega.$$

If, for arbitrarily chosen initial data (f_0, g_0) and final data (f_1, g_1) taken from some *specified* dense subset (the set of controllable states) of $(L^2(\Omega))^2$, a control F exists such that (1.4) is satisfied for *some* finite time T , the operator L_n is said to be exactly controllable in time T .

The pioneering work on problems of this type was done by D. L. Russell [13]–[16] who considered the question of boundary control (both exact and approximate) of ordinary wave operators. For a general region Ω , Russell proved in [16] that the wave equation is exactly controllable in some finite time T exceeding the diameter of Ω and that the set of controllable states includes $H^2(\Omega) \times H^1(\Omega)$. T is unspecified if n is even but, if n is odd, exact control is possible in any time $T > \text{diam } \Omega$. In the

* Received by the editors October 13, 1977.

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

special case when Ω is a sphere and $\alpha = 0$ in (1.3), Graham and Russell [5] proved that exact control is possible in any time $T > \text{diam } \Omega$ regardless of the parity of n , and that the set of controllable states includes $H^1(\Omega) \times L^2(\Omega)$.

For certain hyperbolic operators other than the ordinary wave operators, exact boundary control was established in [10]. The class of operators consisted of those of the form (1.1) having the property that L_N satisfies Huygens' principle in the sense of Hadamard's "minor premise" for some N . In addition to the ordinary wave operators, examples are the EPD (Euler–Poisson–Darboux) operators in self-adjoint form, defined by

$$(1.5) \quad \frac{\partial^2 u}{\partial t^2} - \Delta_n u + \frac{\lambda}{(t+1)^2} u,$$

$$(1.6) \quad \lambda = m(m+1), \quad m = 1, 2, \dots$$

(1.5), (1.6) do not exhaust the class of admissible operators (see [8]). For such operators, and for a general region Ω , it was proved that the set of controllable states includes $H^2(\Omega) \times H^1(\Omega)$ and that exact control can always be achieved in any time $T > \text{diam } \Omega$, regardless of the parity of n .

In this paper we will extend the results of [10] to a much larger class of operators. This class is identified in terms of the asymptotic behavior as $t \rightarrow +\infty$ of the fundamental solution of L_n . We will show that the set of controllable states includes $H^2(\Omega) \times H^1(\Omega)$ and that exact control can be effected in any time $T > \text{diam } \Omega$. Example 4.1 will show that the class of operators under consideration includes those treated in [10]. Additional examples are

$$\frac{\partial^2 u}{\partial t^2} - \Delta_n u + c^2 u, \quad c = \text{const.} > 0,$$

and the EPD operators (1.5) for general values of $\lambda > -\frac{1}{4}$, $\lambda \neq m^2 - \frac{1}{4}$, $m = 0, 1, \dots$.

The method of proof consists of first establishing exact controllability in *some* finite time T , and then using analytic continuation to show that this implies exact controllability for any $T > \text{diam } \Omega$. It is the second part of the proof where the analyticity of $c(t)$ is essential, since we have to effect an analytic continuation of the fundamental solution of L_n . The method used to establish exact controllability in some finite time T is similar to that introduced by Russell in [16] and later utilized in [17], but with one significant difference: In [16], [17], it was necessary to establish that the energy

$$E(u, t, S) = \int_S \left[\left| \frac{\partial u}{\partial t}(t, x) \right|^2 + \sum_{i=1}^n \left| \frac{\partial u}{\partial x_i}(t, x) \right|^2 \right] dx$$

converges to zero as $t \rightarrow +\infty$ for each bounded set S and each solution u on $[0, \infty) \times R^n$ having finite initial energy. We make no such assumption and, in Example 4.4, show that our results apply to operators for which $E(u, t, S) \not\rightarrow 0$ as $t \rightarrow +\infty$.

We should also mention the work of Russell [17] in which the control F acts on only a portion of the boundary. If this portion is "star-complemented," then the wave equation is exactly controllable in some finite time T . How large T must be is not determined, but it is likely that there is no simple connection between T and Ω as in the case when the control acts over the entire boundary. (See also Fattorini [3] for an interesting example concerning the "star-complemented" condition.)

Our main results are stated and proved in § 3, and examples are discussed in § 4. Section 2 is devoted to preliminary material dealing with fundamental solutions and representation of solutions.

2. Preliminaries. We set $R_+ = [0, \infty)$, $R_+^{n+1} = R_+ \times R^n$. Under the assumption that $c(t)$ is analytic on R_+ , (1.1) admits a fundamental solution (in the sense of Hadamard [6]) of the form

$$v(t, x; \tau, \eta) = \begin{cases} U\Gamma^{-p} & n \text{ even,} \\ U\Gamma^{-p} + W \log \Gamma, & n \text{ odd,} \end{cases}$$

where $p = (n-1)/2$ and

$$\Gamma = (t-\tau)^2 - \|x-\eta\|^2, \quad \|x-\eta\|^2 = \sum_{i=1}^n (x_i - \eta_i)^2.$$

U and W are analytic functions of $(t, x; \tau, \eta)$ on $(R_+^{n+1})^2$. They admit expansions of the form

$$(2.1) \quad U = \begin{cases} \sum_{\nu=0}^{\infty} U_{\nu}(t, \tau) \Gamma^{\nu}, & n \text{ even,} \\ \sum_{\nu=0}^{p-1} U_{\nu}(t, \tau) \Gamma^{\nu}, & n \text{ odd,} \end{cases}$$

$$(2.2) \quad W = \sum_{\nu=0}^{\infty} W_{\nu}(t, \tau) \Gamma^{\nu}.$$

The coefficients in the expansions are analytic on $(R_+)^2$ and are determined as the unique solutions of

$$(2.2a) \quad U_0 \equiv 1, \quad (t-\tau) \frac{\partial U_{\nu}}{\partial t} + \nu U_{\nu} = \frac{1}{4(p-\nu)} L_n U_{\nu-1},$$

$$1 \leq \nu < \begin{cases} \infty, & n \text{ even,} \\ p, & n \text{ odd,} \end{cases}$$

$$(2.2b) \quad (t-\tau) \frac{\partial W_0}{\partial t} + p W_0 = \frac{1}{4} L_n U_{p-1},$$

$$(t-\tau) \frac{\partial W_{\nu}}{\partial t} + (p+\nu) W_{\nu} = -\frac{1}{4\nu} L_n W_{\nu-1}, \quad \nu \geq 1,$$

which are smooth at $t = \tau$. If K is any compact subset of R_+^{n+1} , the expansions (2.1), (2.2) converge at least for all $(t, x; \tau, \eta) \in K \times K$ satisfying $|\Gamma| < \delta$ for some $\delta > 0$ depending only on K . Note that $v(t, x; \tau, \eta)$ depends on x, η only in terms of $x - \eta$.

We shall later consider $U(t, x; 0, \eta)$, $U(0, x; t, \eta)$, $W(t, x; 0, \eta)$ and $W(0, x; t, \eta)$ for complex values of t . Because of the analyticity cited above, for each compact $K \subset R^n$ these functions may be analytically continued as functions of (t, x, η) (possibly complex valued) to a region $\Sigma_0 \times K \times K$, where $\Sigma_0 = \Sigma_0(K)$ is an open set in the complex ζ -plane containing the half-line $\text{Re } \zeta = t \geq 0, \text{Im } \zeta = 0$. The same is true, of course, of any derivative of U and W . We also note that if $T > \text{diam } K$ and $\zeta \in S_T$, where

$$S_T = \left\{ \zeta \mid \zeta = T + z, |\arg z| \leq \frac{\pi}{4} \right\},$$

then for all $(x, \eta) \in K \times K$,

$$\text{Re}(\zeta^2 - \|x - \eta\|^2) \geq (1 - \lambda^2) T^2$$

where $0 < \lambda < 1$ satisfies $\lambda T \geq \text{diam } K$. It follows that $(t^2 - \|x - \eta\|^2)^{-p}$ has an analytic continuation from $[T, \infty) \times K \times K$ into $S_T \times K \times K$. (When n is even, so that p is half an odd integer, we choose that value of $(\xi^2 - \|x - \eta\|^2)^{1/2}$ which has positive real part.) Thus the functions $v(t, x; 0, \eta)$ and $v(0, x; t, \eta)$ have analytic continuations from $[T, \infty) \times K \times K$ into a region $\Sigma_T \times K \times K$, where $\Sigma_T(K)$ is an open set in the complex ζ -plane containing the half-line $\text{Re } \zeta = t \geq T, \text{Im } \zeta = 0$. The same is true of any derivative of v .

With the aid of the fundamental solution for L_n , one can write down the solution to the Cauchy problem

$$(2.3) \quad L_n u = 0 \quad \text{in } R_+^{n+1},$$

$$(2.4) \quad u(0, x) = f(x), \quad \frac{\partial u}{\partial t}(0, x) = g(x) \quad \text{in } R^n.$$

Case I. n even. In this case the solution is given by [6]:

$$(2.5) \quad u(t, x) = \frac{(-1)^{n/2}}{\pi \omega_{n-1}} \text{p.f.} \int_{\|x - \eta\| \leq t} \left[f(\eta) \frac{\partial v}{\partial \tau}(t, x; 0, \eta) - g(\eta) v(t, x; 0, \eta) \right] d\eta$$

provided f and g are sufficiently smooth (depending on n). ω_{n-1} denotes the surface area of the unit sphere in R^{n-1} and p.f. denotes the "finite part" of the divergent integral [2, p. 306].

Suppose that f and g have support in a bounded, open set Ω and that $t > \text{diam } \Omega$. For each $x \in \bar{\Omega}$ the set $\{\eta \mid \|x - \eta\| \leq t\}$ then contains Ω . Thus for such t and x the integral in (2.5) is convergent and (2.5) becomes

$$(2.6) \quad u(t, x) = \frac{(-1)^{n/2}}{\pi \omega_{n-1}} \int_{\Omega} \left[f(\eta) \frac{\partial v}{\partial \tau}(t, x; 0, \eta) - g(\eta) v(t, x; 0, \eta) \right] d\eta, \quad x \in \bar{\Omega}, \quad t > \text{diam } \Omega.$$

It follows that u is analytic on $(d, \infty) \times \bar{\Omega}$ where $d = \text{diam } \Omega$.

We shall also have to solve the backwards Cauchy problem

$$(2.7) \quad L_n \hat{u} = 0 \quad \text{in } [0, T] \times R^n,$$

$$(2.8) \quad \hat{u}(T, x) = \hat{f}(x), \quad \frac{\partial \hat{u}}{\partial t}(T, x) = \hat{g}(x), \quad x \in R^n.$$

If \hat{f}, \hat{g} are sufficiently smooth, the unique solution is

$$(2.9) \quad \hat{u}(t, x) = \frac{(-1)^{n/2+1}}{\pi \omega_{n-1}} \text{p.f.} \int_{\|x - \eta\| \leq T-t} \left[\hat{f}(\eta) \frac{\partial v}{\partial \tau}(t, x; T, \eta) - \hat{g}(\eta) v(t, x; T, \eta) \right] d\eta.$$

If \hat{f} and \hat{g} have support in a bounded, open set Ω , if $T > \text{diam } \Omega$ and $t < T - \text{diam } \Omega$, and if $x \in \bar{\Omega}$, then (2.9) becomes

$$(2.10) \quad \hat{u}(t, x) = \frac{(-1)^{n/2+1}}{\pi \omega_{n-1}} \int_{\Omega} \left[\hat{f}(\eta) \frac{\partial v}{\partial \tau}(t, x; T, \eta) - \hat{g}(\eta) v(t, x; T, \eta) \right] d\eta, \quad x \in \bar{\Omega}, \quad t < T - \text{diam } \Omega,$$

hence \hat{u} is analytic on $[0, T - d) \times \bar{\Omega}$, $d = \text{diam } \Omega$.

Case II. n odd. The solution $u(t, x)$ of (2.3), (2.4) satisfies [4]

$$(2.11) \quad \begin{aligned} & (-1)^{(n-1)/2} \omega_{n-1} u(t, x) \\ &= \int_{\|x-\eta\| \leq t} \left[f(\eta) \frac{\partial W}{\partial \tau}(t, x; 0, \eta) - g(\eta) W(t, x; 0, \eta) \right] d\eta \\ & \quad - \text{p.l.} \int_{\|x-\eta\| \leq t} \left[f(\eta) \frac{\partial v}{\partial \tau}(t, x; 0, \eta) - g(\eta) v(t, x; 0, \eta) \right] d\eta \end{aligned}$$

provided f and g are sufficiently smooth. p.l. denotes the ‘‘logarithmic part’’ of the divergent integral. It is important to remark that the p.l. of a divergent integral depends on the values of the integrand only in an arbitrarily small neighborhood of the boundary of the region of integration. Thus if f and g have support in a bounded, open set Ω , then the p.l. of the divergent integral vanishes whenever $t > \text{diam } \Omega$ and $x \in \bar{\Omega}$. In this case

$$(2.12) \quad \begin{aligned} u(t, x) = \frac{(-1)^{(n-1)/2}}{\omega_{n-1}} \int_{\Omega} \left[f(\eta) \frac{\partial W}{\partial \tau}(t, x; 0, \eta) \right. \\ \left. - g(\eta) W(t, x; 0, \eta) \right] d\eta, \quad x \in \bar{\Omega}, \quad t > \text{diam } \Omega, \end{aligned}$$

and u is analytic in $(d, \infty) \times \bar{\Omega}$, $d = \text{diam } \Omega$.

Similarly, for the problem (2.7), (2.8), assuming \hat{f} and \hat{g} have support in Ω and $T > \text{diam } \Omega$,

$$(2.13) \quad \begin{aligned} \hat{u}(t, x) = \frac{(-1)^{(n+1)/2}}{\omega_{n-1}} \int_{\Omega} \left[\hat{f}(\eta) \frac{\partial W}{\partial \tau}(t, x; T, \eta) \right. \\ \left. - \hat{g}(\eta) W(t, x; T, \eta) \right] d\eta, \quad x \in \bar{\Omega}, \quad t < T - \text{diam } \Omega. \end{aligned}$$

The representations (2.5), (2.11) are valid only if the data (f, g) is sufficiently smooth. However, we shall have to solve the Cauchy problem for data in $H^2(R^n) \times H^1(R^n)$. One has the following existence-uniqueness theorem (see, e.g. [1, p. 669] or [18, p. 200]).

THEOREM. Assume $f \in H^r(R^n)$, $g \in H^{r-1}(R^n)$, where $r \geq 2$ is an integer. There is a unique function u defined on R_+^{n+1} such that

- (i) $(\partial^j u)/(\partial t^j) \in L^2([0, t], H^{r-j}(R^n))$ for every $t > 0$, $j = 0, 1, \dots, r$.
- (ii) The map $t \rightarrow (u(t, \cdot), (\partial u/\partial t)(t, \cdot))$ is continuous from R^+ to $H^r(R^n) \times H^{r-1}(R^n)$.
- (iii) u satisfies (2.3), (2.4).

In addition, for each $T > 0$ there is a constant $K = K(r, T)$ such that for $0 \leq t \leq T$,

$$(2.14) \quad \sum_{j=0}^r \left\| \frac{\partial^j u}{\partial t^j}(t, \cdot) \right\|_{r-j, R^n} \leq K (\|f\|_{r, R^n} + \|g\|_{r-1, R^n}).$$

Remark. $\|\cdot\|_{k, \Omega}$ denotes the norm in $H^k(\Omega)$:

$$\|u\|_{k, \Omega}^2 = \sum_{|\alpha| \leq k} \int_{\Omega} |D^\alpha u|^2 dx.$$

If f, g have support in a bounded, open set Ω then, by approximating (f, g) in $H^r(R^n) \times H^{r-1}(R^n)$ with functions in $C_0^\infty(\Omega) \times C_0^\infty(\Omega)$ and then passing to the limit using the energy inequality (2.14), one sees that the function u of the Theorem satisfies (2.6) or (2.12), depending on the parity on n , and is therefore analytic in $(d, \infty) \times \bar{\Omega}$, $d = \text{diam } \Omega$.

Clearly analogous results hold also for the backwards Cauchy problem and, in particular, (2.10), (2.13) are valid if $(\hat{f}, \hat{g}) \in H^r(R^n) \times H^{r-1}(R^n)$ has support in Ω .

3. Results and proofs. Let $v(t, x; \tau, \eta)$ be the fundamental solution for L_n . We introduce a family of functions $\{v_T | T \geq 0\}$ defined on $(R_+^{n+1})^2 \times R^n$ by

$$v_T(t, x; \tau, \eta; \xi) = \begin{cases} v(T, \xi; \tau, \eta) \frac{\partial v}{\partial \tau}(t, x; T, \xi) - \frac{\partial v}{\partial t}(T, \xi; \tau, \eta) v(t, x; T, \xi), & n \text{ even,} \\ W(T, \xi; \tau, \eta) \frac{\partial W}{\partial \tau}(t, x; T, \xi) - \frac{\partial W}{\partial t}(T, \xi; \tau, \eta) W(t, x; T, \xi), & n \text{ odd.} \end{cases}$$

The basic hypothesis for the main results of this section is:

(H) Let $D_i = \partial/(\partial x_i)$. There is an integer $k \geq 0$ and a bounded, open set $\Omega_1 \supset \bar{\Omega}$ such that, if $V_{T,k}$ denotes any one of the functions

$$D_i^k v_T, D_i^k \frac{\partial v_T}{\partial t}, D_i^k \frac{\partial v_T}{\partial \tau} \quad \text{or} \quad D_i^k \frac{\partial^2 v_T}{\partial t \partial \tau}, \quad i = 1, 2, \dots, n,$$

then

$$\lim_{T \rightarrow +\infty} \int_{\Omega_1^3} |V_{T,k}(0, x; 0, \eta; \xi)|^2 dx d\eta d\xi = 0.$$

THEOREM 3.1. Let $(f_0, g_0), (f_1, g_1)$ be given in $H^r(\Omega) \times H^{r-1}(\Omega)$, $r \geq 2$. Suppose $c(t)$ is analytic for $t \geq 0$ and that (H) is satisfied. Let T be any number greater than the diameter of Ω . Then there exists a control function $F \in H^s([0, T] \times \partial\Omega)$ ($s = r - \frac{3}{2}$ if $\beta \neq 0$, $s = r - \frac{1}{2}$ if $\beta = 0$) such that the unique solution of (1.1)–(1.3) in $[0, T] \times \Omega$ satisfies (1.4). Moreover, there is a constant $K = K(r, T)$ such that

$$(3.1) \quad \|F\|_{s, [0, T] \times \partial\Omega} \leq K(\|f_0\|_{r, \Omega} + \|f_1\|_{r, \Omega} + \|g_0\|_{r-1, \Omega} + \|g_1\|_{r-1, \Omega}).$$

The idea of the proof is to ignore the boundary conditions and to solve the initial value problem (2.3), (2.4) where (f, g) is some extension to R^n of (f_0, g_0) . One then tries to find an extension so that the solution satisfies (1.4). The control F is obtained as the restriction to $[0, T] \times \partial\Omega$ of this solution. This idea for proving exact controllability was introduced by Russell [16] in connection with the ordinary wave equation.

Theorem 3.1 will be deduced from the following result.

PROPOSITION. Suppose that $c(t)$ is analytic for $t \geq 0$ and that (H) is satisfied. Let $d = \text{diam } \Omega$. There is an at most countable set $E \subset (d, \infty)$ having d as its only possible limit point such that the following holds: for each $T \in (d, \infty)/E$ and each initial state (f_0, g_0) and final state (f_1, g_1) in $H^r(\Omega) \times H^{r-1}(\Omega)$, there is an extension $(f, g) \in H^r(R^n) \times H^{r-1}(R^n)$ of (f_0, g_0) such that the unique solution to the Cauchy problem (2.3), (2.4) satisfies (1.4). Moreover, there is a constant $K = K(r, T)$ such that

$$(3.2) \quad \|f\|_{r, R^n} + \|g\|_{r-1, R^n} \leq K(\|f_0\|_{r, \Omega} + \|f_1\|_{r, \Omega} + \|g_0\|_{r-1, \Omega} + \|g_1\|_{r-1, \Omega}).$$

Remark. The exceptional set E arises from our method of proof, and we are unable to rule out the possibility of such exceptional points or to show that values of T may exist for which there is no such extension.

Proof of Theorem 3.1. We set

$$V^r(\Omega) = H^r(\Omega) \times H^{r-1}(\Omega), \quad V_0^r(\Omega) = H_0^r(\Omega) \times H_0^{r-1}(\Omega).$$

If T is not one of the exceptional points, we may obtain the control F as the restriction to $[0, T] \times \partial\Omega$ of $\alpha u + \beta(\partial u/\partial \nu)$ where u is the solution given by the proposition. The inequality (3.1) is obtained from (2.14), (3.2) and the trace theorem [12, p. 39].

Suppose that T is one of the exceptional values, and let $\tilde{T} \in (d, T)/E$. Let $(\hat{f}, \hat{g}) \in V'(R^n)$ be an extension of (f_1, g_1) which satisfies

$$(3.3) \quad \|\hat{f}\|_{r,R^n} + \|\hat{g}\|_{r-1,R^n} \leq \text{const.} (\|f_1\|_{r,\Omega} + \|g_1\|_{r-1,\Omega}),$$

and let \hat{u} be the solution to the backwards Cauchy problem (2.7), (2.8). Let $(\hat{f}_0, \hat{g}_0) \in V'(\Omega)$ be the restriction to Ω of $(\hat{u}(0, \cdot), (\partial \hat{u}/\partial t)(0, \cdot))$. There is an extension $(f, g) \in V'(R^n)$ of $(f_0 - \hat{f}_0, g_0 - \hat{g}_0)$ such that

$$(3.4) \quad \|f\|_{r,R^n} + \|g\|_{r-1,R^n} \leq K(r, \tilde{T})(\|f_0 - \hat{f}_0\|_{r,\Omega} + \|g_0 - \hat{g}_0\|_{r-1,\Omega}),$$

and such that the solution \hat{u} of the Cauchy problem (2.3), (2.4) satisfies

$$(3.5) \quad \hat{u}(\tilde{T}, x) = \frac{\partial \hat{u}}{\partial t}(\tilde{T}, x) = 0, \quad x \in \Omega.$$

By repeatedly differentiating the equation (2.3) with respect to t we can, in fact, conclude from (3.5) that

$$(3.6) \quad \frac{\partial^j \hat{u}}{\partial t^j}(\tilde{T}, x) = 0, \quad x \in \Omega, \quad j = 0, 1, \dots, r.$$

Define

$$u(t, x) = \begin{cases} \tilde{u}(t, x) & \text{in } [0, \tilde{T}] \times \Omega, \\ 0 & \text{in } [\tilde{T}, T] \times \Omega. \end{cases}$$

(3.6) implies that $u \in H'([0, T] \times \Omega)$. Set $v = \hat{u} + u$ and let F be the restriction of $\alpha v + \beta(\partial v/\partial \nu)$ to $[0, T] \times \partial\Omega$. Then v satisfies (1.1)–(1.4) in $[0, T] \times \Omega$. From (3.3), (3.4), and the energy inequalities (2.14) applied to the backwards Cauchy problem, it follows that, for $0 \leq t \leq T$,

$$\begin{aligned} & \|v(t, \cdot)\|_{r,\Omega} + \left\| \frac{\partial v}{\partial t}(t, \cdot) \right\|_{r-1,\Omega} \\ & \leq \|\hat{u}(t, \cdot)\|_{r,R^n} + \left\| \frac{\partial \hat{u}}{\partial t}(t, \cdot) \right\|_{r-1,R^n} + \|\tilde{u}(t, \cdot)\|_{r,R^n} + \left\| \frac{\partial \tilde{u}}{\partial t}(t, \cdot) \right\|_{r-1,R^n} \\ & \leq K(r, T)[\|\hat{f}\|_{r,R^n} + \|\hat{g}\|_{r-1,R^n} + \|f\|_{r,R^n} + \|g\|_{r-1,R^n}] \\ & \leq K(r, T)[\|f_0\|_{r,\Omega} + \|g_0\|_{r-1,\Omega} + \|f_1\|_{r,\Omega} + \|g_1\|_{r-1,\Omega}]. \end{aligned}$$

It follows from the trace theorem that F satisfies (3.1).

Proof of Proposition. We assume that $k > 0$ in (H). The proof if $k = 0$ is similar (and simpler; cf. [10], [16]). Also, by using an argument entirely analogous to that used in the proof of Theorem 3.1, we need only consider the case $f_1 = g_1 \equiv 0$.

Let $T_0 > \text{diam } \Omega$ and $\delta > 0$ be such that

$$\Omega_\delta = \{x \in R^n \mid \exists \hat{x} \in \Omega \text{ with } \|x - \hat{x}\| < \delta\}$$

satisfies $\bar{\Omega}_\delta \subset \Omega_1$ and $\text{diam } \Omega_\delta < T_0$. There is a linear operator $E_\delta \in \mathcal{L}(V^p(\Omega), V_0^p(\Omega_\delta))$, every $p = 1, 2, \dots, r+2k$, such that $E_\delta(u_0, v_0) = (u_0, v_0)$ a.e. in Ω for every $(u_0, v_0) \in V^1(\Omega)$ [12, p. 38]. Let $(u_0, v_0) \in V^{r+2k}(\Omega)$ and set

$$(3.7) \quad (u_\delta, v_\delta) = E_\delta(u_0, v_0).$$

Extend (u_δ, v_δ) by 0 to R^n . Then

$$(3.8) \quad \|u_\delta\|_{p,R^n} \leq C \|u_0\|_{p,\Omega}, \quad p = 0, 1, \dots, r+2k,$$

$$(3.9) \quad \|v_\delta\|_{p-1,R^n} \leq C \|v_0\|_{p-1,\Omega}, \quad p = 1, 2, \dots, r+2k,$$

where C only depends on r, k, Ω and δ .

Let $w(t, x)$ be the solution of the Cauchy problem

$$\begin{aligned} L_n w &= 0 \quad \text{in } R_+^{n+1}, \\ w(0, x) &= u_\delta(x), \quad \frac{\partial w}{\partial t}(0, x) = v_\delta(x) \quad \text{in } R^n. \end{aligned}$$

$w \in H^{r+2k}([0, t] \times R^n)$, every $t > 0$, and is analytic on $[T_0, \infty) \times \bar{\Omega}_\delta$. Let $\phi \in C_0^\infty(\Omega_\delta)$ satisfy $\phi \equiv 1$ in Ω , let $T \geq T_0$ and v denote the solution of the backwards problem

$$\begin{aligned} Lv &= 0 \quad \text{in } [0, T] \times R^n, \\ v(T, x) &= \phi(x)w(T, x), \quad \frac{\partial v}{\partial t}(T, x) = \phi(x)\frac{\partial w}{\partial t}(T, x) \quad \text{in } R^n. \end{aligned}$$

Then $v \in C^\infty([0, T] \times R^n)$ and is analytic on $[0, \hat{T}_0] \times \bar{\Omega}_\delta$, $\hat{T}_0 < T - \text{diam } \Omega_\delta$. Let (\hat{u}_0, \hat{v}_0) be the restriction to $\bar{\Omega}$ of $(v(0, \cdot), (\partial v / \partial t)(0, \cdot))$ and set

$$z(t, x) = w(t, x) - v(t, x).$$

Then $z \in H^{r+2k}([0, T] \times R^n)$ satisfies

$$\begin{aligned} L_n z &= 0 \quad \text{in } [0, T] \times R^n, \\ z(0, x) &= u_0(x) - \hat{u}_0(x), \quad \frac{\partial z}{\partial t}(0, x) = \hat{v}_0(x) - v_0(x) \quad \text{in } \Omega, \\ z(T, x) &= \frac{\partial z}{\partial t}(T, x) = 0 \quad \text{in } \Omega. \end{aligned}$$

We now introduce the elliptic operator

$$A = (-1)^k \sum_{i=1}^n D_i^{2k},$$

and set $u = Az \in H^r([0, T] \times R^n)$. Since the coefficients in L_n are independent of x , u satisfies

$$\begin{aligned} L_n u &= 0 \quad \text{in } [0, T] \times R^n, \quad u(0, x) = Au_0(x) - A\hat{u}_0(x) \quad \text{in } \Omega, \\ \frac{\partial u}{\partial t}(0, x) &= Av_0(x) - A\hat{v}_0(x) \quad \text{in } \Omega, \quad u(T, x) = \frac{\partial u}{\partial t}(T, x) = 0 \quad \text{in } \Omega. \end{aligned}$$

We are going to prove that there is a finite set $E_{T_0} \subset [T_0, \infty)$ such that, for each $T \in [T_0, \infty) \setminus E_{T_0}$ and each $(f_0, g_0) \in V^r(\Omega)$, it is possible to choose $(u_0, v_0) \in V^{r+2k}(\Omega)$ such that $Au_0 - A\hat{u}_0 = f_0$, $Av_0 - A\hat{v}_0 = g_0$, and

$$(3.10) \quad \|u_0\|_{r+2k,\Omega} + \|v_0\|_{r+2k-1,\Omega} \leq K(r, T)(\|f_0\|_{r,\Omega} + \|g_0\|_{r-1,\Omega}).$$

Once this is proved, the extension (f, g) can be obtained as

$$f(x) = u(0, x), \quad g(x) = \frac{\partial u}{\partial t}(0, x),$$

and one has

$$\begin{aligned}
 & \|f\|_{r,R^n} + \|g\|_{r-1,R^n} \\
 & \leq \|w(0, \cdot)\|_{r+2k,R^n} + \left\| \frac{\partial w}{\partial t}(0, \cdot) \right\|_{r+2k-1,R^n} + \|v(0, \cdot)\|_{r+2k,R^n} + \left\| \frac{\partial v}{\partial t}(0, \cdot) \right\|_{r+2k-1,R^n} \\
 & \leq \|u_\delta\|_{r+2k,R^n} + \|v_\delta\|_{r+2k-1,R^n} + K(r, T) \left[\|w(T, \cdot)\|_{r+2k,R^n} + \left\| \frac{\partial w}{\partial t}(T, \cdot) \right\|_{r+2k-1,R^n} \right] \\
 & \leq K(r, T) [\|u_\delta\|_{r+2k,R^n} + \|v_\delta\|_{r+2k-1,R^n}] \\
 & \leq K(r, T) [\|f_0\|_{r,\Omega} + \|g_0\|_{r-1,\Omega}]
 \end{aligned}$$

in view of (3.8)–(3.10). The set E is given by $\bigcup_{m=1}^\infty E_{T_m}$ where $T_m \downarrow \text{diam } \Omega$.

In order to establish the existence of (u_0, v_0) , we are going to obtain an explicit representation for (\hat{u}_0, \hat{v}_0) . This is done using (2.6) and (2.10), or (2.12) and (2.13), depending on the parity of n . In either case we find, after a lengthy but straightforward calculation, that

$$\hat{u}_0(x) = -\frac{1}{(\pi\omega_{n-1})^2} \int_{\Omega_\delta} \left[u_\delta(\eta) \frac{\partial \hat{v}_T}{\partial \tau}(0, x; 0, \eta) - v_\delta(\eta) \hat{v}_T(0, x; 0, \eta) \right] d\eta, \quad x \in \bar{\Omega}, \quad (3.11)$$

$$\hat{v}_0(x) = -\frac{1}{(\pi\omega_{n-1})^2} \int_{\Omega_\delta} \left[u_\delta(\eta) \frac{\partial^2 \hat{v}_T}{\partial t \partial \tau}(0, x; 0, \eta) - v_\delta(\eta) \frac{\partial \hat{v}_T}{\partial t}(0, x; 0, \eta) \right] d\eta, \quad x \in \bar{\Omega}, \quad (3.12)$$

where

$$\hat{v}_T(t, x; \tau, \eta) = \int_{\Omega_\delta} \phi(\xi) v_T(t, x; \tau, \eta; \xi) d\xi.$$

According to our discussion of § 2, $\hat{v}_T(0, x; 0, \eta)$ and its derivatives appearing in (3.11), (3.12) are analytic functions of (T, x, η) on $[T_0, \infty) \times K \times K$ for any compact $K \subset \mathbb{R}^n$ and can be analytically continued into a region $\Sigma_{T_0} \times K \times K$, where $\Sigma_{T_0} = \Sigma_{T_0}(K)$ is an open set in the complex ζ -plane containing the half-line $\text{Re } \zeta = T \geq T_0$, $\text{Im } \zeta = 0$. We choose $K = \bar{\Omega}_1$ and imagine that such an analytic continuation has been carried out. Then (3.7) and (3.11), (3.12), with T replaced by ζ , define a family of linear mappings

$$\{K_\zeta | \zeta \in \Sigma_{T_0}\}$$

from every $V^p(\Omega)$ into $A(\bar{\Omega})$, the set of pairs of functions analytic on $\bar{\Omega}$.

LEMMA 3.1. $\{K_\zeta | \zeta \in \Sigma_{T_0}\}$ is a holomorphic family of compact operators from $V^p(\Omega)$ into $V^s(\Omega)$, $p, s = 1, 2, \dots$. Furthermore,

$$\|\hat{u}_0\|_{s,\Omega} + \|\hat{v}_0\|_{s-1,\Omega} \leq C(s, \zeta) (\|u_0\|_{0,\Omega} + \|v_0\|_{0,\Omega}),$$

where $C(s, \zeta)$ is continuous in $\zeta \in \Sigma_{T_0}$.

Proof. Setting

$$D_x^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}},$$

one has for $x \in \bar{\Omega}$

$$\begin{aligned} |D_x^\alpha \hat{u}_0(x)|^2 \leq & \frac{2}{(\pi\omega_{n-1})^4} \left\{ \int_{\Omega_\delta} |u_\delta(\eta)|^2 d\eta \int_{\Omega_\delta} \left| D_x^\alpha \frac{\partial \hat{v}_\zeta}{\partial \tau}(0, x; 0, \eta) \right|^2 d\eta \right. \\ & \left. + \int_{\Omega_\delta} |\hat{v}_\delta(\eta)|^2 d\eta \int_{\Omega_\delta} |D_x^\alpha \hat{v}_\zeta(0, x; 0, \eta)|^2 d\eta \right\} \end{aligned}$$

so that

$$\begin{aligned} \|\hat{u}_0\|_{s,\Omega}^2 & \leq C(s, \zeta) (\|u_\delta\|_{0,\Omega_\delta}^2 + \|v_\delta\|_{0,\Omega_\delta}^2) \\ & \leq C(s, \zeta) (\|u_0\|_{0,\Omega}^2 + \|v_0\|_{0,\Omega_\delta}^2), \end{aligned}$$

and similarly for $\|\hat{v}_0\|_{s-1,\Omega}$. Thus K_ζ is a bounded linear map from $V^p(\Omega)$ into $V^s(\Omega)$ for arbitrary $p, s = 1, 2, \dots$, and is therefore compact.

To verify that $\{K_\zeta | \zeta \in \Sigma_{T_0}\}$ is holomorphic, we note that K_ζ is the composition $K_\zeta = R_\delta L_\zeta E_\delta$ where L_ζ is the bounded linear operator from $V_0^p(\Omega_\delta)$ into $V^s(\Omega_\delta)$ defined by (3.11), (3.12) for $x \in \Omega_\delta$, and $R_\delta \in \mathcal{L}(V^s(\Omega_\delta), V^s(\Omega))$ is the restriction to Ω of functions in $V^s(\Omega_\delta)$. To show that $\{L_\zeta | \zeta \in \Sigma_{T_0}\}$ is holomorphic, it suffices to prove that this family is weakly holomorphic, that is,

$$F(\zeta) \equiv \sum_{|\alpha| \leq s} \int_{\Omega_\delta} D_x^\alpha \hat{u}(x) \overline{D_x^\alpha u(x)} dx + \sum_{|\alpha| \leq s-1} \int_{\Omega_\delta} D_x^\alpha \hat{v}(x) \overline{D_x^\alpha v(x)} dx$$

is holomorphic in Σ_{T_0} for each $(u_\delta, v_\delta) \in V_0^p(\Omega_\delta)$ and each $(u, v) \in V^s(\Omega_\delta)$, where $(\hat{u}, \hat{v}) = L_\zeta(u_\delta, v_\delta)$. However, this is immediate because of the analyticity of $D_x^\alpha \hat{v}_\zeta(0, x; 0, \eta)$, etc., in $\Sigma_{T_0} \times \bar{\Omega}_1 \times \bar{\Omega}_1$.

Let $(u_0, v_0) \in V^p(\Omega)$, $(\hat{u}_0, \hat{v}_0) = K_\zeta(u_0, v_0)$. Define the family $\{AK_\zeta | \zeta \in \Sigma_{T_0}\}$ by

$$AK_\zeta(u_0, v_0) = (A\hat{u}_0, A\hat{v}_0).$$

It follows from Lemma 3.1 that $\{AK_\zeta\}$ is also a holomorphic family of compact operators from $V^p(\Omega)$ into $V^s(\Omega)$ for $p, s = 1, 2, \dots$, and that

$$\|A\hat{u}_0\|_{s,\Omega} + \|A\hat{v}_0\|_{s-1,\Omega} \leq C(s, \zeta) (\|u_0\|_{0,\Omega} + \|v_0\|_{0,\Omega}).$$

$\{AK_\zeta\}$ converges to zero in $\mathcal{L}(V^{r+2k}(\Omega), V^r(\Omega))$ as $\zeta \rightarrow \infty$ along the positive real axis, as we shall now prove.

LEMMA 3.2. Let $(u_0, v_0) \in V^{r+2k}(\Omega)$ and set $(\hat{u}_0, \hat{v}_0) = K_T(u_0, v_0)$. Then

$$\|A\hat{u}_0\|_{r,\Omega} + \|A\hat{v}_0\|_{r-1,\Omega} \leq \alpha(r, k, T) (\|u_0\|_{r+2k,\Omega} + \|v_0\|_{r+2k-1,\Omega})$$

where $\alpha(r, k, T) \rightarrow 0$ as $T \rightarrow +\infty$.

Proof. The proof is given for n even. The proof for n odd is the same except for the notational change $v(t, x; \tau, \eta) \rightarrow W(t, x; \tau, \eta)$.

For any α with $|\alpha| \leq r$ and for $x \in \Omega$,

$$\begin{aligned} (3.13) \quad D_x^\alpha A\hat{u}_0(x) = & \frac{(-1)^{k+1}}{(\pi\omega_{n-1})^2} \sum_{i=1}^n \int_{\Omega_\delta} \left\{ u_\delta(\eta) \cdot D_x^\alpha D_i^{2k} \frac{\partial \hat{v}_T}{\partial \tau}(0, x; 0, \eta) \right. \\ & \left. - v_\delta(\eta) D_x^\alpha D_i^{2k} \hat{v}_T(0, x; 0, \eta) \right\} d\eta. \end{aligned}$$

Let $\beta = \alpha + k e^i$, $e^i = (0, \dots, 0, 1, 0, \dots, 0)$, the 1 appearing in the i th place. One has

$$D_x^\alpha D_i^{2k} \hat{v}_T(0, x; 0, \eta) = \int_{\Omega_\delta} \phi(\xi) \left\{ v(T, \xi; 0, \eta) D_x^\beta \left[D_i^k \frac{\partial v}{\partial \tau}(0, x; T, \xi) \right] - \frac{\partial v}{\partial t}(T, \xi; 0, \eta) D_x^\beta [D_i^k v(0, x; T, \xi)] \right\} d\xi.$$

Since $v(0, x; T, \xi)$ depends on x, ξ only as $x - \xi$, we may replace D_x^β in the above integral by $(-1)^{|\beta|} D_\xi^\beta$. We can then integrate by parts, apply Leibniz' rule and use $D_\xi^\gamma v(T, \xi; 0, \eta) = (-1)^{|\gamma|} D_\eta^\gamma v(T, \xi; 0, \eta)$. There results

$$D_x^\alpha D_i^{2k} \hat{v}_T(0, x; 0, \eta) = \sum_{\gamma \leq \beta} (-1)^{|\gamma|} \binom{\beta}{\gamma} D_\eta^\gamma \int_{\Omega_\delta} D_\xi^{\beta-\gamma} \phi(\xi) D_i^k v_T(0, x; 0, \eta; \xi) d\xi.$$

We substitute this expression, and a similar one for $D_x^\alpha D_i^{2k} (\partial \hat{v}_T / \partial \tau)(0, x; 0, \eta)$, into (3.13). Differentiations with respect to η may be transferred to u_δ and v_δ , and we obtain

$$\begin{aligned} D_x^\alpha A \hat{u}_0(x) &= \frac{(-1)^{k+1}}{(\pi \omega_{n-1})^2} \sum_{i=1}^n \sum_{\gamma \leq \beta} \binom{\beta}{\gamma} \\ &\cdot \left\{ \int_{\Omega_\delta} D_\eta^\gamma u_\delta(\eta) \left[\int_{\Omega_\delta} D_\xi^{\beta-\gamma} \phi(\xi) D_i^k \frac{\partial v_T}{\partial \tau}(0, x; 0, \eta; \xi) d\xi \right] d\eta \right. \\ &\quad \left. + \int_{\Omega_\delta} D_\eta^\gamma v_\delta(\eta) \left[\int_{\Omega_\delta} D_\xi^{\beta-\gamma} \phi(\xi) D_i^k v_T(0, x; 0, \eta; \xi) d\xi \right] d\eta \right\}. \end{aligned}$$

Therefore, for $|\alpha| \leq r$,

$$\begin{aligned} \|D_x^\alpha A \hat{u}_0\|_{0,\Omega}^2 &\leq \text{const.} \sum_{i=1}^n \sum_{\gamma \leq \beta} \int_{\Omega_\delta} |D_\xi^{\beta-\gamma} \phi(\xi)|^2 d\xi \\ &\cdot \left\{ \int_{\Omega_\delta} |D_\eta^\gamma u_\delta(\eta)|^2 d\eta \int_{\Omega_\delta} \left| D_i^k \frac{\partial v_T}{\partial \tau}(0, x; 0, \eta; \xi) \right|^2 dx d\eta d\xi \right. \\ &\quad \left. + \int_{\Omega_\delta} |D_\eta^\gamma v_\delta(\eta)|^2 d\eta \int_{\Omega_\delta} |D_i^k v_T(0, x; 0, \eta; \xi)|^2 dx d\eta d\xi \right\}. \end{aligned}$$

Hence

$$\begin{aligned} \|A \hat{u}_0\|_{r,\Omega} &\leq \text{const.} \left[\|D_i^k v_T(0, \cdot; 0, \cdot; \cdot)\|_{L^2(\Omega_1^3)} \right. \\ &\quad \left. + \left\| D_i^k \frac{\partial v_T}{\partial \tau}(0, \cdot; 0, \cdot; \cdot) \right\|_{L^2(\Omega_1^3)} \right] (\|u_\delta\|_{r+k,\Omega_\delta} + \|v_\delta\|_{r+k,\Omega_\delta}) \\ &\leq \alpha(r, k, T) (\|u_0\|_{r+2k,\Omega} + \|v_0\|_{r+2k-1,\Omega}) \end{aligned}$$

where $\alpha(r, k, T) \rightarrow 0$ as $T \rightarrow +\infty$. Similarly,

$$\begin{aligned} \|A \hat{v}_0\|_{r-1,\Omega} &\leq \text{const.} \left[\left\| D_i^k \frac{\partial v_T}{\partial t}(0, \cdot; 0, \cdot; \cdot) \right\|_{L^2(\Omega_1^3)} \right. \\ &\quad \left. + \left\| D_i^k \frac{\partial^2 v_T}{\partial t \partial \tau}(0, \cdot; 0, \cdot; \cdot) \right\|_{L^2(\Omega_1^3)} \right] (\|u_\delta\|_{r+k-1,\Omega_\delta} + \|v_\delta\|_{r+k-1,\Omega_\delta}) \\ &\leq \alpha(r, k, T) (\|u_0\|_{r+2k,\Omega} + \|v_0\|_{r+2k-1,\Omega}). \end{aligned}$$

The proof of the proposition is completed with the proof of the following result.

LEMMA 3.3. Let $(f_0, g_0) \in V'(\Omega)$. For each $T \geq T_0$, with the possible exception of a finite number of values, there is a unique $(u_0, v_0) \in V^{r+2k}(\Omega) \cap (H_0^k(\Omega))^2$ such that

$$(3.14) \quad (Au_0, Av_0) - AK_T(u_0, v_0) = (f_0, g_0).$$

Moreover,

$$(3.15) \quad \|u_0\|_{r+2k, \Omega} + \|v_0\|_{r+2k-1, \Omega} \leq C(r, T)(\|f_0\|_{r, \Omega} + \|g_0\|_{r-1, \Omega}).$$

Proof. Let $(\hat{u}_0, \hat{v}_0) = K_T(u_0, v_0)$. (3.14) is equivalent to

$$\begin{aligned} Au_0 - A\hat{u}_0 &= f_0, & u_0 &\in H^{r+2k}(\Omega) \cap H_0^k(\Omega), \\ Av_0 - A\hat{v}_0 &= g_0, & v_0 &\in H^{r+2k-1}(\Omega) \cap H_0^k(\Omega). \end{aligned}$$

Let $u_0^{(0)} = v_0^{(0)} \equiv 0$ and define $(u_0^{(m)}, v_0^{(m)})$, $m \geq 1$, recursively by

$$\begin{aligned} Au_0^{(m+1)} &= A\hat{u}_0^{(m)} + f_0, & u_0^{(m+1)} &\in H^{r+2k}(\Omega) \cap H_0^k(\Omega), \\ Av_0^{(m+1)} &= A\hat{v}_0^{(m)} + g_0, & v_0^{(m+1)} &\in H^{r+2k-1}(\Omega) \cap H_0^k(\Omega). \end{aligned}$$

These equations have unique solutions since the bilinear form

$$B(u, v) = \sum_{i=1}^n \int_{\Omega} D_i^k u \overline{D_i^k v} dx$$

associated with A satisfies

$$|B(u, u)| \geq c \|u\|_{k, \Omega}^2, \quad u \in H_0^k(\Omega)$$

for some constant $c > 0$. We apply the a priori inequality

$$(3.16) \quad \|u\|_{p+2k, \Omega} \leq C \|Au\|_{p, \Omega}, \quad u \in H^{p+2k}(\Omega) \cap H_0^k(\Omega),$$

to $u_0^{(m+1)}$ and $v_0^{(m+1)}$ to obtain

$$\begin{aligned} \|u_0^{(m+1)}\|_{r+2k, \Omega} &\leq C(\|A\hat{u}_0^{(m)}\|_{r, \Omega} + \|f_0\|_{r, \Omega}), \\ \|v_0^{(m+1)}\|_{r+2k-1, \Omega} &\leq C(\|A\hat{v}_0^{(m)}\|_{r-1, \Omega} + \|g_0\|_{r-1, \Omega}). \end{aligned}$$

C depends only on Ω , r , k and n . We next apply Lemma 3.2 and suppose that T is so large that $\rho = C\alpha(r, k, T) < 1$. For such T we have

$$\begin{aligned} (3.17) \quad & \|u_0^{(m+1)}\|_{r+2k, \Omega} + \|v_0^{(m+1)}\|_{r+2k-1, \Omega} \\ & \leq \rho(\|u_0^{(m)}\|_{r+2k, \Omega} + \|v_0^{(m)}\|_{r+2k-1, \Omega}) + C(\|f_0\|_{r, \Omega} + \|g_0\|_{r-1, \Omega}) \\ & \leq \frac{C}{1-\rho}(\|f_0\|_{r, \Omega} + \|g_0\|_{r-1, \Omega}). \end{aligned}$$

We may therefore extract a subsequence of $(u_0^{(m)}, v_0^{(m)})$ which will converge weakly in $V^{r+2k}(\Omega)$ to some $(u_0, v_0) \in V^{r+2k}(\Omega) \cap (H_0^k(\Omega))^2$. We keep the same notation for the subsequence. We have $(Au_0^{(m)}, Av_0^{(m)}) \rightarrow (Au_0, Av_0)$ weakly in $V'(\Omega)$ and, since AK_T is compact from $V^{r+2k}(\Omega)$ to $V'(\Omega)$, $(A\hat{u}_0^{(m)}, A\hat{v}_0^{(m)}) \rightarrow (A\hat{u}_0, A\hat{v}_0)$ strongly in $V'(\Omega)$. Hence $(Au_0^{(m)}, Av_0^{(m)}) \rightarrow (A\hat{u}_0, A\hat{v}_0) + (f_0, g_0)$ strongly in $V'(\Omega)$, $(u_0^{(m)}, v_0^{(m)}) \rightarrow (u_0, v_0)$ strongly in $V^{r+2k}(\Omega)$ and $(Au_0, Av_0) = (A\hat{u}_0, A\hat{v}_0) + (f_0, g_0)$. The inequality (3.15) follows from (3.17) and clearly is valid for any solution in $V^{r+2k}(\Omega) \cap (H_0^k(\Omega))^2$ of (3.14) provided T is sufficiently large (Lemma 3.2 and (3.16)). Lemma 3.3 is therefore proved for such values of T .

To obtain the result for other values of T we define an operator \mathcal{A} in $V^r(\Omega)$ by

$$D(\mathcal{A}) = \{(u, v) \in V^r(\Omega) \mid (u, v) \in V^{r+2k}(\Omega) \cap (H_0^k(\Omega))^2\},$$

$$\mathcal{A}(u, v) = (Au, Av), \quad (u, v) \in D(\mathcal{A}).$$

(3.16) implies that \mathcal{A} is a closed operator in $V^r(\Omega)$. Consider the family $\{\mathcal{A}_\zeta = \mathcal{A} - AK_\zeta \mid \zeta \in \Sigma_{T_0}\}$ of closed operators. Since $\{AK_\zeta \mid \zeta \in \Sigma_{T_0}\}$ is a holomorphic family of compact operators from $V^{r+2k}(\Omega)$ into $V^r(\Omega)$, $\{\mathcal{A}_\zeta\}$ is a holomorphic family of type (A) in the sense of [7, p. 375], that is $D(\mathcal{A}_\zeta) \equiv D(\mathcal{A})$ is independent of ζ and $\mathcal{A}_\zeta u = \mathcal{A}u - AK_\zeta u$ is holomorphic for every $u \in D(\mathcal{A})$ and $\zeta \in \Sigma_{T_0}$. From the first part of the proof we know that \mathcal{A}_ζ has a compact inverse (hence compact resolvent) for all sufficiently large real, positive values of ζ . Thus \mathcal{A}_ζ has compact resolvent for all $\zeta \in \Sigma_{T_0}$ and hence \mathcal{A}_ζ has a bounded inverse for all values of $\zeta \in \Sigma_{T_0}$ with the possible exception of a finite number of values in each compact subset of Σ_{T_0} ([7, Thms. VII-2.4 and VII-1.10]). In particular, \mathcal{A}_T has a compact inverse for all $T \geq T_0$ except, possibly, for a finite number of exceptional points. Assuming that T is not one of these points, we have the existence of a unique $(u_0, v_0) \in D(\mathcal{A})$ satisfying (3.14) and

$$(3.18) \quad \|u_0\|_{r,\Omega} + \|v_0\|_{r-1,\Omega} \leq C(r, T)(\|f_0\|_{r,\Omega} + \|g_0\|_{r-1,\Omega}).$$

The inequality (3.15) follows from (3.14), (3.16), (3.18) and Lemma 3.1, and the proof of Lemma 3.3 is complete.

4. Examples.

Example 4.1. Suppose the series expansions for U (n even) and W (n odd) contain only finitely many terms. This occurs exactly when L_n has the property that L_N satisfies Huygens' principle in the sense of Hadamard's "minor premise" for some (necessarily odd) N . The complete class of such operators has been explicitly determined in [8], [11].

Assume first of all that n is odd. Since its series expansion terminates, $W(t, x; \tau, \xi)$ is a polynomial in x and the same is therefore true of v_T and any of its derivatives. Thus for a sufficiently large integer k we will have $V_{T,k}(0, x; 0, \eta; \xi) \equiv 0$ in (H).

If n is even we will have

$$v(t, x; \tau, \eta) = \sum_{\nu=0}^q U_\nu(t, \tau) \Gamma^{\nu-p}, \quad p = \frac{n-1}{2},$$

for some integer q . We need the fact that

$$\sup_{(t,\tau)} \left\{ |U_\nu|, \left| \frac{\partial U_\nu}{\partial t} \right|, \left| \frac{\partial U_\nu}{\partial \tau} \right|, \left| \frac{\partial^2 U_\nu}{\partial t \partial \tau} \right| \right\} < +\infty, \quad \nu = 0, 1, \dots, q.$$

(For the class of operators under consideration in this example, it is known, in fact, that $U_\nu(t, \tau)$ is a symmetric, rational function which is $O(t^{-\nu})$ as $t \rightarrow +\infty$ for each fixed τ . See [9].) It then follows, for example, that

$$\begin{aligned} & \left| D_i^k v(0, x; T, \xi), \left| D_i^k \frac{\partial v}{\partial \tau}(0, x; T, \xi) \right| \right| \\ & = O(T^{2q-2p-k}) \quad \text{as } T \rightarrow +\infty, \end{aligned}$$

uniformly for x, ξ in a bounded set in R^n . Thus

$$|D_i^k v_T(0, x; 0, \eta; \xi)| = O(T^{4q-4p-k})$$

as $T \rightarrow +\infty$, uniformly for x, ξ, η in a bounded set in R^n . Similar growth estimates are

easily seen to hold for $(\partial v_T / \partial t)$, $(\partial v_T / \partial \tau)$, and $(\partial^2 v_T / \partial t \partial \tau)$, and so (H) is satisfied for a sufficiently large integer k . The conclusions of Theorem 3.1 therefore apply to the class of operators L_n having the property that L_N satisfies Huygens' principle for some N . This result was first obtained in [10] by a completely different method.

Example 4.2. We consider the operator defined by

$$(4.1) \quad L_n u = \frac{\partial^2 u}{\partial t^2} - \Delta_n u + c^2 u$$

where c is a positive constant. We are going to construct the fundamental solution $v(t, x; \tau, \eta)$ explicitly.

Case 1. n even. The coefficients $\{U_\nu\}$ in (2.1) are calculated to be

$$(4.2) \quad U_0 \equiv 1, \quad U_\nu(p) = \left(\frac{c^2}{4}\right)^\nu \frac{1}{\nu!(p-1)(p-2)\cdots(p-\nu)}, \quad \nu = 1, 2, \dots; \quad p = \frac{n-1}{2}.$$

Thus

$$v(t, x; \tau, \eta) \equiv v_p(\Gamma_n) = \frac{1}{\Gamma_n^p} \sum_{\nu=0}^{\infty} U_\nu(p) \Gamma_n^\nu,$$

$$\Gamma_n = (t - \tau)^2 - \sum_{i=1}^n (x_i - \eta_i)^2.$$

If $n = 2$, so that $p = \frac{1}{2}$, then

$$(4.3) \quad v(t, x; \tau, \xi) = \frac{1}{\Gamma_2^p} \sum_{\nu=0}^{\infty} \frac{(-1)^\nu c^{2\nu}}{(2\nu)!} \Gamma_2^\nu = \frac{1}{\Gamma_2^{\frac{1}{2}}} \cos c\sqrt{\Gamma_2}.$$

The fundamental solution for $n = 4, 6, \dots$, is now constructed in a recursive manner. From (4.2) we see that

$$U_\nu(p+1) = \left(1 - \frac{\nu}{p}\right) U_\nu(p).$$

Set

$$U_p(\Gamma) = \sum_{\nu=0}^{\infty} U_\nu(p) \Gamma^\nu.$$

We have

$$\begin{aligned} v_{p+1}(\Gamma) &= \frac{U_{p+1}(\Gamma)}{\Gamma^{p+1}} = \frac{1}{\Gamma^{p+1}} \sum_{\nu=\nu}^{\infty} \left(1 - \frac{\nu}{p}\right) U_\nu(p) \Gamma^\nu \\ &= \frac{1}{\Gamma^{p+1}} \left[U_p(\Gamma) - \frac{\Gamma}{p} \frac{d}{d\Gamma} U_p(\Gamma) \right] \\ &= -\frac{1}{p} \frac{d}{d\Gamma} \left[\frac{U_p(\Gamma)}{\Gamma^p} \right] = -\frac{1}{p} \frac{d}{d\Gamma} [v_p(\Gamma)]. \end{aligned}$$

Thus the relation between the fundamental solutions $v_{p+1}(\Gamma_{n+2})$ and $v_p(\Gamma_n)$ is

$$(4.4) \quad v_{p+1}(\Gamma_{n+2}) = -\frac{1}{p} \frac{d}{d\Gamma} [v_p(\Gamma)]|_{\Gamma=\Gamma_{n+2}}.$$

Case 2. n odd. From (2.2a), (2.2b) we calculate

$$U_0 \equiv 1, \quad U_\nu(p) = \left(\frac{c^2}{4}\right)^\nu \frac{1}{\nu!(p-1)(p-2)\cdots(p-\nu)}, \quad \nu = 1, \dots, p-1,$$

$$W_\nu(p) = \frac{1}{(p-1)!} \left(\frac{c^2}{4}\right)^p \left[\left(\frac{c^2}{4}\right)^\nu \frac{(-1)^\nu}{\nu!(p+\nu)!} \right].$$

Therefore

$$(4.5) \quad \begin{aligned} W(t, x; \tau, \eta) &\equiv W_p(\Gamma) \\ &= \frac{1}{(p-1)!} \left(\frac{c}{2\sqrt{\Gamma}}\right)^p \sum_{\nu=0}^{\infty} \left(\frac{c\sqrt{\Gamma}}{2}\right)^{p+2\nu} \frac{(-1)^\nu}{\nu!(p+\nu)!} \\ &= \frac{1}{(p-1)!} \left(\frac{c}{2\sqrt{\Gamma}}\right)^p J_p(c\sqrt{\Gamma}), \end{aligned}$$

where $\Gamma = \Gamma_n$ and J_p is the Bessel coefficient of order p . We recall that [19, p. 362]

$$(4.6) \quad \left| \frac{d^l}{dz^l} J_p(z) \right| \leq 1, \quad z \text{ real}, \quad l = 0, 1, \dots.$$

It follows from (4.3), (4.4), (4.5) and (4.6) that (H) holds with $k = 0$, and therefore the conclusions of Theorem 3.1 apply to the operator (4.1).

Example 4.3. We consider the general EPD operator in the form

$$L_n = \frac{\partial^2}{\partial t^2} - \Delta_n - \frac{\lambda}{(t+1)^2}$$

for real values of λ . We are going to prove that (H) holds for every value of $\lambda > -\frac{1}{4}$ with the possible exception of the values $\lambda = m^2 - \frac{1}{4}$, $m = 1, 2, \dots$.

Case 1. n even. From (2.2a) we calculate for $\nu = 1, 2, \dots$,

$$(4.7) \quad U_0 \equiv 1, \quad U_\nu(t, \tau) = \left(\frac{1}{4}\right)^\nu \frac{[\lambda, \nu]}{\nu!(p-1)(p-2)\cdots(p-\nu)} \cdot \frac{1}{(t+1)^\nu(\tau+1)^\nu},$$

where

$$[\lambda, \nu] = -\lambda(1 \cdot 2 - \lambda)(2 \cdot 3 - \lambda) \cdots (\nu(\nu-1) - \lambda), \quad \nu \geq 1.$$

If we now set

$$\lambda = \alpha(\alpha-1), \quad \alpha > \frac{1}{2},$$

we obtain

$$\begin{aligned} \Gamma^p v(t, x; \tau, \eta) &= \sum_{\nu=0}^{\infty} U_\nu(t, \tau) \Gamma^\nu \\ &= F(\alpha, 1-\alpha; 1-p; z), \quad |z| < 1, \end{aligned}$$

where

$$(4.8) \quad z = -\frac{\Gamma}{4(t+1)(\tau+1)}$$

and F is the hypergeometric series. This series converges if $|z| < 1$ and may be analytically continued to all z such that $|\arg(-z)| < \pi$. For $|z| > 1$ and $|\arg(-z)| < \pi$,

one has [19, p. 289]

$$(4.9) \quad v(t, x; \tau, \eta) = \frac{1}{\Gamma^p} \left\{ \frac{\Gamma(1-p)\Gamma(2\alpha-1)}{\Gamma(1-\alpha)\Gamma(\alpha+p-1)} (-z)^{-\alpha} F\left(\alpha, p+\alpha; 2\alpha; \frac{1}{z}\right) \right. \\ \left. + \frac{\Gamma(1-p)\Gamma(1-2\alpha)}{\Gamma(\alpha)\Gamma(p-\alpha)} (-z)^{\alpha-1} F\left(1-\alpha, p+1-\alpha; 2-2\alpha; \frac{1}{z}\right) \right\}$$

provided $\alpha \neq 1, \frac{3}{2}, 2, \frac{5}{2}, \dots$. (The case $\alpha = 1, 2, \dots$, is included in Example 4.1.) Since, as $T \rightarrow +\infty$,

$$(4.10) \quad \left(\frac{d}{dz}\right)^j F\left(a, b; c; \frac{1}{z}\right) = O(T^{-j-1}), \quad j = 1, 2, \dots,$$

it follows from (4.9) that as $T \rightarrow +\infty$,

$$v(T, \xi; 0, \eta) = O(T^{\alpha-2p-1}), \quad \frac{\partial v}{\partial t}(T, \xi; 0, \eta) = O(T^{\alpha-2p-2}), \\ D_i^k v(0, x; T, \xi) = O(T^{\alpha-2p-k-1}), \quad D_i^k \frac{\partial v}{\partial \tau}(0, x; T, \xi) = O(T^{\alpha-2p-k-2}),$$

and therefore

$$D_i^k v_T(0, x; 0, \eta; \xi) = O(T^{2\alpha-4p-k-3})$$

uniformly for x, η, ξ is bounded sets in R^n . Similarly

$$D_i^k \frac{\partial v_T}{\partial t}(0, x; 0, \eta; \xi) = O(T^{2\alpha-4p-k-4}), \\ D_i^k \frac{\partial v_T}{\partial \tau}(0, x; 0, \eta; \xi) = O(T^{2\alpha-4p-k-4}), \\ D_i^k \frac{\partial^2 v_T}{\partial t \partial \tau}(0, x; 0, \eta; \xi) = O(T^{2\alpha-4p-k-5})$$

as $T \rightarrow +\infty$, uniformly for x, η, ξ in bounded sets in R^n . Therefore (H) is satisfied for all $\lambda > -\frac{1}{4}$,

$$\lambda \neq \frac{1}{4}(2m-1)(2m+1), \quad m = 1, 2, \dots$$

Case 2. n odd. The coefficients U_ν are given by (4.7) for $\nu = 0, 1, \dots, p-1$, and from (2.2b) we calculate

$$W_\nu(t, \tau) = \frac{(-1)^\nu}{4^{\nu+p}} \frac{[\lambda, p+\nu]}{(p-1)!(p+\nu)!\nu!(t+1)^{p+\nu}(\tau+1)^{p+\nu}}, \quad \nu \geq 0.$$

Again we set $\lambda = \alpha(\alpha-1)$, $\alpha > \frac{1}{2}$. Then

$$(4.11) \quad W = \sum_{\nu=0}^{\infty} W_\nu \Gamma^\nu = \frac{[\lambda, p]}{4^p (p-1)! p! (t+1)^p (\tau+1)^p} F(p+\alpha, p-\alpha+1; p+1; z)$$

provided $|z| < 1$, where z is given by (4.8). For $|z| > 1$, $|\arg(-z)| < \pi$, one has the

analytic continuation

$$\begin{aligned}
 &F(p+\alpha, p-\alpha+1; p+1, z) \\
 (4.12) \quad &= \frac{\Gamma(p+1)\Gamma(2\alpha-1)}{\Gamma(p-\alpha+1)\Gamma(\alpha-1)}(-z)^{-p-\alpha}F\left(p+\alpha, \alpha; 2\alpha; \frac{1}{z}\right) \\
 &\quad + \frac{\Gamma(p+1)\Gamma(1-2\alpha)}{\Gamma(p+\alpha)\Gamma(-\alpha)}(-z)^{\alpha-p-1}F\left(p-\alpha+1, 1-\alpha; 2-2\alpha; \frac{1}{z}\right)
 \end{aligned}$$

provided $\alpha \neq 1, \frac{3}{2}, 2, \frac{5}{2}, \dots$. As in Case 1, one may conclude from (4.10)–(4.12) that (H) is satisfied for sufficiently large k .

Example 4.4. We give an example of an equation satisfying (H) and having the property that the energy in bounded sets does not necessarily tend to zero as $t \rightarrow +\infty$, even for solutions having compactly supported initial data. The equation is

$$(4.13) \quad Lu = \frac{\partial^2 u}{\partial t^2} - \Delta_3 u - \frac{20}{(t+1)^2} u = 0.$$

The energy in a set S at time t associated with a solution of (4.13) is

$$E(u, t, S) = \int_S \left\{ \left| \frac{\partial u}{\partial t}(t, x) \right|^2 + \sum_{i=1}^3 \left| \frac{\partial u}{\partial x_i}(t, x) \right|^2 \right\} dx.$$

The fundamental solution for L is

$$v(t, x; \tau, \eta) = \frac{1}{\Gamma} + W \log \Gamma$$

where

$$\begin{aligned}
 W &= W_0 + W_1 \Gamma + W_2 \Gamma^2 + W_3 \Gamma^3 \\
 &= -\frac{5}{(t+1)(\tau+1)}(1-9z+21z^2-14z^3),
 \end{aligned}$$

and where z is given by (4.8). Let u be the solution of (4.13) in R_+^4 satisfying

$$(4.14) \quad u(0, x) \equiv 0, \quad \frac{\partial u}{\partial t}(0, x) = g(x), \quad x \in R^3,$$

and assume g has support in a bounded, open set Ω . If x is in a bounded, open set S and if t is sufficiently large (depending on $\text{diam } \Omega$, $\text{diam } S$ and $\text{dist}(S, \Omega)$) the solution of (4.13), (4.14) is given by (2.12):

$$\begin{aligned}
 u(t, x) &= \frac{1}{2\pi} \int_{\Omega} g(\eta) W(t, x; 0, \eta) d\eta \\
 &= -\frac{5}{2\pi(t+1)} \int_{\Omega} g(\eta)(1-9\gamma+21\gamma^2-14\gamma^3) d\eta
 \end{aligned}$$

where

$$\gamma = -\frac{t^2 - \|x - \eta\|^2}{4(t+1)}.$$

Thus

$$\frac{\partial u}{\partial x_i}(t, x) = \frac{15}{4\pi(t+1)^2} \int_{\Omega} g(\eta)(x_i - \eta_i)(3-14\gamma+14\gamma^2) d\eta, \quad x \in S, \quad t \text{ large},$$

hence

$$\lim_{t \rightarrow +\infty} \frac{\partial u}{\partial x_i}(t, x) = \frac{105}{32\pi} \int_{\Omega} g(\eta)(x_i - \eta_i) d\eta, \quad x \in S,$$

and so $\lim_{t \rightarrow +\infty} E(u, t, s)$ cannot be zero for arbitrary $g \in C_0^\infty(\Omega)$.

REFERENCES

- [1] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, vol. II, Wiley-Interscience, New York, 1962.
- [2] R. E. EDWARDS, *Functional Analysis*, Holt, Rinehart and Winston, New York, 1965.
- [3] H. FATTORINI, *Estimates for sequences biorthogonal to certain complex exponentials and boundary control of the wave equation*, to appear.
- [4] K. O. FRIEDRICHS, *Die Verallgemeinerung der Riemannschen Methode auf eine beliebige gerade Anzahl von Dimensionen*, Nachr. Gesellschaft Wiss. Göttingen, Math.-Phys. Kl. II, (1928), pp. 172–177.
- [5] K. D. GRAHAM AND D. L. RUSSELL, *Boundary value control of the wave equation in a spherical region*, this Journal, 13 (1975), pp. 174–196.
- [6] J. HADAMARD, *Le Problème de Cauchy et les Équations aux Dérivées Partielles Lineaires Hyperboliques*, Hermann, Paris, 1932.
- [7] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
- [8] J. LAGNESE, *A solution of Hadamard's problem for a restricted class of operators*, Proc. Amer. Math. Soc., 19 (1968), pp. 981–988.
- [9] ———, *The structure of a class of Huygens' operators*, J. Math. Mech., 18 (1969), pp. 1195–1201.
- [10] ———, *Boundary value control of a class of hyperbolic equations in a general region*, this Journal, 15 (1977), pp. 973–983.
- [11] J. LAGNESE AND K. L. STELLMACHER, *A method of generating classes of Huygens' operators*, J. Math. Mech., 17 (1967), pp. 461–472.
- [12] J. L. LIONS AND E. MAGENES, *Non-homogeneous Boundary Value Problems and Applications*, vol. I, Springer-Verlag, New York, 1972.
- [13] D. L. RUSSELL, *On boundary-value controllability of linear symmetric hyperbolic systems*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967.
- [14] ———, *Boundary value control of higher dimensional wave equations*, Parts I and II, this Journal, 9 (1971), pp. 29–42 and pp. 401–419.
- [15] ———, *Control theory of hyperbolic equations related to certain questions in harmonic analysis and spectral theory*, J. Math. Anal. Appl., 40 (1972), pp. 336–368.
- [16] ———, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Studies in Appl. Math., 52 (1973), pp. 189–211.
- [17] ———, *Exact boundary value controllability theorems for wave and heat processes in star-complemented regions*, Differential Games and Control Theory, Roxin, Liu and Steinberg, eds., Marcel Dekker, New York, 1974.
- [18] S. L. SOBOLEV, *Applications of Functional Analysis in Mathematical Physics*, Translations of Mathematical Monographs, vol. 7, American Mathematical Society, Providence, RI, 1963.
- [19] E. T. WHITTAKER AND G. N. WATSON, *A Course in Modern Analysis*, 4th ed., University Press, Cambridge, England, 1927.