SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 1, 1979

A NOTE ON THE EXISTENCE OF A NASH EQUILIBRIUM POINT IN STOCHASTIC DIFFERENTIAL GAMES*

KENKO UCHIDA†

Abstract. Using the implicit function lemma of Beneš [SIAM J. Control, 1970] we derive a sufficient condition for the existence of a Nash equilibrium point in a feedback form in N-person stochastic differential games, which is an addendum to the existence theorem given by the author [SIAM J. Control, 1978].

1. Introduction. In the recent paper [1], we have proved that, if the Nash condition stated below is satisfied, there is a Nash equilibrium point in a feedback form in N-person nonzero sum stochastic differential games. The purpose of this paper is to give a necessary and sufficient condition for the Nash condition to hold, which can be regarded as the natural extension of the Issacs condition to the nonzero sum case.

2. The main theorem. We use the same notation as [1]. Let $\mathscr{C}$ be the space of continuous functions from $[0, 1]$ to R^m. x denotes a member of $\mathscr{C}$ and x_t denotes the value of x at t. $\mathscr{F}_t$ is the sigma field of $\mathscr{C}$ generated by $\{x_s; x \in \mathscr{C}, s \leqq t\}$, $\mathscr{R}^m$ is the Borel sigma field of R^m, and $\mathscr{U}_i$ is the Borel sigma field of the compact metric space U_i, $i = 1, \cdots, N$. $\mathscr{D}$ is the sigma field of the subset D of $[0, 1]\times\mathscr{C}$ having the property that the section of D at time t is in $\mathscr{F}_t$ for each t and the section of D at x is Lebesgue measurable for each x.

Now, let the functions H_i, $i = 1, \cdots, N$, which correspond to the Hamiltonians in the N-person stochastic differential games under consideration, be given by

$$H_i: [0, 1]\times\mathscr{C}\times R^m\times U_1\times\cdots\times U_N \to R,$$

such that for each i,

(i) H_i is measurable with respect to the sigma field $\mathscr{D}\otimes\mathscr{R}^m\otimes\mathscr{U}_1\otimes\cdots\otimes\mathscr{U}_N$,

(ii) $H_i(t, x, p_i, \cdot, \cdots, \cdot)$ is continuous on $U_1\times\cdots\times U_N$ for each $(t, x, p_i)\in [0, 1]\times\mathscr{C}\times R^m$.

Then the Nash condition [1] is stated as follows:

DEFINITION. We say the Nash condition holds if there exists a function

$$u_i^*: ([0, 1]\times\mathscr{C}\times R^{mN}, \mathscr{D}\otimes\mathscr{R}^{mN})\to(U_i, \mathscr{U}_i)$$

for all $i = 1, \cdots, N$ such that for each $(t, x, p)\in [0, 1]\times\mathscr{C}\times R^{mN}$ and for all $v_i\in U_i$,

$$H_i(t, x, p_i, u_1^*(t, x, p), \cdots, u_N^*(t, x, p))$$
$$\leqq H_i(t, x, p_i, u_1^*(t, x, p), \cdots, u_{i-1}^*(t, x, p), v_i, u_{i+1}^*(t, x, p), \cdots, u_N^*(t, x, p)),$$

where $p = (p_1, \cdots, p_N)$.

Let us introduce the following function:

$$\Phi(t, x, p, u, v) = \sum_{i=1}^{N} H_i(t, x, p_i, u_1, \cdots, u_{i-1}, v_i, u_{i+1}, \cdots, u_N)$$
$$-\sum_{i=1}^{N} H_i(t, x, p_i, u_1, \cdots, u_N)$$

where $p = (p_1, \cdots, p_N)$, $u = (u_1, \cdots, u_N)$ and $v = (v_1, \cdots, v_N)$. Using this function we can now state the main result.

* Received by the editors February 10, 1978.

† Department of Electrical Engineering, Waseda University, Tokyo 160, Japan.

THEOREM. *The Nash condition holds if and only if*

$$\max_{u\in U}\min_{v\in U}\Phi(t,x,p,u,v)=0 \tag{1}$$

for all $(t,x,p)\in[0,1]\times\mathscr{C}\times R^{mN}$, *where* $U=U_1\times\cdots\times U_N$.

Proof. First note that for fixed (u,v), $\Phi(\cdot,\cdot,\cdot,u,v)$ is measurable with respect to $\mathscr{D}\otimes\mathscr{R}^{mN}$ and for fixed (t,x,p), $\Phi(t,x,p,\cdot,\cdot)$ is continuous on $U\times U$. Note also that $U=U_1\times\cdots\times U_N$ is compact.

Suppose S is a countable dense subset of U. Then for fixed (t,x,p) and u,

$$\min_{v\in U}\Phi(t,x,p,u,v)=\inf_{v\in S}\Phi(t,x,p,u,v),$$

so that for any $a\in R$,

$$\{(t,x,p)\colon \min_{v\in U}\Phi(t,x,p,u,v)<a\}=\bigcup_{v\in S}\{(t,x,p)\colon \Phi(t,x,p,u,v)<a\}.$$

Hence, for fixed u, $\min_{v\in U}\Phi(t,x,p,u,v)$ is measurable with respect to $\mathscr{D}\otimes\mathscr{R}^{mN}$ in (t,x,p) and, further, $\min_{v\in U}\Phi(t,x,p,u,v)$ is continuous in u. Now, for fixed (t,x,p),

$$\max_{u\in U}\min_{v\in U}\Phi(t,x,p,u,v)=\sup_{u\in S}\inf_{v\in S}\Phi(t,x,p,u,v),$$

so that for any $a\in R$,

$$\{(t,x,p)\colon \max_{u\in U}\min_{v\in U}\Phi(t,x,p,u,v)<a\}=\bigcap_{u\in S}\{(t,x,p)\colon \min_{v\in U}\Phi(t,x,p,u,v)<a\}.$$

Hence $\max_{u\in U}\min_{v\in U}\Phi(t,x,p,u,v)$ is measurable with respect to $\mathscr{D}\otimes\mathscr{R}^{mN}$. Then an implicit function lemma of Beneš [2] shows that there is a measurable function $u^*=(u_1^*,\cdots,u_N^*)$,

$$u_i^*\colon([0,1]\times\mathscr{C}\times R^{mN},\mathscr{D}\otimes\mathscr{R}^{mN})\to(U_i,\mathscr{U}_i),$$

$i=1,\cdots,N$, such that for each (t,x,p),

$$\min_{v\in U}\Phi(t,x,p,u^*(t,x,p),v)=\max_{u\in U}\min_{v\in U}\Phi(t,x,p,u,v),$$

where $u^*(t,x,p)=(u_1^*(t,x,p),\cdots,u_N^*(t,x,p))$. Therefore, if $\max_{u\in U}\min_{v\in U}\Phi(t,x,p,u,v)=0$ for all $(t,x,p)\in[0,1]\times\mathscr{C}\times R^{mN}$, we obtain that for each (t,x,p),

$$\Phi(t,x,p,u^*(t,x,p),v)\geqq 0 \tag{2}$$

for all $v\in U$. This inequality is equivalent to

$$\sum_{i=1}^{N}H_i(t,x,p_i,u_1^*(t,x,p),\cdots,u_{i-1}^*(t,x,p),v_i,u_{i+1}^*(t,x,p),\cdots,u_N^*(t,x,p))$$
$$\geqq\sum_{i=1}^{N}H_i(t,x,p_i,u_1^*(t,x,p),\cdots,u_N^*(t,x,p))$$

for each (t,x,p) and all $v\in U$. Consequently, for each i and each (t,x,p), setting $v=(u_1^*(t,x,p),\cdots,u_{i-1}^*(t,x,p),v_i,u_{i+1}^*(t,x,p),\cdots,u_N^*(t,x,p))$ in (2), we obtain the following inequalities: for each $(t,x,p)\in[0,1]\times\mathscr{C}\times R^{mN}$,

$$H_i(t,x,p_i,u_1^*(t,x,p),\cdots,u_{i-1}^*(t,x,p),v_i,u_{i+1}^*(t,x,p),\cdots,u_N^*(t,x,p))$$
$$\geqq H_i(t,x,p_i,u_1^*(t,x,p),\cdots,u_N^*(t,x,p))$$

for all $v_i\in U_i$ and all $i=1,\cdots,N$. This implies that the Nash condition holds.

Conversely suppose that the Nash condition holds. Then, for each (t, x, p), we have

$$\Phi(t, x, p, u^*(t, x, p), v) \geqq \min_{v \in U} \Phi(t, x, p, u^*(t, x, p), v) = 0. \tag{3}$$

On the other hand, it follows from the definition of Φ that for each (t, x, p),

$$\min_{v \in U} \Phi(t, x, p, u, v) \leqq 0 \tag{4}$$

for all $u \in U$. Expressions (3) and (4) imply

$$\max_{u \in U} \min_{v \in U} \Phi(t, x, p, u, v) = \min_{v \in U} \Phi(t, x, p, u^*(t, x, p), v) = 0$$

for all $(t, x, p) \in [0, 1] \times \mathscr{C} \times R^{mN}$.

3. Remarks. A. In the previous paper [1], the following type of N-person stochastic differential game was discussed: The system is described by the stochastic functional differential equation of the form

$$dx_t = f(t, x, u_1, \cdots, u_N)\, dt + \sigma(t, x)\, dB_t \tag{5}$$

where B_t is a Brownian motion, and corresponding to each choice of the feedback strategies, player i incurs a cost of the form

$$P_i(u_1, \cdots, u_N) = E\left\{ g_i(x_1) + \int_0^1 h_i(t, x, u_1, \cdots, u_N)\, dt \right\}, \tag{6}$$

$i = 1, \cdots, N$. In this case, the Hamiltonians are given by

$$H_i(t, x, p_i, u_1, \cdots, u_N) = p_i f(t, x, u_1, \cdots, u_N) + h_i(t, x, u_1, \cdots, u_N),$$

$i = 1, \cdots, N$, and Theorem 2 of [1] asserts that if the Nash condition holds there is a Nash equilibrium point in feedback strategies under the several assumptions [1] to (5) and (6). Therefore the condition (1) becomes sufficient for the existence of a Nash equilibrium point in such stochastic differential games.

B. Suppose the case with the following convexity:

(i) U_i is a convex set for all $i = 1, \cdots, N$,

(ii) H_i is a convex function on U_i for fixed $(t, x, p_i, u_1, \cdots, u_{i-1}, u_{i+1}, \cdots, u_N)$ and all $i = 1, \cdots, N$.

In this case, using the theorem of Nikaido and Isoda [3], we can establish the condition (1). From this fact, we see that the "strict" convexity of h_i in the assumption (H_2) of [1] can be replaced by the convexity.

C. Finally consider the two person zero sum case, that is, $N = 2$ and $H_1(t, x, p_1, u_1, u_2) + H_2(t, x, p_2, u_1, u_2) = 0$ for all $(t, x, p_1, p_2, u_1, u_2) \in [0, 1] \times \mathscr{C} \times R^m \times R^m \times U_1 \times U_2$ such that $p_1 + p_2 = 0$. It is remarkable that in this special case the condition (1) is reduced to the Issacs condition [4]:

$$\min_{u_1 \in U_1} \max_{u_2 \in U_2} H_1(t, x, p_1, u_1, u_2) = \max_{u_2 \in U_2} \min_{u_1 \in U_1} H_1(t, x, p_1, u_1, u_2)$$

for all $(t, x, p_1) \in [0, 1] \times \mathscr{C} \times R^m$.

REFERENCES

[1] K. Uchida, *On existence of a Nash equilibrium point in N-person nonzero sum stochastic differential games*, this Journal, 16 (1978), pp. 142–149.

[2] V. E. BENEŠ, *Existence of optimal strategies based on specific information for a class of stochastic decision problems*, this Journal, 8 (1970), pp. 179–188.

[3] H. NIKAIDO AND K. ISODA, *Note on non-cooperative convex games*, Pacific J. Math., 5 (1955), pp. 807–815.

[4] R. ELLIOTT, *The existence of the value in stochastic differential games*, this Journal, 14 (1976), pp. 85–94.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 1, 1979

0036-0129/79/1701-0002 $01.00/0

SOME PROBLEMS IN THE CONTROL OF DISTRIBUTED SYSTEMS, AND THEIR NUMERICAL SOLUTION*

GREG KNOWLES†

Abstract. The bang-bang control of certain distributed parameter systems is considered, and the relationship between these results and the approximate controllability of the system discussed. Also, a technique for the numerical solution of both fixed time and minimum time problems is given and applied to several control problems governed by the wave equation.

1. Normal Systems.

In this note we consider the control problem whose state equation is given by

$$\dot{y}(t)=Ay(t)+u(t)g, \qquad y(0)=y_0, \tag{1.1}$$

where g, y_0 are fixed elements of X, a Banach space, and $A:X\to X$ is a closed linear operator with domain dense in X. We suppose further that A generates a C_0 semigroup of bounded linear operators $S(t)$: $X\to X$, $t\geqq 0$ [1]. The (scalar) control function u will be restricted to take its values in $\mathcal{U}=\{u: |u(\tau)\leqq 1 \text{ almost everywhere}\}$. For all such admissible controls and $t\geqq 0$, the integral

$$y(t:u)=S(t)y_0+\int_0^t S(t-\tau)gu(\tau)\,d\tau \tag{1.2}$$

(the integration being taken in the sense of Bochner) defines an element of X, which we adopt as the solution of (1.1).

Associated with (1.1) we consider two types of cost functionals

(I) A $y_1\in X$ and $\mathscr{E}>0$ are given, and we attempt to control the system to reach the target set $W=\{x\in X: \|x-y_1\|\leqq \mathscr{E}\}$ in minimum time.

Problem (I) will only be well posed if we assume W is reachable in some finite time, i.e.,

(H1) There exists a $t>0$ and a $u\in\mathcal{U}$ with $y(t: u)\in W$.

The second problem is

(II) A fixed time $T>0$, and a fixed $y_1\in X$ are given and we attempt to minimize $\|y(T:u)-y_1\|$ over all $u\in\mathcal{U}$.

Here we assume

(H2) For all $u\in\mathcal{U}$, $y(T: u)\neq y_1$.

A further concept which will not be investigated here but which plays a central part in this note, is that of approximate controllability of the system (1.1). Namely, we say (1.1) is approximately controllable in $[0, t]$ (in finite time) if $\{y(t:u): u\in L^\infty(0, t)\}$ is dense in X (respectively, is $\bigcup_{t>0}\{y(t: u): u\in L^\infty(0, t)\}$ is dense in X). It is then an easy consequence of the Hahn–Banach theorem that (1.1) is approximately controllable in $[0, t]$ if and only if

$$\langle x', S(\tau)g\rangle=0, \quad \tau\in[0, t], \quad \text{implies} \quad x'=0.$$

Further (1.1) is approximately controllable in finite time if and only if

$$\langle x', S(\tau)g\rangle=0, \quad \tau>0, \quad \text{implies} \quad x'=0,$$

(e.g., [4]).

* Received by the editors July 25, 1977, and in revised form April 3, 1978.

† Mathematics Department, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213. This research was supported in part by the Office of Naval Research under Grant N00014-76-C-0369.

A vector $g \in X$ is called an analytic vector for $\{S(t)\}$ if the function $\tau \to S(\tau)g$, $\tau > 0$ is analytic. For this it is necessary and sufficient that the function $\tau \to \langle x', S(\tau)g\rangle$, $\tau > 0$ be analytic, for every $x' \in X'$ [13]. (Notice that this definition differs slightly from the one used in other areas of mathematics, where it is usually assumed that g also belongs to $\bigcap_{n=1}^{\infty} D(A^n)$. In the case $\{S(t)\}$ is a group, these two definitions coincide: e.g., [23, footnote on page 315].) In the cases of main interest here (1.1) will represent either a parabolic equation, in which case every $g \in X$ is an analytic vector for $\{S(t)\}$, ($\{S(t)\rangle$, $t > 0$, is an analytic semigroup [13]), or (1.1) will represent a wave equation and $\{S(t)\}$ will be a group and then the analytic vectors will be dense in X [19]. Notice also, that when g is an analytic vector for a group $\{S(t)\}$, then the function $\tau \to \langle x', S(\tau)g\rangle$ is analytic for all $-\infty < \tau < \infty$.

The existence of optimal controls for problems (I) and (II) is well known. (See, for example [1], [14].) The purpose of this paper is to determine when these optimal controls are bang-bang and unique, and to investigate methods for their numerical solution. We will proceed via the following two propositions. The first was proven in ([16]), the second is well known (see e.g., [14, Thm. I.6.3]).

PROPOSITION 1.1. *If* (H2) *holds and $\bar{u}$ is an optimal control for problem* (II) *with $\|y(T, \bar{u}) - y_1\| = E$, then there exists an $z' \in X'$ for which*

$$(1.4) \qquad \int_0^T \langle z', S(T-\tau)g\rangle u(\tau)\, d\tau \leqq \int_0^T \langle z', S(T-\tau)g\rangle \bar{u}(\tau)\, d\tau \leqq \langle z', w\rangle$$

for all w with $\|w - y_1\| \leqq E$, and $u \in \mathcal{U}$. Further, $\bar{u}(\tau) = \operatorname{sgn}\{\langle z', S(T-\tau)g\rangle\}$.

PROPOSITION 1.2. *If* (H1) *holds then an optimal control u^* for problem* (I) *exists, and if the minimum time $t^* > 0$, then there exists a nonzero $x' \in X'$, for which*

$$(1.3) \qquad \int_0^{t^*} \langle x', S(t^*-\tau)g\rangle u(\tau)\, d\tau \leqq \int_0^{t^*} \langle x', S(t^*-\tau)g\rangle u^*(\tau)\, d\tau \leqq \langle x', w\rangle$$

for all $w \in W$, $u \in \mathcal{U}$. In particular $u^*(\tau) = \operatorname{sgn}\{\langle x', S(t^*-\tau)g\rangle\}$, $\tau \in [0, t^*]$.

Remark. If X is a Hilbert space in problem (I), as x' supports W in $y(t^*, u^*)$, then we must have $x' = \lambda(y_1 - y(t^*, u^*))$ for some $\lambda \neq 0$. Clearly by dividing (1.3) through by λ, we can take $x' = y_1 - y(t^*, u^*)$. Similarly in problem (II), $z' = y_1 - y(T, \bar{u})$.

From Propositions 1.1 and 1.2 it can be seen that the bang-bangness of the optimal control depends on whether the function $\tau \to \langle x', S(t^*-\tau)g\rangle$, $0 \leqq \tau \leqq t^*$, is nonzero almost everywhere. Accordingly, we call the system (1.1) normal in X, if, for all $t > 0$,

$$\langle x', S(t-\tau)\rangle = 0 \text{ for } \tau \in F \subset (0, t),\ F \text{ nonnull, implies that } x' = 0$$

(see [4] and [3]). It then follows from Propositions 1.1 and 1.2 that if (1.1) is normal in X, the optimal controls for problems (I) and (II) are bang-bang, unique and uniquely determined by (1.3) or (1.4) [4, Thm. 4]. Also, from our earlier remarks, any normal system is approximately controllable in $[0, t]$, for any $t > 0$. Conversely, we have

THEOREM 1.1. *If g is an analytic vector for $\{S(t)\}$ and* (1.1) *is approximately controllable in finite time, then* (1.1) *is normal, and hence u^** (*respectively $\bar{u}$*) *is bang-bang and unique. Further this optimal control has at most a countable number of switches accumulating at t^** (*respectively, T*). *In the case $\{S(t)\}$ is a group, the optimal control has at most a finite number of switches.*

Proof. We show first that (1.1) is normal in X. Suppose there exists a $t > 0$, and an $x' \in X'$ such that

$$\langle x', S(t-\tau)g\rangle = 0 \quad \text{for} \quad \tau \in F \subset [0, t],$$

and F is nonnull. As g is an analytic vector, this function is analytic in τ, and hence identically zero. That is, $\langle x', S(\tau)g\rangle = 0$ for all $\tau > 0$, and so $x' = 0$, as (1.1) is approximately controllable in finite time.

For the second part of the theorem we remark that by analyticity the function

$$\tau \to \langle x', S(t^* - \tau)g\rangle \tag{1.5}$$

can have at most a finite number of zeros in any interval $(0, t^* - \delta)$, for any $\delta > 0$, sufficiently small. Consequently the only possible point of accumulation of the zeros of (1.5) is at t^*. Clearly the zeros of (1.5) are the switching times. Finally, in the group case (1.5) is analytic for all $\tau \in (+\infty, \infty)$ and so can have at most finitely many zeros.

The relation between controllability and normality for lumped systems is well known; it is inherent in the discussion of proper and normal systems in [11]. We see from Theorem 1.1, that if g is an analytic vector for $\{S(t)\}$, then this relation carries over to infinite dimensional situations. For problem (II) for distributed systems similar ideas are discussed in [7]. Unfortunately, the assumption that the system be approximately controllable need not always hold (e.g., [4], [23]). In fact, in general to make (1.1) controllable one needs at least as many scalar controls as the largest multiplicity of the eigenvalues of A [23]. However, if the system is not approximately controllable in X (and hence not normal) is could still be normal in a subspace of X, and this is often enough to prove the bang-bangness and uniqueness of the optimal control for the time optimal problem (problem (I)). We will do this, for the case $y_0 = 0$, under the following slight strengthening of (H1).

(H3) There exists a $t_1 > 0$ and $u \in \mathcal{U}$ such that $\|y_1 - y(t_1 : u)\| < \mathscr{E}$.

LEMMA. *Suppose* (H3) *holds and let* $\tilde{X}$ *be the closure of* $\{\int_0^{t_1} S(t_1 - \tau)gu(\tau)\,d\tau : u \in L^\infty(0, t_1)\}$ *(given the norm topology induced by* X*). If* u^* *is an optimal control for* (I) *and* $t^* > 0$ *the minimum time, then there exists a nonzero* $\tilde{x}' \in \tilde{X}'$ *with*

$$\int_0^{t^*} \langle \tilde{x}', S(t^* - \tau)g\rangle u(\tau)\,d\tau \leqq \int_0^{t^*} \langle \tilde{x}', S(t^* - \tau)g\rangle u^*(\tau)\,d\tau, \qquad u \in \mathcal{U}.$$

Proof. Denote by $B(x, \rho)$ the closed ball in X with radius ρ. Set $\tilde{B} = \tilde{X} \cap B(y_1, \mathscr{E})$ and $\mathcal{A}(t) = \{\int_0^t S(t - \tau)gu(\tau)\,d\tau : u \in \mathcal{U}\}$, the attainable set of the control system (1.1) in time t, for $0 \leqq t \leqq t_1$. Assumption (H3) implies that $\tilde{B}$ is nonempty, and by setting u to be zero on $(0, t_1 - t)$, we can see that $\mathcal{A}(t) \subset \tilde{X}$, for $t \in (0, t_1)$. The sets $\mathcal{A}(t)$ are weakly compact and convex in X, and hence also in $\tilde{X}$, and $\tilde{B}$ is closed and convex in $\tilde{X}$. By (H3) we can choose a $p \in B(y_1, \delta) \cap \tilde{X}$, for some $0 < \delta < \mathscr{E}$, and then $B(p, (\mathscr{E} - \delta)/2) \cap \tilde{X} \subset \tilde{B}$, or $\tilde{B}$ has nonempty interior in $\tilde{X}$. Since t^* is the minimum time, and $\mathcal{A}(t) \subset \tilde{X}$ for $0 \leqq t \leqq t_1$, it follows by the usual arguments that $\mathcal{A}(t^*) \cap \tilde{B} \neq \emptyset$, $\mathcal{A}(t^*) \cap \operatorname{int} \tilde{B} = \emptyset$, and so by [3] there must exist a nonzero $\tilde{x}' \in \tilde{X}'$ separating $\mathcal{A}(t^*)$ and $\tilde{B}$. Consequently,

$$\langle \tilde{x}', \int_0^{t^*} S(t^* - \tau)gu(\tau)\,d\tau\rangle \leqq \langle \tilde{x}', \int_0^{t^*} S(t^* - \tau)gu^*(\tau)\,d\tau\rangle, \qquad u \in \mathcal{U}. \tag{1.6}$$

The lemma will follow (by transferring $\tilde{x}'$ under the integral in (1.6)), if we can show that $S(\tau)g \in \tilde{X}$, for $\tau \in (0, t^*)$. By setting $u = 1$, and using the fact that $\tilde{X}$ is a linear space, we have that

$$x_h = \frac{1}{h}\left(\int_0^{t+h} S(\tau)g\,d\tau - \int_0^t S(\tau)g\,d\tau\right) \in \tilde{X}$$

for $t \in (0, t^*)$ and h sufficiently small. However, the function $\tau \to S(\tau)g$, $\tau \in (0, t^*)$ is continuous into $X(\{S(t)\}$ is a C_0 semigroup), and so

$$\lim_{h\to 0} x_h = S(t)g,$$

i.e., $S(t)g \in \tilde{X}$, for $t \in (0, t^*)$.

THEOREM 1.2. *If* (H3) *holds, g is an analytic vector for* $\{S(t)\}$, *and* $t^* > 0$; *then* u^*, *the optimal control for problem* (I), *is bang-bang, unique, and has the switching properties refered to in Theorem* 1.1.

Proof. From the lemma $u^*(\tau) = \operatorname{sgn}\{\langle \tilde{x}', S(t^*-\tau)g\rangle\}$. Since $\tilde{x}'$ is a continuous linear functional on $\tilde{X}$, it has a continuous extension to all of X; denote this extension by x'. As g is an analytic vector the function $\tau \to \{x', S(t^*-\tau)g\rangle$, $0<\tau<t^*$, is analytic; however since $S(t^*-\tau)g \in \tilde{X}$, for $\tau \in (0, t^*)$ (see Lemma 1), $\langle x', S(t^*-\tau)g\rangle = \langle \tilde{x}', S(t^*-\tau)g\rangle$ for all $0<\tau<t^*$. The proof then follows as in Theorem 1, since (1.1) is automatically approximately controllable in $\tilde{X}$, in time t_1.

Notice that (in the notation of § 3) if A is a normal operator with compact resolvent on a Hilbert space X, the above proof carries through for all those initial conditions $y_0 \in X$ for which $(g, \varphi_{kj}) = 0$ implies $(y_0, \varphi_{kj}) = 0$, $k = 1, \cdots, r_j$, $j = 1, 2, \cdots$. The proof follows as before, as this assumption implies $\mathscr{A}(t) \subset \tilde{X}$ for $0 \leqq t \leqq t_1$.

Finally we remark that these results all extend to the case where the control appears nonlinearly, and the set of admissible controls need not be convex. Namely, suppose the state equation takes the form

$$\dot{y}(t) = Ay(t) + h(t, u(t))g, \qquad y(0) = y_0$$

where h is a bounded Carathéodory function (i.e., measurable in the first variable and continuous in the second), and the set of admissible controls is now $\mathscr{U} = \{u : u(t) \in U \text{ a.e.}\}$, where U is a fixed compact set in $\mathbb{R}$. Set $h(t, U) = \{h(t, v) : v \in U\}$, $\mathscr{L} = \{f : f$ is Borel measurable, and $f(\tau) \in \text{co}\, h(t, U)$ a.e.$\}$ (co denotes convex hull), and denote by $y(t:f)$ the solution of (1.1) with f in place of u, for any $f \in \mathscr{L}$. Then, for problem (I) we have

THEOREM 1.3. *If there exists a time* $t_1 > 0$ *and* $f \in \mathscr{L}$ *such that* $y(t:f) \in W$, *g is an analytic vector for* $\{S(t)\}$, *and* (1.1) *is approximately controllable in finite time*; *then W is reached in minimum time* t^* *by an admissible control* $u^* \in \mathscr{U}$, *and* $h(t, u^*(t))$ *belongs to the extreme points of* co $h(t, U)$ *for almost all* $t \in (0, t^*)$.

The analogue of Theorem 1.2 can be stated similarly, and for problem (II)

THEOREM 1.4. *If* $y_1 \neq y(T:f)$ *for any* $f \in \mathscr{L}$, *g is an analytic vector for* $\{S(t)\}$, *and* (1.1) *is approximately controllable in finite time*; *then an optimal control* $u^* \in \mathscr{U}$ *for problem* (II) *exists and* $h(t, u^*(t))$ *belongs to the extreme points of* co $h(t, U)$ *a.e.*

The proof of these theorems follows as in [15] by defining the vector measure $m_t : \mathscr{B}(0, t) \to X$ ($\mathscr{B}(0, t)$ is the Borel σ-algebra on $(0, t)$) by,

$$m_t(E) = \int_E S(t-\tau)g\, d\tau$$

$E \in \mathscr{B}(0, t)$, $t > 0$. The output of (1.1) can then be represented as

$$\int_0^t h(\tau, u(\tau))\, dm_t(\tau)$$

and the results follow as in [15, Thms 2, 3 and 4].

2. Applications. Suppose Ω is a bounded n-dimensional domain, with sufficiently smooth boundary $\partial\Omega$; $x = (x_1, \cdots, x_n)$ is a point in Ω; $L(x, \partial/\partial x)$ is an

$N \times N$ matrix with each element $L_{ij}(x, \partial/\partial x)$ a linear differential operator of order 2m of the form

$$L_{ij}\left(x, \frac{\partial}{\partial x}\right) = \sum_{|\alpha| \leqq 2m} a_{ij}^{(\alpha)}(x) D^\alpha \qquad (i, j = 1, 2, \cdots, N) \tag{2.1}$$

where $\alpha = (\alpha_1, \cdots, \alpha_n)\,(\alpha_i \geqq 0)$, $|\alpha| = \alpha_1 + \cdots + \alpha_n$, and $D^\alpha = (\partial^{\alpha_1 + \alpha_2 + \alpha_n})/(\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n})$. We assume the coefficients $a_{ij}^{(\alpha)}(x)$ are real and sufficiently smooth, and $L(x, \partial/\partial x)$ is strongly elliptic i.e.,

$$(-1)^m \sum_{i,j=1}^{N} \sum_{|\alpha|=2m} a_{ij}^{(\alpha)}(x) \xi_1^{\alpha_1} \cdots \xi_n^{\alpha_n} \eta_i \eta_j > 0$$

for arbitrary $x \in \Omega$, and real ξ_k, η_l with $\sum \xi_k^2 \neq 0$, $\sum \eta_l^2 \neq 0$, $L(x, \partial/\partial x)$ generates a linear operator in the space $X = \bar{L}^p(\Omega)\,(=\{y : y = (y_1, \cdots, y_n)$ and $\sum_{i=1}^n \int_\Omega |y_i|^p \, d\Omega < \infty\})$, $p > 1$, defined on those smooth functions satisfying

$$y|_{\partial\Omega} = \frac{\partial y}{\partial n}\Big|_{\partial\Omega} = \cdots = \frac{\partial^{m-1} y}{\partial n^{m-1}}\Big|_{\partial\Omega} = 0 \tag{2.2}$$

where n is a normal to $\partial\Omega$. This operator allows a closure; denote it by A. Then (2.1) represents the strongly parabolic system

$$\frac{\partial y}{\partial t} = -L\left(x, \frac{\partial}{\partial x}\right) y + g(x) u(t), \qquad x \in \Omega, \quad t > 0, \tag{2.3}$$

$$y(x, 0) = y_0(x), \qquad x \in \Omega, \tag{2.4}$$

with boundary conditions (2.2), where g is any fixed element of $\bar{L}^p(\Omega)$, and u is the control function. It is known, that A generates an analytic semigroup [16, I.8.3], and hence every $g \in \bar{L}^p(\Omega)$ is analytic for this semigroup. Consequently for (2.2), (2.3), (2.4), Theorem 1.1 becomes

THEOREM 2.1. *If W is reached in some time by an admissible control, and system* (2.1), (2.2), (2.3) *is approximately controllable in finite time, then W is reached in minimum time t^* by a unique, bang-bang optimal control u^*, which has at most a countable number of switches accumulating at t^*.*

Theorems 1.2, 1.3 can be similarly stated.

We remark that analogous results hold for more general boundary conditions of the Sapiro–Lopatinski type [16, I.8.3]. When $m = 1$, $N = 1$, similar results hold for the second boundary value problem

$$\frac{\partial y}{\partial \nu}(x, t) - a(x) y(x, t) = 0, \qquad x \in \partial\Omega,$$

where $a(x) \geq 0$.

Now let $X = \mathring{H}_1(\Omega) \oplus L_2(\Omega)$ endowed with the energy inner product, i.e.,

$$([f_1, h_1], [f_2, h_2])_X = \int_\Omega \{\nabla f_1 \cdot \overline{\nabla f_2} + h_1 \cdot \bar{h}_2\} \, d\Omega.$$

If

$$A = \begin{bmatrix} 0 & I \\ \Delta & 0 \end{bmatrix} \quad \text{and} \quad D(A) = \{H_2(\Omega) \cap \mathring{H}_1(\Omega)\} \oplus \mathring{H}_1(\Omega),$$

and $g=[0, b] \in X$, $y_0=[Q_0, Q_1] \in X$ then (1.1) represents

$$\begin{aligned} &\frac{\partial^2 p}{\partial t^2}=\Delta p+b(x)u(t) && x \in \Omega, \quad t>0 \\ &p(x, 0)=Q_0(x) && x \in \Omega \\ &p_t(x, 0)=Q_1(x) && \\ &p(x, t)=0 && \text{for } x \in \partial\Omega \end{aligned}$$

(where $y(t)=[p(\cdot, t), p_t(\cdot, t)]$). The operator A is normal (in fact skew adjoint), with compact resolvent, and generates a group $\{S(t): t \in (-\infty, \infty)\}$. Consequently the analytic vectors for $\{S(t)\}$ are dense in X. Denoting the eigenvalues of A by λ_j, $j=1, 2, \cdots$, and its normalized eigenvectors by φ_{jk}, $k=1, \cdots, r_j(r_j$ is the multiplicity of $\lambda_j)$, $j=1, 2, \cdots$, then we can represent $S(t)$ as,

$$S(t)=\sum_{j=1}^{\infty} e^{\lambda_j t} \sum_{k=1}^{r_j} (g, \varphi_{jk})\varphi_{jk}, \qquad g \in X. \tag{2.5}$$

If $\{l_j\}$ and $\{\Psi_{jk}\}$ denote the eigenvalues and eigenvectors of Δ on $L^2(\Omega)$ (with homogeneous Dirichelet boundary conditions), then $\lambda_j=\sqrt{l_j}$, and $\varphi_{jk}=[\Psi_{jk}, -\sqrt{l_j}\,\Psi_{jk}]$, $k=1, \cdots, r_j$, $j=1, 2, \cdots$.

It is known that g is an analytic vector for $\{S(t)\}$ if and only if $g \in D(A^n)$ for each $n=1, 2$, and

$$\sum_{n=0}^{\infty} \frac{\|A^n g\|}{n!} t^n < \infty \qquad \text{for some } t>0. \tag{2.6}$$

Since in this case

$$A^n g=\sum_{j=1}^{\infty} \lambda_j^n \sum_{k=1}^{r_j} (g, \varphi_{jk})\varphi_{jk} \quad \text{for } g \in D(A^n),$$

$n=1, 2, \cdots$, we have

$$\begin{aligned} \|A^n g\|^2 &= \sum_{j=1}^{\infty} \lambda_j^{2n} \sum_{k=1}^{r_j} |(g, \varphi_{jk})|^2 \\ &= \sum_{j=1}^{\infty} \lambda_j^{2n+2} \sum_{k=1}^{r_j} |b_{jk}|^2 \end{aligned} \tag{2.7}$$

where $b_{jk}=(b\ , \Psi_{jk})_{L^2(\Omega)}$. Hence, by applying the root test to (2.6), g is an analytic vector for $\{S(t)\}$ if and only if

$$\limsup_{n \to \infty} \left[\frac{\|A^n g\|}{n!}\right]^{1/n} < \infty$$

that is, by (2.7), if and only if

$$\limsup_{n \to \infty} \left[\sum_{j=1}^{\infty} \frac{|\lambda_j|^{2n+2}}{(n!)^2} \sum_{k=1}^{r_j} |b_{jk}|^2\right]^{1/2n} < \infty.$$

In particular, in the case $\Omega=(0, \pi)$, $\lambda_j=\pm ij$, and $\Psi_j=\sin(jx)$, $x \in (0, \pi)$, $j=1, 2, \cdots$, $g=[0, b]$ will be an analytic vector for $\{S(t)\}$ if and only if

$$\limsup_{n \to \infty} \left[\sum_{j=1}^{\infty} \frac{j^{2n+2}}{(n!)^2} |b_j|^2\right]^{1/(2n)} < \infty \tag{2.8}$$

where $b_j = \int_0^\pi b(x) \sin(j, x)\, dx$, $j = 1, 2, \cdots$. In particular, this holds if

$$|b_j| \leqq M e^{-Cj}$$

for some positive constants M, C, all j (c.f. [24, Proposition 1]). It is also known that (for $\Omega = (0, \pi)$) this problem is approximately controllable in finite time if and only if $b_j \neq 0$ $j = 1, 2, \cdots$, [23, Example 4.4], [21, Prop. 1].

3. Numerical approximation. In this section we develop further the method in [8], [9] for the numerical computation of a sequence of suboptimal controls for the minimum time problem. Specifically suppose the state equation is

$$\dot{y}(t) = Ay(t) + \sum_{i=1}^{m} g_i u_i(t), \qquad y(0) = y_0 \in X \tag{3.1}$$

where $g_i \in X$, $\mathbf{u} = (u_i)$, $i = 1, \cdots, m$, and we wish to control (3.1) to the origin in minimum time, with admissible controls $\mathcal{U} = \{\mathbf{u} : |u_i| \leqq 1, i = 1, 2, m\}$. In the remarks following we shall show that this choice of end-point $y_1 = 0$ places no restriction on the generality of the method.

In [22] the problem of finding the control of minimum L^2 norm transferring certain systems from initial to final state in a given time T was considered. Here we suppose (as is more often the case) that the controls are restricted a priori, and we seek to effect this transfer in minimum time. In the final section we indicate extensions of this method to problem (II).

Clearly, the results of the first two sections indicate (at least in certain cases) the feasibility of approximating this problem by bang-bang controls. Following [8], we propose a simple algorithm for doing this, and show its applicability on a control problem governed by the wave equation.

For this section assume that X is a Hilbert space, and A is a normal operator with compact resolvent, having eigenvalues φ_{jk}, $k = 1, \cdots, r_j$, $j = 1, 2, \cdots$, r_j is the multiplicity of λ_j. It is known that the $\{\lambda_j\}$ are isolated, if A is unbounded, so are $\{\lambda_j\}$, and $\{\mathrm{Re}(\lambda_j)\}$ are uniformly bounded from above. If each $r_j = 1$, we simply write $\{\varphi_j\}$ instead of $\{\varphi_{jk}\}$. With these assumptions we can write the solution of (3.1) as

$$y(t : \mathbf{u}) = \sum_{j=1}^{\infty} \sum_{k=1}^{r_j} \left\{ e^{\lambda_j t} y_0^{jk} + \sum_{i=1}^{m} g_i^{jk} \int_0^t e^{\lambda_j (t-\tau)} u_i(\tau)\, d\tau \right\} \varphi_{jk} \tag{3.2}$$

where $y_0^{jk} = (y_0, \varphi_{jk})$ and $g_i^{jk} = (g_i, \varphi_{jk})$, $i = 1, \cdots, m$. Then the minimum time, t^*, for which $y(t^*) = 0$ (in X) will be, by orthonormality of $\{\varphi_{jk}\}$, the smallest time for which

$$\sum_{i=1}^{m} g_i^{jk} \left(\int_0^{t^*} e^{-\lambda_j \tau} u_i(\tau)\, d\tau \right) = -y_0^{jk} \quad \text{for } k = 1, \cdots, r_j, \quad j = 1, 2 \cdots. \tag{3.3}$$

Now defining

$$G_j = \begin{bmatrix} g_1^{j1} & & g_m^{j1} \\ g_1^{j2} & \cdots & g_m^{j2} \\ \vdots & & \vdots \\ g_1^{jr_j} & \cdots & g_m^{jr_j} \end{bmatrix}, \qquad \mathbf{y}_j = \begin{bmatrix} y_0^{j1} \\ y_0^{j2} \\ \vdots \\ y_0^{jr_j} \end{bmatrix}$$

and

$$\mathbf{h}_j(u_1\cdots,u_m;t)=\begin{bmatrix}\int_0^t e^{-\lambda_j\tau}u_1(\tau)\,d\tau\\ \vdots\\ \int_0^t e^{-\lambda_j\tau}u_m(\tau)\,d\tau\end{bmatrix},\quad j=1,2,\cdots,$$

we see that (3.3) can be written as

$$(3.3)'\qquad G_j\cdot\mathbf{h}_j(u_1,\cdots,u_m;t^*)=-\mathbf{y}_j,\qquad j=1,2,\cdots.$$

Even if the optimal controls were bang-bang, in general it would require an infinite number of switchings to satisfy the infinite system of equations (3.3) or ((3.3)′), and so such controls would be impossible to calculate numerically.

An approximating sequence of controls

$$\{\mathbf{u}^n\}=\left\{\begin{bmatrix}u_1^n\\ u_2^n\\ \vdots\\ u_m^n\end{bmatrix}\right\}$$

could be computed by choosing $\mathbf{u}^n$ so that it removes the first n modes in (3.2) i.e., so that

$$(3.4)\qquad G_j\mathbf{h}_j(u_1^n,\cdots,u_m^n:t^n)=-\mathbf{y}_j\quad\text{for } j=1,\cdots,n$$

where t^n is the smallest time for which these equations are satisfied. (However, if coefficients in (3.3) are known analytically, then convergence could be improved by modifying this procedure; we return to this point later.) If the ith control u_i^n is bang-bang with switches at $t_{i1},\cdots,t_{is_i}=t^n$, then $\mathbf{h}_j(u_1^n,\cdots,u_m^n,t^n)=$

$$(3.5)\quad \mathbf{F}_j(t_{il})\triangleq\begin{bmatrix}\pm\frac{1}{\lambda_j}(1-2\exp(-\lambda_jt_{11})+2\exp(-\lambda_jt_{12})\cdots(-1)^{s_1}\exp(-\lambda_jt^n))\\ \vdots\qquad\vdots\qquad\vdots\\ \pm\frac{1}{\lambda_j}(1-2\exp(-\lambda_jt_{m1})+2\exp(-\lambda_jt_{m2})\cdots(-1)^{s_m}\exp(-\lambda_jt^n))\end{bmatrix}$$

$$j=1,2,\cdots,$$

where the + sign is taken if the control starts with value +1, and similarly for the − sign. Combining (3.5) and (3.4), we find that under the above assumption the computation of $\mathbf{u}^n$ reduces to solving the following set of nonlinear equations:

$$(3.6)\qquad G_j\mathbf{F}_j(t_{il})=-\mathbf{y}_j\quad\text{for } j=1,\cdots,n,$$

$l=1,\cdots,s_i$, $i=1,\cdots,m$. We now give conditions under which the set of equations (3.6) has a solution, and when that solution will be unique.

Consider the following controllability assumptions

(H4) There exist $u_i\in\mathcal{U}$, $i=1,\cdots,m$, transferring (3.1) to zero in finite time.

(H5) Rank $G_j=r_j$, and Re $(\lambda_j)<0$ for $j=1,2,\cdots n$.

(H6) For the case $m=1$, and A has no multiple eigenvalues ($r_j=1$), suppose $(g, \varphi_j)\neq 0$ and $\operatorname{Re}(\lambda_j)\leqq 0$, $j=1, 2, \cdots$.

THEOREM 3.1. *If either* (H4), (H5) *or* (H6) *holds the system of equations* (3.6) *has a solution with t^n minimal and $0<t_{i1}<t_{i2}<t_{is_i}(=t^n)$, $i=1, 2, \cdots, m$. Further this solution is unique, if A has no multiple eigenvalues and $(g_i, \varphi_j)\neq 0$ for all $i=1, 2, \cdots, m$, $j=1, 2, \cdots$.* (This is automatically true for (H6).)

Proof. Consider the finite-dimensional control problem of steering the system whose state at time t is given by

$$y^n(t, u)=\sum_{j=1}^{n}\sum_{k=1}^{r_j}\left[e^{\lambda_j t}y_0^{jk}+\sum_{i=1}^{m} g_i^{jk}\int_0^t e^{\lambda_j 0t-\tau)}u_i(\tau)\right]\varphi_{jk}, \tag{3.7}$$

with initial condition

$$y_0^n=\sum_{j=1}^{n}\sum_{k=1}^{r_j} y_0^{jk}\varphi_{jk}, \tag{3.8}$$

and controls $|u_i|\leqq 1$, $i=1, 2, \cdots, m$, to the origin in the space $X_n=\operatorname{sp}[\varphi_{11}, \cdots, \varphi_{1r_1}, \cdots, \varphi_{n1}, \cdots, \varphi_{nr_n}]$, in minimum time. (sp denotes closed linear span). Any one of the assumptions (H4), (H5), (H6) guarantees the existence of an admissible control steering the initial state to zero in finite time, ((H4) is self-evident, for (H5) or (H6) see [23, p. 324] and [16, Cor. I.3.3, II.5.17]); consequently an optimal control, $(\bar{u}_1^n, \cdots, \bar{u}_m^n)$, steering y_0^n to the origin in minimum time, must exist. We can choose this optimal control bang-bang, with each component having a finite number of switches [10]. The switching times of the optimal control provide a desired solution to (3.6); it will be the unique solution if and only if $(\bar{u}_1^n, \cdots, \bar{u}_m^n)$ is the unique optimal control for the problem (3.7), (3.8). The added assumptions of the theorem imply (3.7) is normal, [11, Thm. 16.1], and so this optimal control is unique.

Remark. In the case $\{\lambda_1, \cdots, \lambda_n\}$ are real (for example, if A is self-adjoint), and the normality conditions in the theorem hold, then s_i, the number of switches of the ith control, must be less than or equal to n[20, III.17.10].

The next theorem guarantees the convergence of the minimum times for the approximating problems (3.6) to the minimum time for the original problem (3.1).

THEOREM 3.2. *If the minimum time, t^n, for each of the approximate problems* (3.6) *exists, $n=1, 2, 3, \cdots$, then the sequence $\{t^n\}$ is increasing. If the limit $\lim_{n\to\infty} t^n=t^*$ is finite, t^* is the minimum time for the original problem* (3.1). *On the other hand if $\{t^n\}$ diverges, there is no admissible control transferring y_0 to 0 (in X) in finite time.*

Conversely, suppose (H4) *holds, and t^* is the minimum time for* (3.1). *Then each t^n exists, the sequence $\{t^n\}$ is increasing and $\lim_{n\to\infty} t^n=t^*$.*

Proof. By construction t^{n+1} is the minimum time for which the system of equations (3.6) has a solution (in the sense of Theorem 3.1) for $j=1, 2, \cdots, n+1$. Since t^n is just the smallest time for which (3.6) has a solution for $j=1, 2, \cdots, n$, by minimality we must have $t^n\leqq t^{n+1}$, $n=1, 2, \cdots$.

Suppose $\lim_{n\to\infty} t^n=t^*$ is finite. We shall next construct an admissible control transferring y_0 to zero in time t^*; accordingly t^* must be an upper bound for the minimal solution of the original problem. However, if the minimum time for the original problem, $\bar{t}$, say, is strictly less than t^*, then each of the approximate problems can be solved in time $t^n\leqq\bar{t}$, and consequently $\lim_{n\to\infty} t^n\leqq\bar{t}<t^*$, a contradiction.

Suppose $\mathbf{u}^n\neq(u_1^n, \cdots, u_m^n)$, $|u_i|\leqq 1$, $i=1, 2, \cdots, m$, transfers y_0^n to zero in X_n in time t^n. By extending $\mathbf{u}^n$ to be zero on (t^n, t^*), each $\mathbf{u}^n$ is contained in the unit ball in

$$\underbrace{L^\infty(0, t^*)\times L^\infty(0, t^*)\times\cdots\times L^\infty(0, t^*)}_{m \text{ times}}$$

and consequently we can extract a weak-star convergent subnet $\{\mathbf{u}^\alpha\}$, $\alpha \in A$, with $\mathbf{u}^\alpha \to \mathbf{u}^*$, $\mathbf{u}^* = (u_1^*, \cdots, u_m^*)$, $|u_i^*| \leqq 1$, $i = 1, 2, \cdots, m$. For fixed j, k consider the difference between the jkth terms in the expansions of $y(t^*, \mathbf{u}^*)$ and $y(t^\alpha, \mathbf{u}^\alpha)$,

$$\Big| e^{\lambda_j t^*} y_0^{jk} + \sum_{i=1}^m g_i^{jk} \int_0^{t^*} e^{-\lambda_j (t^* - \tau)} u_i^*(\tau)\, d\tau - e^{\lambda_j t\alpha} y_0^{jk}$$

$$- \sum_{i=1}^m g_i^{jk} \int_0^{t\alpha} e^{-\lambda_j (t\alpha - \tau)} u_i^\alpha(\tau)\, d\tau \Big|$$

$$\leqq |y_0^{jk}|\, |e^{\lambda_j t^*} - e^{\lambda_j t\alpha}| + \sum_{i=1}^m |g_i^{jk}| \Big\{ \Big| \int_0^{t^*} e^{-\lambda_j (t^* - \tau)} (u_i^* - u_i^\alpha)(\tau)\, d\tau \Big| \tag{3.9}$$

$$+ \Big| \int_0^{t^*} (e^{-\lambda_j (t^* - \tau)} - e^{-\lambda_j (t\alpha - \tau)}) u_i^\alpha(\tau) d\tau \Big| \Big\}$$

since $\mathbf{u}^\alpha = 0$ on (t^α, t^*).

Since $t^\alpha \to t^*$, $|u_i^\alpha| \leqq 1$, and $\mathbf{u}^\alpha \to \mathbf{u}^*$ in the weak-star topology it is not difficult to see that (3.9) converges to zero. However we chose $\mathbf{u}^n$ so that

$$y_0^{jk} e^{\lambda_j t^n} + \sum_{j=1}^m g_i^{jk} \int_0^{t^n} e^{-\lambda_j (t^n - \tau)} u_i^n(\tau)\, d\tau = 0 \quad \text{for } k = 1, \cdots, r_j$$

$$j = 1, \cdots, n.$$

Consequently, taking the limit in α in (3.9) we will find that each of the coefficients of the eigenfunction expansion of $y(t^*, u^*)$ must be zero, or $y(t^*, u^*) = 0$.

From our earlier remarks, any time in which we can transfer y_0 to 0 using an admissible control is an upper bound for the sequence $\{t^n\}$, and so if this sequence diverges, the original problem cannot have a solution.

The converse to the theorem can be proven similarly.

We now consider the problem of estimating the norm (in X) of, $(y(t^n, \mathbf{u}^n) - y^n(t^n, \mathbf{u}^n))$, the remaining components of the state variable after time t^n.

$$\|y(t^n, \mathbf{u}^n) - y^n(t^n, \mathbf{u}^n)\|^2 = \sum_{j=n+1}^\infty \sum_{k=1}^{r_j} \Big| y_0^{jk} e^{\lambda_j t^n} + \sum_{i=1}^m g_i^{jk} \int_0^{t^n} e^{\lambda_j (t^n - \tau)} u_i^n(\tau)\, d\tau \Big|^2,$$

and since $|u_i^n| \equiv 1$ all i, n,

$$\|y(t^n, \mathbf{u}^n) - y^n(t^n, \mathbf{u}^n)\|^2 \leqq \sum_{j=n+1}^\infty \sum_{k=1}^{r_j} \{ |y_0^{jk}|^2 e^{2\mathrm{Re}(\lambda_j t^n)} + C_j^2 \sum_{i=1}^m |g_i^{jk}|^2 \}$$

where

$$C_j = \begin{cases} \dfrac{|e^{\mathrm{Re}\, \lambda_j t^n} - 1|}{|\mathrm{Re}\, (\lambda_j)|} & \text{if } \mathrm{Re}\, (\lambda_j) \neq 0 \\ t^n & \text{if } \mathrm{Re}\, (\lambda_j) = 0 \end{cases} \quad j = 1, 2, \cdots.$$

So in the case Re $(\lambda_j) = 0$ (e.g., the second example in section 2) we have

$$\|y(t^n, \mathbf{u}^n) - y^n(t^n : \mathbf{u}^n)\|^2 \leqq \|y_0 - y_0^n\|^2 + (t^n)^2 \sum_{i=1}^m \|g_i - g_i^n\|^2 \tag{3.10}$$

where y_0^n, g_i^n are the approximations to y_0 and g_i at the nth step, $i = 1, 2, \cdots, m$, $n = 1, n = 1, 2, \cdots$. (Alternately they can be considered as the projections of y_0 and g_i onto the subspace of X spanned by the eigenfunctions of $\lambda_1, \cdots, \lambda_n$).

In the case Re $(\lambda_j) \neq 0$ (e.g., if A is self-adjoint), setting

$$\sigma_n = 2 \sup \{\text{Re}\,(\lambda_j) \colon j = n+1, n+2, \cdots\}$$

we have $\sigma_n \to -\infty$, $n \to \infty$, and

$$e^{2\,\text{Re}(\lambda_j)t^n} \leqq e^{\sigma_n t^n} \qquad \text{all } j, n$$

Consequently,

$$\|y(t^n, \mathbf{u}^n) - y_n(t^n, \mathbf{u}^n)\|^2 \leqq e^{\sigma_n t^n} \|y_0 - y_0^n\|^2 + \sum_{j=n+1}^{\infty} \sum_{k=1}^{r_j} C_j^2 \sum_{i=1}^{m} |g_i^{jk}|^2 \tag{3.11}$$

and in the case n is sufficiently large so that Re $(\lambda_j) < 0$ for $j = n+1, n+2, \ldots,$ we have

$$\|y(t^n, \mathbf{u}^n) - y_n(t^n, \mathbf{u}^n)\|^2 \leqq e^{\sigma_n t^n} \|y_0 - y_0^n\|^2 + \left(\frac{16}{|\sigma_n|^2}\right) \sum_{i=1}^{m} \|g_i - g_i^n\|^2. \tag{3.12}$$

Remark. From the estimate (3.12) if follows that for Re $(\lambda_j) \neq 0$, $j = 1, 2, \cdots$, the best convergence will be obtained by solving the equations (3.6) for $j = j_1, j_2, \cdots, j_n$, where Re $(\lambda_{j_1}) \geqq$ Re $(\lambda_{j_2}) \geqq \cdots \geqq$ Re (λ_{j_n}); that is, by removing the modes corresponding to the larger eigenvalues first. In the case the eigenvalues are purely imaginary, it would be best to first remove the modes for which $|y_0^{jk}|$, and $|g_i^{jk}|$ are largest.

Finally in the case $y_1 \neq 0$, setting $y_1^{jk} = (y_1, \varphi_{jk})$, we see, by (3.2), the system of equations (3.3) becomes

$$\sum_{i=1}^{m} g_i^{jk} \int_0^{t^*} e^{-\lambda_j \tau} u_i(\tau)\, d\tau - y_1^{jk} e^{-\lambda_j t^*} = -y_0^{jk} \qquad k = 1, \cdots, r_j, \quad j = 1, 2, \cdots, \tag{3.13}$$

and the equations (3.56) become

$$G_j \mathbf{F}_j(t_{il}) - \mathbf{y}_j^1 e^{-\lambda_j t^n} = -\mathbf{y}_0^j, \qquad j = 1, 2, \cdots, \quad l = 1, 2, \cdots, s_i, \quad i = 1, 2, \cdots, m \tag{3.14}$$

where

$$\mathbf{y}_j^1 = \begin{bmatrix} y_1^{j1} \\ \vdots \\ y_1^{jr_j} \end{bmatrix}.$$

The system (3.14) again gives a set of nonlinear equations for the switching times (albeit, more complex) which can be solved by a suitable modification of the methods of § 4.

4. Numerical solution. In this section we return to the set of nonlinear equations (3.6) which determine the switching times, and consider methods for their numerical solution. We will assume that rank $G_j = r_j$ for each $j = 1, 2, \cdots, n$. (This assumption occurred in the controllability assumption (H5).) If it does not hold, it means that there are not enough controls to force the system to zero from all the possible starting points. The following is well known.

LEMMA. *If* rank $G_j = r_j$, $j = 1, 2, \cdots, n$, *there exist unique vectors* $\mathbf{a}_j = (a_j^1, \cdots, a_j^m)$ *such that* $G_j \mathbf{a}_j = \mathbf{y}_j\, j = 1, 2, \cdots, n$.

The problem of solving (3.6) then reduces to the nm nonlinear equations in $\sum_{i=1}^{m} s_i$ unknowns (t_{il}) $l=1,2,\cdots,s_i$, $i=1,2,\cdots,m$,

$$
\begin{array}{ll}
 & \pm[1-2\exp(-\lambda_j t_{11})+2\exp(-\lambda_j t_{12})\cdots(-1)^{s_1}\exp(-\lambda_j t^n)]=\lambda_j a_j^1 \\
 & \vdots \\
(4.1) & \pm[1-2\exp(-\lambda_j t_{i1})+2\exp(-\lambda_j t_{i2})\cdots(-1)^{s_i}\exp(-\lambda_j t^n)]=\lambda_j a_j^i \\
 & \vdots \\
 & \pm[1-2\exp(-\lambda_j t_{m1})+2\exp(-\lambda_j t_{m2})\cdots(-1)^{s_m}\exp(-\lambda_j t^n)]=\lambda_j a_j^m
\end{array}
$$

for $j=1,2,\cdots,n$, $i=1,2,\cdots,m$.

We now investigate methods for the numerical solution of (4.1). For simplicity consider a particular case of the second example in § 2, with $\Omega=[0,1]$, $y_0=[Q_0,Q_1]$ $g=[0,b]$. The problem, then, is to bring the string with initial position Q_0 and velocity Q_1 to rest (i.e., position and velocity zero) in minimum time. In this case $r_j=1$ for each $j=1,2,\cdots$, and so we consider the solution of (4.1) for $m=1$. The methods developed will extend to cover any finite number of controls. We have

$$
\begin{aligned}
y_0^j &= \left([Q_0(x),Q_1(x)],\left[\frac{\sin(jx)}{j\sqrt{\pi}}, i\frac{\sin(jx)}{\sqrt{\pi}}\right]\right)_X \qquad j=1,2,\cdots,\\
&= \frac{j}{\sqrt{\pi}}Q_0^j-\frac{i}{\sqrt{\pi}}Q_1^j
\end{aligned}
$$

where

$$
Q_0^j=\int_0^\pi Q_0(x)\sin(jx)\,dx, \qquad Q_1^j=\int_0^\pi Q_1(x)\sin(jx)\,dx
$$

and

$$
|y_0^j|^2=\frac{j^2}{\pi^2}(Q_0^j)^2+\frac{1}{\pi}(Q_1^j)^2, \qquad j=1,2,\cdots.
$$

Similarly $g_j=ib_j/\sqrt{\pi}$ and $|g_j|^2=(1/\pi)b_j^2$ where

$$
b_j=\int_0^\pi b(x)\sin(jx)\,dx,
$$

$j=1,2,\cdots$, and $\lambda_j=\pm ij$, $j=1,2,\cdots$.

The equation (4.1) with $m=1$ for the switching times $t_1, t_2,\cdots,t_{s_1}$ becomes

$$
\pm(1-2\exp(-\lambda_j t_1)+2\exp(-\lambda_j t_2)\cdots(-1)^n\exp(-\lambda_j t_{s_1}))=\frac{\lambda_j y_0^j}{g_j}
$$

or, equivalently, when separated into real and imaginary parts,

$$
(4.2)\qquad
\begin{aligned}
\sin(jt_1)-\sin(jt_2)+\cdots+\frac{(-1)^{n+1}}{2}\sin(jt_{s_1}) &= \pm\left(-\frac{jQ_1^j}{b_j}\right)\\
\cos(jt_1)-\cos(jt_2)+\cdots+\frac{(-1)^{n+1}}{2}\cos(jt_{s_1}) &= \pm\left(-\frac{1}{2}-\frac{j^2Q_0^j}{2b_j}\right)
\end{aligned}
$$

for $j=1,2,\cdots,n$. (+ indicates the control starts with $+1$; $-$ it starts with -1.)

As a particular example take $b(x)=(2/\pi)x$, $Q_1(x)=0$, and

$$Q_0(x)=\frac{2}{\pi}\sum_{j=1}^{\infty}\frac{1+(-1)^{j+1}}{j^3}\sin(jx),\qquad 0\leqq x\leqq\pi.$$

It is easy to verify that the control $u=-\frac{1}{2}$ transfers $[Q_0, Q_1]$ to $[0, 0]$ in π seconds, and so for this example $t^*\leqq\pi$. Also

$$|g_j|^2=\frac{4}{\pi j^2},\quad\text{and}\quad |y_0^j|^2=\frac{1}{\pi}\left(\frac{1+(-1)^{j+1}}{j^4}\right),\qquad j=1,2,3,\cdots;$$

accordingly the equations (4.2) should be solved for $j=1,2,\cdots,n$. In this case equations (4.2) become

$$\begin{aligned}
&\cos(t_1)-\cos(t_2)+\cdots+\frac{(-1)^{s_1+1}}{2}\cos(t_{s_1})=1\\
&\sin(t_1)-\sin(t_2)+\cdots+\frac{(-1)^{s_1+1}}{2}\sin(t_{s_1})=0\\
&\quad\vdots\qquad\quad\vdots\qquad\qquad\quad\vdots\qquad\qquad\vdots\\
&\cos(jt_1)-\cos(jt_2)+\cdots+\frac{(-1)^{s_1+1}}{2}\cos(jt_{s_1})=\frac{3-(-1)^j}{4}\\
&\sin(jt_1)-\sin(jt_2)+\cdots+\frac{(-1)^{s_1+1}}{2}\sin(jt_{s_1})=0\\
&\quad\vdots\qquad\quad\vdots\qquad\qquad\quad\vdots\qquad\qquad\vdots\\
&\cos(nt_1)-\cos(nt_2)+\cdots+\frac{(-1)^{s_1+1}}{2}\cos(nt_{s_1})=\frac{3-(-1)^n}{4}\\
&\sin(nt_1)-\sin(nt_2)+\cdots+\frac{(-2)^{s_1+1}}{2}\sin(nt_{s_1})=0,\qquad j=1,2,\cdots,n
\end{aligned}\tag{4.3}$$

and so the problem is reduced to the solution of a set of n nonlinear equations in s_1 unknowns. Unfortunately we cannot always expect $s_1=n$ (as was done in [11]), and even when s_1 was equal to n, we found the Newton–Raphson scheme used in [9] was very sensitive to the initial guess and difficult to use (particularly for large n). Clearly given any solution to (4.3), we can generate an infinite number of different solutions by adding multiples of 2π to any of the unknowns, and so this instability is to be expected. To understand the key to a successful numerical scheme we return to Theorem 3.1, where it was shown that the equations (4.3) have a solution under the restrictions that $0<t_1<t_2<\cdots<t_{s_1}$, and that t_{s_1} is minimal. (In this particular case, $g_j\neq 0$ all j, and by Theorem 3.1 this solution is unique.) For $\mathbf{t}=(t_1,\cdots,t_{s_1})$ define

$$f_j(\mathbf{t})=\cos(jt_1)-\cos(jt_2)+\cdots+\frac{(-1)^{s_1+1}}{2}\cos(jt_{s_1})+\frac{(-1)^j-3}{4}$$

$$g_j(\mathbf{t})=\sin(jt_1)-\sin(jt_2)+\cdots+\frac{(-1)^{s_1+1}}{2}\sin(jt_{s_1}),\qquad j=1,\cdots,n;$$

then the desired solution of (4.3) is the unique solution of the following nonlinear programming problem,

$$\begin{gathered}\mathrm{P}_1\text{: minimize } t_{s_1} \text{ subject to } 0<t_1<t_2<\cdots<t_{s_1}\\ f_j(\mathbf{t})=0,\qquad g_j(\mathbf{t})=0\quad\text{for } j=1,\cdots,n.\end{gathered}$$

Of course, there are many ways of solving P_1 numerically. One, which proved successful here, was to first remove the inequality constraints, by means of the transformation:

$$\begin{aligned} t_1 &= y_1^2 \\ t_2 &= y_1^2 + y_2^2 \\ &\vdots \\ t_{s_1} &= y_1^2 + y_2^2 + \cdots + y_{s_1}^2 \end{aligned}$$

Then when we set $\mathbf{y} = (y_1, \cdots, y_{s_1})$, and

$$\bar{f}_j(\mathbf{y}) = f_j(\mathbf{t}), \quad \bar{g}_j(\mathbf{y}) = g_j(\mathbf{t}), \quad j = 1, 2, \cdots, n$$

P_1 is equivalent to,

$$\begin{aligned} &P_2\text{: minimize } y_1^2 + y_2^2 + \cdots + y_{s_1}^2 \text{ subject to} \\ &\bar{f}_j(\mathbf{y}) = 0, \quad \bar{g}_j(\mathbf{y}) = 0, \quad j = 1, 2, \cdots, n. \end{aligned} \tag{4.4}$$

Various methods for solving the equality constrained optimization problem P_2 were used, amongst them (i) the SUMPT (penalty function) method [5], (ii) Lagrange multipliers, and (iii) the Powell–Hestenes method of multipliers [12], [21]. Of these the Powell–Hestenes method proved the most satisfactory and accurate. Both methods (i) and (iii) proved relatively insensitive to the initial guess for $\mathbf{y}$. The disadvantage of these methods is that for high accuracy they can require many unconstrained minimizations.

The method finally decided upon, was to firstly solve P_2 by the SUMPT method for moderate values of the penalty term. That is, we solved

$$P_3\text{: minimize } y_1^2 + y_2^2 + \cdots + y_{s_1}^2 + r \sum_{j=1}^{n} \{[\bar{f}_j(\mathbf{y})]^2 + [\bar{g}_j(\mathbf{y})]^2\},$$

for increasing values of r, until $|\bar{f}_j|$, $|\bar{g}_j| \approx 0.1$. This usually only required one or two unconstrained minimizations. The resultant solution $\mathbf{y}$ was then used as the initial guess for the problem

$$P_4\text{: minimize } \bar{f}_1^2(\mathbf{y}) + \cdots + \bar{f}_n^2(\mathbf{y}) + \bar{g}_1^2(\mathbf{y}) + \cdots + \bar{g}_n^2(\mathbf{y}).$$

Since the initial guess was already reasonably close, convergence to the true solution was rapid, and of course only one unconstrained minimization is required. In this way the locally quadratic convergence of Newton type methods could be taken advantage of, with relative insensitivity in the choice of the starting point. (Notice that if $s_1 = 2n$, Newton's method could be used for the solution of $\bar{f}_j(\mathbf{y}) = \bar{g}_j(\mathbf{y}) = 0$, $j = 1, 2, \cdots, n$, instead of P_4.) Finally both P_3 and P_4 are just the minimization of a sum of squares, and were solved using a modification due to Brown and Dennis [2] of the Levenberg–Marquardt algorithm [17], [18].

Unless the eigenvalues $\{\lambda_1, \cdots, \lambda_n\}$ are real the number of switches s_1 cannot in general be determined a priori. (If the eigenvalues are real then $s_1 \leqq n$; see the remark following Theorem 3.1). However if too few switches are used, increasing the number of switches decreases the calculated final time t_{s_1}. If too many switches are used, the solution of P_3 has that extra number of components zero. In this way the actual number of switches can be found by starting with, say, $s_1 = 2n$, and increasing s_1 until t_{s_1} stops decreasing or the solution of P_3 is returned with some components zero.

The numerical results were obtained on a UNIVAC 1108. The computation time required for the solution of (4.3) (with the residuals $<10^{-4}$), for, say, $n = 8$, $s_1 = 14$ was ~ 20 seconds.

Table 1 gives the values of the switching times for various values of n, and Fig. 1 gives the corresponding output distributions at the final times. For $n = 8$ the maximum magnitude of the output was less than $(.015)\times(\pi/2)$, and was too small to be accurately plotted. Finally, in the case $n = 8$, the energy norm was reduced by a factor of eight, from 1.6 (to 2 s.f) for the initial distribution to 0.2 for the final distribution.

TABLE 1

Switching times for nth approximate control minimum time problem.

n	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}
1	0.50536	1.8235												
2	0.20163	1.2696	1.6941	2.6462										
3	0.10098	0.79505	1.0855	1.9731	2.2622	2.8960								
4	0.06000	0.53088	0.74085	1.4284	1.6655	2.3523	2.5615	2.9960						
5	0.04000	0.37700	0.53502	1.0738	1.2651	1.8490	2.0413	2.5798	2.7365	3.0465				
8	0.02412	0.23328	0.33370	0.68560	0.81389	1.2159	1.3534	1.7680	1.9054	2.3072	2.4354	2.7867	2.8869	3.0822

As a final remark we observe that the numerical solution of problem (II) can be handled with only minor changes. Namely, we construct an approximating sequence of controls $\mathbf{u}^n$, where $\mathbf{u}^n$ minimizes

$$(4.5) \qquad \sum_{j=1}^{n} \sum_{k=1}^{r_j} \left| y_0^{jk} e^{\lambda_j T} + \sum_{i=1}^{m} g_i^{jk} \int_0^T e^{\lambda_j (T-\tau)} u_i(\tau)\, d\tau - y_1^{jk} \right|^2, \quad |u_i| \leqq 1$$

$i = 1, 2, \cdots, m$, and $y_1^{jk} = (y_1, \varphi_{jk})$. For this problem the theorem of Halkin [10] automatically guarantees $\mathbf{u}^n$ may be taken bang-bang and finite switching (of course, no controllability assumptions are needed here), and the same assumptions as in Theorem 3.1 guarantee uniqueness of the switching times. Accordingly, when we evaluate the integrals in (4.5) in terms of these switching times (c.f., (3.5)), (4.5)

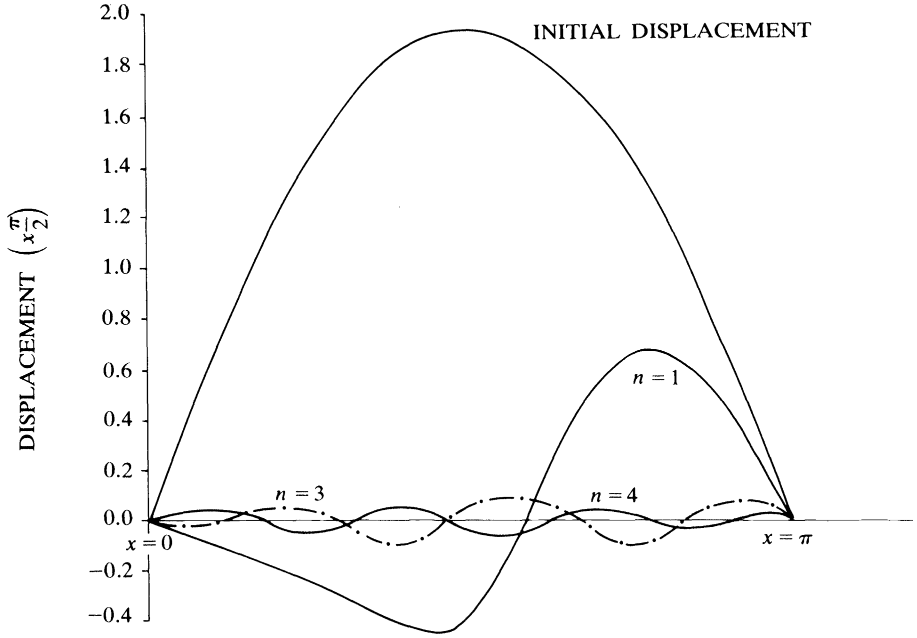

FIG. 1. *Wave displacement at final time for min. time problem* (n = 1, 3, 4).

becomes a minimization of a sum of squares of functions of the switching times. This is just a problem of the type P_4, except now the final time is fixed.

For instance, when we take the previous example with $y_1=0$, and the switching times of u^n be $\{t_1, \cdots, t_s, T\}$, u^n can be computed from the unconstrained minimization,

$$P_4': \quad \text{minimize } \tilde{f}_1(\mathbf{y})^2+\cdots+\tilde{f}_n(\mathbf{y})^2+\tilde{g}_1(\mathbf{y})^2+\cdots+\tilde{g}_n(\mathbf{y})^2$$

$$\mathbf{y}=(y_1, \cdots, y_s), \quad \text{where}$$

$$\tilde{f}_j(\mathbf{y})=jQ_0^j \pm\left(\frac{-b_j}{j}\right)[1-2\cos(jy_1^2)+\cdots+2(-1)^s\cos j(y_1^2+\cdots+y_s^2)$$
$$+(-1)^{s+1}\cos(jT)]$$

$$\tilde{g}_j(\mathbf{y})=-Q_1^j \pm\left(\frac{-b_j}{j}\right)[2\sin(jy_1^2)+\cdots+2(-1)^{s-1}\sin j(y_1^2+\cdots+y_s^2)$$
$$+(-1)^s\sin(jT)]$$

$j=1, \cdots, n$. The switching times of u^n are then calculated from

$$t_1=y_1^2$$
$$t_2=y_1^2+y_2^2$$
$$\vdots \quad \vdots$$
$$t_s=y_1^2+y_2^2+\cdots+y_s^2.$$

With the previous initial conditions Q_0 and $Q_1=0$, the switching time for $n=8$, and various final times T, are given in Table 2 and Fig. 2.

TABLE 2

Switching times for $n = 8$ fixed time problem.

T	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}
2.0	0.164	0.475	0.609	0.906	1.040	1.359	1.475	2.000				
2.5	0.105	0.414	0.530	0.838	0.968	1.351	1.505	1.905	2.031	2.500		
3.0	0.030	0.266	0.377	0.766	0.908	1.348	1.498	1.950	2.100	2.537	2.674	3.000

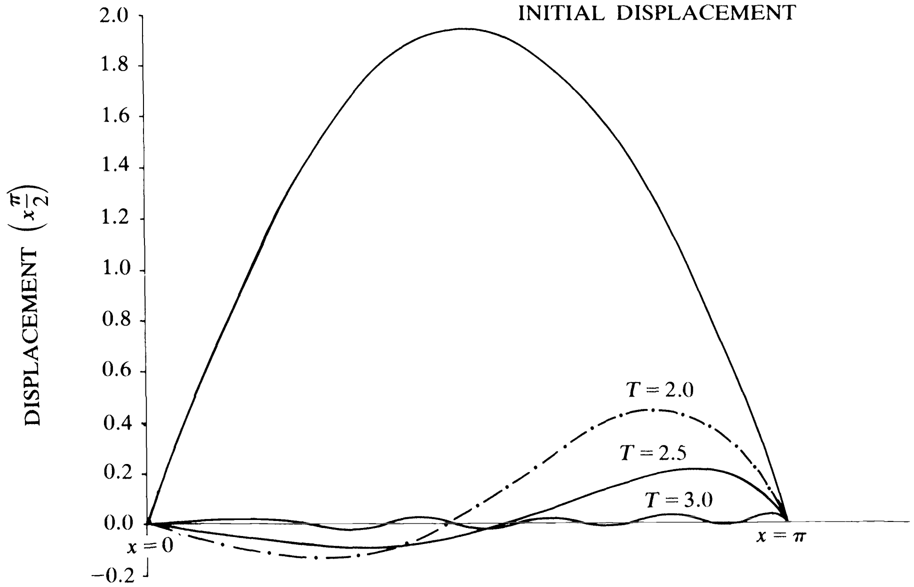

FIG 2. *Wave displacement for fixed time problem at various final times* ($n = 8$).

REFERENCES

[1] A. V. BALAKRISHNAN, *Optimal control problems in Banach spaces*, this Journal, 3 (1965), pp. 152–180.

[2] K. M. BROWN AND J. E. DENNIS, *Derivative free analogues of the Levenberg–Marquardt and Gauss algorithms for non-linear least squares approximations*, Numer. Math., 18 (1972), pp. 289–297.

[3] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, vol. I, Interscience, New York, 1964.

[4] H. FATTORINI, *Some remarks on complete controllability*, this Journal, 4 (1966), pp. 686–694.

[5] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming*: *Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.

[6] A. FRIEDMAN, *Optimal control in Banach spaces*, J. Math. Anal. Appl., 18 (1967), pp. 469–491.

[7] K. GLASHOFF AND N. WECK, *Boundary control of parabolic differential equations in arbitrary dimensions*: *Supremum-norm problems*, this Journal, 14 (1976), pp. 662–681.

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

[9] ——— *Time optimal control of a linear hyperbolic system*, Internat. J. Control, 12 (1970), pp. 645–656.

[10] H. HALKIN, *A generalization of LaSalle's bang-bang principle*, this Journal, 2 (1965), pp. 199–202.

[11] H. HERMES AND J. P. LASALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.

[12] M. R. HESTENES, *Multiplier and gradient methods*, J. Optimization Theory Appl., 4 (1969), pp. 303–320.

[13] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semi-groups*, American Mathematical Society. Providence, RI, 1958.

[14] G. KNOWLES, *Time optimal control in infinite-dimensional spaces*, this Journal, 14 (1976), pp. 919–933.

[15] ——— *Some remarks on infinite dimensional nonlinear control without convexity*, this Journal, 15 (1977), pp. 830–840.

[16] S. G. KREIN, *Linear Differential Operators in a Banach Space*, Translations of Mathematical Monographs, vol. 29, American Mathematical Society, Providence, RI, 1971.

[17] K. LEVENBERG, *A method for the solution of certain nonlinear problems in least squares*, Quart. Appl. Math., 2 (1944), pp. 164–168.

[18] D. W. MARQUARDT, *An algorithm for least squares estimation of non-linear parameters*, this Journal, 2 (1963), No. 2.

[19] E. NELSON, *Analytic vectors*, Ann. of Math., 70 (1959), pp. 572–615.

[20] L. PONTRYAGIN, V. BOLTYANSKII, R. GRAMKRELIDZE AND E. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[21] M. J. D. POWELL, *A method of non-linear constraints in minimization problems*, Optimization. Academic Press, London, 1969.

[22] T. SEIDMAN AND W. CHEWING, *A convergent scheme for boundary control of the heat equation*, this Journal, 15 (1977), pp. 64–72.

[23] R. TRIGGIANI, *Extensions of rank conditions for controllability and observability to Banach spaces and unbounded operators*, this Journal, 14 (1976), pp. 313–338.

[24] K. TSUJIOKA, *Remarks on controllability of second order evolution equations in Hilbert spaces*, this Journal, 8 (1970), pp. 90–99.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 1, 1979

0036–0129/79/1701–0003 $01.00/0

STRONG STABILIZABILITY OF LINEAR CONTRACTIVE CONTROL SYSTEMS ON HILBERT SPACE*

N. LEVAN† AND L. RIGBY‡

Abstract. Strong stabilizability of linear "contractive" systems on Hilbert space, that is those systems denoted by (A, B) and described by the equation $\dot{x} = Ax + Bu$, in which A generates a semigroup of Hilbert space contraction operators, is studied. Necessary and sufficient conditions are given, which are shown to depend on controllability of the system (A, B), and that of (A^*, B) also. Our technique is based on some rather simple properties of invariant and reducing subspaces of Hilbert space operators, and on a canonical decomposition of contraction semigroups due to B. Sz-Nagy and C. Foias.

1. Introduction. Given a system (A, B), that is the equation

$$\dot{x} = Ax + Bu$$

on some abstract space, the problem of finding a state feedback operator $F: u = Fx$, so that the feedback system $(A + BF, B)$ is stable in a suitable sense, is called the stabilizability problem. Such a problem was considered by R. Datko [1] and M. Slemrod [2a], [2b] for the case of Hilbert space, and by R. Triggiani [3] for the Banach space case. In [2b] Slemrod studied the case in which A generates a strongly continuous semigroup of contraction operators. He obtained, among other results, conditions for strong stabilizability of unitary groups, and for weak stabilizability of contraction semigroups. Recently R. E. O'Brien [4] and C. D. Benchimol [5] also studied weak stabilizability of contraction semigroups. Their methods were quite different from those of Slemrod.

This paper will study only strong stabilizability of contractive systems on Hilbert space. The main feature of our work is that, for these systems, controllability properties of the systems (A, B) and (A^*, B) play an equally important role in the problem. This was not pointed out before.

In § 2 we give some basic facts about the controllable and uncontrollable subspaces of the systems (A, B) and (A^*, B). Section 3 will be concerned with some basic properties of contraction semigroups, and with the canonical decomposition of these semigroups. This decomposition will be the basic tool for our stabilizing procedure, which will be given in § 4.

2. Mathematical preliminaries. All spaces that we deal with will be separable complex Hilbert spaces with inner product $[\cdot, \cdot]$ and norm $\|\cdot\|$. Operator will always mean linear, but not necessarily bounded Hilbert space operator, while semigroup will always be a (C_0) strongly continuous semigroup of bounded linear operators over a Hilbert space [6].

By a system (A, B) on a Hilbert space H we mean the state-control equation

$$\dot{x} = Ax + Bu \tag{2.1}$$

where x belongs to H—the state space—and u belongs to U—the control space. The operator A is closed with dense domain $\mathcal{D}(A)$ in H, and it is always taken to be the generator of a (C_0) semigroup, denoted by $[T(t); t \geqq 0]$, over H.

* Received by the editors July 20, 1977, and in revised form March 7, 1978. This work was supported in part by the National Science Foundation under Grant Eng 75-11876.

† Department of System Science, University of California, Los Angeles, California 90024.

‡ Department of Computing and Control, Imperial College, London SW7 2BZ, Great Britain.

By a feedback operator F we mean a bounded linear operator from H to U, defined by $u = Fx$. Consequently, the feedback system $(A+BF, B)$ is characterized by the equation

$$\dot{x} = (A+BF)x + Bu \tag{2.2}$$

where, of course, $(A+BF)$ is also the generator of a (C_0) semigroup, denoted by $[S(t);\, t \geqq 0]$, on H [6].

Consider the "mild" solution [6] of (2.1):

$$x(t) = T(t)x(0) + \int_0^t T(t-\sigma)Bu(\sigma)\, d\sigma, \qquad t > 0, \tag{2.3}$$

for $x(0)$ in H. Following [6], given $x(0) = 0$, a state x in H is called controllable if for an $\varepsilon > 0$, there is a $u(\cdot)$ in $L_2[(0, t);\, U]$ so that

$$\left\| x - \int_0^t T(t-\sigma)Bu(\sigma)\, d\sigma \right\| < \varepsilon \quad \text{for some } t > 0.$$

Thus, the set of all controllable states of (A, B), denoted by M_c, is [6]

$$M_c = \overline{\bigcup_{t \geqq 0} T(t)BU} \tag{2.4}$$

where, as always, $\overline{\qquad}$ denotes the closure.

The orthogonal complement in H of M_c, denoted by M_{uc}, is then a set of uncontrollable states of the system, and

$$M_{uc} = \bigcap_{t \geqq 0} \operatorname{Ker}\, [B^*T(t)^*]. \tag{2.5}$$

If $M_c = H$, equivalently, $M_{uc} = \{0\}$, then (A, B) is said to be (approximate) controllable on H. A subspace M of H is said to be controllable for (A, B) if $M \subseteq M_c$.

Clearly M_c is a closed subspace of H and it is invariant for $[T(t);\, t \geqq 0]$, $T(t)M_c \subseteq M_c$ for all $t \geqq 0$. Similarly, M_{uc} is invariant for $[T(t)^*;\, t \geqq 0]$.

We associate with (A, B) the "adjoint system" (A^*, B):

$$\dot{y} = A^*y + Bu. \tag{2.6}$$

Then since A generated $[T(t);\, t \geqq 0]$, its adjoint A^* generates the adjoint semigroup $[T(t)^*;\, t \geqq 0]$. We have for (A^*, B)

$$M_{*c} = \overline{\bigcup_{t \geqq 0} T(t)^*BU} \tag{2.7}$$

and

$$M_{*uc} = \bigcap_{t \geqq 0} \operatorname{Ker}\, [B^*T(t)]. \tag{2.8}$$

M_{*c} is now invariant for $[T(t)^*;\, t \geqq 0]$ while M_{*uc} is invariant under $[T(t);\, t \geq 0]$.

It is clear from the above that $\overline{BU}$ is contained in M_c and in M_{*c}. Therefore M_{uc} and M_{*uc} are subspaces of $\operatorname{Ker} B^*$. We now prove

LEMMA 2.1. (i) *Any proper subspace of* Ker B^* *which is invariant for* $[T(t)^*;\, t \geqq 0]$ (*resp.* $[T(t);\, t \geqq 0]$) *is contained in* M_{uc} (*resp.* M_{*uc}). *Equivalently* M_c (*resp.* M_{*c}) *has no proper subspace containing* $\overline{BU}$ *and is invariant for* $[T(t);\, t \geqq 0]$ (*resp.* $[T(t)^*;\, t \geqq 0]$).

(ii) *Any reducing subspace of* $[T(t);\, t \geqq 0]$ *in* Ker B^* *is contained in* $M_{uc} \cap M_{*uc}$.

Proof. Suppose M_{uc} is not trivial, i.e., the system (A, B) is uncontrollable. Let $M \subset \operatorname{Ker} B^*$ be invariant for $[T(t)^*; t \geqq 0]$; then $T(t)^* M \subseteq M \subset \operatorname{Ker} B^*$ for all $t \geqq 0$. Hence

$$[T(t)^* M, BU] = 0 = [M, T(t)BU], \qquad t \geqq 0.$$

Therefore $M \perp M_c$ which implies that $M \subseteq M_{uc}$. The proof is similar for $M \subset \operatorname{Ker} B^*$ and is invariant under $[T(t); t \geqq 0]$.

For the equivalent statement we have only to note that if N is an invariant subspace of $[T(t); t \geqq 0]$ and $M_c \supset N \supset \overline{BU}$, then $M_{uc} \subset N^\perp \subset \operatorname{Ker} B^*$ and $N^\perp$ is invariant for $[T(t)^*; t \geqq 0]$. But this contradicts the previous assertion. This completes the proof of part (i).

Part (ii) is an easy consequence of part (i).

It is clear from the above that we have

THEOREM 2.1. *The system (A, B) (resp. (A^*, B)) is uncontrollable if and only if $[T(t)^*; t \geq 0]$ (resp. $[T(t); t \geqq 0]$) has an invariant subspace in* Ker B^*.

Finally, we recall the following definitions which will be needed in subsequent sections.

DEFINITION 2.1. A system (A, B) is said to be *s(strong)-stable* if: $T(t)x \to 0$, $t \to \infty$, for all x in H.

If (A, B) is not s-stable and if a feedback F can be found so that the system $(A + BF, B)$ is s-stable, then (A, B) is said to be *s-stabilizable.*

3. Contractive systems. We study in this section some basic properties with respect to s-stabilizability of contractive systems.

A system (A, B) is called contractive if its operator A generates a semigroup of contraction operators, or simply a contraction semigroup.

First, let us recall the following important properties of contraction semigroups, due originally to R. S. Phillips [7].

LEMMA 3.1. *Let $[T(t); t \geqq 0]$ be a semigroup on H, with generator A. Then,*

(i) *$[T(t); t \geqq 0]$ is a contraction semigroup: $\|T(t)\| \leqq 1$ for all $t \geqq 0$, if and only if A is dissipative:* Re $[Ax, x] \leqq 0$ *for all x in $\mathcal{D}(A)$, and furthermore it does not admit any dissipative extension in H. Hence A is called maximal dissipative.*

(ii) *$[T(t); t \geqq 0]$ is an isometric semigroup: $\|T(t)x\| = \|x\|$ for all $t \geqq 0$ and all x in H, if and only if A is maximal dissipative and* Re $[Ax, x] = 0$ *for all x in $\mathcal{D}(A)$.*

(iii) *$[T(t); t \geqq 0]$ is a unitary semigroup—or what amounts to the same thing $[T(t); -\infty < t < \infty]$ is a unitary group—$\|T(t)x\| = \|x\| = \|T(t)^* x\|$ for all $t \geqq 0$ and all x in H, if and only if A is skewadjoint: $A = -A^*$.*

We now recall the notion of a completely nonunitary (cnu) semigroup of B. Sz-Nagy and C. Foias [8].

DEFINITION 3.1. A semigroup $[T(t); t \geqq 0]$ on H is cnu if for each nonzero x in H there is a $t > 0$ such that, either $\|T(t)x\| \neq \|x\|$ or $\|T(t)^* x\| \neq \|x\|$.

It is evident that the only subspace which reduces a cnu semigroup to a unitary one is the trivial subspace $\{0\}$.

The following canonical decomposition of contraction semigroups of Nagy and Foias will be the main tool of this paper.

THEOREM 3.1 [8], [9]. *To every contraction semigroup $[T(t); t \geqq 0]$ on H, there are reducing subspaces $H_{cnu}(T)$ and $H_u(T)$ such that:*

$$H = H_{cnu}(T) \oplus H_u(T) \tag{3.1}$$

uniquely. Hence $[T(t);\, t \geqq 0]$ *admits the decomposition*

$$T(t) = T_{\mathrm{cnu}}(t) \oplus T_{\mathrm{u}}(t), \qquad t \geqq 0, \tag{3.2}$$

where $T_{\mathrm{cnu}}(t) = T(t)|_{H_{\mathrm{cnu}}(T)}$ *is cnu and* $T_{\mathrm{u}}(t) = T(t)|_{H_{\mathrm{u}}(T)}$ *is unitary.*

$H_{\mathrm{u}}(T)$ *is the maximal reducing subspace on which the semigroup is unitary, and*

$$H_{\mathrm{u}}(T) = \{x \text{ in } H;\, \|T(t)x\| = \|x\| = \|T(t)^*x\|,\, t > 0\}. \tag{3.3}$$

It follows from this theorem that we have

COROLLARY 3.1. (i) *A contraction semigroup* $[T(t);\, t \geqq 0]$ *and its adjoint* $[T(t)^*;\, t \geqq 0]$ *can either be both cnu or both unitary.*

(ii) *The subspaces* $\mathscr{D}(A) \cap H_{\mathrm{u}}(T)$ *and* $\mathscr{D}(A^*) \cap H_{\mathrm{u}}(T)$ *are dense in* $H_{\mathrm{u}}(T)$, *and* $\mathscr{D}(A) \cap H_{\mathrm{u}}(T) \equiv \mathscr{D}(A^*) \cap H_{\mathrm{u}}(T)$.

(iii) *If M is a reducing subspace of* $[T(t);\, t \geqq 0]$, *then the semigroup is unitary on M if and only if*

$$\operatorname{Re}[Ax, x] = 0 = \operatorname{Re}[A^*x, x], \qquad x \text{ in } \mathscr{D}(A) \cap M. \tag{3.4}$$

Proof: Part (i) is trivial. We only have to note that $[T(t)^*;\, t \geqq 0]$ is also a contraction semigroup, and $H_{\mathrm{u}}(T) = H_{\mathrm{u}}(T^*)$, by (3.3).

For part (ii), from the canonical decomposition (3.2), $T_{\mathrm{u}}(t)$ is the restriction of $T(t)$ to its reducing subspace $H_{\mathrm{u}}(T)$; hence the subspace $\mathscr{D}(A) \cap H_{\mathrm{u}}(T)$ is the domain of $A|_{H_{\mathrm{u}}(T)}$ which in turn is the generator of $T_{\mathrm{u}}(t)$. Hence $\mathscr{D}(A) \cap H_{\mathrm{u}}(T)$ is dense in $H_{\mathrm{u}}(t)$. The proof is similar for the subspace $\mathscr{D}(A^*) \cap H_{\mathrm{u}}(T)$. Next, since $[T_{\mathrm{u}}(t);\, t \geqq 0]$ is unitary on $H_{\mathrm{u}}(T)$, therefore $A|_{H_{\mathrm{u}}(T)} = -A^*|_{H_{\mathrm{u}}(T)}$ by Lemma 3.1(iii). Consequently $\mathscr{D}(A) \cap H_{\mathrm{u}}(T)$ is identical to $\mathscr{D}(A^*) \cap H_{\mathrm{u}}(T)$ as expected.

Finally, part (iii) is just a direct application of Lemma 3.1(iii) to the semigroup $[T(t)|_M;\, t \geqq 0]$.

Remark. It is clear that $\mathscr{D}(A) \cap H_{\mathrm{cnu}}(T)$ and $\mathscr{D}(A^*) \cap H_{\mathrm{cnu}}(T)$ are also dense in $H_{\mathrm{cnu}}(T)$. However, they are not in general identical. Also, we do not have in this case the equivalent of Corollary 3.1(iii). For example, for an isometric semigroup, we always have $\|T(t)x\| = \|x\|$, $t \geqq 0$ and x in H, while $\|T(t)^*x\| \neq \|x\|$, $t \geqq 0$ and x in $H_{\mathrm{cnu}}(T)$.

It is evident from the above that, if the subspace $H_{\mathrm{u}}(T)$ of a system (A, B) is *not* trivial, then of course both (A, B) and (A^*, B) can never be s-stable! Thus the very first step in s-stabilizing contractive systems should be that of "converting a contraction semigroup into a cnu one, by means of a suitable feedback operator." This is what we are going to discuss next.

THEOREM 3.2. *Let* $[T(t);\, t \geqq 0]$ *be a contraction semigroup on H with generator A, and B be a bounded linear operator from U to H. Then,*

(i) $A - BB^*$ *generates a contraction semigroup* $[S(t);\, t \geqq 0]$ (*say*) *on H, and*

(ii) $H_{\mathrm{u}}(S) \subseteq H_{\mathrm{u}}(T) \cap \operatorname{Ker} B^*$.

Proof. Part (i) is trivial. We only have to note that $A - BB^*$ generates a (C_0) semigroup which is also contractive, since A is maximal dissipative by Lemma 3.1(i), and $-BB^*$ is bounded dissipative. In fact it is a negative operator on H.

To show part (ii), since $[S(t);\, t \geqq 0]$ is unitary on $H_{\mathrm{u}}(S)$ and $\mathscr{D}(A - BB^*) = \mathscr{D}(A)$, we have

$$\frac{d}{dt}\|S(t)x\|^2 = 0 = 2 \operatorname{Re}[(A - BB^*)S(t)x, S(t)x], \qquad t \geqq 0, \quad x \text{ in } \mathscr{D}(A) \cap H_{\mathrm{u}}(S)$$

or

$$\operatorname{Re}[AS(t)x, S(t)x] = \|B^*S(t)x\|^2, \qquad t \geqq 0, \quad x \text{ in } \mathscr{D}(A) \cap H_{\mathrm{u}}(S).$$

But if x is in $\mathscr{D}(A)$, $S(t)x$, $t \geqq 0$, lies in $\mathscr{D}(A)$, and since A is dissipative, this last equation implies that

$$\text{(3.5a)} \qquad \operatorname{Re}[AS(t)x, S(t)x] = 0, \qquad t \geqq 0, \quad x \text{ in } \mathscr{D}(A) \cap H_{\mathrm{u}}(S)$$

and

$$\text{(3.6a)} \qquad B^*S(t)x = 0, \qquad t \geqq 0, \quad x \text{ in } \mathscr{D}(A) \cap H_{\mathrm{u}}(S).$$

Similarly, from $(d/dt)\|S(t)^*x\|^2$ and since $\mathscr{D}(A) \cap H_{\mathrm{u}}(S) = \mathscr{D}(A^*) \cap H_{\mathrm{u}}(S)$ by Corollary 3.1(ii), we obtain

$$\text{(3.5b)} \qquad \operatorname{Re}[A^*S(t)^*x, S(t)^*x] = 0, \qquad t \geqq 0, \quad x \text{ in } \mathscr{D}(A) \cap H_{\mathrm{u}}(S)$$

and

$$\text{(3.6b)} \qquad B^*S(t)^*x = 0, \qquad t \geqq 0, \quad x \text{ in } \mathscr{D}(A) \cap H_{\mathrm{u}}(S).$$

It then follows from (3.6a) that $H_{\mathrm{u}}(S) \subseteq \operatorname{Ker} B^*$, since $\mathscr{D}(A) \cap H_{\mathrm{u}}(S)$ is dense in $H_{\mathrm{u}}(S)$.

It remains to show that $H_{\mathrm{u}}(S) \subseteq H_{\mathrm{u}}(T)$. This can be seen in two different ways. First, on $\mathscr{D}(A) \cap H_{\mathrm{u}}(S) \subseteq \operatorname{Ker} B^*$ we have $A - BB^* = A$. Hence $S(t)x = T(t)x$ and $S(t)^*x = T(t)^*x$, for $t \geqq 0$ and x in $H_{\mathrm{u}}(S)$, again by the denseness of $\mathscr{D}(A) \cap H_{\mathrm{u}}(S)$. These then imply that the subspace $H_{\mathrm{u}}(S)$ also reduces $[T(t); t \geqq 0]$ since it already reduced $[S(t); t \geqq 0]$. Consequently, by (3.5a), (3.5b) and Corollary 3.1(iii), $[T(t); t \geqq 0]$ is unitary on $H_{\mathrm{u}}(S)$—or better still, by the fact that $[S(t); t \geqq 0]$ is already unitary on $H_{\mathrm{u}}(S)$. Thus $H_{\mathrm{u}}(S) \subseteq H_{\mathrm{u}}(T)$ by the maximality property of $H_{\mathrm{u}}(T)$ from Theorem 3.1. This completes the proof.

Another way of showing that $H_{\mathrm{u}}(S) \subseteq H_{\mathrm{u}}(T)$ is to use (3.6a) and (3.6b) in the following identities [6]:

$$\text{(3.7a)} \qquad S(t)x = T(t)x - \int_0^t T(t-\sigma)BB^*S(\sigma)x \, d\sigma$$

$$\text{(3.7b)} \qquad S(t)^*x = T(t)^*x - \int_0^t T(t-\sigma)^*BB^*S(\sigma)^*x \, d\sigma.$$

Then we obtain as before, $S(t)x = T(t)x$, and $S(t)^*x = T(t)^*x$, for $t \geqq 0$ and x in $\mathscr{D}(A) \cap H_{\mathrm{u}}(S)$, hence for x in $H_{\mathrm{u}}(S)$.

Remark 1. The subspace $\operatorname{Ker} B^*$ is also the unitary subspace $H_{\mathrm{u}}(R)$ of the selfadjoint contraction semigroup $[R(t) = e^{-BB^*t}; t \geqq 0]$. To see this we use (3.4)

$$\operatorname{Re}[-BB^*x, x] = -\|B^*x\|^2 = 0, \qquad x \text{ in } \mathscr{D}(BB^*) \cap H_{\mathrm{u}}(R) = H_{\mathrm{u}}(R)$$

which implies that $H_{\mathrm{u}}(R) \subseteq \operatorname{Ker} B^*$. On the other hand, it is evident that $e^{-BB^*t}x = x$ for $t \geqq 0$ and for all x in $\operatorname{Ker} B^*$. This implies that $\operatorname{Ker} B^*$ reduces $[R(t); t \geqq 0]$ to a unitary semigroup; hence $H_{\mathrm{u}}(R) \supseteq \operatorname{Ker} B^*$; therefore $H_{\mathrm{u}}(R) = \operatorname{Ker} B^*$ as expected. This certainly makes sense since $-BB^*$ is selfadjoint; the only subspace of H on which $-BB^*$ can be skewadjoint is either $\{0\}$ or $\operatorname{Ker} B^*$.

Remark 2. It can be easily seen that Theorem 3.2(i) holds for any dissipative operator C, instead of just for $-BB^*$. If C is also selfadjoint, then $-C$ is positive. In this case we also have as in Remark 1, $H_{\mathrm{u}}(S) \subseteq H_{\mathrm{u}}(T) \cap \operatorname{Ker} C$, where $\operatorname{Ker} C$ is again the unitary subspace $H_{\mathrm{u}}(R)$ of $[R(t) = e^{Ct}; t \geqq 0]$.

If C is just dissipative, then besides (3.5a) and (3.5b), we have

$$\operatorname{Re}[CS(t)x, S(t)x] = 0, \qquad t \geqq 0, \quad x \text{ in } \mathscr{D}(A) \cap H_{\mathrm{u}}(S)$$

$$\operatorname{Re}[C^*S(t)^*x, S(t)^*x] = 0, \qquad t \geqq 0, \quad x \text{ in } \mathscr{D}(A) \cap H_{\mathrm{u}}(S)$$

instead of (3.6a) and (3.6b).

Thus, by Corollary 3.1(iii), if the subspace $H_u(S)$ *reduces* $[T(t); t \geqq 0]$ and $[R(t); t \geqq 0]$, then these semigroups are unitary on $H_u(S)$ and therefore $H_u(S) \subseteq H_u(T) \cap H_u(R)$.

Applying Theorem 3.2 to contractive systems, we have

PROPOSITION 3.1. *Let (A, B) be a contractive system on H. Then the system $(A - BB^*, B)$ is a cnu contractive system if any one of the following conditions holds*:

(i) *The range space BU is dense in H.* [Condition C1]

(ii) *The unitary subspace $H_u(T)$ is controllable for (A, B) or for (A^*, B), i.e., $H_u(T)$ is contained in M_c or in M_{*c}.* [Condition C2]

(iii) *(A, B) and (A^*, B) do not admit any common uncontrollable states, or if they are controllable.* [Condition C3]

Proof. We have from Theorem 3.2(ii)

$$H_u(\mathrm{S}) \subseteq H_u(T) \cap \operatorname{Ker} B^*.$$

Hence $H_u(S)$ is a subspace of $\operatorname{Ker} B^*$. It also reduces $[S(t); t \geqq 0]$, and since controllability is invariant under any bounded feedback—i.e., (A, B) and $(A + BF, B)$ share the same controllable states [6]—we also have $H_u(S) \subseteq M_{uc} \cap M_{*uc}$ by Lemma 2.1(ii). Hence $H_u(S) \subseteq H_u(T) \cap M_{uc} \cap M_{*uc}$. This can also be seen from (3.6a) and (3.6b), i.e.,

$$B^*S(t)x = 0, \qquad t \geqq 0, \quad x \text{ in } \mathscr{D}(A) \cap H_u(S)$$

$$B^*S(t)^*x = 0, \qquad t \geqq 0, \quad x \text{ in } \mathscr{D}(A) \cap H_u(S).$$

Thus from the definitions of M_{uc} and M_{*uc}, equations (2.5) and (2.8), the above imply that $\mathscr{D}(A) \cap H_u(S) \subseteq M_{uc}$ and $\mathscr{D}(A) \cap H_u(S) \subseteq M_{*uc}$. Hence, by the denseness of $\mathscr{D}(A) \cap H_u(S)$, we have $H_u(S) \subseteq M_{uc} \cap M_{*uc}$ as expected. This completes the proof of the proposition.

We note from the above that if $H_u(T)$ is a subspace of $\overline{BU}$, then $H_u(S)$ is also trivial. But this implies that $H_u(T)$ is controllable for (A, B) or for (A^*, B)—which is condition (ii) of the proposition.

We conclude the section by noting that if $[T(t); t \geqq 0]$ is a unitary semigroup, then $H_u(T) \equiv H$, and by Lemma 1.1(iii), $A = -A^*$. Hence $H_u(S) \subseteq M_{uc}$, and we state

PROPOSITION 3.2. *If (A, B) is contractive and $[T(t); t \geqq 0]$ is unitary, then $(A - BB^*, B)$ is a cnu contractive system if the system (A, B) is controllable.*

4. Strong stabilizability of contractive systems. Using the results of § 3 we now show a procedure for s-stabilizing contractive systems.

As in the above, let (A, B) be a given contractive system with A generating a contractive semigroup $[T(t); t \geqq 0]$ on H. Let $M_s(T)$ be the set of s-stable states of the system,

$$M_s(T) = \{x \text{ in } H;\ T(t)x \to 0,\ t \to \infty\}. \tag{4.1}$$

Then clearly $M_s(T)$ is a closed invariant subspace of $[T(t); t \geqq 0]$. Furthermore the semigroup $[T(t); t \geqq 0]$ is cnu on $M_s(T)$, and from Theorem 3.1 we must have $M_s(T) \subseteq H_{cnu}(T)$. We now find conditions for semigroup to be s-stable on $H_{cnu}(T)$.

Now, since $[T(t); t \geqq 0]$ is a contraction semigroup, $\|T(t)x\|^2$ is a nonincreasing function of t. Therefore $\lim \|T(t)x\|^2$, $t \to \infty$, always exists. It then follows that the positive contractions $T(t)^*T(t)$ for $t > 0$ converge in the strong operator topology to a positive contraction P^2 say,

$$P^2 = \lim_{t\to\infty} T(t)^* T(t). \tag{4.2}$$

Similarly, let

$$Q^2 = \lim_{t\to\infty} T(t)T(t)^*. \tag{4.3}$$

Then, it follows that

$$\|Px\|^2 = \lim_{t\to\infty} \|T(t)x\|^2 \tag{4.4}$$

$$\|Qx\|^2 = \lim_{t\to\infty} \|T(t)^*x\|^2 \tag{4.5}$$

and for all $t \geqq 0$,

$$T(t)^* P^2 T(t) = P^2 \tag{4.6}$$

$$T(t) Q^2 T(t)^* = Q^2. \tag{4.7}$$

We prove

THEOREM 4.1. *The contraction semigroups* $[T(t);\, t\geqq 0]$ *and* $[T(t)^*;\, t\geqq 0]$ *over* H *are s-stable on* $H_{\text{cnu}}(T)\,(=H_{\text{cnu}}(T^*))$ *if and only if* $P \equiv Q$ *is a (orthogonal) projection on* H.

Proof. Suppose that $P \equiv Q = P^2$. Then for any x in H, let $y = x - Px$; we have $Py = Qy = Px - P^2x = 0$. Therefore

$$\|Py\|^2 = \|Qy\|^2 = 0 = \lim_{t\to\infty} \|T(t)y\|^2 = \lim_{t\to\infty} \|T(t)^*y\|^2$$

which shows that

$$\begin{aligned}(I-P)H &= \{y \text{ in } H;\ T(t)y \to 0,\, t\to\infty\}\\ &= \{y \text{ in } H;\ T(t)^*y \to 0,\, t\to\infty\}.\end{aligned}$$

Hence $(I-P)H = M_s(T) = M_s(T^*)$; consequently it reduces $[T(t);\ t\geqq 0]$ (to a cnu semigroup). Then so does its orthogonal complement PH. We therefore have

$$T(t)P = PT(t), \quad \text{for } t \geqq 0 \tag{4.8}$$

$$T(t)^*P = PT(t)^*, \quad \text{for } t \geqq 0. \tag{4.9}$$

Combining (4.6), (4.7), (4.8) and (4.9) we find, for all $t \geqq 0$:

$$[T(t)P]^*[T(t)P] = P^2 = P$$

$$[T(t)P][T(t)P]^* = P^2 = P.$$

Hence the semigroup $[T(t);\, t\geqq 0]$ is unitary on PH, while it is cnu on $(I-P)H$. From the uniqueness of the Nagy–Foias canonical decomposition we conclude that $(I-P)H$ $(=M_s(T)=M_s(T^*)) = H_{\text{cnu}}(T)$ and $PH = H_u(T)$. The semigroups $[T(t);\, t\geqq 0]$ and $[T(t)^*;\, t\geqq 0]$ are therefore s-stable on $H_{\text{cnu}}(T)$.

Conversely if $H_{\text{cnu}}(T) = M_s(T) = M_s(T^*)$, then $Px = Qx = 0$ for x in $H_{\text{cnu}}(T)$. Moreover, from the definition of $H_u(T)$ and from (4.4) we find $\|Py\| = \|y\|$ for all y in $H_u(T)$. This of course is equivalent to $Py = y$, since P is a positive contraction. Similarly, $Qy = y$ for all y in $H_u(T)$. Therefore, since $H_{\text{cnu}}(T) \perp H_u(T)$, P and Q are indeed orthogonal projections with range $H_u(T)$. This completes the proof of the theorem.

We note that when $[T(t); t \geqq 0]$ is normal, $T(t)T(t)^* = T(t)^*T(t)$ for $t \geqq 0$. Then $P \equiv Q$ and of course $M_s(T) = M_s(T^*)$. Furthermore, from the spectral theory of normal operators [10] we have readily $H_{cnu}(T) = M_s(T) = M_s(T^*)$. Hence, we have

COROLLARY 4.1. *A normal contraction semigroup $[T(t); t \geqq 0]$ is s-stable on $H_{cnu}(T)$ and $P \equiv Q$ is the projection with range $H_u(T)$.*

We are now ready to state our first s-stabilizing result.

THEOREM 4.2. *A contractive system (A, B) whose semigroup $[T(t); t \geqq 0]$ is normal, and $\mathscr{D}(A) = \mathscr{D}(A^*)$, is s-stabilizable by the feedback $-B^*$*

(i) *if any one of the Conditions* C1, C2, *or* C3 *of Proposition* 3.1 *is satisfied, and*

(ii) *only if $H_u(T)$ is controllable for (A, B) or for (A^*, B).*

Proof. First, we note that if $[T(t); t \geqq 0]$ is normal and $\mathscr{D}(A) = \mathscr{D}(A^*)$, then so is the semigroup $[S(t); t \geqq 0]$ generated by $A - BB^*$. To see this we use the fact that $[T(t); t \geqq 0]$ is normal $\Leftrightarrow \|T(t)x\|^2 = \|T(t)^*x\|^2$, $t \geqq 0$, x in $H \Leftrightarrow \mathrm{Re}\,[Ax, x] = \mathrm{Re}\,[A^*x, x]$, x in $\mathscr{D}(A) \equiv \mathscr{D}(A^*)$. Consequently, we have for $[S(t); t \geqq 0]$

$$\mathrm{Re}\,[(A - BB^*)x, x] = \mathrm{Re}\,[(A - BB^*)^*x, x], \qquad x \text{ in } \mathscr{D}(A).$$

Hence $[S(t); t \geqq 0]$ is also normal.

Now if the conditions of Proposition 3.1 are satisfied, then $[S(t); t \geqq 0]$ is a cnu normal contraction semigroup. Hence by Corollary 4.1 it is s-stable, i.e. (A, B) is s-stabilizable.

Suppose now that the system $(A - BB^*, B)$ is s-stable. Then since $[S(t); t \geqq 0]$ is normal, the adjoint semigroup $[S(t)^*; t \geqq 0]$ is also s-stable when $[S(t); t \geqq 0]$ is s-stable. Using (2.5) in (3.7b) we have $T(t)^*x = S(t)^*x$, $t \geqq 0$, x in M_{uc}—which is just the fact that controllability is invariant under feedback [6]. Hence, from the above, and by assumption we can conclude that $T(t)^*x \to 0$, $t \to \infty$, for all x in M_{uc}. That is, $M_{uc} \subseteq H_{cnu}(T) \Leftrightarrow H_u(T) \subseteq M_c$. Similarly, $M_{*uc} \subseteq H_{cnu}(T) \Leftrightarrow H_u(T) \subseteq M_{*c}$. This completes the proof.

Theorem 4.2 holds in particular for selfadjoint contraction semigroups. In this case $M_c \equiv M_{*c}$; therefore we have

COROLLARY 4.2. *A contractive system (A, B) with $[T(t); t \geqq 0]$ selfadjoint is s-stabilizable by the feedback $-B^*$, if and only if $H_u(T)$ is controllable for (A, B).*

This result was given by Benchimol in [5] using weak stability properties of contraction semigroups.

Let us now s-stabilize the contraction semigroups which satisfied Theorem 4.1. First we prove

LEMMA 4.1. *If $[T(t); t \geqq 0]$ and $[T(t)^*; t \geqq 0]$ are s-stable on $H_{cnu}(T)$, then $[S(t); t \geqq 0]$ generated by $A - BB^*$ and $[S(t)^*; t \geqq 0]$ are s-stable on $H_{cnu}(S)$.*

Proof. Let P^2 be as in Theorem 4.1 and define

$$J^2 = \lim_{t \to \infty} S(t)^*S(t)$$

$$K^2 = \lim_{t \to \infty} S(t)S(t)^*.$$

Then since $H_u(S)$ $(\subseteq H_u(T) \cap \operatorname{Ker} B^*)$ reduces $[T(t); t \geqq 0]$, $S(t)x = T(t)x$, and $S(t)^*x = T(t)^*x$, for $t \geqq 0$ and all x in $H_u(S)$, by Theorem 3.2. It is evident that

$$J^2x = P^2x = K^2x, \quad \text{for } x \text{ in } H_u(S).$$

Now, by assumption and by Theorem 4.1, P is the orthogonal projection with range

$H_u(T)$. Thus the above implies that

$$J^2x = K^2x = x, \quad \text{for } x \text{ in } H_u(S). \tag{4.10}$$

Again, as in the proof of Theorem 4.1 we also have

$$Jx = Kx = x, \quad \text{for } x \text{ in } H_u(S). \tag{4.11}$$

Equations (4.10) and (4.11) and the fact that J and K are positive imply that these operators are projections with range $H_u(S)$. Thus by Theorem 4.1, the lemma is proved.

Combining this lemma and Theorem 4.1, we have

THEOREM 4.3. *If for the semigroup* $[T(t); t \geqq 0]$ *of a contractive system* (A, B)

$$\begin{aligned} H_{\mathrm{cnu}}(T) &= \{x \textit{ in } H;\ T(t)x \to 0, t \to \infty\} \\ &= \{x \textit{ in } H;\ T(t)^*x \to 0, t \to \infty\}, \end{aligned}$$

then the system is s-stabilizable by the feedback $-B^*$

(i) *if any one of the Conditions* C1, C2, *or* C3 *is satisfied, and*

(ii) *only if* $H_u(T)$ *is controllable for* (A, B) *or for* (A^*, B).

Proof. The proof is the same as that of Theorem 4.2 and will be omitted.

Remark. It is of interest to note that in Theorems 4.2, 4.3 and Corollary 4.2, the subspace $H_u(T)$ can be regarded as a set of "s-unstable states" of (A, B). Thus controllability of this set is necessary and sufficient for s-stabilizability of the system. This is an analogue of the well known finite dimensional result of Wonham [11], namely, a finite dimensional state space system is stabilizable if and only if its unstable modes are controllable. For w(weak)-stabilizability of contractive systems, another analogue of Wonham's result was obtained by Benchimol [5]. For a contractive system (A, B) define

$$M_w(T) = \{x \text{ in } H;\ T(t)x \to 0 \text{ weakly}, t \to \infty\}.$$

Then it was shown in [5] that controllability of the w-unstable states $M_w(T)^\perp$ is necessary and sufficient for (A, B) to be w-stabilizable by the feedback $-B^*$. This was obtained by means of a result of Foguel [14], namely, for a contraction semigroup $[T(t); t \geq 0]$, $M_w(T)^\perp = M_w(T^*)^\perp$ and it reduces the semigroup to a unitary one; hence $M_w(T)^\perp \subseteq H_u(T)$ by the Nagy–Foias theorem. It is evident from this and from our results above that if $M_w(T)^\perp = H_u(T)$ then w-stabilizability and s-stabilizability are equivalent. This is indeed the case when $T(t)$ is selfadjoint and when its generator A has a compact resolvent, as is shown in [5] using semigroup theoretic techniques. Our Corollary 4.2 is equivalent to Corollary 3.2 of [5], except that there $M_w(T)$—which in fact becomes $H_u(T)$ as soon as the semigroup is selfadjoint— was required to be controllable. Our results were obtained by means of Theorem 4.1, and in fact there are no results in [5] which are equivalent to Theorems 4.2 and 4.3. We further note that in both Theorem 4.2 and Theorem 4.3 controllability of (A^*, B) also plays a role in s-stabilizability of the system. This is something which does not occur in the finite dimensional case.

In the above we considered the class of contraction semigroups which together with their adjoints were s-stable on the cnu subspace. Now let us consider the case in which only the semigroup itself is s-stable on this subspace. We prove,

LEMMA 4.2. *A contraction semigroup* $[T(t); t \geqq 0]$ *on H is s-stable on* $H_{\mathrm{cnu}}(T)$ *if and only if* $[T(t)^*; t \geqq 0]$ *is a semigroup of isometries on H.*

Proof. Suppose that $[T(t)^*; t \geqq 0]$ is isometric. Then of course $T(t)T(t)^* = I$, $t > 0$, and the operators $T(t)^*T(t)$, $t > 0$, are orthogonal projections with ranges $T(t)^*H$. Therefore from (4.2)

$$P^2 = \lim_{t \to \infty} T(t)^*T(t) = P.$$

Thus as in the proof of Theorem 4.1

$$(I - P)H = \{y \text{ in } H;\ T(t)y \to 0,\ t \to \infty\}.$$

This shows that PH is invariant for $[T(t)^*; t \geqq 0]$. From (4.6) and the fact that the adjoint semigroup is isometric, we find $PT(t) = T(t)P$, $t > 0$. Thus P commutes with $T(t)$, and therefore with $T(t)^*$ also. Hence PH is reducing for $[T(t)^*; t \geqq 0]$. Then as in the proof of Theorem 4.1, $[T(t)^*; t \geqq 0]$ is unitary on PH, while for x in $(I - P)H$, $\|T(t)^*x\| = \|x\|$ and $\|T(t)x\| \neq \|x\|$ for all $t > 0$. Hence $[T(t)^*; t \geqq 0]$ is cnu on $(I - P)H$, so that

$$H_{\text{cnu}}(T) = H_{\text{cnu}}(T^*) = (I - P)H = \{y \text{ in } H;\ T(y) \to 0,\ t \to \infty\}$$

and one half of the theorem is proved.

For the other half, it is well known that if $[T(t); t \geqq 0]$ is s-stable on $H_{\text{cnu}}(t)$, then $T(t)|_{H_{\text{cnu}}(T)}$ is unitarily equivalent to the backward translation semigroup [12]—which is the adjoint of a semigroup of isometries called the forward translation semigroup. This together with the Nagy–Foias decomposition of $[T(t); t \geqq 0]$ completes the proof of the lemma.

This lemma suggests that for this class of semigroups, if a feedback F can be found so that $A + BF$ will generate a contraction semigroup $[S(t); t \geqq 0]$ whose adjoint $[S(t)^*; t \geqq 0]$ is isometric, then as in the previous cases, the s-stabilization problem simply becomes that of "trivializing" the subspace $H_{\text{u}}(S)$.

It follows from Lemma 3.1 that $[S(t); t \geqq 0]$ is contractive if BF is dissipative, equivalently if $-(BF + F^*B^*)$ is positive. Then $[S(t)^*; t \geqq 0]$ is isometric if and only if $\operatorname{Re}[(A + BF)^*x, x] = 0$ for x in $\mathscr{D}(A^*)$ by Lemma 3.1(ii), therefore, if and only if $BF = -F^*B^*$, since $\operatorname{Re}[A^*x, x] = 0$ for x in $\mathscr{D}(A^*)$. This shows that the feedback $F = -B^*$ will *not* work in this case! However we have,

THEOREM 4.4. *Let (A, B) be a contractive system such that A^* generates an isometric semigroup $[T(t)^*; t \geqq 0]$. Then*

(i) *$A - iBB^*$ generates a contraction semigroup $[S(t); t \geqq 0]$ whose adjoint $[S(t)^*; t \geqq 0]$ is isometric;*

(ii) *if neither (A, B) nor (A^*, B) is controllable and if $H_{\text{u}}(S)$ is contained in* Ker B^*, *then (A, B) is s-stabilizable by the feedback $-iB^*$*

a) *if $H_{\text{u}}(T)$ is controllable for (A, B) or for (A^*, B), and*

b) *only if $H_{\text{u}}(T)$ is controllable for (A^*, B).*

Proof. Part (i) is trivial.

For part (ii), we note that the operator $-iBB^*$ is dissipative and skewadjoint. There from the second remark following Theorem 3.2, if $H_{\text{u}}(S)$ reduces $[T(t); t \geqq 0]$ and $[R(t) = e^{-iBB^*t}; t \geqq 0]$, then $H_{\text{u}}(S) \subseteq H_{\text{u}}(T) \cap H_{\text{u}}(R)$. The semigroup $[R(t); t \geqq 0]$ in this case is clearly unitary on all of H, so that $H_{\text{u}}(S) \subseteq H_{\text{u}}(T)$.

Now if $H_{\text{u}}(S)$ is a subspace of Ker B^*, then $H_{\text{u}}(S) \subseteq M_{\text{uc}} \cap M_{*\text{uc}}$ by Lemma 2.1(ii). But as we have seen in the proof of Theorem 3.2(ii), $H_{\text{u}}(S) \subseteq \text{Ker } B^*$ also implies that $H_{\text{u}}(S)$ is a reducing subspace of $[T(t); t \geqq 0]$. Therefore $H_{\text{u}}(S) \subseteq \text{Ker } B^* \Rightarrow H_{\text{u}}(S) \subseteq H_{\text{u}}(T) \cap M_{\text{uc}} \cap M_{*\text{uc}}$ gives sufficient proof. The necessary proof is the same as that of Theorem 4.2.

Finally we close the section with some illustrative examples. Consider the heat equation

$$\frac{\partial x}{\partial t}=\frac{\partial^2 x}{\partial^2 \zeta}, \qquad \zeta \text{ in } [0, 2\pi], \quad t>0$$

$$x(0)=x(2\pi), \qquad \acute{x}(0)=\acute{x}(2\pi).$$

Let $H=L_2[0, 2\pi]$, $A=\partial^2/\partial^2\zeta$ and

$$\mathscr{D}(A)=\{x \text{ in } H;\, x, \acute{x} \text{ absolutely continuous}, \acute{x}, \acute{\acute{x}} \text{ in } H, \text{ and } x(0)=x(2\pi), \acute{x}(0)=\acute{x}(2\pi)\}.$$

Then A is selfadjoint and dissipative on $\mathscr{D}(A)$ [6]. Hence it generates a selfadjoint contraction semigroup $[T(t);\, t\geqq 0]$ which is given by

$$T(t)x=\sum_{-\infty}^{\infty} e^{-n^2t}[x, \phi_n]\phi_n, \qquad \phi_n=e^{in\zeta}/\sqrt{2\pi}, \quad t\geqq 0 \text{ and } x \text{ in } H.$$

Let us characterize the subspaces $H_{\mathrm{u}}(T)$ and $H_{\mathrm{cnu}}(T)$. We have

$$\|T(t)x\|^2=\sum_{-\infty}^{\infty} e^{-2n^2t}|[x, \phi_n]|^2=\|T(t)^*x\|^2, \qquad t\geqq 0.$$

Therefore,

$$H_{\mathrm{u}}(T)=\{\phi_0\}$$

and

$$H_{\mathrm{cnu}}(T)=\overline{\operatorname{span}}\{\phi_n, n=\pm 1, \pm 2, \cdots\}.$$

Thus the system is s-stabilizable by a feedback $-B^*$ as soon as the state ϕ_0 is controllable. Let B be an element $b(\zeta)$ of H. Then

$$M_{\mathrm{uc}}=\{x \text{ in } H;\, [T(t)b, x]=0, t\geqq 0\}$$

and b^* is a bounded linear functional on H. Thus let $[S(t);\, t\geqq 0]$ be the semigroup generated by $A-bb^*$: $(A-bb^*)x=Ax-b[x, b]$, x in $\mathscr{D}(A)$. Then

$$H_{\mathrm{u}}(S)\subseteq\{\phi_0\}\cap\{x \text{ in } H;\, x\perp T(t)b, t\geqq 0\}.$$

Therefore if $[T(t)b, \phi_0]\neq 0$, $t\geqq 0$, then $H_{\mathrm{u}}(S)$ is trivial. But this also implies and is implied that ϕ_0 is controllable. Thus any element b of H such that $[b, \phi_0]\neq 0$ will result in an s-stabilizing feedback $-b^*$ for the system (A, b).

As a second example, consider the equation

$$\frac{\partial x}{\partial t}=-\frac{\partial x}{\partial \zeta}, \qquad \zeta \text{ in } [0, 2\pi], \quad t>0$$

$$x(0)=x(2\pi),$$

again with $H=L_2[0, 2\pi]$, $A=-\partial/\partial\zeta$ and

$$\mathscr{D}(A)=\{x \text{ in } H;\, x \text{ absolutely continuous}, \acute{x} \text{ in } H \text{ and } x(0)=x(2\pi)\}.$$

Then $A=-A^*$ and $[T(t);\, t\geqq 0]$ is therefore unitary and is given by

$$T(x)=\sum_{-\infty}^{\infty} e^{int}[x, \phi_n]\phi_n, \qquad \phi_n=\frac{e^{in\zeta}}{\sqrt{2\pi}}, \quad t\geqq 0 \text{ and } x \text{ in } H.$$

Let us now reduce $[T(t); t \geqq 0]$ to a cnu contraction semigroup. As in the previous example, let B be an element $b(\zeta)$ of H. Then for the semigroup $[S(t); t \geqq 0]$ generated by $A - bb^*$, we have

$$H_u(S) \subseteq M_{uc} = \{x \text{ in } H; [T(t)b, x] = 0, t \geqq 0\}.$$

Hence if (A, b) is controllable then $H_u(S)$ is trivial, i.e., $[S(t); t \geqq 0]$ is cnu. We have $[T(t)b, x] = \sum_{-\infty}^{\infty} e^{int}[b, \phi_n][x, \phi_n]$. Therefore, if $[b, \phi_n] \neq 0$ for any n, then $[T(t)b, x] = 0$, $t \geqq 0$ implies $[x, \phi_n] = 0$ for all n, i.e., $x = 0$, and (A, b) is indeed controllable [13].

The reduction of a unitary semigroup to a cnu contraction one, as shown in this example, is also applicable to other complicated situations. For instance, for the wave equation with homogeneous Dirichlet boundary conditions:

$$\dot{x} = Ax, \qquad A = \begin{bmatrix} 0 & I \\ \Delta & 0 \end{bmatrix}$$

where Δ is the Laplacian. If Ω is a bounded smooth domain of R^n and $H = \mathring{H}_1(\Omega) \oplus H_0(\Omega)$, then A is skewadjoint on $\mathcal{D}(A) = \{H_2(\Omega) \cap \mathring{H}_1(\Omega)\} \oplus \mathring{H}_1(\Omega)$. Hence it generates a unitary semigroup on H. This was considered by Slemrod in [2b]. According to his Theorem 3.6, if (A, B) is controllable, A has compact resolvent and generates a unitary semigroup. Then the system is s-stabilized by the feedback $-B^*$. The wave operator A does indeed have a compact resolvent and therefore the system is s-stabilizable as soon as it is controllable. It is of interest to note that controllability in this case implies that $A - BB^*$ generates a cnu contraction semigroup. Then by a theorem of Foguel [14], the cnu contraction semigroup is also weakly stable. Hence, if A is compact then this implies s-stabilizability. For weak stabilizability of contraction semigroups using Foguel's theorem we refer to [4] and particularly [5].

Finally, although we study only s-stabilizability of contractive systems as stated in the Introduction, it is of interest to note that the heat equation example is also exponentially stabilizable [3] and so is the wave equation example [5].

5. Conclusion. We have seen in this paper the important role of the Nagy–Foias canonical decomposition in the s-stabilization problem of contractive systems. Indeed an s-stable contraction semigroup is necessarily cnu.

Proposition 3.1 showed that a contraction semigroup can be reduced to a cnu one by means of the feedback $-B^*$. It was here that controllability of (A, B) and (A^*, B) came into play. This is due to the fact that the unitary subspace $H_u(T)$ is invariant for the semigroup, as well as for its adjoint. Proposition 3.2 explains why for unitary semigroups, the system has to be controllable before it can be stabilizable. This agrees with a result established by Slemrod [2b]. Indeed controllability in this case is to insure that $A - BB^*$ will generate a cnu contraction semigroup, which together with Slemrod's compactness condition, will also be s-stable.

In this paper we have avoided imposing any compactness condition, but instead rely on the strong convergence property of contraction semigroups. This allows us to s-stabilize a class of contraction semigroups which includes the normal and selfadjoint ones as special cases.

Acknowledgment. We wish to thank the referees for their helpful comments and suggestions.

REFERENCES

[1] R. DATKO, *A linear control problem in an abstract Hilbert space*, J. Differential Equations, 9 (1971), pp. 346–359.

[2a] M. SLEMROD, *The linear stabilization problem on Hilbert space*, J. Functional Analysis, 2 (1972), pp. 334–345.

[2b] ———, *A note on complete controllability and stabilizability for linear control systems in Hilbert space*, this Journal, 12 (1974), pp. 500–508.

[3] R. TRIGGIANI, *On the stabilizability problem in Banach space*, J. Math. Anal. Appl., 9 (1975), pp. 383–405.

[4] R. E. O'BRIEN, *Controllability, stabilization and mean ergodic theorem*, Report X-470-76-165 Goddard Space Flight Center, Greenbelt, MD, 1976.

[5] C. D. BENCHIMOL, *A note on weak stabilizability of contraction semigroups*, this Journal, 16 (1978), pp. 373–379.

[6] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Applications of Mathematics Vol. 3, Springer-Verlag, New York, 1976.

[7] R. S. PHILLIPS, *Dissipative operators and hyperbolic systems of partial differential equations*, Trans. Amer. Math. Soc., 90 (1959), pp. 193–254.

[8] B. SZ-NAGY AND FOIAS, *Sur les contractions de l'espace de Hilbert IV*, Acta Sci. Math. (Szeged), 21 (1960), pp. 251–259.

[9] ———, *Harmonic Analysis of Operators on Hilbert Space*, American Elsevier, New York, 1970.

[10] F. RIESZ AND B. SZ-NAGY, *Functional Analysis*, Frederick Ungar, New York, 1955.

[11] W. M. WONHAM, *On pole assignment in multi-input controllable linear systems*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 660–665.

[12] P. A. FILLMORE, *Notes on Operator Theory*, Van Nostrand Reinhold Mathematical Studies Vol. 30, Van Nostrand, New York, 1970.

[13] R. TRIGGIANI, *Extensions of rank conditions for controllability and observability to Banach spaces and unbounded operators*, this Journal, 14 (1976), pp. 313–338.

[14] S. R. FOGUEL, *Powers of a contraction in Hilbert space*, Pacific J. Math., 13 (1963), pp. 551–562.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 1, 1979

0036-0129/79/1701-0004 $01.00/0

SOLVING THE NONLINEAR COMPLEMENTARITY PROBLEM BY A HOMOTOPY METHOD*

LAYNE T. WATSON†

Abstract. Let F be a C^2 map from n-dimensional Euclidean space into itself. It is proved that, under some mild conditions on F, the complementarity problem $z \geqq 0$, $F(z) \geqq 0$, $zF(z) = 0$ can be solved by a homotopy algorithm developed by Chow, Mallet-Paret, Yorke, and Watson. The algorithm is *globally convergent* with probability one, and uses Mangasarian's nonlinear system equivalent to the complementarity problem. Convergence theorems for the algorithm simultaneously prove existence of a solution, although existence is already well known. Some computational results are included.

Introduction. Let F be a map from n-dimensional Euclidean space E^n into itself. The (nonlinear) complementarity problem is to find a vector z in E^n such that

$$z \geqq 0, \quad F(z) \geqq 0, \quad zF(z) = 0;$$

i.e., z and $F(z)$ are orthogonal and have nonnegative components. The nonlinear complementarity problem and the linear complementarity problem, where $F(z)$ is an affine map, have been studied extensively, for example [5], [7], [9], [11], [15], [16]. Typically the linear problem is solved by algebraic methods based on "complementary pivoting" [7], [11], [15], and the nonlinear problem is solved by simplicial fixed point methods [2], [3], [7], [10], [12], [13]. Murty [11] has shown that complementary pivot methods can have exponential computational complexity, and Watson [15] showed that they fail for many classes of problems. The simplicial methods of Eaves [2], [3], Merrill [10], and Saigal [12] are reasonably efficient, when implemented properly, but the supporting theory is rather complex.

Recently some completely different approaches to computing fixed points have been advanced by Chow, Mallet-Paret, Yorke [1], Kellog, Li, Yorke [6], Li [8], and Watson [17]. All these new approaches, though, require the function to be C^1 or C^2, and the standard formulation of the nonlinear complementarity problem as a fixed point problem results in a function which is not even C^1 [10], [15]. However, there is a very clever reformulation of the complementarity problem, due to Mangasarian [9], as a zero finding problem

$$G(z) = 0,$$

where $G(z)$ can be made as smooth as desired. The intent of this paper is to show that the complementarity problem can be solved by the homotopy method of Chow [1] and Watson [17], by use of the equivalent formulation $G(z) = 0$.

The equation $G(z) = 0$ is useful only in a local sense, in that if an initial approximation to the solution $\bar{z}$ is known, then locally convergent iterative techniques will compute $\bar{z}$. If no good estimate of the solution is available, then the problem $G(z) = 0$ is just a different, and equally hard, version of the complementarity problem. For the linear complementarity problem $F(z) = Mz + q$, a good initial estimate amounts to knowing which complementary cone [15] contains q, but this is tantamount to knowing the solution. The local nature of the nonlinear system $G(z) = 0$ is overcome by a homotopy method: It will be proved that, under certain mild conditions, the Chow–Yorke algorithm is *globally convergent* with probability one.

* Received by the editors August 27, 1977, and in revised form May 18, 1978.

† Department of Mathematics, Michigan State University, East Lansing, Michigan 48824.

As a historical note, homotopy-type methods were applied to the linear complementarity problem with moderate success in [15] and [16], but were not extensively developed because sufficiently powerful theoretical tools (now provided by Chow [1]) were not available then. See also Davidenko [18].

One-dimensional linear case. To motivate the general case and gain some insight into homotopy methods, consider the one-dimensional linear case

$$z \geqq 0, \quad mz+q \geqq 0, \quad z(mz+q)=0,$$

where m and q are real numbers. z solves this linear complementarity problem if and only if $G(z)=0$, where

$$G(z)=|mz+q-z|^3-(mz+q)^3-z^3.$$

This corresponds to taking $\theta(t)=t^3$ in Mangasarian's Theorem 1 [9]. $\theta(t)=t^3$ was chosen because it is the simplest function which is strictly increasing, satisfies $\theta(0)=0$, and for which $\theta(|t|)$ is a C^2 function. Homotopy theorems typically require C^2 differentiability [1], [17], and thus $G(z)$ must be C^2. Consider the following homotopy map, which has a strong supporting theory [1] and is easy to work with:

$$\varphi_a(\lambda, z)=\lambda G(z)+(1-\lambda)(z-a).$$

The idea is to track in $\lambda - z$ space the zero curve of $\varphi_a(\lambda, z)$ emanating from $(0, a)$. Hopefully this zero curve reaches a zero $\bar{z}$ of $G(z)$ (at $\lambda = 1$) after traveling a finite distance. The parametric equations $\lambda = \lambda(s)$, $z = z(s)$ of this zero curve, where s is arc length, are defined by

$$\varphi_a(\lambda(s), z(s))=0, \qquad \lambda(0)=0, \quad z(0)=a.$$

Thus

$$\frac{d\varphi_a(\lambda(s), z(s))}{ds}=D\varphi_a(\lambda, z)\begin{pmatrix}\dfrac{d\lambda}{ds}\\ \dfrac{dz}{ds}\end{pmatrix}=[G(z)-z+a, 1-\lambda+\lambda G'(z)]\begin{pmatrix}\dfrac{d\lambda}{ds}\\ \dfrac{dz}{ds}\end{pmatrix}$$

$$=0, \qquad \lambda(0)=0, \quad z(0)=a.$$

For concreteness, take $m=1$, $q=-1$, so $D\varphi_a(\lambda, z)=[1-(z-1)^3-z^3-z+a,\ 1-\lambda+\lambda(-3(z-1)^2-3z^2)]$. Now $d\lambda/ds=0$ when $D_z\varphi_a(\lambda, z)=1-\lambda+\lambda(-6z^2+6z-3)=0$, which happens at the curve $z=(1\pm\sqrt{(2-5\lambda)/(3\lambda)})/2$. By analyzing the signs of the quantities involved, it is easily verified that this curve (see Fig. 1) contains the locus of turning points for the zero curves of φ_a, and constitutes a "barrier" between $\lambda = 0$ and $\lambda = 1$. When any zero curve (except the trivial one starting at $a = 1$) hits this barrier, it turns back and slides along beside the barrier curve toward $-\infty$ (if $a < 1$) or ∞ (if $a > 1$).

Observe that if the homotopy map uses $-G(z)$ instead of $G(z)$, then $D_z\varphi_a(\lambda, z) > 0$ for *any* (λ, z), $0 \leqq \lambda \leqq 1$. Therefore $d\lambda/ds$ is never zero, which means that the zero curve does not turn back for any a. A straightforward analysis of the signs of the various quantities shows that for *every* a the zero curve of φ_a reaches the solution $\bar{z}=1$.

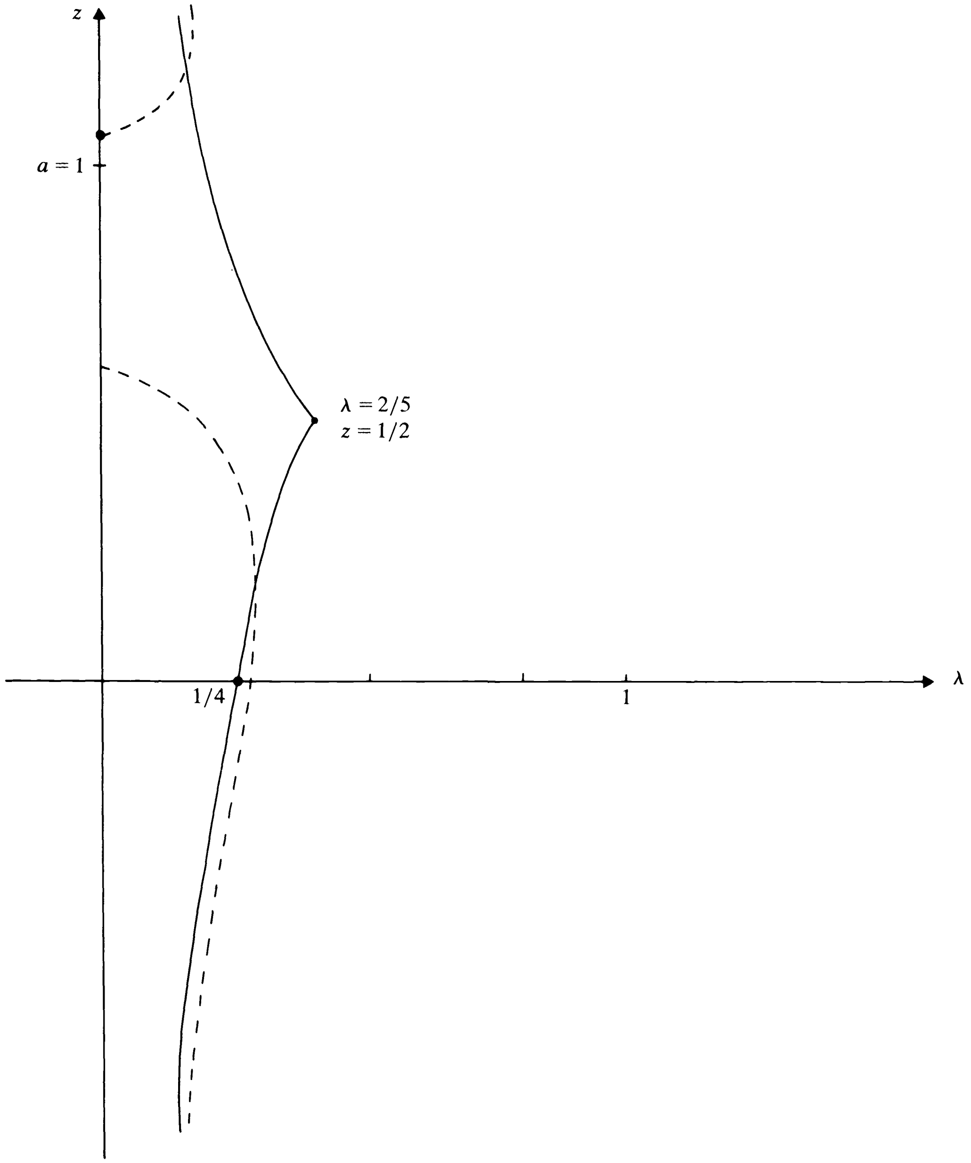

FIG. 1. *Barrier curve (solid line) and zero curves (dashed lines) of* $\varphi_a(\lambda, z)$.

The general nonlinear case. Let $F: E^n \to E^n$ be a C^2 map. Keep in mind that z solves the complementarity problem

$$z \geqq 0, \quad F(z) \geqq 0, \quad zF(z) = 0 \tag{1}$$

if and only if

$$G(z) = 0, \tag{2}$$

where

$$G_i(z) = |F_i(z) - z_i|^3 - (F_i(z))^3 - z_i^3.$$

This corresponds to taking $\theta(t)=t^3$ in Mangasarian's theorem 1 [9]. Let

$$H(z)=-G(z) \tag{3}$$

and

$$\rho_a(\lambda, z)=\lambda H(z)+(1-\lambda)(z-a). \tag{4}$$

The use of $H(z)$ instead of $G(z)$ was motivated by the one-dimensional example. For completeness the following very important result, due to Chow et al. [1], is stated here:

LEMMA 1. *Let $f: E^n \to E^n$ be a C^2 map such that $zf(z)\geqq 0$ on some sphere $\|z\|=r$. Then $f(z)$ has a zero in the ball $\|z\|\leqq r$, and for almost all a in the interior of this ball there is a zero curve of the homotopy map*

$$\Psi_a(\lambda, x)=\lambda f(x)+(1-\lambda)(x-a)$$

leading from $(0, a)$ to a zero of $f(x)$. Furthermore, rank *$D\Psi_a(\lambda, x)=n$ along this zero curve.*

THEOREM 1. *Let $F(z)=Mz+q$, $M=I$. Then for almost all $a\in E^n$, there is a zero curve of $\rho_a(\lambda, z)$ joining $(0, a)$ to $(1, \bar{z})$, where $\bar{z}$ solves the complementarity problem. $D\rho_a(\lambda, z)$ has full rank along this zero curve.*

Proof. This theorem is a corollary of some later theorems in this paper, but it also illustrates an application of Lemma 1.

$$H_i(z)=-|q_i|^3+(z_i+q_i)^3+z_i^3.$$

Now $zH(z)=2\sum_{j=1}^n z_j^4+\text{lower order terms}\geqq 0$ for $\|z\|=r$ sufficiently large. Therefore by Lemma 1 the result follows. Q.E.D.

LEMMA 2. *Let the map $\rho: E^n\times[0, 1)\times E^n \to E^n$ be defined by*

$$\rho(a, \lambda, z)=\lambda H(z)+(1-\lambda)(z-a).$$

Then ρ is transversal to zero.

LEMMA 3. *For almost all $a\in E^n$, the map ρ_a: $[0, 1)\times E^n \to E^n$ defined by*

$$\rho_a(\lambda, z)=\lambda H(z)+(1-\lambda)(z-a)$$

is transversal to zero (i.e., for almost all a the Jacobian matrix $D\rho_a$ has full rank on $\rho_a^{-1}(0)$).

A proof and discussion of Lemmas 2 and 3 can be found in [1]. Lemma 3, known as a "parameterized Sard's theorem", is the theoretical foundation of the homotopy methods developed in this paper. The application of Lemma 3 to computing fixed points of C^2 maps was given in detail in [1] and [17]. Note that $\rho_a(\lambda, z)$ and $H(z)$ are C^2 maps since $F(z)$ is C^2.

LEMMA 4. *Let the Jacobian matrix $DH(z)$ be nonsingular at every zero of $H(z)$. Then for almost all $a\in E^n$, there exists a zero curve γ of $\rho_a(\lambda, z)$ emanating from $(0, a)$ along which $D\rho_a(\lambda, z)$ has full rank. γ either has finite arc length and reaches a zero of $H(z)$ (at $\lambda=1$) or wanders off to infinity.*

Proof. The existence of γ and full rank of $D\rho_a(\lambda, z)$ along γ are just a restatement of Lemma 3. Suppose γ remains bounded. Extend ρ_a and γ to $[0, 1]\times E^n$ in the obvious fashion. Let $(\bar{\lambda}, \bar{z})$ be any point on γ in $[0, 1]\times E^n$. Then

$$D\rho_a(\bar{\lambda}, \bar{z}) = (1-\bar{\lambda})I + \bar{\lambda}Dh(\bar{z})$$

has full rank, and by the implicit function theorem, γ has finite arc length in a neighborhood of $(\bar{\lambda}, \bar{z})$. Therefore by compactness γ has finite arc length. The implicit function theorem also proves that γ cannot intersect itself and must reach $\lambda = 1$. Q.E.D.

Note that γ is a C^1 curve.

LEMMA 5. *Under the hypotheses of Lemma* 4, *let* $a > 0$ *be such that the conclusions of Lemma* 4 *hold. If* $(\bar{\lambda}, \bar{z})$ *is on the zero curve* γ *of* $\rho_a(\lambda, z)$ *emanating from* $(0, a)$, *then* $\bar{z} > 0$.

Proof. The ith component of $\rho_a(\lambda, z)$ is given by

$$\lambda(-|F_i(z)-z_i|^3 + (F_i(z))^3 + z_i^3) + (1-\lambda)(z_i - a_i).$$

Suppose that $\bar{z}_i \leqq 0$. Then $-|F_i(\bar{z})-\bar{z}_i|^3 + (F_i(\bar{z}))^3 + \bar{z}_i^3 \leqq 0$ and $\bar{z}_i - a_i < 0$. Since $0 \leqq \bar{\lambda} < 1$, the ith component of $\rho_a(\bar{\lambda}, \bar{z})$ is negative, which contradicts $\rho_a(\bar{\lambda}, \bar{z}) = 0$. Therefore $\bar{z} > 0$. Q.E.D.

THEOREM 2. *Let the Jacobian matrix* $DH(z)$ *be nonsingular at every zero of* $H(z)$. *Suppose that there exists* $r > 0$ *such that* $z > 0$ *and* $z_k = \|z\|_\infty \geqq r$ *imply* $F_k(z) > 0$. *Then for almost all* $a > 0$ *there exists a zero curve* γ *of* $\rho_a(\lambda, z)$, *along which* $D\rho_a(\lambda, z)$ *has full rank, having finite arc length and connecting* $(0, a)$ *to* $(1, \bar{z})$, *where* $\bar{z}$ *is a zero of* $H(z)$.

Proof. The existence of the zero curve γ along which rank $D\rho_a(\lambda, z) = n$ follows from Lemma 4, and if it can be shown that γ remains bounded, then the rest of the theorem also follows from Lemma 4. By Lemma 5, γ lies in $K = [0, 1)\times\{z \in E^n | z > 0\}$. There is no harm in assuming $r > \|a\|_\infty$. Let $(\bar{\lambda}, \bar{z}) \in K$ be any point with $\bar{z}_k = \|\bar{z}\|_\infty \geqq r$. Then $\bar{z}_k - a_k > 0$ and $-|F_k(\bar{z}) - \bar{z}_k|^3 + (F_k(\bar{z}))^3 + \bar{z}_k^3 > 0$ since $\bar{z}_k \geqq r > \|a\|_\infty \geqq a_k$ and $F_k(\bar{z}) > 0$ by hypothesis. Therefore

$$\bar{\lambda}(-|F_k(\bar{z})-\bar{z}_k|^3 + (F_k(\bar{z}))^3 + \bar{z}_k^3) + (1-\bar{\lambda})(\bar{z}_k - a_k) > 0,$$

which means that $\rho_a(\lambda, z) \neq 0$ for $0 \leqq \lambda < 1$ and $\|z\|_\infty \geqq r$. Hence γ is contained in the box

$$[0, 1]\times\{z | z \geqq 0, \|z\|_\infty \leqq r\},$$

and the theorem follows from Lemma 4. Q.E.D.

Note that Theorem 2 proves the *existence* of a solution to the complementarity problem (1) under certain conditions on F, besides providing an algorithm for *computing* the solution. Before applying Theorem 2 to the linear case $F(z) = Mz + q$, some definitions will be needed. Let M be a real $n \times n$ matrix and q a real n-vector. M is *strictly row diagonally dominant* if $|M_{ii}| > \sum_{j\neq i} |M_{ij}|$ for $i = 1, \cdots, n$. M is *positive definite* if $x^tMx > 0$ for all $x \neq 0$. M is called *nondegenerate* if all its principal minors are nonzero, and a *P-matrix* if all its principal minors are positive. M is *nonnegative* if each element of M is nonnegative, and strictly *copositive* if $x^tMx > 0$ for all $x \geqq 0$, $x \neq 0$. q is *nondegenerate* with respect to M if q is not a linear combination of any $n-1$ columns of $(I, -M)$. M is *strictly semimonotone* if for each nonzero $x \geqq 0$, there exists an index k such that $x_k(Mx)_k > 0$.

COROLLARY 1. *The conclusion of Theorem* 2 *holds for the linear case* $F(z) = Mz + q$, *where* M *is strictly row diagonally dominant with positive diagonal elements, and* q *is nondegenerate with respect to* M.

Proof. Assuming that q is nondegenerate means that any solution $\tilde{z}$ to the linear complementarity problem

$$z \geqq 0, \quad Mz+q \geqq 0, \quad z(Mz+q)=0$$

satisfies

$$\tilde{z}+F(\tilde{z})>0.$$

Furthermore $DF(\tilde{z})=M$ is a P-matrix [15], hence nondegenerate. Therefore by Corollary 1 of Theorem 1 in Mangasarian [9], $DH(z)$ is nonsingular at every zero of $H(z)$. (This can also be easily proved by explicitly writing out $DH(z)$.)

Since M is strictly row diagonally dominant, it is possible to choose $r>0$ such that

$$M_{ii}-\sum_{j\neq i}|M_{ij}|+\frac{q_i}{r}>0, \qquad i=1,\cdots,n.$$

Then for $z_k=\|z\|_\infty \geqq r$, $F_k(z)=(Mz+q)_k=z_k(M_{kk}+\sum_{j\neq k}M_{kj}z_j/z_k+q_k/z_k) \geqq z_k(M_{kk}-\sum_{j\neq k}|M_{kj}|+q_z/z_k)>0$. Q.E.D.

COROLLARY 2. *The conclusion of Theorem 2 holds for the linear case $F(z)=Mz+q$, where M is a nondegenerate nonnegative matrix with positive diagonal elements, and q is nondegenerate with respect to M.*

Proof. The nonsingularity of $DH(z)$ follows as in Corollary 1. Choose r such that $r>|q_i|/M_{ii}$, $i=1,\cdots,n$. Then for $z>0$ and $z_k=\|z\|_\infty \geqq r$, $F_k(z)=(Mz+q)_k=M_{kk}z_k+\sum_{j\neq k}M_{kj}z_j+q_k \geqq M_{kk}r+q_k>0$. Q.E.D.

For technical reasons, it was convenient earlier to define ρ_a on $[0, 1)\times E^n$, leaving out $\lambda=1$. For Lemmas 6, 7, 8 and Theorem 3, assume that $\rho_a(\lambda, z)$ given by (4) is defined on $[0, 1]\times E^n$.

LEMMA 6. *For $a \geqq 0$, any zero of $\rho_a(\lambda, z)$ satisfies $z \geqq 0$.*

Proof. Suppose $z_k<0$. Then the kth component of $\rho_a(\lambda, z)$ satisfies $\lambda(-|F_k(z)-z_k|^3+(F_k(z))^3+z_k^3)+(1-\lambda)(z_k-a_k)<0$, and $\rho_a(\lambda, z)\neq 0$. Hence $\rho_a(\lambda, z)=0$ implies $z \geqq 0$. Q.E.D.

LEMMA 7. *Suppose there exists an $r>0$ such that $z \geqq 0$ and $\|z\|_\infty \geqq r$ imply $z_kF_k(z)>0$ for some index k. Then the set of zeros of $\rho_0(\lambda, z)$ is contained in*

$$[0, 1]\times\{z \mid z \geqq 0, \|z\|_\infty<r\},$$

and hence is bounded.

Proof. Let $\rho_0(\tilde{\lambda}, \tilde{z})=0$. By Lemma 6, $\tilde{z} \geqq 0$. If $\|\tilde{z}\|_\infty \geqq r$, then $\tilde{z}_k>0$ and $F_k(\tilde{z})>0$ for some k. This implies that $(\rho_0(\tilde{\lambda}, \tilde{z}))_k=\tilde{\lambda}(-|F_k(\tilde{z})-\tilde{z}_k|^3+(F_k(\tilde{z}))^3+\tilde{z}_k^3)+(1-\tilde{\lambda})\tilde{z}_k>0$, a contradiction. Therefore $\|\tilde{z}\|_\infty<r$ and the result follows. Q.E.D.

LEMMA 8. *Under the hypothesis of Lemma 7, there exists $\delta>0$ such that $a \geqq 0$ and $\|a\|_\infty<\delta$ imply $\rho_a(\lambda, z)\neq 0$ for $0 \leqq \lambda \leqq 1$, $z \geqq 0$, $\|z\|_\infty=r$.*

Proof. $\|\rho_0(\lambda, z)\|$ is a continuous function on the compact set $K=[0, 1]\times\{z \mid z \geqq 0, \|z\|_\infty=r\}$, and therefore has a minimum value on K. By Lemma 7, $\min_K \|\rho_0(\lambda, z)\|=\alpha>0$. Now $\Psi(a)=\min_K \|\rho_a(\lambda, z)\|$ is a continuous function of a, $\Psi(0)\neq 0$, and therefore $\Psi(a)\neq 0$ in some neighborhood of 0, say $\|a\|_\infty<\delta$. The lemma now follows since $\Psi(a)\neq 0$ implies $\rho_a(\lambda, z)\neq 0$ on K. Q.E.D.

THEOREM 3. *Let the Jacobian matrix $DH(z)$ be nonsingular at every zero of $H(z)$. Suppose there exists an $r>0$ such that $z \geqq 0$ and $\|z\|_\infty \geqq r$ imply $z_kF_k(z)>0$ for some index k. Then there exists $\delta>0$ such that for almost all $a \geqq 0$ with $\|a\|_\infty<\delta$ there is a zero curve γ of $\rho_a(\lambda, z)$, along which $D\rho_a(\lambda, z)$ has full rank, having finite arc length and connecting $(0, a)$ to $(1, \bar{z})$, where $\bar{z}$ is a zero of $H(z)$.*

Proof. The theorem will follow from Lemma 4 in the same fashion as Theorem 2, if the zero curve γ remains bounded. Choose $\delta>0$ according to Lemma 8 (and assume $\delta<r$). Then by Lemma 8, $\rho_a(\lambda, z)\neq 0$ on the surface of the cube

$$Q=[0,1]\times\{z \mid z\geqq 0, \|z\|_\infty\leqq r\}$$

corresponding to $\|z\|_\infty=r$, and therefore by Lemma 6, γ is contained in Q. Since γ is bounded, the result follows. Q.E.D.

Remark 1. Note that Theorem 3 considerably generalizes Theorem 2, but a price is paid, namely the permissable starting points are restricted. As explained later, the restriction of a in Theorem 3 has little, if any, significance for practical numerical computations.

Remark 2. Theorems 2 and 3 could be generalized by removing the assumption on the Jacobian matrix $DH(z)$. Without this assumption, Theorems 2 and 3 have exactly the same conclusions as before, except that the arc length of γ may not be finite. Thus if one is only interested in the *existence* of a zero, the Jacobian assumption is superfluous. However, for practical numerical computations, it is desirable for $DH(z)$, or more generally $D\rho_a(\lambda, z)$, to have full rank at the zeros of $H(z)$.

Remark 3. Theorem 3 could be further generalized by replacing the condition $z_kF_k(z)>0$ by $z_k>0$ and $F_k(z)\geqq 0$, and assuming the solution set of (1) is discrete. However, the intent here is *not* to present the most general homotopy theorem possible, but rather to merely justify the application of homotopy methods to the complementarity problem (1).

COROLLARY 1. *The conclusion of Theorem* 3 *holds for the linear case* $F(z)=Mz+q$, *where* q *is nondegenerate with respect to* M, *and* M *is any one of the following*:

(A) *positive definite,*

(B) *a P-matrix,*

(C) *nondegenerate strictly copositive,*

(D) *nondegenerate strictly semimonotone.*

Proof. The nonsingularity of the Jacobian matrix $DH(z)$ at zeros of $H(z)$ follows from the nondegeneracy of M and q exactly as in the proof of Corollary 1 to Theorem 2. A positive definite matrix is a P-matrix, and a P-matrix is strictly semimonotone by the sign-reversal property of P-matrices [4]. Also a strictly copositive matrix is clearly strictly semimonotone. Therefore it is sufficient to prove just case (D).

For nonzero $x\geqq 0$ define

$$\Psi(x)=\max_{x_i>0}(Mx)_i.$$

$\Psi(x)$ is continuous and satifies $\Psi(\alpha x)=\alpha\Psi(x)$ for real $\alpha>0$. Let $\mu(r)=\min\Psi(x)$ over the compact set $\{x \mid x\geqq 0, \|x\|_\infty=r\}$, $r>0$. Since M is strictly semimonotone, $\Psi(x)>0$ and $\mu(r)>0$. $\mu(r)$ also has the property that $\mu(\alpha r)=\alpha\mu(r)$ for real $\alpha>0$. Therefore $\mu(r)>\|q\|_\infty$ for $r\geqq\|q\|_\infty/\mu(1)+1=\tilde{r}$. It now follows that for $z\geqq 0$, $\|z\|_\infty\geqq\tilde{r}$, there is an index k such that $z_k(Mz+q)_k>0$ since $(Mz)_k=\Psi(z)\geqq\mu(\|z\|_\infty)\geqq\mu(\tilde{r})>\|q\|_\infty\geqq|q_k|$. Q.E.D.

Remark 4. If Theorem 3 were generalized as indicated in Remarks 2 and 3, Corollary 1 would generalize to include semimonotone matrices M.

Remark 5. The question of the arc length of γ will be considered in a future paper.

The previous theorems seem to flow so easily that one might think almost any homotopy method would work. Actually it is quite tricky to get a homotopy method to work. It is easy to get a zero curve leading to the solution, but *not* so easy to get a zero curve which can be tracked. Some examples for the linear case $F(z)=Mz+q$, M a P-matrix, will be given. Let

$$G(\lambda, w, z)=\begin{pmatrix} |w_1-z_1|^3-w_1^3-z_1^3 \\ \vdots \\ |w_n-z_n|^3-w_n^3-z_n^3 \\ w-[(1-\lambda)I+\lambda M]z-q \end{pmatrix}.$$

Since M is a P-matrix, $(1-\lambda)I+\lambda M$ is also a P-matrix for $0\leqq\lambda\leqq 1$. The linear complementarity problem has a unique solution for a P-matrix, and thus $G(\lambda, w, z)$ has a *unique* zero for each λ, $0\leqq\lambda\leqq 1$. Hence there is a zero curve γ of $G(\lambda, w, z)$ leading from $(0, q^+, q^-)$ to the desired solution. Unfortunately DG will have less than full rank at many points along γ, and thus γ cannot be easily tracked. The homotopy defined by

$$H_i(\lambda, z)=|[(1-\lambda)I+\lambda M]_{i\cdot}z+q_i-z_i|^3-([(1-\lambda)I+\lambda M]_{i\cdot}z+q_i)^3-z_i^3$$

has essentially the same characteristics as the $G(\lambda, w, z)$ above.

Computational results. The power of Theorems 2 and 3 is that they simultaneously prove the existence of a solution to the complementarity problem (1) *and* provide an algorithm for computing a solution to (1). The algorithm is beautifully simple: just follow the zero curve of $\rho_a(\lambda, z)$ emanating from $(0, a)$ until a zero of $H(z)$ is reached at $\lambda=1$. This algorithm is guaranteed globally convergent with probability one, in the sense that it works for almost every starting point a. In practice one chooses $a=0$ and starts computing. Due to roundoff error, what is actually computed are points on zero curves of *nearby* homotopy maps $\rho_c(\lambda, z)$, $c\neq 0$. Now every such computed point could correspond to an inadmissable starting point c (a mathematical possibility, since the floating point numbers have measure zero), but this

TABLE 1

q	Arc length	Number of Jacobian evaluations	Execution time
$q^{(1)}$	4.35	231	10.6
$q^{(2)}$	9.21	224	10.3
$q^{(3)}$	14.02	218	9.8
$q^{(4)}$	17.62	349	15.9
$q^{(5)}$	19.43	234	10.8
$q^{(6)}$	19.43	234	10.9
$q^{(7)}$	17.62	349	16.1
$q^{(8)}$	14.02	218	9.9
$q^{(9)}$	9.21	224	10.2
$q^{(10)}$	4.35	231	10.4

is hardly likely. The point is that floating point arithmetic blurs the mathematical distinction between admissable and inadmissable starting points, and in practice the phrase "almost all a" has no significance. Similarly the requirement that $\|a\| < \delta$ is not a serious restriction, since a computer program would start with $a = 0$ anyway.

The numerical algorithm for following the zero curve of $\rho_a(\lambda, z)$ emanating from $(0, a)$ will be sketched only briefly here, since it is presented in detail in [17]. The zero curve is parameterized by arc length, and coincides with the trajectory of the initial value problem

$$\begin{gathered} \frac{d}{ds}\rho_a(\lambda(s), z(s)) = 0, \qquad \lambda(0) = 0, \quad z(0) = a, \\ \left(\frac{d\lambda}{ds}\right)^2 + \left\|\frac{dz}{ds}\right\|^2 = 1. \end{gathered} \tag{5}$$

The crucial property of the zero curve is that rank $D\rho_a(\lambda, z) = n$ along it, which implies that (5) has a unique solution. Thus the zero curve is tracked by solving (5), which can be done accurately and efficiently [14]. Note that λ is *not* a parameter, and hence the zero curve may "turn back" with no adverse effect.

The numerical results given here were obtained using a version of the algorithm in [17], modified to compute zeros of (3) via the homotopy (4). Execution times are in seconds on a CDC 6500, and all answers were obtained accurate to 8 decimal places.

Example 1. $F(z) = Az + q$, where A is a 10×10 symmetric, positive definite matrix given by $A_{ii} = 6$, $A_{ij} = -4$ for $|i-j| = 1$, $A_{ij} = 1$ for $|i-j| = 2$, and $A_{ij} = 0$ otherwise. Let $q^{(i)}$ be the vector with -1 in the ith coordinate and zeros elsewhere. The rather long arc lengths in Table 1 are because the solution vectors z have large components (~10). By scaling the q vectors, the arc length and execution time can be cut by a factor of 4.

Example 2. $F(z) = Az + q$, where

$$A = \begin{pmatrix} 1 & -5 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ -3 & -3 & 1 & 2 & -1 \\ -4 & -4 & 2 & 1 & 2 \\ -5 & -5 & -1 & 4 & 3 \end{pmatrix}.$$

A is neither a P-matrix nor strictly copositive, but is strictly semimonotone. With $q^{(i)}$ the same as in Example 1, the results are given in Table 2. Of course this problem

TABLE 2

q	Arc length	Number of Jacobian evaluations	Execution time
$q^{(1)}$	8.36	132	1.94
$q^{(2)}$	8.33	193	2.88
$q^{(3)}$	1.91	76	1.14
$q^{(4)}$	1.45	58	.84
$q^{(5)}$	1.35	80	1.18

could be solved much more efficiently by simply checking all $2^5 = 32$ possibilities. The point is that the computational effort for the homotopy method varies as (number of Jacobian evaluations)$\times n^3$, and this is better than 2^n for n large.

Example 3. $F(z) = Mz + q$, where

$$M = \begin{pmatrix} 0 & 0 & -1 & -1 & -1 & 1 & 1 & 0 & 1 & 1 \\ -2 & -1 & 0 & 1 & 1 & 2 & 2 & 0 & -1 & 0 \\ 1 & 0 & 1 & -2 & -1 & -1 & 0 & 2 & 0 & 0 \\ 2 & 1 & -1 & 0 & 1 & 0 & -1 & -1 & -1 & 1 \\ -2 & 0 & 1 & 1 & 0 & 2 & 2 & -1 & 1 & 0 \\ -1 & 0 & 1 & 1 & 1 & 0 & -1 & 2 & 0 & 1 \\ 0 & -1 & 1 & 0 & 2 & -1 & 0 & 0 & 0 & -1 \\ 0 & -2 & 2 & 0 & 0 & 1 & 2 & 2 & -1 & 0 \\ 0 & -1 & 0 & 2 & 2 & 1 & 1 & 1 & -1 & 0 \\ 2 & -1 & -1 & 0 & 1 & 0 & 0 & -1 & 2 & 2 \end{pmatrix}.$$

M is not strictly semimonotone, and none of the standard algebraic techniques will solve this problem (although it has been solved by the heuristic hybird n-cycle algorithm [15]). Table 3 shows the results for $q^{(i)}$.

TABLE 3

q	Arc length	Number of Jacobian evaluations	Execution time
$q^{(1)}$		did not converge	
$q^{(2)}$	4.90	402	18.38
$q^{(3)}$	6.64	620	28.65
$q^{(4)}$	5.49	264	11.96
$q^{(5)}$		did not converge	
$q^{(6)}$	6.50	241	11.17
$q^{(7)}$	11.95	303	14.04
$q^{(8)}$		did not converge	
$q^{(9)}$	9.34	628	29.86
$q^{(10)}$	2.72	181	8.56

Example 4 (nonlinear programming problem). Consider the convex programming problem

$$\min \theta(x) = \exp\left(\sum_{i=1}^{5} (x_i - i + 2)^2\right) \quad \text{subject to} \quad x \geqq 0.$$

The Kuhn–Tucker optimality conditions applied to this problem result in a complementarity problem with

$$F(x)=\nabla\theta(x)=2\exp\left(\sum_{i=1}^{5}(x_i-i+2)^2\right)\begin{pmatrix}x_1+1\\x_2\\x_3-1\\x_4-2\\x_5-3\end{pmatrix}.$$

To obtain the solution (0, 0, 1, 2, 3) the homotopy method required 377 Jacobian evaluations and 5.62 seconds of CPU time, with an arc length of 4.537.

Acknowledgment. The proof of Theorem 3 was the result of collaboration with Lee Sonneborn. The author is also indebted to S. N. Chow and K. G. Murty for encouragement and many helpful discussions.

REFERENCES

[1] S. N. Chow, J. Mallet-Paret and J. A. Yorke, *Finding zeros of maps: homotopy methods that are constructive with probability one*, Math. Comput., to appear.

[2] B. C. Eaves, *Homotopies for the computation of fixed points*, Math. Programming, 3 (1972), pp. 1–22.

[3] B. C. Eaves and R. Saigal, *Homotopies for computation of fixed points on unbounded regions*, Ibid., 3 (1972), pp. 225–237.

[4] M. Fiedler and V. Ptak, *On matrices with non-positive off-diagonal elements and positive principal minors*, Czechoslovak Math. J., 12 (1962), pp. 382–400.

[5] S. Karamardian, *The complementarity problem*, Math. Programming, 2 (1972), pp. 107–129.

[6] R. B. Kellogg, T. Y. Li and J. Yorke, *A constructive proof of the Brouwer fixed-point theorem and computational results*, SIAM J. Numer. Anal., 13 (1976), pp. 473–483.

[7] M. M. Kostreva, *Direct algorithms for complementarity problems*, Ph.D. thesis, Rensselaer Polytechnic Institute, Troy, New York, 1976.

[8] T. Y. Li, Private communication, September, 1976.

[9] O. L. Mangasarian, *Equivalence of the complementarity problem to a system of nonlinear equations*, SIAM J. Appl. Math., 31 (1976), pp. 89–92.

[10] O. Merrill, *Applications and extensions of an algorithm to compute fixed points of upper semicontinuous mappings*, Doctoral thesis, I.O.E. Dept., Univ. of Michigan, Ann Arbor, 1972.

[11] K. G. Murty, *Computational complexity of complementary pivot methods*, Math. Programming, to appear.

[12] R. Saigal, *On the convergence rate of algorithms for solving equations that are based on methods of complementary pivoting*, Math. Operations Res., 2 (1977), pp. 108–124.

[13] H. Scarf, *The approximation of fixed points of a continuous mapping*, SIAM J. Appl. Math., 15 (1967), pp. 1328–1343.

[14] L. F. Shampine and M. K. Gordon, *Computer Solution of Ordinary Differential Equations: The Initial Value Problem*, W. H. Freeman, San Francisco, 1975.

[15] L. T. Watson, *A variational approach to the linear complementarity problem*, Doctoral dissertation, Dept. of Mathematics, Univ. of Michigan, Ann Arbor, 1974.

[16] ———, *Some perturbation theorems for Q-matrices*, SIAM J. Appl. Math., 31 (1976), pp. 379–384.

[17] ———, *A globally convergent algorithm for computing fixed points of C^2 maps*, Appl. Math. Comput., to appear.

[18] D. F. Davidenko, *On the approximate solution of systems of nonlinear equations*, Ukrain. Mat. Ž., 5 (1953), pp. 196–206.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 1, 1979

0036–0129/79/1701–0005 $01.00/0

FINITE PRIMAL CUTTING PLANE ALGORITHMS FOR INTEGER PROGRAMS*

GRAHAM LINKS† AND DAVID S. RUBIN‡

Abstract. Several primal cutting plane algorithms have been proposed for the pure integer program. They all require that the entire tableau remain integral at every iteration. We show how to relax this condition and require integrality of only the right hand side. Several cuts are discussed, and sufficient conditions to guarantee finiteness are developed.

1. Introduction. We are interested in the pure integer program

$$\text{IP}\qquad \begin{aligned} \max \quad & c^{\mathrm{T}}x = z \\ \text{subject to} \quad & Ax = b \\ & x \geqq 0 \text{ and integer} \end{aligned}$$

where A is $m \times (m+n)$ and has rank m, and b, c, and x are appropriately dimensioned. We assume that the elements of A, b, and c are integers. This assumption entails no loss of generality if A, b, and c are rational or if $L = \{x \mid Ax = b, x \geqq 0\}$ is bounded; see references [12] and [14]. If the objective value of IP is bounded, there is no loss of generality in assuming that L is bounded, so we will make that assumption henceforth. Relaxing the integrality requirement on x gives us the associated linear program LP. We will add cuts to the constraints of IP, but we shall continue to refer to the resulting problem as IP, and to its linear relaxation as LP.

Primal cutting plane algorithms successively find basic feasible solutions to LP which are all-integer and hence feasible in IP. These successive solutions yield nondecreasing objective values, and optimality is reached when the current solution is optimal in LP. Earlier work on primal algorithms has been done by Ben-Israel and Charnes [4], Young [15], [16], [17], Glover [7], and Arnold and Bellmore [1], [2], [3].

If we partition A as (B, N) (rearranging columns if necessary), where B is $m \times m$ and nonsingular, and similarly partition x as $\begin{pmatrix} x_B \\ x_N \end{pmatrix}$ and c as $\begin{pmatrix} c_B \\ c_N \end{pmatrix}$, then we may represent LP in the usual tableau form:

$$\begin{array}{r|c} & -x_N \\ \hline z = c_B^{\mathrm{T}}B^{-1}b & c_B^{\mathrm{T}}B^{-1}N - c_N^{\mathrm{T}} \\ x_B = B^{-1}b & B^{-1}N \\ x_N = 0 & -I \end{array} \qquad \text{or} \qquad \begin{array}{r|ccc} & -x_{N1} & & -x_{Nn} \\ \hline z = y_{00} & y_{01} & \cdots & y_{0n} \\ x_1 = y_{10} & y_{11} & \cdots & y_{1n} \\ \vdots & \vdots & & \vdots \\ x_{m+n} = y_{m+n,0} & y_{m+n,1} & \cdots & y_{m+n,n} \end{array}$$

In the latter representation of the tableau, we keep the variables in their natural order, so the rows corresponding to basic and nonbasic variables are intermixed. We let y_j denote the vector $(y_{1j}, y_{2j}, \cdots, y_{m+n,j})^{\mathrm{T}}$ and $\bar{y}_j$ denote the vector $\begin{pmatrix} y_{0j} \\ y_j \end{pmatrix}$, for $j = 0, 1, \cdots, n$.

* Received by the editors October 18, 1977.

† College of Commerce, University of Saskatchewan, Saskatoon, Saskatchewan, Canada S7N 0W0.

‡ School of Business Administration and Curriculum in Operations Research and Systems Analysis, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27514.

If $B^{-1}b \geqq 0$ and integer, then the current basic feasible solution for LP is feasible for IP. In all of the earlier work cited above, it is also required that B be unimodular, so the entire tableau is all-integer at each iteration. We will relax this requirement, and require only that $B^{-1}b \geqq 0$ and integer; the other elements in the tableau may be noninteger. We will show that finite algorithms can be obtained for large classes of cuts if we impose a certain weak restriction on the absolute value of the determinant of B.

In § 2 we give some preliminary results. Section 3 states a general primal cutting plane algorithm, and proves that it is finite. Section 4 briefly discusses some of the cuts which can be used in the context of this general algorithm. Section 5 gives some concluding remarks.

2. Preliminary results. Our development will closely parallel that given by Garfinkel and Nemhauser [6] in their proof of the finiteness of the SPA ("simplified Primal Algorithm") of Glover [7] and Young [16]. Our results are generalizations of these earlier results, and the SPA is a special case of the algorithm we present in the next section.

Given two vectors α and β with the same number of components, let $\alpha \prec \beta$ mean that α is lexicographically smaller than β, i.e., $\alpha \neq \beta$ and for the first i such that $\alpha_i \neq \beta_i$, we have $\alpha_i < \beta_i$. We assume that the initial tableau has a special row (or "reference row") with the following two properties:

(p_1) If $\bar{y}_j \prec 0$, then $y_{sj} > 0$.

(p_2) For $j = 1, \cdots, n$ with $y_{sj} \neq 0$, let $R_j = \bar{y}_j / y_{sj}$, and let

$$R_k = \text{lex min}\,\{R_j | y_{sj} > 0\}$$

If $y_{sj} < 0$, then $R_j \prec R_k$.

Given the assumption that $\{x | Ax = b, x \geqq 0\}$ is bounded, we can initially choose $y_{sj} = 1$ for $j = 1, \cdots, n$, and let y_{s0} be a suitably large integer. We note that the columns $\bar{y}_1, \cdots, \bar{y}_n$ are linearly independent (since they contain a negative identity matrix), and hence the index k in (p_2) is unambiguously defined. Notice also that $\bar{y}_k \prec 0$ in all tableaux prior to the optimal tableau.

We move from one tableau to the next by the usual simplex step; however, we will only allow pivots to be done in column k defined in (p_2). If the pivot element is $y_{rk} \neq 0$, then the new tableau is given by

$$\bar{y}'_j = \bar{y}_j - \frac{y_{rj}}{y_{rk}} \bar{y}_k, \qquad j = 0, 1, \cdots, n; \quad j \neq k,$$

$$\bar{y}'_k = -\frac{1}{y_{rk}} \bar{y}_k.$$

LEMMA 1.1. *If* (p_1) *and* (p_2) *hold in the current tableau, then*

$$\bar{y}_j \succ \frac{y_{sj}}{y_{sk}} \bar{y}_k \quad \textit{for all } j = 1, \cdots, n; \quad j \neq k.$$

Proof. If $y_{sj} = 0$, then $\bar{y}_j \succ 0$ by (p_1). If $y_{sj} > 0$ (< 0), then $R_j \succ R_k$ ($\prec R_k$), and multiplying through by y_{sj} yields the result. Q.E.D.

LEMMA 1.2. *If* (p_1) *and* (p_2) *hold in the current nonoptimal tableau, and we pivot on any* $y_{rk} \neq 0$, *then* (p_1) *holds in the next tableau.*

Proof. For column k, $\bar{y}_k < 0$ and hence $y_{sk} > 0$ by (p_1). Thus

$$\frac{1}{y'_{sk}}\bar{y}'_k = -\frac{y_{rk}}{y_{sk}}\left(-\frac{\bar{y}_k}{y_{rk}}\right) = \frac{\bar{y}_k}{y_{sk}} < 0$$

by (p_1). Hence $y'_{sk} > 0$ if $\bar{y}'_k < 0$.

For column $j \neq k$, suppose

$$y'_{sj} = y_{sj} - \frac{y_{rj}}{y_{rk}} y_{sk} \leqq 0.$$

Then

$$\frac{y_{sj}}{y_{sk}} \leqq \frac{y_{rj}}{y_{rk}},$$

and so

$$\bar{y}'_j = \bar{y}_j - \frac{y_{rj}}{y_{rk}}\bar{y}_k \geqq \bar{y}_j - \frac{y_{sj}}{y_{sk}}\bar{y}_k > 0,$$

by Lemma 1.1. The result now follows by contraposition. Q.E.D.

The next two results are based on the following observation. If we pivot on any $y_{rk} \neq 0$, then for any $j \neq k$ we have:

$$\begin{aligned} R'_j - R_k &= \frac{1}{y'_{sj}}\left(\bar{y}_j - \frac{y_{rj}}{y_{rk}}\bar{y}_k\right) - \frac{1}{y_{sk}}\bar{y}_k \\ &= \frac{1}{y'_{sj}y_{sk}}\left(y_{sk}\bar{y}_j - \frac{y_{sk}y_{rj}}{y_{rk}}\bar{y}_k - \left(y_{sj} - \frac{y_{rj}y_{sk}}{y_{rk}}\right)\bar{y}_k\right) \qquad (1) \\ &= \frac{1}{y'_{sj}}\left(\bar{y}_j - \frac{y_{sj}}{y_{sk}}\bar{y}_k\right). \end{aligned}$$

Also, the proof of Lemma 1.2 shows that $R'_k = R_k$, so equation (1) holds for $j = k$ as well.

LEMMA 1.3. *If* (p_1) *and* (p_2) *hold in the current nonoptimal tableau, if we pivot on any* $y_{rk} \neq 0$, *and if* $y'_{sj} < 0$, *then* $R'_j \leqq R_k$, *and the inequality is strict if* $j \neq k$.

Proof. The result is immediate from Lemma 1.1 and equation (1). Q.E.D.

LEMMA 1.4. *If* (p_1) *and* (p_2) *hold in the current tableau, if we pivot on* $y_{rk} \neq 0$, *and if the next tableau is not optimal, then* $R'_{k'} \geqq R_k$ (*with the inequality strict if* $y_{rk} > 0$), *where*

$$R'_{k'} = \operatorname{lex\,min}\{R'_j \mid y'_{sj} > 0\}.$$

Proof. From Lemma 1.1 and equation (1), if $y'_{sj} > 0$, then $R'_j > R_k$. Now if $y_{rk} > 0$, then $y'_{sk} = -y_{sk}/y_{rk} < 0$ by (p_1). Thus $k' \neq k$, so $R'_{k'} > R_k$. However, if $y_{rk} < 0$, then $y'_{sk} > 0$, and so $R'_{k'} = R_k = R'_k$ by equation (1). The result is now immediate. Q.E.D.

We may put all these results together to get

THEOREM 1. *If* (p_1) *and* (p_2) *hold in the initial nonoptimal tableau and we always pivot in column* k, *then* (p_1) *and* (p_2) *hold in all subsequent nonoptimal tableaux, and furthermore, the sequence* $\{R_k\}$ *is lexicographically nondecreasing.*

Proof. All we need show is that (p_2) holds in the second tableau if it is nonoptimal, and then the result follows from Lemmas 1.2 and 1.4 by induction. Now $R'_j - R'_{k'} \leqq R'_j - R_k$ by Lemma 1.4. Thus if $y'_{sj} < 0$, it follows from Lemma 1.1 and equation (1) that $R'_j < R'_{k'}$. Q.E.D.

The Glover–Young SPA always chooses a pivot element $y_{rk} = 1$, and for it we can strengthen Theorem 1 to say that the sequence $\{R_k\}$ is lexicographically increasing. Thus allowing pivots on arbitrary nonzero elements in column k weakens the result only slightly, and it turns out that the weakened version will still support a finiteness proof. It is somewhat remarkable that this crucial result depends only on the choice of the pivot column, and has nothing whatsoever to do with the source row for the cut or type of cut which is added.

3. A general finite primal cutting plane algorithm. If the vector R_k ever satisfies $y_{0k} \geqq 0$, then optimality has been reached. In order to satisfy this condition, attention must be directed to the nature of the cut. Consider the condition

$\mathrm{C}_1(q)$: $$y_{qk} \leqq y_{q0}.$$

The SPA uses Gomory all-integer cuts [9], and can be shown finite provided that the source row of the cut is chosen in such a way as to guarantee that $\mathrm{C}_1(q)$ holds for each row q at finite intervals. There are several source row choice rules which guarantee that this condition will be met; see references [7] and [16]. The proof of finiteness of the SPA depends on the integrality of the tableau. When this integrality is relaxed, finite reoccurence of $\mathrm{C}_1(q)$ no longer suffices to guarantee finiteness.

Let δ be the absolute value of the determinant of B, and let Δ be an arbitrary positive integer. Consider the conditions:

C_2: $$\delta \leqq \Delta$$

C_3: $\bar{y}_k$ is all-integer.

THEOREM 2. *Suppose* IP *has a bounded feasible region. Consider any primal cutting plane algorithm which*

a) *always pivots in column k,*

b) *pivots on positive y_{rk} at finite intervals, and*

c) *guarantees for each row q that at finite intervals* $\mathrm{C}_1(q)$ *will hold in conjunction with either* C_2 *or* C_3.

That algorithm is finite.

Before proving the theorem, we note that in the SPA, $y_{rk} = 1$ at all iterations, so $\delta = 1$ throughout the algorithm, and C_2 and C_3 are always satisfied. Thus we see that the SPA is just one member of a large class of finite primal algorithms.

Proof. We know that

i) $\{R_k\}$ is lexicographically nondecreasing (by Theorem 1), and

ii) at finite intervals $\mathrm{C}_1(s)$ holds in conjunction with either C_2 or C_3.

Since $\{y_0\}$ is bounded, $\{y_{20}\}$ is also bounded, say by M_s. At those tableaux where ii) holds, y_{sk} is a nonnegative rational number, bounded above by M_s and having denominator no larger than Δ. Thus at those tableaux y_{sk} can assume only a finite set of values. From condition i), the sequence $\{y_{0k}/y_{sk}\} = \{R_{0k}\}$ is nondecreasing, and this same sequence is bounded above by 0 prior to optimality. Thus at tableaux where ii) holds, R_{0k} can only assume a finite number of values. Hence after a finite number of iterations the sequence $\{R_{0k}\}$ must be constant.

Suppose $\{R_{0k}\}$ has reached its limit at iteration t_0. Let M_1 be an upper bound for $\{y_{10}\}$. Consider the next nonoptimal tableau at which $\mathrm{C}_1(1)$ holds in conjunction with either C_2 or C_3; denote its elements by $\hat{y}_{ij}$. Then $\hat{y}_{1k} \leqq \hat{y}_{10}$ and $\hat{y}_{sk} > 0$, so

$$\frac{\hat{y}_{1k}}{\hat{y}_{sk}} \leqq \frac{\hat{y}_{10}}{\hat{y}_{sk}} \leqq M_1\Delta.$$

Now the sequence $\{y_{1k}/y_{sk}\} = \{R_{1k}\}$ is nondecreasing for all iterations from t_0 on, and is bounded above by $M_1\Delta$. Consider all subsequent tableaux satisfying ii). Repeating the earlier argument, we see that R_{1k} can only assume a finite number of values, and hence there is an iteration t_1 after which $\{R_{1k}\}$ is constant.

Repeating this argument for each row in turn, we see that after a finite number of nonoptimal tableaux $\{R_k\}$ becomes constant. But at this point, the tableau must be optimal, for if not, then $y_{0k} < 0$ and we would continue. However after a finite number of additional iterations, condition b) guarantees that we would pivot on a positive element. This (by Lemma 1.4) would cause R_k to increase which is impossible. Thus the algorithm is finite. Q.E.D.

We must now consider how to get condition c) of the theorem to hold. It turns out that C_2 and C_3 are relatively easy to establish, so we shall discuss them first. In the next section we discuss cuts which enable us to establish the conditions $C_1(q)$.

LEMMA 3.1. *In the current nonoptimal tableau, suppose that $\bar{y}_k$ is not all-integer. Then we may add a Gomory fractional cut* [8] *and pivot on it in column k to*

a) *reduce the determinant, and*

b) *retain the same pivot column at the next iteration.*

Proof. If $\bar{y}_k$ is not all-integer, there exists a row r

$$x_r = y_{r0} - \sum_{j=1}^{n} y_{rj}x_{Nj} = \frac{n_{r0}}{\delta} - \sum_{j=1}^{n} \frac{n_{rj}}{\delta}x_{Nj}$$

such that y_{rk} is not an integer. Here we have expressed y_{rj} as a ratio of integers n_{rj}/δ where δ is the absolute value of the current basis determinant. Let $n_{fj} = n_{rj} \pmod{\delta}$ and add the cut

$$x_f = -\frac{n_{f0}}{\delta} - \sum_{j=1}^{n}\left(-\frac{n_{fj}}{\delta}\right)x_{Nj} = y_{f0} - \sum_{j=1}^{n} y_{fj}x_{Nj} \geqq 0.$$

Notice that $-1 < y_{fk} < 0$. Hence if we pivot on y_{fk}, part a) follows because $\delta' = \delta|y_{fk}|$, and b) follows from Lemma 1.4. Q.E.D.

Thus if $C_1(q)$ holds, but neither C_2 nor C_3 holds, we can use fractional cuts to try to satisfy C_2 and/or C_3. Of course in the process y_{qk} will increase (since $y'_{qk} = -y_{qk}/y_{fk}$), y_{q0} will be unchanged (since $y_{f0} = 0$), and thus $C_1(q)$ may cease to hold. The following result shows that this difficulty can be overcome.

LEMMA 3.2. *In the current nonoptimal tableau, suppose that* $C_1(q)$ *does not hold. Then we may derive a Gomory all-integer cut and pivot on it in column k to*

a) *retain the current determinant,*

b) *change the pivot column at the next iteration, and*

c) *make $y'_{qk'} \leqq y_{qk} - 1/\delta$ and $y'_{q0} = y_{q0}$.*

Proof. Let $[\lambda]$ denote the greatest integer not exceeding λ. Since $C_1(q)$ does not hold, $0 \leqq y_{q0} < y_{qk}$ and so $[y_{q0}/y_{qk}] = 0$. Suppose we add the all-integer cut

$$x_c = 0 - \sum_{j=1}^{n} [y_{qj}/y_{qk}]x_{Nj} \geqq 0$$

and pivot in the cut row and column k. The pivot element is $+1$. Then a) follows since $\delta' = \delta \cdot 1$, and b) follows from Lemma 1.4. For c) we note that

$$y'_{qj} = y_{qj} - \left[\frac{y_{qj}}{y_{qk}}\right]y_{qk} = y_{qk}\left(\frac{y_{qj}}{y_{qk}} - \left[\frac{y_{qj}}{y_{qk}}\right]\right) \quad \text{for all } j \neq k.$$

In particular $y'_{q0} = y_{q0}$. Also, since $0 \leqq \lambda - [\lambda] < 1$ for all real numbers λ, and since $y_{qk} \geqq 0$, it follows that $y'_{qj} < y_{qk}$ for all $j \neq k$, and c) now follows from b). Q.E.D.

It follows immediately that if $0 \leqq y_{q0} < y_{qk}$ at any iteration, we can force $C_1(q)$ to hold after the addition of no more than $\delta(y_{qk} - y_{q0})$ all-integer cuts. Thus the last two lemmas tell us that we can always get condition c) of Theorem 2 to hold by the application of a finite number of fractional and/or all-integer cuts at appropriate intervals. In the meantime, we are free to use any valid cuts whatsoever which will retain the primal feasibility of the tableau. Of course it is desirable to use cuts which help achieve the conditions $C_1(q)$.

4. Cuts for the general algorithm. The current extreme point is $x = y_0$, and the line $\{x = y_0 - \lambda y_k | \lambda \geqq 0\}$ contains an edge of L, the *LP* feasible region. The last lattice point along that edge can be found by solving the "edge congruency problem"

$$\lambda_k = \max \lambda$$

$$\text{subject to} \quad x = y_0 - \lambda y_k \equiv 0 \ (\text{mod } 1)$$

$$0 \leqq \lambda \leqq \theta_k = \min \left\{ \frac{y_{i0}}{y_{ik}} \middle| y_{ik} > 0 \right\}.$$

Since the primal simplex algorithm moves from one vertex to an adjacent vertex, it is desirable to add cuts that go through the point $y_0 - \lambda_k y_k$. The edge congruency problem is simple to solve, and the interested reader can find the details in reference [13].

In his dissertation [13], Links discusses a class of cuts that he calls "facial". These are determined by pivoting on some y_{rk} to an IP infeasible point, looking at the corner polyhedron [10] at that point, and determining a facet of that polyhedron which goes through the point $y_0 - \lambda_k y_k$. He proves the following result:

THEOREM 3. *Determine the value*

$$\mu_k = \min \lambda$$

$$\textit{subject to} \quad x = y_0 - \lambda y_k \equiv 0 \ (\textit{mod } 1)$$

$$\lambda \geqq 1.$$

Consider any row with $y_{qk} > 0$ and $y_{q0}/y_{qk} < \mu_k$. If we add the facial cut through $y_0 - \lambda_k y_k$ obtained from the corner polyhedron where x_{Nk} replaces x_q in the basis, and pivot on that cut in column k, then $y'_{q0} \geqq y'_{qk'}$. That is to say, if $C_1(q)$ does not hold in the current tableau, it is possible to add a single facial cut and have $C_1(q)$ hold in the next tableau.

The proof of this result is quite lengthy and involves a careful examination of facial cuts. We refer the interested reader to reference [13]. A great deal of work is involved in determining facial cuts, so we would now like to look at other cuts through the point $y_0 - \lambda_k y_k$. These other cuts are related to Dantzig cuts [5].

Again consider any row with $y_{qk} > 0$. If $y_{q0}/y_{qk} = \lambda_k$, then a pivot on y_{qk} leads to a lattice point. If $y_{q0}/y_{qk} > \lambda_k$ then pivoting on y_{qk} leads to a nonlattice point; in this case the Dantzig cut

$$\sum_{\substack{j=1 \\ j \neq k}}^{n} x_{Nj} + x_q \geqq 1$$

is a valid cut. Consider the strengthened Dantzig cut

$$\sum_{\substack{j=1 \\ j \neq k}}^{n} x_{Nj} + \left(\frac{1}{y_{q0} - \lambda_k y_{qk}} \right) x_q \geqq 1.$$

The only lattice points it excludes besides those excluded by the Dantzig cut are elements of $\{x \geqq 0 | 1 < x_q < y_{q0} - \lambda_k y_{qk},\ x_{Nj} = 0 \text{ for } j \neq k\}$, but this set contains no feasible lattice points by virtue of the definition of λ_k. Hence the strengthened cut is valid. If we clear fractions, we may write it as

$$\sum_{\substack{j=1\\ j\neq k}}^{n} (y_{q0} - \lambda_k y_{qk}) x_{Nj} + x_q \geqq y_{q0} - \lambda_k y_{qk}.$$

Recalling that $x_q = y_{q0} - \sum_{j=1}^{n} y_{qj} x_{Nj}$, we may express the strengthened cut in terms of the current nonbasic variables as

$$x_c = \lambda_k y_{qk} - \sum_{\substack{j=1\\ j\neq k}}^{n} (y_{qj} - y_{q0} + \lambda_k y_{qk}) x_{Nj} - y_{qk} x_{Nk} \geqq 0,$$

and clearly this cut goes through the point $y_0 - \lambda_k y_k$.

Now we may distinguish between transition cycles, in which the objective value increases, and stationary cycles, in which the objective value is unchanged. In a stationary cycle, there is at least one row q with $y_{q0} < y_{qk}$, i.e., for which $C_1(q)$ does not hold. For this row $\lambda_k = 0$, and it yields the strengthened Dantzig cut

$$x_c = 0 - \sum_{\substack{j=1\\ j\neq k}}^{n} (y_{qj} - y_{q0}) x_{Nj} - y_{qk} x_{Nk} \geqq 0.$$

If we add this cut and pivot in the cut row on column k, then the pivot element is $y_{qk} > 0$, so $k' \neq k$ by Lemma 1.4. Furthermore, $y'_{q0} = y_{q0}$ and $y'_{qj} = y_{q0}$ for all $j \neq k$. Thus $C_1(q)$ holds in the next tableau.

It is well-known that ordinary Dantzig cuts do not lead to a finite dual cutting plane algorithm [11], but the above result suggests that these strengthened Dantzig cuts might yield a finite primal algorithm. Unfortunately, note that the pivot element y_{qk} can be large and so these cuts can cause the determinant to grow. In fact, Links has shown [13] that these cuts do not yield a finite algorithm when they are the only cuts which are used. However, the above result does show that they have the desirable property of forcing the condition $C_1(q)$ to hold after a single iteration.

Let us again consider a row with $y_{qk} > 0$ and $\lambda_k < y_{q0}/y_{qk} < \mu_k$. Note that if we pivot on y_{qk},

$$y'_0 = y_0 - \frac{y_{q0}}{y_{qk}} y_k$$

is not a lattice point, and so the vector $y'_k = -y_k/y_{qk}$ is not all integer. Charnes and Ben-Israel [4] have shown how to strengthen the ordinary Dantzig cut to

$$\sum_{j\in\hat{J}} x_{Nj} + x_q \geqq 1 \quad \text{where } \hat{J} = \{j \in \{1, \cdots, n\} \backslash \{k\} | y'_j \not\equiv 0 \pmod 1\}.$$

Analogous to our strengthening of the ordinary Dantzig cut, we may strengthen this cut to

$$\sum_{j\in\hat{J}} (y_{q0} - \lambda_k y_{qk}) x_{Nj} + x_q \geqq y_{q0} - \lambda_k y_{qk}$$

which may be expressed in terms of the current nonbasic variables as

$$x_c = \lambda_k y_{qk} - \sum_{j\in\hat{J}} (y_{qj} - y_{q0} + \lambda_k y_{qk}) x_{Nj} - \sum_{\substack{j=1\\ j\notin\hat{J}\\ j\neq k}}^{n} y_{qj} x_{Nj} - y_{qk} x_{Nk} \geqq 0.$$

If we choose a row for which $C_1(q)$ does not hold, and generate this strengthened cut and pivot on it, then $k' \neq k$, $y'_{q0} = y_{q0}$, $y'_{qj} = y_{q0}$ for all $j \in \hat{J}$, and $y'_{qj} = 0$ for all $j \notin \hat{J} \cup \{k\}$. Thus $C_1(q)$ holds in the next tableau.

5. Conclusions. We have shown how to relax the integrality requirement on the entire tableau and still get a finite primal cutting plane algorithm. The finiteness proof essentially permits the use of any cuts whatsoever that guarantee that a certain set of conditions will hold at finite intervals. Subsequent results showed how to use Gomory's fractional and all integer cuts to establish these conditions, and we also discussed several new classes of cuts which forced a subset of these sufficient conditions to hold.

We have done no computational testing of algorithms based on these results. There are several interesting questions which should be investigated, such as:

1. What effect does the arbitrary parameter Δ have on computation times?
2. Does it make sense to bring the determinant back below Δ whenever it gets higher, or would it be better to restore the condition at fixed intervals?
3. Are the relatively weak Dantzig-like cuts superior to the stronger facial cuts because of their relative ease of computation?

We hope to address these questions in future work.

6. Other related work. In [14a], Salkin et al. also discuss relaxing the condition that the entire tableau be all-integer. They show that the Glover and Young algorithms are also finite even if the data are rational and not necessarily integral. However, their discussion applies only to a problem of the form

$$\text{IP}' \qquad \begin{aligned} \max \quad & c^{\mathrm{T}} x = z \\ \text{subject to} \quad & Ax \leqq b \\ & x \geqq 0 \text{ and integer,} \end{aligned}$$

where the slack variables which constitute the initial basis need not be integer valued. It is not valid for the general equality constrained problem which we have discussed.

REFERENCES

[1] L. R. ARNOLD AND M. BELLMORE, *Iteration skipping in primal integer programming*, Operations Res., 22 (1974), pp. 129–136.

[2] ———, *A generated cut for primal integer programming*, Ibid., 22 (1974), pp. 137–143.

[3] ———, *A bounding minimization problem for primal integer programming*, Ibid., 22 (1974), pp. 383-392.

[4] A. BEN-ISRAEL AND A. CHARNES, *On some problems of Diophantine programming*, Cahiers Centre d'Études Recherche Opér., 4 (1962), pp. 215–280.

[5] G. B. DANTZIG, *Note on solving linear programs in integers*, Naval Res. Logist. Quart., 6 (1959), pp. 75–76.

[6] R. S. GARFINKEL AND G. L. NEMHAUSER, *Integer programming*, John Wiley, New York, 1972.

[7] F. GLOVER, *A new foundation for a simplified primal integer programming algorithm*, Operations Res., 16 (1968), pp. 727–740.

[8] R. E. GOMORY, *An algorithm for integer solutions to linear programs*, Recent Advances in Mathematical Programming, R. L. Graves and P. Wolfe, eds., McGraw-Hill, New York, 1963.

[9] ———, *All-integer integer programming algorithm*, Industrial Scheduling, J. F. Muth and G. L. Thompson, eds., Prentice-Hall, Englewood Cliffs, NJ, 1963.

[10] ———, *Some polyhedra related to combinatorial problems*, J. Linear Algebra Appl., 2 (1969), pp. 451–558.

[11] R. E. GOMORY AND A. J. HOFFMAN, *On the convergence of an integer programming process*, Naval Res. Logist. Quart., 10 (1963), pp. 121–123.

[12] F. J. GOULD AND D. S. RUBIN, *Rationalizing discrete programs*, Operations Res., 21 (1973), pp. 343–345.

[13] G. LINKS, *A finite primal integer programming algorithm using facial cuts*, unpublished Ph.D. dissertation, Curriculum in Operations Research and Systems Analysis, Univ. of North Carolina at Chapel Hill, 1976.

[14] R. R. MEYER, *On the existence of optimal solutions to integer and mixed integer programming problems*, Math. Programming, 7 (1974), pp. 223–235.

[14a] H. M. SALKIN, S. MEHTA AND P. H. SHROFF, *All-integer programming algorithms applied to tableaux with rational coefficients*, Tech. Memorandum No. 298, Dept. of Operations Research, Case Western Reserve Univ., Cleveland, OH, April 1973.

[15] R. D. YOUNG, *A primal (all-integer) integer programming algorithm*, J. Res. Nat. Bureau Standards, Sect. B, 69 (1965), pp. 213–249.

[16] ———, *A simplified primal (all-integer) integer programming algorithm*, Operations Res., 16 (1968), pp. 750–782.

[17] ———, *The eclectic primal algorithm*, Math. Programming, 9 (1975), pp. 294–312.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 1, 1979

0036-0129/79/1701-0006 $01.00/0

D-STABILITY AND MULTI-PARAMETER SINGULAR PERTURBATION*

HASSAN K. KHALIL† AND PETAR V. KOKOTOVIC†

Abstract. A new multi-parameter singular perturbation problem is formulated. Sufficient conditions for uniform asymptotic stability are derived, and asymptotic behavior of solution is investigated.

1. Introduction. Single parameter singular perturbations have been extensively used in analysis and control of dynamic systems [1]. Even if they possess several small parameters, electrical networks with parasitics and control systems with small time constants, masses, etc., are modeled as single parameter problems. This is done by expressing small parameters as known multiples of a particular parameter μ, such as $m=\alpha_1\mu$, $T=\alpha_2\mu$, where m is a small mass and T is a small time constant. A characteristic of this approach is that its results depend on the scaling coefficients α_i which are assumed to be known. In many cases of practical interest such an assumption cannot be justified. In multi-controller problems and differential game problems small parameters may represent different independent ways in which individual control agents simplify the model of the overall system, and therefore the relation between the small parameters must remain arbitrary [2]. It may be argued that a more realistic study of parasitics should also allow for the ignorance of the ratios of small parameters.

The purpose of this paper is to examine the vector singular perturbation problem when all the small parameters are of the same order of magnitude, but can have arbitrary bounded ratios. This problem is different from the multiple time scale problem [3], [4] when the parameters are of different orders of magnitude. We treat the uniform asymptotic stability and initial value problems for multi-parameter singular perturbations. In contrast to the boundary layer system stability requirement of the single parameter case [1], we employ a generalization of D-stability. Several tests are given delineating important classes of systems satisfying this condition.

2. Multiparameter perturbations. Linear systems with N singular perturbation parameters $\varepsilon_1,\cdots,\varepsilon_N$ have the general form

$$\dot{x}=A_0(t)x+\sum_{j=1}^{N}A_{0j}(t)z_j,\qquad x(t_0)=x_0, \tag{1a}$$

$$\varepsilon_i\dot{z}_i=A_{i0}(t)x+\sum_{j=1}^{N}A_{ij}(t)z_j,\qquad z_i(t_0)=z_{i0}, \tag{1b}$$

where $x\in R^{n_0}$, $z_i\in R^{n_i}$, that is the system dimension is $n=n_0+\sum_{i=1}^N n_i$. The small positive scalars $\varepsilon_1,\cdots,\varepsilon_N$ represent time constants, inertias, masses and similar physical parameters [1]. They are ordered as components of a vector $\varepsilon\in R^N$. System (1) satisfies

Assumption I. For all $t\geqq t_0$, all the matrices on the right hand side of (1) are continuous, bounded and have bounded first derivatives.

A characteristic of singularly perturbed systems is that the variables z_i are fast since their derivatives are $1/\varepsilon_i$ large. Under the additional assumption that $\varepsilon_{i+1}/\varepsilon_i\to 0$

* Received by the editors November 10, 1977. This work was supported in part by the Department of Energy, Electric Energy Systems Division, under Contract U.S. ERDA EX-76-C-01-2088, and in part by the National Science Foundation under Grant NSF ENG-74-20091.

† Decision and Control Laboratory, Coordinated Science Laboratory and Department of Electrical Engineering, University of Illinois, Urbana, IL 61801.

as $\varepsilon_i \to 0$, the system (1) exhibits N time scales, that is z_{i+1} is fast relative to z_i. In [3], [4] such multi-time scale systems are treated by nested single parameter perturbations. However, in many real systems the parameters are of the same order and do not allow the multi-time scale assumption. We therefore assume that the ratios of $\varepsilon_1, \cdots, \varepsilon_N$ are bounded by some positive constants $\bar{m}$ and $\bar{M}$

$$\text{(2)} \qquad \bar{m} \leqq \frac{\varepsilon_i}{\varepsilon_j} \leqq \bar{M}, \qquad i, j = 1, \cdots, N,$$

that is the possible values of ε are restricted to a cone $H \subset R^N$. In contrast to the multi-time scale systems, in our case all z_i's are in the same time scale. We call this case the multi-parameter problem. A fundamental requirement for every multi-parameter perturbation result is to hold for all sufficiently small $\varepsilon \in H$, that is as $\varepsilon \to 0$ along any arbitrary path in H.

System (1) is rewritten in a form resembling a single parameter perturbation problem

$$\text{(3a)} \qquad \dot{x} = A_0(t)x + A_{0f}(t)z, \qquad x(t_0) = x_0,$$

$$\text{(3b)} \qquad \mu\dot{z} = DA_{f0}(t)x + DA_f(t)z, \qquad z(t_0) = z_0.$$

However, it is not a single parameter problem because both

$$\text{(4)} \qquad \mu = (\varepsilon_1\varepsilon_2 \cdots \varepsilon_N)^{1/N}$$

and

$$\text{(5)} \qquad D = \text{Block diag}\left[\frac{\mu}{\varepsilon_1}I_1, \cdots, \frac{\mu}{\varepsilon_N}I_N\right]$$

depend on all ε_i's. The above form is convenient since, in view of (2), the matrix D is bounded for all $\varepsilon \in H$,

$$\text{(6)} \qquad m \leqq \frac{\mu}{\varepsilon_i} \leqq M$$

where m, M depend on $\bar{m}$, $\bar{M}$. The matrices A_{0f}, A_{f0} and A_f are formed of the submatrices A_{0i}, A_{i0} and A_{ij}, $i, j = 1, \cdots, N$, respectively, and $z' = [z_1', \cdots, z_N']$. A reduced system is now formally obtained by setting $\varepsilon = 0$ in (3),

$$\text{(7a)} \qquad \dot{\bar{x}} = A_0(t)\bar{x} + A_{0f}(t)\bar{z}, \quad \geqq \quad \bar{x}(t_0) = x_0,$$

$$\text{(7b)} \qquad 0 = A_{f0}(t)\bar{x} + A_f(t)\bar{z}.$$

Assuming that $\det A_f(t) \geqq k > 0$ for all $t \geqq t_0$, (7) can be rewritten as

$$\text{(8)} \qquad \dot{\bar{x}} = [A_0(t) - A_{0f}(t)A_f^{-1}(t)A_{f0}(t)]\bar{x} \triangleq A_r(t)\bar{x}, \qquad \bar{x}(t_0) = x_0.$$

We also define a boundary layer system

$$\text{(9)} \qquad \frac{d\tilde{z}}{d\tau} = DA_f(t_0)\tilde{z}(\tau), \qquad \tilde{z}(0) = z_0 - \bar{z}(t_0),$$

where $\tau = (t - t_0)/\mu$ is the "stretched" time scale.

We are concerned with two problems. First, we seek conditions for the uniform asymptotic stability of (1) for all sufficiently small $\varepsilon \in H$. Second, we want to approximate the solution of the initial value problem (1) in terms of the solution of the reduced problem (8) and the boundary layer problem (9).

For the first problem we make the following

Assumption II. The reduced system (8) is uniformly asymptotically stable.

3. Main results. Our crucial assumption is a generalization of the so called D-stability property of the boundary layer system.

Assumption III. For all $t \geqq t_0$, the matrix $A_f(t)$ has the property that

$$\operatorname{Re} \lambda\{DA_f(t)\} \leqq -2\sigma < 0 \tag{10}$$

where σ is a fixed scalar independent of ε, possibly depending on the bounds m and M.

The main results of this paper are summarized in the following

THEOREM 1. *Under Assumptions* I, II *and* III *there exists a positive scalar ν such that for all $\varepsilon \in H$, $0 < \|\varepsilon\| \leqq \nu$, system* (1) *is uniformly asymptotically stable.*

THEOREM 2. *If Assumptions* I *and* III *are satisfied then for every finite $T > t_0$ there exists a positive scalar ν such that for all $t \in [t_0, T]$ and all $\varepsilon \in H$, $0 < \|\varepsilon\| \leqq \nu$, the solution of the initial value problem* (1) *is approximated by the solution of the reduced problem* (8) *and the boundary layer problem* (9), *that is,*

$$x(t) = \bar{x}(t) + O(\|\varepsilon\|) \tag{11a}$$

$$z(t) = -A_f^{-1}(t)A_{f0}(t)\bar{x}(t) + \tilde{z}(\tau) + O(\|\varepsilon\|). \tag{11b}$$

Moreover, for all $t \in [t_1, T]$, $t_0 < t_1 < T$

$$x(t) = \bar{x}(t) + O(\|\varepsilon\|) \tag{12a}$$

$$z(t) = -A_f^{-1}(t)A_{f0}(t)\bar{x}(t) + O(\|\varepsilon\|). \tag{12b}$$

If in addition Assumption II *is satisfied then* (11) *and* (12) *hold for all $T \in (t_0, \infty)$.*

Our Assumption III has a general form, but it is not verifiable by an algorithm with a finite number of steps. It is satisfied in special cases such as when $A_f(t)$ is block diagonal or block triangular with the on-diagonal matrices satisfying the condition

$$\operatorname{Re} \lambda\{A_{ii}(t)\} \leqq -c_i, \quad \text{for all } t \geqq t_0, \quad i = 1, \cdots, N. \tag{13}$$[1]

Another special case is when A_f is constant and the z_i's are scalars. Then Assumption III means that A_f is D-stable, that is DA_f is a stable matrix for all diagonal matrices D with positive elements. Several D-stability conditions have been investigated in the economic literature [5]. Recently this concept has been used in large scale system analysis [6], [7].

Our Assumption III can be considered as an extension of the notion of D-stability to matrices depending on t and to vector rather than scalar subsystems, that is when $n_i > 1$. In this more general framework we now examine several conditions allowing us to test Assumption III. The first condition is the following:

(i) There exists a block diagonal positive definite matrix $P(t)$,

$$P(t) = \text{Block diag}\,[P_1(t), \cdots, P_N(t)] \tag{14}$$

satisfying

$$c_2\|x\|^2 \leqq x'P(t)x \leqq c_3\|x\|^2 \quad \text{for all } x \in R^{\Sigma_i n_i}, \quad t \geqq t_0, \tag{15}$$

such that $Q(t)$ given by

$$P(t)A_f(t) + A_f'(t)P(t) = -Q(t) \tag{16}$$

[1] In this section $c_1, c_2, \cdots$ are used to denote various fixed positive constant scalars.

is bounded from below by

$$x'Q(t)x \geqq c_4\|x\|^2, \quad \text{for all } x \in R^{\Sigma_i n_i}, \quad t \geqq t_0. \tag{17}$$

This condition implies (10) since the Lyapunov function $v(x) = x'P(t)D^{-1}x$ for the system $dx/ds = DA_f(t)x$ has the negative definite derivative $dv/ds = -x'Q(t)x$. Although this condition does not require the knowledge of D, it is still not finitely verifiable. However, it can be used to generate classes of matrices satisfying (10). An example is the case when $A_f(t)$ is symmetric with $\lambda\{A_f(t)\} \leqq -c_5$ for all $t \geqq t_0$. Then condition (i) is satisfied by $P = I$, while $c_2 = c_3 = 1$, $c_4 = 2c_5$ satisfy (15), (17).

The next condition involves two different conditions introduced in [8], [9] as sufficient conditions for stability of matrices with dominating diagonal blocks.

(ii) The matrices $A_{ii}(t)$ are symmetric with

$$\lambda\{A_{ii}(t)\} \leqq -c_6 \quad \text{for all } t \geqq t_0, \qquad i = 1, \cdots, N, \tag{18}$$

and either

$$\sum_{k \neq i} \|A_{ik}(t)\| < c_6 \quad \text{for all } t \geqq t_0, \qquad i = 1, \cdots, N. \tag{19}$$

or

$$\sum_{k \neq i} \|A_{ii}^{-1}(t)A_{ik}(t)\| < 1 \quad \text{for all } t \geqq t_0, \qquad i = 1, \cdots, N. \tag{20}^{2}$$

If $A_f(t)$ satisfies (18) with (19) or (20) then $DA_f(t)$ satisfies the same condition with c_6 replaced by mc_6 where m is the lower bound in (6).

The last two conditions are due to Siljak [6] and Michel [7] who derived them using the decomposition aggregation method to test the stability of interconnected systems when the isolated subsystems are stable. In these conditions the matrices $A_{ii}(t)$ satisfy (13) and symmetric positive definite $P_i(t)$, $Q_i(t)$ are such that

$$P_i(t)A_{ii}(t) + A'_{ii}(t)P_i(t) = -Q_i(t), \qquad i = 1, \cdots, N. \tag{21}$$

Then there exist positive constants ξ_{ij}, π_{i1}, π_{i2}, π_{i3} and π_{i4} satisfying

$$\|A_{ij}(t)\| \leqq \xi_{ij}, \quad \text{for all } t \geqq t_0, \tag{22}$$

$$\pi_{i1}\|x\|^2 \leqq x'P_i(t)x \leqq \pi_{i2}\|x\|^2, \quad \text{for all } x \in R^{n_i}, t \geqq t_0, \tag{23}$$

$$\pi_{i3}\|x\|^2 \leqq x'Q_i(t)x \leqq \pi_{i4}\|x\|^2, \quad \text{for all } x \in R^{n_i}, t \geqq t_0. \tag{24}$$

In both Siljak's and Michel's condition an $N \times N$ aggregation is formed and tested for the stability of $A_f(t)$. The elements of Siljak's aggregation matrix S are

$$s_{ij} = \begin{cases} -\eta_{i3}, & i = j, \\ \xi_{ij}\eta_{j1}^{-1}\eta_{i4}, & i \neq j \end{cases} \tag{25}$$

where

$$\eta_{i1} = \sqrt{\pi_{i1}}, \qquad \eta_{i3} = \frac{\pi_{i3}}{2\pi_{i2}}, \qquad \eta_{i4} = \frac{\pi_{i2}}{\sqrt{\pi_{i1}}}$$

and those of Michel's matrix T are

$$t_{ij} = \begin{cases} -d_i\pi_{i3}, & i = j, \\ d_i\pi_{i2}\xi_{ij} + d_j\pi_{j2}\xi_{ji}, & i \neq j, \end{cases} \tag{26}$$

[2] The matrix norm in (20) is defined as $\|A\| = [\lambda_{\max}(AA')]^{1/2}$.

for some positive numbers $d_1, \cdots, d_N$.

The Siljak condition is the following:

(iii) The matrices $A_{ii}(t)$ satisfy (13) and the principal minors M_k of S have alternating signs, that is

$$M_k = (-1)^k \det \begin{bmatrix} s_{11} & \cdots & s_{1k} \\ \vdots & & \\ s_{k1} & \cdots & s_{kk} \end{bmatrix} > 0, \qquad k = 1, \cdots, N. \tag{27}$$

To show that his condition implies Assumption III we consider the Lyapunov function

$$v(x) = \sum_{i=1}^{N} \frac{\varepsilon_i \delta_i}{\mu} v_i(x_i), \qquad x' = (x_1', \cdots, x_N'), \quad x_i \in R^{n_i} \tag{28}$$

with

$$v_i(x_i) = (x_i' P_i(t) x_i)^{1/2}, \tag{29}$$

where $\delta_i > 0$, $i = 1, \cdots, N$, are yet unspecified numbers. By derivation similar to that in [6] it can be shown that the derivative of v with respect to the system

$$\frac{dx}{ds} = DA_f(t)x \tag{30}$$

satisfies the inequality

$$\frac{dv}{ds} \leqq \sum_{i=1}^{N} \delta_i \sum_{j=1}^{N} s_{ij} v_j. \tag{31}$$

It is shown in [6] that when inequalities (27) are satisfied there exist numbers $\delta_i > 0$ $(i = 1, \cdots, N)$ and $\pi > 0$ such that

$$\frac{dv}{ds} \leqq -\pi \sum_{i=1}^{N} \delta_i v_i. \tag{32}$$

Hence

$$\frac{dv}{ds} \leqq -\pi m v \triangleq -c_7 v. \tag{33}$$

The last condition is that of Michel:

(iv) The matrices $A_{ii}(t)$ satisfy (13) and there exist numbers d_i, $i = 1, \cdots, N$, such that the matrix T is negative definite.

To show that this condition implies Assumption III we consider the Lyapunov function (28) with δ_i replaced by d_i and $v_i(x_i)$ given by

$$v_i(x_i) = x_i' P_i(t) x_i. \tag{34}$$

In a way similar to [7] it can be shown that its derivative with respect to (30) satisfies the inequality

$$\frac{dv}{ds} \leqq \sum_{i,j} t_{ij} \|x_i\| \, \|x_j\|. \tag{35}$$

Since T is negative definite, let $\lambda = \lambda_{\max}(T) < 0$; thus

$$\frac{dv}{ds} \leqq -\lambda \sum_{i=1}^{N} \|x_i\|^2 = -\lambda \|x\|^2. \tag{36}$$

Using (22) and (36) we obtain

$$\frac{dv}{ds} \leqq \frac{-\lambda m}{\max_i d_i \max_i \pi_{i2}} v \triangleq -c_8 v. \tag{37}$$

Michel's condition is not finitely verifiable since it requires the existence of positive numbers $d_1, \cdots, d_N$. However, a more conservative, finitely verifiable, condition implying Michel's condition can be obtained [7] by writing the matrix T as

$$T = \bar{D}W + W\bar{D} \tag{38}$$

where

$$\bar{D} = \operatorname{diag}[d_1, \cdots, d_N], \qquad d_i > 0, \tag{39}$$

and W is given by

$$w_{ij} = \begin{cases} -\frac{1}{2}\pi_{i3}, & i = j, \\ \pi_{i2}\xi_{ij}, & i \neq j. \end{cases} \tag{40}$$

Then, if the principal minors of W have alternating signs, that is, satisfy (27), there exists matrix $\bar{D}$ such that T is negative definite.

It is important to notice that Siljak's and Michel's conditions are not equivalent. In fact examples can be constructed for matrices which satisfy one of them and do not satisfy the other, and vice versa [9]. These two conditions are particularly important since they are applicable to large scale systems. They are also applicable to nonlinear systems. Grujic [11] has used the decomposition-aggregation method to test the stability of a class of nonlinear singularity perturbed systems. The motive to look at these two conditions and study their implication to our Assumption III was that the aggregation matrices S and T which satisfy the respective condition are D-stable. However, as we have shown, the proof that Siljak's or Michel's condition implies Assumption III does not rely upon the D-stability of S or T, because we have chosen the Lyapunov function in either case in such a way that we obtain the aggregation matrix independent of D.

The above discussion of Assumption III shows that the class of matrices $A_f(t)$ satisfying Assumption III contains important subclasses. However a complete characterization of that class is yet to be made by further studies.

4. Proof. We follow [12] to separate the fast and slow modes of (3). Using

$$\begin{bmatrix} y \\ v \end{bmatrix} = \begin{bmatrix} I - \mu M D^{-1} L & -\mu M D^{-1} \\ L & I \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} \tag{41}$$

the system (3) is transformed into

$$\dot{y} = (A_0(t) - A_{0f}(t)L(t))y, \tag{42a}$$

$$\mu\dot{v} = (DA_f(t) + \mu L(t)A_{0f}(t))v, \tag{42b}$$

where $L(t)$ and $M(t)$ satisfy

$$\mu L = DA_f L - DA_{f0} - \mu L A_0 + \mu L A_{0f} L, \tag{43}$$

$$\mu \dot{M} D^{-1} = -MA_f + A_{0f} - \mu M D^{-1} L A_{0f} + \mu A_0 M D^{-1} - \mu A_{0f} L M D^{-1}, \tag{44}$$

with the initial conditions

$$L(t_0)=A_f^{-1}(t_0)A_{f0}(t_0), \tag{45}$$

$$M(t_0)=A_{0f}(t_0)A_f^{-1}(t_0). \tag{46}$$

We first observe that the fast subsystem (42b) is of the form

$$\mu\dot{z}=(DA_f(t)+\mu\Gamma(t,\varepsilon))z \tag{47}$$

whose properties we examine in Lemmas 1 and 2. Then in Lemmas 3 and 4 we establish the existence and convergence of solutions $L(t)$ and $M(t)$ of (43) and (44). Lemmas 1, 2, 3 and 4 are stated under the Assumptions I and III.

LEMMA 1. *There exist positive scalars ν, K_1 and γ_1 such that for all $\varepsilon\in H$, $0<\|\varepsilon\|\leqq\nu$ and $t\geqq s$, the state transition matrix $\varphi_1(t,s)$ of the system* (47) *with* $\Gamma=0$ *has the property that*

$$\|\varphi_1(t,s)\|\leqq K_1\exp\left[-\frac{\gamma_1}{\mu}(t-s)\right]. \tag{48}$$

Proof. By Assumption I and (6) we have $\|DA_f(t)\|\leqq K_2$, for all $t\geqq t_0$ and $\varepsilon\in H$. Using (10) and Lemma 4 of [13, p. 116] we get for all $\theta\geqq 0$, $\varepsilon\in H$

$$\left\|\exp\left[\frac{\theta}{\mu}DA_f(t)\right]\right\|\leqq K_3\exp\left(-\frac{\sigma}{\mu}\theta\right), \tag{49}$$

where K_3 depends only on σ and K_2. Also there exists $\beta>0$ such that $\|DA_f(t_2)-DA_f(t_1)\|\leqq\beta|t_2-t_1|$, $t_1, t_2\geqq t_0$. Then by Theorem 12 of [13, p. 117] there exists $\mu^*>0$ such that for all $\mu<\mu^*$, $\varphi_1(t,s)$ satisfies (48) with $K_1=K_3^2$ and $\gamma_1<\sigma$; and ν can be chosen to be the radius of the largest ball centered at the origin with $\mu<\mu^*$.

LEMMA 2. *If $\|\Gamma(t,\varepsilon)\|\leqq K_4$, for all $t\geqq t_0$, $\varepsilon\in H$, then there exist positive scalars ν, $\gamma_2<\gamma_1$, such that for all $\varepsilon\in H$, $0<\|\varepsilon\|\leqq\nu$ and $t\geqq s$, the state transition matrix $\varphi_2(t,s)$ of* (47) *satisfies*

$$\|\varphi_2(t,s)\|\leqq K_1\exp\left[-\frac{\gamma_2}{\mu}(t-s)\right]. \tag{50}$$

Moreover, there exists $\nu>0$, $K_5>0$ such that for all $\varepsilon\in H$, $0<\|\varepsilon\|\leqq\nu$, $t\geqq t_0$, the matrix $\varphi_3(t,t_0)=\varphi_2(t,t_0)-\exp\left[DA_f(t_0)((t-t_0)/\mu)\right]$ satisfies

$$\|\varphi_3(t,t_0)\|\leqq K_5\|\varepsilon\|. \tag{51}$$

Proof. Inequality (50) follows from Lemma 1 and Theorem 9 of [13, p. 70]. To prove (51) we notice that $\varphi_3(t,t_0)$ satisfies the equation

$$\dot{\varphi}_3(t,t_0)=\frac{1}{\mu}[DA_f(t)+\mu\Gamma(t,\varepsilon)]\varphi_3(t,t_0)$$
$$+\frac{1}{\mu}[DA_f(t)-DA_f(t_0)+\mu\Gamma(t,\varepsilon)]\exp\left[DA_f(t_0)\left(\frac{t-t_0}{\mu}\right)\right].$$

Noting that $\varphi_3(t_0,t_0)=0$, we obtain

$$\varphi_3(t,t_0)=\int_{t_0}^{t}\varphi_2(t,\tau)\frac{1}{\mu}[DA_f(\tau)-DA_f(t_0)+\mu\Gamma(\tau,\varepsilon)]\exp\left[DA_f(t_0)\frac{(\tau-t_0)}{\mu}\right]d\tau.$$

Using Lemma 1, (50), and the fact that $\|DA_f(t)-DA_f(t_0)\| \leqq \beta(t-t_0)$, we obtain

$$\begin{aligned}
\|\varphi_3(t, t_0)\| &\leqq K_1^2 \int_{t_0}^{t} e^{-(\gamma_2/\mu)(t-\tau)} \frac{1}{\mu}[\beta(\tau-t_0)+\mu K_4]\, e^{-(\gamma_1/\mu)(\tau-t_0)}\, d\tau \\
&\leqq \frac{K_1^2}{\mu} e^{-(\gamma_2/\mu)(t-t_0)} \int_{t_0}^{t} [\beta(\tau-t_0)+\mu K_4]\, d\tau \\
&\leqq \mu K_1^2 e^{-(\gamma_2/\mu)(t-t_0)} \left[\frac{\beta}{2}\left(\frac{t-t_0}{\mu}\right)^2 + K_4\left(\frac{t-t_0}{\mu}\right)\right] \\
&\leqq \mu K_1^2 \left[\frac{2\beta}{\gamma_2^2 e^2} + \frac{K_4}{\gamma_2 e}\right] \\
&\leqq \|\varepsilon\| \frac{K_1^2}{m\sqrt{N}} \left[\frac{2\beta}{\gamma_2^2 e^2} + \frac{K_4}{\gamma_2 e}\right] = K_5 \|\varepsilon\|.
\end{aligned}$$

Next we establish the existence of solutions of (43) and (44). Let us first remark that the state transition matrix of the system $\dot{\eta} = A_0(t)\eta$ satisfies

$$\|\varphi_0(t, s)\| \leqq K_6 \exp[\gamma_3 |t-s|], \quad \text{for all } t, s \geqq t_0, \tag{52}$$

for some positive constants K_6, γ_3, (see [14, p. 287]).

LEMMA 3. *There exists a positive scalar ν such that for all $\varepsilon \in H$, $0 < \|\varepsilon\| \leqq \nu$, $t \geqq t_0$, there exists a continuously differentiable bounded solution $L(t)$ of (43) and (45), satisfying*

$$L(t) = A_f^{-1}(t) A_{f0}(t) + O(\|\varepsilon\|). \tag{53}$$

Proof.[3] Every solution of the integral equation $L(t) = SL(t)$, where

$$\begin{aligned}
SL(t) &= \varphi_1(t, t_0) A_f^{-1}(t_0) A_{f0}(t_0) \varphi_0(t_0, t) \\
&\quad + \int_{t_0}^{t} \varphi_1(t, s) \left[\frac{-1}{\mu} DA_{f0}(s) + L(s) A_{0f}(s) L(s)\right] \varphi_0(s, t)\, ds
\end{aligned} \tag{54}$$

is a solution of (43) with initial condition (45). Hence it is sufficient to prove the existence of a solution of this integral equation. Using the identity

$$LA_{0f}L - \tilde{L}A_{0f}\tilde{L} = (L-\tilde{L})A_{0f}L + \tilde{L}A_{0f}(L-\tilde{L}) \tag{55}$$

and expressions (48) and (52) we obtain

$$\|SL(t) - S\tilde{L}(t)\| \leqq \frac{K_1 K_6 \mu}{\gamma_1 - \mu\gamma_3} \|L-\tilde{L}\|\, \|A_{0f}\| (\|L\| + \|\tilde{L}\|), \tag{56}$$

and

$$\begin{aligned}
\|SL(t)\| &\leqq \frac{K_1 K_6}{\gamma_1 - \mu\gamma_3} (\|DA_{f0}\| + \mu \|A_{0f}\|\, \|L\|^2) \\
&\quad + K_1 K_6 \|A_f^{-1}(t_0)\|\, \|A_{f0}\| \exp\left[-\left(\frac{\gamma_1}{\mu} - \gamma_3\right)(t-t_0)\right].
\end{aligned} \tag{57}$$

[3] In this proof L belongs to the space of bounded continuous $\sum_i n_i \times n_0$ matrix functions on the interval $[t_0, \infty)$ with the norm $\|L\| = \sup_{t \geqq t_0} \|L(t)\|$ where the matrix norm can be any norm. This space is a Banach space.

Letting

$$\rho = 2K_1K_6\left(\frac{2}{\gamma_1}\|DA_{f0}\| + \|A_f^{-1}(t_0)\|\,\|A_{f0}\|\right) \tag{58}$$

we choose $\mu^* > 0$ so small that

$$\mu^*\gamma_3 \leqq \frac{\gamma_1}{2} \quad \text{and} \quad \frac{4K_1K_6\mu^*}{\gamma_1}\|A_{0f}\|\rho \leqq \tfrac{1}{2}. \tag{59}$$

If $\|L\| \leqq \rho$, $\|\tilde{L}\| \leqq \rho$, then for $0 < \mu \leqq \mu^*$ we get

$$\|SL - S\tilde{L}\| \leqq \tfrac{1}{2}\|L - \tilde{L}\| \tag{60}$$

and

$$\|SL(t)\| \leqq \tfrac{3}{4}\rho. \tag{61}$$

By the contraction principle the solution $L(t)$ exists and is unique in $\|L\| \leqq \rho$. To prove (53) we let

$$L(t) = A_f^{-1}(t)A_{f0}(t) + \Delta L(t) \triangleq L_0(t) + \Delta L(t). \tag{62}$$

We note that $\Delta L(t_0) = 0$ and that $\Delta L(t)$ satisfies

$$\Delta\dot{L} = \frac{1}{\mu}(DA_f + \mu L_0A_0)\Delta L - \Delta LA_r + \Delta LA_{0f}\Delta L + R_1, \tag{63}$$

where $R_1 = L_0A_{0f}L_0 - L_0A_0 - L_0$. Let $\varphi_r(t, s)$ and $\varphi_4(t, s)$ be the state transition matrices of equation (8) and $\mu\dot{\xi} = (DA_f(t) + \mu A_f^{-1}(t)A_{f0}(t)A_{0f}(t))\xi$, respectively. The norm of $\varphi_r(t, s)$ satisfies an inequality similar to (52) with constants K_7 and γ_4. By Lemma 2, the norm of $\varphi_4(t, s)$ satisfies an inequality similar to (50) with constants K_1 and $\gamma_5 < \gamma_1$. Then from the form of the solution of (63)

$$\Delta L(t) = \int_{t_0}^{t} \varphi_4(t, s)[\Delta L(s)A_{0f}(s)\Delta L(s) + R_1(s)]\varphi_r(s, t)\,ds, \tag{64}$$

it follows that

$$\begin{aligned} \|\Delta L(t)\| &\leqq \frac{K_1K_7\mu}{\gamma_5 - \mu\gamma_4}(\|A_{0f}\|\,\|\Delta L\|^2 + \|R_1\|) \\ &\leqq \mu\frac{K_1K_7}{\gamma_5}(\|A_{0f}\|(\|L_0\| + \|L\|)^2 + \|R_1\|) \leqq \mu K_8 \\ &\leqq \frac{K_8}{m\sqrt{N}}\|\varepsilon\|, \end{aligned} \tag{65}$$

for some positive constant K_8, which proves (53), and ν can be chosen in a way similar to that in Lemma 1.

LEMMA 4. *There exists a positive scalar ν such that for all $\varepsilon \in H$, $0 < \|\varepsilon\| \leqq \nu$, $t \geqq t_0$; there exists a continuously differentiable bounded solution $M(t)$ of* (44) *and* (46), *satisfying*

$$M(t) = A_{0f}(t)A_f^{-1}(t) + O(\|\varepsilon\|). \tag{66}$$

The proof of this lemma is similar to that of Lemma 3. Based on Lemmas 3 and 4 the matrices of the transformed system (42) can be written as $O(\|\varepsilon\|)$ perturbations of

$A_r(t)$, $DA_f(t)$, that is (42) becomes

$$\dot{y}=(A_r(t)+O(\|\varepsilon\|))y, \qquad y(t_0)=x_0+O(\|\varepsilon\|) \tag{67a}$$

$$\mu\dot{v}=(DA_f(t)+O(\|\varepsilon\|))v, \qquad v(t_0)=z_0+L_0(t_0)x_0+O(\|\varepsilon\|). \tag{67b}$$

Proof of Theorem 1. Since the transformation (41) is nonsingular for all sufficiently small $\varepsilon \in H$ and $L(t)$, $M(t)$ are bounded for all $t \geqq t_0$, it is sufficient to show that each subsystem (67a) and (67b) is uniformly asymptotically stable. This immediately follows from Lemma 2 and Theorem 9 of [13, p. 70].

Proof of Theorem 2. The uniform convergence $y(t)\to\bar{x}(t)$ as $\|\varepsilon\|\to 0$ follows from the continuous dependence of the solution of (67a) on the right-hand side and the initial conditions. Lemma 2 guarantees the uniform convergence $v(t)\to\tilde{z}((t-t_0)/\mu)=\tilde{z}(\tau)$. Using the inverse transformation of (41), we obtain

$$x(t)=y(t)+\mu M(t)D^{-1}v(t)=\bar{x}(t)+O(\|\varepsilon\|), \tag{68a}$$

$$z(t)=-L(t)y(t)+(I-\mu L(t)M(t)D^{-1})v(t)=-A_f^{-1}(t)A_{f0}(t)\bar{x}(t)+\tilde{z}(\tau)+O(\|\varepsilon\|) \tag{68b}$$

which proves (11).

Acknowledgment. The authors are grateful to Professor D. D. Siljak for his fruitful discussions during the course of this work.

REFERENCES

[1] 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.

[2] H. K. Khalil and P. V. Kokotovic, *Control strategies for decision makers using different models of the same systems*, IEEE Trans. Automatic Control, Special Issue of Decentralized Control and Large Scale Systems, AC-23 (1978), pp. 289–298.

[3] R. E. O'Malley, Jr., *Introduction to Singular Perturbations*, Academic Press, New York, 1974.

[4] F. Hoppensteadt, *Properties of solution of ordinary differential equations with small parameters*, Comm. Pure Appl. Math., 24 (1971), pp. 807–840.

[5] C. R. Johnson, *Sufficient conditions for D-stability*, J. Economic Theory, 9 (1974), pp. 53–62.

[6] D. D. Siljak, *Large-Scale Dynamic Systems: Stability and Structure*, Elsevier North-Holland, New York, 1977.

[7] A. N. Michel and R. K. Miller, *Qualitative Analysis of Large Scale Dynamical Systems*, Academic Press, New York, 1977.

[8] D. G. Feingold and R. S. Varga, *Block diagonally dominant matrices and gerneralization of the Gerchgorin circle theorem*, Pacific J. Math., 12 (1962), pp. 1241–1250.

[9] I. F. Pearce, *Matrices with dominating diagonal blocks*, J. Economic Theory, 9 (1974), pp. 159–170.

[10] N. R. Sandell, Jr., P. Varaiya and M. Athans, *A survey of decentralized control methods for large scale systems*, Proc. Systems Engineering for Power, ERDA Conference (Henniker, New Hampshire, August 17–22, 1975).

[11] Lj. T. Grujic, *Converse lemma and singularly perturbed large scale systems*, Proc. JACC (San Francisco, June 22–24, 1977).

[12] K. W. Chang, *Singular perturbations of a general boundary value problem*, SIAM J. Math. Anal., 3 (1972), pp. 520–526.

[13] W. A. Coppel, *Stability and Asymptotic Behavior of Differential Equations*, Heath, Boston, 1965.

[14] W. Hahn, *Stability of Motion*, English translation, Springer-Verlag, New York, 1967.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 1, 1979

0036–0129/79/1701–0007 $01.00/0

CONTROL AND STABILIZATION FOR THE WAVE EQUATION IN A BOUNDED DOMAIN*

GOONG CHEN†

Abstract. Exact controllability problem of the wave equation is studied by the stabilizability method of D. L. Russell. It is shown that if we use the linear "velocity feedback," then the energy of the system will decay uniformly exponentially and exact controllability can be achieved. In case the velocity feedback contains a certain nonlinear dissipative term, then we prove that this system evolves as a nonlinear semigroup with respect to time. Certain estimates for the nonlinear equation are also obtained.

1. Introduction. In this paper we shall be concerned with the control problem for the wave equation

$$\frac{\partial^2 w}{\partial t^2} - \Delta w = f(x, t), \qquad x \in \Omega, \quad t \geqq 0, \tag{1.1}$$

in a suitable function space, where Ω is a bounded domain in $\mathbb{R}^N$ and $f(x, t)$ is a distributed parameter control.

Control problems for the wave equation in a bounded domain have been discussed by very many mathematicians, notably, [10], [11], [19], [20], [21], [22], [23], [28], [29], to mention a few. In comparison with the existing literature, the controls we have obtained in this paper have the following features:

i) they are distributed parameters of the "velocity feedback" form;
ii) they provide us with better regularity property (Theorem 2.5);
iii) exact controllability is achieved.

Basically, this paper consists of two parts: the linear part § 2 and the nonlinear part § 3. In § 2, the linear theory of controllability and stabilizability is discussed; the main theorems are Theorems 2.4 and 2.5. There we show the exponential decay of a semigroup associated with a wave equation with dissipation; thus, it provides us with the feedback controllers to achieve exact controllability, by D. L. Russell's "controllability via stabilizability" principle. In § 3, we consider the case when the feedback controllers are of a specific dissipative nonlinear form. We prove that this system evolves as a nonlinear semigroup (Theorem 3.3). Certain estimates for that nonlinear equation are obtained (Theorem 3.5). Unfortunately we have not been able to derive a nonlinear version of Russell's "controllability via stabilizability" argument.

Throughout this paper, Ω denotes a bounded, open, connected subset of $\mathbb{R}^N$ with piecewise C^∞ regular boundary Γ $(=\partial\Omega)$. There is no restriction on the geometry of the domain Ω, as is in contrast to the case of the boundary value controls [4], [17], [18]. We use $H^m(\Omega)$ to denote the real Sobolev space of order $m > 0$. We use $H_0^m(\Omega)$ to denote the completion of $C_0^\infty(\bar{\Omega})$ in $H^m(\Omega)$.

From now on, $\mathcal{H}_1$ and $\mathcal{H}_2$ will denote the spaces $H_0^1(\Omega) \oplus H^0(\Omega)$ and $(H^2(\Omega) \cap H_0^1(\Omega)) \oplus H_0^1(\Omega)$, respectively. They are equipped with the inner products

$$\begin{aligned} \langle (w_1, v_1), (w_2, v_2) \rangle_{\mathcal{H}_1} &= \int_\Omega (\operatorname{grad} w_1 \cdot \operatorname{grad} w_2 + v_1 \cdot v_2)\, dx, \\ \langle (w_1, v_1), (w_2, v_2) \rangle_{\mathcal{H}_2} &= \langle w_1, w_2 \rangle_{H^2(\Omega)} + \langle v_1, v_2 \rangle_{H^1(\Omega)}. \end{aligned} \tag{1.2}$$

The inner product (1.2) is the "energy inner product" on $\mathcal{H}_1$. By Poincaré's inequality, (1.2) induces a norm on $\mathcal{H}_1$.

* Received by the editors November 22, 1977, and in revised form February 10, 1978.

† Department of Mathematics, Southern Illinois University, Carbondale, Illinois 62901. Now at Department of Mathematics, Pennsylvania State University, University Park, Pennsylvania 16802.

The homogeneous wave equation

$$\frac{\partial^2 w}{\partial t^2} - \Delta w = 0 \tag{1.3}$$

can be written in the form of a system

$$\frac{d}{dt}\begin{bmatrix} w \\ v \end{bmatrix} = \begin{bmatrix} 0 & I \\ \Delta & 0 \end{bmatrix}\begin{bmatrix} w \\ v \end{bmatrix} \equiv A\begin{bmatrix} w \\ v \end{bmatrix} = \begin{bmatrix} v \\ \Delta w \end{bmatrix}, \tag{1.4}$$

with A defined as above and $D(A) = \mathcal{H}_2$.

We are now in a position to pose the

EXACT CONTROLLABILITY PROBLEM (ECP): *For the system*

$$\text{(CS)}\qquad \begin{aligned} &\frac{\partial^2 y(x,t)}{\partial t^2} - \Delta y(x,t) = f(x,t), \qquad x \in \Omega, \quad t \geqq 0, \\ &y(x,t) = 0, \qquad x \in \partial\Omega, \quad 0 \leqq t \leqq T, \end{aligned}$$

given any initial state

$$\begin{bmatrix} y(x,0) \\ z(x,0) \end{bmatrix} \equiv \begin{bmatrix} y(x,0) \\ y_t(x,0) \end{bmatrix} = \begin{bmatrix} y_0 \\ z_0 \end{bmatrix} \in \mathcal{H}_1 \tag{1.5}$$

find a control $f \in L^2(\Omega \times [0, T])$ *so that the solution of* (CS) *and* (1.5) *satisfies the zero terminal condition*

$$y(x,T) \equiv 0, \qquad z(x,T) \equiv y_t(x,T) \equiv 0, \qquad x \in \Omega.$$

This is the problem of "null" controllability. The general controllability problem can be solved by using the time reversibility of the wave equation.

2. Controllability via stabilizability principle applied to the wave equation. The main idea in this paper is "controllability via stabilizability" principle. It can be said to be the most efficient method to obtain exact controllability for hyperbolic partial differential equations. For a finite dimensional control system

$$\text{(2.1)}\qquad \begin{aligned} &\frac{d}{dt}x = Ax + Bf, \qquad x \in \mathbb{R}^N, \quad f \in \mathbb{R}^M, \\ &A, B \text{ are constant } N \times N, N \times M \text{ matrices,} \end{aligned}$$

Russell's principle can be stated as follows:

THEOREM 2.1 (Russell [18]). *Let the system* (2.1) *be both* (+) *and* (−) *stabilizable,* (*or, completely stabilizable*), *namely, there exist two* $M \times N$ *constant matrices* K^+ *and* K^- *such that*

$$A^+ \equiv A + BK^+, \tag{2.2}$$

$$A^- \equiv A + BK^- \tag{2.3}$$

have, respectively, only eigenvalues with negative and positive real parts. Then the system (2.1) *is controllable.*

The main idea in the proof of the above theorem is to use feedback signals

$$f^+(t) = K^+x^+(t), \qquad f^-(t) = K^-x^-(t) \tag{2.4}$$

from auxiliary systems. The control f is obtained by "blending" the above signals, i.e.,

$$f(t) \equiv f^+(t) + f^-(t). \tag{2.5}$$

In the proof, it is essential that we have

$$\|e^{A^+t}\| \leqq K_1 e^{-\alpha t}, \qquad t \geqq 0, \tag{2.6}$$

$$\|e^{-A^-t}\| \leqq K_2 e^{\alpha t}, \qquad t \leqq 0, \tag{2.7}$$

for some K_1, K_2, $\alpha > 0$, in order to guarantee the invertibility of a certain transformation when t is large.

Theorem 2.1 can readily be extended to an infinite-dimensional control system after a suitable modification. Let X be a Banach space of functions and let

$$\frac{d}{dt}x = Ax + Bf, \qquad x(0) = x_0 \in X \tag{IDCS}$$

be a linear evolution equation in X, where

$x = x(t)$: the state of the system at time t;

A: an (unbounded, densely defined, closed) operator from $D(A)$ into X;

f: control, an element in the control space C;

B: an (usually bounded) operator from C into X.

Suppose we can find two operators K^+, K^- from $D(K^+) \subseteq X$, $D(K^-) \subseteq X$ into C satisfying the following requirements:

R1) $D(K^+) \supseteq D(A)$, $D(K^-) \supseteq D(A)$;

R2) $A^+ \equiv A + BK^+$ and $A^- \equiv A + BK^-$ generate two strongly continuous semigroups $S^+(t)$ $(t \geqq 0)$ and $S^-(t)$ $(t \leqq 0)$ which decay forward and backward with respect to time, namely,

$$\|S^+(t)\| \leqq M_1 e^{-\omega_1 t}, \qquad M_1, \omega_1 \geqq 0, \quad t \geqq 0, \tag{2.8}$$

$$\|S^-(t)\| \leqq M_2 e^{\omega_2 t}, \qquad M_2, \omega_2 \geqq 0, \quad t \leqq 0. \tag{2.9}$$

Then the exact controllability result for the control system (IDCS) follows immediately.

Let us return to the distributed parameter control problem of the wave equation. First, we consider the damped wave equation

$$\begin{aligned} &\frac{\partial^2 w}{\partial t^2} + \alpha(x)\frac{\partial w}{\partial t} - \Delta w = 0 \quad \text{in } \Omega, \\ &w|_\Gamma = 0 \quad \text{(boundary condition)}; \\ &w(x,0) = w_0(x) \in H^2(\Omega) \cap H_0^1(\Omega), \qquad w_t(x,0) = v_0(x) \in H_0^1(\Omega) \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(initial conditions)}, \end{aligned} \tag{2.10}$$

where

$$\alpha(x) \geqq a > 0 \quad \text{a.e. on } \Omega, \quad \alpha \in L^\infty(\Omega). \tag{2.11}$$

It can be written in the following form as a system

$$\frac{d}{dt}\begin{bmatrix} w \\ v \end{bmatrix} = \begin{bmatrix} 0 & I \\ \Delta & \alpha(x)I \end{bmatrix}\begin{bmatrix} w \\ v \end{bmatrix} \equiv \hat{A}\begin{bmatrix} w \\ v \end{bmatrix}, \quad \text{with } D(\hat{A}) = \mathcal{H}_2. \tag{2.12}$$

We summarize the functional properties of the operator $\hat{A}$ below.

THEOREM 2.2. i) *$\hat{A}$ is a densely defined, closed, dissipative linear operator on $\mathcal{H}_1$.*

ii) *$\hat{A}^{-1}$ exists and is a compact operator on $\mathcal{H}_1$. Furthermore, the resolvent operator $R(\lambda; \hat{A}) \equiv (\lambda - \hat{A})^{-1}$ is compact for every λ belonging to the resolvent set of $\hat{A}$.*

iii) *$\hat{A}$ is the infinitesimal generator of a strongly continuous semigroup $S(t)$ of contractions on $\mathcal{H}_1$.*

Proof. Without loss of generality, we shall assume that $\alpha(x) \equiv 2\gamma > 0$.

i) it is a routine procedure to verify that $\hat{A}$ is densely defined, closed and linear. To show that $\hat{A}$ is dissipative, let $(w_1, v_1) \in D(\hat{A})$.

$$\begin{aligned}\left\langle \hat{A}\begin{bmatrix} w_1 \\ v_1 \end{bmatrix}, \begin{bmatrix} w_1 \\ v_1 \end{bmatrix} \right\rangle &= \left\langle \begin{bmatrix} v_1 \\ w_1 - 2\gamma v_1 \end{bmatrix}, \begin{bmatrix} w_1 \\ v_1 \end{bmatrix} \right\rangle \\ &= \int_\Omega [\operatorname{grad} v_1 \cdot \operatorname{grad} w_1 + (\Delta w_1 - 2\gamma v_1) v_1]\, dx \\ &= -\int v_1 \cdot \Delta w_1 \, dx + \int \Delta w_1 \cdot v_1 \, dx - 2\gamma \int v_1^2 \, dx \\ &= -2\gamma \int v_1^2 \, dx \leqq 0.\end{aligned} \tag{2.13}$$

ii) Let (f, g) be any given element $\mathscr{H}_1$. Consider the equation

$$\hat{A}\begin{bmatrix} w \\ v \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}. \tag{2.14}$$

This is equivalent to solving

$$v = f, \tag{2.15}$$

$$\Delta w - 2\gamma v = g. \tag{2.16}$$

Substituting (2.15) into (2.16), we obtain

$$\Delta w = 2\gamma f + g, \qquad w|_\Gamma = 0.$$

The above inhomogeneous boundary value problem has a unique solution $w \in H^2(\Omega) \cap H_1^0(\Omega)$ with the property that

$$\|w\|_{H^2(\Omega)} \leqq K_1 \|2\gamma f + g\|_{L^2(\Omega)}. \tag{2.17}$$

Thus (2.14) is completely solvable and

$$\left\| \begin{bmatrix} w \\ v \end{bmatrix} \right\|_{\mathscr{H}_2} \leqq K_2 \left\| \begin{bmatrix} f \\ g \end{bmatrix} \right\|_{\mathscr{H}_1}.$$

Since $\mathscr{H}_2$ is compact in $\mathscr{H}_1$, we conclude that $\hat{A}^{-1}$ exists and is a compact operator on $\mathscr{H}_1$.

From a theorem of classical boundary value problems, we know that $\hat{A}$ can have at most a point spectrum and that $(\lambda - \hat{A})^{-1}$ is a compact operator for every λ in the resolvent set of $\hat{A}$.

iii) This follows immediately from (i) and (ii) and Lumer–Phillips' theorem [16]. This proof also follows from Theorem 3.1.1 [16] because $\hat{A}$ is a bounded perturbation of a skew-adjoint operator A (which is the generator of a group of isometries on $\mathscr{H}_1$).

From the analytic theory of semigroups, we obtain the following theorem.

THEOREM 2.3. *The abstract Cauchy problem*

$$\frac{d}{dt}\begin{bmatrix} w(\cdot, t) \\ v(\cdot, t) \end{bmatrix} = \begin{bmatrix} 0 & I \\ \Delta & \alpha(x) I \end{bmatrix} \begin{bmatrix} w(\cdot, t) \\ v(\cdot, t) \end{bmatrix} \equiv \hat{A}\begin{bmatrix} w(\cdot, t) \\ v(\cdot, t) \end{bmatrix} \tag{2.18}$$

has a unique solution

$$\begin{bmatrix} w(\cdot, t) \\ v(\cdot, t) \end{bmatrix} = S(t)\begin{bmatrix} w_0 \\ v_0 \end{bmatrix} \in C([0, \infty); \mathscr{H}_1) \tag{2.19}$$

for every initial state $(w_0, v_0) \in \mathcal{H}_1$. *We have, furthermore, the classical solution*

$$\begin{bmatrix} w(\cdot, t) \\ v(\cdot, t) \end{bmatrix} = S(t)\begin{bmatrix} w_0 \\ v_0 \end{bmatrix} \in C([0, \infty); [\mathcal{H}_2]) \cap C^1([0, \infty), \mathcal{H}_1) \tag{2.20}$$

if the initial state (w_0, v_0) *belongs to* $D(\hat{A}) = \mathcal{H}_2$, *where* $[\mathcal{H}_2]$ *is normed with the graph norm of* $\hat{A}$.

In the following theorem, we give a proof of exponential decay for the contraction semigroup in Theorem 2.3, based upon the energy method (cf., e.g., [24]).

THEOREM 2.4. *Let* $w(x, t)$ *be the solution of the equation*

$$\frac{\partial^2 w}{\partial t^2} + \alpha(x)\frac{\partial w}{\partial t} - \Delta w = 0 \tag{2.21}$$

with initial state $(w_0, v_0) \in \mathcal{H}_1$. *Then there exist* K, $\beta > 0$ *such that*

$$\left(\int_\Omega \left[\left(\frac{\partial w}{\partial t}\right)^2 + |\text{grad } w|^2\right] dx\right)^{1/2} \equiv \left\| S(t)\begin{bmatrix} w_0 \\ v_0 \end{bmatrix} \right\|_{\mathcal{H}_1} \leqq K e^{-\beta t} \left\| \begin{bmatrix} w_0 \\ v_0 \end{bmatrix} \right\|_{\mathcal{H}_1} \tag{2.22}$$

for all $(w_o, v_0) \in \mathcal{H}_1$.

Proof. We first consider those $(w_0, v_0) \in \mathcal{H}_2$. We multiply (2.21) by $\partial w/(\partial t)$ and integrate by parts to obtain

$$\frac{1}{2}\frac{d}{dt}\int \left[\left(\frac{\partial w}{\partial t}\right)^2 + |\text{grad } w|^2\right] dx + \int \alpha(x)\left(\frac{\partial w}{\partial t}\right)^2 dx = 0. \tag{2.23}$$

Similarly, we use λw and obtain

$$\lambda\left[\frac{d}{dt}\int \frac{\partial w}{\partial t} w \, dx - \int \left(\frac{\partial w}{\partial t}\right)^2 dx + \frac{1}{2}\frac{d}{dt}\int \alpha(x) w^2 \, dx + \int |\text{grad } w|^2 \, dx\right] = 0. \tag{2.24}$$

Adding (2.23) and (2.24), we obtain

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}\int \left[\left(\frac{\partial w}{\partial t}\right)^2 + |\text{grad } w|^2 + 2\lambda \frac{\partial w}{\partial t} w + \lambda \alpha(x) w^2\right] dx \\ &\qquad + \int \left[\alpha(x)\left(\frac{\partial w}{\partial t}\right)^2 + \lambda |\text{grad } w|^2 - \lambda\left(\frac{\partial w}{\partial t}\right)^2\right] dx = 0. \end{aligned} \tag{2.25}$$

We write the above simply as

$$\frac{d}{dt}\bar{P} + \bar{Q} = 0. \tag{2.26}$$

By Poincaré's inequality, there exists a constant C_1 such that

$$\int w^2 \, dx \leqq C_1 \int |\text{grad } w|^2 \, dx. \tag{2.27}$$

Let C_2 be a constant such that

$$\operatorname*{ess\,sup}_{x \in \Omega} \alpha(x) \leqq C_2. \tag{2.28}$$

Now if we take $0 < \lambda \leqq \min\left(\frac{1}{2}, 1/(2(C_1 + C_1 C_2))\right)$, from (2.25) we have

$$\frac{1}{4}\int \left[\left(\frac{\partial w}{\partial t}\right)^2 + |\text{grad } w|^2\right] dx \leqq \bar{P}(t) \leqq \int \left[\left(\frac{\partial w}{\partial t}\right)^2 + |\text{grad } w|^2\right] dx.$$

On the other hand, if we take $0<\lambda \leqq \min(\frac{1}{2}a, 1/(2c_1))$

$$\bar{Q}(t)\geqq \int \left[\frac{1}{2}a\left(\frac{\partial w}{\partial t}\right)^2+\lambda\,|\mathrm{grad}\; w|^2\right] dx \geqq C_3 \int \left[\left(\frac{\partial w}{\partial t}\right)^2+|\mathrm{grad}\; w|^2\right] dx \geqq C_3\bar{P}(t),$$

where $C_3=\min(\frac{1}{2}a, \lambda)$. Thus (2.26) implies

$$\frac{d}{dt}\bar{P}+C_3\bar{P}\leqq\frac{d}{dt}\bar{P}+\bar{Q}=0.$$

Hence

$$\bar{P}(t)\leqq\bar{P}(0)\,e^{-C_3t},$$

so we obtain

$$\int \left[\left(\frac{\partial w}{\partial t}\right)^2+|\mathrm{grad}\; w|^2\right] dx \leqq 4\bar{P}(t)\leqq 4\bar{P}(0)\,e^{-C_3t}\leqq 4\,e^{-C_3t}\left\|\begin{bmatrix} w_0 \\ v_0 \end{bmatrix}\right\|^2_{\mathscr{H}_1}. \tag{2.29}$$

Since (2.29) is independent of the $\mathscr{H}_2$-norm of (w_0, v_0) by the continuity of the semigroup $S(t)$, (2.29) remains true for (w_0, v_0) in $\mathscr{H}_1$. Q.E.D.

Now, the exact controllability problem (ECP) posed in § 1 is readily solvable.

THEOREM 2.5. *There exists a control* $f(x, t)\in L^2(\Omega\times[0, T])$ *such that for any given initial state* $(w_0, v_0)\in\mathscr{H}_1$, *f steers the system* (CS) *to the state* (0, 0) *at time T if T is long enough. Furthermore, f can be chosen in* $C([0, T], L^2(\Omega))$.

Proof. The method we use here is standard, [17], [18], [21], indeed, it is the infinite-dimensional version of "controllability via stabilizability" principle. We let $\tilde{\eta}$ be the solution of

$$\frac{\partial^2\tilde{\eta}}{\partial t^2}-\Delta\tilde{\eta}=-\alpha(x)\frac{\partial\tilde{\eta}}{\partial t}, \qquad \alpha(x) \text{ satisfies (2.12)}, \tag{2.30}$$

$$\begin{bmatrix}\tilde{\eta}_0\\ \tilde{\zeta}_0\end{bmatrix}\in\mathscr{H}_1 \quad \text{(initial state)}. \tag{2.31}$$

An energy decay result of the form (2.22) then applies. We take T large enough so that

$$K\,e^{-\beta T}<1 \tag{2.32}$$

and obtain

$$\left\|\begin{bmatrix}\tilde{\eta}(\cdot\,, T)\\ \tilde{\zeta}(\cdot\,, T)\end{bmatrix}\right\|_{\mathscr{H}_1}\leqq K\,e^{-\beta T}\left\|\begin{bmatrix}\tilde{\eta}_0\\ \tilde{\zeta}_0\end{bmatrix}\right\|_{\mathscr{H}_1}. \tag{2.33}$$

For $0\leqq t\leqq T$ we let

$$\tilde{f}(x, t)\equiv-\alpha(x)\frac{\partial\tilde{\eta}}{\partial t}(x, t). \tag{2.34}$$

Now let $\hat{\eta}$ be the solution of

$$\frac{\partial\hat{\eta}}{\partial t}-\Delta\hat{\eta}=\alpha(x)\frac{\partial\hat{\eta}}{\partial t}, \tag{2.35}$$

$$\begin{bmatrix}\hat{\eta}_T\\ \hat{\zeta}_T\end{bmatrix}=\begin{bmatrix}\tilde{\eta}(\cdot\,, T)\\ \tilde{\zeta}(\cdot\,, T)\end{bmatrix}\in\mathscr{H}_1 \quad \text{(terminal state)}. \tag{2.36}$$

The term $\alpha(x)\partial\hat{\eta}/(\partial t)$ in (2.35) causes energy to decay as t decreases, an estimate of the form (2.22) now applies to give

$$\left\| \begin{bmatrix} \hat{\eta}(\cdot, 0) \\ \hat{\zeta}(\cdot, 0) \end{bmatrix} \right\|_{\mathscr{H}_1} \leqq K e^{-\beta T} \left\| \begin{bmatrix} \tilde{\eta}(\cdot, T) \\ \tilde{\zeta}(\cdot, T) \end{bmatrix} \right\|_{\mathscr{H}_1}.$$

For $0 \leqq t \leqq T$ we let

$$\hat{f}(x, t) \equiv \alpha(x) \frac{\partial \hat{\eta}}{\partial t}(x, t). \tag{2.37}$$

We let

$$(\hat{\eta}_0, \hat{\zeta}_0) = (\hat{\eta}(\cdot, 0), \hat{\zeta}(\cdot, 0)) \tag{2.38}$$

and observe that

$$\eta = \tilde{\eta} - \hat{\eta} \tag{2.39}$$

is a solution of

$$\frac{\partial^2 \eta}{\partial t^2} - \Delta \eta = f \tag{2.40}$$

if we take

$$f = \tilde{f} - \hat{f}. \tag{2.41}$$

We have

$$\eta(\cdot, 0) = \tilde{\eta}_0 - \hat{\eta}_0, \tag{2.42}$$

$$\zeta(\cdot, 0)\left(\equiv \frac{\partial \eta}{\partial t}(\cdot, 0)\right) = \tilde{\zeta}_0 - \hat{\zeta}_0, \tag{2.43}$$

and (by (2.36) and (2.39))

$$\begin{bmatrix} \eta(\cdot, T) \\ \zeta(\cdot, T) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Thus the control f steers the initial state (2.42), (2.43) to (0, 0) during $(0, T)$.

From (2.35), (2.36), it is clear that $(\hat{\eta}_0, \hat{\zeta}_0)$ depends linearly on $(\tilde{\eta}_0, \tilde{\zeta}_0)$ so we can write

$$(\hat{\eta}_0, \hat{\zeta}_0) = F(\tilde{\eta}_0, \tilde{\zeta}_0), \qquad F: \mathscr{H}_1 \to \mathscr{H}_1, \quad \|F\| \leqq K e^{-\beta T} < 1. \tag{2.44}$$

Then (2.42), (2.43) become

$$(\eta_0, \zeta_0) = (I - F)(\tilde{\eta}_0, \tilde{\zeta}_0),$$

and since $\|F\| < 1$, we can solve for arbitrary $(\eta_0, \zeta_0) \in \mathscr{H}_1$ to give (w_0, v_0) in Theorem 2.5. Thus f solves the exact controllability problem.

Since the steering function f is defined by (2.34), (2.37) and (2.41), we obtain $f \in L^2(\Omega \times [0, T])$. By choosing $\alpha(x)$ a positive constant, we obtain $f \in C([0, T], L^2(\Omega))$ since both $\partial\tilde{\eta}/\partial t$ and $\partial\tilde{\eta}/\partial t$ are in $C([0, T], L^2(\Omega))$. Q.E.D.

Remark 2.6. Theorem 2.5 was obtained independently by D. L. Russell by using the controllers (2.34), (2.37) with $\alpha(x) \equiv C > 0$. His result was unpublished.

Remark 2.7. From (2.32) and (2.44), we easily see that the control time T cannot be made as short as we wish, due to the presence of the constant K. The reason is simple: because we have used velocity feedback as our controls, the wave propagation speed in a bounded domain is finite; therefore it takes a while to spread the control effect all over the domain Ω. In order to obtain exact controllability as fast as we wish, we must include a *compensation* factor in the feedback. See [5] for the details.

Likewise, the decay rate coefficient β in (2.22) cannot be magnified by magnifying the damping coefficient $\alpha(x)$ in (2.11). This can be shown by an eigenfunction argument as follows. Let $\{\phi_K\}$ be a complete orthonormal set of eigenfunctions of $(-\Delta)$ in $L^2(\Omega)$, with corresponding (strictly positive) eigenvalues $\{\lambda_K^2\}$. For any $w(x, t)$ which is the solution of the damped wave equation

$$\frac{\partial^2 w}{\partial t^2}+2\gamma\frac{\partial w}{\partial t}-\Delta w=0, \tag{2.45}$$

with initial state $(w_0, v_0)\in \mathcal{H}_2$, we can represent it by

$$w(x,t)=\sum_{k=1}^{\infty} a_k e^{i\mu_k t}\phi_k(x) \tag{2.46}$$

and the μ_K's are determined by (2.45):

$$\sum (i\mu_k)^2 a_k e^{i\mu_k t}\phi_k+\sum (i\mu_k)2\gamma a_k e^{i\mu_k t}\phi_k+\sum \lambda_k^2 a_k e^{i\mu_k t}\phi_k=0. \tag{2.47}$$

The L^2-convergence of each of the above series is ensured by the regularity of the initial condition. Thus each μ_k must satisfy

$$-\mu_k^2+i2\gamma\mu_k+\lambda_k^2=0$$

so

$$\mu_k=i\gamma\pm\sqrt{\lambda_k^2-\gamma^2}. \tag{2.48}$$

A formula in [7], [12] informs us that

$$\lambda_k\sim 2\pi\left(\frac{k}{\omega_N V}\right)^{1/N}, \qquad k \text{ large}, \tag{2.49}$$

where N is the dimension of $\mathbb{R}^N$, ω_N is the volume of the N-dimensional unit ball and V is the volume of Ω. Thus (2.49) can be used to obtain asymptotic values for μ_k. Hence

$$\|w(\cdot,t)\|_{H^1}\leqq K_1 e^{-\beta t}\|(w_0, v_0)\|_{\mathcal{H}_1}, \qquad K_1>0,$$

where

$$\beta\equiv -\sup \operatorname{Re}\{i\mu_k\}_{k=1}^{\infty}>0. \tag{2.50}$$

Similarly,

$$\|v(\cdot,t)\|_{H^0}\equiv\left\|\frac{\partial w}{\partial t}(\cdot,t)\right\|_{H^0}\leqq K_2 e^{-\beta t}\|(w_0, v_0)\|_{\mathcal{H}_1}, \qquad K_2>0.$$

Thus

$$\left\|\begin{bmatrix} w(\cdot,t)\\ v(\cdot,t)\end{bmatrix}\right\|_{\mathcal{H}_1}\leqq K e^{-\beta t}\left\|\begin{bmatrix} w_0\\ v_0\end{bmatrix}\right\|_{\mathcal{H}_1}. \tag{2.51}$$

By continuity of the semigroup, (2.51) remains true for $(w_0, v_0)\in\mathcal{H}_1$. From (2.48) and

(2.51), we see that as γ becomes large,

$$\beta \equiv -\sup \operatorname{Re}\{i\mu_k\} = -\sup \operatorname{Re}\{-\gamma \pm \sqrt{\gamma^2 - \lambda_k^2}\}$$

$$= \gamma(1-\sqrt{1-(\lambda_1/\gamma)^2}) \sim \frac{\lambda_1^2}{2\gamma} \downarrow 0 \quad \text{as } \gamma \uparrow \infty.$$

Therefore magnifying the damping coefficient 2γ will have the opposite effect of diminishing β. This phenomenon can also be overcome by adding a compensation term [5].

Remark 2.8. Theorem 4.2 can be proved without using the energy method. An alternative proof by control-theoretic method (first developed by Quinn and Russell in [17]) proceeds as follows. First, we notice that we have proved (2.51) in Remark 2.7. Hence Theorem 2.5 is true (just think of $\alpha(x) \equiv 2\gamma > 0$). Let $w(x, t)$ be the solution of (2.21) and let $y(x, t)$ be the solution of

$$\frac{\partial^2 y}{\partial t^2} - \Delta y = f, \qquad y|_{\partial\Omega} = 0,$$

$$y(x, 0) = 0, \qquad z(x, 0) \equiv y_t(x, 0) = 0 \quad \text{(initial conditions)},$$

steered by a control f from the initial state $(y(x, 0), z(x, 0)) = (0, 0)$ to the final state $(y(x, T), z(x, T)) = (w(x, T), v(x, T))$ $(\equiv (w(x, T), w_t(x, T)))$.

Applying the divergence theorem, we have

$$u = \int_0^T \int_\Omega \left[\frac{\partial y}{\partial t}\left(\frac{\partial^2 t}{\partial t^2} + \alpha(x)\frac{\partial w}{\partial t} - \Delta w\right) + \frac{\partial w}{\partial t}\left(\frac{\partial^2 y}{\partial t^2} - \Delta y - f\right)\right] dx\, dt$$

$$= \int_0^T \int \left[\frac{\partial}{\partial t}\left(\frac{\partial y}{\partial t}\frac{\partial w}{\partial t} + \operatorname{grad} y \cdot \operatorname{grad} w\right)\right] dx\, dt + \int_0^T \int_\Omega \left[\alpha(x)\frac{\partial y}{\partial t} - f\right]\frac{\partial w}{\partial t}\, dx\, dt.$$

Therefore

$$\int_\Omega \left[\left(\frac{\partial w}{\partial t}(x, T)\right)^2 + |\operatorname{grad} w(x, T)|^2\right] dx = \int_0^T \int_\Omega \left[f - \alpha(x)\frac{\partial y}{\partial t}\right]\frac{\partial w}{\partial t}\, dx\, dt.$$

By the Cauchy–Schwartz inequality,

$$\text{(2.52)} \qquad \int_0^T \int \left(\frac{\partial w}{\partial t}\right)^2 dx\, dt \geqq \frac{\{\int [(\partial w/(\partial t)(x, T))^2 + |\operatorname{grad} w(x, T)|^2]\, dx\}^2}{\int_0^T \int [f - \alpha(x)(\partial y/\partial t)]^2\, dx\, dt}.$$

In the controllability proof of Theorem 2.5, we know that the denominator of the right-hand side of (2.51) can be bounded by a positive multiple of $\int [(\partial w/(\partial t)(x, T))^2 + |\operatorname{grad} w(x, T)|^2]\, dx$, i.e.,

$$\text{(2.53)} \qquad \int_0^T \int \left[f - \alpha(x)\frac{\partial y}{\partial t}\right]^2 dx\, dt \leqq K_1 \int \left[\left(\frac{\partial w}{\partial t}(x, T)\right)^2 + |\operatorname{grad} w(x, T)|^2\right] dx.$$

It is implied by (2.21) that

$$\int \left[\left(\frac{\partial w}{\partial t}(x, 0)\right)^2 + |\operatorname{grad} w(x, 0)|^2\right] dx - \int \left[\left(\frac{\partial w}{\partial t}(x, T)\right)^2 + |\operatorname{grad} w(x, T)|^2\right] dx$$

$$= \int_0^T \int \alpha(x)\left(\frac{\partial w}{\partial t}(x, t)\right)^2 dx\, dt$$

$\Rightarrow$ by (2.12), (2.52), (2.53),

$$\geqq a\int_0^T\int\left(\frac{\partial w}{\partial t}(x,t)\right)^2 dx\,dt \geqq \frac{a}{K_1}\int\left[\left(\frac{\partial w}{\partial t}(x,T)\right)^2+|\text{grad}\,w(x,T)|^2\right]dx.$$

Hence

$$\left\|\begin{bmatrix} w(x,T)\\ v(x,T)\end{bmatrix}\right\|_{\mathscr{H}_1} \leqq \frac{1}{1+K}\left\|\begin{bmatrix} w(x,0)\\ v(x,0)\end{bmatrix}\right\|_{\mathscr{H}_1} \qquad (K\equiv a/K_1).$$

Repeating the above reasoning, we have

$$\begin{aligned}\left\|\begin{bmatrix} w(x,(k+1)T)\\ v(x,(k+1)T)\end{bmatrix}\right\|_{\mathscr{H}_1} &\leqq \frac{1}{1+K}\left\|\begin{bmatrix} w(x,kT)\\ v(x,kT)\end{bmatrix}\right\|_{\mathscr{H}_1}\\ &\leqq \cdots\\ &\leqq \frac{1}{(1+K)^{k+1}}\left\|\begin{bmatrix} w(x,0)\\ v(x,0)\end{bmatrix}\right\|_{\mathscr{H}_1}.\end{aligned}$$

Hence the exponential decay result is proved.

Remark 2.9. Comparing $\hat{A}$ in (2.13) with A in (1.4) as well as with A^+ in (R2), we notice that $\hat{A}$ plays the role of A^+ with

$$\hat{A}=A^+=A+BK^+,$$

i.e.,

$$\begin{bmatrix} 0 & I\\ \Delta & \alpha(x)I\end{bmatrix}=\begin{bmatrix} 0 & I\\ \Delta & 0\end{bmatrix}+\begin{bmatrix} 0 & 0\\ 0 & \alpha(x)I\end{bmatrix}$$

so $\begin{bmatrix} 0 & 0\\ 0 & \alpha(x)I\end{bmatrix}$ plays the role of BK^+. In [5], a more general form of BK^+ with the appearance

$$BK^+=\begin{bmatrix} aI & bI\\ cI & dI\end{bmatrix}, \qquad a, b, c, d \text{ complex numbers},$$

is discussed. This, in particular, means feedback signals of the second order (i.e., involving $\partial^2 w/(\partial t^2)$).

3. The case when the feedback control contains a nonlinear term. In the previous section, we see that the steering function $f(x,t)$ is obtained by "blending" two feedback signals. In this section we shall study the evolution of the system when the feedback signal contains a nonlinear dissipative term. Let $f(x,t)$ have the form

$$f(x,t)=-\gamma_1\frac{\partial w}{\partial t}-\gamma_2\left(\frac{\partial w}{\partial t}\right)^3, \qquad \gamma_1,\gamma_2>0. \tag{3.1}$$

We are interested in studying the equation

$$\frac{\partial^2 w}{\partial t^2}-\Delta w=f(x,t), \tag{3.2}$$

which is a nonlinear wave equation. Following § 2, we can also write (3.2) in the form of a system

$$\frac{d}{dt}\begin{bmatrix} w\\ v\end{bmatrix}=\begin{bmatrix} v\\ \Delta w-\gamma_1 v-\gamma_2 v^3\end{bmatrix}\equiv\bar{A}\begin{bmatrix} w\\ v\end{bmatrix}, \tag{3.3}$$

where $\bar{A}$ is a nonlinear operator on $\mathscr{H}_1$.

In general, $\bar{A}$ is a well-defined operator on $\mathscr{H}_1$ with domain $[H^2(\Omega)\cap H_0^1(\Omega)]\oplus(H_0^1(\Omega)\cap L^6(\Omega))$. However, if the space dimension N is less than or equal to 3, we have

$$H_0^1(\Omega)\subseteq L^p(\Omega), \qquad 1\leqq p<\infty, \quad N=2, \tag{3.4}$$

$$H_0^1(\Omega)\subseteq L^6(\Omega), \qquad N=3, \tag{3.5}$$

by the Sobolev imbedding theorem. Therefore for the sake of simplicity we shall assume $N\leqq 3$ and this will be sufficient for all practical purposes. Thus $\bar{A}$ becomes a nonlinear operator on $\mathscr{H}_1$ with domain $\mathscr{H}_2$.

Our main tool in studying the system (3.3) is nonlinear functional analysis. Some nonlinear terminology such as maximal dissipative sets, nonlinear semigroups, monotonicity, hemicontinuity, etc., will be needed. We refer the readers to [2], [8], [9], [13], [26], [27] for their definitions.

PROPOSITION 3.1. *The operator $\bar{A}$ on $\mathscr{H}_1$ defined by*

$$\bar{A}\begin{bmatrix} w \\ v \end{bmatrix}=\begin{bmatrix} v \\ \Delta w-\gamma_1 v-\gamma_2 v^3 \end{bmatrix} \tag{3.6}$$

with domain $D(\bar{A})=\mathscr{H}_2$ is a maximal dissipative operator on $\mathscr{H}_1$.

Proof. For $(w, v)\in D(\bar{A})$, we have

$$\left\langle \bar{A}\begin{bmatrix} w \\ v \end{bmatrix}, \begin{bmatrix} w \\ v \end{bmatrix}\right\rangle=-\int(\gamma_1 v^2+\gamma_2 v^4)\,dx\leqq 0, \tag{3.7}$$

where every term in this integral is well-defined because $v\in H_0^1(\Omega)\subseteq L^6(\Omega)\subseteq L^4(\Omega)$. Thus $\bar{A}$ is dissipative.

In order to prove that $\bar{A}$ is maximal dissipative, we must, by Minty's theorem, show that $R(I-\bar{A})=\mathscr{H}_1$, [2], [8]. Let (f, g) be any given element in $\mathscr{H}_1$. We want to show that

$$(I-\bar{A})\begin{bmatrix} w \\ v \end{bmatrix}=\begin{bmatrix} f \\ g \end{bmatrix} \tag{3.8}$$

is always solvable. This means

(3.I)

$$w-v=f, \tag{3.9}$$

$$v-\Delta w+\gamma_1 v+\gamma_2 v^3=g. \tag{3.10}$$

Formally, we can substitute $w=v+f$ into (3.10) and obtain

(3.II)

$$-\Delta v+(1+\gamma_1)v+\gamma_2 v^3=g+\Delta f, \tag{3.11}$$

$$-\Delta w=g-(1+\gamma_1)v-\gamma_2 v^3, \tag{3.12}$$

where Δf is a distribution in the space $(H_0^1(\Omega))'\equiv H^{-1}(\Omega)$. It is easy to see that (3.I) and (3.II) are equivalent because the Laplacian Δ is an isomorphism from $H_0^1(\Omega)$ onto $H^{-1}(\Omega)$.

We first solve (3.11), which is a nonlinear elliptic equation of the function v only. The method is by variational inequalities. We use the following theorem:

PROPOSITION 3.2. *Let E be a reflexive Banach space, T a monotone hemicontinuous operator from E into E'. Suppose that the coercivity condition*

$$\frac{(Tu, u)}{\|u\|_E} \to \infty \quad \textit{as } \|u\|_E \to \infty \tag{3.13}$$

holds. Then T is surjective from E onto E'.

We define

$$T(z) \equiv -\Delta z + (1+\gamma_1)z + \gamma_2 z^3;$$

then T maps $H_0^1(\Omega)$ into $H^{-1}(\Omega)+L^2(\Omega)=H^{-1}(\Omega)$, since $z \in L^6(\Omega)$ implies $z^3 \in L^2(\Omega)$. Now, with slight modifications of Lemmas VI.8, VI.9, Theorem VI.1 and Example VII.1 presented in [27], we summarize the properties of T below:

1) T is strictly monotone from $H_0^1(\Omega)$ into $H^{-1}(\Omega)$.

2) T is hemicontinuous, indeed, T is continuous from $H_0^1(\Omega)$ strong into $H^{-1}(\Omega)$ strong.

3) The coercivity condition (3.13) is satisfied.

Thus the assumptions of Proposition 3.2 are satisfied. Hence T is surjective. T is also injective due to the strict monotonicity. Indeed, T^{-1} exists and is locally Hölder continuous with exponent (2/3), i.e.,

$$\|u_1 - u_2\|_{H_0^1(\Omega)} \leqq K(\|f_1 - f_2\|_{H^{-1}(\Omega)})^{2/3}$$

for $f_1 = Tu_1$, $f_2 = Tu_2$ and $\|f_1 - f_2\|_{H^{-1}(\Omega)}$ small.

Therefore the solution of (3.11) exists and is unique. Now w can be solved from (3.12) (with the Dirichlet boundary condition), since the right hand side of (3.12) is in $L^2(\Omega)$. A unique $w \in H^2(\Omega) \cap H_0^1(\Omega)$ solves (3.12). Thus (3.8) is solved and the proof is complete. Q.E.D.

We readily conclude the following:

THEOREM 3.3. *The nonlinear operator $\bar{A}$ in* (3.6) *generates a strongly continuous nonlinear semigroup of contractions on $\mathscr{H}_1$.*

This theorem follows from Proposition 3.1 and a theorem of Crandall and Pazy [9]:

> Let S be a maximal dissipative set in $H \times H$, where H is a Hilbert space. Then there is a unique strongly continuous semigroup of contractions $S(t)$ on $D(\bar{T})$ such that T^0, the minimal section of T, is the generator of $S(t)$.

Because the operator $\bar{A}$ is single-valued, we know that $\bar{A} = \bar{A}^0$ is the generator of $S(t)$. We know [9] further that for any $x \in D(\bar{A}) = \mathscr{H}_2$:

(i) $S(t)x \in D(\bar{A})$ for all $t \geqq 0$ and the function $t \to \bar{A}S(t)x$ is continuous from the right on $[\cdot, \infty)$.

(ii) $S(t)x$ has a right derivative $(d^+/dt)A(t)x$ at every $t \geqq 0$ and $(d^+/dt)S(t)x = \bar{A}S(t)x$ for all $t \geqq 0$.

(iii) $(d/dt)S(t)x = \bar{A}S(t)x$ exists and is continuous except at a countable number of values $t \geqq 0$.

From the above we see that $S(t)x$ is "less smooth" than a classical solution because $(d/dt)S(t)x = \bar{A}S(t)x$ is only right continuous on $[0, \infty)$.

Remark 3.4. From [13], we have learned that if we use the "method of compactness," then the solution $w(x, t)$ of the nonlinear equation

$$\frac{\partial^2 w}{\partial t^2} + \gamma_1 \frac{\partial w}{\partial t} + \gamma_2 \left(\frac{\partial w}{\partial t}\right)^3 - \Delta w = 0, \qquad w \in H_0^1(\Omega), \tag{3.14}$$

$$w(\cdot,0)=w_0\in H^2(\Omega)\cap H_0^1(\Omega), \tag{3.15}$$

$$w_t(\cdot,0)=v_0\in H_0^1(\Omega), \tag{3.16}$$

exists and is unique, with the following properties:

(i) $w\in L^\infty(0,T;H^2(\Omega)\cap H'_0(\Omega))$;

(ii)

$$\partial w/\partial t\in L^\infty(0,T;H'_0(\Omega)); \tag{3.17}$$

(iii) $\partial^2 w/\partial t^2\in L^\infty(0,T;L^2(\Omega))$;

(iv) $\partial w/\partial t\in L^4(\Omega\times(0,T))$ (implied by (3.17) in this case).

If the initial conditions (3.15), (3.16) are less smooth, say, we have

$$(3.14),\qquad w(\cdot,0)=w_0\in H_0^1(\Omega),\qquad w_t(\cdot,0)=v_0\in L^2(\Omega),$$

and if we use a combination of the method of "monotonicity" and "compactness," then the solution also exists and is unique, with

(v) $w\in L^\infty(0,T;H_0^1(\Omega))$;

(vi) $\partial w/\partial t\in L^\infty(0,T;L^2(\Omega))$.

In the following theorem we show some growth rate estimates for solutions of (3.2) with smooth initial condition $(w_0,v_0)\in\mathcal{H}_2$. Unfortunately, due to the presence of the nonlinear term, we cannot obtain a strong exponential decay result. Neither can we derive a nonlinear version of the "controllability via stabilizability" principle.

THEOREM 3.5. *Let $w(x,t)$ be the solution of*

$$\frac{\partial^2 w}{\partial t^2}+\gamma_1\frac{\partial w}{\partial t}+\gamma_2\left(\frac{\partial w}{\partial t}\right)^3-\Delta w=0,\qquad (\gamma_2\ \textit{small}) \tag{3.18}$$

with initial conditions

$$w(\cdot,0)=w_0\in H^2(\Omega)\cap H_0^1(\Omega), \tag{3.19}$$

$$v(\cdot,0)\left(\equiv\frac{\partial w}{\partial t}(\cdot,0)\right)=v_0\in H_0^1(\Omega), \tag{3.20}$$

and boundary condition

$$w(x,t)|_{x\in\Gamma}=0,\qquad t\geqq 0;$$

then there exist constants K, $\alpha>0$, which are independent of (w_0,v_0), such that

$$\left\|\begin{bmatrix} w(\cdot,t)\\ v(\cdot,t)\end{bmatrix}\right\|^2_{\mathcal{H}_1}\leqq K\left\{e^{-\alpha t}\left\|\begin{bmatrix} w_0\\ v_0\end{bmatrix}\right\|^2_{\mathcal{H}_1}+\int_0^t\int_\Omega e^{\alpha(\tau-t)}\left(\frac{\partial w}{\partial t}\right)^6 dx\,d\tau\right\}, \tag{3.21}$$

$$\left\|\begin{bmatrix} w(\cdot,t)\\ v(\cdot,t)\end{bmatrix}\right\|^2_{\mathcal{H}_1}\leqq K\left\{e^{-\alpha t}\left\|\begin{bmatrix} w_0\\ v_0\end{bmatrix}\right\|^2_{\mathcal{H}_1}+\int_0^t\int_\Omega e^{\alpha(\tau-t)}w^4\,dx\,d\tau\right\}, \tag{3.22}$$

$$\left\|\begin{bmatrix} w(\cdot,t)\\ v(\cdot,t)\end{bmatrix}\right\|^2_{\mathcal{H}_1}\leqq K\left\{e^{-\alpha t}\left\|\begin{bmatrix} w_0\\ v_0\end{bmatrix}\right\|^2_{\mathcal{H}_1}+\left\|\begin{bmatrix} w_0\\ v_0\end{bmatrix}\right\|^4_{\mathcal{H}_1}\cdot\frac{1}{\alpha}(1-e^{-\alpha t})\right\}. \tag{3.23}$$

Proof. Again, we use the energy method. Multiplying (3.18) by $\partial w/(\partial t)+\lambda w$ and integrating by parts over Ω, we obtain

$$\begin{aligned}(3.24)\qquad &\frac{\partial}{\partial t}\int_\Omega \frac{1}{2}\left[\left(\frac{\partial w}{\partial t}\right)^2+|\text{grad } w|^2+\lambda\left(2\frac{\partial w}{\partial t}w+w^2\right)\right]dx\\ &+\int_\Omega\left\{\left[\gamma_1\left(\frac{\partial w}{\partial t}\right)^2+\gamma_2\left(\frac{\partial w}{\partial t}\right)^4\right]+\lambda\left[|\text{grad } w|^2-\left(\frac{\partial w}{\partial t}\right)^2+\gamma_2\left(\frac{\partial w}{\partial t}\right)^3 w\right]\right\}dx=0.\end{aligned}$$

We write (3.24) simply as

$$(3.25)\qquad \frac{dP}{dt}+Q=0.$$

By Poincaré's inequality, there exists a constant $K>0$ such that

$$\int w^2\,dx \leqq K\int |\text{grad } w|^2\,dx.$$

If we choose $0<\lambda<\min\left(\frac{1}{3}, 1/(9C)\right)$, then

$$(3.26)\qquad \frac{1}{6}\int\left[|\text{grad } w|^2+\left(\frac{\partial w}{\partial t}\right)^2\right]dx\leqq P(t)\leqq\frac{5}{6}\int\left[|\text{grad } w|^2+\left(\frac{\partial w}{\partial t}\right)^2\right]dx$$

i.e.,

$$(3.27)\qquad \frac{1}{6}\left\|\begin{bmatrix} w(\cdot,t)\\ v(\cdot,t)\end{bmatrix}\right\|^2_{\mathcal{H}_1}\leqq P(t)\leqq\frac{5}{6}\left\|\begin{bmatrix} w(\cdot,t)\\ v(\cdot,t)\end{bmatrix}\right\|^2_{\mathcal{H}_1}.$$

On the other hand, using Young's inequality,

$$(3.28)\qquad \int\lambda\gamma_2\left(\frac{\partial w}{\partial t}\right)^3 w\,dx\leqq\frac{\lambda\gamma_2}{2}\int\left(\frac{\partial w}{\partial t}\right)^6dx+\frac{\lambda\gamma_2}{2}\int w^2\,dx,$$

we obtain

$$\begin{aligned}Q(t)&\geqq\int\left\{\gamma_1\left(\frac{\partial w}{\partial t}\right)^2+\gamma_2\left(\frac{\partial w}{\partial t}\right)^4+\lambda|\text{grad } w|^2-\lambda\left(\frac{\partial w}{\partial t}\right)^2-\frac{\lambda\gamma_2}{2}w^2-\frac{\lambda\gamma_2}{2}\left(\frac{\partial w}{\partial t}\right)^6\right\}dx\\ &\geqq\int\left\{\gamma_1\left(\frac{\partial w}{\partial t}\right)^2+\gamma_2\left(\frac{\partial w}{\partial t}\right)^4+\lambda|\text{grad } w|^2-\lambda\left(\frac{\partial w}{\partial t}\right)^2-\frac{\lambda\gamma_2}{2}K|\text{grad } w|^2\right.\\ &\qquad\left.-\frac{\lambda\gamma_2}{2}\left(\frac{\partial w}{\partial t}\right)^6\right\}dx,\end{aligned}$$

so if we choose γ_2 such that $0<\gamma_2<2/K$ and choose λ such that $0<\lambda<\gamma_1$, we will have

$$(3.29)\qquad Q(t)\geqq\delta\int\left[|\text{grad } w|^2+\left(\frac{\partial w}{\partial t}\right)^2\right]dx-\frac{\lambda\gamma_2}{2}\int\left(\frac{\partial w}{\partial t}\right)^6dx,$$

where $\delta=\min\left(\gamma_1-\lambda, \lambda\left(1-(\gamma_2/_2)K\right)\right)$.

From (3.25), (3.26) and (3.29) we deduce that

$$(3.30)\qquad \frac{dP}{dt}+\frac{6}{5}\delta P-\frac{\lambda\gamma_2}{2}\int\left(\frac{\partial w}{\partial t}\right)^6dx\leqq\frac{dP}{dt}+Q=0.$$

By (3.27) and (3.30), we conclude that (3.21) has been proved. We remark here that $\partial w/\partial t\in L^6(\Omega)$ because Theorem 3.3 implies $\partial w/\partial t\in H_0^1(\Omega)$ and (3.5) applies.

If we apply Young's inequality in the following form instead of (3.28), we have

$$\int \lambda\gamma_2\left(\frac{\partial w}{\partial t}\right)^3 w\,dx \leqq \lambda\gamma_2\left[\frac{3}{4}\int\left(\frac{\partial w}{\partial t}\right)^4 dx+\frac{1}{4}\int w^4\,dx\right]. \tag{3.31}$$

Following similar procedures as above, we can obtain

$$\frac{dP}{dt}+\frac{5}{6}\delta P-\frac{\lambda\gamma^2}{4}\int w^4 \leqq \frac{dP}{dt}+Q=0.$$

Thus (3.22) follows immediately.

Since the imbedding

$$H_0^1(\Omega)\rightarrow L^4(\Omega)$$

is continuous (indeed, compact), there exists a constant C such that

$$\left(\int w^4\,dx\right)^{1/4} \leqq C\left(\int |\text{grad } w|^2\,dx\right)^{1/2}.$$

Thus

$$\int w^4\,dx \leqq C^4\left(\int |\text{grad } w|^2\,dx\right)^2 \leqq C^4\left\|\begin{bmatrix} w(\cdot,t)\\ v(\cdot,t)\end{bmatrix}\right\|_{\mathscr{H}_1}^4$$
$$\leqq C^4\left\|\begin{bmatrix} w_0\\ v_0\end{bmatrix}\right\|^4.$$

Using this in relation in (3.22), we see that (3.23) is proved. Q.E.D.

Note added in proof. After the present paper was submitted, the author received the following paper from H. O. Fattorini: *The time optimal problem for distributed control of systems described by the wave equation*, Control Theory of Systems Governed by Partial Differential Equations, Aziz, Wingate, Balas, eds., Academic Press, New York 1977. Fattorini obtains exact controllability by a method using sine and cosine operators. He does not approach the problem from the stabilizability point of view, as we do in this paper.

Acknowledgments. The author wishes to express his thanks to M. G: Crandall and L. Tartar for teaching him nonlinear functional analysis and partial differential equations; to D. L. Russell for many stimulating discussions and a control theory seminar in the summer of 1977 at the University of Wisconsin-Madison, which contributed to a large part of this paper. Finally, the author thanks the referees for their valuable criticism which aided him in improving this paper.

REFERENCES

[1] S. Agmon, *Elliptic Boundary Value Problems*, Van Nostrand, New York, 1965.

[2] V. Barbu, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoof, Leyden, the Netherlands, 1976.

[3] G. Chen and R. S. Millman, *Control theory for the wave equation in compact Riemannian manifolds*, revised preprint.

[4] G. Chen, *Energy decay estimates and exact boundary value controllability for the wave equation in a bounded domain, Part I*, J. Math. Pures Appl., to appear.

[5] ———, *Energy decay estimates and exact boundary value controllability for the wave equation in a bounded domain, Part II*, in preparation.

[6] G. Chen and D. L. Russell, *Asymptotic stability and energy decay rates for the wave equation with boundary dissipation in a bounded domain*, in preparation.

[7] R. COURANT AND D. HILBERT, *Method of Mathematical Physics*, vol. II, Wiley-Interscience, New York, 1962.

[8] M. G. CRANDALL, *Lecture notes in nonlinear analysis*, University of Wisconsin-Madison, Spring, 1977, unpublished.

[9] M. G. CRANDALL AND A. PAZY, *Semigroup of nonlinear contractions and dissipative sets*, J. Functional Analysis, 3 (1969), pp. 376–418.

[10] H. O. FATTORINI, *Some remarks on complete controllability*, this Journal, 4 (1966), pp. 686–694.

[11] ———, *Estimates for sequences biorthogonal to certain complex exponentials and boundary control of the wave equation*, preprint.

[12] P. D. LAX AND R. S. PHILLIPS, *Decaying modes for the wave equation in the exterior of an obstacle*, Comm. Pure Appl. Math., 22 (1969), pp. 737–787.

[13] J. L. LIONS, *Quelques méthodes de resolution des problèmes aux limites non linéaires*, Dunod-Gauthier-Villars, Paris, 1969.

[14] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non-homogènes et applications*, vol. I, Dunod, Paris, 1968.

[15] ———, Ibid., vol. II, Dunod, Paris, 1968.

[16] A. PAZY, *Semigroup of Linear Operators and Applications to Partial Differential Equations*, Lecture note #10, Dept. of Math., University of Maryland, College Park, 1974.

[17] J, QUINN AND D. L. RUSSELL, *Asymptotic stability and energy decay rates for solutions of hyperbolic equations with boundary damping*, MRC Technical Summary Report #1575, University of Wisconsin-Madison, Nov., 1975.

[18] D. L. RUSSELL, *Exact boundary value controllability theorems for wave and heat processes in star-complemented regions*, Differential Games and Control Theory, Roxin, Liu, Sternberg, eds., Marcel Dekker, New York, 1974.

[19] ———, *Boundary value control of the higher dimensional wave equation*, this Journal, 9 (1971), pp. 29–42.

[20] ———, part II, this Journal, 9 (1971), pp. 401–419.

[21] ———, *A unified boundary value controllability theory for hyperbolic and parabolic partial differential equations*, Studies in Appl. Math., 52 (1973), pp. 189–211.

[22] M. SLEMROD, *Stabilization of boundary control systems*, J. Differential Equations, to appear.

[23] ———, *A note on complete controllability and stabilizability for linear control systems in Hilbert space*, this Journal, 12 (1974), pp. 500–509.

[24] W. A. STRAUSS, *The energy method in nonlinear partial differential equations*, Instituto de Mathematica Pura e Applicada, Brazil, 1969.

[25] L. TARTAR, *Lecture notes in partial differential equations*, University of Wisconsin-Madison, Fall, 1974, unpublished.

[26] ———, *Evolution equations in infinite dimensions*, MRC Technical Summary Report #1485, University of Wisconsin-Madison, Dec., 1974.

[27] ———, *Variational methods and monotonicity*, MRC Technical Summary Report #1571, University of Wisconsin-Madison, Oct., 1975.

[28] R. TRIGGIANI, *Extensions of rank conditions for controllability and observability to Banach spaces and unbounded operators*, this Journal, 14 (1976), pp. 313–338.

[29] ———, *On the stabilizability problem in Banach space*, J. Math. Anal. Appl., 52 (1975), pp. 383–403.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 1, 1979

0036-0129/79/1701-0008 $01.00/0

AN ADAPTIVE PRECISION METHOD FOR NONLINEAR OPTIMIZATION PROBLEMS*

K. SCHITTKOWSKI†

Abstract. Consider the problem of minimizing a real-valued Gateaux-differentiable function φ over some subset of a normed linear space. The main advantage of the algorithm presented will be that it is possible to vary the accuracy of the evaluation of φ and its derivative. It is allowed to use low precision while far from a solution and to improve it step by step as a solution is approached. In order to enlarge the application of the algorithm, these approximations are defined by approximating sets and not by functions. Constrained problems are transformed into unconstrained ones using a penalty approach, where the algorithm controls automatically the degree of penalizing the function φ. The minimization algorithm requires a general class of gradient-type search directions, restricted only by the condition that they are bounded away from orthogonality with the gradient. The steplength procedure is a simple reduction method as proposed by Armijo (1966). It is possible to derive global convergence results in the sense that starting with arbitrary initial values a critical point is approximated.

1. Introduction.

Consider the general minimization problem

$$\inf \varphi(u), \qquad u \in C, \tag{1}$$

where φ is a real-valued function defined on a normed linear space E, and C is a subset of E. It is assumed that

$$\inf \{\varphi(u) : u \in C\} > -\infty.$$

In many applications, however, it is not possible to compute φ exactly or a precise evaluation is extremely expensive. Let us consider some examples:

Example 1.1. Solve the min-max problem

$$\min_{x \in C} \max_{y \in D} f(x, y)$$

with $C \subset \mathbb{R}^n$, $D \subset \mathbb{R}^m$. The function $\varphi(x) := \max \{f(x, y) : y \in D\}$ is (with suitable assumptions) continuously Gateaux-differentiable, but must be evaluated by a second maximization procedure, cf. Klessig and Polak [9].

Example 1.2. Consider the problem of finding an $x \in \mathbb{R}^n$ with

$$f(x) \in K,$$

where K is a closed, convex subset of $\mathbb{R}^m$. A solution can be derived from minimizing the function $\varphi(x) := d(f(x), K)^2$, where $d(\cdot, K)$ is the distance function. If K is a complicated set, the function φ can be computed only approximately, cf. [19].

Example 1.3. Let E be the set of all continuous, piecewise continuously differentiable functions with $u(0) = u(1) = 0$, $C = E$, and φ given by

$$\varphi(u) := \int_0^1 f(t, u(t), \dot{u}(t))\, dt.$$

For the numerical solution of such variational problems with an adaptive precision steepest descent method see [21], [22].

* Received by the editors February 28, 1977, and in final revised form March 13, 1978.

† Institut für Angewandte Mathematik und Statistik, Universität Würzburg, Würzburg, West Germany.

Example 1.4. A similar approach for unconstrained control problems was used by Klessig and Polak [10], where $E = L_2^m[0, T]$, $C = E$, and φ is given by

$$\varphi(u) := g(x(T, u)).$$

$x(\cdot, u)$ is the solution of the differential equation

$$\dot{x} = f(x, u(t), t), \qquad x(0, u) = x_0.$$

For some further comments about this example see § 2.

Example 1.5. A time-optimal heat diffusion process leads to the problem of heating a thin rod at one end-point such that a given temperature distribution $k_0 \in L_2[0, 1]$ will be approximated as soon as possible, subject to the L_2-norm, with a given accuracy ε, see [23]. Using the bang-bang principle and the assumption that the optimal solution has at most k switching times, we get a finite dimensional optimization problem of the kind (1) with

$$\varphi(T, t) := T$$

for all $(T, t) = (T, t_1, \cdots, t_k) \in \mathbb{R}^{k+1}$ and

$$C := \{(T, t) : \|y(\cdot, T, u(T, t)) - k_0(\cdot)\|_2 \leqq \varepsilon, 0 \leqq t_1 \leqq \cdots \leqq t_k \leqq T\},$$

where the control $u(T, t)$ is bang-bang with jumps at $t_1, \cdots, t_k$ and the temperature distribution $y(s, T, u)$ is given by

$$y(s, T, u) := \sum_{j=1}^{\infty} \alpha_j \cos(\mu_j s) \int_0^T u(\tau) \exp(-\mu_j^2 (T - \tau))\, d\tau$$

for all $s \in [0, 1]$, $T > 0$, $u \in L_2[0, T]$ and with suitable constants α_j, μ_j.

It will be our special aim to overcome the difficulty that the cost function φ and its derivative must be evaluated exactly to guarantee convergence. If the numerical computation of φ is not accurate enough, the convergence of the minimization algorithm cannot be ensured in general. The main advantage of the subsequent algorithm will be that it is possible to vary the accuracy of the evaluation of φ or its derivative. In particular, it is allowed to use low accuracy while far from a solution and improve it step-by-step as a solution is approached. Klessig and Polak [9], [10], [16] developed general algorithms and applied them to the special problems mentioned above. Now we want to use a different approach.

The minimization procedure is derived from an unconstrained version defined in $\mathbb{R}^n$ [19] and in arbitrary normed linear spaces [20], which allows very general gradient-type search directions restricted only by the condition that they are bounded away from orthogonality with the gradient. Furthermore the steplength procedure is the simple reduction method of Armijo [1] leading to short programming codes.

If restrictions appear, i.e. if $C \neq E$, then they are handled by defining certain penalty functions φ^n. The degree of penalizing the function φ, i.e. the index n, is controlled by the algorithm. This degree can be low, while far from a solution, but will be raised infinitely, if a solution is approached.

So we get an algorithm which controls both the accuracy of the approximations and the degree of penalizing φ. It is possible to develop global convergence theorems in the sense that a critical point is approximated.

In the next section, Example 1.4 is used to motivate the assumptions necessary to construct the method and to illustrate the fundamental ideas of the algorithm. In § 3 all assumptions are gathered to define the penalty functions, the accuracy of the approximations, and to present the algorithm. Some preliminary results are established in § 4 in order to achieve (in the following section) a global convergence theorem for the unconstrained case. Some convergence results for constrained problems are given in the last section.

2. A motivating example. In order to motivate the subsequent algorithm and its assumptions about the structure and convergence of the approximations, consider the following unrestricted optimal control problem: Let E be identical with $L_\infty^m[0, T]$, the space of all essentially bounded measurable functions from $[0, T]$ into $\mathbb{R}^m$ subject to the norm

$$\|u\|_\infty := \operatorname*{ess\,sup}_{0 \leqq t \leqq T} \|u(t)\|_2,$$

where $u \in E$ and $\|\cdot\|_2$ denotes the Euclidean norm in $\mathbb{R}^m$. The cost function φ is defined by

$$\varphi(u) := g(x(T, u)).$$

$x(\cdot, u)$ satisfies the differential equation

$$\dot{x} = f(x, u(t), t)$$

with $x(0, u) := x_0$, and continuously differentiable functions $g: \mathbb{R}^n \to \mathbb{R}$ and $f: \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^n$. Let us assume that this trajectory exists and is uniquely determined for all $u \in E$. Leese [12] stated that the well known costate equations define the continuous Fréchet differential operator $D\varphi$ in the following way:

$$D\varphi(u)v := \int_0^T p(t, u)^T \frac{\partial}{\partial u} f(x(t, u), u(t), t)v(t)\, dt$$

for all $u, v \in E$. $p(\cdot, u)$ is the solution of the differential equation

$$\dot{p} = -\frac{\partial}{\partial x} f(x(t, u), u(t), t)^T p$$

with end condition $p(T, u) := (\partial/\partial x)g(x(T, u))^T$.

A gradient-type descent method for this unrestricted optimal control problem is the following: Starting from an initial point u_0 and a $\gamma > 0$, the algorithm constructs an iteration sequence $\{u_k\}$ such that $u_{k+1} = u_k - \beta_k s_k$. For the search direction s_k we require that $\|s_k\| = 1$ and $D\varphi(u_k)s_k \geqq \gamma\|D\varphi(u_k)\|$; the steplength β_k has to satisfy at least the condition $\varphi(u_{k+1}) < \varphi(u_k)$.

In order to realize this algorithm computationally first we have to approximate the controls or, more precisely, we have to determine a transformation from E into the finite dimensional space $E_l := \mathbb{R}^m \times \mathbb{R}^l$ and back using some interpolation rule. Various methods are used in practical implementations of methods solving optimal control problems: Hull [7] proposed polynomials and Chebyshev polynomials, Johnson [8] preferred cubic splines, Kraft [11] continuous piecewise linear functions. Furthermore step functions with predetermined mesh points, cf. Klessig and Polak [10], or with variable switching times, cf. Sargent and Sullivan [18] are frequently used. But in all cases this process serves to identify special finite dimensional subsets $E_l^* \subset E$ with the space E_l. For our general purpose we denote this relationship by a so-called interpolation operator $I_l: E_l \to E_l^*$ and a so-called restriction operator $R_l: E \to E_l$, and require that I_l, R_l are linear, bounded and satisfy the condition

$$I_l R_l u = u, \quad R_l I_l z = z \quad \text{for all } u \in E_l^* \text{ and } z \in E_l.$$

Given now any control $u \in E$, we have to integrate the state and costate equations to compute the cost function φ and its derivative. For this we need a differential equation solver of the following more general type: If $\dot{x} = F(x, t)$ defines a differential equation subject to the boundary conditions $x(0) \in A_0$, $x(T) \in A_T$, where $A_0, A_T \subset \mathbb{R}^{n'}$, $F: \mathbb{R}^{n'} \times \mathbb{R} \to \mathbb{R}^{n'}$, then we get subsets $\chi_l(F, A_0, A_T)$ of $\mathbb{R}^{n'} \times \mathbb{R}^l$ such that for each $\varepsilon > 0$ there is a l_ε with

$$\|I_l z(\cdot) - x(\cdot)\|_\infty \leqq \varepsilon$$

for all $z \in \chi_l(F, A_0, A_T)$ and $l \geqq l_\varepsilon$. Here $x(\cdot)$ is the exact solution of the differential equation.

Application to the control problem leads to the definitions

$$F_u\left(\begin{pmatrix} x \\ p \end{pmatrix}, t\right) := \begin{pmatrix} f(x, u(t), t) \\ -\dfrac{\partial}{\partial x} f(x, u(t), t)^T p \end{pmatrix},$$

$x, p \in \mathbb{R}^n$, $t \in [0, T]$, and

$$A_0 := \left\{ \begin{pmatrix} x_0 \\ p \end{pmatrix} : p \in \mathbb{R}^n \right\}, \qquad A_T := \left\{ \begin{pmatrix} x \\ \dfrac{\partial}{\partial x} g(x)^T \end{pmatrix} : x \in \mathbb{R}^n \right\}.$$

There are various numerical realizations to determine $\chi_l(F_u, A_0, A_T)$ of $\mathbb{R}^{2n} \times \mathbb{R}^l$ using forward and backward integration. The most simple methods are based on one-step integration functions, for example the Euler–Cauchy method as realized by Klessig and Polak [10]. One-step and multi-step schemes of higher order are used by Hager [6] for the discretization of optimal control problems. He also presents the corresponding convergence rates. Furthermore extrapolation methods as originally proposed by Bulirsch and Stoer [3] can be used for the practical implementation.

Finally we have to use the integration and restriction operators, and also the differential equation solver to compute approximations of the cost functions φ and its derivative $D\varphi$. For each $X \in E_l$ and $l \in N$ we determine $\begin{pmatrix} Y \\ Q \end{pmatrix} \in \chi_l(F_{I_l X}, A_0, A_T)$, $Y, Q \in \mathbb{R}^n \times \mathbb{R}^l$, and let

$$\varphi_l(X) := g(y_l),$$

and also a matrix $\varphi_l'(X) \in E_l$ with the columns

$$\frac{\partial}{\partial u} f(y_j, x_j, t_j)^T q_j,$$

where y_j, q_j, x_j are the jth columns of Y, Q, and X, respectively, and t_j is given by the jth element of the first row of $R_l e$ with $e(t) := (t, \cdots, t)^T$, $j = 1, \cdots, l$.

The subsequent algorithm proceeds from the gradient-type descent method given in the beginning of this section. The iterates are elements of the finite dimensional spaces E_l or, equivalently, are elements of finite dimensional subspaces of E. The cost function φ and its derivative $D\varphi$ are approximated by an integration rule. But the index l is not fixed, the algorithm raises it whenever advantageous. It is possible to handle additional constraints via penalty functions.

3. The algorithm. Let $L(E, \mathbb{R})$ be the set of all bounded linear operators from E into $\mathbb{R}$. The norm in E and the corresponding operator norm of its dual space $L(E, \mathbb{R})$

is denoted by $\|\cdot\|$. Sometimes we need the distance function $d(\cdot, U)$ for some subset U of E, which is defined by

$$d(v, U) := \inf_{u\in U} \|v-u\|$$

for each $v \in E$.

The approximations are defined on a family of normed linear spaces E_l, $l \in \mathbb{N}_0 := \mathbb{N}\cup\{0\}$. To simplify the notations, denote each norm in E_l by $\|\cdot\|$, too. Suppose that there is a monotone increasing family of subsets E_l^* of E, i.e. $E_l^* \subset E_{l+1}^* \subset E$ for all $l \in \mathbb{N}_0$, and that there are two sequences of continuous linear operators $R_l: E \to E_l$, $I_l: E_l \to E_l^*$, $l \in \mathbb{N}_0$. Let each R_l be bijective on E_l^* and its inverse be identical with I_l, more precisely

$$(2) \qquad I_lR_lu = u, \qquad R_lI_lx = x$$

for all $u \in E_l^*$ and $x \in E_l$. R_l and I_l will be called restriction or interpretation operator, respectively. As a special seminorm we introduce for $l \in \mathbb{N}_0$ and $f \in L(E, \mathbb{R})$

$$\|f\|_l := \sup\{|f(u)|: \|u\| \leqq 1,\ u \in E_l^*\}.$$

The subsequent algorithm is based on the idea to penalize the function φ, if an iterate leaves the set C.

DEFINITION 3.1. A sequence of real-valued functions $\{\varphi^n\}_{n\in\mathbb{N}}$ is called a *sequence of penalty functions* for problem (1), if the following statements are valid:

a) Each φ^n is continuous and continuously Gateaux-differentiable on E; more precisely, the linear and bounded differential operator $D\varphi^n(u)$ defines a continuous function $D\varphi^n: E \to L(E, \mathbb{R})$.

b) Let L be a closed and bounded subset of E, $\varepsilon > 0$, and C_ε be the neighborhood $C_\varepsilon := \{z \in E: d(z, C) \leqq \varepsilon\}$. Then there is a $\delta_\varepsilon > 0$, l_ε, $n_\varepsilon \in \mathbb{N}$ with

$$(3) \qquad \|D\varphi^n(u)\|_l \geqq \delta_\varepsilon$$

for all $n \geqq n_\varepsilon$, $l \geqq l_\varepsilon$, $u \in L\backslash C_\varepsilon$.

Let $\{\varphi^n\}$ be a sequence of penalty functions. Now it is possible to state the assumption about the accuracy of the approximations of φ^n, an important part of our algorithm.

ASSUMPTION 3.2. For each $n \in \mathbb{N}$, $l \in \mathbb{N}_0$, and $x \in E_l$ there are subsets $\Phi_l^n(x) \subset \mathbb{R}$ and $\Phi_l^{n\prime}(x) \subset L(E_l, \mathbb{R})$ such that the following conditions are valid:

a) The convergence of the approximations of φ^n and $D\varphi^n$ is uniform on every closed and bounded subset U of E; that means for all $\varepsilon > 0$ there exists a $l_\varepsilon \in \mathbb{N}$ with

$$(4) \qquad |\varphi_l^n(x) - \varphi^n(I_lx)| \leqq \varepsilon, \qquad \|\varphi_l^{n\prime}(x) - D\varphi^n(I_lx)I_l\| \leqq \varepsilon$$

for all $l \geqq l_\varepsilon$, $n \in \mathbb{N}$, $\varphi_l^n(x) \in \Phi_l^n(x)$, $\Phi_l^{n\prime}(x) \in \Phi_l^{n\prime}(x)$, and for all $x \in E_l$ with $I_lx \in U$.

b) There exists positive real numbers R and I with

$$(5) \qquad \|R_l\| \leqq R, \qquad \|I_l\| \leqq I$$

for all $l \in \mathbb{N}_0$.

It is required that the convergence of the approximations is uniform with respect to n. That means in practice that the convergence of the approximations depends on the index n.

The approximations of φ^n and $D\varphi^n$ are defined by subsets and not by functions. This makes sense for the solution of Example 1.4, as mentioned in the last section, and for Example 1.1 or 1.2, too. As the lth approximation at a point $x \in \mathbb{R}^n$ we would use the results of the lth iteration step of an optimization procedure to maximize $f(x, y)$, $y \in D$, or to minimize $\|f(x)-z\|$, $z \in K$. But such an optimization algorithm is in general not unique and the iterates depend on a lot of parameters, for example on the choice of the initial values.

It has been shown by the author [19], [21], [23] that suitable approximations of the problems mentioned in Examples 1.2, 1.3, and 1.5 fulfill Assumption 3.2. Klessig and Polak [9], [10] proved the uniform convergence of the approximations subject to Examples 1.1 and 1.4. But it should be noted that their abstract prototype algorithm as presented in [10] requires an assumption on the rate of improvement in the precision of the approximation to the cost function which is not needed in 3.2.

Now we are able to develop an algorithm for solving the general minimization problem (1). A simpler version for the unconstrained case can be found in [20]. A point $u \in E$ shall be called an optimal solution if the norm of the derivative is "small", more precisely, if $\|D\varphi^n(u)\|_l \leqq \varepsilon$ for an $l \geqq l_\varepsilon$, $n \geqq n_\varepsilon$, and if the distance from the set C, i.e. $d(u, C)$, is less than ε where ε, l_ε, and n_ε are given by the user.

For the implementation of the procedure select positive real numbers γ_j, σ_j with

$$(6) \qquad 1 \geqq \gamma_j \geqq \gamma > 0, \qquad \sigma^* \geqq \sigma_j \geqq \sigma > 0$$

for all $j \in \mathbb{N}_0$; furthermore real numbers μ_j, α_j, $j \in \mathbb{N}_0$, with

$$(7) \qquad \mu_j > 0, \quad \lim_{j\to\infty} \mu_j = 0 \quad \text{and} \quad \alpha_j \geqq \alpha_{j+1} > 0, \quad \lim_{j\to\infty} \alpha_j = 0.$$

The algorithm proceeds as follows:

ALGORITHM 3.3.

0) Initial stage: Choose $q(0)$, $n(0) \in \mathbb{N}_0$, $x_0 \in E_{q(0)}$. For $k = 0, 1, 2, \cdots$ compute $q(k+1)$, $n(k+1)$, and $x_{k+1} \in E_{q(k+1)}$ as follows:

1) Denote $l := q(k)$, $n := n(k)$, $x_k^l := x_k$.

2) Determine $\rho_k^l := \sigma_k \|\varphi_l^{n\prime}(x_k^l)\|$ for a $\varphi_l^{n\prime}(x_k^l) \in \Phi_l^{n\prime}(x_k^l)$.

3) If $\rho_k^l \leqq \mu_n$, let $n := n+1$, otherwise go to step 4.

4) Calculate

$$(8) \qquad x_k^{l+1} := R_{l+1}(I_l x_k^l) \in E_{l+1}, \text{ set } l := l+1.$$

5) Determine a $\varphi_l^n(x_k^l) \in \Phi_l^n(x_k^l)$ and a $\varphi_l^{n\prime}(x_k^l) \in \Phi_l^{n\prime}(x_k^l)$. Let $\rho_k^l := \sigma_k \|\varphi_l^{n\prime}(x_k^l)\|$.

6) Compute a search direction $s_k^l \in E_l$ with $\|s_k^l\| \leqq 1$ and

$$(9) \qquad \varphi_l^{n\prime}(x_k^l) s_k^l \geqq \gamma_k \|\varphi_l^{n\prime}(x_k^l)\|.$$

7) Evaluate the smallest nonnegative integer $j \leqq l$ with

$$(10) \qquad \varphi_l^n(x_k^l - \alpha_j \rho_k^l s_k^l) \leqq \varphi_l^n(x_k^l) - \tfrac{1}{2}\alpha_j \rho_k^l \gamma_k \|\varphi_l^{n\prime}(x_k^l)\|,$$

where $\varphi_l^n(x_k^l - \alpha_j \rho_k^l s_k^l) \in \Phi_l^n(x_k^l - \alpha_j \rho_k^l s_k^l)$ can be chosen arbitrarily.

8) If such a j does not exist, go to step 4, otherwise go to 9.

9) Define the new iterate

$$x_{k+1} := x_k^l - \alpha_j \rho_k^l s_k^l \in E_l$$

and let $j_k := j$, $q(k+1) := l$, $n(k+1) := n$.

In other words, the algorithm works as follows: Starting from an initial point, 3.3 yields an iteration sequence $x_k \in E_{q(k)}$, $k \in \mathbb{N}_0$, and furthermore two monotone increasing sequences of positive integers $\{q(k)\}$ and $\{n(k)\}$ with $\lim_{k\to\infty} q(k) = \infty$ and (this has to be shown) $\lim_{n\to\infty} n(k) = \infty$. In each iteration step the accuracy of the approximations will be raised at least by 1. This might be replaced by a more general condition which only requires that the increase of the accuracy, defined by the integer l, is unbounded.

If we omit all approximations, we see that the algorithm is frequently studied in mathematical programming theory, see Polak, Sargent, Sebastian [17] and Leese [12] for example, and in numerical implementations, see Murtagh and Sargent [14, pp. 215–246]. The computation of the search direction s_k^l can be characterized as a gradient-type method. It is a "downhill" direction only restricted by condition (9) which requires that s_k^l is bounded away from orthogonality with the gradient. It is obvious that (9) can be replaced by a more general condition using the concept of forcing functions as was done by Ortega and Rheinboldt [15].

The Armijo [1] steplength rule (10) is one of the most simple possibilities from the computational point of view to guarantee global convergence of the underlying algorithm. On the other hand, quadratic or cubic interpolation schemes yield better numerical results. Therefore a combination of the two methods, in the sense that a condition of the kind (10) or a related version is used as a stopping criterion for an interpolation routine, is implemented sometimes in existing programming codes, see, for example, Gill [5] and Biggs [2].

Since the approximations φ_l^n are in general not functions, it is not possible to guarantee the existence of a j implementing inequality (10). Therefore the restriction "$j \leqq l$" is necessary and it will be an important part of the convergence statement to ensure that the loop from step 4 to step 8 is finite. Each index l for which this loop will be left, defines a new integer $q(k+1)$.

Note that the idea to transform the iterations from E to E_l and back by bounded operators is related to a paper of Esser [4, pp. 69–88], who studied the discretization of extremal problems

$$\inf_{x\in E} f(x) = \lim_{k\to\infty} \inf_{x\in E_k} f_k(x).$$

His results are based on the more general discretization theory of Stummel [24], who examined the discrete convergence of mappings.

4. Some preliminary results. In this section some preliminary results are gathered which will be helpful for proving the subsequent convergence theorem. First note that for each iterate x_k^l of Algorithm 3.3

$$(11) \qquad I_{l+1}x_k^{l+1} = I_l x_k^l \quad \text{or} \quad I_{q(k+1)}x_k^{q(k+1)} = I_{q(k)}x_k.$$

This follows from the assumption $E_l^* \subset E_{l+1}^*$ for all $l \in \mathbb{N}_0$. Further, consider an operator $A \in L(E_l, \mathbb{R})$, $l \in \mathbb{N}_0$. Then we get the estimates

$$(12) \qquad \|AR_l\|_l \leqq R\|A\|, \qquad \|A\| \leqq I\|AR_l\|_l.$$

Step 7 of Algorithm 3.3 indicates that the mean value theorem is fundamental:

LEMMA 4.1. *For each $u, v \in E$ and $\lambda > 0$ there exists a $\lambda^* \in (0, \lambda)$ with*

$$\varphi^n(u-\lambda v) - \varphi^n(u) = -\lambda D\varphi^n(u-\lambda^* v)v, \qquad n \in \mathbb{N}.$$

As an immediate consequence of Assumption 3.2 we establish now the following lemma, whose proof is obvious:

LEMMA 4.2. *Assumption* 3.2a) *is equivalent to the following statement: Let U be a closed, bounded subset of E, and let $\varepsilon > 0$. Then there is an $l_\varepsilon \in \mathbb{N}$ with*

$$(13) \qquad |\varphi_l^n(R_l u) - \varphi^n(u)| \leqq \varepsilon, \qquad \|\varphi_l^{n\prime}(R_l u)R_l - D\varphi^n(u)\|_l \leqq \varepsilon,$$

for all $l \geqq l_\varepsilon$, $n \in \mathbb{N}$, $\varphi_l^n(R_l u) \in \Phi_l^n(R_l u)$, $\varphi_l^{n\prime}(R_l u) \in \Phi_l^{n\prime}(R_l u)$, and for all $u \in U \cap E_l^*$.

It is easy to show that the boundedness of $D\varphi^n$ implies the boundedness of its approximations:

LEMMA 4.3. *Assume that the function $D\varphi^n$ is bounded on a closed and bounded subset L of E, $n \in \mathbb{N}$. Then there is an $M_n > 0$ and an $l_0 \in \mathbb{N}$ with*

$$\|\varphi_l^{n\prime}(x)\| \leqq M_n$$

for all $l \geqq l_0$, $x \in E_l$ with $I_l x \in L$, and $\varphi_l^{n\prime}(x) \in \Phi_l^{n\prime}(x)$.

Proof. From Assumption 3.2 there is an $l_0 \in \mathbb{N}$ with

$$\|\varphi_l^{n\prime}(x) - D\varphi^n(I_l x)I_l\| < 1$$

for all $l \geqq l_0$, $x \in E_l$ with $I_l x \in L$, and $\varphi_l^{n\prime}(x) \in \Phi_l^{n\prime}(x)$. $D\varphi^n$ is bounded on L, i.e. there exists an $M_n^* > 0$ with

$$\|D\varphi^n(u)\| < M_n^*$$

for all $u \in L$. Now choose an $l \geqq l_0$, $x \in E_l$ with $I_l x \in L$, and a $\varphi_l^{n\prime}(x) \in \Phi_l^{n\prime}(x)$. Then

$$\|\varphi_l^{n\prime}(x)\| \leqq 1 + \|D\varphi^n(I_l x)I_l\| \leqq 1 + M_n^* I =: M_n. \qquad \text{Q.E.D.}$$

As mentioned in § 3, it is an important part of the convergence analysis to show the finiteness of the loop from step 8 to step 4 of Algorithm 3.3.

THEOREM 4.4. *The loop between step 4 and step 8 of Algorithm 3.3 is finite for every k with $\|D\varphi^n(I_{q(k)}x_k)\|_{q(k)} > 0$, where $n := n(k)$, if $\rho_k^{q(k)} > \mu_{n(k)}$, $n := n(k)+1$ otherwise.*

Proof. Assume that the loop from step 4 to step 8 is not finite for a $k^* \in \mathbb{N}$ with $\|D\varphi^{n^*}(I_{l^*}x_{k^*})\|_{l^*} > 0$, $l^* := q(k^*)$, $n^* := n(k^*)$, if $\rho_{k^*}^{l^*} > \mu_{n(k^*)}$, $n^* := n(k^*)+1$ otherwise. To simplify the notations we omit the index n^*. So we get for each $l > l^*$ a $\varphi_l(x_{k^*}^l) \in \Phi_l(x_{k^*}^l)$, $\varphi_l'(x_{k^*}^l) \in \Phi_l'(x_{k^*}^l)$, and a $s_{k^*}^l \in E_l$ with

$$\varphi_l'(x_{k^*}^l)s_{k^*}^l \geqq \gamma_{k^*}\|\varphi_l'(x_{k^*}^l)\|$$

and

$$\varphi_l(x_{k^*}^l - \alpha_j\rho_{k^*}^l s_{k^*}^l) > \varphi_l(x_{k^*}^l) - \tfrac{1}{2}\alpha_j\rho_{k^*}^l\gamma_{k^*}\|\varphi_l'(x_{k^*}^l)\| \tag{14}$$

for all j, $0 \leqq j \leqq l$.

Let $l > l^*$. For each $\lambda > 0$ the mean value theorem 4.1 yields a $\lambda^* \in (0, \lambda)$ with

$$\varphi(I_l(x_{k^*}^l - \lambda\rho_{k^*}^l s_{k^*}^l)) - \varphi(I_l x_{k^*}^l) = -\lambda\rho_{k^*}^l D\varphi(I_l(x_{k^*}^l - \lambda^*\rho_{k^*}^l s_{k^*}^l))I_l s_{k^*}^l. \tag{15}$$

Define $\varepsilon^* := (\gamma\sigma/(\sigma^* R))\|D\varphi(I_{l^*}x_{k^*})\|_{l^*}$. $D\varphi$ is continuous, i.e. there is a $\delta > 0$ with

$$\|D\varphi(I_l(x_{k^*}^l - \lambda^*\rho_{k^*}^l s_{k^*}^l)) - D\varphi(I_l x_{k^*}^l)\| \leqq \frac{\varepsilon^*}{8I}, \tag{16}$$

if $\|I_l(x_{k^*}^l - \lambda^*\rho_{k^*}^l s_{k^*}^l) - I_l x_{k^*}^l\| \leqq \lambda^*\rho_{k^*}^l I < \delta$. Applying Lemma 4.3 to $L := \{I_{l^*}x_{k^*}\}$ we get an $l_1 > l^*$ and an $M > 0$ with $\|\varphi_l'(x_{k^*}^l)\| \leqq M$ for all $l \geqq l_1$, since $x_{k^*}^l = R_l I_{l^*} x_{k^*}$. Choose a λ with $\lambda\sigma^* MI < \delta$. From

$$\begin{aligned}\lambda^*\rho_{k^*}^l I &\leqq \lambda\sigma_{k^*}\|\varphi_l'(x_{k^*}^l)\|I \\ &\leqq \lambda\sigma^* MI < \delta\end{aligned}$$

it follows that (16) is valid for all $l \geqq l_1$. Assumption 3.2 with $U := \{I_{l^*}x_{k^*}\}$ shows there is an $l_2 \geqq l_1$ with

$$|\varphi_l'(x_{k^*}^l)s_{k^*}^l - D\varphi(I_l x_{k^*}^l)I_l s_{k^*}^l| < \varepsilon^*/8$$

for all $l \geqq l_2$. Combining this estimate with (16) we get

$$\begin{aligned}
&|D\varphi(I_l(x^l_{k^*} - \lambda^* \rho^l_{k^*} s^l_{k^*}))I_l s^l_{k^*} - \varphi'_l(x^l_{k^*})s^l_{k^*}| \\
&\qquad \leqq \|D\varphi(I_l(x^l_{k^*} - \lambda^* \rho^l_{k^*} s^l_{k^*})) - D\varphi(I_l x^l_{k^*})\| \, \|I_l s^l_{k^*}\| \\
&\qquad\qquad + |D\varphi(I_l x^l_{k^*})I_l s^l_{k^*} - \varphi'_l(x^l_{k^*})s^l_{k^*}| \\
&\qquad < \varepsilon^*/4, \quad \text{or} \\
&D\varphi(I_l(x^l_{k^*} - \lambda^* \rho^l_{k^*} s^l_{k^*}))I_l s^l_{k^*} \geqq \varphi'_l(x^l_{k^*})s^l_{k^*} - \varepsilon^*/4.
\end{aligned} \tag{17}$$

From Lemma 4.2, an $l_3 \geqq l_2$ can be found with

$$\begin{aligned}
\|\varphi'_l(x^l_{k^*})R_l - D\varphi(I_{l^*}x_{k^*})\|_1 &= \|\varphi'_l(R_l(I_{l^*}x_{k^*}))R_l - D\varphi(I_{l^*}x_{k^*})\|_l \\
&\leqq \tfrac{1}{4}\|D\varphi(I_{l^*}x_{k^*})\|_{l^*}
\end{aligned}$$

for all $l \geqq l_3$. This yields, together with (12),

$$\begin{aligned}
R\|\varphi'_l(x^l_{k^*})\| &\geqq \|\varphi'_l(x^l_{k^*})R_l\|_l \\
&\geqq \|D\varphi(I_{l^*}x_{k^*})\|_l - \tfrac{1}{4}\|D\varphi(I_{l^*}x_{k^*})\|_{l^*} \\
&\geqq \tfrac{3}{4}\|D\varphi(I_{l^*}x_{k^*})\|_{l^*}
\end{aligned} \tag{18}$$

since $E^*_l \supset E^*_{l^*}$. Therefore

$$\begin{aligned}
\rho^l_{k^*}\varphi'_l(x^l_{k^*})s^l_{k^*} &= \sigma_{k^*}\|\varphi'_l(x^l_{k^*})\|\varphi'_l(x^l_{k^*})s^l_{k^*} \\
&\geqq \sigma\gamma_{k^*}\|\varphi'_l(x^l_{k^*})\|^2 \\
&\geqq \frac{3}{4}\frac{\sigma\gamma}{R}\|D\varphi(I_{l^*}x_{k^*})\|_{l^*}\|\varphi'_l(x^l_{k^*})\|
\end{aligned} \tag{19}$$

and

$$\begin{aligned}
\rho^l_{k^*}\varepsilon^* &= \sigma_{k^*}\|\varphi'_l(x^l_{k^*})\|\frac{\gamma\sigma}{\sigma^* R}\|D\varphi(I_{l^*}x_{k^*})\|_{l^*} \\
&\leqq \frac{\gamma\sigma}{R}\|\varphi'_l(x^l_{k^*})\| \, \|D\varphi(I_{l^*}x_{k^*})\|_{l^*}.
\end{aligned} \tag{20}$$

Summing up, we get for each $\lambda \leqq \delta/(\sigma^* MI)$ and $l \geqq l_3$ a $\lambda^* \in (0, \lambda)$ with

$$\begin{aligned}
&\varphi(I_l(x^l_{k^*} - \lambda\rho^l_{k^*}s^l_{k^*})) - \varphi(I_l x^l_{k^*}) + \tfrac{1}{2}\lambda\rho^l_{k^*}\varphi'_l(x^l_{k^*})s^l_{k^*} \\
&\quad = -\lambda\rho^l_{k^*}D\varphi(I_l(x^l_{k^*} - \lambda^*\rho^l_{k^*}s^l_{k^*}))I_l s^l_{k^*} + \tfrac{1}{2}\lambda\rho^l_{k^*}\varphi'_l(x^l_{k^*})s^l_{k^*} \\
&\quad \leqq -\tfrac{1}{2}\lambda\rho^l_{k^*}\varphi'_l(x^l_{k^*})s^l_{k^*} + \tfrac{1}{4}\lambda\rho^l_{k^*}\varepsilon^* \\
&\quad \leqq -\tfrac{1}{8}\lambda\frac{\sigma\gamma}{R}\|D\varphi(I_{l^*}x_{k^*})\|_{l^*}\|\varphi'_l(x^l_{k^*})\| \\
&\quad \leqq -\tfrac{3}{32}\frac{\lambda\sigma\gamma}{R^2}\|D\varphi(I_{l^*}x_{k^*})\|^2_{l^*}.
\end{aligned} \tag{21}$$

Now choose a $j^* \in \mathbb{N}$ with $\alpha_{j^*} \leqq \delta/(\sigma^* MI)$ and define

$$\varepsilon := \tfrac{1}{32}\alpha_{j^*}\sigma\gamma\frac{1}{R^2}\|D\varphi(I_{l^*}x_{k^*})\|^2_{l^*}.$$

The convergence of the approximations of φ is uniform on

$$S(\delta, I_{l^*}x_{k^*}) := \{u \in E : \|u - I_{l^*}x_{k^*}\| \leqq \delta\}.$$

There is an $l_4 \geqq l_3$ with

$$|\varphi_l(z_l) - \varphi(I_l z_l)| < \varepsilon \tag{22}$$

for all $l \geqq l_4$, $l_l z_l \in S(\delta, I_{l^*}x_{k^*})$, $\varphi_l(z_l) \in \Phi_l(z_l)$. Let $l \geqq l_4$. Since

$$I_l x^l_{k^*} = I_{l^*}x_{k^*} \in S(\delta, I_{l^*}x_{k^*}) \quad \text{and} \quad I_l(x^l_{k^*} - \alpha_{j^*}\rho^l_{k^*}s^l_{k^*}) \in S(\delta, I_{l^*}x_{k^*}),$$

estimate (22) is valid for these points. Finally we get from (21) and (22)

$$\begin{aligned}\varphi_l(x^l_{k^*} - \alpha_{j^*}\rho^l_{k^*}s^l_{k^*}) - \varphi_l(x^l_{k^*}) + \tfrac{1}{2}\alpha_{j^*}\rho^l_{k^*}\varphi'_l(x^l_{k^*})s^l_{k^*} \\ \leqq \varphi(I_l(x^l_{k^*} - \alpha_{j^*}\rho^l_{k^*}s^l_{k^*})) - \varphi(I_l x^l_{k^*}) + \tfrac{1}{2}\alpha_{j^*}\rho^l_{k^*}\varphi'_l(x^l_{k^*})s^l_{k^*} + 2\varepsilon \\ < 0,\end{aligned}$$

or $\varphi_l(x^l_{k^*} - \alpha_{j^*}\rho^l_{k^*}s^l_{k^*}) < \varphi_l(x^l_{k^*}) - \frac{1}{2}\alpha_{j^*}\rho^l_{k^*}\gamma_{k^*}\|\varphi'_l(x^l_{k^*})\|$ for all $l \geqq l_4$. This contradicts assumption (14). Q.E.D.

5. Global convergence for the unconstrained problem. Now consider the unconstrained minimization problem (1), i.e. the case $C = E$. This implies $\varphi^n = \varphi$ for all n. It is our aim to develop a global convergence theorem in the sense that a critical point of φ is approximated. But we cannot ensure that we get for each $\varepsilon > 0$ a $k \in \mathbb{N}$ with $\|D\varphi(I_{q(k)}x_k)\| < \varepsilon$, since we have no assumption of the kind $\overline{\bigcup_l E^*_l} = E$. It is only possible to show the following global convergence result:

THEOREM 5.1. *Let $\{x_k\}$ be an infinite sequence constructed by Algorithm 3.3 with $x_k \in E_{q(k)}$ and $\|D\varphi(I_{q(k)}x_k)\|_{q(k)} > 0$ for all k. Assume that there is a bounded, closed subset L of E with*:

a) *$I_{q(k)}x_k \in L$ for all k.*

b) *φ is bounded below on L.*

c) *$D\varphi$ is bounded on L.*

d) *$D\varphi$ is uniformly continuous on $U(L) := \{y \in E : d(y, L) \leqq \sigma^* IM\}$, where M is the bound defined by Lemma 4.3 with respect to L.*

Then there exists for each $\varepsilon > 0$ a $k \in \mathbb{N}$ with

$$\|D\varphi(I_{q(k)}x_k)\|_{q(k)} \leqq \varepsilon. \tag{23}$$

Proof. Algorithm 3.3 constructs iterates $x_k \in E_{q(k)}$ with

$$\varphi_{q(k+1)}(x_{k+1}) \leqq \varphi_{q(k+1)}(x_k) - \tfrac{1}{2}\alpha_j\rho^{q(k+1)}_k\gamma_k\|\varphi'_{q(k+1)}(x^{q(k+1)}_k)\|,$$

where $0 \leqq j \leqq q(k+1)$, $q(k+1) > q(k)$. Since $I_{q(k+1)}x^{q(k+1)}_k = I_{q(k)}x_k \in L$, Lemma 4.3. guarantees the existence of an $M > 0$ and a $k_1 \in \mathbb{N}$ with

$$\|\varphi'_{q(k+1)}(x^{q(k+1)}_k)\| \leqq M \tag{24}$$

for all $k \geqq k_1$.

The following assumption will lead to a contradiction: There is an $\varepsilon > 0$ such that for all $k \in \mathbb{N}$

$$\|D\varphi(I_{q(k)}x_k)\|_{q(k)} > \varepsilon. \tag{25}$$

A trivial consequence of (25) is $\|D\varphi(I_{q(k)}x_k)\| > \varepsilon$ for all k. $D\varphi$ is uniformly continuous on $U(L)$. There is a $\delta > 0$ with

$$\|D\varphi(x) - D\varphi(y)\| < \frac{1}{8}\frac{\gamma\varepsilon}{IR} \tag{26}$$

for all $x, y \in U(L)$ with $\|x - y\| \leqq \delta$. Denote $l := q(k+1)$ and choose $\lambda \leqq \min(1, \delta/(I\sigma^* M))$. Since $I_l x_k^l = I_{q(k)} x_k \in L$, we have

$$\begin{aligned} d(I_l(x_k^l - \lambda^* \rho_k^l s_k^l), L) &\leqq \|I_l(x_k^l - \lambda^* \rho_k^l s_k^l) - I_l x_k^l\| \\ &\leqq \lambda^* \rho_k^l I \\ &\leqq \lambda \sigma^* I \|\varphi_l'(x_k^l)\| \\ &\leqq \lambda \sigma^* I M \leqq \min(I\sigma^* M, \delta) \end{aligned} \tag{27}$$

for all $k \geqq k_1$, $0 < \lambda^* < \lambda$. Therefore

$$I_l(x_k^l - \lambda^* \rho_k^l s_k^l) \in U(L) \quad \text{and} \quad \|I_l(x_k^l - \lambda^* \rho_k^l s_k^l) - I_l x_k^l\| \leqq \delta. \tag{28}$$

Estimate (26) yields

$$\|D\varphi(I_l x_k^l) - D\varphi(I_l(x_k^l - \lambda^* \rho_k^l s_k^l))\| < \frac{1}{8}\frac{\gamma\varepsilon}{IR} \tag{29}$$

for all $k \geqq k_1$, $0 < \lambda^* < \lambda$. From Assumption 3.2 we get a $k_2 \geqq k_1$ with

$$|\varphi_l'(x_k^l) s_k^l - D\varphi(I_l x_k^l) I_l s_k^l| < \frac{1}{8}\frac{\gamma\varepsilon}{R} \tag{30}$$

for all $k \geqq k_2$, $l := q(k+1)$. Furthermore we conclude from Lemma 4.2 that there is a $k_3 \geqq k_2$ with

$$\begin{aligned} \|\varphi_l'(x_k^l) R_l\|_l &= \|\varphi_l'(R_l I_{q(k)} x_k) R_l\|_l \\ &\geqq \|D\varphi(I_{q(k)} x_k)\|_l - \|\varphi_l'(x_k^l) R_l - D\varphi(I_l x_k^l)\|_l \\ &\geqq \tfrac{3}{4}\varepsilon, \end{aligned} \tag{31}$$

for all $k \geqq k_3$, $l := q(k+1)$. This implies

$$\|\varphi_l'(x_k^l)\| \geqq \frac{1}{R}\|\varphi_l'(x_k^l) R_l\| \geqq \frac{1}{R}\|\varphi_l'(x_k^l) R_l\|_l \geqq \frac{3}{4}\frac{\varepsilon}{R} \tag{32}$$

for all $k \geqq k_3$, $l := q(k+1)$. Combining (29) with (30) we get for all $k \geqq k_3$, $l := q(k+1)$, and $0 < \lambda^* < \lambda \leqq \min(1, \delta/(I\sigma^* M))$

$$\begin{aligned} &|D\varphi(I_l(x_k^l - \lambda^* \rho_k^l s_k^l)) I_l s_k^l - \varphi_l'(x_k^l) s_k^l| \\ &\quad \leqq I \|D\varphi(I_l(x_k^l - \lambda^* \rho_k^l s_k^l)) - D\varphi(I_l x_k^l)\| + |D\varphi(I_l x_k^l) I_l s_k^l - \varphi_l'(x_k^l) s_k^l| \\ &\quad < \frac{1}{4}\frac{\gamma\varepsilon}{R} \end{aligned}$$

or

$$D\varphi(I_l(x_k^l - \lambda^* \rho_k^l s_k^l)) I_l s_k^l \geqq \varphi_l'(x_k^l) s_k^l - \frac{1}{4}\frac{\gamma\varepsilon}{R}. \tag{33}$$

Furthermore

$$\varphi_l'(x_k^l) s_k^l \geqq \gamma_k \|\varphi_l'(x_k^l)\| \geqq \frac{3}{4}\frac{\gamma\varepsilon}{R}. \tag{34}$$

For each $\lambda \leqq \min(1, \delta/(I\sigma^* M))$, $k \geqq k_3$, $l := q(k+1)$ there is a $\lambda^* \in (0, \lambda)$ such that the following estimates are valid:

$$\begin{aligned}
&\varphi(I_l(x_k^l - \lambda\rho_k^l s_k^l)) - \varphi(I_l x_k^l) + \tfrac{1}{2}\lambda\rho_k^l \varphi_l'(x_k^l)s_k^l \\
&\quad = -\lambda\rho_k^l D\varphi(I_l(x_k^l - \lambda^*\rho_k^l s_k^l))I_l s_k^l + \tfrac{1}{2}\lambda\rho_k^l\varphi_l'(x_k^l)s_k^l \\
&\quad \leqq -\lambda\rho_k^l\left(\tfrac{1}{2}\varphi_l'(x_k^l)s_k^l - \tfrac{1}{4}\frac{\gamma\varepsilon}{R}\right) && \text{see (33)} \\
&\quad \leqq -\lambda\rho_k^l\left(\tfrac{3}{8}\frac{\gamma\varepsilon}{R} - \tfrac{1}{4}\frac{\gamma\varepsilon}{R}\right) && \text{see (34)} \\
&\quad = -\tfrac{1}{8}\lambda\sigma_k\|\varphi_l'(x_k^l)\|\frac{\gamma\varepsilon}{R} \\
&\quad \leqq -\tfrac{3}{32}\lambda\sigma\gamma\varepsilon^2/R^2 && \text{see (32)} \\
&\quad < -\tfrac{1}{16}\lambda\sigma\gamma\varepsilon^2/R^2.
\end{aligned} \tag{35}$$

Now choose a $j \in \mathbb{N}_0$ with $\alpha_j \leqq \min(1, \delta/(\sigma^* IM))$ and define

$$\varepsilon^* := \tfrac{1}{32}\alpha_j\sigma\gamma\varepsilon^2/R^2.$$

Assumption 3.2 shows there is an $l_{\varepsilon^*} \in \mathbb{N}$ with

$$|\varphi_r(z_r) - \varphi(I_r z_r)| \leqq \varepsilon^* \tag{36}$$

for all $r \geqq l_{\varepsilon^*}$, $z_r \in E_r$ with $I_r z_r \in U(L)$, and $\varphi_r(z_r) \in \Phi_r(z_r)$. Consider a $k_4 \geqq k_3$ such that $q(k+1) \geqq \max(l_{\varepsilon^*}, j)$ for all $k \geqq k_4$. Then

$$\begin{aligned}
&I_l(x_k^l - \alpha_j\rho_k^l s_k^l) \in U(L), \qquad |\varphi_l(x_k^l) - \varphi(I_l x_k^l)| \leqq \varepsilon^*, \\
&|\varphi_l(x_k^l - \alpha_j\rho_k^l s_k^l) - \varphi(I_l(x_k^l - \alpha_j\rho_k^l s_k^l))| \leqq \varepsilon^*
\end{aligned} \tag{37}$$

for all $k \geqq k_4$, $l := q(k+1)$. Defining $\lambda := \alpha_j$ the following estimates are easily verified:

$$\begin{aligned}
&\varphi_l(x_k^l - \alpha_j\rho_k^l s_k^l) - \varphi_l(x_k^l) + \tfrac{1}{2}\alpha_j\rho_k^l\varphi_l'(x_k^l)s_k^l \\
&\quad \leqq \varphi(I_l(x_k^l - \alpha_j\rho_k^l s_k^l)) - \varphi(I_l x_k^l) + \tfrac{1}{2}\alpha_j\rho_k^l\varphi_l'(x_k^l)s_k^l + 2\varepsilon^* \\
&\quad < 0 \quad \text{(see (37) and (35))} \quad \text{or} \\
&\quad \varphi_l(x_k^l - \alpha_j\rho_k^l s_k^l) \leqq \varphi_l(x_k^l) - \tfrac{1}{2}\alpha_j\rho_k^l\gamma_k\|\varphi_l'(x_k^l)\|.
\end{aligned} \tag{38}$$

Therefore j is feasible with respect to the j_k-computation of step 7 of Algorithm 3.3; more precisely $j_k \leqq j$ for all $k \geqq k_4$. Since $\alpha_{j_k} \geqq \alpha_j$, we conclude for $k \geqq k_4$ and $l := q(k+1)$,

$$\begin{aligned}
\varphi_l(x_{k+1}) - \varphi_l(x_k^l) &\leqq -\tfrac{1}{2}\alpha_{j_k}\rho_k^l\gamma_k\|\varphi_l'(x_k^l)\| \\
&\leqq -\tfrac{1}{2}\alpha_j\rho_k^l\gamma_k\|\varphi_l'(x_k^l)\| \\
&\leqq -\tfrac{1}{2}\alpha_j\sigma\gamma\|\varphi_l'(x_k^l)\|^2 \\
&\leqq -\tfrac{9}{32}\alpha_j\sigma\gamma\varepsilon^2/R^2 && \text{see (32)} \\
&= -9\varepsilon^*.
\end{aligned}$$

Applying (36) we get

$$\varphi(I_{q(k+1)}x_{k+1}) \leqq \varphi_l(x_{k+1}) + \varepsilon^* \leqq \varphi_l(x_k) - 8\varepsilon^* \leqq \varphi(I_l x_k^l) - 7\varepsilon^*$$

or

$$\varphi(I_{q(k+1)}x_{k+1}) \leqq \varphi(I_{q(k)}x_k) - 7\varepsilon^*$$

for all $k \geqq k_4$. Therefore we get

$$\lim_{k\to\infty} \varphi(I_{q(k)}x_k) = -\infty,$$

but the function φ is bounded below on L, leading to a contradiction. Q.E.D.

If L is a ball in E, then the uniform continuity of the Gateaux-differential operator $D\varphi$ implies the existence of the Fréchet-derivative in L; see for example Ljusternik and Sobolev [13, p. 310].

6. Global convergence for the constrained problem. Consider now the general problem (1) with restrictions and let $\{\varphi^n\}$ be a sequence of penalty functions. In § 4 we have seen that the loop from step 8 to step 4 is finite. Using the convergence result for the unconstrained case it is easy to show that $\lim_{k\to\infty} n(k) = \infty$:

THEOREM 6.1. *Let $\{x_k\}$ be an infinite sequence constructed by Algorithm 3.3 with $x_k \in E_{q(k)}$ and $\|D\varphi^{n(k+1)}(I_{q(k)}x_k)\|_{q(k)} > 0$ for all k. Assume that there is a bounded, closed subset L of E with:*

a) *$I_{q(k)}x_k \in L$ for all k.*

b) *φ^n is bounded below on L for each n.*

c) *$D\varphi^n$ is bounded on L for each n.*

d) *$D\varphi^n$ is uniformly continuous on $U_n(L) := \{y \in E : d(y, L) \leqq \sigma^* I M_n\}$, where M_n is the bound defined by Lemma 4.3 with respect to L, $n \in \mathbb{N}$.*

Then we have

$$\lim_{k\to\infty} n(k) = \infty \tag{39}$$

and there is a subsequence of $\{x_k\}$, i.e. an infinite subset S of $\mathbb{N}$ with

$$\lim_{k\in S} \|D\varphi^{n(k)}(I_{q(k)}x_k)\|_{q(k)} = 0, \tag{40}$$

$$\lim_{k\in S} d(I_{q(k)}x_k, C) = 0. \tag{41}$$

Proof. Assume that there is a $k^* \in \mathbb{N}$ with

$$\rho_k^l > \mu_{n^*}$$

for all $k \geqq k^*$, $l := q(k)$, $n^* := n(k^*)$. Then Algorithm 3.3 is a procedure for the unconstrained minimization of the function φ^{n^*} with initial values $q(k^*)$, $n(k^*)$, and x_{k^*}. Since all assumptions of Theorem 5.1 are satisfied, we conclude from the convergence statement that there is a $k_0 \geqq k^*$ with

$$\|D\varphi^{n^*}(I_{q(k_0)}x_{k_0})\|_{q(k_0)} < \mu_{n^*}/(2\sigma^* I). \tag{42}$$

Lemma 4.2 yields a $k_1 \geqq k^*$ with

$$\|\varphi_l^{n^{*\prime}}(R_l I_l x_k)R_l - D\varphi^{n^*}(I_l x_k)\|_l < \mu_{n^*}/(2\sigma^* I)$$

for all $k \geqq k_1$, $l := q(k)$. This implies for each $k \geqq k_1$, $l := q(k)$

$$\begin{aligned}\rho_k^l = \sigma_k\|\varphi_l^{n^{*\prime}}(x_k)\| &\leqq \sigma^* I\|\varphi_l^{n^{*\prime}}(x_k)R_l\|_l \qquad \text{(see (12))}\\ &\leqq \sigma^* I(\|\varphi_l^{n^{*\prime}}(x_k)R_l - D\varphi^{n^*}(I_l x_k)\|_l + \|D\varphi^{n^*}(I_l x_k)\|_l)\\ &< \sigma^* I(\mu_{n^*}/(2\sigma^* I) + \|D\varphi^{n^*}(I_l x_k)\|_l).\end{aligned}$$

Without loss of generality we assume that $k_0 \geqq k_1$, since (42) is valid for infinitely many k. This implies

$$\rho_{k_0}^{q(k_0)} < \mu_{n^*}$$

and leads to a contradiction; we get $\lim_{k\to\infty} n(k) = \infty$. Let $\varepsilon > 0$ be chosen arbitrarily. Step 3 of Algorithm 3.3 and condition (7) ensure that there is a k_0 with

$$\|\varphi_{q(k)}^{n(k)'}(x_k)\| < \varepsilon/(2R)$$

for infinitely many $k \geqq k_0$. From Lemma 4.2 it follows that there is a $k_1 \geqq k_0$ with

$$\|\varphi_{q(k)}^{n'}(R_{q(k)} I_{q(k)} x_k) R_{q(k)} - D\varphi^n(I_{q(k)} x_k)\|_{q(k)} < \varepsilon/2$$

for all $k \geqq k_1$, $n \in \mathbb{N}$. Therefore we get a $k \geqq k_1$ with

$$\begin{aligned} &\|D\varphi^{n(k)}(I_{q(k)} x_k)\|_{q(k)} \\ &\quad \leqq \|\varphi_{q(k)}^{n(k)'}(x_k) R_{q(k)} - D\varphi^{n(k)}(I_{q(k)} x_k)\|_{q(k)} \\ &\qquad\qquad + R\|\varphi_{q(k)}^{n(k)'}(x_k)\| \\ &\quad < \varepsilon/2 + R\varepsilon/(2R) \\ &\quad = \varepsilon. \end{aligned}$$

This shows statement (40).

In order to show (41), assume that there is an $\varepsilon > 0$ and an infinite subset S' of S with

$$d(I_{q(k)} x_k, C) > \varepsilon \tag{43}$$

for all $k \in S'$. Let

$$C_\varepsilon := \{y \in E : d(y, C) \leqq \varepsilon\}.$$

Since $\{\varphi^n\}$ is a sequence of penalty functions (see Definition 3.1) we know that there is a $\delta_\varepsilon > 0$ and l_ε, $n_\varepsilon \in N$ with

$$\|D\varphi^n(u)\|_l \geqq \delta_\varepsilon$$

for all $l \geqq l_\varepsilon$, $n \geqq n_\varepsilon$, and $u \in L \backslash C_\varepsilon$. Now choose a $\bar{k} \in S'$ with $q(\bar{k}) \geqq l_\varepsilon$ and $n(\bar{k}) \geqq n_\varepsilon$. From $I_{q(\bar{k})} x_{\bar{k}} \in L \backslash C_\varepsilon$ we conclude for all $k \geqq \bar{k}$, $k \in S'$

$$\|D\varphi^{n(k)}(I_{q(k)} x_k)\|_{q(k)} \geqq \delta_\varepsilon,$$

but this contradicts (40). Q.E.D.

Since the statement of the last theorem is only concerned with the penalty functions φ^n, we need some stronger assumptions to get convergence results for minimizing the original function φ. Let us first define

$$C_\varepsilon := \{u \in E : d(u, C) \leqq \varepsilon\}$$

for each $\varepsilon > 0$.

THEOREM 6.2. *Let C be bounded, $\overline{\bigcup_l E_l^*} = E$ and φ be uniformly continuous on C_ε for some $\varepsilon > 0$. Furthermore let $\{\varphi^n\}$ be a sequence of convex penalty functions with $\varphi^n(u) \geqq \varphi(u)$ for all $u \notin C$, $\lim_{n\to\infty} \varphi^n(u) = \varphi(u)$ uniformly on C and $\{x_k\}$ an iteration sequence of Algorithm 3.3 both satisfying the assumptions of Theorem 6.1. Then there is an infinite subset S of $\mathbb{N}$ with*

$$\lim_{k\in S} \varphi_{q(k)}^{n(k)}(x_k) = \inf_{u\in C} \varphi(u) =: m. \tag{44}$$

If there exists a minimizing point with respect to each φ^n, then

$$\lim_{n\to\infty}\min_{u\in E}\varphi^n(u)=m \quad \textit{and} \quad \underline{\lim}_{k\to\infty}\varphi^{n(k)}_{q(k)}(x_k)=m. \tag{45}$$

Proof. Consider a sequence $y_k\in C$ with

$$\lim_{k\to\infty}\varphi(y_k)=\inf_{u\in C}\varphi(u)=m. \tag{46}$$

Since $\overline{\bigcup_l E_l^*}=E$, we are able to choose $y_k\in E^*_{l(k)}$ with suitable $l(k)\in\mathbb{N}$. Let S be an infinite subset of $\mathbb{N}$ defined by Theorem 6.1. Because $\lim_{k\in S}d(I_{q(k)}x_k, C)=0$ and $\lim_{k\to\infty}q(k)=\infty$, we assume without loss of generality that $I_{q(k)}x_k\in C_\varepsilon$ and $q(k)\geqq l(k)$ for all $k\in S$. The convexity of the penalty functions implies

$$\varphi^n(v)-\varphi^n(u)\geqq D\varphi^n(u)(v-u) \tag{47}$$

for all $u, v\in E$ and for all n. Since $\lim_{k\in S}d(I_{q(k)}x_k, C)=0$, we get for each k a $w_k\in C$ with

$$\lim_{k\in S}\|I_{q(k)}x_k-w_k\|=0. \tag{48}$$

Therefore we have

$$\begin{aligned} m=\inf_{u\in C}\varphi(u) &\leqq \underline{\lim}_{k\in S}\varphi(w_k)\\ &\leqq \underline{\lim}_{k\in S}\varphi(I_{q(k)}x_k)\\ &\leqq \underline{\lim}_{k\in S}\varphi^{n(k)}(I_{q(k)}x_k)\\ &\leqq \overline{\lim}_{k\in S}\varphi^{n(k)}(I_{q(k)}x_k). \end{aligned} \tag{49}$$

The third inequality follows from $\varphi(u)\leqq\varphi^{n(k)}(u)$ for $u\notin C$ and the uniform convergence of $\{\varphi^n\}$ in C. Inequality (47) yields for all $k\in S$

$$\varphi^{n(k)}(I_{q(k)}x_k)\leqq\varphi^{n(k)}(y_k)-D\varphi^{n(k)}(I_{q(k)}x_k)(y_k-I_{q(k)}x_k).$$

Since $y_k\in E^*_{l(k)}\subset E^*_{q(k)}$, we get

$$\begin{aligned} &|D\varphi^{n(k)}(I_{q(k)}x_k)(y_k-I_{q(k)}x_k)|\\ &\qquad\leqq\|D\varphi^{n(k)}(I_{q(k)}x_k)\|_{q(k)}\|y_k-I_{q(k)}x_k\|. \end{aligned}$$

The boundedness of $\{y_k\}$ and $\{I_{q(k)}x_k\}$ and statement (40) imply

$$\overline{\lim}_{k\in S}\varphi^{n(k)}(I_{q(k)}x_k)\leqq\overline{\lim}_{k\in S}\varphi^{n(k)}(y_k)\leqq\lim_{k\in S}\varphi(y_k)=m.$$

The last estimate results from the fact that $y_k\in C$ and from the uniform convergence of $\{\varphi^n\}$ in C. Therefore we conclude, together with (49),

$$\lim_{k\in S}\varphi^{n(k)}(I_{q(k)}x_k)=m.$$

Equation (44) follows then from Assumption 3.2. It is presumed that there are $z_n\in E$ with

$$\varphi^n(z_n)=\min_{z\in E}\varphi^n(z)=:m_n.$$

Condition (3) of Definition 3.1 indicates that $\lim_{n\to\infty} d(z_n, C)=0$. There is a sequence $u_n \in C$ with $\lim_{n\to\infty}\|u_n - z_n\| = 0$. The convergence of $\{\varphi^n\}$ to φ in C gives $\varepsilon_n > 0$, $\delta_k > 0$ with $\lim_{n\to\infty}\varepsilon_n = 0$, $\lim_{k\to\infty}\delta_k = 0$ such that

$$m \geqq \varphi(y_k) - \delta_k \geqq \varphi^n(y_k) - \varepsilon_n - \delta_k \geqq m_n - \varepsilon_n - \delta_k$$

for all $n, k \in \mathbb{N}$. By picking n large and then k large, we conclude that

$$m \geqq \overline{\lim_{n\to\infty}}\, m_n \geqq \underline{\lim}_{n\to\infty}\, m_n = \underline{\lim}_{n\to\infty}\, \varphi^n(z_n) \geqq \underline{\lim}_{n\to\infty}\, \varphi(z_n).$$

The last inequality follows from the fact that $\varphi^n(z_n) \geqq \varphi(z_n)$, if $z_n \notin C$, and from the uniform convergence of $\{\varphi^n\}$ in C. From

$$\underline{\lim}_{n\to\infty}\, \varphi(z_n) \geqq \underline{\lim}_{n\to\infty}\, \varphi(u_n) \geqq m$$

we get

$$\lim_{n\to\infty} m_n = m.$$

Therefore it is not possible that there is any accumulation point of $\{\varphi^{n(k)}(I_{q(k)}x_k)\}$ below m. This shows statement (45) using Assumption 3.2 again. Q.E.D.

REFERENCES

[1] L. Armijo, *Minimization of functions having Lipschitz-continuous first partial derivatives*, Pacific J. Math., 16 (1966), pp. 1–3.

[2] M. C. Biggs, *A new method of constrained minimisation using recursive equality quadratic programming*, Technical Report No. 24, The Hatfied Polytechnic, Numerical Optimisation Centre, Hatfield, Great Britain, 1972.

[3] R. Bulirsch and J. Stoer, *Numerical treatment of ordinary differential equations by extrapolation methods*, Numer. Math., 8 (1966), pp. 1–13.

[4] H. Esser, *Zur Diskretisierung von Extremalproblemen*, Springer Lecture Notes in Mathematics, No. 333, Springer-Verlag, Berlin, 1973.

[5] P. E. Gill, *The design and implementation of software for unconstrained optimization*, Paper presented at the NATO Advanced Study Institute on "The design and implementation of optimization software", SOGESTA, Urbino, Italy, 1977.

[6] W. H. Hager, *Rates of convergence for discrete approximations to unconstrained control problems*, SIAM J. Numer. Anal., 13 (1976), pp. 449–472.

[7] D. G. Hull, *Application of parameter optimization methods to trajectory optimization*, AIAA Mechanics and Control of Flight Conference (Anaheim, CA, 1974).

[8] I. L. Johnson, *Optimization of the solid-rocket assisted space shuttle ascent trajectory*, J. Spacecraft, 12 (1975), pp. 765–769.

[9] R. Klessig and E. Polak, *A method of feasible directions using function approximations with applications to minimax problems*, J. Math. Anal. Appl., 41 (1973), pp. 583–602.

[10] ———, *An adaptive precision gradient method for optimal control*, this Journal, 11 (1973), pp. 80–93.

[11] D. Kraft, *Optimierung von Flugbahnen mit Zustandsbeschränkungen durch mathematische Programmierung*, 9. DGLR-Jahrestagung (München, 1976), DGLR-JB, 1976, pp. 201-1, 201-23.

[12] S. J. Leese, *Convergence of gradient methods for optimal control problems*, J. Optimization Theory Appl., 21 (1977), pp. 329–338.

[13] L. A. Ljusternik and W. I. Sobolev, *Elemente der Funktionalanalysis*, Akademie-Verlag, Berlin, 1968.

[14] B. A. Murtagh and R. W. H. Sargent, *A constrained minimization method with quadratic convergence*, Optimization, R. Fletcher, ed., Academic Press, London-New York, 1969.

[15] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, Academic Press, New York-San Francisco-London, 1970.

[16] E. Polak, *Computational methods in optimization: A unified approach*, Academic Press, New York, 1971.

[17] E. Polak, R. W. H. Sargent, D. J. Sebastian, *On the convergence of sequential minimization algorithms*, J. Optimization Theory Appl., 14 (1974), pp. 439–442.

[18] R. W. H. Sargent and G. R. Sullivan, *The development of an efficient optimal control package*, Optimization Techniques, Proc. of the 8th IFIP Conf. on Optimization Techniques (Würzburg, 1977), Part 2, Springer-Verlag, Berlin-Heidelberg-New York, 1978, pp. 158–168.

[19] K. Schittkowski, *Algorithmen und Konvergenzsätze für Systeme nichtlinearer Funktionen mit Werten in einger konvexen Menge*, Dissertation, Universität Würzburg, Würzburg, West Germany, 1975.

[20] ———, *A global minimization algorithm in a normed linear space using function approximations with adaptive precision*, Preprint No. 16, Institut für Angewandte Mathematik und Statistik, Universität Würzburg, Würzburg, West Germany, 1976.

[21] ———, *A steepest descent method for the numerical solution of variational problems using adaptive precision*, Preprint No. 17, Institut für Angewandte Mathematik und Statistik, Universität Würzburg, Würzburg, West Germany, 1976.

[22] ———, *Über Minimierungsverfahren bei approximativer Kenntnis der Zielfunktion angewandt auf die numerische Lösung von Variationsproblemen*, Methods of Operations Research XXV, Symposium on Operations Research (Heidelberg, 1976), Verlag Anton Hain, Meisenheim am Glan, 1977.

[23] ———, *An adaptive precision method for the numerical solution of constrained optimization problems applied to a time-optimal heating process*, Optimization Techniques, Proc. of the 8th IFIP Conf. on Optimization Techniques (Würzburg, 1977), Part 2, Springer-Verlag, Berlin-Heidelberg-New York, 1978, pp. 125–135.

[24] F. Stummel, *Weak stability and weak discrete convergence of continuous mappings*, Numer. Math., 26 (1976), pp. 301–316.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 1, 1979

SEARCH GAMES WITH MOBILE AND IMMOBILE HIDER*

SHMUEL GAL†

Abstract. We consider search games in which the searcher moves along a continuous trajectory in a set Q until he captures the hider, where Q is either a network or a two (or more) dimensional region. We distinguish between two types of games; in the first type which is considered in the first part of the paper, the hider is immobile while in the second type of games which is considered in the rest of the paper, the hider is mobile. A complete solution is presented for some of the games, while for others only upper and lower bounds are given and some open problems associated with those games are presented for further research.

1. Introduction. The work on search problems which are considered in this paper has been developed in two directions. The first direction was initiated by Bellman [6], who introduced the following search problem: An immobile hider is located on the real line according to a known probability distribution. A searcher who can move in unit velocity wishes to discover the hider in minimal expected time. It is assumed that the searcher cannot see the hider until he actually reaches the point in which the hider is located and the time elapsed until this moment is the duration of the search.

This problem has been considered by Beck in [2] and [3] and by Franck in [9]. They found some properties of the optimal search trajectory, but a complete solution to the problem as presented in [6] has not yet been found.

In their paper [4], Beck and Newman presented a solution for the search on the real line considered as a game. In this game the hider can choose his hiding point using any probability distribution whose first absolute moment is bounded by a known constant, while the searcher can move along any continuous trajectory. This author considered in [12], and in [13] together with Chezan, some games of this type played in other regions, e.g., a set of rays or a plane and proved that the minimax search trajectories are of the exponential type.

Some other variations of this problem were presented by Beck and Warren [5] and by Fristedt and Heath [10] who used a different cost function, and by McCabe in [15] who considered the search for a random walker. All the above-mentioned problems belonging to the first class consider search in an unbounded region for a hider who is usually immobile. The cases of a mobile hider presented in [15] and in § 4.3 of [13] consider a motion of a very specific type.

The second direction of research was initiated by the search games presented by Isaacs in his book [14, pp. 345–350] and especially by his Princess and Monster game. In this search game, both the searcher (the Monster) and the hider (the Princess) can move in a bounded region Q. Each of the players cannot see the other until the distance between them is less than r, and at this very moment capture occurs. As a stepping stone for the case where Q is a general region, Isaacs proposed the simpler problem where it is the boundary of a circle.

The problem where both the searcher and the hider can move on the boundary of a circle has been solved by Alpern [1], Foreman [7] and Zelikin [19]. Another version of this problem has been solved by Foreman in [8]. In [17] Wilson presented a solution of a discrete version of the problem. Worsham [18] described a numerical algorithm for solving discrete search problems of this type.

* Received by the editors November 14, 1977, and in revised form May 3, 1978.

† IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598 and IBM Israel Scientific Center, Technion City, Haifa, Israel.

In the present paper we shall consider search games of the following type: The search takes place in a bounded set Q which will be called *the searching set.* In some of the games Q will be a two (or higher dimensional) region, while in other cases Q will be a network. The searcher starts moving from a specified point O called the origin, with maximum velocity 1 (the time unit can be always chosen so as to normalize this maximal velocity). The maximal velocity of the hider is w. We shall distinguish between cases where w is equal to zero, i.e., an immobile hider, and cases where $w > 0$. The results we obtained are also valid for the case where the hider's velocity is unbounded. It is assumed that the searcher and the hider cannot see one another until their distance is less than or equal to the discovery radius r and at that very moment capture occurs. In cases where Q is a network, r will be taken as zero while for cases where Q has two or more dimensions, it will be assumed that r is very small in comparison with the "magnitude" of the set Q.

This problem is considered as a two-person zero-sum game with the loss of the searcher (or the gain of the hider) being the expected time elapsed until capture occurs. A pure strategy of the searcher is a trajectory inside Q starting from the origin O; a pure strategy of the hider is a point in Q if he is immobile and a trajectory inside Q if he is mobile. We shall denote a strategy of the searcher by s and a strategy of the hider by h. The strategies will usually be mixed which means that the user determines his strategy according to a probability distribution on a set of pure strategies. Let $v(s, h)$ be the expected capure time when the searcher uses strategy s and the hider uses h. The value of strategy s of the searcher is defined as

$$v(s) = \sup_h v(s, h) \tag{1}$$

and the value of the searcher as

$$\bar{v} = \inf_s v(s). \tag{2}$$

The value of strategy h of the hider is defined as

$$v(h) = \inf_s v(s, h) \tag{3}$$

and the value of the hider as

$$\underline{v} = \sup_h v(h) \qquad (\underline{v} \leqq \bar{v}). \tag{4}$$

Usually we shall have $\bar{v} = \underline{v} = v$, and in this case the value of the game is denoted by v.

An optimal (resp. ε-optimal) strategy of the searcher is a strategy s^* which satisfies for all h:

$$v(s^*, h) \leqq v \qquad (\text{resp.} \leqq v(1+\varepsilon)). \tag{5}$$

An optimal (resp. ε-optimal) strategy of the hider is a strategy h^* which satisfies for all s:

$$v(s, h^*) \geqq v \qquad (\text{resp.} \geqq v(1-\varepsilon)). \tag{6}$$

It should be noted that in order to prove the optimality of s^* it is sufficient to prove (5) only for pure strategies of the hider. A similar statement holds for h^*.

In this paper we present the optimal (or ε-optimal) strategies and the value for several search games belonging to the type previously described. The description of the results will be presented in the next chapter.

2. Description of the results. In order to describe the results we use some additional notations:

μ – The Lebesgue measure of the set Q. Thus if the players move on a network, μ is the sum of the lengths of all the arcs in Q, while if Q is a two or three dimensional region then μ is the area or the volume.

g—The maximal rate of discovery of the searcher, i.e., the maximal measure of points which can be detected in a time unit. Since the maximal velocity of the searcher is 1, it follows that for the case where Q is a network $g = 1$, for the case where Q is a two dimensional region, then the sweep width is $2r$ so that the maximal area g of the strip which can be detected in a unit time interval is $2r$; by a similar reasoning, g is defined as πr^2 for three dimensional regions, etc.

A general description of the results follows:

In § 3 we consider the case of an immobile hider. If Q is two dimensional then it has been demonstrated by Isaacs (see [14, pp. 345–346]), that the value of the game is approximately $\mu/(2g)$. We shall extend Isaacs' result to some networks: For an Eulerian network, the same result is true so that the value is $\frac{1}{2}(\mu/g)$ $(=\frac{1}{2}\mu$ because $g = 1$ for the case of a network).

We shall show that $\frac{1}{2}\mu$ is the lower bound for any network. The other extreme case which represents the upper bound occurs when Q is a tree. In this case we prove that the value is μ. An interesting intermediate case when Q consists of an uneven number of nonintersecting paths of the same length between two points is also considered.

The subsequent chapters are dedicted to search games with a mobile hider. In § 4 we solve such a problem for the case where Q is a set of k nonintersecting paths of equal length, connecting two points. We show that the value of this game is μ/g $(=\mu)$. This result which is an extension of the Princess and Monster game on the circle (see [1], [7], [17] and [19]) is the first step towards § 5 where we consider the case of searching for a mobile hider in a two (or more) dimensional region. We show there that if the searching region is convex then the value v of this search game satisfies $v \sim \mu/g$ and we present ε-optimal strategies for both players.

3. The case of an immobile hider.

3A. General description. The search games considered in this section will usually take place in a network Q, where in this paper a network will mean any connected finite set of arcs. An immobile hider chooses a hiding point in Q and the searcher chooses a searching trajectory s starting from the origin O. It can obviously be assumed that the searcher will move along s with maximal speed because any pure search strategy s_1 which does not use the maximal velocity is dominated by the pure strategy s_1' which uses the maximal velocity along the trajectory visited by s_1. The capture time will be the time elapsed from the beginning of the search until the moment the searcher reaches the hiding point.

It will be assumed that both the hider and the searcher can use mixed strategies: A mixed strategy of the hider is a distribution on Q, while a mixed strategy of the searcher is a probability distribution over a set of possible trajectories. Under the above-mentioned assumptions the search game has a value v. This can be shown as follows: Let $B = \{x_1, x_2, \cdots, x_n, \cdots\}$ be a denumerable set of points which is dense in Q. For all $n = 1, 2, \cdots$ consider the search game G_n in which the hider can choose any point in Q while the searcher can choose any trajectory that starts from the origin and move to a point $x_{i_1} \in \{x_1, \cdots, x_n\}$, then move to another point $x_{i_2} \in \{x_1, \cdots, x_n\}$ etc.

Since the hider is immobile, it is sufficient to take into account only a finite set of pure search strategies and thus G_n has a value v_n. It is obvious that v_n is a monotone decreasing sequence and thus has a limit v. We now show that v is the value of the original game described in this section: The searcher can guarantee a capture time not more than $v_n \leqq v+\varepsilon$ by using an optimal strategy for G_n where n is large enough. On the other hand since B is dense in Q, then any ε-optimal hiding strategy $h_n^*(\varepsilon)$ for G_n where n is large enough, is a 2ε-optimal strategy for the original game, because for any pure search strategy S for the original game, there exists a pure search strategy S_n for G_n such that $v(S, h_n^*(\varepsilon)) > v(S_n, h_n^*(\varepsilon)) - \varepsilon \geqq v_n - 2\varepsilon \geqq v - 2\varepsilon$. Thus, for any $\varepsilon > 0$ the hider can guarantee an expected capture time which exceeds $v - 2\varepsilon$.

3B. An Eulerian network. In our paper, an Eulerian network will mean a connected set of arcs which can be obtained by drawing a continuous *closed* curve S which does not go through any of the arcs more than once (but it may pass through a vertex several times). Such a curve will be called *an Eulerian curve*. A necessary and sufficient condition for a (connected) network Q to have such a property is that each of its vertices has an even degree (i.e. an even number of arcs attached to it).

We assume that the searcher has to start from the origin O, which might be any point in Q not necessarily a vertex, and has to move in Q until he finds the (immobile) hider. The solution of the search problem for Eulerian network can be obtained by the same method used by Isaacs [14, pp. 345–346]. The searcher can assure himself of an expected capture time not more than $\frac{1}{2}\mu$ (μ being the sum of the lengths of the arcs) by using the following strategy: choose any Eulerian curve S^*; there are two directions of encircling S^*; choose each of them with probability 1/2.

The hider can assure himself that the expected capture time will be at least $\frac{1}{2}\mu$ by using a uniform distribution on Q for the hiding point. A simple proof of the above-mentioned statements is given in [14]. These statements can be summed up by the following lemma:

LEMMA 1. *The value of the search game with an immobile hider for an Eulerian network is* $\mu/2$.

3C. General bounds. In this section we consider the case where Q can be any network which satisfies the assumptions of § 3A.

Let S^* be a *closed* curve which passes through all the points of Q having a minimal length (S^* may go more than once through some of the arcs). The length of S^* will be denoted by $L(S^*)$. The following result holds for any network Q:

LEMMA 2. *The value v of the search game with immobile hider satisfies*

$$\frac{1}{2}\mu \leqq v \leqq \frac{1}{2}L(S^*). \tag{7}$$

Proof. The left side of (7) can be proved as follows: the hider can keep the capture time to be at least $\frac{1}{2}\mu$ by using the strategy h_u which chooses the hiding point uniformly on Q. Then for any search trajectory s the rate of coverage of points not encountered before is at most 1 (the maximal velocity of the searcher) thus $v(s, h_u) \geqq \int_0^{\mu} (1/\mu)t\,dt = \mu/2$.

The right side of (7) can be achieved if the searcher chooses the strategy s_1 which chooses each of the two directions of encircling S^* with probability 1/2. Assume that the hider chooses a point h. Let the distance between the origin O and h alongside S^* be d_1 for one direction and d_2 for the second direction. Then clearly $d_1 + d_2 \leqq L(S^*)$; thus for any h

$$v(s_1, h) = \tfrac{1}{2}d_1 + \tfrac{1}{2}d_2 \leqq \tfrac{1}{2}L(S^*) \qquad \text{Q.E.D.}$$

It can be easily seen that for any network, $L(S^*) \leqq 2\mu$; thus Lemma 2 implies that

$$\tfrac{1}{2}\mu \leqq v \leqq \mu. \tag{8}$$

It has been shown in § 3B that for Eulerian networks the value v is equal to the leftside of (8). There exist networks such that v is equal to the rightside of (8) namely μ. A network which has such a property is a tree. This result will be established in the next section.

3D. Searching on a tree. In this section we consider the case when the set Q is a tree. By adding an extra node at the origin O (if it is not a node in the original tree) the tree can be always represented in such a way that the origin (i.e. the starting point) is at the root of the tree. We shall establish the following result:

THEOREM 1. *Let v be the value of the search game for an immobile hider in a tree Q. Then*

$$v = \mu. \tag{9}$$

Proof. It follows from the rightside of (8) that $v \leqq \mu$. We shall establish the fact that $v \geqq \mu$ by proving the following lemma:

LEMMA 3. *Consider two trees Q and Q' as depicted in Fig. 1. The only difference between Q and Q' is that two adjacent terminal branches DA_1 of length a_1 and DA_2 of length a_2, are replaced by one terminal branch DA' of length a_1+a_2. Let v be the value of the search game in Q and let v' be the value of the search game in Q'; then $v \geqq v'$.*

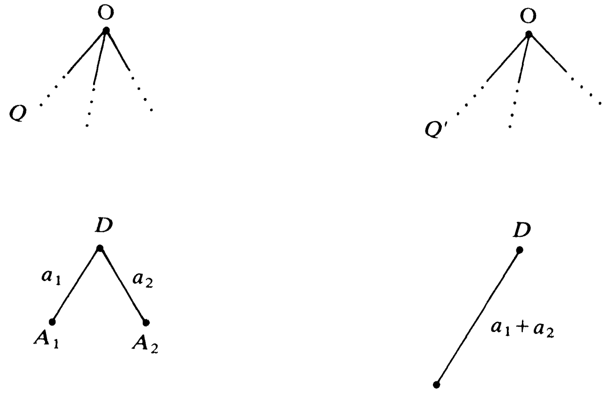

FIG. 1

Proof of Lemma 3. Any hiding strategy in a tree is obviously dominated by a hiding strategy which chooses points only among the terminal nodes of the tree and thus we can assume that an optimal hiding strategy h'^* has this property. Such a hiding strategy h'^* satisfies for any search trajectory s' in Q':

$$v(s', h'^*) \geqq v' \quad \text{(see (6)).} \tag{10}$$

Let us define a hiding strategy h^* in Q by attaching to each terminal node of Q, the same probability of h'^*, except that the probability of choosing the new nodes A_1 and A_2, which will be denoted by p_1 and p_2 respectively, is given by

$$p_1 = \frac{a_1}{a_1+a_2} p' \quad \text{and} \quad p_2 = \frac{a_2}{a_1+a_2} p' \tag{11}$$

where p' is the probability of choosing A' under h'^* and a_1, a_2 are depicted in Fig. 1.

We shall show that $v \geqq v'$ by proving that for any search trajectory s in Q

$$v(s, h^*) \geqq v'. \tag{12}$$

In order to prove (12) we proceed as follows: If the hider uses a strategy h^* which chooses its hiding point at terminal nodes only, then it is best for the searcher to use only search trajectories which have the following characteristics: It starts from the root O, moves in the shortest route to a terminal node, then moves in the shortest route to another terminal node and so on until all the terminal nodes have been visited. Therefore, there exist a one to one correspondence between the relevant search trajectories and the permutations of the terminal nodes. Bearing that in mind and assuming (for convenience) that the search strategy s visits the terminal node A_1 before visiting A_2 (the proof is similar for the case where A_2 is visited before A_1) then s can be represented by the following permutation of terminal nodes:

$$s \sim (A_{i_1}, \cdots, A_{i_I}, A_1, A_{j_1}, \cdots, A_{j_J}, A_2, A_{l_1}, \cdots, A_{l_L}). \tag{13}$$

Let us denote the distance from the origin O to the terminal node A_k along the trajectory s, by d_k and let h^* assign a probability p_k to the hiding point A_k; then inequality (12) is equivalent to

$$\sum_{m=1}^{I} d_{i_m} p_{i_m} + d_1 p_1 + \sum_{m=1}^{J} d_{j_m} p_{j_m} + d_2 p_2 + \sum_{m=1}^{L} d_{l_m} p_{l_m} \geqq v'. \tag{14}$$

In order to prove (14) let us consider two search trajectories in Q':

$$s_1' \sim (A_{i_1}, \cdots, A_{i_I}, A', A_{j_1}, \cdots, A_{j_J}, A_{l_1}, \cdots, A_{l_L}) \tag{15}$$

and

$$s_2' \sim (A_{i_1}, \cdots, A_{i_I}, A_{j_1}, \cdots, A_{j_J}, A', A_{l_1}, \cdots, A_{l_L}). \tag{16}$$

It follows from (10) that

$$v(s_1', h'^*) \geqq v' \tag{17}$$

and

$$v(s_2', h'^*) \geqq v'. \tag{18}$$

Let us denote the distance from the origin O to the terminal node A_k along the trajectory s_1' and s_2' by d_{1k} and d_{2k} respectively and the distance to the terminal node A' by d_1' and d_2' respectively. The following relations hold:

$$d_{1i_m} = d_{2i_m} = d_{i_m}, \tag{19}$$

$$d_1' = d_1 + a_2, \tag{20}$$

$$d_2' \leqq d_2 - a_1, \tag{21}$$

$$d_{1j_m} = d_{j_m} + 2a_2, \tag{22}$$

$$d_{2j_m} \leqq d_{j_m} - 2a_1, \tag{23}$$

$$d_{1l_m} \leqq d_{l_m}, \tag{24}$$

$$d_{2l_m} \leqq d_{l_m}. \tag{25}$$

It follows from (17) and (18) that

$$\frac{a_1}{a_1+a_2}v(s_1', h'^*)+\frac{a_2}{a_1+a_2}v(s_2', h'^*)\geqq v';$$

thus

$$\frac{a_1}{a_1+a_2}\left(\sum_{m=1}^{I} d_{1i_m}p_{i_m}+d_1'p'+\sum_{m=1}^{J} d_{1j_m}p_{j_m}+\sum_{m=1}^{L} d_{1l_m}p_{l_m}\right)$$
$$+\frac{a_2}{a_1+a_2}\left(\sum_{m=1}^{I} d_{2si_m}p_{i_m}+\sum_{m=1}^{J} d_{2j_m}p_{j_m}+d_2'p'+\sum_{m=1}^{L} d_{2l_m}p_{l_m}\right)\geqq v'.$$

Using (19)–(25) we obtain

$$(26)\qquad \sum_{m=1}^{I} d_{i_m}p_{i_m}+\sum_{m=1}^{J} d_{j_m}p_{j_m}+\sum_{m=1}^{L} d_{l_m}p_{l_m}+\frac{a_1}{a_1+a_2}d_1p'+\frac{a_1}{a_1+a_2}d_2p'\geqq v'$$

and now inequality (14) immediately follows from (11) and (26). Q.E.D.

Using Lemma 3 we can readily establish by induction on the number of terminal nodes, that the inequality $v \geqq \mu$ holds for any tree thus completing the proof of Theorem 1.

It is worth noting that Lemma 3 also presents, as a byproduct, a method for finding an optimal hiding strategy (an optimal strategy for the searcher is the one described in the proof of Lemma 2).

3E. The case of *k* parallel paths. In § 3C it has been proved that for any network Q, the value of the search game v satisfies $\frac{1}{2}\mu \leqq v \leqq \frac{1}{2}L(S^*)$ where $L(S^*)$ is the minimal length of a closed curve which passes through every point of Q. For an Eulerian network $L(S^*)=\mu$ while for a tree $L(S^*)=2\mu$ so that Lemma 1 and Theorem 1 establish the fact that for these two cases $v=\frac{1}{2}L(S^*)$. It might seem to the reader that this result holds for all networks but this is not true as will be demonstrated by the examples to be presented in this section.

Each network Q considered in this section consists of a set of k nonintersecting arcs, each of them of equal length a, which join two points O and A. Such networks will be encountered again in the next section which considers search games with a mobile hider but here we are concerned with an immobile hider. If the number of arcs k is even then Q is an Eulerian network and the solution to the search game is simple, but if k is an odd number greater than one, then the search game seems to be quite complicated. We shall not find the complete solution for these games but only establish the strict inequality

$$(27)\qquad v<\tfrac{1}{2}L(S^*)$$

which did not occur in our previous examples (where v is equal to $\frac{1}{2}L(S^*)$).

LEMMA 4. *If Q is a set of k nonintersecting arcs of equal length which join* O *and* A, *and k is odd, then the value of the search game v, satisfies* (27).

Proof. At first we note that any closed curve which visits any point of Q has to repeat itself on one of the arcs; thus $L(S^*)=(k+1)a$ so that (27) is equivalent to

$$(28)\qquad v<\frac{(k+1)a}{2}.$$

Inequality (28) is established by presenting the following mixed search strategy s': In this strategy the searcher starts from O, chooses each of the arcs with equal probability $(1/k)$ and moves to A; then he chooses each of the remaining arcs with equal probability $(1/(k-1))$ and moves back to O and so on until all the arcs have been visited. Let h be any pure strategy (i.e. a point in Q) and assume that the distance of this point from A is d so that the distance from O is $a-d$. Let E_i, $i=1,\cdots,\lfloor k/2\rfloor$ be the event that the hider will be discovered during the time period $(2(i-1)a, 2ia]$ and let E_f be the event that the hider will be discovered during the time period $((k-1)a, ka]$; then

$$\begin{aligned} v(s',h) &= \sum_{i=1}^{\lfloor k/2\rfloor} p(E_i)\cdot(2(i-1)a+a)+p(E_f)\cdot((k-1)a+a-d) \\ &= \sum_{i=1}^{\lfloor k/2\rfloor} \frac{2}{k}(2(i-1)a+a)+\frac{1}{k}((k-1)a+a-d) \qquad (29) \\ &= \left(\frac{k}{2}+\frac{1}{2k}\right)a-\frac{d}{k} \leqq \left(\frac{k}{2}+\frac{1}{2k}\right)a < \frac{k+1}{2}a \quad \text{Q.E.D.} \end{aligned}$$

The case where Q consists of an odd number of arcs joining O and A has been used in order to show that there are cases in which $v<\frac{1}{2}L(S^*)$, but this case is interesting by itself. It is amazing that the solution of the game is simple for any even k (and also, as will be demonstrated in the next chapter, for any k, odd or even, if the hider is mobile) but it may be quite complicated to solve this game for the case where k is equal only to three. The reasonable symmetric search strategy s' which was used in (29) can assure the searcher of losing no more than $(k/2+1/(2k))a$ but it can be shown that the hider cannot achieve (or even ε-achieve) this value. Thus, s' is not an optimal strategy for an odd number of arcs (in spite of its being optimal for an even number of arcs). It seems to us that the optimal strategies for the odd number case may be quite complicated.

4. Mobile hider on k parallel arcs.

4A. General description. In this section we consider the following search game: The searching set Q is a set of k nonintersecting arcs $b_1,\cdots,b_k$ joining two points O and A as depicted in Fig. 2, the length of each arc being equal to a. The searcher has to start moving from O with maximal velocity equal to unity. The hider can choose an arbitrary starting point and from this point he can move along any continuous trajectory in Q with maximal velocity w. In this section we shall usually assume that $w\geqq 1$. The capture time which is the loss of the searcher (or the gain of the hider) is the time elapsed until the searcher reaches a point which is occupied at the same time by the hider. For the case $k>2$ it will be assumed that the searcher can only pass but not stay at the points O or A, or alternatively that the hider can pass from any arc b_i to any other arc b_j not only through O or A but also in an ε distance from them. Such a

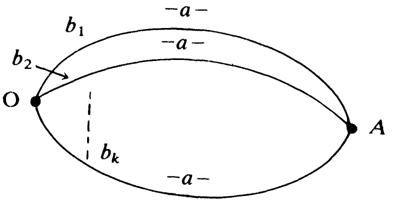

FIG. 2

search game is an extension of the Princess and Monster game on the circle which has been proposed as an open problem in [14, p. 350] and solved in the papers [1], [7], [17] and [19]. The game on the circle is a special case, namely $k=2$, of the game considered in this chapter. The search on k parallel arcs is interesting in itself and in addition will be used in order to obtain an approximation for the value of search games in higher dimensional sets.

4B. Solution of the game. We shall show that the optimal strategies of the searcher and the hider are both of the type $z(X, T)$ defined as follows

DEFINITION 1. Let X be either the point O or the point A and let $0 \leqq T < \infty$ be any real number; then the (random) trajectory $z(X, T)$ is defined by the following rules: At time T starting from point X, choose an integer i, $1 \leqq i \leqq k$ with each of the integers $\{1, 2, \cdots, k\}$ having the same probability of choice $(1/k)$ and move along the arc b_i with unit velocity until arriving to the end point of this arc (O or A); then (at time $T+a$) choose an integer j again with uniform probability on the integers $1, 2, \cdots, k$, independently on i, and move along b_j until reaching the other end point (A or O) and so on. Using Definition 1 we state the following theorem:

THEOREM 2. *For the search game presented in* § 4A: *An optimal strategy for the searcher is* $z(\mathrm{O}, 0)$. *An* ε*-optimal strategy for the hider is to start moving with maximal speed to the point* A, *stay there until time* $a-\varepsilon$ *and then use the strategy* $z(A, a-\varepsilon)$. *The value of the game is* ka (which is equal to μ/g as defined in § 2; this last form of the formula will be encountered in the next section).

The theorem will be proved using three lemmas. The following lemma is the fundamental one and will also be used in the next section.

FUNDAMENTAL LEMMA. *Let* Q *be the set of* k *parallel arcs joining* O *and* A *defined in* § 4A. *Assume that at time* $t=0$, k *horses leave the point* O *and each of them rides in unit velocity along a different arc from* O *to* A. *At the same time* $t=0$ *a giraffe starts moving from any point different from* O *along a trajectory* $h(t)$ *where both the trajectory and the* (*not necessarily constant*) *velocity along this trajectory are arbitrary. The trajectory* $h(t)$ *should be continuous so that in passing from any arc* b_i *to any arc* b_j *the giraffe has to go through either* O *or* A; *then*:

(a) *For any trajectory* $h(t)$, *the giraffe will meet at least one horse in the time period* $0 < t \leqq a$.

(b) *If the giraffe moves with velocity which does not exceed unity then he will not meet more than one horse in the time period* $0 < t < a$.

Proof of part (a). Let $b_i(t)$, $i=1, \cdots, k$, $0 \leqq t \leqq a$, be the point located by the ith horse at time t and let

$$G(t)=\{b_i(t), i=1, 2, \cdots, k\}; \tag{30}$$

then the set $G(t)$ is the boundary of the set

$$G^{+}(t)=\{b_i(u)\, t<u \leqq a, i=1, 2, \cdots, k\}; \tag{31}$$

$G^{+}(0)$ is the whole set Q except for the point O while $G^{+}(a)$ is empty.

Let $h(t)$ be the location of the giraffe at time t; obviously $h(0) \in G^{+}(0)$ while $h(a) \notin G^{+}(a)$. The first instant at which the giraffe leaves the set $G^{+}(t)$ is defined by

$$t_0 = \inf_{0<t \leqq a} \{t: h(t) \notin G^{+}(t)\}. \tag{32}$$

Since $G(t)$ is the boundary of $G^{+}(t)$ it is obvious that $h(t_0) \in G(t_0)$; thus at time t_0 the giraffe meets (at least) one of the horses.

Proof of part (b). Suppose that at time t_0, $0<t_0<a$ the giraffe meets one of the horses. In order to meet another horse he has to pass to another arc which can be done only through O or through A, but since the velocity of the giraffe does not exceed unity (which is the velocity of the horses) then he cannot reach the point A before time a and if he moves through point O then he would not reach another horse until time a. Q.E.D.

LEMMA 5. *If the searcher uses the strategy* $z(\mathrm{O}, 0)$ (see Definition 1) *then for any hiding strategy* h

$$v(z(\mathrm{O},0),h)\leqq ka. \tag{33}$$

Proof. Strategy $z(\mathrm{O},0)$ for the searcher means: At time $t=0$ the searcher chooses one of the horses of the Fundamental Lemma, each with probability $1/k$ and rides on it to A; then at time $t=a$ he chooses one of k such horses which ride from A to O, each one again with equal probability (and independent of previous choice) etc. We may consider the hider as the giraffe and thus, part a of the Fundamental Lemma implies that if the searcher uses $z(\mathrm{O},0)$ and if the hider has not been captured until time $(j-1)a$ then the probability of capture in the period $(j-1)a<t\leqq ja$ is at least $1/k$ (independent of the previous part of the trajectory). Thus,

$$\Pr\,(\text{capture after } t=ja)\leqq\left(\frac{k-1}{k}\right)^j. \tag{34}$$

It follows that for any hiding strategy h,

$$\begin{aligned} v(z(\mathrm{O},0),h)&\leqq\sum_{j=1}^{\infty}\Pr\,(\text{capture at } (j-1)a<t\leqq ja)\cdot ja\\ &=a\sum_{j=1}^{\infty}\Pr\,(\text{capture after } t=(j-1)a)\\ &\qquad\text{(by (34))}\\ &\leqq a\sum_{j=1}^{\infty}\left(\frac{k-1}{k}\right)^{j-1}=ka\quad\text{Q.E.D.}\end{aligned} \tag{35}$$

LEMMA 6. *Let ZA be the strategy of the hider described as follows: Stay at point A until time* $t=a-\varepsilon$ *and then use* $z(A,a-\varepsilon)$ (see Definition 1); *then for any search trajectory*

$$v(s,ZA)\geqq ka-\varepsilon. \tag{36}$$

Proof. Using the strategy $z(A,a-\varepsilon)$ means choosing each of the horses (note that it is the hider who is now riding the horses) with probability $1/k$ (and independently of previous choices) at time periods $a-\varepsilon$, $2a-\varepsilon,\cdots$. Now the searcher takes the role of the giraffe of the Fundamental Lemma. We shall use the assumptions of § 4A and especially the fact that the velocity of the searcher does not exceed the velocity of the hider and the fact that the searcher cannot wait for the hider at any of the points O or A. Thus it can be assumed that at any one of the periods $2a-\varepsilon$, $3a-\varepsilon,\cdots$ the probability that the searcher is either at O or at A is zero (this can be achieved by the searcher by using ε as a random variable uniformly distributed in any small interval). Thus the conditions of part (b) of the Fundamental Lemma are satisfied so that for any trajectory of the searcher, if the hider has not been captured until $t=ja$ then the probability of capture in the time period $ja-\varepsilon<t\leqq(j+1)a-\varepsilon$ is equal to

$1/k$. Thus the probability that capture will occur at the time period $ja-\varepsilon<t\leqq (j+1)a-\varepsilon$ is equal to

$$\left(\frac{k-1}{k}\right)^{j-1}\frac{1}{k};$$

thus

$$v(s, ZA)\geqq \sum_{j=1}^{\infty}\left(\frac{k-1}{k}\right)^{j-1}\frac{1}{k}(ja-\varepsilon)=ka-\varepsilon \quad \text{Q.E.D.}$$

Theorem 2 is an immediate consequence of Lemmas 5 and 6.

It is interesting to note that the value of the game is equal to ka irrespective of the maximal speed of the hider w, as long as it is not less than unity. If w is less than unity then the optimal strategies of the searcher and the hider may be quite complicated. However, using an argument similar to those which shall be presented in § 5C, it seems to us that if k is large and w is not too small then the value of the game should be approximately ka, even for the case where $0<w<1$ because the hider can achieve a value of $ka(1-\varepsilon)$ by randomly choosing one of the arcs and moving along it with speed w from A to O then choosing again each arc with the same probability and moving along it with maximal speed from O to A and so on. If $m=1/w$ and the searcher moves with maximal speed (unity), then when the searcher reaches each of the points O or A, the maximal amount of information which may be available to the searcher is that at that moment the hider is not located on any of the m last arcs visited by the searcher. Thus, even if the searcher could rule out m arcs out of k each time he reaches O or A then his gain would be to increase the probability of capture for each period $ja<t\leqq(j+1)a$ from $1/k$ to $1/(k-m)=1/(k-1/w)$ and thus the expected capture time would decrease at most to $(k-1/w)a-\varepsilon$ so that if k is large in comparison with $1/w$, the value obtained in Theorem 2 would remain about the same even for $w<1$.

An interesting fact which is worth noting is the following: It has been shown in § 3 that for an immobile hider, the value v of the search game on k parallel arcs is equal to $ka/2$ for an even k and is approximately equal to $ka/2$ for an odd k. Thus for an immobile hider $v\approx\mu/(2g)$. For a mobile hider v is doubled and becomes μ/g and this is due to the fact that contrary to the case of an immobile hider, the searcher cannot rule out the arcs previously visited by him. We shall have the same phenomenon for the case of a two (or more) dimensional search set.

4C. Some reflections about general networks. In § 3 where search games with an immobile hider have been considered, we obtained inequality (8) which stated that the value of the search game lies between $\frac{1}{2}\mu$ and μ where μ is the sum of the lengths of the arcs of the network Q. The lower bound is achieved for the case where Q is an Eulerian network while the upper bound is achieved for the case where Q is a tree.

It seems to us that analogous results should hold for the case of a mobile hider. Our conjecture is that in this case the value of the game should be bounded below by μ and above by 2μ. The case considered in § 4B, where Q consists of k parallel arcs is an example in which $v=\mu$. A case where $v\approx 2\mu$ could be of the following type:

Let Q consist of k rays radiating from the origin O, each ray of equal length a. For this case it is obvious that $v<2\mu$ because if at time $t=0, 2a, 4a, \cdots$ the searcher chooses each ray with probability $1/k$ and then moves along it and returns to the origin; then the expected capture time is at most

$$\sum_{j=1}^{\infty}(a+2(j-1)a)\left(\frac{k-1}{k}\right)^{j-1}\frac{1}{k}<2ka=2\mu.$$

On the other hand, if k is large then the hider can achieve a capture time exceeding $2ka-\varepsilon$ by a strategy which is similar to the one which will be presented in § 5B: He uniformly chooses a ray and stays at its terminal point until time Ja where $1 \ll J \ll k$; then he chooses again each ray with probability $1/k$ and moves to its terminal point with maximal speed, and stays there until time $2Ja$, etc. (It should be assumed as in § 4A that the searcher cannot wait at the origin O or alternatively that the hider can pass from each ray to another by moving "near" the origin without actually passing through it.)

Thus we conjecture that these two cases which have been discussed represent the upper and the lower bounds for the value of the search game with mobile hider in a network. Thus it is still an open problem if the inequality $\mu \leqq v \leqq 2\mu$ always holds for all networks.

5. Mobile hider in a two (or more) dimensional region.

5A. General description. In this section the domain of search Q will be a two (or more) dimensional region. The searcher can move along any continuous trajectory which starts from the origin O. The hider can choose his initial location and an arbitrary continuous trajectory starting from that point. The hider can move along his trajectory with maximal speed w. In contrast to § 4, we shall not require that $w \geqq 1$ but we shall assume that w is not too small (the exact formulation of this condition will be presented in § 5C).

The notations that will frequently be used in this section are: μ—the Lebesgue measure of Q, R—the diameter of Q and g—the maximal rate of discovery of the searcher (which is equal to $2r$ for two dimensional sets and πr^2 for three dimensional sets, etc.). The radius of capture r will be assumed to be small with relation to the magnitude of Q or specifically

$$\frac{rR}{\mu}<\varepsilon' \ll 1. \tag{37}$$

For the case of an immobile hider it is mentioned in § 3 that v, the value of the search game, satisfies $v \sim \frac{1}{2}\mu/g$.

For the case of a mobile hider we shall show in § 5B that the searcher can guarantee an expected capture time not exceeding $(\mu/g)(1+\varepsilon)$. The dual result is presented in § 5C. We show there that if Q is convex then the hider can make sure that the expected capture time will exceed $(\mu/g)(1-\varepsilon)$. Thus we demonstrate that the value of the Princess and Monster game in a multidimensional convex set satisfies $v \sim \mu/g$ and we present ε-optimal strategies for both players.

5B. Strategy of the searcher. In this section we shall prove that for any bounded two dimensional region Q which satisfies a rather weak requirement, if the detection radius r is small then the searcher can use a strategy s^* which makes sure that the expected capture time does not exceed $(1+\varepsilon)\mu/(2r)$, where ε is small. We shall assume that:

(38) Q is a simply connected region whose boundary is the union of two single valued continuous functions f_1 and f_2 as depicted in Fig. 3.

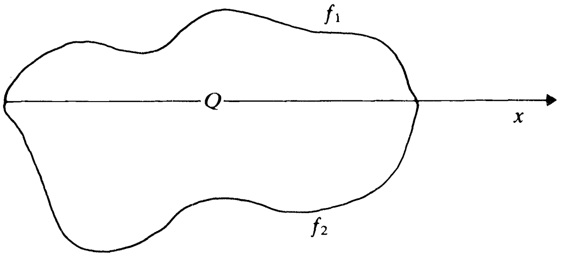

FIG. 3

Actually the proof can be easily extended to a multiply connected region whose boundary is the union of a finite number of single valued continuous functions. However, in this case we shall have to introduce more details which may complicate the proof so that for the simplicity of exposition we shall assume that (38) holds.

The search strategy s^* to be considered is of the following type: It consists of covering Q by a set of parallel and similar narrow rectangles $Q_1, Q_2, \cdots, Q_m, \cdots, Q_M$; randomly choosing a rectangle Q_m, going into it and moving N times forward and backward in Q_m along randomly chosen trajectories; then randomly choosing another set $Q_{m'}$ etc. The number N should be large enough in order to "absorb" the effect of the time spent in going from one rectangle to another, but on the other hand N must not be too large in order not to stay too long in one rectangle. Having this idea in mind we proceed as follows:

Referring to assumption (38) and Fig. 3, let $Q_a \supset Q$ be the set with minimal area which is a union of the rectangles $B_1, \cdots, B_L$ where all these rectangles have the same width a and are parallel to the x-axis as shown in Fig. 4.

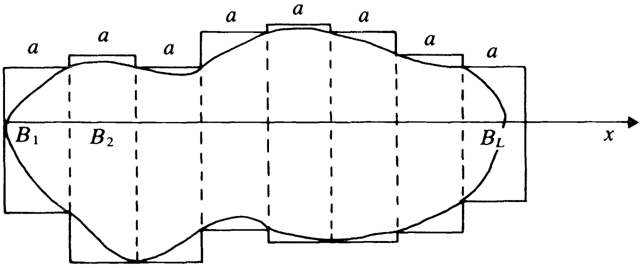

FIG. 4

Let μ be the area of Q and μ_a be the area of Q_a; then

$$\mu_a = (1+\phi_a)\mu \tag{39}$$

where, by assumption (38)

$$\phi_a \downarrow 0 \quad \text{as } a \downarrow 0. \tag{40}$$

For any $\mathbf{z}_1, \mathbf{z}_1 \in Q_a$ let $d(\mathbf{z}_1, \mathbf{z}_2)$ be the minimal length of a path which connects $\mathbf{z}_1$ and $\mathbf{z}_2$ and passes inside Q_a. We define the radius R or Q_a as

$$R = \sup_{\mathbf{z}_1, \mathbf{z}_2 \in Q_a} d(\mathbf{z}_1, \mathbf{z}_2). \tag{41}$$

We shall give a constructive proof of the following theorem:

THEOREM 3. *Let r satisfy*

$$r=\varepsilon_a\frac{a^2}{2R} \tag{42}$$

and assume that

$$\delta=\varepsilon_a^{1/4}\ll 1; \tag{43}$$

then there exists a searching strategy s^* in Q_a such that for any evading trajectory h used by the hider, the expected capture time $v(s^*, h)$ satisfies

$$v(s^*, h)\leqq(1+4\delta)\frac{\mu_a}{2r}=(1+4\delta)(1+\phi_a)\frac{\mu}{2r}. \tag{44}$$

Proof. Let

$$c=\delta^2 a. \tag{45}$$

Divide each rectangle B_l, $l=1,\cdots,2$, (see Fig. 4) into "narrow" rectangles so that all of these rectangles, except possibly one, has a width c while the upper one has a width $\leqq c$, as depicted in Fig. 5. Thus each rectangle Q_m, $m=1,\cdots,M$, has a width $c'\leqq c$ and the number M of such rectangles satisfies

$$\begin{aligned} M&=\sum_{l=1}^{L}\left(\left[\frac{C_l}{c}\right]+1\right)\leqq\sum_{l=1}^{L}\frac{C_l}{c}+L \\ &\leqq\frac{\mu_a}{ac}+\frac{R+a}{ac}=\frac{\mu_a}{ac}\left(1+c\frac{R+a}{\mu_a}\right) \\ &=\frac{\mu_a}{ac}\left(1+\delta^2\frac{a(R+a)}{\mu_a}\right)<\frac{\mu_a}{ac}\left(1+\frac{\delta}{2}\right) \quad \text{(by (45) and (43)).} \end{aligned} \tag{46}$$

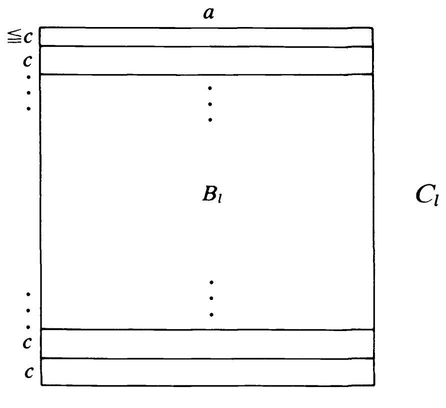

FIG. 5

Let N be a positive integer defined by

$$N=\left[\frac{R}{\delta a}\right]+1. \tag{47}$$

Let c' be a real number. If $c' \geqq 2r$ then we define a random variable y with the following density

$$
(48)\qquad
\begin{aligned}
f(y) &= \frac{2}{c'+2r} && \text{for } 0 \leqq y \leqq r,\\
f(y) &= \frac{1}{c'+2r} && \text{for } r < y < c'-r,\\
f(y) &= \frac{2}{c'+2r} && \text{for } c'-r \leqq y \leqq c',\\
\text{and } f(y) &= 0 && \text{elsewhere.}
\end{aligned}
$$

If $c' < 2r$ then we define y to be identically $c'/2$.

The search strategy s^* is composed of independent repetitions of the following step: At time $t = 0$ make a random choice out of the narrow rectangles $Q_1, Q_2, \cdots, Q_m, \cdots, Q_M$ such that each rectangle Q_m has a probability $1/M$ of being chosen and also make a random choice of N independent random variables $y_1, \cdots, y_N$ where N is given by (47) and all the y_n, $n = 1, \cdots, N$, have the probability density given by (48). In order to describe the motion of the searcher within Q_m we shall use a coordinate system with origin at the lower left corner of Q_m as depicted in Fig. 6. At time $t = 0$, the searcher starts moving as fast as possible to the point $(0, y_1)$. He rests at that point until $t = R$ and then moves with maximal velocity in a straight line to the point (a, y_1) and reaches it at time $t = R + a$; then he moves to the point (a, y_2) and rests there until $t = R + a + c$ and at that moment he starts moving to $(0, y_2)$ etc. The important feature of the movement of the searcher is that at the time segments $T_n = [R + (n-1)(a+c), R + (n-1)(a+c) + a]$, $n = 1, \cdots, N$, he moves along the intervals which join $(0, y_n)$ to (a, y_n). We shall show that for this kind of movement, the following lemma holds:

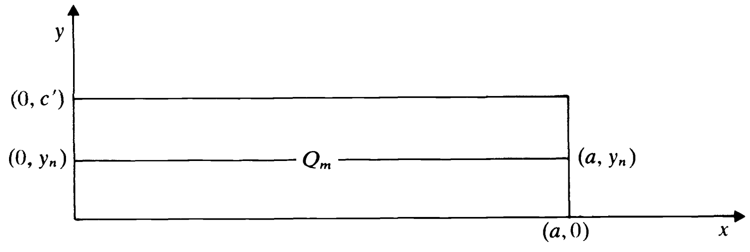

FIG. 6

LEMMA 7. *If the searcher moves in the manner described above, then the probability p of capture during the time segment* $0 < t \leqq R + N(a+c)$ *satisfies*

$$
(49)\qquad p \geqq \frac{1}{M}\left(1 - \left(\frac{c}{c+2r}\right)^N\right).
$$

Proof. Consider a specific time segment T_n given by

$$
(50)\qquad T_n = \{t : R + (n-1)(a+c) \leqq t \leqq R + (n-1)(a+c) + a\} \quad \text{where } 1 \leqq n \leqq N.
$$

We shall distinguish between two cases: If n is odd then for any $t \in T_n$ we define $G_m(t)$, $m = 1, 2, \cdots, M$, as the vertical line segment of length c', which has a distance

$d(t)$ from the *left* vertical edge of Q_m where

$$d(t)=t-[R+(n-1)(a+c)] \tag{51}$$

so that $G_m(t)$ is given by

$$G_m(t)=\{(d(t), y), 0\leqq y\leqq c'\}. \tag{52}$$

If n is even then $G_m(t)$ is the vertical line segment which has the distance $d(t)$ given by (51) from the *right* vertical edge of Q_m, i.e. in this case

$$G_m(t)=\{(a-d(t), y), 0\leqq y\leqq c'\}. \tag{53}$$

In both cases we define

$$G(t)=\bigcup_{m=1}^{M} G_m(t). \tag{54}$$

By an argument similar to the one used in proving the Fundamental Lemma one can show that the following proposition holds:

PROPOSITION. *Let h be any trajectory used by the hider; then for any n there exists at least one time instant $t_n\in T_n$* (see (50)) *such that the point $h(t_n)$ visited by the hider at time t_n satisfies $h(t_n)\in G(t_n)$* (see (54)).

It follows from the proposition that for any n there exists an m such that

$$h(t_n)\in G_m(t_n). \tag{55}$$

Let $I_m(n)=1$ if (55) holds and zero otherwise and define

$$I_m=\sum_{n=1}^{N} I_m(n); \tag{56}$$

then it follows from the above discussion that

$$\sum_{m=1}^{M} I_m\geqq N. \tag{57}$$

If the searcher chooses the rectangle Q_m and if $I_m(n)=1$ then it follows from the definition of the random variable y_n that the probability of capture during the time segment T_n is greater than or equal to the probability that at the time t_n (see (55)) the random interval Y_n given by

$$Y_n=[y_n-r, y_n+r]\cap[0, c'] \tag{58}$$

will contain the yth coordinate of $h(t_n)$. Now, it follows from (48) that for any point b in the interval $[0, c']$ the probability that $b\in Y_n$ is greater than or equal to $2r/(c'+2r)\geqq 2r/(c+2r)$.

Since the random variables $y_1,\cdots, y_N$ are independent, it follows that if Q_m has been chosen then the probability of capture during the time segment $0<t\leqq R+N(a+c)$, is greater than or equal to

$$1-\left(1-\frac{2r}{c+2r}\right)^{I_m}=1-\left(\frac{c}{c+2r}\right)^{I_m} \quad \text{(see (56)).}$$

Since each rectangle Q_m, $m=1,\cdots,M$, is chosen with probability $1/M$, it follows that the probability p of capture during the time segment $0<t\leqq R+N(a+c)$

satisfies

$$(59)\qquad p \geqq \sum_{m=1}^{M} \frac{1}{M}\left(1-\left(\frac{c}{c+2r}\right)^{I_m}\right) = 1-\frac{1}{M}\sum_{m=1}^{M}\left(\frac{c}{c+2r}\right)^{I_m}.$$

Since for any nonnegative integers J, K

$$\left(1-\left(\frac{c}{c+2r}\right)^{J}\right)\left(1-\left(\frac{c}{c+2r}\right)^{K}\right) \geqq 0,$$

it follows that

$$\left(\frac{c}{c+2r}\right)^{J}+\left(\frac{c}{c+2r}\right)^{K} \leqq 1+\left(\frac{c}{c+2r}\right)^{J+K}$$

so that

$$(60)\qquad \sum_{m=1}^{M}\left(\frac{c}{c+2r}\right)^{I_m} \leqq M-1+\left(\frac{c}{c+2r}\right)^{\Sigma_{m=1}^{M} I_m} \leqq M-1+\left(\frac{c}{c+2r}\right)^{N} \quad \text{(by (57))}.$$

Thus, it follows from (59) and (60) that

$$p \geqq \frac{1}{M}\left(1-\left(\frac{c}{c+2r}\right)^{N}\right)$$

so that the proof of Lemma 7 is completed.

We now proceed with the proof of Theorem 3. At first we note that (42), (43), (45) and (47) imply that

$$\frac{2r}{c}=\frac{2\delta^4 a^2}{2R\delta^2 a}=\delta^2\frac{a}{R}\geqq\frac{\delta}{N};$$

thus by (45), the probability p of capture in the time segment $0<t\leqq R+N(a+c)$ satisfies

$$(61)\quad p \geqq \frac{1}{M}\left(1-\left(\frac{1}{1+2r/c}\right)^{N}\right) \geqq \frac{1}{M}\left(1-\frac{1}{(1+\delta/N)^N}\right) \geqq \frac{1}{M}\left(1-\frac{1}{1+\delta}\right)=\frac{\delta}{M(1+\delta)}.$$

Now, since the search strategy s^* is composed of independent repetitions of the step described for the time segment $0\leqq t\leqq R+N(a+c)$, then for any hiding trajectory h, the probability $\bar{p}_K$ of capture after the time instant $t=K(R+N(a+c))$ satisfies

$$(62)\qquad \bar{p}_K \leqq (1-p)^K.$$

Thus the expected capture time $v(s^*, h)$ satisfies

$$\begin{aligned}
v(s^*, h) &\leqq (R+N(a+c))\sum_{K=0}^{\infty}\bar{p}_K \leqq \frac{R+N(a+c)}{p} && \text{(by (62))}\\
&\leqq \frac{M(1+\delta)}{\delta}Na\left(1+\frac{R}{Na}+\frac{c}{a}\right) && \text{(by (61))}\\
&\leqq \frac{\mu_a}{ac}\left(1+\frac{\delta}{2}\right)\frac{1+\delta}{\delta}\left(\frac{1}{\delta}\frac{R}{a}+1\right)a(1+\delta+\delta^2) && \text{(by (45), (46) and (47))}\\
&= \frac{1}{\delta^4}\frac{\mu_a R}{a^2}\left(1+\frac{\delta}{2}\right)(1+\delta)\left(1+\delta\frac{a}{R}\right)(1+\delta+\delta^2)\\
&\leqq \frac{\mu_a}{\varepsilon_a(a^2/R)}(1+4\delta)=\frac{\mu_a}{2r}(1+4\delta) && \text{(by (42) and (43))}\quad \text{Q.E.D.}
\end{aligned}$$

It should be noted that if the searcher uses the strategy s^* presented in the proof, then a part of his trajectory which is near the boundary of Q_a might be a little outside of the original searching set Q. However, if Q is convex then we can make a slight modification in s^* and introduce a search strategy s^{**} which uses trajectories entirely inside Q and still guarantees that the result (44) of Theorem 3 holds. The strategy s^{**} is defined as follows: If the chosen rectangle Q_m is inside Q then the movement is identical to the one in s^*. However if a part of Q_m is outside Q as depicted in Fig. 7 we make the following modification:

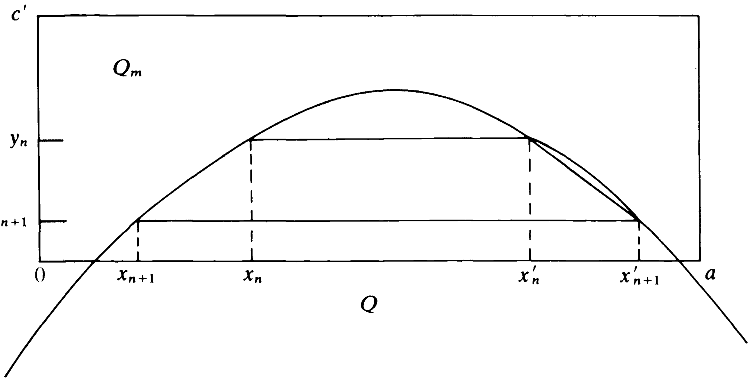

FIG. 7

Assume that in the strategy s^* the searcher moves from the point $(0, y_n)$ to (a, y_n) in the time segment T_n (see (50)) and then moves from the point (a, y_{n+1}) to $(0, y_{n+1})$ in the time segment T_{n+1}. The movement in s^{**} is as follows: The searcher moves from the point (x_n, y_n) to (x'_n, y_n) (see Fig. 7) in the time segment $R+(n-1)(a+c)+x_n \leqq t \leqq R+(n-1)(a+c)+x'_n$, then moves with maximal velocity in a straight line from (x'_n, y_n) to (x'_{n+1}, y_{n+1}) and stays there until the time instant $t' = R+n(a+c)+a-x'_{n+1}$ (the searcher can arrive at (x'_{n+1}, y_{n+1}) before t' because the length of the segment which joins (x'_n, y_n) to (x'_{n+1}, y_{n+1}) does not exceed $2a+c-x'_n-x'_{n+1}$) then moves from (x'_{n+1}, y_{n+1}) to (x_{n+1}, y_{n+1}) in the time segment $t' \leqq t \leqq R+n(a+c)+a-x_n$ etc.

Using exactly the same method of proving (44) it can be established that for any hiding strategy h, the expected capture time satisfies

$$v(s^{**}, h) \leqq (1+4\delta)\frac{\mu_a}{2r}$$

so that for a convex Q, the strategy s^{**} guarantees the desired result by moving only inside Q.

The result stated and proved by Theorem 3 for two dimensional regions can be extended to any number of dimensions. In this case the searcher can use a strategy s^* which guarantees that for any trajectory h used by the hider, the expected capture time satisfies

$$v(s^*, h) \leqq (1+\varepsilon)\frac{\mu}{g}$$

where μ is the Lebesgue measure of Q, g is the maximal rate of discovery of the searcher and ε is small.

For example if Q is a three dimensional region then the construction of s^* which keeps the expected capture time below $(1+\varepsilon)\mu/(\pi r^2)$ can be made as follows: Cover Q by a large number M of boxes of dimension $c\times c\times a$ where $c \ll a$, choose randomly one of the boxes and move along N random horizontal segments which join the two edges of size $c\times c$ etc. The proof of the result can be carried out by the same technique used in Theorem 3.

5C. Strategy of the hider. The strategy h_u of the hider which is considered in this section is defined as follows:

DEFINITION 2. Choose a point $\mathbf{x}_1$ using a uniform probability distribution in Q, and stay there during the time period $0 \leqq t < u$. At the time $t = u$ choose a point $\mathbf{x}_2$ which is uniformly distributed in Q independent of $\mathbf{x}_1$, move towards $\mathbf{x}_2$ with velocity $w_1 = \min[w, 1]$ in a straight line and stay in $\mathbf{x}_2$ for a time period of length u, then choose a point $\mathbf{x}_3$ uniformly distributed in Q and independent of $\mathbf{x}_1$ and $\mathbf{x}_2$, move towards it, again with velocity $w_1 = \min[w, 1]$ in a straight line and stay there for a time period of length u and so on.

The "resting time" u should satisfy two conditions:

1) It should not be too long so that the area covered by the searcher in a time interval of length u would be small relatively to μ; but on the other hand
2) In order to keep the probability of capture during the motion relatively small, the hider should not move too frequently and thus u should not be too small.

Assume that the hider uses a strategy h_u as described by Definition 2. Let

$$E_i\text{—The event: capture occurs at point } \mathbf{x}_i,\ i = 1, 2, \cdots. \tag{63}$$

Now if it were possible to neglect the probability of capture during the motion from $\mathbf{x}_i$ to $\mathbf{x}_{i+1}$ then for any search trajectory s, the expected capture time will approximately satisfy

$$\begin{aligned} v(s, h_u) &\gtrsim \sum_{i=1}^{\infty} u(i-1)\Pr(E_i) \\ &= u \sum_{n=1}^{\infty}\sum_{i=n+1}^{\infty} \Pr(E_i) \geqq u \sum_{n=1}^{\infty} \Pr\left(\overline{\bigcup_{i=1}^{n} E_i}\right). \end{aligned} \tag{64}$$

Since each $\mathbf{x}_i$ is uniformly distributed in Q and since the maximal rate of discovery is $2r$ it then follows that for each $\mathbf{x}_i$, the probability of being discovered at $\mathbf{x}_i$ is at most $2ru/\mu$. Thus, it follows from the independence of $\mathbf{x}_i$ that

$$\Pr\left(\overline{\bigcup_{i=1}^{n} E_i}\right) \geqq \left(1 - \frac{2ru}{\mu}\right)^n. \tag{65}$$

Thus it would follow from (64) and (65) that for all s

$$v(s, h_u) \gtrsim u \sum_{n=1}^{\infty} \Pr\left(\overline{\bigcup_{i=1}^{n} E_i}\right) = u\frac{1-2ru/\mu}{2ru/\mu} = \frac{\mu}{2r} - u \tag{66}$$

so that if the parameter u of the hider's strategy h_u would be chosen to be small in comparison with $\mu/(2r)$ we would get the desired result.

We have presented the previous discussion in order to help the reader to understand the motivation behind the definition of the strategy h_u and the idea of the (rather complicated) proof of Theorem 4 which follows.

We have also to require that the maximal velocity of the hider w should not be too small, because when w reduces to zero we approach the situation of an immobile hider considered in Chapter 3 and the value of the search game should then be approximately $\mu/(2g)$ $(=\mu/(4r))$. Such a condition which also includes condition (37) is the following:

$$\text{(67)} \qquad \frac{rR}{\mu}\max\left[\frac{1}{w}, 1\right] < \varepsilon' \ll 1.$$

In the next theorem we show for any two dimensional convex set, that if (67) holds then the hider can make sure that the capture time will exceed $(\mu/g)(1-\varepsilon)$ (recall that for two dimensions the discovery rate g is equal to $2r$). The theorem is formulated and proved for two dimensions but it can be easily extended to three or more dimensions.

THEOREM 4. *Let the searching set Q be any two dimensional convex set, and let condition* (67) *be satisfied. Denote*

$$\text{(68)} \qquad \delta = \left[\varepsilon' \cdot 37\pi \max\left(\frac{R^2}{\mu}, 1\right)\right]^{1/4}$$

and assume that $\delta \ll 1$.

If the hider uses the strategy h_u presented in Definition 2, *where*

$$\text{(69)} \qquad u = \delta\frac{\mu}{2r}$$

then for any search strategy s, the expected capture time satisfies

$$\text{(70)} \qquad v(s, h_u) \geqq \frac{\mu}{2r}[1-3\delta].$$

Proof. Let $\mathbf{x}_i$, i = 1, 2, · · · , be the hiding points of Definition 2 and denote:

(71)
F_i—The event: Capture occurs while the hider is moving from point $\mathbf{x}_i$ to point $\mathbf{x}_{i+1}$, $i = 1, 2, \cdots$
F_i'—The event: Capture does not occur before the hider leaves point x_i.

Then (see (63) and (71))

$$\begin{aligned} \Pr(F_n') &= \Pr\left(\overline{\bigcup_{i=1}^{n} E_i \bigcup_{i=1}^{n-1} F_i}\right) = 1 - \Pr\left(\bigcup_{i=1}^{n} E_i \bigcup_{i=1}^{n-1} F_i\right) \\ \text{(72)} \qquad &\geqq 1 - \Pr\left(\bigcup_{i=1}^{n} E_i\right) - \Pr\left(\bigcup_{i=1}^{n-1} F_i\right) \\ &\geqq \Pr\left(\overline{\bigcup_{i=1}^{n} E_i}\right) - \sum_{i=1}^{n-1} \Pr(F_i). \end{aligned}$$

The first stage of the proof is to establish an upper bound for Pr (F_i) so as to show that it is negligible in comparison to the other relevant terms.

Since the hider moves from point $\mathbf{x}_i$ to point $\mathbf{x}_{i+1}$ in a straight line with velocity $w_1 = \min[w, 1]$ it follows that the time of movement T_i satisfies

$$\text{(73)} \qquad T_i \leqq \frac{R}{w_1} = \frac{R}{\min[w, 1]}.$$

Since the probability of capture is monotonically increasing in T_i, it will be implicitly assumed that T_i can be replaced by R/w_1 whenever necessary.

Let S_i be that section of the searcher's trajectory which corresponds to the time interval when the hider moves from $\mathbf{x}_i$ to $\mathbf{x}_{i+1}$. Divide S_i into J_i arcs $S_{i1}, \cdots, S_{iJ_i}$ such that

$$J_i = \left[\frac{T_i}{r}\right] + 1 \leqq \frac{R}{w_1 r} + 1 \tag{74}$$

and each arc S_{ij} corresponds to an equal time interval t which satisfies

$$t = \frac{T_i}{J_i} < r. \tag{75}$$

Let F_{ij} be the event:

(76) F_{ij}: At some point of S_{ij}, the distance of the searcher and the hider is less than or equal to r.

Obviously (see (71))

$$\Pr(F_i) = \Pr\left(\bigcup_{j=1}^{J_i} F_{ij}\right) \leqq \sum_{j=1}^{J_i} \Pr(F_{ij}). \tag{77}$$

Let us denote the middle point of the arc S_{ij} by A_{ij} and the time when the searcher reaches this point by t_{ij} and let B_{ij} be a circle of radius c around A_{ij} (as depicted in Fig. 8) where c satisfies

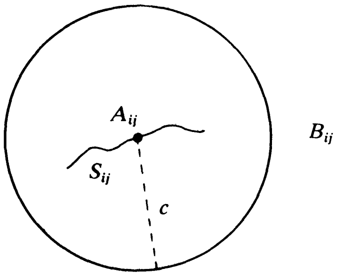

FIG. 8

$$c = \left(1 + \frac{1}{2} + \frac{w_1}{2}\right) r. \tag{78}$$

It follows from (75) that a necessary condition for the validity of the event F_{ij} (see (76)) is the event M_{ij}:

(79) M_{ij}—The event: At the time t_{ij} (corresponding to the point A_{ij}) the hider is inside the circle B_{ij};

and a necessary condition for the validity of M_{ij} is the event $D_{ij} \cap D_{ij}^*$ where

(80) D_{ij}—The event: The distance d_{ij} of $\mathbf{x}_i$ from A_{ij} satisfies: $(j+\frac{1}{2})w_1 r - c \leqq d_{ij} \leqq (j+\frac{1}{2})w_1 r + c$ (see (78))

and

(81) D_{ij}^*—The event: $\mathbf{x}_{i+1}$ lies in the region Y_{ij} depicted in Fig. 9.

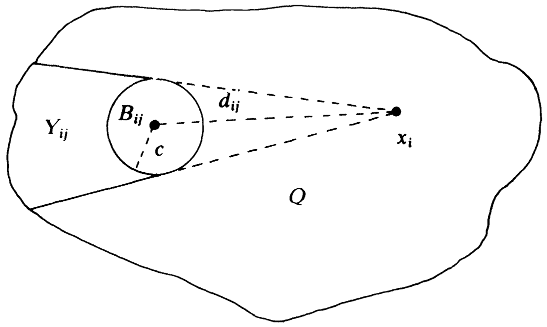

FIG. 9

It follows from (80) that

$$\Pr(D_{ij}) \leqq \frac{\pi(2j+1)w_1 r \cdot 2c}{\mu} = \frac{\pi(2j+1)w_1(3+w_1)r^2}{\mu} \quad \text{(by (78))} \tag{82}$$

and (81) implies that

$$\Pr(D_{ij}^* \mid D_{ij}) \leqq \frac{R}{\mu} \cdot \frac{2cR}{(j+1/2)w_1 r - c} = \frac{(3+w_1)R^2}{(jw_1 - 3/2)\mu}. \tag{83}$$

We are now in a position to obtain an upper bound for Pr (F_i) in (82): Let

$$m = \left\lfloor \frac{\sqrt{T_i}}{\sqrt{r}} \right\rfloor - 1 \quad \text{(see (73))}; \tag{84}$$

then

$$\begin{aligned} \Pr(F_i) &\leqq \sum_{j=1}^{m} \Pr(F_{ij}) + \sum_{j=m+1}^{J_i} \Pr(F_{ij}) \\ &\leqq \sum_{j=1}^{m} \Pr(D_{ij}) + \sum_{j=m+1}^{J_i} \Pr(D_{ij}) \cdot \Pr(D_{ij}^* \mid D_{ij}) \quad \text{(see (80) and (81))}. \end{aligned} \tag{85}$$

Now

$$\begin{aligned} \sum_{j=1}^{m} \Pr(D_{ij}) &\leqq \sum_{j=1}^{m} \frac{\pi(2j+1)w_1(3+w_1)r^2}{\mu} \quad \text{(by (82))} \\ &\leqq \frac{\pi w_1(3+w_1)r^2}{\mu}(m+1)^2 \leqq \frac{\pi w_1(3+w_1)r^2}{\mu}\frac{T_i}{r} \quad \text{(by (84))} \\ &\leqq 4\pi\frac{Rr}{\mu} < 4\pi\varepsilon' \quad \text{(by (73) and (67))}. \end{aligned} \tag{86}$$

In addition, it follows from (74), (82), (83), and (87) that

$$
\begin{aligned}
\sum_{j=m+1}^{J_i} \Pr(D_{ij}) \cdot \Pr(D_{ij}^* \mid D_{ij}) &\leqq \sum_{j=m+1}^{J_i} \frac{\pi(2j+1)w_1(3+w_1)r^2}{\mu^2} \frac{(3+w_1)R^2}{(jw_1-3/2)} \\
&\leqq \frac{16\pi}{\mu^2} \sum_{j=m+1}^{J_i} \frac{2j+1}{j-3/(2w_1)} R^2 r^2 \\
&\leqq \frac{16\pi}{\mu^2}(J_i-1)R^2r^2 \frac{2(m+1)+1}{m+1-3/(2w_1)} \\
&< \frac{16\pi}{\mu^2}\frac{R^3r}{w_1}\frac{2+1/(m+1)}{1-3/(2w_1(m+1))} \quad \text{(by (74))}.
\end{aligned} \tag{87}
$$

Since T_i can be replaced by R/w_1 it follows that

$$
\begin{aligned}
\frac{3}{2w_1(m+1)} = \frac{3}{2w_1\sqrt{R/(w_1r)}} &= \frac{3}{2}\sqrt{r/(w_1R)} \quad \text{(by (73) and (84))} \\
&= \frac{3}{2}\sqrt{Rr/(R^2w_1)} \ll 1 \quad \text{(by (67))}.
\end{aligned}
$$

Thus,

$$
16\frac{2+1/(m+1)}{1-3/(2w_1(m+1))} < 33
$$

so that a bound for (87) can be written as

$$
\frac{33\pi}{\mu^2}\frac{R^3r}{w_1} < \frac{33\pi R^2}{\mu}\varepsilon' \quad \text{(by (67))}. \tag{88}
$$

Combining (86) and (88) together we obtain

$$
\Pr(F_i) \leqq 37\pi\varepsilon' \cdot \max\left[\frac{R^2}{\mu}, 1\right] = \delta^4 \quad \text{(see (68))}. \tag{89}
$$

Now, let $v(s, h_u)$ be the expected capture time where the "resting" parameter u of the hider's strategy h_u is chosen to be $\delta\mu/(2r)$. If F_n' is defined, as in (71) to be the event that capture does not occur before the hider leaves the point $\mathbf{x}_n$, then the same argument used in (66) leads to the following inequality:

$$
\begin{aligned}
v(s, h_n) > u \sum_{n=1}^{\infty} \Pr(F_n') &\geqq u \sum_{n=1}^{N} \Pr(F_n') \quad \text{(where } N \text{ is an integer which will be determined later)} \\
&\geqq u \sum_{n=1}^{N} \left(\Pr\left(\overline{\bigcup_{i=1}^{n} E_i} \right) - \sum_{i=1}^{n-1} \Pr(F_i) \right) \quad \text{(see (72))} \\
&\geqq u \left(\sum_{n=1}^{N} \left(1 - \frac{2ru}{\mu}\right)^n - \sum_{n=1}^{N} (n-1)\delta^4 \right) \quad \text{(by (65) and (89))} \\
&\geqq u \left[\frac{(1-(2ru/\mu)) - (1-(2ru/\mu))^{N+1}}{2ru/\mu} - N^2\delta^4 \right].
\end{aligned} \tag{90}
$$

Now if we choose $N = \lfloor 1/\delta^2 \rfloor$ and use $u = \delta\mu/(2r)$ we obtain

$$
v(s, h_u) \geqq \delta\frac{\mu}{2r}\left(\frac{1-\delta-(1-\delta)^{1/\delta 2}}{\delta} - 1 \right) = \frac{\mu}{2r}[1-2\delta-(1-\delta)^{1/\delta 2}]. \tag{91}
$$

It can be easily seen that for $0<\delta<1$ as assumed,

$$(92)\qquad (1-\delta)^{1/\delta^2}=\exp\left[\frac{1}{\delta^2}\ln(1-\delta)\right]<\exp\left[-\frac{1}{\delta}\right]=\delta\exp\left[-\ln\delta-\frac{1}{\delta}\right]<\delta$$

(since $\delta\cdot\ln\delta\geqq -e^{-1}>-1$ so that $-\ln\delta-1/\delta<0$). Thus it follows from (91) and (92) that $v(s,h_u)\geqq(\mu/(2r))[1-3\delta]$. Q.E.D.

It seems to us that h_u can be modified so as to guarantee at least $(1-\varepsilon)\mu/g$ expected capture time for sets which satisfy a weaker condition like (38). The points $\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_i,\cdots$ should be chosen as before and the only problem is to make the movement from $\mathbf{x}_i$ to $\mathbf{x}_{i+1}$ so that the probability of capture during this movement would still be negligible with respect to the probability of capture at $\mathbf{x}_i$. This requirement should be achievable if the detection radius is small enough. Thus it is conceivable that $v\sim\mu/g$ for all reasonable regions.

REFERENCES

[1] S. ALPERN, *The search game with modile hider on the circle*, Differential Games and Control Theory, E. O. Roxin, P. T. Liu and R. L. Sternberg, eds., Marcel Dekker, New York, 1974, pp. 181–200.

[2] A. BECK, *On the linear search problem*, Israel J. Math., 2 (1964), pp. 221–228.

[3] ———, *More on the linear search problem*, Ibid., 3 (1965), pp. 61–70.

[4] A. BECK AND D. J. NEWMAN, *Yet more on the linear search problem*, Ibid., 8 (1970), pp. 419–429.

[5] A. BECK AND P. WARREN, *The return of the linear search problem*, Ibid., 14 (1973), pp. 169–183.

[6] R. BELLMAN, *An optimal search problem*, SIAM Rev., 5 (1963), p. 274.

[7] J. G. FOREMAN, *The princess and the monster on the circle*, Differential Games and Control Theory, E. O. Rubin, P. T. Liu and R. L. Sternberg, eds., Marcel Dekker, New York, 1974, pp. 231–240.

[8] ———, *Differential search game with mobile hider*, this Journal, 15 (1977), pp. 841–856.

[9] W. FRANCK, *On an optimal search problem*, SIAM Rev., 7 (1965), pp. 503–512.

[10] B. FRISTEDT AND D. HEATH, *Searching for a particle on the real line*, Advances in Appl. Probability, 6 (1974), pp. 79–102.

[11] S. GAL, *A general search game*, Israel J. Math., 12 (1972), pp. 32–45.

[12] ———, *Minimax solutions for linear search problems*, SIAM J. Appl. Math., 27 (1974), pp. 17–30.

[13] S. GAL AND D. CHEZAN, *On the optimality of the exponential functions for some minimax problems*, Ibid., 30 (1976), pp. 324–348.

[14] R. ISAACS, *Differential Games*, John Wiley, New York, 1965.

[15] B. J. MCCABE, *Searching for a one dimensional random walker*, J. Appl. Probability, 11 (1974), pp. 86–93.

[16] L. D. STONE, *Theory Of Optimal Search*, Academic Press, New York, 1975.

[17] D. J. WILSON, *Isaacs' princess and monster game on the circle*, J. Optimization Theory Appl., 9 (1972), pp. 265–288.

[18] R. H. WORSHAM, *A discrete game with a mobile hider*, Differential Games and Control Theory, E. O. Roxin, P. T. Liu and R. L. Sternberg, eds., Marcel Dekker, New York, 1974, pp. 201–230.

[19] M. I. ZELIKIN, *On a differential game with incomplete information*, Soviet Math. Dokl., 13 (1972), pp. 228–231.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 1, 1979

0036-0129/79/1701-0010 $01.00/0

STRONG STRUCTURAL CONTROLLABILITY*

HIROKAZU MAYEDA† AND TAKASHI YAMADA†

Abstract. For linear time-invariant control systems, the system parameters values may vary or be never known precisely with the exception of fixed zeros determined by the physical structure of the system. Dividing the system parameters into two categories, indeterminate parameters and fixed zero parameters, the notion of strong structural controllability is introduced with the following meaning: A system is strongly structurally controllable if, whatever values (other than zero) the indeterminate parameters of the system may take, the system is controllable.

The two necessary and sufficient graph theoretic conditions for linear time-invariant control systems to be strongly structurally controllable are given. The one is fundamental for strong structural controllability and shows what is the essential set of indeterminate parameters the change of whose values may cause a system to be uncontrollable. The other is useful because of its very simple and intuitive form in graph theoretic aspect. For sparse systems, its conditions can be easily examined by inspection.

1. Introduction. Consider a linear time-invariant control system

$$\dot{x} = Ax + bu \tag{1}$$

where A is $n \times n$ and b is $n \times 1$. For convenience, the control system (1) is denoted throughout the paper by the system (A, b).

System parameter values may vary or be never known precisely with the exception of zeros that are fixed by coordination or by the absence of physical connections between certain parts of a system. Therefore assume that the entries of the matrices A and b are fixed (zero) or indeterminate (arbitrary). We shall say that the system $(\bar{A}, \bar{b})$ has the same *structure* as the system (A, b), of the same dimensions, if for every fixed (zero) entry of the matrix $[\bar{A}\ \bar{b}]$, the corresponding entry of the matrix $[A\ b]$ is fixed (zero) and, at the same time, for every fixed (zero) entry of the matrix $[A\ b]$, the corresponding entry of the matrix $[\bar{A}\ \bar{b}]$ is also fixed (zero).

Lin [7] has defined a system (A, b) to be *structurally controllable* if there exists a completely controllable system $(\bar{A}, \bar{b})$ which has the same structure as the system (A, b), and by the graph-theoretic approach he developed necessary and sufficient conditions for a system (A, b) to be structurally controllable. Recently, Shields and Pearson [9] and Glover and Silverman [5] extended Lin's results on single-input systems to multi-input systems by the purely algebraic approach.

Although all the uncontrollable systems which have the same structure as a structurally controllable system are atypical (see [7], [9]), in many cases, the existence of such uncontrollable systems are not allowed. For example, when it is required that a nonlinear system with fixed zero parameters be regulated at various set points which vary in some domain, we adopt a set of same structure linear systems which are obtained by linearizing the nonlinear system at every set point and require that all the systems in the set are controllable. In this case, even if the systems are structurally controllable and almost all systems are controllable, there remain the set points whose corresponding systems are uncontrollable. And if there exists such a set point, all the systems whose corresponding set points are in the neighborhood of this set point are physically uncontrollable since it may require an unreasonably vast amplitude of the inputs to regulate the systems. We cannot neglect the existence of such a neighborhood. Therefore there arises the problem that under what condition a controllable system remains controllable for any changes in its indeterminate parameters.

* Received by the editors April 29, 1977, and in revised form March 21, 1978.

† Department of Control Engineering, Faculty of Engineering Science, Osaka University, Toyonaka, Osaka, Japan.

From this point of view, the notion of *strong structural controllability* is introduced as follows: The system (A, b) is *strongly structurally controllable* if any system $(\bar{A}, \bar{b})$ which has the same structure as the system (A, b) is completely controllable as long as every indeterminate entry of the matrix $[\bar{A}\ \bar{b}]$ is not zero. It immediately follows that every strongly structurally controllable system is also structurally controllable. If the value of an indeterminate parameter is zero, this means the corresponding connection in the system is cut off. So this case is excluded.

Our approach is graph theoretic. In § 2 we define the graph of a system (A, b) and the related terms. In § 3 we develop the two necessary and sufficient graph theoretic conditions (Theorem 1 and Theorem 2) for systems to be strongly structurally controllable. Theorem 1 shows what is the essential set of indeterminate parameters the change of whose values may cause a system to be uncontrollable. This problem is not only fundamental for strong structural controllability but also may give some insight to structural controllability of systems with dependent indeterminate parameters. Theorem 2 is given in a very simple and intuitive form in graph theoretic aspect. Its conditions can be examined by inspection and are easier to examine than those of Theorem 1 for sparse systems. Theorem 2 is also useful to investigate the strong structural controllability when some subsystems are added to or deleted from a system.

2. The graph of a system (A, b). The graph of a system (A, b), which will be denoted by $G(A, b)$ hereinafter, is a graph that contains exactly $n+1$ nodes, $1, 2, \cdots, n+1$, and all of whose edges are obtained as follows: For every indeterminate entry c_{ij} of the $n \times (n+1)$ matrix $[A\ b]$, the graph contains the oriented edge (j, i) (an arrow going from the node j to the node i). The node $n+1$, which corresponds to the $n+1$st column of $[A\ b]$, will be called the *origin* of $G(A, b)$. The set with all the nodes in $G(A, b)$ except the origin will be denoted by Z. For any oriented edge (j, i) in $G(A, b)$, the node j (node i) will be called the *initial* (*final*) node of the edge (j, i).

A graph, which consists of a subset of the set of all the edges in $G(A, b)$ and all the initial or final nodes of the edges in the subset, will be called a subgraph of $G(A, b)$. For any subgraph H of $G(A, b)$, $V(H)$ $(E(H))$ denotes the set of all the nodes (edges) in H and the subgraph H will be said to *span* the graph $G(A, b)$ if $V(H) = V(G(A, b))$ is satisfied. For any two subgraphs H_1, H_2, $H_1 \cup H_2$ denotes the subgraph which consists of $E(H_1) \cup E(H_2)$ and $V(H_1) \cup V(H_2)$.

Consider a subgraph H of $G(A, b)$ which consists of a sequence of edges, $(i_1, i_2), (i_2, i_3), \cdots, (i_{k-1}, i_k)$, and nodes $i_1, i_2, \cdots, i_k$. The subgraph H will be called a *path* for $k \geqq 2$ if no pair of nodes in H are coincident. The node i_1 (node i_k) will be called the *initial* (*final*) node of the path. Further, the path is said to *reach* the node i_k from the node i_1. The subgraph H will be called a *cycle* for $k \geqq 1$ if only the pair of nodes i_1 and i_k are coincident. Moreover, the subgraph H will be called a *bud* for $k \geqq 2$ if only the pair of nodes i_2 and i_k are coincident. In the bud, the node i_1 or the edge (i_1, i_2) will be called the *origin* or the *distinguished edge* of the bud respectively. The subgraph obtained by removing the distinguished edge and the origin from a bud B will be called the *cycle part* of the bud B and will be denoted by B^*. In certain cases, a path whose initial node is the origin of $G(A, b)$ will be called a *stem*. See Fig. 1.

For a path and a cycle C (set $\mathscr{C}$ of cycles), we shall say that the path *reaches* the cycle C (set $\mathscr{C}$ of cycles) from node i if the node i is the initial node of the path and only the final node of the path belongs at the same time to the path and the cycle C (some cycle in the set $\mathscr{C}$).

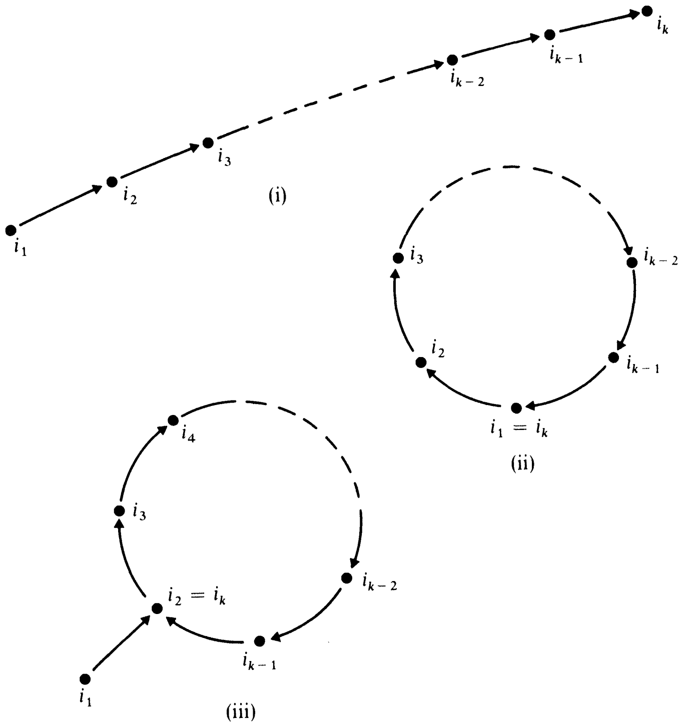

FIG. 1. (i) *Path.* (ii) *Cycle.* (iii) *Bud.*

The graph $G(A, b)$ is defined to be *accessible* if for any node i in Z, there exists at least one path whose initial node is the origin of $G(A, b)$ and whose final node is the node i. For any set N of nodes, $|N|$ denotes the number of distinct nodes in N and $T(N)$ denotes the set containing all the nodes i with the property that there is an edge in $G(A, b)$ going from the node i to some node in N. The graph $G(A, b)$ is said to contain a *dilation* if there exists a subset N of Z which satisfies $|T(N)| < |N|$.

Given a stem B_0 and p buds, $B_1, B_2, \cdots, B_p$, the subgraph $B_0 \cup B_1 \cup B_2 \cup \cdots \cup B_p$ will be called a *cactus* if for every $j = 1, 2, \cdots, p$, the origin of B_j is not the final node of B_0 and is the only node which belongs at the same time to B_j and $B_0 \cup B_1 \cup B_2 \cup \cdots \cup B_{j-1}$. Moreover the bud B_j will be called the jth *bud* of the cactus for $1 \leqq j \leqq p$. In the above definition of a cactus, if the origin of B_1 belongs to B_0 and the origin of B_j belongs to B_{j-1}^* for $2 \leqq j \leqq p$, then the subgraph $S = B_0 \cup B_1 \cup B_2 \cup \cdots \cup B_p$ will be called a *serial buds cactus* and abbreviated to s.b.c. Here let us define a subgraph S^* of S as $S^* = B_0 \cup B_1^* \cup B_2^* \cup \cdots \cup B_p^*$. Then, for any subset N of $V(S^*)$, $T_{S^*}(N)$ denotes the set containing all the node i with the property that there is an edge in $E(S^*)$ going from the node i to some node in N.

Suppose $G(A, b)$ is spanned by an s.b.c. $B_0 \cup B_1 \cup B_2 \cup \cdots \cup B_p$ where B_0 is the stem and B_j is the jth bud for $1 \leqq j \leqq p$; then all the nodes of $G(A, b)$ are labeled associated with this s.b.c. as follows: Starting from the initial node of B_0, assign labels, $1(B_0), 2(B_0), \cdots, d_0(B_0), \cdots, m_0(B_0)$, to the nodes in B_0 in order, where $m_0 = |V(B_0)|$ and the node labeled $d_0(B_0)$ is the origin of B_1. And for every $j = 1, 2, \cdots, p$, starting from the final node of the distinguished edge of B_j, assign labels,

$1(B_j), 2(B_j), \cdots, d_j(B_j), \cdots, m_j(B_j)$, to the nodes in the cycle part of B_j in order, where $m_j = |V(B_j)| - 1$ and the node labeled $d_j(B_j)$ is the origin of B_{j+1}.[1] This will be called the labeling associated with the s.b.c. See Fig. 2.

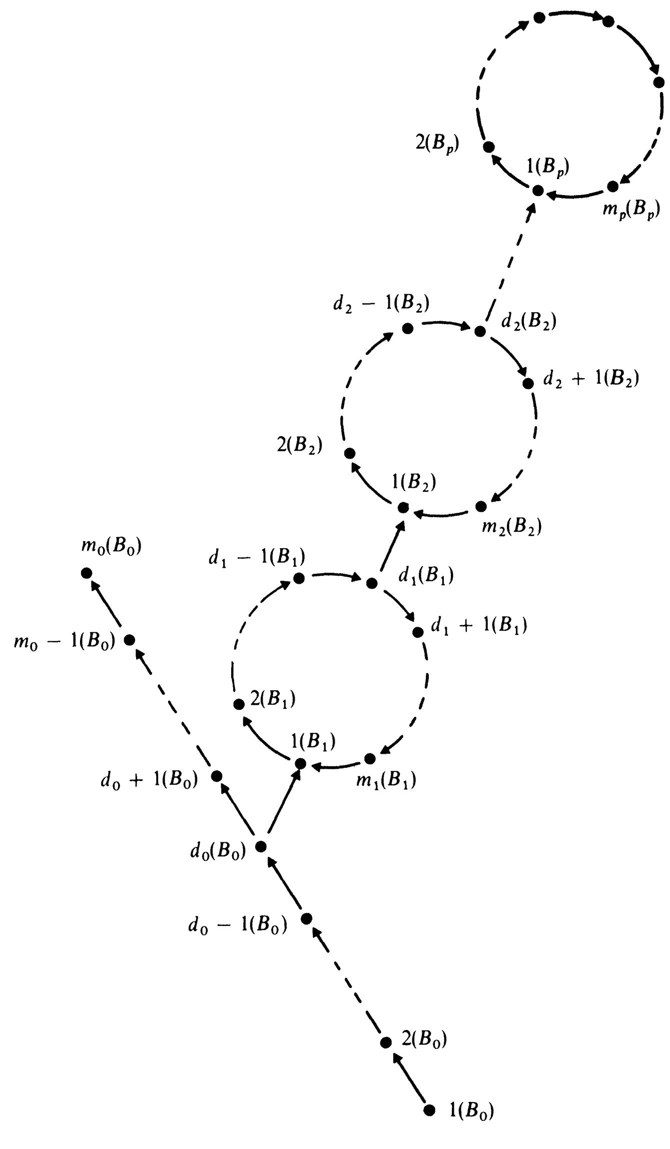

FIG. 2. *Labeling associated with an s.b.c.*

3. Strong structural controllability. In this section, the two necessary and sufficient conditions for a system (A, b) to be strongly structurally controllable will be given. Before stating our results, we should show some results of Lin concerning structural controllability.

[1] In the case of $j = p$, the label $d_j(B_j)$ is not defined. And in certain cases, $i(B_j)$ denotes the node labeled $i(B_j)$ for convenience.

LEMMA 1 (Lin [7]). *The system (A, b) is structurally controllable if and only if the graph $G(A, b)$ satisfies either of the following properties:*

(a1) *The graph $G(A, b)$ is accessible and contains no dilation.*

(a2) *The graph $G(A, b)$ is spanned by a cactus.*

Our first results are as follows:

THEOREM 1. *The system (A, b) is strongly structurally controllable if and only if the graph $G(A, b)$ satisfies the following three properties:*

(b1) *The graph $G(A, b)$ is accessible.*

(b2) *For any subset N of Z, there exists at least one node i in $T(N)$ such that the number of the edges going from the node i to some nodes in N is one.*

(b3) *For any subset N of Z which satisfies $T(N) \supset N$, there exists at least one node i in $T(N)-N$ such that the number of the edges going from the node i to some nodes in N is one.*

First, we will show the following two lemmas:

LEMMA 2. *If the graph $G(A, b)$ is assumed accessible, then the property* (b3) *of Theorem* 1 *is satisfied if and only if whatever nonzero real value each indeterminate entry of the matrix $[A\ b]$ may take, there exist no complex number $\alpha(\neq 0)$ and n-vector $q(\neq 0)$ with complex entries which satisfy $q^T[A\ b]=\alpha[q^T\ 0]$.*

Proof. Define c_{ij} as the (i, j) entry of $[A\ b]$ and q_i as the ith entry of the vector q. Since $G(A, b)$ is accessible, $T(N)-N$ is not empty for any subset N of Z.

Sufficiency. Suppose that there exists a subset N of Z such that $T(N) \supset N$ and each node i in $T(N)-N$ has more than one edge going from the node i to some nodes in N. By a permutation of the coordinates, it can be assumed without loss of generality that the first k rows (columns) of $[A\ b]$ correspond to the nodes in N. (Here $k=|N|$.)

Now choose an α and a vector q so that $\alpha=1$, $q_1=q_2=\cdots=q_k=1$ and $q_{k+1}=\cdots=q_n=0$. Since $T(N)\supset N$, each jth column in the first k columns of $[A\ b]$, which corresponds to some node in N, contains at least one indeterminate entry in its first k entries from the definition of $T(N)$. So, if the indeterminate entries in the first k entries of the jth column take suitable nonzero real values,

$$\sum_{i=1}^{n} q_i c_{ij} = \sum_{i=1}^{k} c_{ij} = 1 = \alpha q_j \tag{2}$$

can be satisfied for $1 \leqq j \leqq k$. Moreover, since each node in $T(N)-N$ has more than one edge going from the node to some nodes in N and there is no edge going from any node in $V(G(A, b))-T(N)$ to any node in N, any rth column in the last $n+1-k$ columns of $[A\ b]$ contains more than one or no indeterminate entry in its first k entries. Therefore if the indeterminate entries in the last $n+1-k$ columns take suitable nonzero real values, the equation

$$\sum_{i=1}^{n} q_i c_{ir} = \sum_{i=1}^{k} c_{ir} = 0 \tag{3}$$

can be satisfied for $k+1 \leqq r \leqq n+1$. From (2) and (3), it can be concluded that if each indeterminate entry of $[A\ b]$ takes a suitable nonzero real value, there exist a complex number $\alpha(\neq 0)$ and a vector $q(\neq 0)$ which satisfy $q^T[A\ b]=\alpha[q^T\ 0]$.

Necessity. Suppose that each indeterminate entry of $[A\ b]$ takes a certain nonzero real value and there exist a complex number $\alpha(\neq 0)$ and a vector $q(\neq 0)$ which satisfy $q^T[A\ b]=\alpha[q^T\ 0]$. By a permutation of the coordinates, it can be assumed without loss of generality that all the nonzero entries of the vector q are located in its first k

entries. Then

$$\sum_{i=1}^{n} q_i c_{ij} = \sum_{i=1}^{k} q_i c_{ij} = \alpha q_j \tag{4}$$

is satisfied for $1 \leqq j \leqq k$, and

$$\sum_{i=1}^{n} q_i c_{ir} = \sum_{i=1}^{k} q_i c_{ir} = 0 \tag{5}$$

is satisfied for $k+1 \leqq r \leqq n+1$.

Let us choose N as the set which contains all the nodes corresponding to the first k rows (columns) of $[A\ b]$; then since $q_j \neq 0$ for $1 \leqq j \leqq k$ in (4), any jth column in the first k columns of $[A\ b]$ contains at least one indeterminate entry in its first k entries. So $T(N) \supset N$ can be concluded. Since, in (5), q_i is a nonzero complex number and c_{ir} is a real number for $1 \leqq i \leqq k$ and $k+1 \leqq r \leqq n+1$, the first k entries of any rth column in the last $n+1-k$ columns of $[A\ b]$ contain more than one indeterminate (nonzero) entry in its first k entries or are all fixed (zero) entries. So any node s in $T(N)-N$ which corresponds to some column in the last $n+1-k$ columns of $[A\ b]$ has more than one edge going from the node s to some nodes in N. Therefore the existence of the set N contradicts the property (b3) of Theorem 1. Q.E.D.

LEMMA 3. *Assuming that the graph $G(A, b)$ is accessible, the property* (b2) *of Theorem* 1 *is satisfied if and only if whatever nonzero real value each indeterminate entry of the matrix $[A\ b]$ may take, there exists no n-vector $q(\neq 0)$ with complex entries which satisfies $q^T[A\ b] = 0$.*

Lemma 3 is easily proved in the way similar to the way Lemma 2 is proved.

Proof of Theorem 1. Since the fact the system (A, b) is strongly structurally controllable implies that the system (A, b) is structurally controllable, the necessity of property (b1) of Theorem 1 is apparent from Lemma 1. Recall that the system (A, b) is completely controllable (in the usual sense) if and only if the relation $q^TA = \alpha q^T$ implies $q^Tb \neq 0$ where α is a complex number and $q \neq 0$ is an n-vector with complex entries (see V. M. Popov [8] for reference). Thus Theorem 1 is proved immediately from Lemmas 2 and 3. Q.E.D.

From Theorem 1 and its proof, we can identify the essential sets of indeterminate entries, the change of whose values may cause the system to be uncontrollable even if the system is structurally controllable. That is, if N is a subset of Z which does not satisfy property (b2) or (b3) of Theorem 1, only the indeterminate entries which correspond to the edges whose final nodes are in N can take certain values such that the system is uncontrollable, and the values of other indeterminate entries have no concern with it.

Next, we shall show our second result which is useful because of its simple and intuitive form in graph theoretic aspect.

THEOREM 2. *The system (A, b) is strongly structurally controllable if and only if the graph $G(A, b)$ satisfies the following two properties*:

(c1) *The graph $G(A, b)$ is spanned by a unique s.b.c.*

(c2) *The graph $G(A, b)$ contains no set of cycles which is reached from the origin of $G(A, b)$ by more than one path.*

For sparse systems, Theorem 2 can be examined by inspection and is easier to test than Theorem 1. The simple instructions to check the uniqueness of the s.b.c. which spans $G(A, b)$ will be given in the Remarks. Moreover Theorem 2 is very useful to investigate the strong structural controllability when some edges or subsystems are added to or deleted from a system.

We shall show the following lemmas. They are also interesting in themselves.

LEMMA 4. *If the graph $G(A, b)$ is assumed accessible, then the property* (b3) *of Theorem* 1 *is equivalent to the property* (c2) *of Theorem* 2.

Proof. (b3) ⇒ (c2): Suppose that the property (c2) is not satisfied, that is, $G(A, b)$ contains a set $\mathscr{C}$ of cycles which is reached from the origin of $G(A, b)$ by exactly h distinct paths, $P_1, P_2, \cdots, P_h$ ($h \geqq 2$). The initial node‸ of every path P_j ($j = 1, 2, \cdots, h$) is the origin of $G(A, b)$, so there exists a positive integer f such that starting from the origin of $G(A, b)$, the first f node sequences of all the paths $P_1, P_2, \cdots, P_h$ are the same node sequence, $i_1, i_2, \cdots, i_f$, and all the $f+1$st nodes of the paths are not coincident. Let F be the set of nodes, $i_1, i_2, \cdots, i_f$, and $Q = V(\bigcup_{j=1}^{h} P_j) - F$. Moreover, define the set R of nodes as follows: For every node i in R, there exists a path which contains no node in F and reaches some node in Q from the node i.

Now let us choose the set N as

$$N = Q \cup R; \tag{6}$$

then it is obvious that

$$T(N) = T(Q) \cup T(R). \tag{7}$$

Since there do not exist more than h distinct paths each of which reaches the set $\mathscr{C}$ from the origin of $G(A, b)$, there exists no edge going from some node in $F - \{\text{node } i_f\}$ to some node in Q or R. So, both $T(Q)$ and $T(R)$ are included in $R \cup Q \cup \{\text{node } i_f\}$ from the definition of Q and R. Thus

$$T(N) \subseteq N \cup \{\text{node } i_f\} \tag{8}$$

is satisfied from (6) and (7).

Consider an arbitrary node t in R; then the node t has an edge going from the node t to some node in R or Q from the definition of R. So,

$$R \subseteq T(R) \cup T(Q) = T(N) \tag{9}$$

is satisfied.

Consider an arbitrary node t in Q. If the node t is not the final node of any path among $P_1, P_2, \cdots, P_h$, it is obvious that the node t is contained in $T(Q)$. If the node t is the final node of some path among $P_1, P_2, \cdots, P_h$, the node t is contained in a cycle C in $\mathscr{C}$ which contains no node in F. So it can be easily obtained that the node t is contained in $T(Q) \cup T(R)$. Thus

$$Q \subseteq T(Q) \cup T(R) = T(N) \tag{10}$$

can be concluded.

It is obvious that the node i_f is contained in $T(Q)$ ($\subset T(N)$), so

$$T(N) \supseteq N \cup \{\text{node } i_f\} \tag{11}$$

is obtained from (6), (9) and (10). Thus $T(N) = N \cup \{\text{node } i_f\}$ is satisfied from (8) and (11). Moreover $T(N) - N = \{\text{node } i_f\}$ is satisfied since the node i_f is contained neither in Q nor in R. Therefore the existence of the set N contradicts the property (b3) since the node i_f has more than one edge going from the node i_f to some nodes in $Q (\subset N)$.

(c2) ⇒ (b3): Suppose that the property (b3) is not satisfied, that is, there exists a subset N of Z such that $T(N) \supset N$ and each node i in $T(N) - N$ has more than one edge going from the node i to some nodes in N. For any node i_1 in N, there exists an edge going from the node i_1 to some node i_2 in N since the node i_1 is contained in

$T(N)$, and since the node i_2 is also contained in $T(N)$, there exists an edge going from the node i_2 to some node i_3 in N. Continuing this procedure, since $|N|$ is finite, we can obtain a sequence of distinct nodes, $i_1, i_2, \cdots, i_k$ in N such that there exist an edge going from the node i_j to the node i_{j+1} for $1 \leqq j \leqq k-1$ and an edge going from the node i_k to some node among the nodes, $i_1, i_2, \cdots, i_k$. Therefore each node i_1 in N is contained in some cycle in N or has a path which is contained in N and reaches some node in some cycle contained in N from the node i_1.

From $N \subset Z$, the origin of $G(A, b)$ is contained in $V(G(A, b)) - T(N)$ or $T(N) - N$. If the origin of $G(A, b)$ is contained in $T(N) - N$, the origin of $G(A, b)$ has more than one edge going from the origin of $G(A, b)$ to some node in N. If the origin of $G(A, b)$ is contained in $V(G(A, b)) - T(N)$, since $G(A, b)$ is accessible and every path which reaches some node in N from the origin of $G(A, b)$ contains some nodes in $T(N) - N$, there exists a path which reaches some node t in $T(N) - N$ from the origin of $G(A, b)$ and contains no node in N. The node t in $T(N) - N$ also has more than one edge going from the node t to some nodes in N.

Thus, in both cases,' we can easily find a set of cycles which is reached from the origin of $G(A, b)$ by more than one path. This fact contradicts the property (c2). Q.E.D.

LEMMA 5. *If the graph $G(A, b)$ satisfies property* (c2) *of Theorem* 2 *and is spanned by an s.b.c. $B_0 \cup B_1 \cup \cdots \cup B_p$ where B_0 is the stem and B_j is the j-th bud for $1 \leqq j \leqq p$, then the graph $G(A, b)$ satisfies the following properties when all the nodes of $G(A, b)$ are labeled associated with the s.b.c.*:

(d1) *If $1 \leqq i_1 < i_1 + 2 \leqq i_2 \leqq d_j$ is satisfied for some j $(0 \leqq j \leqq p-1)$, there does not exist an edge $(i_1(B_j), i_2(B_j))$ in $G(A, b)$.*

(d2) *If $d_j + 1 \leqq i_1 \leqq i_2 \leqq m_j$ is satisfied for some j $(0 \leqq j \leqq p-1)$, there does not exist an edge $(i_2(B_j), i_1(B_j))$ in $G(A, b)$.*

(d3) *For $0 \leqq j \leqq p-1$, there does not exist an edge going from any node in B_j to any node in $V(B^*_{j+1}) \cup V(B^*_{j+2}) \cup \cdots \cup V(B^*_p)$ other than the edge $(d_j(B_j), 1(B_{j+1}))$.*

(d4) *If $1 \leqq i_1 < i_1 + 2 \leqq i_3 \leqq i_4 \leqq m_j$ and $i_1 + 1 \leqq i_2 \leqq i_4$ are satisfied for some j $(0 \leqq j \leqq p)$, there do not exist an edge $(i_1(B_j), i_3(B_j))$ and an edge $(i_4(B_j), i_2(B_j))$ at the same time.* (See Fig. 3.)

It is easy to prove Lemma 5 since $G(A, b)$ satisfies property (c2).

LEMMA 6. *Assume that the graph $G(A, b)$ satisfies property* (c2) *of Theorem* 2 *and is spanned by an s.b.c. S. Then the graph $G(A, b)$ does not satisfy property* (b2) *of Theorem* 1, *if there exists a subset N of Z such that each node in $T_{S^*}(N)$ has more than one edge in $E(G(A, b))$ going from that node to some nodes in N.*

Proof. Let the s.b.c. S which spans $G(A, b)$ be $S = B_0 \cup B_1 \cup \cdots \cup B_p$, where B_0 is the stem and B_j is the jth bud for $1 \leqq j \leqq p$, and label all the nodes of $G(A, b)$ associating with the s.b.c. S.

If $T(N) = T_{S^*}(N)$, $G(A, b)$ does not satisfy the property (b2). So suppose that $T(N) - T_{S^*}(N)$ is not empty. In this case, $m_0(B_0) \notin T(N)$ is shown as follows: If $m_0(B_0) \in T(N)$, there exists at least one edge by $E(G(A, b))$ going from the node $m_0(B_0)$ to some node in $N \cap \{\text{node } k(B_0) | 2 \leqq k \leqq d_0\}$ by properties (d2) and (d3). Next, define i_m as $i_m = \min \{i \,|\, \text{node } i(B_0) \in N, 2 \leqq i \leqq d_0\}$. Then it follows that $i_m - 1(B_0) \in T_{S^*}(N)$. From the assumption concerning N, there exists another edge than $(i_m - 1(B_0), i_m(B_0))$ in $E(G(A, b))$ such that the initial node is $i_m - 1(B_0)$ and the final node is contained in N. But, whatever node is chosen as the final node, the existence of such an edge contradicts the definition of i_m or property (d3) or (d4).

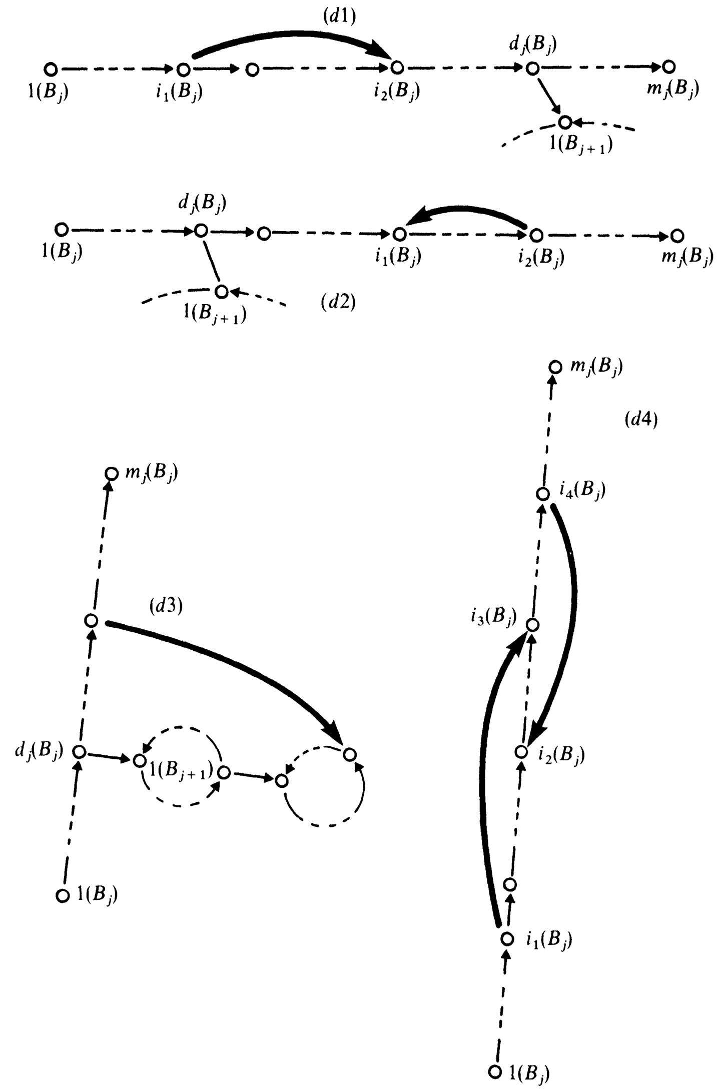

FIG. 3. *Illustrations for* (d1)–(d4).

Since $m_0(B_0) \notin T(N)$, we can choose the subset N_1 of Z which satisfies $T_{S^*}(N_1) = T(N) - T_{S^*}(N)$. Then it is evident that each node in $T_{S^*}(N \cup N_1)$ has more than one edge in $E(G(A, b))$ going from the node to some nodes in $N \cup N_1$.

If $T(N \cup N_1) = T_{S^*}(N \cup N_1)$, $G(A, b)$ does not satisfy property (b2). If $T(N \cup N_1) - T_{S^*}(N \cup N_1)$ is not empty, repeating the above procedure, we can construct a subset of Z which does not satisfy property (b2) since $|Z|$ is finite. Q.E.D.

LEMMA 7. *If the graph $G(A, b)$ is spanned by an s.b.c. and satisfies properties* (c2) *of Theorem* 2 *and* (b2) *of Theorem* 1, *then the s.b.c. which spans $G(A, b)$ is unique.*

Proof. Induction on the dimension n of the state space is used for the proof. Lemma 7 is evidently true when $n = 1$. Assume that Lemma 7 is true when $n = k$ and that there exist more than one s.b.c. which spans $G(A, b)$ when $n = k + 1$. Then we can derive contradictions as follows: When $n = k + 1$, choose two s.b.c.'s $S = B_0 \cup B_1 \cup \cdots \cup B_p$ and $\bar{S} = \bar{B}_0 \cup \bar{B}_1 \cup \cdots \cup \bar{B}_{\bar{p}}$ both of which span $G(A, b)$, where

B_0 ($\bar{B}_0$) is the stem of S ($\bar{S}$) and B_j ($\bar{B}_j$) is the jth bud of S ($\bar{S}$) for $1 \leqq j \leqq p$ ($1 \leqq j \leqq \bar{p}$). Let us label all the nodes of $G(A, b)$ associated with S and $\bar{S}$ at the same time.[2]

In the case that S and $\bar{S}$ are not stems, we define positive integers h and $\bar{h}$ as follows:

$$(12)\quad h = \begin{cases} 1 & \text{if } p = 1, \\ p & \text{if } p > 1 \text{ and } d_j = m_j,\ j = 1, 2, \cdots, p-1, \\ \min\{j \mid d_j \neq m_j,\ 1 \leqq j \leqq p-1\} & \text{otherwise.} \end{cases}$$

$\bar{h}$ is defined in the same way as (12) replacing h, p, d_j, m_j with $\bar{h}$, $\bar{p}$, $\bar{d}_j$, $\bar{m}_j$.

Deleting the node $m_0(B_0)$ and the edge $(m_0 - 1(B_0), m_0(B_0))$ from S when S is a stem or S is not a stem and $m_0 > d_0 + 1$, or deleting the node $m_0(B_0)$ and the edges $(m_0 - 1(B_0), m_0(B_0))$ and $(m_j(B_j), 1(B_j))$ for $1 \leqq j \leqq h$ from S when S is not a stem and $m_0 = d_0 + 1$, we can obtain an s.b.c., which will be denoted by S'. From the s.b.c. $\bar{S}$, we can obtain an s.b.c. in the same way as S'. This will be denoted by $\bar{S}'$. Moreover, define the graph G' as the graph obtained by deleting from $G(A, b)$ the node $m_0(B_0)$ and all the edges whose initial nodes or final nodes are $m_0(B_0)$. Then G' is spanned by the s.b.c. S' from the definition of S'. And it is easily shown from Lemma 6 that G' satisfies properties (b2) and (c2).

From the above, the following can be assumed without loss of generality (see Appendix A).

$$(13)\qquad p \geqq 1, \text{ that is, } S \text{ is not a stem;}$$

$$(14)\qquad m_0(B_0) = \bar{m}_0(B_0);$$

$$(15)\qquad m_0(B_0) = d_0 + 1(B_0).$$

G' is also spanned by the s.b.c. $\bar{S}'$ from (14) and the definition of $\bar{S}'$. Since the s.b.c. which spans G' is unique from the assumption of the induction, $S' = \bar{S}'$ is concluded.

Contradictions will be derived in the following two cases:

(i) $\bar{S}$ is a stem or $\bar{S}$ is not a stem and $\bar{m}_0 > \bar{d}_0 + 1$.

(ii) $\bar{S}$ is not a stem and $\bar{m}_0 = \bar{d}_0 + 1$.

In case (i), the final node of the stem of S' or $\bar{S}'$ is the node $m_h(B_h)$ or $\bar{m}_0 - 1(\bar{B}_0)$ respectively. Since $S' = \bar{S}'$, it follows that $m_h(B_h) = \bar{m}_0 - 1(\bar{B}_0)$. Therefore there exists the edge $(m_h(B_h), d_0 + 1(B_0))$ in $E(\bar{S})$ from (14) and (15). Choose the subset N_1 of Z as $N_1 = \{d_0 + 1(B_0)\} \cup \{1(B_j) | 1 \leqq j \leqq h\}$;[3] then it is easily shown from the definition of h that each node in $T_{S^*}(N_1)$ has more than one edge in $E(G(A, b))$ going from that node to some nodes in N_1. This contradicts the property (b2) from Lemma 6.

In case (ii), the final node of the stem S' or $\bar{S}'$ is the node $m_h(B_h)$ or $\bar{m}_{\bar{h}}(\bar{B}_{\bar{h}})$. Since $S' = \bar{S}'$, $m_h(B_h) = \bar{m}_{\bar{h}}(\bar{B}_{\bar{h}})$ is satisfied and we can assume that $\bar{d}_0(\bar{B}_0) \in V(B_1 \cup B_2 \cup \cdots \cup B_h)$ without loss of generality from the definitions of S' and $\bar{S}'$. If $\bar{d}_0(\bar{B}_0) = m_r(B_r)$ for some $1 \leqq r \leqq h$, contradiction is derived in the same way as case (i).

If $\bar{d}_0(\bar{B}_0) \neq m_r(B_r)$ for all $1 \leqq r \leqq h$, define a positive integer t as

$$(16)\quad t = \min\{i \mid \bar{m}_i(\bar{B}_i) = m_j(B_j) \text{ and } 1(\bar{B}_i) \neq 1(B_j) \text{ for some } 1 \leqq j \leqq h,\ 1 \leqq i \leqq \bar{h}\}.^4$$

[2] When all the nodes of $G(A, b)$ are labeled associated with $\bar{S}$, $\bar{m}_0$ or $\bar{m}_j$ ($1 \leqq j \leqq p$) is equal to $|V(\bar{B}_0)|$ or $|V(\bar{B}_j)| - 1$ respectively, and the origin of $\bar{B}_{j+1}$ is denoted by $\bar{d}_j(\bar{B}_j)$ for $0 \leqq j \leqq \bar{p} - 1$.

[3] $h = p$ is satisfied if $\bar{S}$ is a stem, since $\bar{S}'$ is a stem and $h < p$ means that S' is not a stem.

[4] $\bar{m}_{\bar{h}}(\bar{B}_{\bar{h}}) = m_h(B_h)$ is satisfied. If $\bar{d}_0(\bar{B}_0) \neq d_0(B_0)$ and $\bar{d}_0(\bar{B}_0) \neq m_r(B_r)$ for all $1 \leqq r \leqq h$, $1(\bar{B}_1) \neq 1(B_j)$ for all $1 \leqq j \leqq h$, and if $\bar{d}_0(\bar{B}_0) = d_0(B_0)$, there exists at least one edge $(\bar{m}_i(\bar{B}_i), 1(\bar{B}_1))$ ($1 \leqq i \leqq \bar{h}$) which differs from $(m_j(B_j), 1(B_j))$ for all $1 \leqq j \leqq h$ since $\bar{S}$ differs from S. Thus the definition of t makes sense.

Then, from the definitions of S' and $\bar{S}'$, the edges $(\bar{m}_j(\bar{B}_j), 1(\bar{B}_j))$, $1 \leqq j \leqq t$ in $\bar{S}$ are represented by the labeling associated with S as follows:

$$(17) \qquad (m_{j_t}(B_{j_t}), i_t(B_{j_t})), (i_t - 1(B_{j_t}), i_{t-1}(B_{j_{t-1}})), \cdots, (i_2 - 1(B_{j_2}), i_1(B_{j_1}))$$

(if $i_k = 1$ $(2 \leqq k \leqq t)$, $i_k - 1(B_{j_k})$ should be exchanged for $m_{j_k-1}(B_{j_k-1}))$ where $\bar{m}_t(\bar{B}_t) = m_{j_t}(B_{j_t})$, $1(\bar{B}_k) = i_k(B_{j_k})$ for $1 \leqq k \leqq t$, $\bar{m}_{k-1}(\bar{B}_{k-1}) = i_k - 1(B_{j_k})$ $(m_{j_k-1}(B_{j_k-1})$ if $i_k = 1)$ for $2 \leqq k \leqq t$, $\bar{d}_0(\bar{B}_0) = i_1 - 1(B_{j_1})$ $(m_{j_1-1}(B_{j_1-1})$ if $i_1 = 1)$. Moreover $j_{k-1} \leqq j_k$ and if $j_{k-1} = j_k$, $i_{k-1} \leqq i_k - 1$.[5] Here define the subset N_2 of Z as $N_2 = \{d_0 + 1(B_0)\} \cup \{1(B_j) | 1 \leqq j \leqq j_t\} \cup \{i_k(B_{j_k}) | 1 \leqq k \leqq t\}$.

In the case that $\bar{d}_0(\bar{B}_0) \neq d_0(B_0)$ and $\bar{d}_0(\bar{B}_0) \neq m_r(B_r)$ for all $1 \leqq r \leqq h$, $\bar{d}_0(\bar{B}_0) = i_1 - 1(B_{j_1})$ is satisfied[6] and there exists the edge $(i_1 - 1(B_{j_1}), d_0 + 1(B_0))$ which is the edge $(\bar{d}_0(\bar{B}_0), \bar{d}_0 + 1(\bar{B}_0))$ in $\bar{S}$. Taking account of the edge $(i_1 - 1(B_{j_1}), d_0 + 1(B_0))$ and the edges in (17), it is evident that each node in $T_{S^*}(N_2)$ has more than one edge ir $E(G(A, b))$ going from that node to some nodes in N_2. This contradicts the propert (b2) from Lemma 6. See Fig. 4.

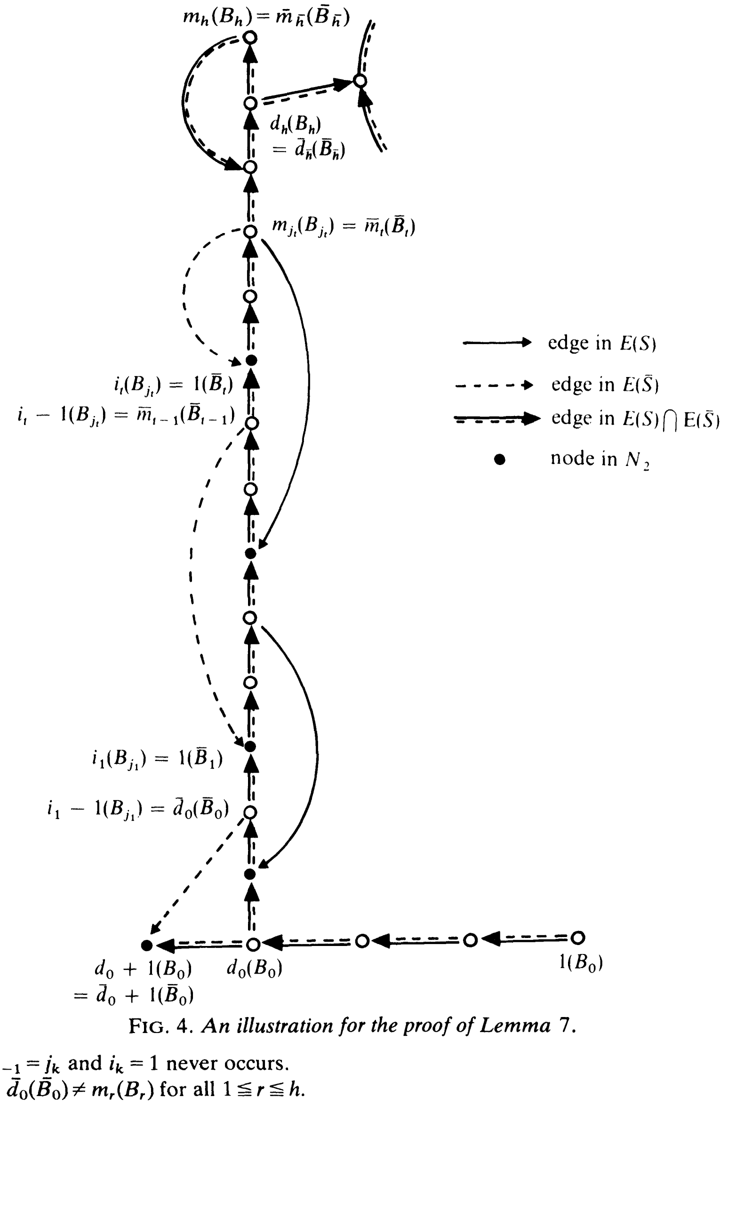
$m_h(B_h) = \bar{m}_{\bar{h}}(\bar{B}_{\bar{h}})$
$d_h(B_h)$
$= \bar{d}_{\bar{h}}(\bar{B}_{\bar{h}})$
$m_{j_t}(B_{j_t}) = \bar{m}_t(\bar{B}_t)$
$i_t(B_{j_t}) = 1(\bar{B}_t)$
$i_t - 1(B_{j_t}) = \bar{m}_{t-1}(\bar{B}_{t-1})$
$i_1(B_{j_1}) = 1(\bar{B}_1)$
$i_1 - 1(B_{j_1}) = \bar{d}_0(\bar{B}_0)$
$d_0 + 1(B_0)$
$= \bar{d}_0 + 1(\bar{B}_0)$
$d_0(B_0)$
$1(B_0)$
edge in $E(S)$
edge in $E(\bar{S})$
edge in $E(S) \cap E(\bar{S})$
● node in N_2

FIG. 4. *An illustration for the proof of Lemma 7.*

[5] The case $j_{k-1} = j_k$ and $i_k = 1$ never occurs.

[6] $i_1 \neq 1$ since $\bar{d}_0(\bar{B}_0) \neq m_r(B_r)$ for all $1 \leqq r \leqq h$.

In the case of $\bar{d}_0(\bar{B}_0)=d_0(B_0)$, $i_1(B_{j_1})=1(B_1)$ is satisfied. Taking account of the edges in (17), the contradiction is also derived in the same way as in the above case. Q.E.D.

LEMMA 8. *If the graph $G(A, b)$ satisfies properties* (c1) *and* (c2) *of Theorem* 2, *it satisfies property* (b2) *of Theorem* 1.

Proof. In the following, the contradiction will be derived by assuming that there exists a subset N of Z such that each node i in $T(N)$ has more than one edge going from the node i to some nodes in N. Since $G(A, b)$ is spanned by a unique s.b.c. S, let us represent the s.b.c. S by $B_0\cup B_1\cup\cdots\cup B_p$ where B_0 is a stem and $B_1, B_2, \cdots, B_p$ are buds, and label all the nodes in $G(A, b)$ associating with S.

Now let us define a nonnegative integer s as

$$s=\begin{cases}0 & \text{if } V(B_0)\cap N\neq\varnothing,\\ \min\{j\mid V(B_j^*)\cap N\neq\varnothing, 1\leqq j\leqq p\} & \text{if } V(B_0)\cap N=\varnothing,\end{cases}\tag{18}$$

and positive integers i_m and i_M as

$$i_m=\min\{i\mid \text{node } i(B_s)\in N, 1\leqq i\leqq m_s\},\tag{19}$$

$$i_M=\max\{i\mid \text{node } i(B_s)\in N, 1\leqq i\leqq m_s\}.\tag{20}$$

Then we can show (Appendix B) the following:

$$\text{node } 1(B_s)\notin N,\tag{21}$$

$$s<p \text{ (i.e. } B_{s+1} \text{ exists.)},\tag{22}$$

$$N\cap\{\text{node } k(B_s)\mid d_s+2\leqq k\leqq m_s\}=\varnothing \quad \text{if } m_s\geqq d_s+2,\tag{23}$$

$$m_s\geqq d_s+1 \text{ and node } d_s+1(B_s)\in N\cap V(B_s^*),\tag{24}$$

$$\text{node } 1(B_{s+1})\in N \text{ (this implies node } m_{s+1}(B_{s+1})\in T(N).).\tag{25}$$

If node $1(B_j)\in N$ and node $m_j(B_j)\in T(N)$ are satisfied for some j $(s+1\leqq j\leqq p-1)$, there exists, in addition to the edge $(m_j(B_j), 1(B_j))$, at least one edge going from the node $m_j(B_j)$ to some node in $N\cap(V(B_s^*)\cup V(B_{s+1}^*)\cup\cdots\cup V(B_j^*))$ or $d_j=m_j$ and node $1(B_{j+1})\in N$ are satisfied from the assumption concerning N, the property (d3) and (18). By the iterative use of above fact, since the number of buds in S is finite and (25) is satisfied, it can be shown that there exists a positive integer t such that $m_{s+j}=d_{s+j}$ is satisfied for $1\leqq j\leqq t-1$ if $t\geqq 2$, node $1(B_{s+j})\in N$ is satisfied for $1\leqq j\leqq t$ and there exists at least one edge going from the node $m_{s+t}(B_{s+t})$ to some node $i_1(B_{j_1})$ in $N\cap(V(B_s^*)\cup V(B_{s+1}^*)\cup\cdots\cup V(B_{s+1}^*))-\{\text{node } 1(B_{s+t})\}$.

In the above, if the node $i_1(B_{j_1})$ is contained in $N\cap\{\text{node } k(B_s)\mid i_m\leqq k\leqq d_s\}$, since the node $i_1-1(B_s)$ is contained in $T(N)$,[7] using (23), (24) and properties (d1) and (d3), we can show that there exists an edge $(i_1-1(B_s), i_2(B_s))$ such that $i_2\leqq i_1-1$ or $i_2=d_s+1$ is satisfied and the node $i_2(B_s)$ is contained in N. If $i_2\leqq i_1-1$ (i.e. $i_2\neq d_s+1$) is satisfied, there exists an edge $(i_2-1(B_s), i_3(B_s))$ such that $i_3\leqq i_2-1$ or $i_3=d_s+1$ is satisfied and the node $i_3(B_s)$ is contained in N from the same reason. Repeating this procedure, we can conclude that there exist edges, $(m_{s+t}(B_{s+t}), i_1(B_s))$, $(i_1-1(B_s), i_2(B_s)), \cdots, (i_{f-1}-1(B_s), i_f(B_s))$, $(i_f-1(B_s), d_s+1(B_s))$ which satisfy $d_s\geqq i_1>i_2>\cdots>i_f\geqq i_m$. In this case and in the case that the node $i_1(B_{j_1})$ coincides with the node $d_s+1(B_s)$ or some node in $\{\text{node } 1(B_{s+j})\mid 1\leqq j\leqq t-1\}$, we can find another

[7] $i_1-1\geqq 1$ is satisfied from (21).

s.b.c. than S which spans $G(A, b)$, and this contradicts the uniqueness of S. Thus it is shown from the above and (23) that

$$\text{node } i_1(B_{j_1}) \in N \cap \bigcap_{j=1}^{t} (V(B^*_{s+j}) - \{\text{node } 1(B_{s+j})\}). \tag{26}$$

Since the node $i_1 - 1(B_{j_1})$ is contained in $T(N) \cap V(B^*_{j_1})$ from (26), there exists an edge going from the node $i_1 - 1(B_{j_1})$ to some node $i_2(B_{j_2})$ in $N \cap (V(B^*_s) \cup V(B^*_{s+1}) \cup \cdots \cup V(B^*_p)) - \{\text{node } i_1(B_{j_1})\}$ from (18). Moreover $i_1 - 1 \neq d_{j_1}$ is satisfied for $s+1 \leqq j_1 \leqq s+t-1$ from (26) and $m_{s+j} = d_{s+j}$ for $1 \leqq j \leqq t-1$; besides even if $j_1 = s+t$ and $s+t \leqq p-1$ are satisfied, $i_1 - 1 < d_{s+t}$ can be obtained from (d2). Thus the node $i_2(B_{j_2})$ is not contained in $\{\text{node } k(B_{j_1}) \mid i_1 \leqq k \leqq m_{j_1}\} \cup (V(B^*_{j_1+1}) \cup V(B^*_{j_1+2}) \cup \cdots \cup V(B^*_p))$ from properties (d1), (d3) and (d4). And by the same argument as in the proof of (26), it can be shown that

(27)

$$\text{node } i_2(B_{j_2}) \in N \cap \left(\bigcup_{j=1}^{j_1-s-1} (V(B^*_{s+j}) - \{\text{node } 1(B_{s+j})\}) \cup \{\text{node } k(B_{j_1}) | 2 \leqq k \leqq i_1 - 1\} \right)$$

where $\{\text{node } k(B_{j_1}) \mid 2 \leqq k \leqq i_1 - 1\} = \varnothing$ in the case of $i_1 = 2$.

Since $|\bigcup_{j=1}^{t} V(B^*_{s+j})|$ is finite, by the iterative use of the same argument which derived (27) from (26), we can finally obtain a node $i_q(B_{j_q})$ in N such that the node $i_q - 1(B_{j_q}) (\in T(N))$ has no edge going from the node $i_q - 1(B_{j_q})$ to some node in N other than the edge $(i_q - 1(B_{j_q}), i_q(B_{j_q}))$. This contradicts the assumption concerning N. Q.E.D.

Proof of Theorem 2. *Necessity.* If the system (A, b) is strongly structurally controllable, properties (b1), (b2) and (b3) are satisfied from Theorem 1. By the use of Lemma 4, property (c2) is derived from properties (b1) and (b3). Since a strongly structurally controllable system (A, b) is also structurally controllable, $G(A, b)$ is spanned by a cactus from (a2) of Lemma 1. This cactus should be an s.b.c. since $G(A, b)$ satisfies property (c2). Thus using Lemma 7, we can derive property (c1) from properties (b2) and (c2).

Sufficiency. Property (b1) is derived immediately from property (c1). And property (b3) is derived from properties (b1) and (c2) by Lemma 4. Moreover property (b2) is derived from properties (c1) and (c2) by the use of Lemma 8. So by Theorem 1, we can conclude that the system (A, b) is strongly structurally controllable. Q.E.D.

Remarks. If the graph $G(A, b)$ satisfies property (c2) of Theorem 2 and is spanned by two different s.b.c.'s S_1 and S_2, then the relation between S_1 and S_2 is reduced to the two cases, the typical examples of which are illustrated in Fig. 5. From this, we can examine the uniqueness of the s.b.c. which spans $G(A, b)$. That is, if an s.b.c. S which spans $G(A, b)$ is given, regarding S as S_1 or S_2, check the existence of S_2 or S_1 which satisfies the relation, respectively. For sparse systems, this procedure also will be performed easily by inspection. Although we give no proof here, we can guess the relation between S_1 and S_2 from the proof of Lemma 7.

Conclusions. The two necessary and sufficient graph theoretic conditions for single-input linear control systems to be strongly structurally controllable have been developed.

Theorem 1 shows the fundamental conditions for strong structural controllability and the essential sets of indeterminate entries, the change of whose values may cause a system to be uncontrollable. These results may give some insight to structural controllability of systems with dependent indeterminate parameters, which is left as a further research problem. Theorem 1 can be extended to the multi-input case directly.

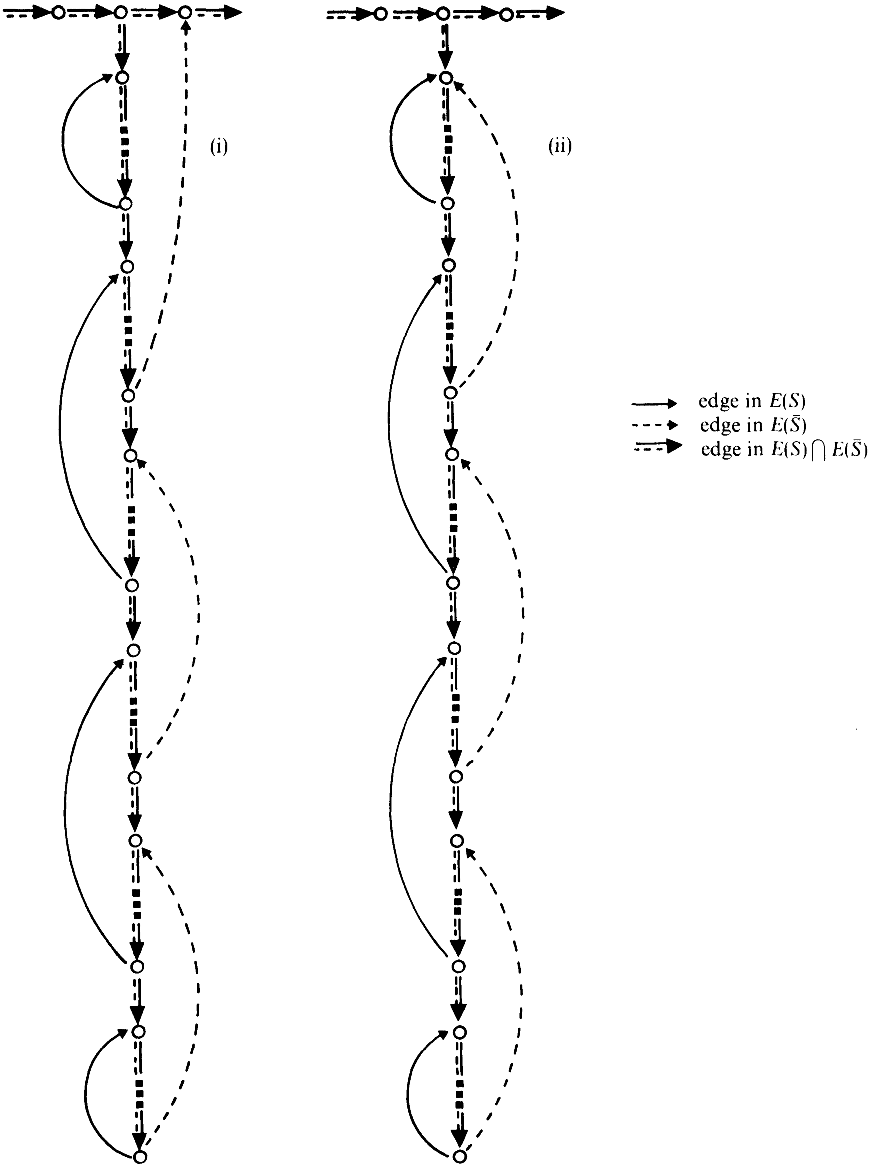

FIG. 5. *The typical examples in which two different s.b.c.'s span the graph $G(A, b)$ at the same time.*

Theorem 2 is useful because of its simple and intuitive form in graph a theoretic aspect. For sparse systems, the conditions of Theorem 2 can be easily examined by inspection. The complete and efficient computer algorithms to test the strong structural controllability in general case are currently under investigation.

Appendix A. We shall verify (13)–(15). First, let us assume that both S and $\bar{S}$ are stems. From the definition of a stem, it is satisfied that $1(B_0)=1(\bar{B}_0)=$ the origin of $G(A, b)$. If $2(B_0)\neq 2(\bar{B}_0)$, the edge $(1(\bar{B}_0), 2(\bar{B}_0))$ is represented as the edge $(1(B_0), r_0(B_0))$ $(3\leqq r_0\leqq m_0)$. Since $\bar{S}$ is a stem, there exists the edge represented as $(r_1(B_0), r_2(B_0))$ $(2\leqq r_2<r_0\leqq r_1\leqq m_0)$ which is contained in $\bar{S}$, but this contradicts

property (d4). Hence $2(B_0)=2(\bar{B}_0)$. By the use of the same arguments, we can show that $S=\bar{S}$, but this contradicts the assumption that $S\neq\bar{S}$.

Next, if $m_0(B_0)\neq\bar{m}_0(\bar{B}_0)$, there exists at least one edge contained in $\bar{S}$ whose initial node is $m_0(B_0)$. And we may assume from (13) and properties (d2) and (d3) that there exists the edge in $\bar{S}$ represented as $(m_0(B_0), s_0(B_0))$ where $2\leqq s_0\leqq d_0$. Then it is easily shown from (d1)–(d4) that the stem $\bar{B}_0$ of $\bar{S}$ does not contain the edge $(m_0(B_0), s_0(B_0))$. Now, if $\bar{m}_0(\bar{B}_0)\notin\{2(B_0), 3(B_0), \cdots, s_0(B_0)\}$, then it follows from the above that there exists a path $\bar{B}_0$ from $1(B_0)$ $(=1(\bar{B}_0))$ to $\bar{m}_0(\bar{B}_0)$ which does not contain the node $s_0(B_0)$, but the presence of such a path contradicts the properties (d1)–(d4). So the node $\bar{m}_0(\bar{B}_0)$ is contained in $\{2(B_0), 3(B_0), \cdots, s_0-1(B_0)\}$. Since the node $\bar{m}_0(\bar{B}_0)$ is the final node of $\bar{B}_0$, from the definition of an s.b.c., there exists the path in $\bar{S}$ from the node $1(B_0)$ to the node $s_0(B_0)$ which does not contain the node $\bar{m}_0(\bar{B}_0)$, but it contradicts property (d3) or (d4).

Last, assume that $m_0>d_0+1$. $\bar{S}'$ also spans G' from (14) and the definition of $\bar{S}'$. Since the s.b.c. which spans G' is unique, $S'=\bar{S}'$ is satisfied. If $\bar{S}$ is a stem, $\bar{S}'$ is also a stem. But S' is not a stem from $m_0>d_0+1$. Therefore $\bar{S}$ is not a stem. If $\bar{m}_0>\bar{d}_0+1$ is satisfied, $S=\bar{S}$ is derived from $S'=\bar{S}'$ and the definitions of S and $\bar{S}$. This contradicts the assumption $S\neq\bar{S}$. Therefore we conclude $m_0=d_0+1$ or $\bar{m}_0=\bar{d}_0+1$. So, $m_0=d_0+1$ can be assumed without loss of generality.

Appendix B. We shall verify (21)–(25). If $1\leqq s\leqq p$, node $1(B_s)\in N$ implies node $d_{s-1}(B_{s-1})\in T(N)$. But the edge $(d_{s-1}(B_{s-1}), 1(B_s))$ is the only edge going from the node $d_{s-1}(B_{s-1})$ to some node in N because of property (d3) and (18). This contradicts the assumption concerning the set N. Moreover N does not contain the node $1(B_0)$ (= the origin of $G(A, b)$). Thus we can obtain (21), that is,

$$\text{(B.1)} \qquad \text{node } 1(B_s)\notin N.$$

If $s=p$ is satisfied, node $i_M-1(B_p)\in T(N)$ is satisfied from (20) and (B.1). So there exists an edge going from the node $i_M-1(B_p)$ to some node $i_1(B_p)$ in $N\cap\{\text{node } k(B_p)\,|\, i_m\leqq k\leqq i_M-1\}$ from (18), (20) and the assumption concerning the set N. It can be shown similarly as above that there exists an edge going from the node $i_1-1(B_p)$ to some node $i_2(B_p)$ in $N\cap\{\text{node } k(B_p)\,|\, i_m\leqq k\leqq i_M, k\neq i_1\}$. If $i_2=i_M$ is satisfied, another s.b.c. than S which spans $G(A, b)$ can be easily constructed. This contradicts the uniqueness of S. If the node $i_2(B_p)$ is contained in $\{\text{node } k(B_p)\,|\, i_1+1\leqq k\leqq i_M-1\}$; property (d4) is contradicted since $i_1-1<i_1<i_2\leqq i_M-1$ is satisfied. Thus the node $i_2(B_p)$ should be contained in $N\cap\{\text{node } k(B_p)\,|\, i_m\leqq k\leqq i_1-1\}$. Continuing this procedure, we can show that the node $i_m-1(B_p)$ is contained in $T(N)$ and the edge $(i_m-1(B_p), i_m(B_p))$ is the only edge going from the node $i_m-1(B_p)$ to some node in N. This contradicts the assumption concerning the set N. So we can derive (22), that is,

$$\text{(B.2)} \qquad s<p \text{ (i.e. } B_{s+1} \text{ exists.).}$$

Suppose $N\cap\{\text{node } k(B_s)\,|\, d_s+2\leqq k\leqq m_s\}\neq\emptyset$; then node $i_M-1(B_s)\in T(N)$. and $i_M-1\geqq d_s+1$ are satisfied. The node $i_M-1(B_s)$ has an edge going from the node $i_M-1(B_s)$ to some node $i_1(B_s)$ in $\{\text{node } k(B_s)\,|\, i_m\leqq k\leqq i_M-1\}$ from (18), and property (d3) and the assumption concerning the set N. Moreover the node $i_1(B_s)$ is not contained in $\{\text{node } k(B_s)\,|\, d_s+1\leqq k\leqq i_M-1\}$ from (B.2) and property (d2). So the node $i_1(B_s)$ should be contained in $\{\text{node } k(B_s)\,|\, i_m\leqq k\leqq d_s\}$. Here applying the similar argument in the proof of (B.2) and property (d3), we can derive the contradiction.

Thus if $m_s \geqq d_s + 2$ is satisfied, (23), that is,

$$N \cap \{\text{node } k(B_s) \mid d_s + 2 \leqq k \leqq m_s\} = \varnothing \tag{B.3}$$

can be proved.

If node $d_s + 1(B_s) \notin N$ or $d_s(B_s) = m_s(B_s)$ is satisfied, $N \cap \{\text{node } k(B_s) \mid i_m \leqq k \leqq d_s\} \neq \varnothing$ is derived from (19) and (B.3). Since the node $i_m(B_s)$ is contained in N and $i_m \geqq 2$ is satisfied from (B.1), the node $i_m - 1(B_s)$ exists in $V(B_s^*) \cap T(N)$. So there exists an edge going from the node $i_m - 1(B_s)$ to some node in $\{\text{node } k(B_s) \mid i_m + 1 \leqq k \leqq d_s\}$ from the assumption concerning N, property (d3) and (18). This contradicts property (d1) from (B.2). Thus we can obtain (24), that is,

$$m_s \geqq d_s + 1 \text{ and node } d_s + 1(B_s) \in N \cap V(B_s^*). \tag{B.4}$$

The node $d_s(B_s)$ is contained in $T(N)$ from (B.4). If the node $1(B_{s+1})$ is not contained in N, a contradiction is derived from property (d3) and the use of the same procedure in the proof of (B.2). So we can derive (25), that is,

$$\text{node } 1(B_{s+1}) \in N. \tag{B.5}$$

Acknowledgment. The authors wish to express their gratitude to Prof. T. Fujisawa, Prof. N. Suda and Assoc. Prof. Y. Yasuda of Osaka University for their helpful suggestions and encouragement.

REFERENCES

[1] W. K. CHEN, *Applied Graph Theory*, North-Holland, Amsterdam, 1976.

[2] J. P. CONFMAT AND A. S. MORSE, *Structurally controllable and structurally canonical systems*, IEEE Trans. Automatic Control, AC-21 (1976), pp. 129–131.

[3] E. J. DAVISON, *Connectability and structural controllability of composite systems*, Automatica, 13 (1977), pp. 109–123.

[4] F. R. GANTMACHER, *The Theory of Matrices*, Chelsea, New York, 1960.

[5] K. GLOVER AND L. M. SILVERMAN, *Characterization of structural controllability*, IEEE Trans. Automatic Control, AC-21 (1976), pp. 534–537.

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

[7] C.-T. LIN, *Structural Controllability*, IEEE Trans. Automatic Control, AC-19 (1974), pp. 201–208.

[8] V. M. POPOV, *Hyperstability of Control Systems*, Springer-Verlag, New York, 1973.

[9] R. W. SHIELDS AND J. B. PEARSON, *Structural controllability of multiinput linear systems*, IEEE Trans. Automatic Control, AC-21 (1976), pp. 203–212.

[10] L. A. ZADEH AND C. A. DESOER, *Linear System Theory*, McGraw-Hill, New York, 1963.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 1, 1979

0036-0129/79/1701-0011 $01.00/0

ON THE OBSERVABILITY OF POLYNOMIAL SYSTEMS, I: FINITE-TIME PROBLEMS*

EDUARDO D. SONTAG†

Abstract. Different notions of observability are compared for systems defined by polynomial difference equations. The main result states that, for systems having the standard property of (multiple-experiment initial-state) observability, the response to a generic input sequence is sufficient for final-state determination. Some remarks are made on results for nonpolynomial and/or continuous-time systems. An identifiability result is derived from the above.

Introduction. This paper deals with observability problems for (deterministic) control systems defined by simultaneous polynomial difference equations, and for other related classes of systems. These problems are natural from a (mathematical) system-theoretic viewpoint, and a strong motivation for their study is also provided by the goal of obtaining explicit solutions to filtering and regulation problems for rather general, yet tractable, classes of nonlinear systems.

Roughly, questions of observability deal with determining the internal state of a (known) dynamical system on the basis of available input/output data. "Observability" is a fundamental system property, due, among others, to the following reasons:

(a) The modern "state-variable" approach to regulator construction is based upon the possibility of feeding back a function of (good estimates of) the state, which must be obtained via "observers" operating on input/output data (in the linear case, "Luenberger observers").

(b) In the stochastic version of the above, the only known effective solution of the optimal nonlinear filtering problem, the Kalman filter, consists precisely of an effective observer construction (for a deterministic system), with parameters optimized on the basis of the available statistical data. This view of Kalman filtering as "deterministic system theory plus elementary theory of Gaussian processes" strongly suggests that a solution in the nonlinear case may be conditional upon a better understanding of nonlinear observers. Moreover, for the known cases, estimation is feasible (in the sense that the error covariance can be made small) only when the system has suitable observability characteristics, as is known for finite-dimensional linear systems (see, e.g., Kwakernaak and Sivan (1972, § 4.4)) and as recently found for infinite-dimensional linear systems (Vinter (1977)).

(c) Observability is one of the main concepts in realization theory, where it appears, under various technical variants, as a characterizing property of canonical systems.

(d) Even in problems not explicitly involving outputs, observability may appear as an important question. To insure the stability of the optimal state regulator, the unstable states must be "observed" by the performance index, as explained intuitively—and proved rigorously in the linear case—in Anderson and Moore (1971, § 3.2).

(e) Problems of identification, i.e., the possibility of determining the input/output behavior of an unknown system on the basis of a limited number of experiments, are closely related to observability questions, as further discussed below.

* Received by the editors April 24, 1977, and in revised form January 31, 1978. This research was supported in part by U.S. Army Grant DA-ARO-D-31-124-72-G114 and by U.S. Air Force Grant 72-2268 through the Center for Mathematical System Theory, University of Florida.

† Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903.

The above rough description of observability as a specific property of systems is highly ambiguous, even at an intuitive, nontechnical, level. This ambiguity arises mainly in the following senses: it is not clear whether the state to be determined is that which existed *before* or *after* experimentation, nor whether *simple* or *multiple* experimentation is allowed, nor whether the steps in the experiments can be modified according to partial information (*open*- vs. *closed-loop* observation). Finally, other, rather different, interpretations are possible; for instance, state determination may be only "asymptotic" in that an infinite procedure permits obtaining progressively better estimates of the internal state, as opposed to the above "finite-time" interpretation, where states are precisely determined after experimentation.

As an example of the different possibilities, the canonical realization of any given input/output behavior is multiple-experiment initial-state observable, while an observer is a device solving a single-experiment final-state problem. Thus, for instance, regulator synthesis via the design philosophy "obtain a canonical realization/build an observer/feed-back 'observed' variables" presupposes a positive answer to the question: "is a canonical realization necessarily final-state observable ("reconstructible")?"

Possible observability notions. The main variations on the notions of observability to be studied and compared are, at an intuitive level:

(a) *Observability*: this terminology is reserved for the standard multiple-experiment initial-state notion. A system is observable when any two states can be distinguished by some input/output experiment. Since the experiment (i.e., the input to be applied) depends on the pair of states to be distinguished, practical determination of an initial state assumes the possibility of somehow resetting the system to this (unknown) state after experimentation, or alternatively having a number of copies of the original system, all in the same initial state. This notion of observability appears naturally in realization theory, since "canonical" or "minimal" realizations usually exhibit technical variants of this property (e.g., "algebraic observability," when each coordinate of the initial state can be obtained by *algebraic* manipulations—additions and multiplications—of input data; this property characterizes "canonical" polynomial systems, as discussed in Sontag and Rouchaleau (1975), Sontag (1976a)).

(b) *Single-experiment observability*: there exists a single input (over some finite time interval) which by itself permits the determination, through measurement of ensuing outputs, of the initial state. Clearly this is a much more desirable property than (a); it turns out to be, however, rather restrictive for *discrete*-time systems. (This is not surprising; already Moore (1956) showed that (a) and (b) are far from equivalent, at least for finite automata. For *linear* systems (a) is equivalent to (b), and in fact any long-enough input distinguishes any pair of states, as discussed for instance in Kalman (1968) or Wonham (1974).)

(c) *Final-state determinability*: there is an input sequence w which permits determination of the state of the system resulting *after* w is applied. (In other words, if two states produce the same output sequence under input w, then these two states are necessarily sent into the same state under the action of w.) This property is of interest from a control viewpoint, since control actions can be taken after the state of the system is determined, independently of the state before experimentation. Of course, (b) implies (c). What is not clear is what are the relations, if any, between (a) and (c), since in the former case multiple experimentation is required. It is known that, for finite automata, (a) (called in automata theory a "diagnosing" problem) implies (c) ("homing" problem). This was proved by Moore (1956); expositions are given by Gill

(1962) and Conway (1971); applications to regulation are given by Gatto and Guardabassi (1976). The same result holds for certain types of finite-dimensional systems—e.g., Theorem 4.8 below—; proofs are in fact totally analogous to the finite automata case, with a new type of finiteness (algebraic, linear, or analytic) replacing a set-theoretic finiteness.

(d) *Generic final-state determinability*: while (c) concerns the *existence* of an input such that final states can be determined by testing the system with this input, (d) concerns the much more desirable case when no "experimentation" is needed, but, strictly speaking, "observation" of the input/output behavior is enough. The extreme case of (d) would correspond to that case in which *any* (long-enough) input permits final-state determination. This extreme case is easily seen to be too restrictive, but it may be weakened to only requiring that "almost any" (i.e., a "generic") long-enough input permits this determination. (The rigorous definition of "generic" is a purely technical question, to be discussed later.) In other words, real-time observation of a system, not influencing it in any way (or even, observation of data from past behavior) should be enough for final-state determination. This property is totally different from (c), except in the very special case of linear systems, where (c) = (d). In the automata-theoretic case, "genericity" cannot be even *defined* in a satisfactory way, so this is a genuinely *new* system-theoretic concept.

The main result of this paper states that (a) implies (d) for polynomial systems. Thus, for instance, final states can be determined for canonical realizations of polynomial systems, just observing the "generic" input/output behavior. The proof of the main result uses some elementary notions from algebraic geometry. Since all results remain true when system parameters are not necessarily real or complex but belong to an arbitrary field, everything is stated for arbitrary infinite fields (the finite field case belongs properly to finite automata theory; infinite fields permit identifying polynomials and polynomial functions). Some technical variants of the above observability properties are also discussed and relations between all such notions are clarified.

The last section deals with (i) the particular case of state-affine systems, (ii) generalizations to related classes of systems, in particular state-analytic and continuous-time analytic, and (iii) a restatement of the main result as a system identification problem.

This paper does not treat questions of closed-loop and/or asymptotic observability (closely related to problems of stability), nor the effective construction of "observers." Another interesting set of problems left open is that of finding numerical values for smallest lengths of observability experiments; except for the state-affine case, only qualitative results are given (even for the case of finite automata many of these problems are still unresolved; see Conway (1971)).

The results of this paper strongly suggest that the proper definition of "observer" in the nonlinear context may be that of a dynamical system which determines the state of the "observed" system on the basis of a *generic* set of data.

1. Definitions and characterizations. Let k be an arbitrary but fixed infinite field, and m, n, p arbitrary positive integers. Recall that an *algebraic subset* S of the affine space k^q, $q \geqq 0$, is a set defined by polynomial equations $S = \{Q_i(x_1, \cdots, x_q) = 0\}$. An *irreducible* algebraic set is one which cannot be expressed as the union of two proper algebraic subsets. In this context, a subset R of an irreducible algebraic set S is *generic* when its complement is contained in a proper algebraic subset of S. (These definitions are justified by the fact that for $k = \mathbb{R}$ or $\mathbb{C}$, a proper algebraic set is "thin" in most possible senses, including Baire category and measure-theoretic.)

DEFINITION 1.1. A (discrete-time) *polynomial system* Σ is given by a set of equations

$$x(t+1)=P(x(t),u(t)),\qquad y(t)=h(x(t)),\qquad t=0,1,2,\cdots,$$

where *inputs* $u(t)$, *states* $x(t)$ and *outputs* $y(t)$ belong to algebraic subsets U of k^m, X of k^n, and Y of k^p respectively, U is irreducible, and $P:X\times U\to X$ and $h:X\to Y$ are polynomial maps.

Allowing proper algebraic subsets, rather than insisting on finite dimensional spaces, for U, X, Y, permits increasing the generality of the results to include input or state constraints of a polynomial type. The irreducibility assumption on U is made purely for technical convenience. For instance, the unit real circle $U=\{x^2+y^2-1=0\}$, as well as any space k^m, are admissible input sets. Nonpolynomial systems will be considered later.

Some extra notation will be useful. The extension of P to input sequences is also denoted by $P:X\times U^*\to X$ (for the empty sequence e, $P(x,e)=x$). Applying an input sequence $w=u_1\cdots u_r$ to a system in state x produces an output sequence

$$H^w(x)=(h(x),h(P(x,u_1)),\cdots,h(P(x,w)))$$

in Y^{r+1}.

In what follows, Σ is a fixed polynomial system. The input sequence w *distinguishes* between the states x and z iff $H^w(x)\neq H^w(z)$. The following are several possible definitions of "observability":

(A) *Single-experiment observability*: there exists an input sequence w which distinguishes every pair of states.

(B) *Single-experiment observability with a generic input*: there are a positive integer r and a generic subset R of U^r such that any w in R distinguishes every pair of states.

(C) *Observability*: each pair of states can be distinguished by some input sequence.

(D) *Finite observability*: there are a positive integer r and input sequences $w_1,\cdots,w_s$ of length r such that each pair of states x, z is distinguished by some w_i.

(E) (*Finite*) *observability with generic inputs*: there are integers r, s and a proper generic subset R of U^{rs} such that (D) holds for any set $w_1,\cdots,w_s$ of inputs of length r for which $(w_1,\cdots,w_s)$ is in R.

(F) *Algebraic observability*: for each polynomial function $\hat{q}:X\to k$ there are input sequences $w_1,\cdots,w_s$ and a polynomial function $q:Y^s\to k$ such that $\hat{q}(x)=q(h(P(x,w_1)),\cdots,h(P(x,w_s)))$ for all x in X.

(G) *Final-state determinability*: there is an input sequence w such that for each pair of states x, z either $H^w(x)\neq H^w(z)$ or $P(x,w)=P(z,w)$.

(H) *Final-state determinability with generic inputs*: there are a positive integer r and a generic subset R of U^r such that (G) holds for all w in R.

The characterizations below are useful in checking observability. They are stated in terms of the polynomial functions h_{ij} defined as follows by induction on j. First, an (infinite) basis B is chosen for the vector space of all polynomial functions on U. (If $U=k^m$, the natural choice is the set of all m-variable monomials; if U is a proper algebraic set one may choose a linearly independent subset of such monomials.) The polynomial map $h:X\to Y\subseteq k^p$ gives rise to p polynomial functions

$$h_{01},\cdots,h_{op}$$

by composing with the coordinate projections. If the h_{ij} have been defined for some i

and $j = 1, \cdots, q_i$, one may express

$$h_{ir}(P(x, u)) = \sum_s a_{rs}(x)g_s(u), \qquad r = 1, \cdots, q_i \tag{1.2}$$

for some finite subset $g_1, \cdots$ of B. The $h_{i+1,j}$ are then given by the a_{rs}, $r = 1, \cdots, q_i$, all s, listed in any order except that an a_{rs} is dropped if it is redundant, i.e., if a_{rs} is in the algebra generated by the previous h_{ij}'s.

LEMMA 1.3. (a) Σ *is observable iff the map*

$$x \to (h_{11}(x), h_{12}(x), \cdots, h_{21}(x), \cdots) \tag{1.4}$$

is one-to-one.

(b) Σ *is algebraically observable iff each coordinate function* $x_i : X \to k$, $i = 1, \cdots, n$, *is a polynomial combination of the* $h_{ij}(x)$.

Proof. Observability clearly implies that (1.4) is one-to-one, since the functions $x \mapsto h(P(x, w))$ are combinations of the h_{ij}. Conversely, from Sontag (1976a, "Main lemma" (10.7)), the $h_{ij}(\cdot)$ are linear combinations of the functions $h(P(\cdot, w))$; it follows that if x, z are indistinguishable then $h_{ij}(x) = h_{ij}(z)$ for all i, j. The proof of (b) is similar.

The above result permits checking observability without having to consider, for each pair of states, if there is an input sequence separating them. The result can be tightened considerably, in that it is theoretically possible to specify an integer s (which depends only on the degrees of the polynomials defining Σ) such that it is enough to check, in order to determine (algebraic) observability, if the map

$$X \to Y^{sr_s} : x \to (h_{11}(x), \cdots, h_{sr_s}(x)) \tag{1.5}$$

is one-to-one (or if each coordinate function is a combination of the h_{ij}'s); this follows from the decidability theory in commutative algebra, as remarked in Sontag and Rouchaleau (1975). The problem of checking if (1.5), or a general polynomial map, is one-to-one is very difficult, and it appears also in trying to determine if a system is observable with respect to a *fixed* input $w = u_1 \cdots u_r$, since one must then check

$$x \mapsto (h(x), h(P(x, u_1)), \cdots, h(P(x, w)));$$

in that context, sufficient conditions for one-to-oneness (with k = reals) were surveyed by Fitts (1972).

As a very simple illustration of Lemma 1.3, take the polynomial system Σ_1 with equations

$$x_1(t+1) = x_2(t), \qquad x_2(t+1) = x_1(t), \qquad x_3(t+1) = x_3(t),$$
$$x_4(t+1) = x_1(t)u_1^2(t) + x_2(t)u_2^2(t) + x_3(t),$$
$$y(t) = x_4(t),$$

where $U = k^2$, $X = k^4$, $Y = k$. Then h_{01} = the coordinate function x_4. From the fourth equation, and noting that u_1^2, u_2^2, 1 are linearly independent functions, one has x_1, x_2, x_3 for the h_{1j}. Thus Σ is algebraically observable, and in particular observable.

If, instead, now U is the circle $u_1^2 + u_2^2 = 1$, then $u_2^2 = 1 - u_1^2$ as functions on U, so

$$x_1u_1^2 + x_2u_2^2 + x_3 = (x_1 - x_2)u_1^2 + (x_2 + x_3),$$

so $h_{11} = x_2 + x_3$, $h_{12} = x_1 - x_2$. Now, in obtaining the h_{2j}, $x_1 - x_2$ yields $x_2 - x_1$ (from the first two equations), which is $-(x_1 - x_2)$ and hence belongs to the algebra generated by previous h_{ij}'s. On the other hand, $x_2 + x_3$ yields $x_1 + x_3$, which is equal to $(x_1 - x_2) +$

(x_2+x_3), hence in the algebra generated by previous h_{ij}'s. Thus no h_{ij} are added for $i=2, 3, \cdots$. The system is therefore *not* observable, since

$$(x_1, x_2, x_3, x_4) \mapsto (x_4, x_1-x_2, x_2+x_3)$$

is not one-to-one. In fact, the indistinguishable pairs of states are those in the lines parallel to $\{x_4=0, x_1-x_2=0, x_2+x_3=0\}$.

When $k=$ reals or complexes, observability can be checked using only inputs of arbitrarily small amplitude; this is easily derived from the above characterization using Sontag (1976a, Lemma (2.11)).

2. Implications among observability notions.

THEOREM 2.1. *With the notations in the previous section, the only implications are those indicated by the following diagram*:

(2.2)
$$\begin{tikzcd}
 & & F \arrow[d] & & & & \\
B \arrow[r] & A \arrow[r] & C \arrow[r, equal] & D \arrow[r, equal] & E \arrow[r] & H \arrow[r] & G.
\end{tikzcd}$$

Proof. The following implications are immediate from the definitions: E → D → C, B → A → C, H → G, and F → C. That C → D is proved in Sontag and Rouchaleau (1975, Prop. 7.2). Proofs are given below for D → E (2.4) and C → H (Theorem 3.5). To complete the proof of 2.1, counterexamples must be given to B → F, A → B, F → A, G → H and H → C. For the latter it is sufficient to consider the trivial system with both transition and output maps equal to zero: after one step, the state is known (zero), no matter which input was "applied", but the initial state cannot be determined. The remaining counterexamples are given by:

B → F: let $k=\mathbb{R}$, $X=Y=k$, $U=$ arbitrary, $P(x, u)=0$ for all x, u, and $h(x)=x^3$.

A → B: let $U=Y=k$, $X=k^2$, and Σ_2 given by

$$x_1(t+1)=0, \qquad x_2(t+1)=x_1(t)+x_1^2(t)u(t), \qquad y(t)=x_2(t).$$

An input $w=u_1 \cdots u_r$ distinguishes initial states if and only if $u_1=0$. But the set of all such inputs is not generic in U^r, for any r.

F → A: let $U=Y=k$, $X=k^2$, and Σ_3 given by

$$x_1(t+1)=0, \qquad x_2(t+1)=x_1(t)u(t)-x_1^2(t), \qquad y(t)=x_2(t).$$

Algebraic observability follows from criterion 1.3: recursively, one generates x_2 and then x_1 (and x_1^2, which is redundant). But no single sequence w serves to distinguish every pair of states: let $w=uw'$, with u in U; if $u=0$ then $(1\ 0)'$ and $(-1\ 0)'$ are not distinguished by w, while if $u \neq 0$ then $(u\ 0)'$ is indistinguishable from $(0\ 0)'$.

G → H: let $U=X=k$ and

$$x(t+1)=x(t)u(t), \qquad u(t)=0.$$

Then $w=u_1 \cdots u_r$ determines the final state if and only if some $u_i=0$. The set of all such w is not generic.

It will be now proved that finite observability (D) implies, for polynomial systems, generic finite observability (E). This is somewhat surprising because the corresponding implication for single-experiment observability (A → B) is false. (Σ_3 above is, however, generically finitely observable: any two length-one inputs u, v permit observing $x_1+x_1^2u$, $x_1+x_1^2v$, hence also

$$x_1=[(x_1+x_1^2u)v-(x_1+x_1^2v)u](v-u)^{-1}$$

is known. Thus the generic set R of all (u, v) in $U^2=k^2$ with $u-v \neq 0$ satisfies definition E.)

The following algebraic result is needed; its proof is essentially the same as that in Sontag (1976a, Lemma (10.6)):

LEMMA 2.3. *Let V, W be algebraic sets, W irreducible, and $f: V\times W\to k$ a polynomial function. There exists then an integer s and a nonzero polynomial function $d: W^s\to k$ such that, for each $w, w_1, \cdots, w_s$ in W there are $a_1, \cdots, a_s$ in k with*

$$d(w_1, \cdots, w_s)f(v, w)=\sum_i a_i f(v, w_i).$$

One can now give the

(2.4) *Proof of $D\to E$.* Assume that Σ is finitely observable, and let $\bar{w}_1, \cdots, \bar{w}_t$ be such that $x\neq z$ implies $h(P(x, \bar{w}_i))\neq h(P(z, \bar{w}_i))$ for some i. For each i, let $f_i = h\circ P: x\times U^{r_i}\to k$. Applying 2.3 with $V=X$, $W=U^{r_i}$, $f=f_i$, a $d_i: U^{s_ir_i}\to k$ is obtained for each i. Let q:=largest of the s_i. In the definition of generic observability, take r:=largest of the r_i and $s:=t.q$. An element of U^{rs} can be written as

$$(w_{11}, \cdots, w_{t1}, w_{12}, \cdots, w_{t2}, \cdots, w_{tq}),$$

with each w_{ij} in U. Define the proper algebraic subset F of U^{rs} by the equations

$$d_i(w_{i1}, \cdots, w_{is_i})=0, \qquad i=1, \cdots, t.$$

Then generic observability holds with R = complement of F.

3. Proof of the main result.

LEMMA 3.1. *For any polynomial system Σ there exists an integer $r\geqq 0$ and a proper algebraic subset F of U^r such that, for every $w=(u_1, \cdots, u_r)$ not in F, and for any x, z in X,*

$$H^w(x)=H^w(z)$$

implies that

$$P(x, w) \textit{ is indistinguishable from } P(z, w).$$

Proof. Since $Y\subseteq k^p$ for some integer p and since a union of proper algebraic subsets of U^r is again a proper algebraic subset, it is sufficient to prove the lemma with $Y=k$. The general case can be reduced to this by considering the p projections $Y\to k$.

Let $s\geqq 0$ be such that any pair of distinguishable states is already distinguished by inputs of length $\leqq s$ (Sontag and Rouchaleau (1975, Cor. 7.3)).

For any algebraic set Z, let $A(Z)$ denote the algebra of polynomial functions on Z. Irreducibility of U means that $A(U^t)$ is an integral domain for all t. Let D be the direct limit of the sequence of k-algebras

$$A(U)\to\cdots\to A(U^t)\to A(U^{t+1})\to\cdots,$$

where

$$A(U^t)\to A(U^{t+1})=A(U^t)\otimes A(U): f\to f\otimes 1.$$

Let K be the quotient field of D (which is an integral domain, being a direct limit of integral domains); K contains all $A(U^t)$.

Since $Y=k$, a polynomial map $X\times X\times U^t\to Y$ is an element of $A(X\times X)\oplus A(U^t)$; in particular the functions h_t defined by

$$h_t(x, z, u_1, \cdots, u_t):=h(P(x, u_1, \cdots, u_t))-h(P(z, u_1, \cdots, u_t))$$

are in $A(X\times X)\otimes K$. The latter is a finitely generated algebra over the field K, hence Noetherian. Thus there is some integer r such that all h_t are in the ideal of $A(X\times X)\otimes K$ generated by $h_0, \cdots, h_r$. In particular, there are therefore equations

$$ch_{r+j}=\sum_{t=0}^{r} a_{jt}h_t, \qquad j=1, \cdots s, \tag{3.2}$$

with all a_{jt} in $A(X\times X)\otimes D$ and c a nonzero element of D. Since D is the union of the $A(U^t)$, there is some integer q such that all a_{jt} are in $A(X\times X)\otimes A(U^q)$ and c is in $A(U^q)$. Without loss of generality, we shall assume that $q \geqq r+s$.

Define the proper algebraic set

$$F := \{(u_1, \cdots, u_r) \text{ in } U^r \text{ such that } c(u_1, \cdots, u_r, \cdots, u_q) = 0 \text{ for all } (u_{r+1}, \cdots, u_q)\}.$$

Claim: F satisfies the requirements of the lemma. Indeed, assume that $\underset{\sim}{w} = (u_1, \cdots, u_r)$ is not in F. Take x, z in X such that $h(P(x, u_1, \cdots, u_t)) = h(P(z, u_1, \cdots, u_t))$ for all $t = 0, \cdots, r$, i.e.,

$$h_t(x, z, u_1, \cdots, u_t) = 0, \qquad t = 0, \cdots, r. \tag{3.3}$$

Denote $\underset{\sim}{x} := P(x, \underset{\sim}{w})$, $\underset{\sim}{z} := P(z, \underset{\sim}{w})$. It must be proved that $\underset{\sim}{x}$, $\underset{\sim}{z}$ are indistinguishable.

Assume that $\underset{\sim}{x}$, $\underset{\sim}{z}$ are distinguished by an input sequence v, which can be taken of length j, $0 \leqq j \leqq s$, by definition of s. let

$$F_1 := \{w \text{ in } U^j \text{ such that } h_{r+j}(x, z, \underset{\sim}{w}, w) = 0\};$$

this is an algebraic set, proper because v is not in F_1. Let

$$F_2 := \{w \text{ in } U^j \text{ such that } c(\underset{\sim}{w}, w, w') = 0 \text{ for all } w' \text{ in } U^{q-r-j}\};$$

this is also an algebraic set, and it is proper because $\underset{\sim}{w}$ was taken not in F.

It follows that $F_1 \cup F_2$ is also a proper algebraic set. Let then w be in neither F_1 nor F_2. Then $c(\underset{\sim}{w}, w, w') \neq 0$ for some w', so

$$c(\underset{\sim}{w}, w, w')h_{r+j}(x, z, \underset{\sim}{w}, w) \neq 0. \tag{3.4}$$

But (3.2), (3.3) and (3.4) taken together are contradictory.

THEOREM 3.5. *Observability implies, for polynomial systems, final-state determinability with generic inputs.*

Proof. Immediate from the lemma.

Remark 3.6. As shown in Sontag (1976), canonical realizations Σ_f of polynomial response maps are not, in general, polynomial systems. So the Theorem above is not applicable directly ($A(X_f)$ is not Noetherian). However, if f admits a polynomial realization Σ, then the reachable states of Σ_f form a set which is a quotient of the reachable set of Σ. Then Lemma 3.1 can be applied to Σ, implying that the reachable part of Σ_f does satisfy Theorem 3.5. Another generalization regards the case in which X is a nonaffine variety: taking an affine cover of X, equations as in (3.2) result on each piece of the corresponding decomposition of $X\times X$, and Lemma 3.1 is again true. This generalization is of interest in identifiability questions, with nonaffine parameter spaces.

4. Particular cases, applications, generalizations.

(Polynomial) State-affine systems. For this class of systems, whose realization theory was studied in Sontag (1976b), most of the implications among observability properties are easy generalizations of the linear case.

DEFINITION 4.1. A polynomial system Σ is *state-affine* iff $X = k^n$, $U = k^m$, P is affine (linear + translation) in states, and h is linear.

Fixing a basis in X, the equations for a state-affine system have the form

$$x(t+1) = F(u(t))x(t) + G(u(t)),$$
$$y(t) = Hx(t),$$

where $F(\cdot)$ is a (polynomial) matrix function of u, $G(\cdot)$ is a vector function of u, and H

is a constant matrix. A particular case is that of internally-bilinear systems (see, e.g., Brockett (1972), D'Alessandro, Isidori and Ruberti (1974), Fliess (1973)), when F and G are themselves linear or affine in u.

For state-affine systems the table of implications given in § 2 collapses to

$$\mathrm{A}=\mathrm{B}\rightarrow\mathrm{C}=\mathrm{D}=\mathrm{E}=\mathrm{F}\rightarrow\mathrm{H}\rightarrow\mathrm{G}.$$

It must be proved that C → F and A → B. That C → F is clear from (1.3), since the h_{ij} are linear functions of x, observability thus meaning that the coordinates x_i are *linear* combinations of the h_{ij}. (An explicit matrix criterion for observability is described in Sontag (1976b, Lemma 1.32).) That A → B follows from the following characterization, which can be also generalized to the case $U=$ proper algebraic set by considering a basis of functions $U\rightarrow k$ instead of all monomials u^{α_i}:

PROPOSITION 4.2. *The state-affine system Σ is single-experiment observable iff*

$$\text{(4.3)}\qquad \operatorname{rank}\begin{bmatrix} H & & \\ HF(U_1) & & \\ \vdots & & \\ HF(U_{n-1}) & \cdots & F(U_1) \end{bmatrix}=n$$

over the field $K=k(U_1,\cdots,U_{n-1})$ of rational functions in $m(n-1)$ variables. Moreover, if (4.3) *holds, then any $w=u_1\cdots u_{n-1}$ such that the rank in* (4.3) *remains n after specializing $U_1=u_1,\cdots,U_{n-1}=u_n$ solves the single-experiment observation problem.* (The set of all such w is generic.)

For example, consider the three-dimensional state-affine system Σ_4:

$$x_1(t+1)=x_1(t)u_1(t)+x_2(t)u_2(t)+x_3(t)u_3(t)\quad x_2(t+1)=0,\quad x_3(t+1)=0,$$

$$y(t)=x_1(t).$$

This system is observable, but (with $U_1=[U_{11},U_{21},U_{31}]$, $U_2=[U_{12},U_{22},U_{32}]$) the matrix in (b) is

$$\begin{bmatrix} 1 & 0 & 0 \\ U_{11} & U_{21} & U_{31} \\ U_{12}U_{11} & U_{12}U_{21} & U_{12}U_{31} \end{bmatrix},$$

which has rank two, so the system is not single-experiment observable.

Proof of Proposition 4.2. Single-experiment observability with an input $w=u_1\cdots u_t$ is equivalent to the map $x\rightarrow H^w(x)$ being one-to-one. Since

$$H^w(x)=(Hx,\cdots,HF(u_t)\cdots F(u_1)x)+\text{translation},$$

H^w is one-to-one if and only if the rank of

$$\begin{bmatrix} H & & \\ HF(u_1) & & \\ \vdots & & \\ HF(u_t) & \cdots & F(u_1) \end{bmatrix}$$

is n. Being full rank means that some $n\times n$ minor is nonzero, so the same minor is nonzero as a polynomial in $u_1=U_1,\cdots,u_t=U_t$ (variables) over K. Thus the rank of this matrix is also n. Consider the chain of subspaces V_r of K^n defined by

$$V_r:=\text{span over } K \text{ of } F'(U_1)\cdots F'(U_i)H_j,\qquad i<r,\ j=1,\cdots,p,$$

where H_j is the jth column of H. It is easy to see that if $V_r = V_{r-1}$ for some r then $V_r = V_{r+1} = \cdots$. Thus $V_n = V_{n+1} = \cdots = V_t$. This proves that the rank in (4.3) is n. The rest of the statement is clear from the proof.

The proof of Lemma 3.1 can be rederived for the state-affine case, using only linear-algebraic methods (over rational function fields). A constructive proof is thus obtained, with the precise value $r = n$.

Parametric identification. The result in § 3 can be applied to the following identification problem: a family of polynomial systems is given, parametrized by polynomial functions. It follows that, if the output is known for a generic input, then the future input/output behavior of the system is completely determined. Specifically, considering a family (or "structure"—see Bellman and Aström (1970)):

$$\Sigma_\lambda : \begin{cases} x(t+1) = P(\lambda, x(t), u(t)) = P_\lambda(x(t), u(t)). \\ y(t) = h(\lambda, x(t)) = h_\lambda(x(t)), \end{cases} \tag{4.4}$$

where $P: \Lambda \times X \times U \to x$ and $h: \Lambda \times X \to y$ are polynomial maps and Λ, X, U, Y are algebraic subsets of k^q, k^n, k^m, k^p respectively, U irreducible. The input/output map of Σ_λ for initial state x is

$$f_{\lambda,x} : w \mapsto H_\lambda^w(x).$$

THEOREM 4.5. *There is a positive integer r and a generic subset R of U^r such that, for each input sequence $w = u_1 \cdots u_r$ in R,*

$$f_{\lambda,x}(w) = f_{\mu,z}(w)$$

implies that

$$f_{\lambda,x}(wv) = f_{\mu,z}(wv) \quad \textit{for all input sequences } v.$$

Proof. Let $\hat{\Sigma}$ be the polynomial system with $\hat{X} := \Lambda \times X$, $\hat{U} = U$, $\hat{Y} = Y$ and equations

$$\lambda(t+1) = \lambda(t), \qquad x(t+1) = P(\lambda(t), x(t), u(t)),$$
$$y(t) = h(\lambda(t), x(t)).$$

Then Lemma 3.1 applied to $\hat{\Sigma}$ gives an r and an R such that $\hat{H}^w(\lambda, x) = f_{\lambda,x}(w)$ determines the final state $(\lambda, x(r))$ up to indistinguishability, i.e., all future outputs coincide.

For instance, the future input/output behavior of the system Σ_5 (with $U = Y = k$, $X = k^3$):

$$x_1(t+1) = x_3(t), \quad x_2(t+1) = \lambda x_1(t) + x_2(t), \quad x_3(t+1) = x_2(t)u(t) + x_2(t),$$
$$y(t) = x_3(t)$$

is uniquely determined once that the output corresponding to a $w = u_1u_2u_3$, $u_i \neq -1$ is known, since $x_3(0)$, $x_2(0)$, $\lambda x_1(0)$, and $\lambda x_3(0)$ are successively obtained. (Note that the parameter λ itself is in general not determinable, for instance if $x_3(0)$ is zero; an additional "parameter-identifiability" condition is needed on the given family in order to determine λ.)

The above definition of family of systems includes the case in which the identification is desired of a system of which one only has a bound on dimension and a bound on the degree of the polynomials in its defining equations: it is then obviously enough to add one parameter for each unknown coefficient and one for each coordinate of the (unknown) initial state. (Such a parametrization is of course highly redundant; realization theory may give lower order ones; see Remark 3.6.).

Nonpolynomial systems. Many of the remarks and results of previous sections apply to more general finite-dimensional systems than polynomial systems, i.e., systems

$$x(t+1)=P(x(t),u(t)), \qquad y(t)=h(x(t)) \tag{4.6}$$

where, say, U, X, Y are subsets of $\mathbb{R}^m$, $\mathbb{R}^n$, $\mathbb{R}^p$ (or corresponding complex spaces) and P, h are analytic, infinitely differentiable, or just continuous, either in both x and u, or only in x. The "generic" conditions, defined in terms of algebraic sets, should of course, be redefined according to the category to be worked on (analytic sets, nowhere dense sets, etc.). We conjecture, but have not yet proved, that Theorem 3.5 is true in the analytic case. (In certain cases this is trivially true, e.g. for "analytic state-affine systems," when P, G are analytic in u and linear in x.) The weaker result C→G: observability implies final-state determinability, holds for the following kind of system (analogous definitions for the complex case):

DEFINITION 4.7. A *state-analytic system* Σ has equations (4.6) with X an open subset of $\mathbb{R}^n$, Y a subset of $\mathbb{R}^p$, and both P and h analytic in x.

(U, and the dependence of P on inputs u, are completely arbitrary.)

THEOREM 4.8. *Let the state-analytic system Σ be observable. Let K be any compact subset of X. Then there exists an input sequence w such that, for each pair of states x, z in K, either $H^w(x)\neq H^w(z)$ or $P(x,w)=P(x,z)$.*

Proof. For each input sequence w, let

$$K_w := \{(x,z) \text{ in } K\times K \mid H^w(x)=H^w(z)\}.$$

Each K_w is a subset of the compact set $K\times K$, defined by analytic equations in $X\times X$. It can be proved, using compactness and applying the generalized form of the Weirstrass preparation theorem given by Hervé [1963, Thm. 2.7, Cor. 3], that sets defined by analytic equations satisfy a descending chain condition on compact sets. Thus, there is a minimal K_w.

Then w satisfies the conclusion of the theorem. Indeed, assume that, on the contrary, there is a pair (x,z) in $K\times K$ with $H^w(x)=H^w(z)$ but $P(x,w)\neq P(z,w)$. By observability of Σ, there is an input sequence v such that

$$H^{wv}(x)=H^w(P(x,v))\neq H^w(P(z,v))=H^{wv}(z).$$

So K_{wv} is properly contained in K_w, contradicting minimality of the latter.

Except for our use of the result from analytic functions, the above is essentially the standard proof of C→G for automata (all sets finite, so there is again a minimal K_w) and for internally-bilinear systems (all sets are linear subspaces), in particular as given by Muchnik (1973) and independently (strictly speaking, for continuous-time) by Grasselli and Isidori (1977).

The compactness assumption cannot be dropped: the one-dimensional state-analytic system Σ_6 with equations

$$x(t+1)=\tfrac{1}{2}x(t), \qquad y(t)=\sin x(t)$$

is observable but is not final-state determinable with any (finite length) input. Similarly, infinite differentiability (instead of analyticity) will not be sufficient: consider the one-dimensional system Σ_7 with $X:=(-1,1)$ and

$$x(t+1)=a(x(t)), \qquad y(t)=b(x(t)),$$

where a, b are infinitely differentiable with $a(x)=2x$ on $[-\frac{1}{4},\frac{1}{4}]$ (arbitrary otherwise), and $b(x)=0$ on $[-\frac{1}{4},\frac{1}{4}]$ and bijective in the complement. Then Σ_7 is observable, but pairs of states in $K = [-\frac{1}{4}, \frac{1}{4}]$ do not satisfy the conclusion of Theorem 4.8. (It is interesting to remark that in both these examples there is an "asymptotic" final-state determinability; infinite-time conditions are more appropriate for nonpolynomial systems.)

Continuous-time. Many of the previous results can be generalized to continuous-time finite-dimensional systems

$$\dot{x}(t) = P(x(t), u(t)), \qquad y(t) = h(x(t)), \tag{4.9}$$

where appropriate restrictions are placed on the state-space, input set, spaces of input functions, and P, h. The continuous case is simpler than the discrete one, due to the time-reversibility of (finite dimensional) differential equations. This implies that no information is lost when an experiment is performed on such a system, i.e., the maps

$$x \mapsto P(x, w) \tag{$*$}$$

are homeomorphisms for all w ($P(x, w)$ = solution of (4.9) at time T with $x(0) = x$ and input $w(\cdot)$ on $[0, T]$). It follows that final-state-determination is equivalent to single-experiment observability. If P is analytic in both x, u and h is analytic in x (so that the maps (*) are analytic), and under suitable technical assumptions insuring existence and uniqueness of solutions of (4.9) for admissible input functions w, it follows by essentially the same argument as in Theorem 4.8 that *observability implies single-experiment observability*. (The internally-bilinear case of this result was proved via linear-algebraic techniques by Grasselli and Isidori (1977).) When P, h are *polynomial* in x, the methods in Sontag and Rouchaleau (1975) can be applied to jets of outputs corresponding to smooth inputs, resulting in finiteness results for the continuous case.

REFERENCES

B. D. O. ANDERSON AND J. B. MOORE (1971), *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ.

R. BELLMAN AND K. J. ASTRÖM (1970), *On structural identifiability*, Math. Biosci., 7, pp. 329–339.

R. W. BROCKETT (1972), *On the algebraic structure of bilinear systems*, Theory and Applications of Variable Structure Systems, R. Mohler and A. Ruberti, eds., Academic Press, New York.

J. H. CONWAY (1971), *Regular Algebra and Finite Machines*, Chapman and Hall, London.

P. D'ALESSANDRO, A. ISIDORI AND A. RUBERTI (1974), *Realization and structure theory of bilinear systems*, this Journal, 12, pp. 517–535.

S. EILENBERG (1974), *Automata, Languages, and Machines*, vol. A, Academic Press, New York.

J. M. FITTS (1972), *On the observability of non-linear systems with applications to non-linear regression analysis*, Information Sci., 4, pp. 129–156.

M. FLIESS (1973), *Sur la réalisation des systèmes dynamiques bilinéaires*, C. R. Acad. Sci. Paris, Sér. A, 277, pp. 243–247.

M. GATTO AND G. GUARDABASSI (1976), *The regulator theory of finite automata*, Information and Control, 31, pp. 1–16.

A. GILL (1962), *Introduction to the Theory of Finite-State Machines*, McGraw-Hill, new York.

O. M. GRASSELLI AND A. ISIDORI (1977), *Deterministic state reconstruction and reachability of bilinear control processes*, Proc. J.A.C.C. (San Francisco, June 22–25).

M. HERVÉ (1963), *Several Complex Variables, Local Theory*, Oxford University Press, London.

R. E. KALMAN (1968), *Lectures in Controllability and Observability*, Cremonese, Rome.

R. E. KALMAN, P. FALB AND M. A. ARBIB (1969), *Topics in Mathematical System Theory*, McGraw-Hill, New York.

H. KWAKERNAAK AND R. SIVAN (1972), *Linear Optimal Control Systems*, John Wiley, New York.

E. F. MOORE (1956), *Gedanken experiments on sequential machines*, Automata Studies, Princeton University Press, Princeton, NJ.

A. A. MUCHNIK (1973), *General Linear Automata*, Systems Theory Research, A. A. Lyapunov, ed., 23, pp. 179–218, Consultants Bureau, New York.

E. D. SONTAG (1976a), *On the internal realization of polynomial response maps*, Doctoral Dissertation, University of Florida.

——— (1976b), *Realization theory of discrete-time nonlinear systems. I. The bounded case*, Proc. IEEE Dec. and Control Conf. Clearwater, FL, Dec. 1976. Full paper submitted to IEEE Trans. Circuits and Systems.

E. D. SONTAG AND Y. ROUCHALEAU (1975), *On discrete-time polynomial systems*, CNR–CISM Symposium on Algebraic System Theory (Udine, Italy, June 1975). It appeared in revised form in J. Nonlinear Analysis, Methods, Theory Appl., 1 (1976), pp. 55–64.

R. B. VINTER (1977), *Filter stability for stochastic evolution equations*, this Journal, 15, pp. 465–485.

W. M. WONHAM (1974), *Linear multivariable control*, Economics and Math. Systems, Springer, New York.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 1, 1979

0036-0129/79/1701-0012 $01.00/0

DYNAMIC PROGRAMMING APPROACH TO STOCHASTIC EVOLUTION EQUATIONS*

AKIRA ICHIKAWA†

Abstract. In this paper stochastic regulator problems and optimal stationary control as well as stability are studied for infinite dimensional systems with state and control dependent noise. The stochastic model is described by a semigroup and Wiener processes in Hilbert space and Wonham's approach using differential generators and dynamic programming is extended to infinite dimensions.

Introduction. The theory of differential equations, both deterministic and stochastic, optimal control, and filtering has been extended by many authors to infinite dimensions [1]–[7], [10], [11], [14], [15], [16]. This is partly because a wide class of partial differential equations and delay differential equations can be described by differential equations in infinite dimensions using semigroups or evolution operators [5].

In [13] Wonham has developed an extensive study of stochastic processes in control theory including stability, regulator problems, optimal stationary control, invariant measures of a Markov process, filtering and separation principle. He exploits the theory of differential generator and dynamic programming to solve these problems.

In infinite dimensions Wonham's approach does not seem promising, since in general stochastic evolution equations have only so-called mild solutions and hence stochastic differentials do not exist for such systems [1], [3]. In other words we cannot apply Ito's formula to them, which is essential to dynamic programming approach. In [9] we have solved stochastic regulator problems both on finite and infinite horizons. But because of the reason mentioned above we needed an indirect approach and had to restrict ourselves to linear feedback controls. And yet we can formally consider differential generators and Bellman's equations associated with infinite dimensional systems. The aim of this article is to see whether they have any consequences to stability or optimality. In fact we shall show in § 3 that stochastic stability in mean square sense is equivalent to the existence of a solution to a Lyapunov equation as well as to exponential stability of second moments [7], [17]. This is an extension of Datko's results in [6] to stochastic systems. Then we shall show that invariant measures exist for stable processes (in the sense defined in § 3). In § 4 we solve regulator problems and optimal stationary control problems considered by Wonham [13], [14] using dynamic programming. We shall show that the solution of a Bellman equation yields an optimal feedback control and the minimum cost. We introduce approximating systems which have stochastic differentials and apply dynamic programming arguments to them. This is an extension of our previous work [11] to stochastic control.

1. Preliminaries. We summarize results on stability and perturbation of semigroups and stochastic integrals in Hilbert space, which we need in subsequent sections. Let X, Y, U, and H be real Hilbert spaces. We write $\langle \cdot , \cdot \rangle$ for inner products and $|\cdot|$ for norms of elements and operators. We denote by $\mathscr{L}(\cdot)$ and $\mathscr{L}(\cdot,\cdot)$ spaces of bounded linear operators, for example, $\mathscr{L}(X)$, $\mathscr{L}(X, Y)$.

* Received by the editors March 29, 1977, and in revised form January 31, 1978.

† Control Theory Centre, University of Warwick, Coventry, England. Now at Faculty of Engineering, Shizuoka University, Hamamatsu 432, Japan.

1.1. Stability and perturbation of semigroups. Let $T(t)$, $t \geqq 0$, be a strongly continuous semigroup on X with infinitesimal generator A [8]. The domain of A is written $\mathscr{D}(A)$.

DEFINITION 1.1. $T(t)$ (or A) is *stable* if $|T(t)| \leqq a\, e^{-\alpha t}$ for some positive numbers $a \geqq 1$ and α.

THEOREM 1.1 (Datko [6]). *The following statements are equivalent*:

(i) *$T(t)$ is stable,*

(ii) *$\int_0^\infty |T(t)x|^2\, dt < \infty$ for each $x \in X$,*

(iii) *there exists a self-adjoint nonnegative operator P* ($P \geqq 0$ for short) *in $\mathscr{L}(X)$ such that*

$$2\langle Ax, Px\rangle = -\langle x, x\rangle \quad \textit{for each } x \in \mathscr{D}(A). \tag{1.1}$$

Take $B \in \mathscr{L}(U, X)$, $C \in \mathscr{L}(X, Y)$ and recall the definitions:

DEFINITION 1.2.

(i) (A, B) is *stabilizable* if there exists $K \in \mathscr{L}(X, U)$ such that $A - BK$ is stable.

(ii) (C, A) is *detectable* if there exists $J \in \mathscr{L}(Y, X)$ such that A-JC is stable.

Take the control system

$$\dot{x}(t) = Ax(t) + u(t) \tag{1.2}$$

and the observation

$$y(t) = Cx(t); \tag{1.3}$$

then detectability implies stabilizability of (1.2) by a linear feedback law on the observation (1.3). The following results due to Zabczyk [16] are useful in quadratic control and filtering.

LEMMA 1.1. *Take $K \in \mathscr{L}(X, U)$, $0 \leqq Q \in \mathscr{L}(X)$, and $0 < N \in \mathscr{L}(U)$ with bounded inverse $N^{-1} \in \mathscr{L}(U)$. Suppose that (C, A) is detectable and*

$$2\langle (A - BK)x, Qx\rangle + \langle Cx, Cx\rangle + \langle NKx, Kx\rangle \leqq 0 \quad \textit{for each } x \in \mathscr{D}(A). \tag{1.4}$$

Then $A - BK$ is stable.

THEOREM 1.2. *If (A, B) is stabilizable, then the algebraic Riccati equation*

$$2\langle Ax, Qx\rangle + \langle Cx, Cx\rangle - \langle QBN^{-1}B^*Qx, x\rangle = 0, \qquad x \in \mathscr{D}(A) \tag{1.5}$$

has a solution $Q \geqq 0$. If (C, A) is detectable, then (1.5) *has at most one solution and if the solution exists, $A - BN^{-1}B^*Q$ is stable.*

The following results on perturbed semigroups are useful in the sequel. Let $\mathscr{B}_\infty(0, t_1; \mathscr{L}(X))$, $0 < t_1 < \infty$, be the space of strongly measurable essentially bounded $\mathscr{L}(X)$-valued functions. Consider the integral equation

$$x(t) = T(t-s)\xi + \int_s^t T(t-r)K(r)x(r)\, dr, \qquad s \leqq t \leqq t_1, \tag{1.6}$$

where $\xi \in X$, $0 \leqq s < t_1$, and $K \in \mathscr{B}_\infty(0, t_1, \mathscr{L}(X))$. By a standard method one can show that there exists a unique solution, denoted by $x(t; s, \xi)$, which is strongly continuous in t and depends continuously on the initial condition (s, ξ). Define $U(t, s)\xi = x(t; s, \xi)$, then we have:

PROPOSITION 1.1. *$U(t, s) \in \mathscr{L}(X)$ and*

$$\begin{aligned} &\text{(i)} && U(s, s) = I, \textit{ the identity operator}, \quad \textit{for each } 0 \leqq s \leqq t_1, \\ &\text{(ii)} && U(t, r)U(r, s) = U(t, s) \quad \textit{for } 0 \leqq s \leqq r \leqq t \leqq t_1, \\ &\text{(iii)} && U(t, s) \textit{ is strongly continuous in } t \in [s, t_1] \textit{ and } s \in [0, t]. \end{aligned} \tag{1.7}$$

COROLLARY 1.1. *$U(t, s)$ is the unique solution of the operator integral equation*

$$(1.8)\qquad U(t,s)\xi = T(t-s)\xi + \int_s^t T(t-r)K(r)U(r,s)\xi\,dr$$

satisfying (1.7).

DEFINITION 1.3. The operator $U(t, s)$ is called the *perturbation of $T(t)$ by $K(t)$*.

COROLLARY 1.2. *$U(t, s)$ is the unique solution of the integral equation*

$$(1.9)\qquad U(t,s)\xi = T(t-s)\xi + \int_s^t U(t,r)K(r)T(r-s)\xi\,dr$$

with property (1.7).

Proof. Let $U(t, s)$ be the solution of (1.8) and set

$$(1.10)\qquad \tilde{U}(t,s)\xi = T(t-s)\xi + \int_s^t U(t,r)K(r)T(r-s)\xi\,dr.$$

Substitution of (1.8) in (1.10) and Fubini's theorem yield $\tilde{U}(t,s)=U(t,s)$. The uniqueness follows from linearity of (1.9) or by reversing the above process.

The next proposition follows from Corollaries 1.1, 1.2.

PROPOSITION 1.2. *$U(t, s)$ satisfies the following*:

$$(1.11)\qquad U(t,s)x - x = \int_s^t U(t,r)[A+K(r)]x\,dr, \qquad x\in\mathscr{D}(A)$$

or equivalently,

$$(1.12)\qquad \frac{\partial}{\partial s}U(t,s)x = -U(t,s)[A+K(s)]x, \qquad x\in\mathscr{D}(A) \quad \textit{for almost all } s\in[0,t].$$

$$(1.13)\qquad \frac{\partial}{\partial t}\langle U(t,s)\xi, y\rangle = \langle U(t,s)\xi, A^*y+K^*(t)y\rangle, \qquad \xi\in X,\quad y\in\mathscr{D}(A^*)$$

for almost all $t\in[s,t_1]$, where A^* is the adjoint of A.

The equation (1.6) is the integrated version of the differential equation

$$(1.14)\qquad \begin{aligned} \dot{x}(t) &= Ax(t)+K(t)x(t),\\ x(s) &= \xi, \end{aligned}$$

but the solution of (1.6) does not necessarily satisfy (1.14).

DEFINITION 1.4. $U(t, s)\xi$ is called the *mild solution* of (1.14).

We may regard $U(t, s)\xi$ as the weak solution of (1.14) in the sense

$$(1.15)\qquad \frac{d}{dt}\langle x(t), y\rangle = \langle x(t), A^*y+K^*(t)y\rangle, \qquad y\in\mathscr{D}(A^*) \quad \text{for almost all } t\in[s,t_1].$$

We refer the reader to [5] for details and perturbation theory for evolution operators.

1.2. Stochastic calculus in Hilbert space. We introduce Wiener processes and stochastic integrals in Hilbert space [2], [3]. Let (Ω, Σ, σ) be a complete probability space.

DEFINITION 1.5. An H-valued stochastic process $w(t)$ on (Ω, Σ, σ) is a *Wiener process* if

$$w(t) = \sum_{i=1}^{\infty} \beta_i(t)e_i$$

where $\beta_i(t)$ are mutually independent real Wiener processes with

$$E\{\beta_i^2(t)\}=\lambda_i t \quad \text{and} \quad \sum_{i=1}^{\infty} \lambda_i<\infty$$

and $\{e_i\}$ is an orthonormal set of vectors in H. The nonnegative trace class operator $W\in\mathscr{L}(H)$ with $We_i=\lambda_i e_i$, trace $W=\sum_{i=1}^{\infty}\lambda_i$ is called the *(incremental) covariance operator of* $w(t)$.

It is known that

$$\begin{aligned}
&E\{w(t)\}=0,\\
(1.16)\qquad &E\{(w(t)-w(s))\circ(w(t)-w(s))\}=W(t-s),\\
&E\{|w(t)-w(s)|^2\}=\text{trace } W(t-s),\\
&E\{|w(t)-w(s)|^4\}\leqq 3(\text{trace } W)^2(t-s)^2,
\end{aligned}$$

where $\circ$ denotes the tensor product, i.e., $(g\circ h)k\triangleq g\langle h,k\rangle$ for any $g, h, k\in H$.

Let $F_t=\sigma_t\{w(\cdot)\}$, the minimum σ-algebra generated by $w(s)$, $0\leqq s\leqq t$. Take Φ: $[0,t_1]\times\Omega\to\mathscr{L}(H,X)$ a strongly measurable function adapted to F_t with

$$\int_0^{t_1} E|\Phi(t)|^2\,dt<\infty.$$

DEFINITION 1.6. The *stochastic integral with respect to* $w(t)$ is

$$\int_0^{t_1}\Phi(t)\,dw(t)\triangleq\sum_{i=1}^{\infty}\int_0^{t_1}\Phi(t)\,e_i\,d\beta_i(t)$$

where the convergence is in mean square sense.

Results given below are found in [2].

PROPOSITION 1.3. $\int_0^t\Phi(r)\,dw(r)$, $0\leqq t\leqq t_1$, *is a martingale relative to* F_t *and has continuous sample paths. Moreover,*

$$\begin{aligned}
&\text{(i) } E\left\{\int_0^{t_1}\Phi(t)\,dw(t)\right\}=0,\\
&\text{(ii) } E\left\{\left|\int_0^{t_1}\Phi(t)\,dw(t)\right|^2\right\}=\text{trace } E\left\{\int_0^{t_1}\Phi(t)W\Phi^*(t)\,dt\right\}\\
(1.17)\qquad &\qquad\leqq\text{trace } W\int_0^{t_1}E|\Phi(t)|^2\,dt,\\
&\text{(iii) } \textit{if} \int_0^{t_1}E|\Phi(t)|^4\,dt<\infty, \textit{ then}\\
&\qquad E\left\{\left|\int_0^{t_1}\Phi(t)\,dw(t)\right|^4\right\}\leqq ct_1(\text{trace } W)^2\int_0^{t_1}E|\Phi(t)|^4\,dt,\qquad c>0.
\end{aligned}$$

THEOREM 1.3 (Ito's lemma). *Suppose that* $g(t,x):[0,t_1]\times X\to R^1$ *is a continuous map and* $x(t)$ *is an* X*-valued stochastic process with stochastic differential*

$$dx(t)=\phi(t)\,dt+\Phi(t)\,dw(t) \tag{1.18}$$

such that

$$
(1.19)\quad
\begin{aligned}
&\text{(i) } g_t(t,x) \textit{ is continuous on } [0,t_1]\times X,\\
&\text{(ii) } g(t,\cdot\,) \textit{ is twice Fréchet differentiable on } X \textit{ for each } t\in[0,t_1],\\
&\text{(iii) } g_x(t,x),\ g_{xx}(t,x) \textit{ are continuous in } (t,x)\in[0,t_1]\times X,\\
&\text{(iv) } \phi(t) \textit{ is an } X\textit{-valued process adapted to } F_t \textit{ and is integrable on } [0,t_1],\\
&\text{(v) } \Phi(t) \textit{ is an } \mathscr{L}(H,X)\textit{-valued strongly measurable function adapted to } F_t
\end{aligned}
$$

with

$$\int_0^{t_1} E|\Phi(t)|^4\,dt<\infty.$$

Then $z(t)=g(t,x(t))$ *has the stochastic differential*

$$
(1.20)\quad
\begin{aligned}
dz(t)=\{&g_t(t,x(t))+\langle g_x(t,x(t)),\phi(t)\rangle\\
&+\tfrac{1}{2}\operatorname{trace}\Phi(t)W\Phi^*(t)g_{xx}(t,x(t))\}\,dt\\
&+\langle g_x(t,x(t)),\Phi(t)\,dw(t)\rangle.
\end{aligned}
$$

2. Stochastic evolution equations with state dependent noise. Let $X, H_i, i=1,2$, be real Hilbert spaces. Consider the stochastic differential equation in X

$$
(2.1)\quad
\begin{aligned}
&dx(t)=Ax(t)\,dt+D(x(t))\,dw_1(t)+F\,dw_2(t),\\
&x(0)=x_0\in X,
\end{aligned}
$$

where A is the infinitesimal generator of a strongly continuous semigroup $T(t)$ on X, $w_i(t)$ is a Wiener process in H_i with covariance operator $W_i, i=1,2, D\in\mathscr{L}(X,\mathscr{L}(H_1,X)), F\in\mathscr{L}(H_2,X)$ and $w_i(t), i=1,2$, are mutually independent.

2.1. Mild solutions. To establish a solution of (2.1) we need very restrictive assumptions. See Remark 2.2 below. So we shall mainly be concerned with the integral equation associated with (2.1):

$$(2.2)\quad x(t)=T(t)x_0+\int_0^t T(t-r)D(x(r))\,dw_1(r)+\int_0^t T(t-r)F\,dw_2(r).$$

PROPOSITION 2.1. *There exists a unique solution of* (2.2) *which is continuous in mean square and adapted to* $\sigma_t\{w_i(\cdot\,), i=1,2\}$. *The fourth moment is also finite and continuous.*

One can establish a proof similar to that of Theorem 2.1 [2] with slight modification.

DEFINITION 2.1. The unique solution of (2.2) is called the *mild solution* of (2.1).

Remark 2.1. The operator D can be nonlinear in x. If D satisfies, for example,

$$
(2.3)\quad
\begin{aligned}
&|D(x)h|\leqq c(1+|x|)|h|\\
&|D(x)h-D(y)h|\leqq c|x-y||h| \qquad (0<c\text{: generic constant})
\end{aligned}
$$

for any $h\in H_1, x, y\in X$, then Theorem 2.1 remains true. For some class of $T(t)$'s we may take D unbounded in the sense $D\in\mathscr{L}(V,\mathscr{L}(H_1,X)), V\subset X$ [10].

Remark 2.2. If $x_0\in\mathscr{D}(A)$ and D, F satisfy

$$
(2.4)\quad
\begin{aligned}
&|T(t)AD(x)h_1|^2\leqq f_1(t)|x|^2|h_1|^2, \qquad x\in X, \quad h_1\in H_1,\\
&|T(t)AFh_2|^2\leqq f_2(t)|h_2|^2, \qquad h_2\in H_2,
\end{aligned}
$$

for some locally integrable function f_1, f_2, then the mild solution in fact satisfies (2.1). This follows from a stochastic Fubini theorem [3].

Remark 2.3. The initial value x_0 can be random.

More general stochastic evolution equations with martingale noise can be found in [1].

2.2. Stochastic perturbation of semigroups. Consider the stochastic version of the integral equation (1.8):

$$S(t,s)\xi = T(t-s)\xi + \int_s^t T(t-r)D(S(r,s)\xi)\,dw_1(r), \qquad \xi \in X. \tag{2.5}$$

This is an operator integral equation corresponding to (2.2) with $F=0$.

PROPOSITION 2.2. *There exists a unique solution of* (2.5) *in* $\mathscr{L}(X)$ *such that*

$$\begin{aligned} &\text{(i) } S(t,s)\xi \text{ is adapted to } F_{t,s} = \sigma\{w_1(r), s \leqq r \leqq t\} \text{ for each } \xi \in X\\ &\text{(ii) } S(s,s)=I, \quad s \geqq 0,\\ &\text{(iii) } S(t,r)S(r,s)=S(t,s), \quad .0 \leqq s \leqq r \leqq t,\\ &\text{(iv) } E\{S(t,s)\xi\} = T(t-s)\xi, \quad \xi \in X, \quad 0 \leqq s \leqq t,\\ &\text{(v) } E\{S(t,s)\xi | F_{r,s}\} = T(t-r)S(r,s)\xi, \quad 0 \leqq s \leqq r \leqq t,\\ &\text{(vi) } S(t,s)\xi \text{ is mean square continuous in } t \text{ and } s, \quad 0 \leqq s \leqq t,\\ &\text{(vii) } S(t,s)\xi \text{ is the unique solution of} \end{aligned} \tag{2.6}$$

$$x(t) = T(t-s)\xi + \int_s^t T(t-r)D(x(r))\,dw_1(r). \tag{2.7}$$

Proof. This is an immediate consequence of Proposition 2.1.

Example 2.1. Consider the stochastic heat equation

$$\begin{aligned} dx(t,l) &= \frac{\partial^2}{\partial l^2}x(t,l)\,dt + \delta x(t,l)\,d\beta(t), \qquad \delta > 0,\\ x(t,0) &= x(t,1) = 0, \qquad x(0,l) = x_0(l) \end{aligned} \tag{2.8}$$

where $\beta(t)$ is a real Brownian motion. In this case $X = L_2(0,1)$, $T(t)$ is generated by

$$A = \frac{d^2}{dl^2}, \qquad \mathscr{D}(A) = \left\{x(\cdot) \in L_2(0,1): \frac{dx}{dl}, \frac{d^2x}{dl^2} \in L_2(0,1), x(0)=x(1)=0\right\}$$

and $D(x) = \delta x$. Then

$$S(t,s)\xi = e^{-(\delta^2/2)(t-s)+\delta(\beta(t)-\beta(s))}T(t-s)\xi.$$

The mild solution is given by

$$x(t) = e^{-(\delta^2/2)t+\delta\beta(t)}T(t)x_0 \tag{2.9}$$

which has continuous sample paths. If $x_0 \in \mathscr{D}(A)$, then (2.9) is the solution of

$$\begin{aligned} dx(t) &= Ax(t)\,dt + \delta x(t)\,d\beta(t),\\ x(0) &= x_0. \end{aligned}$$

2.3. Ito's lemma and its application. Since the stochastic differential equation (2.1) involves an unbounded operator A, assumption (1.19)(i) turns out to be too strong. So we present a modified Ito's lemma which is appropriate for (2.1).

THEOREM 2.1. *Suppose that the mild solution* $x(t)$ *satisfies the stochastic differential equation* (2.1) *and that* $g(t, x): [0, t_1] \times X \to R^1$ *is a continuous map satisfying*

(i) $g(t, x)$ *is differentiable in* t *for each* $x \in \mathscr{D}(A)$ *and the derivative* $g_t(t, x)$ *is continuous in* t *with estimate*

$$|g_t(t, x)| \leqq c(1+|x|)(1+|x|+|Ax|), \qquad c>0$$

and (1.19)(ii), (iii). *Then* $x(t) = g(t, x(t))$ *has the stochastic differential*

$$\begin{aligned} dz(t) = \{&g_t(t, x(t)) + \langle g_x(t, x(t)), Ax(t)\rangle \\ &+\tfrac{1}{2}\operatorname{trace} D(x(t))W_1D^*(x(t))g_{xx}(t, x(t)) \\ &\qquad +\tfrac{1}{2}\operatorname{trace} FW_2F^*g_{xx}(t, x(t))\}\,dt \\ &+\langle g_x(t, x(t)), D(x(t))\,dw_1(t) + F\,dw_2(t)\rangle. \end{aligned} \tag{2.10}$$

Proof. The new assumption (i) guarantees the integrability of $g_t(t, x(t))$ and the proof [2] of Theorem 1.3 goes through.

Using this theorem we shall calculate

$$\int_0^{t_1} E\langle Mx(t), x(t)\rangle\,dt + E\langle Gx(t_1), x(t_1)\rangle \tag{2.11}$$

where $0 \leqq M$, $0 \leqq G \in \mathscr{L}(X)$. For this purpose consider the linear operator differential equation

$$\begin{aligned} &\frac{d}{dt}\langle P(t)x, x\rangle + 2\langle Ax, P(t)x\rangle + \langle [M+\Delta(P(t))]x, x\rangle = 0, \qquad x \in \mathscr{D}(A), \\ &P(t_1) = G \end{aligned} \tag{2.12}$$

or its integrated version,

$$\begin{aligned} P(t)x = \int_t^{t_1} &T^*(r-t)[M+\Delta(P(r))]T(r-t)x\,dr \\ &+T^*(t_1-t)GT(t_1-t)x, \qquad x \in X, \end{aligned} \tag{2.13}$$

where $\langle \Delta(R)x, y\rangle = \operatorname{trace} D^*(y)RD(x)W_1$, $x, y \in X$, $R \in \mathscr{L}(X)$.

PROPOSITION 2.3. *There exists a unique solution satisfying* (2.12), (2.13) *in the class of linear self-adjoint nonnegative strongly continuous operators on* X.

Proof. The existence and uniqueness of a solution to (2.13) is shown in [9]. To show the equivalence of (2.12) and (2.13), suppose first that $P(t)$ satisfies (2.13). Then differentiating $\langle P(t)x, x\rangle$, $x \in \mathscr{D}(A)$ yields (2.12). Conversely, suppose $P(t)$ satisfies (2.12). Let $x \in \mathscr{D}(A)$, $t > s \geqq 0$. Then $\langle P(t)T(t-s)x, T(t-s)x\rangle$ is differentiable in t and

$$\begin{aligned} \frac{d}{dt}\langle P(t)T(t-s)x, T(t-s)x\rangle &= -2\langle AT(t-s)x, P(t)T(t-s)x\rangle \\ &\quad -\langle [M+\Delta P(t)]T(t-s)x, T(t-s)x\rangle \\ &\quad +2\langle P(t)T(t-s)x, AT(t-s)x\rangle \\ &= -\langle [M+\Delta P(t)]T(t-s)x, T(t-s)x\rangle. \end{aligned}$$

Integrating this from s to t_1 we obtain

$$\langle P(s)x, x\rangle = \langle GT(t_1-s)x, T(t_1-s)x, T(t_1-s)x\rangle$$
$$+\int_s^{t_1} \langle [M+\Delta(P(t))]T(t-s)x, T(t-s)x\rangle\, dt.$$

Since $\mathscr{D}(A)$ is dense in X, (2.13) follows easily.

PROPOSITION 2.4. *Suppose that the mild solution* $x(t)$ *of* (2.1) *has the stochastic differential* (2.1). *Then*

$$\int_s^{t_1} E\langle Mx(t), x(t)\rangle\, dt + E\langle Gx(t_1), x(t_1)\rangle$$
$$= E\langle P(s)x(s), x(s)\rangle + \int_s^{t_1} \text{trace}\, F^*P(t)FW_2\, dt. \tag{2.14}$$

Proof. Take $g(t,x)=\langle P(t)x, x\rangle$; then $g(t,x)$ satisfies the assumptions in Theorem 2.1. So Ito's lemma yields

$$d\langle P(t)x(t), x(t)\rangle = \{-\langle Mx(t), x(t)\rangle + \text{trace}\, F^*P(t)FW_2\}\, dt$$
$$+\langle 2P(t)x(t), D(x(t))\, dw_1(t)+F\, dw_2(t)\rangle.$$

Integrating from s to t_1 and taking expectations we obtain (2.14).

To calculate (2.11) we only need to solve (2.12). Note that both sides of (2.14) are well-defined for the mild solution. Thus the assumptions that the mild solution $x(t)$ satisfies (2.1) seems unnecessary. In fact this is the case and we shall establish (2.14) for the mild solution. Let $0<\lambda\in\rho(A)$, the resolvent set of A and $R(\lambda, A)$ the resolvent of A. Introduce the approximation of (2.1):

$$dx(t) = Ax(t)\, dt + \lambda R(\lambda, A)[D(x(t))dw_1(t)+F\, dw_2(t)],$$
$$x(0)=x_0. \tag{2.15}$$

Take $x_0\in\mathscr{D}(A)$. Since $A\lambda R(\lambda, A)=\lambda-\lambda^2R(\lambda, A)$ is bounded, the assumptions in Remark 2.2 are satisfied. So there exists a unique solution of (2.15) and we denote it by $x(t,\lambda)$. It is also the mild solution. Then applying Ito's lemma to $\langle P(t)x(t,\lambda), x(t,\lambda)\rangle$ we obtain

$$\int_s^{t_1} E\langle Mx(t,\lambda), x(t,\lambda)\rangle\, dt + E\langle x(t_1,\lambda), x(t_1,\lambda)\rangle$$
$$= E\langle P(s)x(s,\lambda), x(s,\lambda)\rangle + \int_s^{t_1} \text{trace}\, (\lambda R(\lambda, A)F)^*P(t)\lambda R(\lambda, A)FW_2\, dt. \tag{2.16}$$

The next lemma allows us to pass to the limit $\lambda\to+\infty$.

LEMMA 2.1. $x(t,\lambda)\to x(t)$ *in mean square uniformly on* $[0, t_1]$ *as* $\lambda\to+\infty$, *where* $x(t,\lambda)$, $x(t)$ *are the mild solutions of* (2.15), (2.1) *respectively.*

Proof. Form the difference

$$x(t,\lambda)-x(t) = \int_0^t T(t-r)[\lambda R(\lambda, A)D(x(r,\lambda))-D(x(r))]\, dw_1(r)$$
$$+\int_0^t T(t-r)[\lambda R(\lambda, A)-I]F\, dw_2(r).$$

So

$$E|x(t,\lambda)-x(t)|^2 \leqq 3E\left\{\int_0^t T(t-r)\lambda R(\lambda,A)D(x(r,\lambda)-x(r))\,dw_1(r)\right\}^2$$
$$+3E\left\{\int_0^t T(t-r)[\lambda R(\lambda,A)-I]D(x(r))\,dw_1(r)\right\}^2$$
$$+3E\left\{\int_0^t T(t-r)[\lambda R(\lambda,A)-I]F\,dw_2(r)\right\}^2$$
$$\leqq c\int_0^t E|x(r,\lambda)-x(r)|^2\,dr+\varepsilon(t,\lambda)$$

where $c>0$ and

$$\varepsilon(t,\lambda)=3E\left\{\int_0^t T(t-r)[\lambda R(\lambda,A)-I]D(x(r))\,dw_1(r)\right\}^2$$
$$+3E\left\{\int_0^t T(t-r)[\lambda R(\lambda,A)-I]F\,dw_2(r)\right\}^2.$$

By Gronwall's inequality we obtain

$$E|x(t,\lambda)-x(t)|^2 \leqq \varepsilon(t,\lambda)+c\int_0^t e^{c(t-r)}\varepsilon(r,\lambda)\,dr.$$

But we know [8] that $\lim_{\lambda\to+\infty}\lambda R(\lambda,A)x=x$ for any $x\in X$. So $\varepsilon(t,\lambda)\to 0$ uniformly on $[0,t_1]$ as $\lambda\to+\infty$, which proves the lemma.

THEOREM 2.2. *Let $x(t)$ be the mild solution of* (2.1); *then* (2.14) *holds*:

$$\int_s^{t_1} E\langle Mx(t),x(t)\rangle\,dt+E\langle Gx(t_1),x(t_1)\rangle$$
$$=E\langle P(s)x(s),x(s)\rangle+\int_s^{t_1}\operatorname{trace} F^*P(t)FW_2\,dt.$$

Proof. If $x_0\in\mathscr{D}(A)$, then we can pass to the limit $\lambda\to\infty$ in (2.16) to obtain (2.14). But since $\mathscr{D}(A)$ is dense in X and $x(t)$ depends continuously on x_0 (2.14) holds for any $x_0\in X$.

3. Stochastic stability and invariant measures.

Our basic system is

$$x(t)=T(t)x_0+\int_0^t T(t-r)D(x(r))\,dw_1(r)+\int_0^t T(t-r)F\,dw_2(r). \tag{3.1}$$

3.1. Stochastic stability.

We first prove a stochastic version of Datko's result Theorem 1.1. Set $F=0$ and consider

$$x(t)=T(t)x_0+\int_0^t T(t-r)D(x(r))\,dw_1(r). \tag{3.2}$$

DEFINITION 3.1. The system (3.2) (or (A,D)) is *stable* if the solution $x(t)$ satisfies

$$\int_0^\infty E|x(t)|^2\,dt<\infty \quad \text{for each } x_0\in X.$$

THEOREM 3.1. *The following statements are equivalent*:

(i) (A, D) *is stable,*

(ii) *there exists* $0 \leqq P \in \mathscr{L}(X)$ *such that*

$$2\langle Ax, Px\rangle + \langle \Delta(P)x, x\rangle = -\langle x, x\rangle \quad \textit{for any } x \in D(A), \tag{3.3}$$

(iii) *there exists positive numbers* $a \geqq 1$, $\alpha > 0$ *such that*

$$E|S(t, s)|^2 \leqq a\, e^{-\alpha(t-s)}$$

or equivalently

$$E|x(t)|^2 \leqq a\, e^{-\alpha t}|x_0|^2$$

where $S(t, s)$ *is the stochastic perturbation of* $T(t)$ *given in* § 2.2.

Proof. The equivalence of (i) and (ii) is shown in [17] (see also [9]). Suppose (i) holds and let $P_{t_1}(t)$ be the solution of

$$\begin{aligned} &\frac{d}{dt}\langle P(t)x, x\rangle + 2\langle Ax, P(t)x\rangle + \langle [I + \Delta(P(t))]x, x\rangle = 0, \qquad x \in \mathscr{D}(A),\\ &P(t_1) = 0. \end{aligned} \tag{3.4}$$

Then specializing Theorem 2.2 and (2.13) we obtain

$$\int_0^{t_1} |T(t)x_0|^2\, dt \leqq \langle P_{t_1}(0)x_0, x_0\rangle = \int_0^{t_1} E|x(t)|^2\, dt < \int_0^{\infty} E|x(t)|^2\, dt < \infty.$$

By Theorem 1.1 $T(t)$ is stable. $P_{t_1}(0)$ is clearly monotone increasing in t_1 and uniformly bounded (Banach–Steinhaus theorem). Thus there exists a limit $P \geqq 0$. Then in view of (3.4) we can conclude that P satisfies (3.3). So (i) implies (ii). Suppose (ii) holds. With the aid of Ito's lemma for $g(x) = \langle Px, x\rangle$ we can establish (see Theorem 2.2)

$$\langle Px_0, x_0\rangle = E\langle Px(t), x(t)\rangle + \int_0^t E|x(r)|^2\, dr, \qquad t > 0. \tag{3.5}$$

Hence

$$\infty > \langle Px_0, x_0\rangle \geqq \int_0^{\infty} E|x(t)|^2\, dt \geqq \int_0^{\infty} |T(t)x_0|^2\, dt,$$

which implies (i) as well as the stability of $T(t)$. Then (iii) follows [7]. In fact from (3.5) we obtain

$$\frac{d}{dt}E\langle Px(t), x(t)\rangle = -E|x(t)|^2 \leqq -|P|^{-1}E\langle Px(t), x(t)\rangle,$$

from which follows

$$E\langle Px(t), x(t)\rangle \leqq e^{-pt}\langle Px_0, x_0\rangle, \qquad p = |P|^{-1}.$$

Since $T(t)$ is stable, $|T(t)| \leqq be^{-\beta t}$ for some $b \geqq 1, \beta > 0$. Now from (3.2) follows the estimate

$$\begin{aligned}
E|x(t)|^2 &= |T(t)x_0|^2 + \text{trace} \int_0^t E\{T(t-r)D(x(r))W_1D^*(x(r))T^*(t-r)\}\,dr \\
&\leqq b^2\,e^{-2\beta t}|x_0|^2 + c\int_0^t e^{-2\beta(t-r)}E|x(r)|^2\,dr \qquad (c>0) \\
&= b^2\,e^{-2\beta t}|x_0|^2 - c\int_0^t \; e^{-2\beta(t-r)}\frac{d}{dr}E\langle Px(t), x(r)\rangle\,dr \\
&= b^2\,e^{-2\beta t}|x_0|^2 - cE\langle Px(t), x(t)\rangle + c\,e^{-2\beta t}\langle Px_0, x_0\rangle \\
&\qquad + \frac{c}{2\beta}\int_0^t e^{-2\beta(t-r)}E\langle Px(r), x(r)\rangle\,dr \\
&\leqq b^2\,e^{-2\beta t}|x_0|^2 + c\,e^{-2\beta t}\langle Px_0, x_0\rangle + \frac{c}{2\beta}\int_0^t e^{-2\beta(t-r)}\,e^{-pr}\langle Px_0, x_0\rangle\,dr \\
&= b^2\,e^{-2\beta t}|x_0|^2 + c\,e^{-2\beta t}\langle Px_0, x_0\rangle + \frac{c}{2\beta}\frac{1}{2\beta-p}(e^{-pt}-e^{-2\beta t})\langle Px_0, x_0\rangle \qquad (p \neq 2\beta) \\
&\leqq b^2\,e^{-2\beta t}|x_0|^2 + c\,e^{-2\beta t}\langle Px_0, x_0\rangle + \frac{c}{2\beta(2\beta-p)}\,e^{-pt}\langle Px_0, x_0\rangle \qquad (p \neq 2\beta).
\end{aligned}$$

If $p = 2\beta$, the last term above is replaced by $(c/(2\beta))t\,e^{-2\beta t}\langle Px_0, x_0\rangle$. This implies (iii) i.e.,

$$E|x(t)|^2 \leqq a\,e^{-\alpha t}|x_0|^2 \quad \text{for some } a \geqq 1, \quad \alpha > 0.$$

But obviously (iii) implies (i).

COROLLARY 3.1. *The equation* (3.3) *has at most one solution.*

Proof. If there exists a solution $P \geqq 0$, then (3.5) yields

$$\langle Px_0, x_0\rangle = \int_0^\infty E|x(t)|^2\,dt.$$

If (A, D) is stable, then the average of second moments of the solution $x(t)$ of (3.1) is finite. In fact we have:

COROLLARY 3.2. *Let $x(t)$ be the solution of* (3.1). *If (A, D) is stable, then*

$$\lim_{t\to\infty}\frac{1}{t}\int_0^t E|x(r)|^2\,dr = \text{trace}\,F^*PFW_2$$

where P is the solution of (3.3).

Proof. Let $P_{t_1}(\cdot)$ be the solution of (3.4); then in view of Theorem 2.2

$$\int_0^{t_1} E|x(t)|^2\,dt = \langle P_{t_1}(0)x_0, x_0\rangle + \int_0^{t_1}\text{trace}\,F^*P_{t_1}(t)FW_2\,dt.$$

But we know by Theorem 3.1 that $P_{t_1}(t)$ converges monotonically to P as $t_1 \to +\infty$. Hence the assertion follows.

3.2. Invariant measures. We examine invariant measures of the Markov process $x(t)$ associated with (3.1):

$$x(t)=T(t)x_0+\int_0^t T(t-r)D(x(r))\,dw_1(r)+\int_0^t T(t-r)F\,dw_2(r).$$

We need the transition function of the process $x(t)$ defined by

$$P_t(\xi,\theta))=\text{Probability }\{x(t)\in\theta|x(0)=\xi\}$$

where θ is an arbitrary Borel set of H and $x(t)$ is the solution of (3.1) with $x(0)=\xi$. Let μ be a probability measure on X.

DEFINITION 3.2. *μ is an invariant measure of $P_t(\cdot,\cdot)$ if $P_t(\mu,\cdot)=\mu(\cdot)$, where $P_t(\mu,\cdot)\triangleq\int_X\mu(d\xi)P_t(\xi,\cdot)$.*

The next theorem is our main result in this subsection.

THEOREM 3.2. *Suppose that (A,D) is stable; then there exists an invariant measure μ of $P_t(\cdot,\cdot)$ and*

$$\int_X |\xi|^2\mu(d\xi)=\text{trace }F^*PFW_2$$

where P is the unique solution of (3.3).

The following series of lemmas will establish the theorem.

LEMMA 3.1. *If f is a bounded weakly continuous function on X, then so is*

$$P_t(\xi,f)\triangleq\int_X f(\eta)P_t(\xi,d\eta)=Ef(x(t,\xi))$$

where $x(t,\xi)$ is the solution of (3.1) *with* $x(0)=\xi$.

Proof. Let $\xi,\eta\in X$, then by Proposition 2.2

$$x(t,\xi)-x(t,\eta)=S(t,0)(\xi-\eta).$$

Hence if $\xi\to\eta$ weakly, $x(t,\xi)\to x(t,\eta)$ weakly with probability one. Since f is weakly continuous, so is $P_t(\xi,f)$.

COROLLARY 3.3. *If ν_n, ν are probability measures on X such that $\nu_n\to\nu$ weakly* [12] *as $n\to\infty$ with respect to the weak topology of X* (we shall always take weak topology of X as the underlying topology), *then $P_t(\nu_n,\cdot)\to P_t(\nu,\cdot)$ weakly as $n\to\infty$.*

Proof. Let f be an arbitrary real bounded weakly continuous function; then

$$\begin{aligned}\int_X P_t(\nu_n,d\eta)f(\eta)&=\int_X f(x(t,\xi))\nu_n(d\xi)\\&=\int_X P_t(\xi,f)\nu_n(d\xi)\\&\to\int_X P_t(\xi,f)\nu(d\xi)=\int_X P_t(\nu,d\eta)f(\eta)\end{aligned}$$

since $P_t(\xi,f)$ is a real bounded weakly continuous function.

LEMMA 3.2. *$(1/t)\int_0^t P_r(\xi,\cdot)\,dr$ is weakly convergent to some probability measure μ_ξ.*

Proof. Since $(1/t)\int_0^t E|x(t,\xi)|^2\,dr$ is finite for any ξ and t, we can show as in [15] that $(1/t)\int_0^t P_r(\xi,\cdot)\,dr$ is uniformly tight [12]. So we can extract a subsequence $t_n\uparrow+\infty$ such that

$$\frac{1}{t_n}\int_0^{t_n}P_r(\xi,\cdot)\,dr\to\mu_\xi\quad\text{weakly}$$

for some μ_ξ. Note that if $g:[0,\infty)\to R^1$ is bounded measurable function, then $(1/t)\int_0^t g(r)\,dr$ is convergent as $t\to\infty$. Hence

$$\lim_{t\to\infty}\frac{1}{t}\int_0^t\left[\int_X P_r(\xi, d\eta)f(\eta)\right]dr=\int_X f(\eta)\mu_\xi(d\eta),$$

which completes the proof.

LEMMA 3.3. *μ_ξ is an invariant measure of $P_t(\cdot\,,\cdot\,)$ and*

$$\int_X|\eta|^2\mu_\xi(d\eta)=\text{trace } F^*PFW_2$$

where P is the unique solution of (3.3).

Proof. Take f a bounded weakly continuous real function. Then

$$\begin{aligned}\int_X P_t(\mu_\xi, d\eta)f(\eta)&=\int_X\mu_\xi(d\zeta)\int_X f(\eta)P_t(\zeta, d\eta)\\&=\lim_{s\to\infty}\frac{1}{s}\int_0^s\int_X P_r(\xi, d\zeta)\left[\int_X f(\eta)P_t(\zeta, d\eta)\right]dr\\&=\lim_{s\to\infty}\frac{1}{s}\int_0^s\int_X\int_X f(\eta)P_r(\xi, d\zeta)P_t(\zeta, d\eta)\,dr\\&=\lim_{s\to\infty}\frac{1}{s}\int_0^s\int_X f(\eta)P_{r+t}(\xi, d\eta)\,dr\\&=\lim_{s\to\infty}\frac{1}{s}\int_t^{s+t}\int_X f(\eta)P_r(\xi, d\eta)\,dr\\&=\lim_{s\to\infty}\frac{1}{s}\int_0^s\int_X f(\eta)P_r(\xi, d\eta)\,dr=\int_X f(\eta)\mu_\xi(d\eta),\end{aligned}$$

where we have used the Chapman–Kolmogorov equation. Then μ_ξ is an invariant measure of $P_t(\cdot\,,\cdot\,)$. Now consider

$$\begin{aligned}\int_X|\eta|^2\mu_\xi(d\eta)&=\frac{1}{t}\int_0^t\int_X P_r(\mu_\xi, d\eta)|\eta|^2\,dr\\&=\int_X\mu_\xi(d\zeta)\left[\frac{1}{t}\int_0^t\int_X P_r(\zeta, d\eta)|\eta|^2\,dr\right]\\&=\int_X\mu_\xi(d\zeta)\frac{1}{t}\int_0^t E|x(r,\zeta)|^2\,dr\\&=\int_X u_\xi(d\zeta)\lim_{t\to\infty}\frac{1}{t}\int_0^t E|x(r,\zeta)|^2\,dr\\&=\int_X\mu_\zeta(d\zeta)\,\text{trace } F^*PFW_2=\text{trace } F^*PFW_2.\end{aligned}$$

As for the uniqueness of an invariant measure we have the following.

PROPOSITION 3.1. *If for any $x_0\in X$ the solution $x(t)$ of* (3.2) *converges to zero with probability one as $t\to\infty$, then $P_t(\cdot\,,\cdot\,)$ has at most one invariant measure.*

Proof. The proof is similar to that of Proposition 4.1 [15].

Example 3.1. Consider again Example 2.1. Recall the solution

$$x(t)=e^{-(\delta 2/2)t+\delta\beta(t)}T(t)x_0=S(t,0)x_0.$$

We have an estimate $|T(t)| \leqq e^{-\pi^2 t}$ and the direct calculation yields

$$E|S(t,0)|^2 \leqq e^{-2(\pi^2-\delta^2/2)t}.$$

Hence (2.8) is stable if $|\delta| < \sqrt{2}\pi$. It is known that

$$\text{Probability}\left\{\lim_{t\to\infty}\frac{\beta(t)}{t}=0\right\}=1.$$

So $x(t)\to 0$ with probability one as $t\to+\infty$ and the assumption in Proposition 3.1 is satisfied.

4. Regulator problems. Consider the control system

$$\begin{aligned} x(t) &= T(t)x_0+\int_0^t T(t-r)D(x(r))\,dw_1(r)+\int_0^t T(t-r)F\,dw_2(r) \\ &\quad+\int_0^t T(t-r)Bu(r)\,dr+\int_0^t T(t-r)C(u(r))\,dw_3(r), \qquad x_0\in X, \end{aligned} \tag{4.1}$$

where $w_3(t)$ is a Weiner process in a real Hilbert space H_3 with covariance operator W_3, $u(t)$ is a control with values in a real Hilbert space U, $B\in\mathscr{L}(U,X)$, $C\in\mathscr{L}(U,\mathscr{L}(H_3,X))$ and we assume that $w_3(t)$ is independent of $w_i(t)$, $i=1,2$.

The stochastic differential equation corresponding to (4.1) is

$$\begin{aligned} &dx(t)=(Ax(t)+Bu(t))\,dt+D(x(t))\,dw_1(t)+F dw_2(t)+C(u(t))\,dw_3(t), \\ &x(0)=x_0. \end{aligned} \tag{4.2}$$

We have solved regulator problems for (4.1) in [9], but admissible controls are of linear feedback type. Here we employ the dynamic programming method which enables us to take feedback controls of Lipschitz type. Using results in § 3 we can also formulate an optimal stationary control problem involving invariant measures.

4.1. Optimal control over a finite horizon. Consider (4.1) on a finite interval $[0,t_1]$. For admissible controls we take $u(t)$'s which are adapted $\sigma_t\{w_i(\cdot), i=1,2,3\}$ and satisfy $\int_0^{t_1} E|u(t)|^2\,dt<\infty$. The cost functional to be minimized is

$$\mathscr{C}(u)=E\langle Gx(t_1),x(t_1)\rangle+\int_0^{t_1} E\{\langle Mx(t),x(t)\rangle+\langle Nu(t),u(t)\rangle\}\,dt \tag{4.3}$$

where $0\leqq G$, $0\leqq M\in\mathscr{L}(X)$ and $0<N\in\mathscr{L}(U)$ with $N^{-1}\in\mathscr{L}(U)$. Note that feedback controls of the form

$$u=K(t,x)$$

where $K(t,x):[0,t_1]\times X\to U$ is measurable and

$$|K(t,x)|\leqq c(1+|x|), \qquad |K(t,x)-K(t,y)|\leqq c|x-y|, \qquad c>0, \quad x,y\in X$$

are admissible. In fact for such a feedback control (4.1) has a unique solution $x(t)$ adapted to $\sigma_t\{w_i(\cdot), i=1,2,3\}$ with continuous second moments and $u(t)=K(t,x(t))$ is admissible.

Now define $\Gamma(\cdot)$ and the differential generator L_u of (4.1):

$$\langle\Gamma(S)u,v\rangle=\operatorname{trace} C^*(v)SC(u)W_3, \qquad S\in\mathscr{L}(U), \quad u,v\in U, \tag{4.4}$$

$$\begin{aligned} L_uV(x)&=\langle V_x(x),Ax+Bu\rangle \\ &\quad+\tfrac{1}{2}\{\langle\Gamma(V_{xx}(x))u,u\rangle+\langle\Delta(V_{xx}(x))x,x\rangle+\operatorname{trace} F^*V_{xx}(x)FW_2\} \end{aligned} \tag{4.5}$$

for each $x\in\mathscr{D}(A)$, $u\in U$ and twice Fréchet differentiable real function $V(x)$ on X.

The next lemma gives sufficient conditions for optimality.

LEMMA 4.1 (Optimality lemma). *Suppose there exist a feedback control $\bar{u} = \bar{K}(t, x)$ and a real function $V(t, x): [0, t_1] \times X \to R^1$ with properties*

$$
(4.6)\quad
\begin{aligned}
&\text{(i) } V(t,x) \textit{ is twice Fréchet differentiable in } x \textit{ for each } t \in [0, t_1] \textit{ and } V(t,x),\ V_x(t,x),\ V_{xx}(t,x) \textit{ are continuous;}\\
&\text{(ii) } V(t,x) \textit{ is differentiable in } t \textit{ for each } x \in \mathscr{D}(A) \textit{ and}\\
&\qquad |V_t(t,x)| \leqq c(1+|x|)(1+|x|+|Ax|), \qquad x \in \mathscr{D}(A), \quad c>0;\\
&\text{(iii) } |V(t,x)| + |x|\,|V_x(t,x)| + |x|^2 |V_{xx}(t,x)| \leqq c(1+|x|^2), \qquad x \in X, \quad c>0;\\
&\text{(iv) } V(t_1, x) = \langle Gx, x\rangle, \qquad x \in X;\\
&\text{(v) } 0 = V_t(t,x) + L_{\bar{u}} V(t,x) + \langle Mx, x\rangle + \langle N\bar{u}, \bar{u}\rangle\\
&\qquad \leqq V_t(t,x) + L_u V(t,x) + \langle Mx, x\rangle + \langle Nu, u\rangle\\
&\qquad \textit{for each } x \in \mathscr{D}(A) \textit{ and } u \in U;\\
&\text{(vi) } |\bar{K}(t,x)| \leqq c(1+|x|), \qquad |\bar{K}(t,x) - \bar{K}(t,y)| \leqq c|x-y|, \qquad c>0, \quad x, y \in X;
\end{aligned}
$$

then $\bar{u} = \bar{K}(t, x)$ is optimal and the minimum cost is $\mathscr{C}(\bar{u}) = V(0, x_0)$.

Proof. Let $\bar{x}(t)$ be the solution of (4.1) with $\bar{u} = \bar{K}(t, x)$.

Introducing an approximation of the form (2.15) to (4.1) and applying Ito's lemma to $V(t, x)$, we can show as Theorem 2.2

$$EV(t_1, \bar{x}(t)) - V(0, x_0) = -\int_0^t E\{\langle M\bar{x}(t), \bar{x}(t)\rangle + \langle N\bar{u}(t), \bar{u}(t)\rangle\}\, dt.$$

Hence

$$V(0, x_0) = E\langle G\bar{x}(t_1), \bar{x}(t_1)\rangle + \int_0^{t_1} E\{\langle M\bar{x}(t), \bar{x}(t)\rangle + \langle N\bar{u}(t), \bar{u}(t)\rangle\}\, dt = \mathscr{C}(\bar{u}).$$

Repeating the same procedure for the solution $x(t)$ of (4.1) with arbitrary admissible control $u(t)$, we obtain

$$V(0, x_0) \leqq E\langle Gx(t_1), x(t_1)\rangle + \int_0^{t_1} E\{\langle Mx(t), x(t)\rangle + \langle Nu(t), u(t)\rangle\}\, dt = \mathscr{C}(u).$$

Here the inequality is due to the one in (4.6)(v).

In order to solve the regulator problem (4.1), (4.3) we seek a function $V(t, x)$ of the form

$$V(t, x) = \langle Q(t)x, x\rangle + q(t), \qquad Q(t) \in \mathscr{L}(X).$$

Then (4.8)(v) yields the following Riccati equation:

$$
(4.7)\quad
\begin{aligned}
&\frac{d}{dt}\langle Q(t)x, x\rangle + 2\langle Ax, Q(t)x\rangle + \langle Mx, x\rangle + \langle \Delta(Q(t))x, x\rangle\\
&\quad - \langle Q(t)B[N + \Gamma(Q(t))]^{-1} B^* Q(t)x, x\rangle = 0, \qquad x \in \mathscr{D}(A),\\
&Q(t_1) = G,
\end{aligned}
$$

$$(4.8)\qquad q(t) = \int_t^{t_1} \operatorname{trace} F^* Q(r) F W_2\, dr$$

and the feedback control:

$$\bar{u} = -[N + \Gamma(Q(t))]^{-1} B^* Q(t)x. \tag{4.9}$$

So all we need is to establish a solution of the Riçcati equation (4.7).

THEOREM 4.1. *The Riccati equation* (4.7) *has a unique solution in the class of self-adjoint nonnegative strongly continuous $\mathscr{L}(X)$-valued functions. The control law* (4.9) *is optimal and the minimum cost is*

$$\mathscr{C}(\bar{u}) = \langle Q(0)x_0, x_0\rangle + \int_0^{t_1} \operatorname{trace} F^* Q(t) F W_2 \, dt. \tag{4.10}$$

Proof. In a manner parallel to the finite dimensional case, take the sequence of linear differential equations:

$$\begin{aligned} &\frac{d}{dt}\langle Q_0(t)x, x\rangle + 2\langle Ax, Q_0(t)x\rangle + \langle [M + \Delta(Q_0(t))]x, x\rangle = 0, \qquad x \in \mathscr{D}(A), \\ &Q_0(t_1) = G; \end{aligned} \tag{4.11}$$

$$\begin{aligned} &\frac{d}{dt}\langle Q_n(t)x, x\rangle + 2\langle [A - BK_{n-1}(t)]x, Q_n(t)x\rangle + \langle [M + \Delta(Q_n(t))]x, x\rangle \\ &\qquad + \langle K_{n-1}^*(t)[N + \Gamma(Q_n(t))]K_{n-1}(t)x, x\rangle = 0, \qquad x \in \mathscr{D}(A), \\ &Q_n(t_1) = G; \end{aligned} \tag{4.12}$$

$$K_n(t) = [N + \Gamma(Q_n(t))]^{-1} B^* Q_n(t). \tag{4.13}$$

They have a unique solution [9]. Moreover, they are equivalent to the integral equations

$$Q_0(t)x = \int_t^{t_1} T(r-t)[M + \Delta(Q_0(r))]T(r-t)x \, dr + T^*(t_1 - t)GT(t_1 - t)x, \qquad x \in X, \tag{4.14}$$

$$\begin{aligned} Q_n(t)x = \int_t^{t_1} U_n^*(r, t)\{&M + \Delta(Q_n(r)) \\ &+ K_{n-1}^*(r)[N + \Gamma(Q_{n-1}(r))]K_{n-1}(r)\} U_n(r, t)x \, dr \\ &+ U_n^*(t_1, t)GU_n(t_1, t)x, \qquad x \in X, \end{aligned} \tag{4.15}$$

where $U_n(t, s)$ is the perturbation of $T(t)$ by $-BK_{n-1}(t)$. As in Theorem 3.2 we can show

$$\mathscr{C}(u_n) = \langle Q_n(0)x_0, x_0\rangle + \int_0^{t_1} \operatorname{trace} F^* Q_n(t) F W_2 \, dt$$

where u_n is the control law

$$u_n = -K_n(t)x, \qquad n = 1, 2, \cdots, \quad u_0 = 0.$$

Next we shall prove $Q_{n-1}(t) \geqq Q_n(t) \geqq 0$, $n = 1, 2, \cdots$ (see [9], [11]). Set $R_n(t) = Q_{n-1}(t) - Q_n(t)$; then it satisfies

$$\begin{aligned} &\frac{d}{dt}\langle R_n(t)x, x\rangle + 2\langle [A - BK_{n-1}(t)]x, R_n(t)x\rangle + \langle \Delta(R_n(t))x, x\rangle \\ &\qquad + \langle K_{n-1}^*(t)[N + \Gamma(Q_{n-1}(t))]K_{n-1}(t)x, x\rangle = 0, \qquad x \in \mathscr{D}(A), \\ &R_n(t_1) = 0; \end{aligned} \tag{4.16}$$

$$R_n(t)x = \int_t^{t_1} U_n^*(r, t)\{\Delta(R_n(r)) + K_{n-1}^*(r)[N + \Gamma(Q_{n-1}(r))]K_{n-1}(r)\}U_n(r-t)x\,dr, \qquad x \in \mathscr{D}(A). \tag{4.17}$$

But (4.17) has a unique solution $R_n(t) \geqq 0$, thus necessarily $Q_{n-1}(t) \geqq Q_n(t)$. Since $Q_n(t)$, $n = 0, 1, 2, \cdots$, is the sequence of monotone decreasing nonnegative operators, there exists a limit $Q(t)$. Passing to the limit $n \to \infty$ in the integrated version of (4.12) and then differentiating it, we can show that $Q(t)$ satisfies (4.7). Letting $n \to \infty$ in (4.15) yields

$$\begin{aligned} Q(t)x = &\int_t^{t_1} U^*(r, t)\{M + \Delta(Q(r)) + K^*(r)[N + \Gamma(Q(r))]K(r)\}U(r, t)x\,dr \\ &+ U^*(t_1, t)GU(t_1, t)x, \\ K(t) = &[N + \Gamma(Q(t))]^{-1}B^*Q(t), \end{aligned} \tag{4.18}$$

where $U(t, s)$ is the perturbation of $T(t)$ by $-BK(t)$ and we have used the strong convergence of $U_n(t, s)$ [5]. The uniqueness of a solution of (4.7) (and hence (4.18)) and the rest of the theorem follows from Lemma 4.2.

4.2. Optimal control over an infinite horizon. Take $F = 0$ in (4.1):

$$\begin{aligned} x(t) = T(t)x_0 &+ \int_0^t T(t-r)D(x(r))\,dw_1(r) \\ &+ \int_0^t T(t-r)Bu(r)\,dr + \int_0^t T(t-r)C(u(r))\,dw_3(r) \end{aligned} \tag{4.19}$$

and the cost functional

$$\mathscr{C}(u) = \int_0^\infty E\{\langle Mx(t), x(t)\rangle + \langle Nu(t), u(t)\rangle\}\,dt. \tag{4.20}$$

For admissible controls we take the class of feedback controls $u = K(t, x)$ such that

(i) $K(t, x): [0, \infty] \times X \to U$ is measurable and

$$|K(t, x)| \leqq c(1 + |x|), \qquad |K(t, x) - K(t, y)| \leqq c|x - y|, \qquad x, y \in X,$$

(ii) $E|x(t)|^2 \to 0$ as $t \to \infty$, where $x(t)$ is the solution of (4.19) with $u = K(t, x)$.

We extend Definition 1.2(i) to the stochastic case.

DEFINITION 4.1. The system (4.19) (or $(A, B; C, D)$) is *stabilizable* if there exists $K_1 \in \mathscr{L}(X, U)$ such that the feedback law $u = -K_1x$ yields a stable solution $x(t)$, i.e.

$$\int_0^\infty E|x(t)|^2\,dt < \infty.$$

In this case we say that $(A - BK_1, C, D)$ is stable.

If $(A, B; C, D)$ is stabilizable, then the control problem (4.19), (4.20) is a meaningful one.

With slight modification in Theorem 3.1 we obtain:

LEMMA 4.2. *$(A, B; C, D)$ is stabilizable iff there exists $K_1 \in \mathscr{L}(X, U)$ and $0 \leqq P_1 \in \mathscr{L}(x)$ such that*

$$2\langle (A - BK_1)x, P_1x\rangle + \langle [K_1^*\Gamma(P_1)K_1 + \Delta(P_1)]x, x\rangle = -\langle x, x\rangle, \qquad x \in \mathscr{D}(A). \tag{4.21}$$

LEMMA 4.3. *If $(A - BK_1, C, D)$ is stable, then there exists $0 \leqq Q_1 \in \mathscr{L}(X)$ such that*

$$\begin{aligned} &2\langle (A - BK_1)x, Q_1, x\rangle + \langle (M + K_1^*NK_1)x, x\rangle \\ &\quad + \langle [\Delta(Q_1) + K_1^*\Gamma(Q_1)K_1]x, x\rangle = 0, \qquad x \in \mathscr{D}(A). \end{aligned} \tag{4.22}$$

Proof. Let $Q^1_{t_1}(t)$ be the unique solution of

$$\begin{aligned}&\frac{d}{dt}\langle Q(t), x, x\rangle+\langle (A-BK_1)x, Q(t)x\rangle+\langle (M+K_1^*NK_1)x, x\rangle\\ (4.23)\qquad &+\langle [\Delta(Q(t))+K_1^*\Gamma(Q(t))K_1]x, x\rangle=0, \qquad x\in\mathscr{D}(A)\\ &Q(t_1)=0.\end{aligned}$$

Then from a variation of Theorem 2.2 we have

$$\langle Q^1_{t_1}(0)x_0, x_0\rangle=\int_0^{t_1}E\{\langle Mx(t), x(t)\rangle+\langle NK_1x(t), K_1x(t)\rangle\}\,dt$$

where $x(t)$ is the solution of (4.1) with $u=-K_1x$. Since $(A-BK_1, C, D)$ is stable, $Q^1_{t_1}(0)$ is uniformly bounded in t_1. But $Q^1_{t_1}(0)$ is monotone increasing and nonnegative. So there exists a limit $Q_1\geqq 0$ and it satisfies (4.22).

Denote by L^0_u the differential generator (4.5) with $F=0$.

LEMMA 4.4 (Optimality lemma). *Suppose that there exists an admissible control $\bar{u}=-\bar{K}(x)$, and a real-valued function $V(x)$ on X such that*

(i) $V(x)$ *is twice Fréchet differentiable and* $V(x)$, $V_x(x)$, $V_{xx}(x)$ *are continuous,*

(ii) $|V(x)|+|x||V_x(x)|+|x|^2|V_{xx}(x)|<c|x|^2, \qquad x\in X, \quad c>0,$

(4.24) (iii)
$$\begin{aligned}0&=L^0_{\bar{u}}V(x)+\langle Mx, x\rangle+\langle N\bar{u}, \bar{u}\rangle\\ &\leqq L^0_uV(x)+\langle Mx, x\rangle+\langle Nu, u\rangle \quad \textit{for any } x\in\mathscr{D}(A) \textit{ and } u\in U,\end{aligned}$$

(iv) $|\bar{K}(x)-\bar{K}(y)|\leqq c|x-y|, \qquad x, y\in X, \quad c>0.$

Then $\bar{u}=-\bar{K}(x)$ is optimal and $\mathscr{C}(\bar{u})=V(x_0)$

Proof. As in Lemma 4.1 one can show

$$EV(\bar{x}(t))-V(x_0)=-\int_0^t E\{\langle M\bar{x}(r), \bar{x}(r)\rangle+\langle N\bar{u}(r), \bar{u}(r)\rangle\}\,dr$$

where $\bar{x}(t)$ is the solution of (4.19) with $u=\bar{u}$. Note that $|V(x)\leqq c|x|^2$ and $E|\bar{x}(t)|^2\to 0$ as $t\to+\infty$. So

$$V(x_0)=\int_0^\infty E\{\langle M\bar{x}(t), \bar{x}(t)\rangle+\langle N\bar{u}(t), \bar{u}(t)\rangle\}\,dt=\mathscr{C}(\bar{u}).$$

Similarly for any admissible control u we obtain

$$V(x_0)\leqq\int_0^\infty E\{\langle Mx(t), x(t)\rangle+\langle Nu(t), u(t)\rangle\}\,dt=\mathscr{C}(u).$$

Now we seek a function $V(x)$ of the form

$$V(x)=\langle Qx, x\rangle, \qquad 0\leqq Q\in\mathscr{L}(X).$$

Then (4.24)(iii) yields an algebraic Riccati equation

$$(4.25)\qquad 2\langle Ax, Qx\rangle+\langle\{M+\Delta(Q)-QB[N+\Gamma(Q)]^{-1}B^*Q\}x, x\rangle=0, \qquad x\in\mathscr{D}(A)$$

and the control law

$$(4.26)\qquad u=-Kx, K=[N+\Gamma(Q)]^{-1}B^*Q.$$

Hence to solve the regulator problem (4.19), (4.20) we need only establish conditions which guarantee the existence of a solution to (4.25) and the admissibility of $\bar{u} = -Kx$.

LEMMA 4.5. *If there exist operators* $K_1 \in \mathscr{L}(X, U)$, $0 \leqq Q_1 \in \mathscr{L}(X)$ *satisfying* (4.22), *then the Riccati equation* (4.25) *has a solution* $Q \geqq 0$.

Proof [9]. Let $Q_{t_1}(t)$, $Q_{t_1}^{\infty}(t)$ be the solution of (4.23) and of (4.7) with $G = 0$ respectively. Then

$$0 \leqq Q_{t_1}^{\infty}(t) \leqq Q_{t_1}(t) \leqq Q_1.$$

$Q_{t_1}^{\infty}(t)$ is monotone increasing in t_1 and thus there exists a limit of $Q_{t_1}^{\infty}(t)$ as $t_1 \to +\infty$ which is independent of t and satisfies (4.25).

The next lemma gives a sufficient condition for $(A - BK, C, D)$ to be stable. It is an extension of Lemma 1.1 to the stochastic case.

LEMMA 4.6. *Let* $K \in \mathscr{L}(X, U)$. *Suppose that there exist* $J \in \mathscr{L}(X)$ *and* $0 \leqq Q \in \mathscr{L}(X)$ *such that*

(i) $A - JM^{1/2}$ *generates a stable semigroup* $\tilde{T}(t)$ *with* $|\tilde{T}(t)| \leqq \tilde{a}\, e^{-\tilde{\alpha}t}$, *and*

$$\frac{\tilde{a}^2 d^2 \text{ trace } W_1}{\tilde{\alpha}} < 1, \qquad d = |D| \text{ the norm of } D(\,\cdot\,), \tag{4.27}$$

(ii) $2\langle (A - BK)x, Qx\rangle + \langle (M + K^*NK)x, x\rangle + \langle [\Delta(Q) + K^*\Gamma(Q)K]x, x\rangle \leqq 0,$

$x \in \mathscr{D}(A)$.

Then $(A - BK, C, D)$ *is stable.*

Proof. Take $x(t)$ the solution of (4.19) with $u = -Kx$ and write (4.19) in the form

$$x(t) = \tilde{T}(t)x_0 + \int_0^t \tilde{T}(t-r)(JM^{1/2} - BK)x(r)\,dr$$
$$+ \int_0^t \tilde{T}(t-r)D(x(r))\,dw_1(r) + \int_0^t \tilde{T}(t-r)C(Kx(r))\,dw_3(r).$$

From this follows

$$\begin{aligned}
E|x(t)|^2 &\leqq 2E\left|\int_0^t \tilde{T}(t-r)D(x(r))\,dw_1(r)\right|^2 + 6E|\tilde{T}(t)x_0|^2 \\
&\quad + 6E\left|\int_0^t \tilde{T}(t-r)(JM^{1/2} - BK)x(r)\,dr\right|^2 \\
&\quad + 6E\left|\int_0^t \tilde{T}(t-r)C(Kx(r))\,dw_3(r)\right|^2 \\
&\leqq 2\tilde{a}^2 d^2 \text{ trace } W_1 \int_0^t e^{-2\tilde{\alpha}(t-r)}E|x(r)|^2\,dr + 6\tilde{a}^2\, e^{-2\tilde{\alpha}t}|x_0|^2 \\
&\quad + 6\tilde{a}^2 c \int_0^t e^{-2\tilde{\alpha}(t-r)}E\{\langle Mx(r), x(r)\rangle + \langle Kx(r), Kx(r)\rangle\}\,dr \\
&\quad + 6\tilde{a}^2 c \text{ trace } W_3 \int_0^t e^{-2\tilde{\alpha}(t-r)}E\langle Kx(r), Kx(r)\rangle\,dr
\end{aligned} \tag{4.28}$$

where $c = \max\{|J|^2; |B|^2, |C|^2\}$.

Now recall the following result for real functions: Let f, g, k be real nonnegative continuous functions such that $f, k \in L_1(0, \infty)$ $\|k\|_{L_1(0,\infty)} \leqq 1$ and

$$g(t) \leqq f(t) + \int_0^t k(t-r) g(r)\, dr.$$

Then

$$g \in L_1(0, \infty) \quad \text{and} \quad \|g\|_{L_1(0,\infty)} \leqq \frac{\|f\|_{L_1(0,\infty)}}{1-\|k\|_{L_1(0,\infty)}}.$$

Applying this to the inequality (4.21) we obtain

$$E|x(t)|^2 \in L_1(0, \infty),$$

which completes the proof.

Now we can state the main result in this subsection.

THEOREM 4.2. *Suppose that there exist $K_1 \in \mathscr{L}(X, U)$, $0 \leqq Q_1 \in \mathscr{L}(X)$, and $J \in \mathscr{L}(X)$ satisfying* (4.22) *and* (4.27)(i) *respectively. Then there exists a unique solution $Q \geqq 0$ for the Riccati equation* (4.25). *Moreover, the optimal control for* (4.19), (4.20) *is the feedback law* (4.26) *and $\mathscr{C}(\bar{u}) = \langle Qx_0, x_0 \rangle$.*

Proof. The existence of a solution to (4.25) and the stability of $(A - B\bar{K}, C, D)$ follow from Lemma 4.5, 4.6 respectively. Since $\langle Qx_0, x_0 \rangle$ is the minimum cost, Q is unique.

4.3. Optimal stationary control. We consider the control system (4.1)

$$\begin{aligned} x(t) = T(t)x_0 &+ \int_0^t T(t-r)D(x(r))\, dw_1(r) + \int_0^t T(t-r)F\, dw_2(r) \\ &+ \int_0^t T(t-r)Bu(r)\, dr + \int_0^t T(t-r)C(u(r))\, dw_3(r) \end{aligned}$$

together with the class of feedback controls $K(x)$: $X \to U$ with

$$(4.29) \qquad |K(x) - K(y)| \leqq c|x-y|, \qquad c > 0, \quad x, y \in X.$$

Following Wonham [13], [14] we say that the feedback control $u = -K(x)$ is *admissible* if it satisfies (4.29) and the Markov process given by (4.1) with $u = -K(x)$ has an invariant measure μ_K such that

$$(4.30) \qquad \int_X |x|^2 \mu_K(dx) < \infty.$$

Our control problem is to minimize

$$\mathscr{C}(u) = \int_X \{\langle Mx, x \rangle + \langle NK(x), K(x) \rangle\} \mu_K(dx)$$

over all admissible controls. Clearly the cost $\mathscr{C}(u)$ is independent of the initial value x_0.

Generalizing Theorem 3.2 slightly we can show that $u = -K_1 x$ is admissible if $(A - BK_1, C, D)$ is stable. So stabilizability of $(A, B; C, D)$ is sufficient for the existence of an admissible control.

LEMMA 4.7 (Optimality lemma). *Suppose that there exist an admissible control $\bar{u} = -\bar{K}(x)$, a number γ and a real-valued function $V(x)$ on X such that*

(i) *$V(x)$ is twice Fréchet differentiable and $V(x)$, $V_x(x)$, $V_{xx}(x)$ are continuous,*

(4.31) (ii) $|V(x)|+|x||V_x(x)|+|x|^2|V_{xx}(x)|\leqq c(1+|x|^2), \qquad x \in X, \quad c>0,$

(iii)
$$\begin{aligned}\gamma &= L_{\bar{u}}V(x)+\langle Mx, x\rangle+\langle N\bar{u}, \bar{u}\rangle\\ &\leqq L_uV(x)+\langle Mx, x\rangle+\langle Nu, u\rangle\end{aligned}$$

for any $x \in \mathcal{D}(A)$, $u \in U$, where L_u is the differential generator given by (4.5).

Then $\bar{u} = -\bar{K}(x)$ is optimal and $\mathscr{C}(\bar{u}) = \gamma$.

Proof. Let $\bar{x}(t, \xi)$ be the solution of (4.1) with $u = -\bar{K}(x)$ and $x_0 = \xi$.

As in Lemma 4.1 we can derive

$$EV(\bar{x}(t, \xi))-V(\xi)=\int_0^t \{\gamma-E[\langle M\bar{x}(r, \xi)\bar{x}(r, \xi)\rangle+\langle N\bar{K}(\bar{x}(r, \xi)), \bar{K}(\bar{x}(r, \xi))\rangle]\}\, dr. \tag{4.32}$$

Now let $\bar{\mu}$ be an invariant measure of $\bar{x}(t)$ and $P_t(\cdot, \cdot)$ the transition function. Then

$$\int_X EV(\bar{x}(t, \xi))\bar{\mu}(d\xi)=\int_X\left[\int_X V(\eta)P_t(\xi, d\eta)\right]\bar{\mu}(d\xi)=\int_X V(\xi)\bar{\mu}(\xi)$$

when we have used the fact $P_t(\bar{\mu}, \cdot)=\bar{\mu}(\cdot)$. Take expectations of (4.32) with respect to $\bar{\mu}$; then

$$0=\int_0^t\left\{\gamma-\int_X[\langle M\xi, \xi\rangle+\langle N\bar{K}(\xi), \bar{K}(\xi)\rangle]\bar{\mu}(d\xi)\right\}dr.$$

Hence

$$\gamma=\int_X[\langle Mx, x\rangle+\langle N\bar{K}(x), \bar{K}(x)\rangle]\bar{\mu}(dx)=\mathscr{C}(\bar{u}).$$

Similarly for any admissible control $u = -K(x)$ we have

$$\gamma \leqq \mathscr{C}(u).$$

Remark 4.1. Wonham [14] proved the optimality lemma through $\int_X L_{\bar{u}}V(x)\bar{\mu}(dx)=0$. But in infinite dimensions $L_{\bar{u}}V(x)$ is defined only for $x \in \mathcal{D}(A)$ and $\int_X L_{\bar{u}}V(x)\bar{\mu}(dx)$ does not make sense in general.

Again we seek a function $V(x)$ of the form

$$V(x)=\langle Qx, x\rangle.$$

Then (4.31)(iii) yields the Riccati equation (4.25)

$$2\langle Ax, Qx\rangle+\langle\{M+\Delta(Q)-QB[N+\Gamma(Q)]^{-1}B^*Q\}x, x\rangle=0, \qquad x \in \mathcal{D}(A),$$

the control law (4.26)

$$\bar{u}=-\bar{K}x, \qquad \bar{K}=[N+\Gamma(Q)]^{-1}B^*Q$$

and

$$\gamma = \text{trace } F^*QFW_2 \tag{4.33}$$

By virtue of Lemmas 4.5, 4.6 we can prove:

THEOREM 4.3. *Suppose that there exist $K_1 \in \mathscr{L}(X, U)$, $0 \leqq Q_1 \in \mathscr{L}(X)$ and $J \in \mathscr{L}(X)$ satisfying* (4.22) *and* (4.27)(i) *respectively. Then the optimal control is the feedback law* (4.26) *and the minimum cost is $\mathscr{C}(\bar{u}) = \text{trace } F^*QFW_2$,* where *$0 \leqq Q$ is the unique solution of the Riccati equation* (4.25).

Remark 4.2. Assuming the existence of K_1, Q_1 satisfying (4.22) is weaker than the stabilizability of $(A, B; C, D)$. But (4.27)(i) is stronger than the detectability of $(M^{1/2}, A)$.

Note that d^2 trace $W_1 = |D|^2$ trace W_1 indicates the size of the state dependent noise. So (4.27)(i) is always satisfied if we assume that the noise is sufficiently small. Stabilizability and detectability are the two conditions used in Wonham [14] and [9].

Remark 4.3. The infinite horizon problem with average cost considered in [9] is a variant of the optimal stationary control problems.

5. Final remarks. In [4] we have shown that the separation principle holds for quadratic problems with Gaussian noise disturbance. After reducing the problems to those with complete observation [4] we can use the results in § 4 and obtain an optimal feedback control law on filters. The filtering part can be solved as a dual problem of deterministic regulator problem [11]. Hence quadratic problems with incomplete observation may be solved using our approach here. An extension to time varying systems of some of our results is also possible as far as we can approximate them by more regular systems so that we are able to use dynamic programming arguments.

In [7] and [9] the case of an unbounded operator D has been considered under additional assumptions either on A or on $T(t)$. This is more interesting and probably more important. But most of the results here can be extended to such a case as well. It would be interesting to find a class of permissible D's.

Acknowledgment. I would like to thank the referees for many valuable comments and suggestions and J. Zabczyk for helpful discussions.

REFERENCES

[1] A. Chojnowska-Michalik, *Stochastic differential equations in Hilbert spaces and some of their applications*, Thesis, Inst. of Math. Polish Acad. of Sciences, 1977.

[2] R. F. Curtain and P. L. Falb, *Ito's lemma in infinite dimensions*, J. Math. Anal. Appl., 31 (1970), pp. 434–448.

[3] R. F. Curtain, *Estimation theory for abstract evolution equations excited by general white noise processes*, this Journal, 14 (1976), pp. 1124–1150.

[4] R. F. Curtain and A. Ichikawa, *The separation principle for stochastic evolution equations*, this Journal, 15 (1977), pp. 367–383.

[5] R. F. Curtain and A. J. Pritchard, *The infinite dimensional Riccati equation for systems defined by evolution operators*, this Journal, 14 (1976), pp. 951–983.

[6] R. Datko, *Extending theorem of A. M. Liapunov to Hilbert space*, J. Math. Anal. Appl., 32 (1970), pp. 610–616.

[7] U. G. Haussmann, *Asymptotic stability of the linear Ito equation in infinite dimensions*, Ibid., to appear 1978.

[8] E. Hille and R. S. Phillips, *Functional Analysis and Semigroups*, Colloquium Publications, no. 33, American Mathematical Society, Providence, RI, 1957.

[9] A. Ichikawa, *Optimal control of a linear stochastic evolution equation with state and control dependent noise*, Proc. IMA Conference, "Recent Theoretical Developments in Control" (Leicester, England, 1976), Academic Press.

[10] ———, *Linear stochastic evolution equations in Hilbert space*, Control Theory Centre Rep. 51, Univ. of Warwick, 1976; J. Differential Equations, to appear.

[11] ———, *Dynamic programming approach to infinite dimensional systems*, Control Theory Centre Rep. 57, Univ. of Warwick, England, 1977.

[12] K. R. Parthasarathy, *Probability Measures on Metric Spaces*, Academic Press, New York, 1967.

[13] W. M. Wonham, *Random differential equations in control theory*, Probabilistic Method in Applied Mathematics, vol. 2, A. Bharucha-Reid, ed., Academic Press, New York, 1970, pp. 131–212.

[14] ———, *Optimal stationary control of a linear system with state-dependent noise*, this Journal, 5 (1967), pp. 486–500.

[15] J. ZABCZYK, *On optimal stochastic control of discrete-time systems in Hilbert space*, this Journal, 13 (1975), pp. 1217–1234.

[16] ———, *Remarks on the algebraic Riccati equation in Hilbert space*, Appl. Math. Optimization, 2 (1976), pp. 251–258.

[17] ———, *On stability of infinite dimensional stochastic systems*, Proc. Probability Semester, Banach Centre (Warsaw, 1976), Scientific Publisher, to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 2, March 1979

0363-0129/79/1702-0001 $01.00/0

PERTURBATION OF CONTROLLABLE SYSTEMS*

R. E. O'BRIEN†

Abstract. Let a semi-dynamical system, (A, B):

$$(A,B) \qquad \begin{aligned} \dot{x} &= Ax+Bu \\ x(0) &= 0 \end{aligned}$$

be given on a Hilbert space X. When A is self-adjoint and semi-bounded with spectral measure $E(\cdot)$ it is shown that controllability of the system (A, B) is equivalent to that of $(f(A), B)$ where $f(\cdot)$ is any semi-bounded, Borel measurable $E(\cdot)$-surjective function on $-\infty < t < +\infty$. In particular (A, B) is controllable if and only if (A_α, B) is controllable, $A_\alpha = -(-A)^\alpha$. These results are then extended to the case where A generates a uniformly bounded C_0-semigroup on X and are applied to a system whose dynamics are governed by the singular integral operator generating the Poisson integral semigroup.

Introduction. Establishing the controllability (in the sense of Fattorini [3]) of a given semi-dynamical system on a Hilbert space X is often a difficult task. Few techniques other than direct verification of the basic definition or Kalman's condition ([1] or [3]) are readily available. These remarks point out several cases in which the controllability of the semi-dynamical system, (A, B):

$$(A,B) \qquad \begin{aligned} \dot{x} &= Ax+Bu \\ x(0) &= 0 \end{aligned}$$

is dependent upon (or equivalent to) that of an auxiliary system $(g(A), B)$ for some appropriate function g.

Here A generates a C_0-semigroup, $T(t)$, $t \geqq 0$, on the Hilbert space X (with inner product $\langle \cdot, \cdot \rangle$). B is a bounded operator from a Hilbert space Y to X, and the control function u is any Y-valued, strongly measurable, locally L^2 function defined on $(-\infty, +\infty)$.

These results are then applied to characterize those operators, B, 2-controlling a system whose dynamics are given by a certain singular operator generating the Poisson integral semigroup.

Notation and basic definitions are those of Yosida [8].

Definitions. The system (A, B) defined above is said to be *controllable* if the span of the trajectories of its mild solutions is dense in X, [1]. If in addition Y is finite dimensional, (A, B) is said to be *finitely controllable*. In either case, controllability is equivalent to the condition:

$$B^*T(t)^*x = 0 \text{ for } t \geqq 0 \text{ implies } x = 0$$

(see for example [1] or [3]).

As Fattorini has noted, for A self-adjoint and semi-bounded above (i.e., for some real number a, $\langle Ax, x\rangle \leqq a\|x\|^2$ for all x in $D(A)$) with Borel spectral measure $E(\cdot)$, A generates a C_0-semigroup and the above condition becomes [3]

$$\text{(F)} \qquad B^*E(\delta)x = 0 \text{ for all Borel sets } \delta \text{ implies } x = 0.$$

* Received by the editors October 27, 1977, and in revised form May 15, 1978.

† Aerospace Systems Section, Computer Science Corporation, Falls Church, Virginia 22041.

Let f be any real-valued Borel function defined on $(-\infty, +\infty)$, then $\int_{-\infty}^{+\infty} f(s)E(ds)$ exists and defines a self-adjoint operator, $f(A)$, with domain

$$D(f(A))=\left\{x: \int_{-\infty}^{+\infty} |f(s)|^2\|E(ds)x\|^2<+\infty\right\}.$$

The spectral measure of $f(A)$, $F(\cdot)$, satisfies $F(\delta)=E(f^{-1}(\delta))$ for all Borel sets δ [2].

We shall call $f(A)$ a *semi-bounded perturbation of* A if there is a real number M such that

$$f(s)\leqq M \qquad E(\cdot)\text{ a.e.}$$

on $-\infty<s<+\infty$. In this case $f(A)$ is clearly semi-bounded above.

Results. Consider the perturbed system, $(f(A), B)$:

$$(f(A), B) \qquad \begin{aligned}\dot{x}&=f(A)x+Bu\\ x(0)&=0.\end{aligned}$$

THEOREM 1. *Let A be self-adjoint and semi-bounded above and suppose $f(A)$ is a semi-bounded perturbation of A. If the system $(f(A), B)$ is controllable then the system (A, B) is controllable.*

Proof. Let $E(\cdot)$, $F(\cdot)$ be the spectral measures defined by A and $f(A)$ respectively and assume $(f(A), B)$ is controllable. If for some x in X

$$B^*E(\delta)x=0$$

for all Borel sets δ then (since f is assumed Borel measurable)

$$B^*F(\delta)x=B^*E(f^{-1}(\delta))x=0.$$

A double application of Fattorini's condition (F) establishes the controllability of the system (A, B). Q.E.D.

For any Borel measure, m, a Borel function f is said to be *m-surjective* if for each Borel set δ there is a Borel set η such that $f^{-1}(\eta)=\delta$ except for a set of m-measure 0.

THEOREM 2. *Let A be self-adjoint and semi-bounded above with spectral measure $E(\cdot)$. If f is $E(\cdot)$-surjective and $f(A)$ is a semi-bounded perturbation of A then the system $(f(A), B)$ is controllable if and only if the system (A, B) is controllable.*

Proof. By Theorem 1, we need only show necessity. Let $F(\cdot)$ be the spectral measure of $f(A)$, assume that (A, B) is controllable, that f is $E(\cdot)$-surjective, and that for some x in X

$$B^*F(\delta)x=0$$

for all Borel sets δ. Then for any Borel set δ, since $\delta=f^{-1}(\eta)$ $E(\cdot)$ a.e. for some Borel set η,

$$B^*E(\delta)x=B^*E(f^{-1}(\eta))x=B^*F(\eta)x=0.$$

Another double application of Fattorini's condition (F) shows that $(f(A), B)$ is controllable. Q.E.D.

COROLLARY. *Let A be self-adjoint and dissipative (i.e. $\langle Ax, x\rangle\leqq 0$ for all x in $D(A)$) with spectral measure $E(\cdot)$. For any α, $0<\alpha<+\infty$, define $A_\alpha=-(-A)^\alpha$. Then the system (A_α, B) is controllable if and only if the system (A, B) is controllable.*

Proof. Note that the function defined on $(-\infty, +\infty)$ by

$$f(t)=\begin{cases} t^\alpha, & t\geqq 0,\\ -(-t)^\alpha, & t<0\end{cases}$$

is invertible, bicontinuous, and $E(\cdot)$-surjective. Moreover, since the spectrum of A is contained in $(-\infty, 0]$,

$$f(t) \leqq 0, \qquad E(\cdot) \text{ a.e.},$$

therefore $A_\alpha = f(A)$ is an $E(\cdot)$-surjective, dissipative perturbation of A. Applying Theorem 2 we are finished.

When the Hilbert space X is separable, the result of Theorem 2 is to be expected. For if f is $E(\cdot)$-surjective then it is easy to see that A and $f(A)$ have the same ordered spectral representations and hence are unitarily equivalent. Ordered spectral representations are discussed in [2].

If A is not necessarily self-adjoint but is known to generate a uniformly bounded C_0-semigroup, $T(t)$, $t \geqq 0$, on X, we may still define a "fractional power" of A, A_α, for $0 < \alpha < 1$. (See for example [6, p. 259]). A_α is given by

$$A_\alpha x = -\frac{\sin(\alpha\pi)}{\pi} \int_0^\infty \lambda^{\alpha-1}(\lambda I - A)^{-1}(Ax)\,d\lambda \quad \text{for } x \in D(A)$$

and A_α generates the semigroup $T_\alpha(t)$, $t \geqq 0$, defined by

$$T_\alpha(t)x = \int_0^\infty f_{t,\alpha}(\lambda)T(\lambda)x\,d\lambda.$$

(The kernel $f_{t,\alpha}(\lambda)$ is defined by

$$f_{t,\alpha}(\lambda) = \frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty} \exp(z\lambda - tz^\alpha)\,dz, \qquad \lambda > 0$$

and $f_{t,\alpha}(0) = 0$, for $0 < \sigma$, $0 < \alpha < 1$, $t \geqq 0$ [6]).

THEOREM 3. *Let A generate a uniformly bounded C_0-semigroup $T(t)$, $t \geqq 0$, on X and take $0 < \alpha < 1$. If the system (A_α, B) is controllable then for each δ, $\alpha < \delta < 1$, the system (A_δ, B) is controllable. Moreover, the system (A, B) will be controllable.*

Proof. Suppose $0 < \alpha < 1$, $\alpha < \delta < 1$ and (A_α, B) is controllable. If for some $x \in X$, $B^* T_\delta^*(t)x = 0$ for $t \geqq 0$ then

$$B^* T_\alpha^*(t)x = \int_0^\infty f_{t,\alpha/\delta}(s) B^* T_\delta^*(s)x\,ds = 0$$

for $t \geqq 0$. But since (A_α, B) is controllable x must be 0. Similarly $B^* T^*(t)x = 0$ for $t \geqq 0$ implies $B^* T_\alpha^*(t)x = 0$ for $t \geqq 0$, and hence in both cases the controllability of (A_α, B) implies that of (A_δ, B) (respectively (A, B)). Q.E.D.

Given a C_0-semigroup $T(t)$, $t \geqq 0$, on X, the function $\omega(t) = \log\|T(t)\|$ is lower semi-continuous on $[0, \infty]$, subadditive, while $\omega(0) = 0$ and $\lim_{t\downarrow 0}\omega(t) < +\infty$. Let $S(\omega)$ be the Banach algebra (under convolution) of all complex Borel measures, m on $0 \leqq t < +\infty$ such that

$$\text{(1)} \qquad \int_0^\infty \|T(t)\|\,|dm(t)| < +\infty.$$

$S(\omega)$ can be identified with the Banach algebra of functions of bounded variation, continuous on the left and satisfying condition (1) [4, § 4.16]. Let $m(t, \cdot)$, $t \leqq 0$, form a semigroup of functions in $S(\omega)$ whose Laplace transforms $\hat{m}(t, \cdot)$, $t \geqq 0$, satisfy (for $\omega_0 = \inf_{t>0}(1/t\omega(t))$

$$\text{(P)} \qquad \frac{1}{t}\log[\hat{m}(t, \lambda)] = q\lambda + \int_0^\infty (\exp\{s(\lambda - \omega_0)\} - 1)\,d\psi(s) + a$$

where $q \geqq 0$, a is real, $\operatorname{Re} \lambda \leqq \omega_0$ and $\psi(s)$ is a monotone nondecreasing function such that

$$\int_0^1 s\, d\psi(s) < +\infty \quad \text{and} \quad \int_0^\infty \exp\,[\omega(s) - \omega_0 s]\, d\psi(s) < +\infty;$$

then Phillips [6] has shown that

$$S(t) = \int_0^\infty T(s)\, dm(t, s), \qquad t \geqq 0,$$

defines a C_0-semigroup on X whose generator, C, is a function of A in the sense that $D(C) \supset D(A)$ and for x in $D(A)$

$$Cx = qAx + \int_0^\infty [e^{-\omega_0 s} T(s)x - x]\, d\psi(s) + ax.$$

We shall say the C_0-semigroups $S(t)$, $t \geqq 0$, and $T(t)$, $t \geqq 0$, are *ω-equivalent* if $\|S(t)\| = \|T(t)\|$, $t \geqq 0$, and there exist semigroups, $m(t, \cdot)$ and $n(t, \cdot)$, $t \geqq 0$, in $S(\omega)$ satisfying condition (P) such that

$$S(t) = \int_0^\infty T(s)\, dm\,(t, s), \qquad T(t) = \int_0^\infty S(s)\, dn\,(t, s), \qquad t \geqq 0.$$

Clearly ω-equivalence forms an equivalence relation on the collection of all C_0-semigroups on X. (Note that each semigroup $T(t)$, $t \geqq 0$, is ω-equivalent to itself, for if $e(t, \cdot)$ is the Borel measure defined by

$$e(t, \delta) = \begin{cases} 1, & t \in \delta \\ 0, & t \notin \delta \end{cases} \quad \text{for any Borel set } \delta,\ t \geqq 0,$$

then $e(0, \cdot)$ is the identity of $S(w)$, $e(t, \cdot)$, $t \geqq 0$, clearly satisfies (P) and

$$\int_0^\infty T(s)\, de\,(t, s) = T(t), \qquad t \geqq 0.)$$

For $S(t)$, $t \geqq 0$, and $T(t)$, $t \geqq 0$, ω-equivalent,

$$B^*S(t)^*x = \int_0^\infty B^*T^*(s)x\, d\bar{m}\,(t, s), \qquad t \geqq 0,$$

and

$$B^*T(t)^*x = \int_0^\infty B^*S^*(s)x\, d\bar{n}\,(t, s), \qquad t \geqq 0.$$

Hence $B^*T(t)^*x = 0$ for $t \geqq 0$ implies $B^*S(t)^*x = 0$ for $t \geqq 0$ and conversely. The conclusion of these remarks is the analogue of Theorem 2:

THEOREM 4. *Let $S(t)$, $t \geqq 0$, and $T(t)$, $t \geqq 0$, be ω-equivalent C_0-semigroups with respective generators C and A. Then the system (C, B) is controllable if and only if the system (A, B) is controllable.* (Note that C and A are necessarily bounded here.)

Theorems 3 and 4 remain true when X is replaced by a general Banach space. The present Hilbert space setting exhibits the fundamental ideas.

An example. Take $X = L^2(-\infty, +\infty)$, $A = d^2/ds^2$, U to be the rapidly decreasing functions [7, p. 146] on $(-\infty, +\infty)$ and define x to be in $D(A)$ if and only if there exists a

u in $L^2(-\infty, +\infty)$ such that for all g in U,

$$\int_{-\infty}^{+\infty} u(s)\bar{g}(s)\,ds = \int_{-\infty}^{+\infty} x(s)\frac{d^2}{ds^2}\bar{g}(s)\,ds.$$

Then $D(A)$ is dense in X, and it is easy to see (via the Fourier transform) that A is self-adjoint, dissipative, and generates the Gauss–Weierstrass semigroup on X:

$$(G) \qquad [T(t)x](s) = \begin{cases} (2\pi t)^{-1/2} \displaystyle\int_{-\infty}^{+\infty} x(s-u)\exp(-u^2/(2t))\,du, & t>0, \\ x(s), & t=0. \end{cases}$$

Note that (G) with x in $D(A)$ is the solution to the heat equation, $u(s,t) = [T(t)x](s)$,

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial s^2} \quad \text{a.e. in } -\infty < s < +\infty,$$

$$\lim_{t\downarrow 0} \|u(\cdot\,, t) - x(\cdot)\|_2 = 0 \qquad [5, \text{p. } 578].$$

Then $A_{1/2} = -(-A)^{1/2}$ exists (since A is self-adjoint and dissipative) and is given by the singular integral operator

$$[A_{1/2}x](s) = \underset{h\downarrow 0}{\text{l.i.m.}}\, \frac{1}{\pi}\int_{-\infty}^{+\infty} \frac{x(s-u)-x(s)}{u^2+h^2}\,du \quad \text{a.e. in } -\infty < s < +\infty \qquad [8, \text{p. } 268].$$

Hence by Theorem 2, $(A_{1/2}, B)$ is finitely controllable if and only (A, B) is finitely controllable, and therefore by [3], $(A_{1/2}, B)$ is controllable if and only if the dimension of Y is at least two, and in this case the operator B

$$B(y_1, y_2) = y_1 U_1(\cdot) + y_2 U_2(\cdot); \qquad U_1, U_2 \text{ in } L^2(-\infty, +\infty)$$

must satisfy ($\hat{U}$ denotes the Fourier transform of U, $U \in L^2$)

$$\hat{U}_1(s)\hat{U}_2(-s) - \hat{U}_1(-s)\hat{U}_2(s) \neq 0 \qquad \text{a.e. in } s \geqq 0.$$

Acknowledgment. This work benefited from several discussions with C. T. Taam and expands some remarks made in the author's Dissertation presented to George Washington University, January 1977.

REFERENCES

[1] A. V. Balakrishnan, *Introduction to Optimization Theory in Hilbert Space*, Springer-Verlag, Berlin, 1973.

[2] N. Dunford and J. T. Schwartz, *Linear Operators*, vol. II, Interscience, New York, 1958.

[3] H. O. Fattorini, *On complete controllability of linear systems*, J. Differential Equations, 3 (1967), pp. 391–402.

[4] ———, *Some remarks on complete controllability*, this Journal, 4 (1967), pp. 686–694.

[5] E. Hille and R. S. Phillips, *Functional Analysis and Semigroups*, vol. XXXI, American Mathematical Society, Providence, RI, 1957.

[6] R. S. Phillips, *On the generation of semigroups of linear operators*, Pacific J. Math., 2 (1952), pp. 343–369.

[7] R. Triggiani, *Extensions of rank conditions for controllability and observability to Banach spaces and unbounded operators*, this Journal, 14 (1976), pp. 313–338.

[8] K. Yosida, *Functional Analysis*, Springer-Verlag, Berlin, 1966.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 2, March 1979

VALUE CONVERGENCE IN A GENERALIZED MARKOV DECISION PROCESS*

GARY J. KOEHLER†

Abstract. We consider generalized Markov decision problems formed from the duals to Leontief substitution systems and relate several properties of the associated polyhedron to value iterative solution procedures used on the induced generalized Markov decision problems. For example, under two assumptions the associated polyhedron is bounded if and only if the spectral radius of some transition matrix is greater than one. The polyhedron has an interior and is unbounded if and only if the spectral radius of each transition matrix is less than or equal to one and no transition matrix P having a spectral radius of one gives $I-P$ spans its corresponding reward vector. These results and others are then used to show that the value iterative procedure is convergent for all v if and only if the dual set has a nonempty interior and is unbounded.

1. Introduction. We consider here a generalized Markov decision process where the transition matrices are nonnegative rather than stochastic. Related models can be found in [6], [9], [12]. Our main concern is to answer: when is value iteration convergent irrespective of the starting vector? Unlike the discounted Markov decision process [5], value iteration in the generalized Markov decision process may not yield a convergent sequence even starting with the zero vector.

A concomitant objective is to relate the geometric properties of a polyhedron associated with the decision process to the spectral properties of the transition matrices and, in turn, to the convergence properties of the value iteration procedure.

In § 2 we introduce the notation to be used throughout this paper. In addition, we define precisely what we mean by our generalized Markov process. In § 3 we derive properties of C (the set of convergent points). In § 4 we relate some geometric properties of D, the associated polyhedron, to the spectral radii of the generalized transition matrices. These results are sharpened and extended in § 5 after adding one additional assumption. This assumption is motivated by computational considerations. Here also we completely characterize D in terms of the spectral radii of the P_δ's and determine when value iteration converges irrespective of the starting vector. In doing so we generalize some results presented in the literature.

2. Notation and preliminary results. Let x and y be two vectors. Write $x \geqq y$ (resp., $x > y$) if $x_i \geqq$ (resp., $>$)y_i for every i. Also, write $x \geq y$ if $x \geqq y$ but $x \neq y$. Also, let $L(x)=\{z | z \leqq x\}$ and if T is a set let $L(T)=\bigcup_{x\in T} L(x)$. We define $G(x)=\{z|z \geqq x\}$ and $G(T)$ in a manner analogous to $L(T)$. We say a set is bounded from above if there is a u such that $T \subseteq L(u)$. Finally, a matrix B is said to be Leontief [13] if it has exactly one positive element in each column and there is some $x \geqq 0$ for which $Bx > 0$.

Consider the problem

$$\begin{array}{lll} & \max & c'x \\ (2.1) & \text{subject to} & Bx=b \\ & & x \geqq 0, \end{array}$$

where B is an $m \times k$ Leontief matrix and $b \geq 0$.

We impose the following on (2.1).

Assumption A. Problem 2.1 has a bounded objective and the columns of B are scaled so that the positive elements of B are not greater than one.

* Received by the editors February 4, 1977, and in final revised form January 30, 1978.

† Department of Management, University of Florida, Gainesville, Florida 32611.

Let $A_i = \{j | B_{ij} > 0\}$ for $i = 1, \cdots, m$ and $\Delta = \prod_{i=1}^{m} A_i$. Observe that $A_i \neq \varnothing$ since each row of B must have a positive element for B to be Leontief. For $\delta \in \Delta$ let B^δ be the corresponding submatrix of B, and let $Q^\delta = I - B^\delta$ and $P_\delta = (Q^\delta)'$. Of course $P_\delta \geqq 0$.

Since the objective in (2.1) is bounded, we have that the problem has an optimal solution at an extreme point of the feasible set. By Veinott [13, Thm. 6] this extreme point corresponds to a basis B^{δ^*} where $\delta^* \in \Delta$, B^{δ^*} is invertible, $(B^{\delta^*})^{-1} \geqq 0$ and $\rho(P_{\delta^*}) < 1$, where $\rho(P)$ denotes the spectral radius of the matrix P.

Let $D = \{y | B'y \geqq c\}$ represent the dual feasible set. From Cottle and Veinott [2, Thm. 1] D has a least element v^* and $v^* = [(B^{\delta^*})']^{-1} c^{\delta^*}$. Since $b \geq 0$ and v^* is the least element of D, v^* solves the dual of (2.1).

We define the following operators on R^m:

$$\mathscr{L}_\delta(v) = P_\delta v + c^\delta, \qquad \delta \in \Delta.$$

Also, let

$$\mathscr{L}(v) = \max_{\delta \in \Delta} \mathscr{L}_\delta(v).$$

Note that the "max" operator above is well defined. Also, as $P_\delta \geqq 0$, one has that for every $\delta \in \Delta$, $\mathscr{L}_\delta$ is isotone, i.e., $\mathscr{L}_\delta(y) \leqq \mathscr{L}_\delta(z)$ whenever $y \leqq z$. Consequently, $\mathscr{L}$ is also isotone.

It is easy to see that since $\rho(P_{\delta^*}) < 1$, for every $v \in R^m$

$$\lim_{n \to \infty} \mathscr{L}_{\delta^*}^n(v) = \sum_{i=0}^{\infty} P_{\delta^*}^i c^{\delta^*} = (I - P_{\delta^*})^{-1} c^{\delta^*} = v^*.$$

We will study conditions under which $\lim_{n \to \infty} \mathscr{L}^n(v) = v^*$. These conditions will depend on v and the spectral radii of the P_δ matrices. Let $C = \{v | \lim_{n \to \infty} \mathscr{L}^n(v) = v^*\}$. Of course, always $v^* \in C$. In order to compute v^* by successive approximation, i.e., by iterating $\mathscr{L}$, we have to start with $v \in C$. Unlike discounted Markov processes, the zero vector may not be in C, even though $C \neq \varnothing$, and $C \neq R^m$ may occur.

3. Some properties of *C*. In this section we derive some properties of C that will be used in subsequent sections.

The following lemma was first established in [8].

LEMMA 3.1. *For every* $v \in R^m$, $\liminf_{n \to \infty} \mathscr{L}^n(v) \geqq v^*$.

Proof. Obviously, $\mathscr{L}(v) \geqq \mathscr{L}_{\delta^*}(v)$ and by induction, for every $n = 1, 2, \cdots, \mathscr{L}^n(v) \geqq (\mathscr{L}_{\delta^*})^n(v)$. Since $\rho(P_{\delta^*}) < 1$, we have that $\lim_{n \to \infty} (\mathscr{L}_{\delta^*})^n(v) = v^*$. So, $\liminf_{n \to \infty} \mathscr{L}^n(v) \geqq v^*$, completing the proof. □

The above lemma implies that $v \in C$ if and only if $\limsup_{n \to \infty} \mathscr{L}^n(v) \leqq v^*$. We use this fact and the isotonicity of $\mathscr{L}$ to obtain two important properties of C.

PROPOSITION 3.2.

(a) *$L(C) \subseteq C$.*

(b) *C is convex.*

Proof. Assume that $y, z \in C$, $v \leqq y$ and $0 < \lambda < 1$. The isotonicity of $\mathscr{L}$ implies that

$$\limsup_{n \to \infty} \mathscr{L}^n(v) \leqq \limsup_{n \to \infty} \mathscr{L}^n(y) \leqq v^*,$$

completing the proof of (a). Next notice that $\mathscr{L}[\lambda y + (1 - \lambda) z] \leqq \lambda \mathscr{L}(y) + (1 - \lambda) \mathscr{L}(z)$. A simple inductive argument shows that for $n = 1, 2, \cdots, \mathscr{L}^n[\lambda y + (1 - \lambda) z] \leqq \lambda \mathscr{L}^n(y) + (1 - \lambda) \mathscr{L}^n(z)$. Taking limit superiors of both sides shows that

$$\limsup_{n \to \infty} \mathscr{L}^n[\lambda y + (1 - \lambda) z] \leqq \lambda \limsup_{n \to \infty} \mathscr{L}^n(y) + (1 - \lambda) \limsup_{n \to \infty} \mathscr{L}^n(z)$$

$$\leqq \lambda v^* + (1 - \lambda) v^* = v^*. \qquad \square$$

Part (a) implies that $L(v^*)\subseteqq C$. A somewhat stronger result was derived by Denardo [4, Thm. 4] under stronger conditions.

A point v is called *excessive* (resp., *fixed point*) of $\mathscr{L}$ if $\mathscr{L}(v)\leq$ (resp., $=$) v. Of course, D is the set of excessive and fixed points of $\mathscr{L}$. The set of fixed points of $\mathscr{L}$ will be denoted F. Naturally $F\subseteq D$ and it is easy to verify that $v^*\in F$. So v^* is also the least element of F.

PROPOSITION 3.3. $D\backslash G(F\backslash\{v^*\})\subseteq C$.

Proof. Let $v^*\neq v\in D\backslash G(F\backslash\{v^*\})$. As $v\in D$, $v\geqq\mathscr{L}(v)$. Iterating this inequality, it follows that $v\geqq\mathscr{L}(v)\geqq\mathscr{L}^2(v)\geqq\cdots$. This shows that $\mathscr{L}^n(v)\in D$ for $n=0,1,\cdots$; so $\mathscr{L}^n(v)\geqq v^*$. Since the sequence $\{\mathscr{L}^n(v)\}_{n=0,1,\cdots}$ is decreasing and bounded from below, it has a limit, say w. We will show that $v\in C$ by establishing that $w=v^*$. A simple continuity argument shows that $\mathscr{L}(w)=w$, i.e., $w\in F$. But $w\leqq v$ and $v\ngeqq f$ for every $f\in F\backslash\{v^*\}$, so $w=v^*$, completing the proof. □

COROLLARY 3.4. *If $F=\{v^*\}$, then $L(D)\subseteq C$. In addition, if for some $d>0$ $P_\delta d\leqq d$ for all $\delta\in\Delta$, then $C=R^m$.*

Proof. The first result follows directly from Proposition 3.3. Next note that if $P_\delta d\leqq d$ for all $\delta\in\Delta$, then $v^*+\lambda d\in D$ for all $\lambda\geqq 0$, so $L(D)=R^m$. □

Remark. Corollary 3.4 establishes some relation between unboundedness of D and the case where $C=R^m$.

The next result establishes a sufficient condition that $F=\{v^*\}$.

PROPOSITION 3.5. *If $\rho(P_\delta)\leqq 1$ for all $\delta\in\Delta$ and $c^\delta\notin$ range $(B^\delta)'$ whenever $\rho(P_\delta)=1$, then $F=\{v^*\}$.*

Proof. Obviously $v^*\in F$. Let $f\in F$, i.e., $\mathscr{L}(f)=f$. Then there exists $\gamma\in\Delta$ with $(I-P_\gamma)f=c^\gamma$. This implies that $\rho(P_\gamma)<1$, so $f=(I-P_\gamma)^{-1}c^\gamma$. As $v^*\in D$, $(I-P_\gamma)v^*\geqq c^\gamma$, so $v^*\geqq(I-P_\gamma)^{-1}c^\gamma=f$. But as v^* is the least element of F, we have that $f=v^*$. □

THEOREM 3.6. *Assume that $C=R^m$. Then for every $\delta\in\Delta$, $\rho(P_\delta)\leqq 1$ and for δ having $\rho(P_\delta)=1$ there exists no $v\in R^m$ such that $v\leqq P_\delta v+c^\delta$.*

Proof. First observe that for $\delta\in\Delta$, $v, x\in R^m$ and $\alpha\in R$

$$\begin{aligned}\mathscr{L}^n(v+\alpha x)\geqq\mathscr{L}^n_\delta(v+\alpha x)&=\sum_{i=0}^{n-1}P^i_\delta c^\delta+P^n_\delta(v+\alpha x)\\&=\mathscr{L}^n_\delta(v)+\alpha P^n_\delta x.\end{aligned}$$

Assume $\rho\equiv\rho(P_\delta)>1$ for some $\delta\in\Delta$. By the Perron–Frobenius theorem (see [10] or [11]) there exists a row vector $y\geq 0$ having $y'P_\delta=\rho y'$. Let $x>0$. Premultiplying (3.7) by y gives

$$\begin{aligned}y'\mathscr{L}^n(v+\alpha x)&\geqq\sum_{i=0}^{n-1}\rho^i y'c^\delta+\rho^n y'(v+\alpha x)\\&=\rho^n(y'c^\delta/(1-\rho)+y'v+\alpha y'x)-y'c^\delta/(1-\rho).\end{aligned}$$

The above and the fact that $y'x>0$ imply that for every (fixed) v and $\alpha>0$ sufficiently large, $y'\mathscr{L}^n(v+\alpha x)$ does not converge to a finite vector. So $v+\alpha x\notin C$, contradicting the hypothesis that $C=R^m$. Thus $\rho(P_\delta)\leqq 1$ for every $\delta\in\Delta$.

Next assume that $\rho(P_\delta)=1$ and that $v\leqq P_\delta v+c^\delta$. Iterating this inequality we get for $n=0,1,\cdots,\mathscr{L}^n_\delta(v)\geqq v$. By the Perron–Frobenius theorem there exists a vector $z\geq 0$ such that $P_\delta z=z$. Let $\alpha>0$ be large enough so that $v^*\ngeqq v+\alpha z$. Apply $\mathscr{L}$ to $v+\alpha z$ to conclude that for every $n=0,1,\cdots,\mathscr{L}^n(v+\alpha z)\geqq v+\alpha z$. If $v+\alpha z\in C$, this would imply that $v^*\geqq v+\alpha z$, a contradiction. Thus $v+\alpha z\notin C$, proving that $C\neq R^m$. □

4. Spectral properties of the P_δ's and geometric properties of D. In this section we obtain results that relate the spectral radii of the P_δ's to boundedness of D and the property that it has a nonempty interior. We first study boundedness of D from above.

THEOREM 4.1. *Consider the following properties of the $\rho(P_\delta)$'s and D*:

(a) *For some $d>0$, $P_\delta d \leqq d$ for all $\delta \in \Delta$.*

(b) $\rho(P_\delta) \leqq 1$ *for all $\delta \in \Delta$.*

(c) *D is unbounded from above.*

(d) *For some $d \geq 0$, $P_\delta d \leqq d$ for $\delta \in \Delta$.*

Then (a)$\rightarrow$(b)$\rightarrow$(c)$\rightleftarrows$(d).

Proof. (a)$\rightarrow$(b): Assume that for some $d>0$, $P_\delta d \leqq d$ for every $\delta \in \Delta$. By Varga [11, p. 47], for every $\delta \in \Delta$, $\rho(P_\delta) \leqq \max_i (P_\delta d)_i/d_i \leqq 1$.

(c)$\leftrightarrows$(d): The proof is trivial.

(b)$\rightarrow$(c): We will show that there exists a vector $d \geq 0$ such that $B'd \geqq 0$, i.e., $P_\delta d \leqq d$ for every $\delta \in \Delta$. This will be accomplished by showing that the operator $\mathscr{H}$ defined on R^n by $\mathscr{H}v = \max_{\delta\in\Delta} P_\delta v$ has an excessive point $d>0$. This result was proved by Seneta [10, Thm. 3.1, p. 60] for the case where each P_δ is irreducible (see also Howard and Matheson [7] and Veinott [12]). To prove the general case let, for each $\varepsilon>0$, $P_\delta^\varepsilon \equiv P_\delta + \varepsilon J$ where J is the matrix of ones and define the operators $\mathscr{H}^\varepsilon$ on R^n by $\mathscr{H}^\varepsilon v = \max_{\delta\in\Delta} P_\delta^\varepsilon v$. Applying the results for irreducible matrices, we find there exist vectors $d^\varepsilon > 0$ so that $\mathscr{H}^\varepsilon d^\varepsilon \leqq d^\varepsilon$. Obviously one can normalize d^ε so that $\|d^\varepsilon\| = 1$. Since $\{d^\varepsilon \mid \varepsilon > 0\}$ is a bounded set it has a finite limit point d. Obviously, $d \geqq 0$ and $\|d\| = 1$, so $d \geq 0$. A simple continuity argument shows that $\mathscr{H}d \leqq d$. □

To see that (c) does not imply (b) consider the following example.

$$\begin{array}{rl} \max & x_1 - 2x_2 + x_3 - 3x_4 + x_5. \\ \text{subject to} & x_1 + x_2 \qquad - 3x_4 \qquad = 2 \\ & \quad -2x_2 + x_3 + x_4 \qquad = 2 \\ & \qquad\qquad\qquad\qquad x_5 = 2 \\ & x_1, x_2, x_3, x_4, x_5 \geqq 0. \end{array} \tag{4.2}$$

Here $v^{*\prime} = (1,1,1)$ and for $d' = (0,0,1)$ we have that $v^* + \lambda d \in D$ for all $\lambda \geqq 0$. However, for $\delta = \{2,4,5\}$, $\rho(P_\delta) = \sqrt{6}$.

THEOREM 4.3. *Consider the following properties of the P_δ's and D*:

(a) *For some $d>0$, $P_\delta d \leqq d$ for every $\delta \in \Delta$.*

(b) *D has a nonempty interior.*

(c) *If $\rho(P_\delta) = 1$ then $c^\delta \notin$* range $(B^\delta)'$.

Then (a) *and* (c) *together imply* (b), *and* (b)$\rightarrow$(c).

Proof. (b)$\rightarrow$(c): Suppose that $(I - P_\delta)v = c^\delta$ for some $v \in R^m$ and $\delta \in \Delta$. Since B is Leontief, no column vanishes and the interior of D is given by $\{\omega \mid B'\omega > c\}$. So, if the interior of D is empty then for some ω $B_\delta'\omega > c^\delta$. It follows that $(I - P_\delta)(\omega - v) > 0$. By the Perron–Frobenius theorem there exists a vector $y \geq 0$ such that $\rho(P_\delta)y' = y'P_\delta$. It follows that $(1 - \rho(P_\delta))y'(\omega - v) = y'(I - P_\delta)(\omega - v) > 0$ proving that $1 \neq \rho(P_\delta)$.

(a) and (c)$\rightarrow$(b): Assume that (a) and (c) hold and that D has an empty interior. The latter condition implies that the system of linear equations $B'v > c$ is infeasible. Let B_i' be the ith row of B'. Let α be the set of indices i for which

$$c_i = \max_{v \in D} B_i' v.$$

Since $B'v > c$ is infeasible, $\alpha \neq \varnothing$. Thus for $i \in \alpha$ the system $B'v \geqq c$ and $B_i'v > c_i$ is (in-)feasible for ($i \in \alpha$) $i \notin \alpha$. Note for later use that $(B^\alpha)'v^* \geqq c^\alpha$ implies $(B^\alpha)'v^* = c^\alpha$.

It follows from duality and Assumption A that there exists a vector z^i such that

$$z^i \geqq 0, \quad z^i_i > 0, \quad Bz^i = 0, \quad \text{and} \quad c'z^i = 0$$

for $i \in \alpha$ and there exists no such z^i for $i \notin \alpha$. Let $z = \sum_{i \in \alpha} z^i$. We have that $z \geq 0$, $z_i > 0$ if and only if $i \in \alpha$, $Bz = 0$ and $c'z = 0$. It follows that $z^\alpha > 0$, $b^\alpha z^\alpha = 0$ and $c^\alpha z^\alpha = 0$.

By our assumption there exists a vector $d > 0$ such that $d'B^\alpha \geqq 0$. Let $\gamma = \{j | \text{row } j$ of B^α does not vanish$\}$. Since $\alpha \neq \emptyset$ and each column of B has a positive element we have that $\gamma \neq \emptyset$. Also as $B^\alpha z^\alpha = 0$ we have that every row of B^α that is indexed in γ has a positive element. Let B^β_γ be a submatrix of B^δ_γ with exactly one positive element in each row and column. Clearly, from $d'B^\alpha = 0$, $\rho(Q^\beta_\gamma) = 1$. Let δ be a δ^* completion of β. Then, collecting results, we have $\rho(P_\delta) = 1$ and $(B^\delta)'v^* = c^\delta$. Thus we reach a contradiction since $\rho(P_\delta) = 1$ and $c^\delta \in \text{range } (B^\delta)'$. Hence D must have an interior. □

COROLLARY 4.4. *If D has a nonempty interior and for some $d > 0$, $P_\delta d \leqq d$ for all $\delta \in \Delta$, then $C = R^m$.*

Proof. Combine Theorem 4.3 with Proposition 3.5 and Corollary 3.4. □

5. The irreducible case. A partitioned matrix of form

$$\begin{bmatrix} B_{11} & & 0 \\ B_{21} & B_{22} & \\ \vdots & & \ddots \\ B_{p1} & \cdots & B_{pp} \end{bmatrix}$$

is called dynamic Leontief if each B_{ii}, $i = 1, \cdots, p$, is Leontief and $B_{ij} \leqq 0$ for $j < i$. If matrix B in problem (2.1) can be permuted to a dynamic Leontief matrix then, using a procedure given by Dantzig [3], one may solve problem (2.1) by solving a sequence of Leontief problems. Hence computational considerations suggest then that we consider only Leontief systems not permutable to dynamic Leontief systems. Throughout this section we impose the following on (2.1).

Assumption B. B of problem (2.1) is not permutable to a dynamic Leontief matrix.

We would like to point out that Assumption B is weaker than the assumption that every P_δ is irreducible. For example

$$B = \begin{bmatrix} 1 & -1 & 0 & 0 \\ -\frac{1}{2} & 1 & 1 & -1 \\ 0 & 0 & -\frac{1}{2} & 1 \end{bmatrix}$$

is not dynamic Leontief yet each P_δ is reducible. Note that B is Leontief (use $x' = (2, 0, 10, 7)$).

Assumption B may be stated in a computationally verifiable manner. Let $S = \{1, \cdots, m\}$. "State" $i \in S$ is accessible from j, $j \in S$, if there is a $\delta \in \Delta$ and n, where $0 \leqq n \leqq m - 1$, which give $(P^n_\delta)_{ij} > 0$.

Communicating classes are mutually accessible classes of states. Assumption B can be stated as "problem 2.1 has a single class." When there is more than one communicating class, one can determine the appropriate permutations to the dynamic Leontief block triangular form using a procedure given by Bather [1].

We now strengthen results of § 4 using Assumption B.

LEMMA 5.1. *If Assumption* B *holds, $d \geq 0$ and $P_\delta d \leqq d$ for every $\delta \in \Delta$, then $d > 0$.*

Proof. Assume that $P_\delta d \leqq d$ for all $\delta \in \Delta$ where $d \geq 0$. So $d'B \geqq 0$. Let $\cup(d) = \{i | d_i > 0\}$ and $\overline{\cup}(d) = \{i | d_i = 0\}$. We will show that $\overline{\cup}(d) = \emptyset$. Let $j \in A_l$ for some

$l \in \overline{\mathrm{U}}(d)$. The facts that $\sum_i d_iB_{ij} \geqq 0$ and $B_{ij} \leqq 0$ for $i \neq l$ imply that $B_{il} = 0$ for all $i \in \mathrm{U}(d)$. Thus, if $\overline{\mathrm{U}}(d) \neq \emptyset$, B is permutable to a dynamic Leontief matrix, which contradicts Assumption B. Hence, $\overline{\mathrm{U}}(d) = \emptyset$, or equivalently, $d > 0$. □

COROLLARY 5.2. *If Assumption* B *holds, then the following are equivalent*:

(a) *D is unbounded from above,*

(b) *$P_\delta d \leqq d$ for some $d > 0$,*

(c) *$\rho(P_\delta) \leq 1$ for all $\delta \in \Delta$.*

Proof. The implication (b)→(c)→(a) follows from Theorem 4.1. The implication (a)→(b) follows from Theorem 4.1 and Lemma 5.1. □

Notice that example (4.2) was an example of a dynamic Leontief system.

The following result completes the characterization of D under Assumption B in terms of the spectral properties of the P_δ's and shows when $C = R^m$.

THEOREM 5.3. *If Assumption* B *holds, then the following are equivalent*:

(a) *D is unbounded from above and has a nonempty interior.*

(b) $C = R^m$.

(c) *$\rho(P_\delta) \leqq 1$ for all $\delta \in \Delta$ and $c^\delta \notin$ range $(B^\delta)'$ whenever $\rho(P_\delta) = 1$.*

Proof. The fact that (b)→(c) follows from Theorem 3.6. We next show that (c)→(a). Assume (c) holds. It follows from Theorem 4.1 that for some $d \geq 0$, $P_\delta d \leqq d$ for every $\delta \in \Delta$. By Lemma 5.1, $d > 0$. The fact that D has a nonempty interior now follows directly from Theorem 4.3. Finally we prove that (a)→(b). If D is unbounded from above then from Theorem 4.1 and Lemma 5.1, for some $d > 0$, $P_\delta d \leqq d$ for every $\delta \in \Delta$. Corollary 4.4 gives us the rest. □

6. Conclusion.

It was shown under two assumptions that the dual feasible solution set of a Leontief substitution system can be characterized by the spectral radii of the P_δ matrices of the corresponding generalized Markov decision process. By use of these results it was then shown that such processes can be solved by value iterative type methods, independent of the starting vector, if and only if each $\rho(P_\delta) \leqq 1$ for $\delta \in \Delta$ and if $\rho(P_\delta) = 1$ no v solves $v = P_\delta v + c^\delta$.

Acknowledgment. We would like to thank the reviewers A. Federgruen and U. Rothblum for their valuable suggestions and time. In addition we would like to thank one of the reviewers for suggesting substantial organizational and stylistic revisions to the original manuscript. Included in these revisions were some alternative proofs.

We would also like to thank Professor Antal Majthay for his conversations and Professor Christopher Barry and Surya Kumar for several suggestions.

REFERENCES

[1] J. BATHER, *Optimal decision procedures for finite Markov chains. Part III: General convex systems*, Advances in Appl. Probability, 5 (1973), pp. 541–553.

[2] R. W. COTTLE AND A. F. VEINOTT, JR., *Polyhedral sets having a least element*, Math. Programming, 3 (1972), pp. 238–249.

[3] G. B. DANTZIG, *Optimal solution of a dynamic Leontief model with substitution*, Econometrica, 23 (1955), pp. 295–302.

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

[5] C. DERMAN, *Finite State Markovian Decision Process*, Academic Press, New York, 1970.

[6] R. C. GRINOLD, *A generalized discrete dynamic programming model*, Management Sci., 20 (1974), pp. 1092–1103.

[7] R. HOWARD AND J. MATHESON, *Risk sensitive Markov decision processes*, Ibid., 18 (1972), pp. 356–369.

[8] G. J. KOEHLER, *A generalization of Leontief substitution systems and their solution by recursive procedures*, Dept. of Management, Univ. of Florida, Gainesville, January 1975.

[9] U. G. ROTHBLUM, *Normalized Markov decision chains*, Operations Res., 23 (1975), pp. 785–795.

[10] E. SENETA, *Non-negative Matrices*, John Wiley, New York, 1973.

[11] R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[12] A. F. VEINOTT, JR., *Discrete dynamic programming with sensitive discount optimality criteria*, Ann. Math. Statist., 40 (1969), pp. 1635–1660.

[13] ———, *Extreme points of Leontief substitution systems*, Linear Algebra and Appl., 1 (1968), pp. 181–194.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 2, March 1979

RATES OF CONVERGENCE FOR CONDITIONAL GRADIENT ALGORITHMS NEAR SINGULAR AND NONSINGULAR EXTREMALS*

J. C. DUNN†

Abstract. Two conditional gradient algorithms are considered for the problem $\min_\Omega F$, with Ω a bounded convex subset of a Banach space. Neither method requires line search; one method needs no Lipschitz constants. Convergence rate estimates are similar in the two cases, and depend critically on the continuity properties of a set valued operator T whose fixed points, ξ, are the extremals of F in Ω. The continuity properties of T at ξ are determined by the way the function $a(\sigma)=\inf\{\rho=\langle F'(\xi), y-\xi\rangle | y\in\Omega, \|y-\xi\|\geqq\sigma\}$ grows with increasing σ. It is shown that for convex F and Lipschitz continuous F', the algorithms converge like $o(1/n)$, geometrically, or in finitely many steps, according to whether $a(\sigma)>0$ for $\sigma>0$, or $a(\sigma)\geqq A\sigma^2$ with $A>0$, or $a(\sigma)\geqq A\sigma$ with $A>0$. These three abstract conditiions are closely related to established notions of nonsingularity for an important class of optimal control problems with bounded control inputs. The first condition is satisfied (in $\mathcal{L}^1$) when meas $\{t|s(t)=0\}=0$, where $s(\cdot)$ is the switching function associated with the extremal control $\xi(\cdot)$; the second condition is satisfied when $s(\cdot)$ has finitely many zeros, all simple (typical of the bang-bang extremal); the third condition is satisfied when $s(\cdot)$ is bounded away from zero. Strong or uniform convexity assumptions are *not* invoked in the main convergence theorems. One of the theorems can be extended to a large subclass of quasiconvex functionals F.

1. Introduction. In [1], Demyanov and Rubinov consider two basic types of step length rule for conditional gradient processes in convex sets Ω. The first rule is a classic line minimization scheme in which the $(n+1)$st iterate x_{n+1} is gotten by minimizing the payoff functional F over a segment of a conditional gradient descent line issuing from x_n. When F is quadratic, there is a simple formula for x_{n+1}; however, in general it is necessary to approximate x_{n+1} with an inner iterative line search loop which may entail numerous and possibly costly evaluations of F. The second step size rule in [1] avoids this difficulty by minimizing a certain upper bound on the one dimensional section of F, in place of F itself. The bounding function is a simple quadratic expression in the step length parameter, and so it is an easy matter to solve for x_{n+1} (eqs. (4.2)–(4.3), in § 4). Moreover, it is shown in [1] that for convex F, Lipschitz continuous F', and convex, weakly compact Ω, this simple step size rule always produces minimizing sequences with $O(1/n)$ convergence at least, and geometric convergence under certain additional conditions of the uniform convexity type on Ω. However, from a practical standpoint, the utility of (4.2) (and all its variants in [1]) is compromised to some degree because this rule requires explicit knowledge of a Lipschitz constant for the derivative F' of F, i.e. a constant L satisfying

$$L\geqq L_0=\sup_{\substack{x,y\in\Omega\\ x\neq y}}\frac{\|F'(x)-F'(y)\|}{\|x-y\|}. \tag{1.1}$$

Computing a bound on L_0 can be a formidable problem in its own right, but if (4.2) is used with a constant $L<L_0$, the resulting step length parameters ω_n may never become small enough to insure that x_n is a minimizing sequence for F.

Ideally, one would like to have a conditional gradient step size rule which avoids line search, does not require inaccessible "nonlocal" parameters of the Lipschitz type, and yet manages to match the performance of other known rules. The open loop step length parameter sequences devised in [2] are simple in the extreme, but still produce

* Received by the editors October 14, 1977.

† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27607. This investigation was supported by a grant from the North Carolina Engineering Foundation.

the same worst case convergence rate ($O(1/n)$) as the closed loop rules in [1]. On the other hand $O(1/n)$ convergence is essentially the *best* one can expect from open loop schemes, whereas closed loop rules are capable of achieving geometric convergence under certain conditions. In the present paper, an analytical device employed in [2] is modified to produce the one parameter family of simple closed loop step length rules (4.9) which resemble (4.2) but require no Lipschitz constants. Under the same conditions invoked in [1] on F and Ω, *every* member of the family (4.9) generates minimizing sequences with asymptotic properties comparable to sequences obtained with (4.2)–(4.3). In fact, for certain values of the parameter θ, the a priori error estimates for (4.9)–(4.10) are actually *better* than the corresponding estimates for (4.2)–(4.3).

In the analysis to follow, it will be shown that the behavior of conditional gradient sequences near a minimizer ξ of F differs markedly according to whether a certain multivalued mapping $T\colon \Omega \to 2^{\Omega}$ is or is not "continuous" at ξ. The map in question determines the set of all conditional gradient directions emanating from each $x \in \Omega$, and its continuity properties at a minimizer ξ are pivotally dependent upon how $\langle F'(\xi), y\rangle$ grows as y moves away from ξ into Ω. If the monotone nondecreasing function, $a(\cdot)\colon [0, \infty) \to [0, \infty]$ defined by

$$a(\sigma) = \inf_{\substack{y \in \Omega \\ \|y-\xi\| \geqq \sigma}} \langle F'(\xi), y - \xi\rangle, \qquad \sigma \geqq 0,$$

is strictly positive for $\sigma > 0$ and if F' is continuous at ξ, then every single valued branch (i.e., section) of T is strongly continuous at ξ. Furthermore, if $a(\sigma) \geqq A\sigma^2$ for some constant $A > 0$, and if F' is locally Lipschitz continuous at ξ, then every branch of T is locally Lipschitz continuous at ξ. Finally, if $a(\sigma) \geqq A\sigma$, with $A > 0$, and if F' is continuous at ξ, then every branch of T is constant in some neighborhood of ξ. In § 5, these continuity properties support a comprehensive convergence theorem for the algorithm (4.9)–(4.10), along with a comparable theorem for (4.2)–(4.3) which improves significantly on the analysis in [1]. Under the same conditions imposed on F and Ω in [1], it is shown here that $F(x_n) - \inf_{\Omega} F = o(1/n)$ if $a(\sigma) > 0$ for $\sigma > 0$, that $F(x_n) - \inf_{\Omega} F$ and x_n converge geometrically if $a(\sigma) \geqq A\sigma^2$, and that x_n actually terminates at a minimizer of F after finitely many steps if $a(\sigma) \geqq A\sigma$. These conclusions run counter to the impression, created inadvertently by certain remarks in the literature (e.g., [3], [4]), that $O(1/n)$ convergence is the best one can expect of a conditional gradient method *unless* Ω satisfies some sort of uniform convexity condition. Actually, uniform convexity is part of a very strong *sufficient* condition for geometric convergence [1], [4], considerably stronger than the condition, $a(\sigma) \geqq A\sigma^2$, invoked here. This point is developed at length in §§ 3–5, and is made even more forcefully by the results obtained in § 6 for optimal control problems on admissible control sets with *empty interiors* in $\mathscr{L}^1$. It is of further interest that the geometric convergence theorem for the algorithm in [3] does not apply to the class of optimal control problems treated here.

In § 3, an extremal ξ for F in Ω is classified as i) singular, ii) nonsingular, iii) strongly nonsingular, iv) regular, or v) strongly regular, according to whether i) the linear functional $F'(\xi)$ has multiple minimizers in Ω, ii) $F'(\xi)$ has a unique minimizer (namely, ξ itself), iii) $a(\sigma) > 0$ for $\sigma > 0$, iv) $a(\sigma) \geqq A\sigma^2$ with $A > 0$, or v) $a(\sigma) \geqq A\sigma$ with $A > 0$. Within the context of optimal control theory, this general classification scheme is closely related to the Haynes–Hermes notion of singular optimal control [5] and the more comprehensive definitions formulated by Kelley et al. [6], [7], and Dunn [8]. Moreover, the min H method proposed in [9], the averaging methods treated in

[10], [11], [12], and the iteration scheme in [13] are all conditional gradient methods in one form or another, and conversely, the conditional gradient algorithms investigated here are immediately applicable to optimal control problems. In § 6, it is shown that for an important class of bounded optimal control problems, the position of an extremal control within the proposed classification scheme is determined by the behavior of an associated "switching function" $s(\cdot)$ near its zeros. Thus: if the set $\theta = \{t|s(t)=0\}$ has positive measure, the control is singular; if meas $\theta = 0$, the control is *strongly* nonsingular relative to the $\mathscr{L}^1$ norm; if θ consists of finitely many *simple* zeros (typical for "bang-bang" controls), the control is $\mathscr{L}^1$-regular; finally, if $s(\cdot)$ is bounded away from zero, the control is $\mathscr{L}^1$-strongly regular. When placed alongside the general convergence theorems in § 5, these results give added insight into why singular controls are aptly named.

2. Preliminaries. In the following development,

X = a real Banach space with norm $\|\cdot\|$
X^* = the dual space of continuous linear functionals, $x^*: X \to \mathbb{R}^1$
$\langle x^*, x\rangle$ = value of x^* at x
$\|x^*\|$ = the induced norm on X^*, i.e., $\sup_{\|x\|=1}\langle x^*, x\rangle$
Ω = nonempty subset of X
$K_\Omega(x)$ = cone of normals to Ω at x, i.e., $\{x^* \in X^* | \langle x^*, y-x\rangle \leqq 0, \forall y \in \Omega\}$
F = real functional on X
$F'(x)$ = Fréchet derivative of F at x
Ω_F = set of minimizers of F in Ω, i.e., $\{\xi \in \Omega | F(\xi) \leqq F(y), \forall y \in \Omega\}$
$T(x)$ = set of minimizers of $F'(x)$, i.e., $\{\bar{x} \in \Omega | \langle F'(x), \bar{x}\rangle \leqq \langle F'(x), y\rangle, \forall y \in \Omega\}$.

DEFINITION 2.1. $\xi \in \Omega$ is an *extremal* of F in Ω if and only if $F'(\xi)$ exists and

$$\langle F'(\xi), y-\xi\rangle \geqq 0, \quad \forall y \in \Omega. \tag{2.1}$$

THEOREM 2.1. *If Ω is convex, $\xi \in \Omega_F$ and $F'(\xi)$ exists, then ξ is an extremal of F. If F is convex and ξ is an extremal, then $\xi \in \Omega_F$.*

Proof. The proof may be found in standard references, e.g. [1], [14].

Note 2.1. The following statements are equivalent:

i) ξ is an extremal,
ii) $-F'(\xi) \in K_\Omega(\xi)$,
iii) $\xi \in T(\xi)$.

DEFINITION 2.2. $\{x_n\} \subset \Omega$ is a *conditional gradient sequence* if and only if there exist sequences $\{\bar{x}_n\} \subset \Omega$ and $\{\omega_n\} \subset [0, 1]$ such that

$$x_{n+1} = x_n + \omega_n(\bar{x}_n - x_n) \tag{2.2A}$$

$$\bar{x}_n \in T(x_n) \tag{2.2B}$$

for $n = 1, 2, \cdots$.

If Ω is convex and weakly compact, then $T(x)$ is never empty and (2.2) cannot lead out of Ω. However, as in [2], the principal concern here is with the convergence properties of conditional gradient sequences, on the assumption that such sequences exist. Therefore, compactness will be invoked only where it seems to have an essential bearing on the convergence of (2.2).

3. Singular and nonsingular extremals. At an extremal ξ, the set $T(\xi)$ must contain ξ (Theorem 2.2) and consequently is not empty. There are now just two possibilities: either ξ is the *only* element of $T(\xi)$ or else there are *several* elements in $T(\xi)$. The continuity properties of the set valued map $T: \Omega \to 2^\Omega$ are significantly different in the two cases, and this has an important bearing on the convergence of conditional gradient sequences.

DEFINITION 3.1. x is a *singular point in* Ω (relative to F) if and only if $T(x)$ is empty or has more than one member. Otherwise, x is a nonsingular point in Ω.

Note 3.1. An extremal ξ is singular if and only if $T(\xi)$ contains more than just ξ. Every extremal in the interior of Ω is necessarily singular since $F'(\xi)=0$ at such a point and consequently $T(\xi)=\Omega$. More generally, if an extremal ξ is not an extreme point of Ω, then ξ falls in the relative interior of some line segment $\ell \subset \Omega$, in which case $\langle F'(\xi), x-\xi\rangle = 0, \forall x \in \ell$, and therefore $T(\xi) \supset \ell$. Thus every extremal ξ which is not an extreme point of Ω must be singular. However, it is also possible to have a singular extremal at an extreme point. Various cases are illustrated in Fig. 3.1.

The following theorem reveals a fundamental connection between uniqueness conditions on $T(x)$ and the continuity properties of the map T at x.

THEOREM 3.1. *Let F' be continuous on Ω, and let Ω be compact. Then every single-valued branch of T is continuous at nonsingular points $x \in \Omega$. Conversely, at every singular point $x \in \Omega$, at least one single-valued branch of T is discontinuous.*

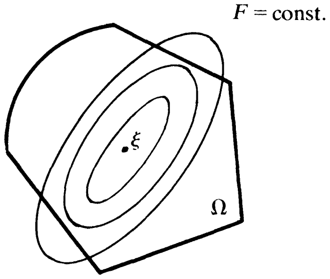

FIG. 3.1A

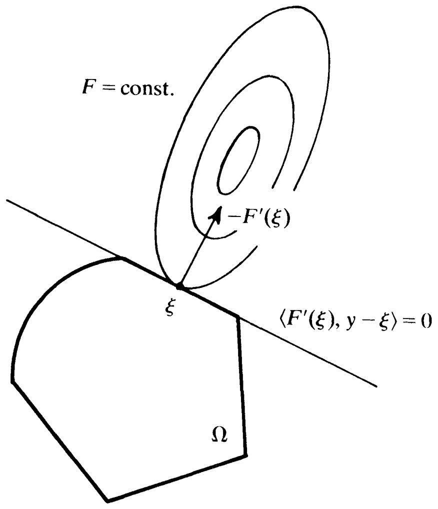

FIG. 3.1B

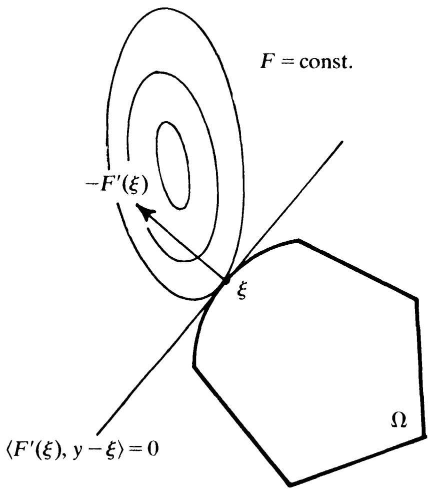

FIG. 3.1C

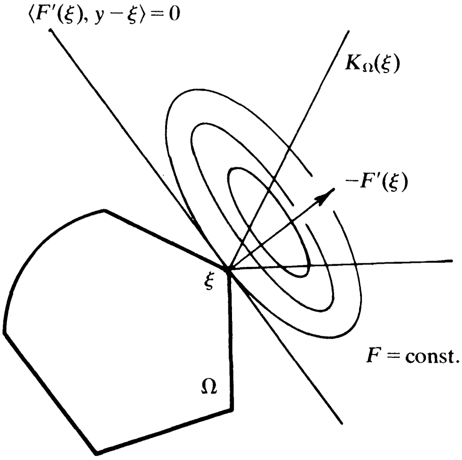

FIG. 3.1D

Proof. Ω compact$\Rightarrow T(x)\neq\emptyset$, $\forall x\in\Omega$, consequently there is at least one single-valued branch $\bar{T}\colon\Omega\to\Omega$ with $\bar{T}(x)\in T(x)$, $\forall x\in\Omega$. For any such $\bar{T}$, suppose that $x_n\to x$ and $\bar{T}(x_{n_k})\to\bar{x}\in\Omega$. For each k and each $y\in\Omega$, one then has

$$\begin{aligned}\langle F'(x_{n_k}), y-\bar{x}\rangle &\geqq \langle F'(x_{n_k}), \bar{T}(x_{n_k})-\bar{x}\rangle\\ &=\langle F'(x_{n_k})-F'(x), \bar{T}(x_{n_k})-\bar{x}\rangle+\langle F'(x), \bar{T}(x_{n_k})-\bar{x}\rangle\\ &\geqq -\|F'(x_{n_k})-F'(x)\|\|\bar{T}(x_{n_k})-\bar{x}\|+\langle F'(x), \bar{T}(x_{n_k})-\bar{x}\rangle.\end{aligned}$$

Fix y and let $k\to\infty$ to obtain

$$\langle F'(x), y-\bar{x}\rangle\geqq 0,\quad \forall y\in\Omega.$$

Thus, $x_n\to x\Rightarrow$ all cluster points $\bar{x}$ of $\{\bar{T}(x_n)\}$ fall in $T(x)$. Since x is nonsingular, $\bar{T}(x)$ is the only element in $T(x)$, and therefore all cluster points of $\{\bar{T}(x_n)\}$ coincide with $\bar{T}(x)$. Because Ω is compact, this means that $\bar{T}(x_n)\to\bar{T}(x)$.

Conversely, suppose $\bar{T}$ is continuous at x. If x is singular, there is an $\bar{x}\in T(x)$ with $\bar{x}\neq\bar{T}(x)$. Put

$$\bar{T}_1(y)=\begin{cases}\bar{T}(y), & \text{if } y\in\Omega \text{ and } y\neq x,\\ \bar{x}, & \text{if } y=x.\end{cases}$$

Then $\bar{T}_1$ is a single-valued branch of T, discontinuous at x. Q.E.D.

Note 3.2. Theorem 3.1 has several straightforward "weak" extensions. For instance, if compactness is replaced by weak compactness, then every branch $\bar{T}$ is strong-weak continuous at nonsingular points. If continuity of F' is replaced by weak-strong continuity, then every $\bar{T}$ is weak-strong (resp., weak-weak) continuous at nonsingular points if Ω is compact (resp., weakly compact). In all such cases, $\bar{T}(x_n)-x_n$ approaches 0 in some sense, as x_n approaches a nonsingular extremal ξ, and this favors the more rapid convergence of conditional gradient sequences. It is also possible to improve and considerably extend the strong continuity result in Theorem 3.1. This is accomplished in Theorems 3.2 and 3.3, after the following preparations.

DEFINITION 3.2. x is a *strongly nonsingular point in* Ω (relative to F and the norm on X) if and only if x is nonsingular in Ω, and in addition

$$\sigma>0\quad\Rightarrow\quad \inf_{\substack{y\in\Omega\\ \|y-\bar{x}\|\geqq\sigma}}\langle F'(x), y-\bar{x}\rangle>0 \tag{3.1}$$

where $\bar{x}$ is the unique element in $T(x)$.

LEMMA 3.1. *For $z\in\Omega$ and $-z^*\in K_\Omega(z)$, put*

$$a(\sigma)=\inf_{\substack{y\in\Omega\\ \|y-z\|\geqq\sigma}}\langle z^*, y-z\rangle,\qquad \sigma\geqq 0, \tag{3.2}$$

and

$$b(\sigma)=\frac{a(\sigma)}{\sigma},\qquad \sigma>0. \tag{3.3}$$

Suppose that Ω is either bounded or convex, and that $\infty\geqq a(\sigma)>0$ for $\sigma>0$. Then for all $\varepsilon>0$, $b(\sigma)$ is bounded away from zero on the half line $\sigma\geqq\varepsilon$.

Proof. It follows immediately from (3.2) and the definition of K_Ω, that $a(\cdot)$ is always monotone nondecreasing, with $\infty\geqq a(\sigma)\geqq 0$, for $\sigma\geqq 0$. If $a(\sigma)$ is strictly positive for $\sigma>0$, then $\sigma\geqq\varepsilon>0\Rightarrow a(\sigma)\geqq a(\varepsilon)>0$.

Case i) (Ω is bounded): Ω bounded $\Rightarrow d=\sup_{y\in\Omega}\|y-z\|<\infty$. Furthermore

$$d\geqq\sigma\geqq\varepsilon>0 \quad\Rightarrow\quad \frac{a(\sigma)}{\sigma}\geqq\frac{a(\varepsilon)}{d} \quad\text{and}\quad \sigma>d\geqq\varepsilon>0 \quad\Rightarrow\quad \frac{a(\sigma)}{\sigma}=\infty\geqq\frac{a(\varepsilon)}{d}.$$

Finally, $\sigma\geqq\varepsilon>d\geqq0\Rightarrow a(\sigma)/\sigma=\infty$.

Case ii) (Ω is convex): For $\sigma>0$, let $B_\sigma=\{u\in X|\exists y\in\Omega,\ u=(y-z)/\sigma\}$. Then

$$b(\sigma)=\frac{a(\sigma)}{\sigma}=\inf_{\substack{y\in\Omega\\ \|y-z\|\geqq\sigma}}\left\langle z^*,\frac{y-z}{\sigma}\right\rangle=\inf_{\substack{u\in B_\sigma\\ \|u\|\geqq1}}\langle z^*,u\rangle.$$

Furthermore, Ω convex $\Rightarrow \omega u\in B_\sigma$, $\forall u\in B_\sigma$, $\forall\omega\in[0,1]$. If $\sigma_2>\sigma_1>0$ and $u\in B_{\sigma_2}$, then $\exists y\in\Omega\ni u=(y-z)/\sigma_2=\omega(y-z)/\sigma_1$, with $\omega=\sigma_1/\sigma_2\in(0,1)$. Since $(y-z)/\sigma_1\in B_{\sigma_1}$, this gives $u\in B_{\sigma_1}$. Consequently, $\sigma_2>\sigma_1>0\Rightarrow B_{\sigma_1}\supset B_{\sigma_2}$, and therefore $\infty\geqq b(\sigma_2)\geqq b(\sigma_1)\geqq0$. It follows that $\sigma\geqq\varepsilon\Rightarrow b(\sigma)\geqq b(\varepsilon)=a(\varepsilon)/\varepsilon>0$. Q.E.D.

LEMMA 3.2. *For $z\in\Omega$ and $-z^*\in K_\Omega(z)$, let $a(\cdot)$ be the corresponding function in* (3.2). *If Ω is convex, then*

$$a(\sigma)=\inf_{\substack{y\in\Omega\\ \|y-z\|=\sigma}}\langle z^*,y-z\rangle,\quad\forall\sigma\geqq0. \tag{3.4}$$

Proof. Let $y\in\Omega$ with $\|y-z\|\geqq\sigma$, and put $u=z+(\sigma/\|y-z\|)(y-z)$. If Ω is convex, then $u\in\Omega$. Also, $\|u-z\|=\sigma$ and $\langle z^*,u-z\rangle\leqq\langle z^*,y-z\rangle$. Consequently,

$$\inf_{\substack{y\in\Omega\\ \|y-z\|\geqq\sigma}}\langle z^*,y-z\rangle\geqq\inf_{\substack{y\in\Omega\\ \|y-z\|=\sigma}}\langle z^*,y-z\rangle\geqq\inf_{\substack{y\in\Omega\\ \|y-z\|\geqq\sigma}}\langle z^*,y-z\rangle.\quad\text{Q.E.D.}$$

THEOREM 3.2. *Let F' be continuous on Ω and suppose that Ω is either bounded or convex. Then at every strongly nonsingular x in Ω, the set valued map $T\colon\Omega\to2^\Omega$ is continuous in the sense that*

$$\forall\varepsilon>0,\exists\delta>0\ni y\in\Omega\text{ and }\|y-x\|\leqq\delta\ \Rightarrow\ \|\bar y-\bar x\|\leqq\varepsilon,\forall\bar y\in T(y) \tag{3.5}$$

where $\bar x$ is the unique element in $T(x)$. In particular, this means that every single-valued branch of T is continuous at x.

Proof. If $\bar x\in T(x)$, then $-F'(x)\in K_\Omega(\bar x)$. Let $a(\cdot)$ and $b(\cdot)$ be given by (3.2) and (3.3), with $z=\bar x$ and $z^*=F'(x)$. If x is strongly nonsingular, then $a(\sigma)>0$ for $\sigma>0$, and it follows from Lemma 3.1, that for all $\varepsilon>0$, $b(\sigma)$ is bounded away from 0 on $\sigma\geqq\varepsilon$. Consequently, $\forall\varepsilon>0$, $\exists\mu>0\ni0\leqq b(\sigma)\leqq\mu\Rightarrow0\leqq\sigma<\varepsilon$. Furthermore, for $\bar y\in T(y)$,

$$\begin{aligned}\|F'(x)-F'(y)\|\|\bar y-\bar x\|&\geqq\langle F'(x)-F'(y),\bar y-\bar x\rangle+\langle F'(y),\bar y-\bar x\rangle\\&=\langle F'(x),\bar y-\bar x\rangle\\&\geqq a(\|\bar y-\bar x\|).\end{aligned}$$

Since F' is continuous, $\exists\delta>0\ni\|y-x\|\leqq\delta\Rightarrow\|F'(x)-F'(y)\|\leqq\mu$. Consequently, $\|y-x\|\leqq\delta\Rightarrow\mu\|\bar y-\bar x\|\geqq a(\|\bar y-\bar x\|)\Rightarrow\bar x=\bar y$ or $\mu\geqq b(\|\bar y-\bar x\|)\geqq0\Rightarrow\|\bar y-\bar x\|\leqq\varepsilon$. Q.E.D.

THEOREM 3.3. *Let Ω be convex and suppose that either one of the following conditions also holds*:

(i) *Ω is compact, or*

(ii) *Ω is closed and X is finite dimensional.*

Then x is a strongly nonsingular point in Ω if and only if x is a nonsingular point in Ω.

Proof. Every strongly nonsingular point is nonsingular. Conversely, suppose that x is nonsingular and that $\bar{x}$ is the unique element in $T(x)$. In Lemma 3.2, put $z=\bar{x}$ and $z^*=F'(x)$, to obtain

$$\inf_{\substack{y\in\Omega\\ \|y-\bar{x}\|\geqq\sigma}} \langle F'(x), y-\bar{x}\rangle = \inf_{\substack{y\in\Omega\\ \|y-\bar{x}\|=\sigma}} \langle F'(x), y-\bar{x}\rangle.$$

By hypothesis, Ω is compact, or the sphere $\{y\in X \mid \|y-\bar{x}\|=\sigma\}$ is compact and Ω is closed. In either case, the intersection $\Omega_\sigma=\{y\in\Omega \mid \|y-\bar{x}\|=\sigma\}$ is compact and the infimum on the right is achieved somewhere in Ω_σ. Moreover, since x is nonsingular, one has $\langle F'(x), y-\bar{x}\rangle>0$ for $y\neq\bar{x}$, consequently the infimum on the right is positive for $\sigma>0$, and so x is strongly nonsingular. Q.E.D.

Note 3.3. Let $\overline{\mathrm{Co}\,\Omega}$ denote the closure of the convex hull of Ω. Then for every $x\in\Omega$, $K_\Omega(x)=K_{\overline{\mathrm{Co}\,\Omega}}(x)$. Let $-z^*\in K_\Omega(z)=K_{\overline{\mathrm{Co}\,\Omega}}(z)$ and put

$$\underline{a}(\sigma)=\inf_{\substack{y\in\overline{\mathrm{Co}\,\Omega}\\ \|y-z\|\geqq\sigma}} \langle z^*, y-z\rangle,\qquad \sigma\geqq 0, \tag{3.6}$$

$$\underline{b}(\sigma)=\frac{\underline{a}(\sigma)}{\sigma},\qquad \sigma>0. \tag{3.7}$$

Then $a(\sigma)\geqq\underline{a}(\sigma)\geqq 0$ and $b(\sigma)\geqq\underline{b}(\sigma)\geqq 0$, where $a(\cdot)$ and $b(\cdot)$ are given by (3.2) and (3.3). Suppose $x\in\Omega$ is strongly nonsingular in $\overline{\mathrm{Co}\,\Omega}$, i.e., $\underline{a}(\sigma)>0$, $\forall\sigma>0$; then it follows at once from Lemma 3.1, that $\underline{b}(\sigma)$ (and, a fortiori, $b(\sigma)$) is bounded away from 0 on half lines $\sigma\geqq\varepsilon>0$. This means that condition (3.5) in Theorem 3.2 actually holds at every x which is strongly nonsingular in $\overline{\mathrm{Co}\,\Omega}$, irrespective of whether Ω is bounded or convex. Theorem 3.3 has a similar extension; thus if $x\in\Omega$ is nonsingular in $\overline{\mathrm{Co}\,\Omega}$ (i.e., $\langle F'(x), y-\bar{x}\rangle>0$, $\forall y\in\overline{\mathrm{Co}\,\Omega}$, $y\neq\bar{x}$) and if $\overline{\mathrm{Co}\,\Omega}$ is compact ($\Leftrightarrow\bar{\Omega}$ is compact) or X is finite dimensional, then x is strongly nonsingular in $\overline{\mathrm{Co}\,\Omega}$. What really matters for the continuity condition (3.5) is not the convexity or boundedness of Ω per se, but rather how the function $b(\cdot)$ behaves near 0 and ∞. One can also see now that the distinction between nonsingularity and strong nonsingularity has little significance in finite dimensional spaces. However, this is not true for infinite dimensional X, as the following example demonstrates.

Example 3.1. In the Hilbert space of ℓ^2 sequences $x=\{x_1, x_2, \cdots, x_n, \cdots\}$, let $e^{(1)}=\{1, 0, 0, \cdots\}$, $e^{(2)}=\{0, 1, 0, \cdots\}$ etc., and let $\Omega=$ closure of the convex hull of the vectors 0 and $e^{(n)}$, $n=1, 2, \cdots$; Ω is a closed bounded convex subset of ℓ^2. At $x=0$, the cone of normals to Ω consists of all vectors in ℓ^2 with nonpositive components, i.e., $K_\Omega(0)=\{x^*\in\ell^2 \mid x_i^*\leqq 0, 1\leqq i<\infty\}$. Suppose that for some Fréchet differentiable functional $F\colon \ell^2\to\mathbb{R}^1$, and for some $x\in\Omega$, one has $F'(x)=w\in\ell^2$, with $w_i>0$ for $1\leqq i<\infty$. Then $-F'(x)\in K_\Omega(0)$. Moreover, since the w_i's are strictly positive, $y\in\Omega$ and $y\neq 0\Rightarrow\langle F'(x), y\rangle=\sum_{i=1}^{\infty} w_i y_i>0$, consequently x is a nonsingular point in Ω and 0 is the unique element in $T(x)$. However, since $e^{(n)}\in\Omega$, $\|e^{(n)}\|=1$, and $\lim_{n\to\infty}\langle F'(x), e^{(n)}\rangle=\lim_{n\to\infty} w_n=0$, it follows that

$$\inf_{\substack{y\in\Omega\\ \|y\|\geqq\sigma}} \langle F'(x), y\rangle=0,\qquad \forall\sigma\in[0, 1].$$

Thus x is nonsingular, but *not* strongly nonsingular (relative to the ℓ^2 norm).

Roughly speaking, x will be strongly nonsingular if the boundary of Ω curves away from the supporting hyperplane $\{y \in X | \langle F'(x), y-\bar{x}\rangle = 0\}$, "uniformly" with respect to directions leading into Ω from $\bar{x}$. But even if the boundary of Ω contains line segments issuing from $\bar{x}$, x will still be strongly nonsingular if $-F'(x)$ lies in the *interior* of the normal cone at $\bar{x}$; in fact this is the ultimate form of strong nonsingularity. These ideas are developed below.

The following theorem establishes an important link between strong nonsingularity and conditions of the uniform convexity type. This connection is implicit in Theorem 4.3 of [1, p. 56] and Theorem 1.8 of [1, p. 131].

THEOREM 3.4. *Suppose that Ω satisfies the uniform convexity condition*

$$x, y \in \Omega \text{ and } \|z\| \leqq \gamma(\|x-y\|) \quad \Rightarrow \quad \frac{x+y}{2}+z \in \Omega \tag{3.8}$$

where $\gamma(\cdot)$ is some monotone nondecreasing function, with $\gamma(0)=0$ and $\gamma(\sigma)>0$, $\forall \sigma>0$. Let F' exist on Ω. Then for $x \in \Omega$, $\bar{x} \in T(x)$, and $\sigma \geqq 0$,

$$a(\sigma) = \inf_{\substack{y \in \Omega \\ \|y-\bar{x}\| \geqq \sigma}} \langle F'(x), y-\bar{x}\rangle \geqq 2\|F'(x)\|\gamma(\sigma). \tag{3.9}$$

Consequently, x is strongly nonsingular in Ω if and only if $T(x)$ is not empty and $F'(x) \neq 0$. In particular, an extremal ξ is strongly nonsingular in Ω if and only if $F'(\xi) \neq 0$.

Proof. For $y \in \Omega$, write $\langle F'(x), y-\bar{x}\rangle = 2\langle F'(x), (y+\bar{x})/2 - z\rangle + 2\langle F'(x), z-\bar{x}\rangle$. If $\|z\| \leqq \gamma(\|y-\bar{x}\|)$, then $(y+\bar{x})/2 - z \in \Omega$, and so $\langle F'(x), (y+\bar{x})/2-z\rangle \geqq \langle F'(x), \bar{x}\rangle$. It follows that $\langle F'(x), y-\bar{x}\rangle \geqq 2\langle F'(x), z\rangle$ if $\|z\| \leqq \gamma(\|y-\bar{x}\|)$. Consequently,

$$\inf_{\substack{y \in \Omega \\ \|y-\bar{x}\| \geqq \sigma}} \langle F'(x), y-\bar{x}\rangle \geqq 2 \sup_{\substack{z \in X \\ \|z\| \leqq \gamma(\sigma)}} \langle F'(x), z\rangle = 2\|F'(x)\|\gamma(\sigma).$$

The rest is now immediate from the definitions. Q.E.D.

Note 3.4. In applications, Ω is typically specified by inequality constraints, $g_i(x) \leqq 0$, $\forall i \in I$. Consequently, it is worth noting that certain uniformly convex functionals $g: X \to \mathbb{R}^1$ have uniformly convex level sets, $\Omega = \{x \in X | g(x) \leqq 0\}$. Thus, if

$$g\left(\frac{x+y}{2}\right) \leqq \frac{g(x)+g(y)}{2} - \delta(\|y-x\|), \quad \forall x, y \in X \tag{3.10}$$

for some positive definite nondecreasing $\delta(\cdot)$, and if g' exists and is bounded on the level set Ω, then Ω is uniformly convex. The prototype for this is $g(x) = \|x\|^2 - R^2$ in Hilbert space; here, the parallelogram law establishes (3.10) with $\delta(\sigma) = \sigma^2/4$, the closed ball $\Omega = \{x \in X | \|x\|^2 - R^2 \leqq 0\}$ satisfies the corresponding uniform convexity condition

$$x, y \in \Omega \text{ and } \|z\| \leqq \frac{1}{8R}\|x-y\|^2 \quad \Rightarrow \quad \frac{x+y}{2}+z \in \Omega,$$

and condition (3.9) becomes

$$a(\sigma) = \inf_{\substack{y \in \Omega \\ \|y-\bar{x}\| \geqq \sigma}} \langle F'(x), y-\bar{x}\rangle \geqq A\sigma^2 \tag{3.11}$$

with $A = \|F'(x)\|/(4R)$. Finally, the intersection of a *family* of uniformly convex sets

Ω_i, $i \in I$, is uniformly convex if $\gamma(\sigma) = \inf_{i\in I} \gamma_i(\sigma) > 0$, $\forall \sigma > 0$; in particular, this is so if the index set I is finite.

Note 3.5. For (3.9) to hold at any one $x \in \Omega$, it is sufficient that Ω is "separated" from the hyperplane $\{y \in X \mid \langle F'(\xi), y - \bar{x}\rangle = 0\}$ by some uniformly convex set Γ, in the sense that $\Gamma \supset \Omega$ and $-F'(x) \in K_\Gamma(\bar{x}) \subset K_\Omega(\bar{x})$ (Fig. 3.2).

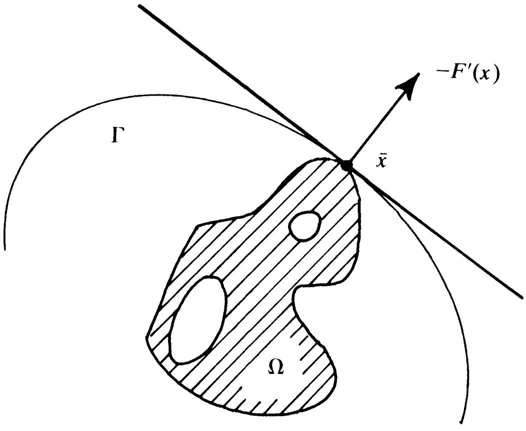

FIG. 3.2

THEOREM 3.5. *$x \in \Omega$ satisfies the strong nonsingularity condition*

$$(3.12) \qquad a(\sigma) = \inf_{\substack{y\in\Omega \\ \|y-\bar{x}\|\geqq\sigma}} \langle F'(x), y - \bar{x}\rangle \geqq A\sigma$$

with $\bar{x} \in T(x)$ and $A > 0$, if and only if $-F'(x)$ lies in the interior of the normal cone $K_\Omega(\bar{x})$.

Proof. For $x^* \in X^*$, $\langle -F'(x) + x^*, y - \bar{x}\rangle = \langle -F'(x), y - \bar{x}\rangle + \langle x^*, y - \bar{x}\rangle \leqq -A\|y - \bar{x}\| + \|x^*\|\|y - \bar{x}\|$. Consequently, $\|x^*\| \leqq A \Rightarrow -F'(x) + x^* \in K_\Omega(\bar{x})$, and therefore $-F'(x) \in \operatorname{Int} K_\Omega(\bar{x})$. Conversely, if $-F'(x) \in \operatorname{Int} K_\Omega(\bar{x})$, then for some $A > 0$, $\|x^*\| \leqq A \Rightarrow -F'(x) + x^* \in K_\Omega(\bar{x})$. For $y \in \Omega$ and $\|x^*\| \leqq A$, one then has $\langle F'(x), y - \bar{x}\rangle = \langle F'(x) - x^*, y - \bar{x}\rangle + \langle x^*, y - \bar{x}\rangle \geqq \langle x^*, y - \bar{x}\rangle$. Consequently, $\langle F'(x), y - \bar{x}\rangle \geqq \sup_{\|x^*\|\leqq A} \langle x^*, y - \bar{x}\rangle = A\|y - \bar{x}\|$. Q.E.D.

Figure 3.3 gives a finite dimensional illustration of the essential content of this proof, and Theorem 6.2 shows that the condition $-F'(x) \in \operatorname{Int} K_\Omega(\bar{x})$ can be realized for nontrivial optimization problems on infinite dimensional spaces. Notice, however, that the normal cone $K_\Omega(0)$ in Example 3.1 has an empty interior.

If F' is locally Lipschitz continuous at $x \in \Omega$, and if $T(x) \neq \emptyset$ and x satisfies the strong nonsingularity condition (3.11), it turns out that the set valued map T is also locally Lipschitz continuous at x in a certain sense; moreover, if F' is merely continuous and x satisfies the more stringent condition (3.12) with $T(x) \neq \emptyset$, then T is actually constant near x. These results (Theorems 3.6 and 3.7) and the associated conditional gradient convergence theorems in § 5, justify the following additional terminology.

DEFINITION 3.3. x is a *regular* (resp., *strongly regular*) point in Ω if and only if x is nonsingular in Ω and satisfies condition (3.11) (resp. (3.12)) for some $A > 0$, with $\bar{x} =$ the unique element in $T(x)$.

Note 3.6. If Ω is bounded, (3.12) $\Rightarrow$ (3.11). Also, (3.11) $\Rightarrow$ Ω is bounded.

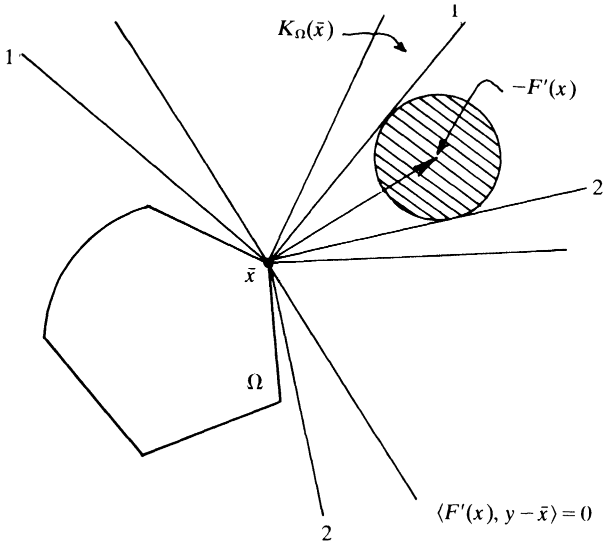

FIG. 3.3

THEOREM 3.6. *Let F' be locally Lipschitz continuous at $x \in \Omega$, i.e.,*

$$(3.13) \qquad \exists L>0, \exists \delta_0>0 \ni \|y-x\| \leqq \delta_0 \quad \Rightarrow \quad \|F'(y)-F'(x)\| \leqq L\|y-x\|.$$

If x is a regular point in Ω, then the set valued map $T: \Omega \to 2^\Omega$ is locally Lipschitz continuous at x in the sense that

$$(3.14) \qquad \|y-x\| \leqq \delta_0 \quad \Rightarrow \quad \|\bar{y}-\bar{x}\| \leqq M\|y-x\|, \forall \bar{y} \in T(y)$$

where $\bar{x}$ is the unique element in $T(x)$, $M = L/A$, and A is the constant in (3.11).

Proof. For all $\bar{y} \in T(y)$, $\|F'(x)-F'(y)\|\|\bar{y}-\bar{x}\| \geqq \langle F'(x)-F'(y), \bar{y}-\bar{x}\rangle + \langle F'(y), \bar{y}-\bar{x}\rangle = \langle F'(x), \bar{y}-\bar{x}\rangle \geqq A\|\bar{y}-\bar{x}\|^2$. Therefore $\|y-x\| \leqq \delta_0 \Rightarrow L\|y-x\|\|\bar{y}-\bar{x}\| \geqq A\|\bar{y}-\bar{x}\|^2 \Rightarrow \|\bar{y}-\bar{x}\| \leqq (L/A)\|y-x\|$. Q.E.D.

THEOREM 3.7. *Let F' be continuous at x. If x is a strongly regular point in Ω, then the set valued map $T: \Omega \to 2^\Omega$ is constant near x; more precisely,*

$$(3.15) \qquad \exists \delta_0>0 \ni \|y-x\| \leqq \delta_0 \quad \Rightarrow \quad T(y) = \{\bar{x}\}$$

where $\bar{x}$ is the unique element in $T(x)$.

Proof. According to Theorem 3.5, $-F'(x) \in \operatorname{Int} K_\Omega(\bar{x})$. Therefore, since F' is continuous at x, $\exists \delta_0>0 \ni \|y-x\| \leqq \delta_0 \Rightarrow -F'(y) \in \operatorname{Int} K_\Omega(\bar{x})$. A second application of Theorem 3.5 now yields $T(y) = \{\bar{x}\}$ (Fig. 3.4). Q.E.D.

Note 3.7. In a general metric space setting, Dantzig et al. [15] investigate upper semicontinuity properties of minimizer sets under perturbations in the payoff *and* the constraint set. Robinson [16] obtains continuity estimates of the Lipschitz type for certain solution branches of perturbed linear programs in $\mathbb{R}^n$.

4. The algorithms. Let Ω be convex, let $\{x_n\} \subset \Omega$ be a conditional gradient sequence, and suppose that F' is Lipschitz continuous on Ω. According to Lemma 1.2 of [1, p. 117], one then has

$$(4.1\text{A}) \qquad 0 \leqq r_{n+1} \leqq r_n - \omega_n \langle F'(x_n), x_n - \bar{x}_n\rangle + \frac{\omega_n^2}{2} L\|x_n - \bar{x}_n\|^2$$

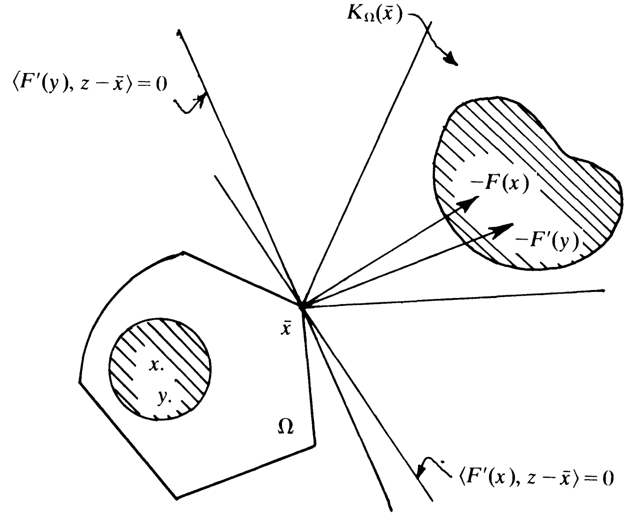
$K_\Omega(\bar{x})$
$\langle F'(y), z-\bar{x}\rangle = 0$
$-F(x)$
$-F'(y)$
$\bar{x}$
$x.$
$y.$
Ω
$\langle F'(x), z-\bar{x}\rangle = 0$

FIG. 3.4

with

$$r_n = F(x_n) - \inf_\Omega F \tag{4.1B}$$

where L is a Lipschitz constant for F' on Ω. In this expression, $\{\omega_n\}$ can be any sequence in $[0, 1]$ which satisfies (2.2) with $\{x_n\}$ and $\{\bar{x}_n\}$. Suppose now that at each n, ω_n is specified by the rule

$$\omega_n = \omega(x_n, \bar{x}_n; F) = \begin{cases} 0, & \text{if } \langle F'(x_n), x_n - \bar{x}_n\rangle = 0 \\ \dfrac{\langle F'(x_n), x_n - \bar{x}_n\rangle}{L\|x_n - \bar{x}_n\|^2}, & \text{if } 0 < \dfrac{\langle F'(x_n), x_n - \bar{x}_n\rangle}{L\|x_n - \bar{x}_n\|^2} < 1 \\ 1, & \text{if } 1 \leqq \dfrac{\langle F'(x_n), x_n - \bar{x}_n\rangle}{L\|x_n - \bar{x}_n\|^2} \end{cases} \tag{4.2}$$

which minimizes the right side of (4.1A) over $\omega_n \in [0, 1]$, given x_n and $\bar{x}_n$, and also insures that $x_n = x_N$, $\forall n \geqq N$, if x_N is an extremal. This is the basic closed loop step length rule proposed and analyzed in [1].

Given any single valued branch $\bar{T}$ of T, (2.2) and (4.2) produce a corresponding recursive algorithm for generating conditional gradient sequences, namely

$$x_{n+1} = x_n + \omega(x_n, \bar{x}_n; F)(\bar{x}_n - x_n); \qquad x_1 = z \in \Omega \tag{4.3A}$$

with

$$\bar{x}_n = \bar{T}(x_n). \tag{4.3B}$$

For each $z \in \Omega$, this scheme produces exactly one corresponding conditional gradient sequence $\{x_n\} \subset \Omega$ with $x_1 = z$ (provided $T(x) \neq \emptyset$ for $x \in \Omega$). Demyanov and Rubinov prove that if, in addition to the assumptions already listed, F is also convex and Ω is bounded, then $\inf_\Omega F > -\infty$, and for any branch $\bar{T}$ of T, the corresponding sequences $\{x_n\}$ generated by (4.3) are always minimizing sequences with $r_n = O(1/n)$; moreover, if Ω satisfies the uniform convexity condition

$$(4.4)\qquad x, y \in \Omega \text{ and } \|z\| \leqq \gamma\|x-y\|^2 \quad\Rightarrow\quad \frac{x+y}{2}+z \in \Omega$$

for some constant $\gamma>0$, and if $\inf_\Omega \|F'(x)\| \geqq \varepsilon > 0$, then $\{r_n\}$ converges to 0 geometrically, i.e., $0 \leqq r_n \leqq \text{const.}\,\lambda^n$, $\exists\lambda \in [0,1)$ [1, Thms. 1.7 and 1.8, Chap. 3]. Worst case convergence rate estimates for the classical line minimization step size rule are no better than this [17], and (4.2) has the important advantage of not requiring a potentially costly inner iterative line search loop. On the other hand, if (4.2)–(4.3) is used with an L which is *not* a Lipschitz constant for F', then (4.3) may generate nonminimizing and divergent sequences $\{x_n\}$. For example, let $X=\mathbb{R}^1$, $\Omega=[-1,1]$, and $F(x)=x^2/2$. Then $F'(x)=x$ and consequently $L_0=1$ in (1.1). If $0<L<\frac{1}{2}$ in (4.2) and if $x_1 \neq 0$, then (4.3) generates a sequence $\{x_n\}$ which terminates in a limit cycle $\cdots,-1,1,-1,\cdots$ beyond some value of n. On the other hand, if $L \geqq \frac{1}{2}$ then $x_n=(1-1/L)^{n-1}x_1$. Therefore, for this problem, (4.2)–(4.3) produces minimizing sequences if and only if $L>\frac{1}{2}=L_0/2$. The minimizer 0 happens to be singular here; however similar behavior is observed for problems with nonsingular solutions in spaces of arbitrary dimension.

The convergence theorems quoted above from [1] for (4.2)–(4.3), assume that F is convex. If this condition is invoked at the outset, the inequality (4.1A) can be carried further to

$$(4.5)\qquad 0 \leqq r_{n+1} \leqq (1-\omega_n)r_n + \frac{\omega_n^2}{2}L\|x_n-\bar{x}_n\|^2$$

since one can readily show that

$$(4.6)\qquad \infty > \langle F'(x_n), x_n-\bar{x}_n\rangle \geqq r_n \geqq 0, \quad \forall \bar{x}_n \in T(x_n)$$

when F is convex (cf. proof of Theorem 1 in [2]). Let θ be an arbitrary constant $\neq 0$, and divide (4.5) by $L\theta^2$ to obtain

$$(4.7\text{A})\qquad 0 \leqq R_{n+1} \leqq (1-\omega_n)R_n + \frac{\omega_n^2}{2\theta^2}\|x_n-\bar{x}_n\|^2$$

with

$$(4.7\text{B})\qquad R_n = \frac{r_n}{L\theta^2} = \frac{1}{L\theta^2}\left(F(x_n) - \inf_\Omega F\right).$$

In effect, θ is a kind of scaling parameter; the reason for introducing it here will become clear in § 5. In any case, if $\{x_n\}$ satisfies (2.2) with $\{\bar{x}_n\}$ and $\{\omega_n\} \subset [0,1]$, then (4.7) holds for all $n \geqq 1$, and a simple inductive argument yields

$$(4.8\text{A})\qquad R_n \leqq B\beta_n, \quad \forall n \geqq 1$$

with

$$(4.8\text{B})\qquad B = \max\{1, R_1\}$$

and $\{\beta_n\} \subset [0,\infty)$ recursively generated by

$$(4.8\text{C})\qquad \beta_{n+1} = (1-\omega_n)\beta_n + \frac{\omega_n^2}{2\theta^2}\|x_n-\bar{x}_n\|^2; \qquad \beta_1 = 1$$

Proceeding as before, let us now choose ω_n according to the rule,

$$(4.9)\quad \omega_n = \omega(x_n, \bar{x}_n; \beta_n) = \begin{cases} 0, & \text{if } \langle F'(x_n), x_n - \bar{x}_n\rangle = 0 \\ \dfrac{\theta^2\beta_n}{\|x_n - \bar{x}_n\|^2}, & \text{if } \langle F'(x_n), x_n - \bar{x}_n\rangle > 0 \text{ and } \dfrac{\theta^2\beta_n}{\|x_n - \bar{x}_n\|^2} < 1 \\ 1, & \text{if } \langle F'(x_n), x_n - \bar{x}_n\rangle > 0 \text{ and } \dfrac{\theta^2\beta_n}{\|x_n - \bar{x}_n\|^2} \geqq 1 \end{cases}$$

which minimizes the right side of (4.8C) over $\omega_n \in [0, 1]$ and insures that $x_n = x_N$, $\forall n \geqq N$, if x_N is an extremal. Given any single valued branch $\bar{T}$ of T, one then has the associated algorithm

$$(4.10\text{A})\qquad x_{n+1} = x_n + \omega_n(\bar{x}_n - x_n); \qquad x_1 = z \in \Omega,$$

$$(4.10\text{B})\qquad \beta_{n+1} = (1-\omega_n)\beta_n + \frac{\omega_n^2}{2\theta^2}\|x_n - \bar{x}_n\|^2; \qquad \beta_1 = 1,$$

$$(4.10\text{C})\qquad \bar{x}_n = \bar{T}(x_n),$$

$$(4.10\text{D})\qquad \omega_n = \omega(x_n, \bar{x}_n; \beta_n).$$

It is shown in § 5 that for *any fixed value of* $\theta \neq 0$, the sequences $\{x_n\}$ generated by (4.10) are always minimizing sequences for convex F on bounded Ω, even though a Lipschitz constant L does not appear explicitly in (4.9); in fact, this is also true for a somewhat larger class of quasiconvex F on bounded Ω (Note 5.4).

5. Convergence theorems.

5. Convergence theorems. The following result is a straightforward modification of Lemma 2 in [18].

LEMMA 5.1. *Let* $\{\beta_n\} \subset (0, \infty)$ *and* $\{q_n\} \subset [0, \infty)$ satisfy

$$\beta_{n+1} \leqq \beta_n - q_n\beta_n^2; \qquad \beta_1 = 1$$

for all $n \geqq 1$. *If* $q_n \geqq q > 0$, *for* $n \geqq 1$, *then* $\beta_n = O(1/n)$; *more precisely*

$$0 < \beta_n \leqq \frac{1}{1+q(n-1)}, \quad \forall n \geqq 1.$$

Furthermore, if $\lim_{n\to\infty} q_n = \infty$, *then* $\beta_n = o(1/n)$.

Proof. Put $\delta_k = 1/\beta_k$. Then for all $k \geqq 1$,

$$\delta_k - q_k > 0 \quad \text{and} \quad \delta_{k+1} - \delta_k \geqq \frac{q_k\delta_k}{\delta_k - q_k} \geqq q_k.$$

Consequently, for all $n \geqq 1$

$$\delta_n = \delta_1 + \sum_{k=1}^{n-1} (\delta_{k+1} - \delta_k) \geqq 1 + \sum_{k=1}^{n-1} q_k.$$

If $q_k \geqq q$ for all k, then

$$\beta_n = \frac{1}{\delta_n} \leqq \frac{1}{1+q(n-1)}, \quad \forall n \geqq 1.$$

Suppose that $q_n \to \infty$. Given $M > 0$, choose K so large that $k \geqq K \Rightarrow q_k \geqq M$. Then, for $n > K$,

$$0 < n\beta_n \leqq n\Big/\left(1 + \sum_{k=1}^{K-1} q_k + \sum_{k=K}^{n-1} q_k\right) \leqq \frac{n}{(n-K)M}$$

and consequently $0 \leqq \underline{\lim}_{n\to\infty} n\beta_n \leqq \overline{\lim}_{n\to\infty} n\beta_n \leqq 1/M$. Since M can be arbitrarily large, this gives $\lim_{n\to\infty} n\beta_n = 0$. Q.E.D.

LEMMA 5.2. *Let F be convex. If ξ is a nonsingular minimizer of F in Ω, then ξ is the* only *minimizer of F in Ω. Moreover, if ξ is strongly nonsingular in Ω, then every minimizing sequence for F converges strongly to ξ.*

Proof. $\xi \in \Omega_F$ and ξ nonsingular $\Rightarrow \langle F'(\xi), y-\xi\rangle > 0$, $\forall y \in \Omega$, $y \neq \xi$. F convex $\Rightarrow$ $F(y)-F(\xi) \geqq \langle F'(\xi), y-\xi\rangle$, $\forall y \in \Omega$. Therefore ξ is unique. Furthermore, $F(y)-F(\xi) \geqq \langle F'(\xi), y-\xi\rangle \geqq a(\|y-\xi\|) \geqq 0$, with

$$a(\sigma) = \inf_{\substack{y\in\Omega \\ \|y-\xi\|\geqq\sigma}} \langle F'(\xi), y-\xi\rangle.$$

Consequently, if $\{y_n\}$ is a minimizing sequence for F in Ω, then $a(\|y_n-\xi\|)\to 0$ as $n\to\infty$. But if ξ is strongly nonsingular, $a(\sigma)$ is strictly positive and monotone nondecreasing for $\sigma>0$. Therefore, $a(\|y_n-\xi\|)\to 0 \Rightarrow \|y_n-\xi\|\to 0$. Q.E.D.

THEOREM 5.1. *Suppose that Ω is convex and bounded with $D = \operatorname{diam}\Omega$, that F is convex, and that F' is Lipschitz continuous on Ω, with L a Lipschitz constant for F'. If the sequences $\{x_n\}\subset\Omega$, $\{\bar{x}_n\}\subset\Omega$, and $\{\beta_n\}\subset[0,\infty)$ are generated by the algorithm* (4.9)–(4.10) *with any $\theta\neq 0$, then*

$$0 \leqq r_n \leqq \max\{r_1, L\theta^2\}\cdot\beta_n, \quad \forall n \geqq 1 \tag{5.1A}$$

where

$$0 \leqq F(x_n) - \inf_{\Omega} F = r_n < \infty. \tag{5.1B}$$

Moreover, one of the following conditions holds: either x_N is a minimizer of F in Ω for some N and

$$x_n = x_N \in \Omega_F, \quad \forall n > N, \tag{5.2}$$

or else

$$0 < \beta_n \leqq \frac{1}{1+q(n-1)}, \quad \forall n \geqq 1 \tag{5.3A}$$

with

$$q = \frac{1}{2}\min\left\{1, \frac{\theta^2}{D^2}\right\}. \tag{5.3B}$$

In either case, $\{\beta_n\}$ is monotone nonincreasing. Finally:

i) *If ξ is a strongly nonsingular minimizer of F in Ω, and if* (5.2) *does not hold, then the sequences $\{x_n\}$ and $\{\bar{x}_n\}$ converge strongly to ξ, and $\beta_n = o(1/n)$.*

ii) *If ξ is a regular minimizer of F in Ω and if* (5.2) *does not hold, then*

$$0 < \beta_n \leqq \lambda^{n-1}, \quad \forall n \geqq 1 \tag{5.4A}$$

with

$$\lambda = \max\left\{\frac{1}{2}, 1-\frac{\mu}{2}\right\} \in \left[\frac{1}{2}, 1\right) \tag{5.4B}$$

and

$$\mu = \frac{1}{(1+L/A)^2}\cdot\frac{A\theta^2}{\max\{r_1, L\theta^2\}} \tag{5.4C}$$

where A is the constant in (3.11). *Moreover*, $\|x_n - \xi\| = O(\lambda^{n/2})$ *and* $\|\bar{x}_n - \xi\| = O(\lambda^{n/2})$.

iii) *If ξ is a strongly regular minimizer of F in Ω, then* (5.2) *always holds for some $N \geqq 1$, with $x_N = \xi$ and $\bar{x}_n = \xi$, $\forall n \geqq N$.*

Proof. (4.6) $\Rightarrow r_n < \infty$ and (4.8) $\Rightarrow$ (5.1). Suppose that $\langle F'(x_n), x_n - \bar{x}_n\rangle > 0$, for $1 \leqq n \leqq N-1$. Then for n in this range, x_n is not an extremal and consequently $r_n > 0$ by Theorem 2.1. It follows from (5.1) that $\beta_n > 0$, and (4.9)–(4.10) then yield

$$0 < \beta_{n+1} \leqq \beta_n - q_n\beta_n^2 < \beta_n \tag{5.5A}$$

with

$$q_n = \frac{1}{2}\min\left\{\frac{1}{\beta_n}, \frac{\theta^2}{\|x_n - \bar{x}_n\|^2}\right\} \geqq q = \frac{1}{2}\min\left\{1, \frac{\theta^2}{D^2}\right\} \tag{5.5B}$$

for $1 \leqq n \leqq N-1$. If $\langle F'(x_N), x_N - \bar{x}_N\rangle = 0$, then $x_N \in \Omega_F$ by Theorem 2.1, and condition (5.2) holds. On the other hand, if $\langle F'(x_n), x_n - \bar{x}_n\rangle > 0$, $\forall n \geqq 1$, then (5.5) holds with $\beta_n > 0$ for all n, and the a priori error estimate (5.1)–(5.3) follows from Lemma 5.1. In either case, $\{\beta_n\}$ is monotone nonincreasing.

i) If ξ is strongly nonsingular, then ξ is unique and $x_n \to \xi$, by Lemma 5.2, (5.1), (5.2) and (5.3). Furthermore, $\bar{x}_n \to \xi$, by Theorem 3.2, and consequently $\|x_n - \bar{x}_n\| \to 0$. If $\langle F'(x_N), x_N - \bar{x}_N\rangle = 0$ for some N, then (5.2) holds with $x_N = \xi$ and $\bar{x}_n = \xi$, $\forall n > N$. Otherwise, if $\langle F'(x_n), x_n - \bar{x}_n\rangle > 0$, $\forall n \geqq 1$, then (5.5B) gives $q_n \to \infty$, since $\beta_n \to 0$ and $\|x_n - \bar{x}_n\| \to 0$ through positive values. Consequently $\beta_n = o(1/n)$ by Lemma 5.1.

ii) If ξ is regular, then for all $n \geqq 1$

$$\beta_n \geqq \frac{r_n}{BL\theta^2} \geqq \frac{\langle F'(\xi), x_n - \xi\rangle}{BL\theta^2} \geqq \frac{A}{\max\{r_1, L\theta^2\}} \cdot \|x_n - \xi\|^2 \tag{5.6}$$

where A is the constant in (3.11). Moreover, by Theorem 3.6,

$$\|\bar{x}_n - \xi\| \leqq \frac{L}{A}\|x_n - \xi\|, \quad \forall n \geqq 1. \tag{5.7}$$

If $\langle F'(x_n), x_n - \bar{x}_n\rangle > 0$ for $1 \leqq n \leqq N-1$, then (5.6) and (5.7) give

$$\frac{\theta^2\beta_n}{\|x_n - \bar{x}_n\|^2} \geqq \frac{\theta^2\beta_n}{(\|x_n - \xi\| + \|\bar{x}_n - \xi\|)^2} \geqq \mu > 0 \tag{5.8}$$

for $1 \leqq n \leqq N-1$, where μ is the constant in (5.4C). It now follows from (5.5) and (5.8) that

$$0 < \beta_{n+1} \leqq \left(1 - \frac{1}{2}\min\{1, \mu\}\right)\beta_n = \max\left\{\frac{1}{2}, 1 - \frac{\mu}{2}\right\}\beta_n \tag{5.9}$$

for $1 \leqq n \leqq N-1$. If $\langle F'(x_N), x_N - \bar{x}_N\rangle = 0$, then $x_N \in \Omega_F$ and (5.2) holds, as in i), with $x_N = \xi$ and $\bar{x}_n = \xi$, $\forall n > N$. Otherwise, if $\langle F'(x_n), x_n - \bar{x}_n\rangle > 0$, $\forall n \geqq 1$, then (5.9) holds for all $n \geqq 1$ and this establishes (5.4). From (5.4), (5.6) and (5.7) one then obtains

$$\|x - \xi\| = O(\lambda^{n/2}) \quad \text{and} \quad \|\bar{x}_n - \xi\| = O(\lambda^{n/2}).$$

iii) If ξ is strongly regular, the inequality

$$\beta_n \geqq \frac{A}{\max\{r_1, L\theta^2\}}\|x_n - \xi\|, \quad \forall n \geqq 1 \tag{5.10}$$

replaces (5.6), where A is now the constant in (3.12). If $\langle F'(x_n), x_n-\bar{x}_n\rangle > 0, \forall n \geqq 1$, then $\|x_n-\xi\|>0, \forall n \geqq 1$, and

$$\frac{\theta^2\beta_n}{\|x_n-\bar{x}_n\|^2} \geqq \frac{A\theta^2}{\max\{r_1, L\theta^2\}} \cdot \frac{\|x_n-\xi\|}{(\|x_n-\xi\|+\|\bar{x}_n-\xi\|)^2}, \quad \forall n \geqq 1.$$

It has already been established that $\|x_n-\xi\| \to 0$; therefore by Theorem 3.6, $\bar{x}_n = \xi$ for all large n, and so,

$$\frac{\theta^2\beta_n}{\|x_n-\bar{x}_n\|^2} \geqq \frac{A}{\max\{r_1, L\theta^2\}} \cdot \frac{1}{\|x_n-\xi\|} \geqq 1 \tag{5.11}$$

for n sufficiently large. However, it follows from (4.9)–(4.10), and (5.11) that $\omega(x_n, \bar{x}_n; \beta_n) = 1$ and $x_{n+1} = \bar{x}_n = \xi$ for large n. This contradiction establishes (5.2) with $x_N = \xi$, and $\bar{x}_n = \xi, \forall n > N$. Q.E.D.

THEOREM 5.2. *Let Ω and F satisfy the conditions of Theorem* 5.1. *If the sequences $\{x_n\} \subset \Omega$ and $\{\bar{x}_n\} \subset \Omega$ are generated by the algorithm* (4.2)–(4.3), *then either* (5.2) *holds or else*

$$0 < r_n \leqq 2\max\{r_1, LD^2\} \cdot \frac{1}{n}, \quad \forall n \geqq 1 \tag{5.12A}$$

with

$$r_n = F(x_n) - \inf_{\Omega} F. \tag{5.12B}$$

In either case $\{r_n\}$ is monotone nonincreasing. Furthermore:

i) *If ξ is a strongly nonsingular minimizer of F in Ω and if* (5.2) *does not hold, then the sequences $\{x_n\}$ and $\{\bar{x}_n\}$ converge strongly to ξ, and $r_n = o(1/n)$.*

ii) *If ξ is a regular minimizer of F in Ω and if* (5.2) *does not hold, then*

$$0 < \frac{r_n}{r_1} \leqq \lambda^{n-1}, \quad \forall n \geqq 1 \tag{5.13A}$$

with

$$\lambda = \max\left\{\frac{1}{2}, 1-\frac{\mu}{2}\right\} \tag{5.13B}$$

and

$$\mu = \frac{1}{(L/A)} \cdot \frac{1}{(1+L/A)^2} \tag{5.13C}$$

where A is the constant in (3.11).

iii) *If ξ is a strongly regular minimizer of F in Ω then* (5.2) *always holds for some $N \geqq 1$, with $x_N = \xi$, and $\bar{x}_n = \xi, \forall n > N$.*

Proof. If $\langle F'(x_n), x_n-\bar{x}_n\rangle > 0$ for $1 \leqq n \leqq N-1$, then for n in this range, x_n is not an extremal and consequently $r_n > 0$, by Theorem 2.1. From (4.1)–(4.2) one then obtains for $1 \leqq n < N-1$

$$r_{n+1} \leqq \begin{cases} r_n - \frac{1}{2}\langle F'(x_n), x_n-\bar{x}_n\rangle, & \text{if } \dfrac{\langle F'(x_n), x_n-\bar{x}_n\rangle}{L\|x_n-\bar{x}_n\|^2} \geqq 1 \\ r_n - \dfrac{\langle F'(x_n), x_n-\bar{x}_n\rangle^2}{2L\|x_n-\bar{x}_n\|^2}, & \text{if } 1 > \dfrac{\langle F'(x_n), x_n-\bar{x}_n\rangle}{L\|x_n-\bar{x}_n\|^2} > 0. \end{cases} \tag{5.14}$$

Since F is convex, the estimate (4.6) holds and therefore

$$(5.15) \qquad 0 \leqq r_n^2 \leqq \langle F'(x_n), x_n - \bar{x}_n \rangle^2.$$

Thus (5.14) can be carried further to

$$(5.16\text{A}) \qquad r_{n+1} \leqq r_n - q_n r_n^2 < r_n$$

with

$$(5.16\text{B}) \qquad q_n = \frac{1}{2} \min \left\{ \frac{1}{r_n}, \frac{1}{L\|x_n - \bar{x}_n\|^2} \right\} \geqq q = \frac{1}{2} \min \left\{ \frac{1}{r_1}, \frac{1}{LD^2} \right\} > 0$$

for $1 \leqq n \leqq N-1$. If $\langle F'(x_N), x_N - \bar{x}_N \rangle = 0$, then $x_N \in \Omega_F$, by Theorem 2.1, and condition (5.2) holds. On the other hand, if $\langle F'(x_n), x_n - \bar{x}_n \rangle > 0$, $\forall n \geqq 1$, then (5.16) holds with $r_n > 0$ for all n, and the estimate (5.12) follows easily from Lemma 5.1. In either case, $\{r_n\}$ is monotone nonincreasing.

The remainder of the proof is obtained from the proof of Theorem 5.1 by putting $\theta^2 = 1/L$ and $B = 1$, and replacing β_n and (5.6) by r_n and (5.16) respectively. Q.E.D.

Note 5.1. It may be possible to sharpen the estimate (5.13) considerably if $\|F'(x)\|$ is bounded away from 0 on Ω, and Ω satisfies the uniform convexity condition (4.4). Thus, from Theorem 3.4 and (4.4), one obtains

$$\langle F'(x), x - \bar{x} \rangle \geqq 2\gamma \|F'(x)\| \|x - \bar{x}\|^2 \geqq A \|x - \bar{x}\|^2$$

for all $\bar{x} \in T(x)$ and $x \in \Omega$, where

$$A = 2\gamma\varepsilon > 0$$

and $\varepsilon > 0$ is a lower bound for $\|F'(x)\|$. It then follows from (4.6) and (5.14) that (5.13A) and (5.13B) hold with

$$\mu = \frac{1}{(L/A)} = \frac{2\gamma\varepsilon}{L}$$

in place of (5.13C) (this result is essentially a corollary of Theorem 1.8, of [1, p. 131]; the estimate (5.12) is contained in the preceding Theorem 1.7). In general (5.4C) and (5.13C) are rather crude estimates; however, the regularity condition (3.11) is considerably weaker than a uniform convexity assumption (see Note 3.4 and § 6). No results comparable to i) and iii) are established for the algorithm (4.2)–(4.3) in [1].

Note 5.2. For the algorithm (4.9)–(4.10), the inequalities (5.1) and (5.3) in Theorem 5.1 combine to give

$$(5.17) \qquad 0 \leqq r_n \leqq C(L, \theta) \cdot \frac{1}{n}$$

with

$$C(L, \theta) = 2 \max \{r_1, L\theta^2\} \cdot \max \left\{ 1, \frac{D^2}{\theta^2} \right\}$$

where L is any Lipschitz constant for F', i.e. $L \geqq L_0$ (see eq. (1.1)). The least value of C is achieved with $\theta = D^2$ and $L = L_0$, i.e. for fixed $r_1 > 0$, one has

$$C(L, \theta) \geqq C(L_0, D) = 2 \max \{r_1, L_0 D^2\}$$

for all $L \geqq L_0$ and $\theta^2 > 0$. Notice that the corresponding estimate (5.17) equals or surpasses (5.12) in Theorem 5.2 (depending on whether $L = L_0$ or $L > L_0$ in (4.2)). Similarly, for $\theta^2 \geqq r_1/L$, the parameters μ coincide in the estimates (5.4) and (5.13).

Note 5.3. $O(1/n)$ convergence for $F(x_n)-\inf_\Omega F$ can be achieved even with open loop conditional gradient methods [2]. Theorems 5.1 and 5.2 show that closed loop methods can improve considerably on this when F has a strongly nonsingular minimizer; however for problems with singular solutions, $O(1/n)$ convergence is typically the *best* one can achieve with any conditional gradient method (e.g. see [17]). Moreover, if F has a singular minimizer ξ, the sequence $\{x_n\}$ may not converge, even weakly, to *any* minimizer of F unless ξ is unique and Ω is weakly compact [2]; and $\{\bar{x}_n\}$ is typically divergent in any case.

Note 5.4. For convex F, the quantity $\rho_n=\langle F'(x_n), x_n-\bar{x}_n\rangle$ always provides an a posteriori upper bound on $r_n=F(x_n)-\inf_\Omega F$ (eq. (4.6)). It is known that $\rho_n\to 0$ for the algorithm (4.2)–(4.3) under the conditions imposed on F and Ω in Theorems 5.1 and 5.2. Under these same conditions, $\rho_n=O(\lambda^{n/2})$ for (4.2)–(4.3) or (4.9)–(4.10) if F has a regular minimizer.

Note 5.5. The essential content of Theorem 5.1 extends readily to a somewhat larger class of quasiconvex functionals F. In general, suppose that

$$F(x)=h(G(x)),\quad \forall x\in X \tag{5.18}$$

where h is a strictly increasing real function with a continuous derivative h'. The chain rule then gives

$$F'(x)=h'(G(x))G'(x)$$

with $h'(G(x))>0$. It follows that F and G have the same minimizers, extremals, strongly nonsingular extremals, regular extremals, and strongly regular extremals in Ω. Moreover, if $T_F(x)=\{\bar{x}\in\Omega|\langle F'(x),\bar{x}\rangle\leqq\langle F'(x), y\rangle, \forall y\in\Omega\}$ and $T_G(x)=\{\bar{x}\in\Omega|\langle G'(x),\bar{x}\rangle\leqq\langle G'(x), y\rangle, \forall y\in\Omega\}$, then $T_F(x)=T_G(x), \forall x\in\Omega$, and the conditional gradient sequences $\{x_n\}$ generated by the algorithm (4.9)–(4.10) for F are indistinguishable from those generated for G. Consequently, if (4.9)–(4.10) is applied to F, and if G is convex and G' Lipschitz continuous, then according to Theorem 5.1, $G(x_n)\to\inf_\Omega G$, which in turn yields $F(x_n)\to\inf_\Omega F$ in view of (5.18). Furthermore, since h' is continuous and $G(\Omega)$ is a bounded set in $\mathbb{R}^1$ under the conditions of Theorem 5.1, it follows that h is Lipschitz continuous and this means that the sequence $F(x_n)-\inf_\Omega F$ is $O(1/n)$, $o(1/n)$, geometrically convergent, or terminates in finitely many steps, according to whether F has a singular, strongly nonsingular, regular, or strongly regular extremal.

6. Some connections with optimal control theory. The results in §§ 3–5 clarify the meaning and significance of the Haynes–Hermes notion of singularity [5] and the related definitions of Kelley et al. [6], [7] and Dunn [8] in optimal control theory. For example, consider the class of Mayer optimal control problems prescribed by

$$\mathring{x}=f(x, t, u(t));\qquad x(0)=0,\quad 0\leqq t\leqq 1, \tag{6.1}$$

$$\Omega=\{u(\cdot)\in\mathscr{L}^1_{[0,1]}||u(t)|\leqq 1, \text{a.e. in } [0, 1]\} \tag{6.2}$$

$$F(u(\cdot))=P(x(1)) \tag{6.3}$$

with $x\in\mathbb{R}^n$, P: $\mathbb{R}^n\to\mathbb{R}^1$, and f: $\mathbb{R}^n\times\mathbb{R}^1\times\mathbb{R}^1\to\mathbb{R}^n$.[1] Suppose that the initial value problem (6.1) corresponding to each $u(\cdot)\in\Omega$ has a unique solution $x(\cdot)$ on [0, 1]. Then

[1] The restriction of $u(t)$ to $\mathbb{R}^1$ is a convenience; all of the conclusions reached in this section can be extended to $\Omega=\{u(\cdot)\in\mathscr{L}^1([0, 1], \mathbb{R}^m)||u_i(t)|\leqq 1$ a.e. in $[0, 1]$, $i=1,\cdots, m\}$.

(6.1) and (6.3) define a functional F on Ω, and the problem is to minimize F over Ω. If f and P are sufficiently smooth, F can be extended to some neighborhood of Ω in $\mathscr{L}^1_{[0,1]}$ and will have a Fréchet derivative F' on Ω. More specifically,

$$\langle F'(u(\cdot)), v(\cdot)\rangle = \int_0^1 s_{u(\cdot)}(t)v(t)\,dt, \quad \forall v(\cdot)\in\mathscr{L}^1_{[0,1]}$$

with $s_{u(\cdot)}(\cdot)\in\mathscr{L}^\infty_{[0,1]}$ and

$$s_{u(\cdot)}(t)=\psi^T(t)f'_u(x(t),t,u(t)) \quad \text{a.e. in } [0,1]$$

where $x(\cdot)$ is the unique solution of (6.1) corresponding to $u(\cdot)$, $\psi(\cdot)$ is the unique solution of the associated final value problem,

$$\mathring{\psi} = -[f'_x(x(t),t,u(t))]^T\psi; \qquad \psi(1)=P'(x(1))$$

and superscript T denotes the transpose operation. In this setting, the first part of Theorem 2.1 is equivalent to the Pontryagin maximum principle [19] when f is linear in u; more generally, when f is not linear in u, the result in question is a corollary of the maximum principle. In any event, an optimal control $\xi(\cdot)$ must be an extremal in the sense that

$$\langle F'(\xi(\cdot)), v(\cdot)-\xi(\cdot)\rangle = \int_0^1 s_{\xi(\cdot)}(t)(v(t)-\xi(t))\,dt \geqq 0$$

for all $v(\cdot)\in\Omega$. The function $s_{\xi(\cdot)}(\cdot)$ is called a "switching function" because $\xi(\cdot)$ typically has jump discontinuities at isolated zeros of $s_{\xi(\cdot)}(\cdot)$.

Let $\theta=\{t\in[0,1] | s_{\xi(\cdot)}(t)=0\}$. In the classification schemes of [5], [6], and [7], an extremal $\xi(\cdot)$ is said to be singular if meas $\theta>0$. It turns out that $\xi(\cdot)$ is then also singular in the sense of Definition 3.1. In the references just cited, all extremals with meas $\theta=0$ are considered to be nonsingular and no further delineations are made within this class. However in [8], reasons are given for distinguishing between extremals with $\theta\neq\varnothing$, meas $\theta=0$, and extremals for which $\theta=\varnothing$. Additional support for this distinction is provided by the results which follow.

THEOREM 6.1. *Let $\Omega\subset\mathscr{L}^1_{[0,1]}$ be the set* (6.2) *and let F be defined on some neighborhood of Ω. Suppose that F has a Fréchet derivative at $u(\cdot)\in\Omega$, with*

$$\langle F'(u(\cdot)), v(\cdot)\rangle = \int_0^1 s(t)v(t)\,dt, \quad \forall v(\cdot)\in\mathscr{L}^1_{[0,1]}$$

and $s(\cdot)\in\mathscr{L}^\infty_{[0,1]}$.[2] *Then $u(\cdot)$ is respectively singular or strongly nonsingular, according to whether* meas $\theta>0$ *or* meas $\theta=0$, *where*

$$\theta=\{t\in[0,1] | s(t)=0\}.$$

Proof. The set $T(u(\cdot))$ consists of all the single-valued branches of $-\text{sgn}(s(\cdot))$ in Ω, where

$$\text{sgn}\, s=\begin{cases}\{1\}, & \text{if } s>0\\ [-1,1], & \text{if } s=0\\ \{-1\}, & \text{if } s<0.\end{cases}$$

[2] F is a general functional here; i.e., F is not necessarily specified by (6.1)–(6.3). Also, the dependence of the representor $s(\cdot)$ on $u(\cdot)$ has been suppressed in order to simplify the notation.

More precisely,

$$\inf_{v(\cdot)\in\Omega} \langle F'(u(\cdot)), v(\cdot)\rangle = -\int_0^1 s(t)\,\mathrm{sgn}\, s(t)\,dt$$
$$= -\int_0^1 |s(t)|\,dt = \langle F'(u(\cdot)), \bar{u}(\cdot)\rangle$$

if and only if

$$\bar{u}(t) \in -\mathrm{sgn}\, s(t), \quad \text{a.e. in } [0, 1]. \tag{6.4}$$

If meas $\theta > 0$, there are infinitely many such $\bar{u}(\cdot)$, pairwise distinct on a set with positive measure. Thus $T(u(\cdot))$ contains more than one member (i.e., more than one a.e. equivalence class) and so $u(\cdot)$ is a singular point in Ω, according to Definition 3.1. Conversely, if meas $\theta = 0$, then any pair of functions $\bar{u}(\cdot)$ satisfying (6.4) must coincide a.e., in which case $T(u(\cdot))$ contains a single element and $u(\cdot)$ is therefore nonsingular. To see that $u(\cdot)$ is in fact *strongly* nonsingular, suppose that $\langle F'(u(\cdot)), v_n(\cdot) - \bar{u}(\cdot)\rangle \to 0$ for some sequence $\{v_n(\cdot)\} \subset \Omega$, with $\bar{u}(\cdot) =$ the unique member of $T(u(\cdot))$. Then

$$\lim_{n\to\infty} \int_0^1 s(t)(v_n(t) + \mathrm{sgn}\, s(t))\,dt = \lim_{n\to\infty} \int_0^1 |s(t)|\,|v_n(t) + \mathrm{sgn}\, s(t)| = 0$$

or equivalently, $\|\rho_n(\cdot)\|_1 \to 0$, with $\rho_n(t) = |s(t)|\,|v_n(t) + \mathrm{sgn}\, s(t)|$ for $0 \leqq t \leqq 1$. It follows that $\rho_n(\cdot)$ converges in measure to 0 and consequently has a subsequence $\rho_{n_k}(\cdot)$ which converges to 0 pointwise a.e. in [0, 1] [20]. Since meas $\theta = 0$, the bounded sequence $v_{n_k}(\cdot) + \mathrm{sgn}\, s(\cdot)$ must therefore converge to 0 pointwise a.e. in [0, 1], and consequently $\|v_{n_k}(\cdot) + \mathrm{sgn}\, s(\cdot)\|_1 = \|v_{n_k}(\cdot) - \bar{u}(\cdot)\|_1 \to 0$, by the dominated convergence theorem. Thus

$$\langle F'(u(\cdot)), v_n(\cdot) - \bar{u}(\cdot)\rangle \to 0 \quad \Rightarrow \quad \exists \{v_{n_k}(\cdot)\} \ni \|v_{n_k}(\cdot) - \bar{u}(\cdot)\|_1 \to 0,$$

and this means that

$$\sigma > 0 \quad \Rightarrow \quad \inf_{\substack{v(\cdot)\in\Omega \\ \|v(\cdot)-\bar{u}(\cdot)\|_1 \geqq \sigma}} \langle F'(u(\cdot)), v(\cdot) - \bar{u}(\cdot)\rangle > 0. \quad \text{Q.E.D.}$$

When meas $\theta = 0$, $u(\cdot)$ may or may not be regular in $\Omega \subset \mathscr{L}^1_{[0,1]}$, depending upon how the function $s(\cdot)$ behaves near its zeros. It will now be shown that if $|s(\cdot)|$ is bounded below a.e. by a certain type of nonnegative continuous function with finitely many zeros, all of which are "simple" in a certain generalized sense, then $u(\cdot)$ will be regular. Moreover, if $|s(\cdot)|$ is bounded away from 0 a.e. by a positive constant function, then $u(\cdot)$ will be strongly regular. These conclusions, taken together with the convergence theorems in § 5, have practical significance for optimal control problems with "bang-bang" solutions (Note 6.4).

LEMMA 6.1. *Given $N \geqq 1$ points $t_i \in [0, 1]$, $i = 1, \cdots, N$, with*

$$0 \leqq t_1 < t_2 < \cdots < t_N \leqq 1,$$

put

$$\phi_i(t) = |t - t_i|, \qquad t \in [0, 1], \quad 1 \leqq i \leqq N.$$

For $m > 0$ and $\bar{\varepsilon} > 0$, put

$$\phi(t) = \min\{\bar{\varepsilon}, m\phi_1(t), m\phi_2(t), \cdots, m\phi_N(t)\} \tag{6.5}$$

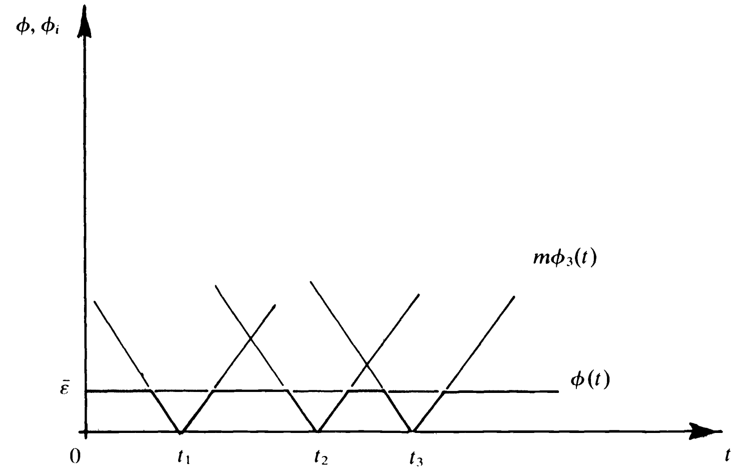

FIG. 6.1

(Fig. 6.1) *and let*

$$\Phi(\varepsilon)=\{t\in[0,1]\mid\phi(t)\leqq\varepsilon\}. \tag{6.6}$$

Then for some sufficiently small $\bar{\varepsilon}>0$,

$$\operatorname{meas}\Phi(\varepsilon)=\frac{C}{m}\varepsilon,\quad \forall\varepsilon\in[0,\bar{\varepsilon}) \tag{6.7}$$

and

$$\int_{\Phi(\varepsilon)}\phi(t)\,dt=\frac{C}{2m}\varepsilon^2,\quad \forall\varepsilon\in[0,\bar{\varepsilon}) \tag{6.8}$$

with

$$C=\sum_{i=1}^{N}C_i$$

and

$$C_i=\begin{cases}2, & \textit{if } t_i\in(0,1)\\ 1, & \textit{if } t_i=0 \textit{ or } 1.\end{cases}$$

Proof. For $\varepsilon\geqq0$ and $1\leqq i\leqq N$, consider the intervals

$$I_i(\varepsilon)=\begin{cases}\left[0,\dfrac{\varepsilon}{m}\right], & \text{if } t_i=0\\[2ex] \left[t_i-\dfrac{\varepsilon}{m},t_i+\dfrac{\varepsilon}{m}\right], & \text{if } t_i\in(0,1)\\[2ex] \left[1-\dfrac{\varepsilon}{m},1\right], & \text{if } t_i=1.\end{cases}$$

Choose $\bar{\varepsilon} > 0$ so that $I_i(\varepsilon) \subset [0, 1], \forall i$, and $i \neq j \Rightarrow I_i(\varepsilon) \cap I_j(\varepsilon) = \varnothing, \forall \varepsilon \in [0, \bar{\varepsilon})$. Then $\Phi(\varepsilon) = \bigcup_{i=1}^{N} I_i(\varepsilon)$ and meas $\Phi(\varepsilon) = \sum_{i=1}^{N}$ meas $I_i(\varepsilon)$, provided $\varepsilon \in [0, \bar{\varepsilon})$. If $t_1 = 0$, then $I_1(\varepsilon) = [0, \varepsilon/m]$ and meas $I_1(\varepsilon) = \varepsilon/m$. If $t_N = 1$, then $I_N(\varepsilon) = [1 - \varepsilon/m, 1]$ and meas $I_N(\varepsilon) = \varepsilon/m$. Otherwise, for all $t_i \in (0, 1)$, meas $I_i(\varepsilon) = 2\varepsilon/m$. This establishes (6.7). Finally, for $\varepsilon \in [0, \bar{\varepsilon})$, $t \in I_i(\varepsilon) \Rightarrow \phi(t) = m\phi_i(t)$, and therefore

$$\int_{\Phi(\varepsilon)} \phi(t)\,dt = m \sum_{i=1}^{N} \int_{I_i(\varepsilon)} |t - t_i|\,dt = \frac{C}{2m}\varepsilon^2. \quad \text{Q.E.D.}$$

THEOREM 6.2. *Let the conditions of Theorem* 6.1 *hold, and suppose that for some* $m > 0$ *and for all sufficiently small* $\bar{\varepsilon} > 0$, *the function* $|s(\cdot)|$ *is bounded below a.e. by the function* $\phi(\cdot)$ *in* (6.5) (Fig. 6.2). *Then* $u(\cdot)$ *is regular in* $\Omega \subset \mathscr{L}^1_{[0,1]}$. *Furthermore, if* $|s(\cdot)|$ *is bounded away from* 0 *a.e., then* $u(\cdot)$ *is strongly regular in* $\Omega \subset \mathscr{L}^1_{[0,1]}$.

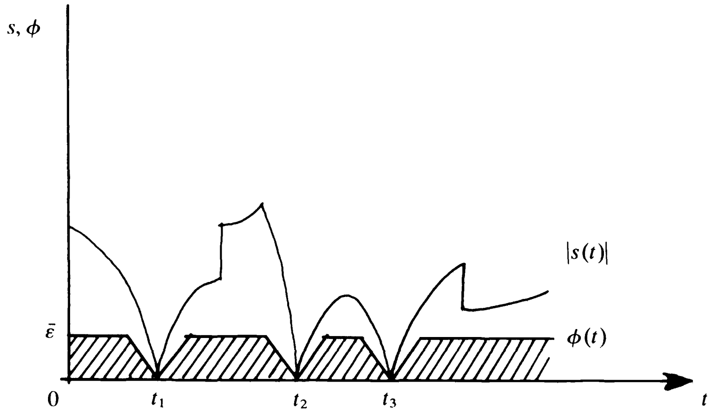

FIG. 6.2

Proof. Since Ω is convex, Lemma 3.2 gives

$$(6.9) \qquad a(\sigma) = \inf_{\substack{v(\cdot) \in \Omega \\ \|v(\cdot) - \bar{u}(\cdot)\|_1 \geqq \sigma}} \langle F'(u(\cdot)), v(\cdot) - \bar{u}(\cdot) \rangle = \inf_{\substack{v(\cdot) \in \Omega \\ \|v(\cdot) - \bar{u}(\cdot)\|_1 = \sigma}} \int_0^1 s(t)(v(t) - \bar{u}(t))\,dt$$

with $\bar{u}(\cdot)$ given by (6.4). If $|s(\cdot)|$ is bounded below a.e. by $\phi(\cdot)$ in (6.5) with $\bar{\varepsilon} > 0$ and $m > 0$, then meas $\theta = 0$, and so $\bar{u}(t) = +1$ or -1 a.e. in $[0, 1]$. Thus, (6.9) can be carried further to

$$a(\sigma) = \inf_{\substack{w(\cdot) \in W \\ \|w(\cdot)\|_1 = \sigma}} \int_0^1 |s(t)| w(t)\,dt$$

with

$$W = \{w(\cdot) \in \mathscr{L}^1_{[0,1]} \,|\, 0 \leqq w(t) \leqq 2, \text{ a.e. in } [0, 1]\}$$

By Lemma 6.1, there is an $\bar{\varepsilon} > 0$ so small that conditions (6.7) and (6.8) hold, while $\phi(\cdot)$ in (6.5) continues to bound $|s(\cdot)|$ from below. For $w(\cdot) \in W$, $\|w(\cdot)\|_1 = \sigma$, $\varepsilon \in$

$[0, \bar{\varepsilon})$, and $\Phi(\varepsilon)$ given by (6.6), one then has

$$\begin{aligned}\int_0^1 |s(t)|w(t)\,dt &\geqq \int_{\Phi(\varepsilon)} \phi(t)w(t)\,dt + \int_{[0,1]-\Phi(\varepsilon)} \phi(t)w(t)\,dt \\ &= \int_{\Phi(\varepsilon)} (\phi(t)-\varepsilon)w(t)\,dt + \int_{[0,1]-\Phi(\varepsilon)} (\phi(t)-\varepsilon)w(t)\,dt + \varepsilon\sigma \\ &\geqq \int_{\Phi(\varepsilon)} (\phi(t)-\varepsilon)w(t)\,dt + \varepsilon\sigma \\ &\geqq 2\int_{\Phi(\varepsilon)} (\phi(t)-\varepsilon)\,dt + \varepsilon\sigma \\ &= 2\int_{\Phi(\varepsilon)} \phi(t)\,dt + \varepsilon(\sigma - 2 \text{ meas } \Phi(\varepsilon)) = \varepsilon\sigma - \frac{C}{m}\varepsilon^2\end{aligned}$$

and therefore

$$a(\sigma) \geqq \varepsilon\sigma - \frac{C}{m}\varepsilon^2, \quad \forall \varepsilon \in [0, \bar{\varepsilon}).$$

For σ in the range $0 \leqq \sigma < \bar{\sigma} = 2C\bar{\varepsilon}/m$, the right side of this expression is maximized over $[0, \bar{\varepsilon})$ at $\varepsilon = m\sigma/(2C)$, where $\varepsilon\sigma - (C/m)\varepsilon^2 = (m/(4C))\sigma^2$. Consequently $0 \leqq \sigma < \bar{\sigma} \Rightarrow a(\sigma) \geqq A\sigma^2$, with $A = m/(4C)$. Furthermore, since Ω is bounded and $a(\sigma)$ is nondecreasing, one actually has $a(\sigma) \geqq \tilde{A}\sigma^2$ for $\sigma \geqq 0$, with $\tilde{A} = A\bar{\sigma}^2/d^2$ and $d = \sup_{v(\cdot)\in\Omega} \|v(\cdot) - \bar{u}(\cdot)\| < \infty$. Thus, $u(\cdot)$ is regular in $\Omega \subset \mathscr{L}^1_{[0,1]}$.

Finally, if $|s(t)| \geqq A$ a.e. in $[0, 1]$ for some $A > 0$, then

$$a(\sigma) = \inf_{\substack{v(\cdot)\in\Omega \\ \|v(\cdot)-\bar{u}(\cdot)\|_1 \geqq \sigma}} \int_0^1 |s(t)|\,|v(t) - \bar{u}(t)| \geqq A\sigma$$

and therefore $u(\cdot)$ is strongly regular in $\Omega \subset \mathscr{L}^1_{[0,1]}$. Q.E.D.

Note 6.1. If $s(\cdot)$ is continuously differentiable and has just finitely many zeros at $t_1, \cdots, t_N$, all of which are simple in the classical sense (i.e., $s(t_i) = 0$ and $(ds/dt)(t_i) \neq 0$) then $|s(\cdot)|$ is bounded below by a $\phi(\cdot)$ in (6.5), with $m < \min_{1\leqq i\leqq N} |(ds/dt)(t_i)|$ and some sufficiently small $\bar{\varepsilon} > 0$. To see what can happen when higher order zeros are present, suppose that $s(\cdot)$ has an isolated zero at $t = 0$ and is increasing near 0. By an argument similar to that used in the proof of Theorem 6.2, one can then show that for sufficiently small $\sigma \geqq 0$

$$a(\sigma) \leqq 2\int_0^{\sigma/2} |s(t)| \leqq \sigma s(\sigma/2).$$

Consequently, if $s(t) = o(t)$ as $t \to 0$, then $a(\sigma) \geqq A\sigma^2$ for $\sigma \geqq 0$ is impossible with any $A > 0$.

Note 6.2. For continuous $s(\cdot)$, $|s(\cdot)|$ is bounded away from 0 a.e. $\Leftrightarrow \theta = \emptyset$.

Note 6.3. The set Ω in (6.2) is not uniformly convex in $\mathscr{L}^1_{[0,1]}$; in fact, $\text{Int}\,\Omega = \emptyset$ in $\mathscr{L}^1_{[0,1]}$. The unit ball in $\mathscr{L}^1_{[0,1]}$ is also not uniformly convex.

Note 6.4. For the set Ω in (6.2) it can be shown that $u(\cdot)$ is strongly nonsingular in Ω relative to the $\mathscr{L}^1$ norm if and only if it is strongly nonsingular in Ω relative to the $\mathscr{L}^p$ norm, with $1 < p < \infty$. Thus, the strong nonsingularity condition in Theorem 6.1 can be utilized in conjunction with the convergence theorems of § 5, under $(\mathscr{L}^p, \mathscr{L}^{p/(1-p)})$ Lipschitz continuity conditions on F'. On the other hand, the regularity conditions in Theorem 6.2 really are tied to $\mathscr{L}^1_{[0,1]}$ and consequently have significance for the convergence theory only when F' is $(\mathscr{L}^1, \mathscr{L}^\infty)$ Lipschitz continuous on Ω.[3] Fortunately, this requirement is satisfied for a nontrivial class of optimal control problems with f linear in u. For instance, in (6.1)–(6.3), let $x = (y_0, y) \in \mathbb{R}^1 \times \mathbb{R}^{n-1}$ and let

$$\dot{y} = A(t)y + B(t)u(t), \qquad \dot{y}_0 = Q(y, t) + B_0(t)u(t),$$

with $A(\cdot)$ and $B(\cdot)$ matrix valued $\mathscr{L}^\infty$ functions, $B_0(\cdot)$ a scalar valued $\mathscr{L}^\infty$ function, $Q(y, \cdot)$ and $Q'_y(y, \cdot)\mathscr{L}^\infty$ functions for each fixed y, and $Q'_y(\cdot, t)$ Lipschitz continuous uniformly in $t \in [0, 1]$. Furthermore, let $P = y_0 + \hat{P}(y)$, with $\hat{P}'_y$ Lipschitz continuous. Then the Fréchet derivative of the corresponding functional F in (6.3) is $(\mathscr{L}^1, \mathscr{L}^\infty)$ Lipschitz continuous. Moreover, if $Q(\cdot, t)$ is convex for $t \in [0, 1]$ and if $\hat{P}$ is convex, then F is convex.

A more thorough working out of the present theory's implications for optimal control problems will be presented elsewhere (also, see [1], [3] and [10]).

REFERENCES

[1] V. F. Demyanov and A. M. Rubinov, *Approximate Methods in Optimization Problems*, American Elsevier, New York, 1970.

[2] J. C. Dunn and S. Harshbarger, *Conditional gradient algorithms with open loop step size rules*, J. Math. Anal. Appl., 62 (1978), pp. 432–444.

[3] E. R. Barnes, *A geometrically convergent algorithm for solving optimal control problems*, this Journal, 10 (1972), pp. 434–443.

[4] E. S. Levitin and B. T. Polyak, *Constrained minimization methods*, U.S.S.R. Computational Math. and Math. Phys., 6 (1966), no. 5, pp. 1–50.

[5] G. Haynes and H. Hermes, *Nonlinear control problems with control appearing linearly*, J. Soc. Indust. Appl. Math. Ser. A: Control, 1 (1963), pp. 85–108.

[6] H. J. Kelley, *A transformation approach to singular sub-arcs in optimal trajectory and control problems*, Ibid., 2 (1964), pp. 234–240.

[7] H. J. Kelley, R. E. Kopp and H. G. Moyer, *Singular extremals*, Optimization Theory and Applications, A Variational Approach, G. Leitmann, ed., Academic Press, New York, 1966.

[8] J. C. Dunn, *On the classificiation of singular and nonsingular extremals for the Pontryagin maximum principle*, J. Math. Anal. Appl., 17 (1967), pp. 1–36.

[9] H. J. Kelley, *Method of gradients, Optimization Techniques*, Academic Press, New York, 1962.

[10] J. C. Dunn, *A simple averaging process for approximating the solutions of certain optimal control problems*, J. Math. Anal. Appl., 48 (1974), pp. 875–894.

[11] J. C. Dunn and V. Kumar, *An averaging technique for solving extremal boundary value problems*, Recent Advances in Engineering Science, vol. 8, Scientific Publishers, Boston, 1975.

[12] V. Kumar, *A control averaging technique for solving a class of singular optimal control problems*, Internat. J. Control, 23 (1967), pp. 361–380.

[13] E. G. Gilbert, *An iterative procedure for computing the minimum of a quadratic form on a convex set*, this Journal, 4 (1966), pp. 61–80.

[14] R. T. Rockafeller, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[15] G. B. Dantzig, J. Folkman and N. Shapiro, *On the continuity of the minimum set of a continuous function*, J. Math. Anal. Appl., 17 (1967), pp. 519–548.

[16] S. M. Robinson, *Bounds for error in the solution set of a perturbed linear program*, Linear Algebra Appl., 6 (1973), pp. 69–81.

[17] M. D. Cannon and C. D. Cullum, *A tight upper bound on the rate of convergence of the Frank–Wolfe algorithm*, this Journal, 6 (1968), pp. 509–516.

[3] E.g., $|s(\cdot)|$ bounded below by ϕ in (6.5) does *not* imply regularity in the $\mathscr{L}^p$ norm for any $p > 1$.

[18] J. C. Dunn, *Iterative construction of fixed points for multi-valued operators of the monotone type*, J. Functional Anal., 27 (1978), pp. 38–50.

[19] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[20] A. N. Kolmogorov and S. V. Fomin, *Elements of the Theory of Functions and Functional Analysis*, vol. 2, Graylock Press, Albany, NY, 1961.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 2, March 1979

DETERMINATION OF THE TRANSITIVITY OF BILINEAR SYSTEMS*

WILLIAM M. BOOTHBY† AND EDWARD N. WILSON†

Abstract. We consider a bilinear control system on $\mathbb{R}_0^n = \mathbb{R}^n - \{0\}$

$$\frac{dx}{dt} = \left(A_0 + \sum_{i=1}^{r} u_i(t)A_i\right)x,$$

where $x \in \mathbb{R}^n$, $A_0, A_1, \cdots, A_r$ are $n \times n$ real matrices, $\mathfrak{g}$ is the Lie algebra they generate, and $u_1(t), \cdots, u_r(t)$ are real valued control functions. Although there exists a standard rank condition in terms of the Lie algebra $\mathfrak{g}$ which Sussman and Jurdjevic have shown to be sufficient to guarantee accessibility, it is primarily of theoretical interest, being essentially impossible to apply to given data. In this paper the authors investigate the possibility of developing algorithms involving only rational computations on the matrices, which would determine whether the rank condition is everywhere satisfied, i.e. whether the bilinear system has the accessibility or, in some instances, controllability property. This is equivalent to determining whether or not the matrix Lie group G generated by $\exp(tA_i)$, $i = 0, 1, \cdots, r$, is or is not transitive on $\mathbb{R}_0^n$. The possible transitive Lie groups have been classified by one of the authors. Using this classification, it is shown that there do exist various sequences of rational operations on the matrices $A_0, A_1, \cdots, A_r$ which enable one in a finite number of steps to decide whether or not the system has the accessibility-transitivity property.

1. Introductory remarks. We will be concerned below with a well-known Lie algebra condition for controllability and accessibility of a bilinear system of differential equations of the form

$$\frac{dx}{dt} = \left(A_0 + \sum_{i=1}^{r} u_i(t)A_i\right)x \tag{1}$$

where $u_1(t), \cdots, u_r(t)$ are admissible (say, piecewise continuous) real-valued control functions, $x(t) = (x_1(t), \cdots, x_n(t))$ is $\mathbb{R}^n$-valued, and $A_0, A_1, \cdots, A_r$ are $n \times n$ real matrices. Such systems have been studied by many people, in particular by Mohler [10] and Brockett [2]. Indeed, Brockett showed that any bilinear system has a realization of form (1).

Before giving details, we establish a few notations and definitions. As usual, the set (and associative algebra) of all real $n \times n$ matrices is denoted by $\mathcal{M}_n(\mathbb{R})$ and the group of all invertible elements in $\mathcal{M}_n(\mathbb{R})$ is denoted by $GL(n, \mathbb{R})$. The Lie algebra of $GL(n, \mathbb{R})$ is denoted by $\mathfrak{gl}(n, \mathbb{R})$ and is identified with the set $\mathcal{M}_n(\mathbb{R})$ equipped with the bracket operation $[A, B] = AB - BA$. For any subalgebra $\mathfrak{g}$ of $\mathfrak{gl}(n, \mathbb{R})$, the *rank of* $\mathfrak{g}$ *at* x (denoted $\operatorname{rank}_x \mathfrak{g}$) is the dimension of the subspace $\{Ax : A \in \mathfrak{g}\} \subseteq \mathbb{R}^n$. We say that $\mathfrak{g}$ is *transitive* if $\operatorname{rank}_x \mathfrak{g} = n$ for all $x \neq 0$. To understand the reason for this terminology, let G be the unique connected Lie subgroup of $GL(n, \mathbb{R})$ with Lie algebra $\mathfrak{g}$. Then $\operatorname{rank}_x \mathfrak{g}$ coincides with the dimension of the orbit Gx of x under the natural action of G on $\mathbb{R}^n$. It follows that $\operatorname{rank}_x \mathfrak{g} = n$ if and only if Gx is an open subset of $\mathbb{R}^n$. Since $\mathbb{R}_0^n = \mathbb{R}^n - \{0\}$ is connected and is the disjoint union of its G-orbits, it is then immediate that $\mathfrak{g}$ is transitive if and only if G is transitive on $\mathbb{R}_0^n$ in the usual sense that $Gx = \mathbb{R}_0^n$ for all $x \neq 0$.

Let $\mathfrak{g}$ be the subalgebra generated by the matrices $A_0, A_1, \cdots, A_r$. It has been shown by Elliott [5] that $\mathfrak{g}$ transitive is a necessary condition for controllability of (1) on $\mathbb{R}_0^n$. By controllability we mean here that any two points of $\mathbb{R}_0^n$ may be joined by a solution curve. In the case of (1), this condition is not sufficient in general. Further detailed results concerning accessibility and rank are given by Sussman and Jurdjevic

* Received by the editors November 30, 1977, and in revised form June 8, 1978.

† Department of Mathematics, Washington University, St. Louis, Missouri 63130.

[12]. Systems of the more restrictive type

$$(2) \qquad \frac{dx}{dt}=\left(\sum_{i=1}^{r} u_i(t)A_i\right)x$$

have been studied by Kucěra [7] and Elliott and Tarn [6] who showed that such a system is controllable on $\mathbb{R}_0^n$ with, say, piecewise constant controls having unrestricted values, if and only if $\mathfrak{g}$ is transitive.

Thus, in studying systems of type (1) or (2), it is of interest to determine whether or not the Lie algebra $\mathfrak{g}$ generated by a given finite set of matrices is transitive. From the definition of rank given above, if $B_1, \cdots, B_N$ is a basis of $\mathfrak{g}$, then $\mathfrak{g}$ is transitive if and only if the $n \times N$ matrix whose columns are $B_1x, \cdots, B_Nx$ has rank n for all $x \neq 0$. This requires that the $\binom{N}{n}$ $n \times n$ minor determinants of the matrix, each a polynomial in $x=(x_1, \cdots, x_n)$ of degree n, have no common zeros (except $x=0$) and is thus a difficult criterion to verify. On the other hand, one of the authors [1] has classified all transitive Lie algebras and thus all controllable systems of type (2). The purpose of this paper is to use this information to give a relatively straightforward procedure to determine by purely rational operations on the generating set when the algebra $\mathfrak{g}$ is transitive. In what follows, such an algorithm is given. It is not, in principle, difficult to carry out except when n is a multiple of 4. In only this case (see § 5) is it possible for the semi-simple part of $\mathfrak{g}$ to be both noncompact and nonsimple; this greatly complicates the situation.

Our insistence that the algorithm be limited to rational computations stems not only from the simplicity of such operations, but the problem of loss of accuracy inherent in operations such as polynomial root extraction to find characteristic values. Indeed, as shown in [1], any of the transitive Lie algebras may be generated by two suitably chosen matrices A_1, A_2. Moreover, the set of pairs of $n \times n$ matrices (A_1, A_2) which generate $\mathfrak{gl}(n, \mathbb{R})$ forms an open dense set in the collection of all pairs of $n \times n$ matrices. Hence any small round off error in operations on given matrices (see also Lobry [9] and Levitt, Sussman [8]) is apt to yield generators for $\mathfrak{gl}(n, \mathbb{R})$. The following argument outlines the proof of this assertion. Using the notion of a free Lie algebra on two generators, it is possible to write down once and for all a finite collection of brackets, $A_1, A_2, [A_1, A_2], [A_1[A_1A_2], [A_2[A_1A_2], \cdots, [A_{i_1}[A_{i_2} \cdots [A_{i_{k-1}}A_{i_k}] \cdots]$ (whose form depends only on n and not at all on A_1, A_2!) which contain a basis of $\mathfrak{g}$. Each element of the collection is a matrix whose entries are polynomials in the entries of A_1, A_2, i.e. polynomials on $\mathcal{M}_n(\mathbb{R}) \times \mathcal{M}_n(\mathbb{R})$. The set of pairs A_1, A_2 for which this collection contains fewer than the maximum, n^2, independent elements, is given by the vanishing of certain polynomials and is thus of lower dimension. But such pairs are precisely those which do not generate $\mathfrak{gl}(n, \mathbb{R})$. Hence $\{(A_1, A_2) | \mathfrak{g}=\mathfrak{gl}(n, \mathbb{R})\}$ is open and dense in $\mathcal{M}_n(\mathbb{R}) \times \mathcal{M}_n(\mathbb{R})$.

In § 2, we give a brief description of the most frequently used rational computations. Section 3 supplies a table of the transitive algebras and some data concerning them, including some corrections and improvements to [1]. In § 4, the algorithm is completely described except for the case $n=4k$ which is dealt with in § 5. The last section points out a few of the simplifications possible in special cases.

2. Rational computations. For easy reference, we give here four basic computations, denoted BC_1 to BC_4, each given by a finite sequence of rational computations and hence itself rational. We also illustrate our use of these computations to uncover information about a Lie subalgebra $\mathfrak{g}$ of $\mathfrak{gl}(n, \mathbb{R})$ defined by given generators $A_1, A_2, \cdots, A_r$.

BC$_1$. *Given a finite sequence $v_1, v_2, \cdots, v_r$ of vectors in $\mathbb{R}^d$ for which the first s elements are independent, select a basis of the linear space they span which contains $v_1, v_2, \cdots, v_s$ as the first s elements.*

This is a standard exercise in linear algebra solved by applying Gaussian row reduction to the $d \times r$ matrix whose columns are the given vectors.

We may apply BC$_1$ to find a basis for $\mathfrak{g}$ in the following way. Let $\mathfrak{g}_1$ be the subspace of $\mathcal{M}_n(\mathbb{R})$ spanned by $A_1, A_2, \cdots, A_r$ and use BC$_1$ to obtain a basis of $\mathfrak{g}_1$. Proceeding inductively, for $k > 1$, define $\mathfrak{g}_k$ to be the subspace of $\mathcal{M}_n(\mathbb{R})$ spanned by $\mathfrak{g}_{k-1}$ together with the collection of all matrices of the form $[X, Y]$ for $X, Y \in \mathfrak{g}_{k-1}$. By induction, we may assume a basis for $\mathfrak{g}_{k-1}$ has been found. Since $\mathfrak{g}_k$ is spanned by these basis elements and their brackets,[1] we may use BC$_1$ to extend the basis of $\mathfrak{g}_{k-1}$ to a basis of $\mathfrak{g}_k$. The computation stops when no extension is necessary for then $\mathfrak{g} = \mathfrak{g}_k = \mathfrak{g}_{k-1}$ and the computed basis $\{B_1, B_2, \cdots, B_N\}$ for $\mathfrak{g}_{k-1}$ is also a basis for $\mathfrak{g}$. In general, assuming $A_1, A_2, \cdots, A_r$ linearly independent, at most $n^2 - r$ steps are necessary; use of theorems on free Lie algebras could further simplify this process somewhat. Note that inversion of the row reduction operations in BC$_1$ at the last step yields a description of each term $[B_i, B_j]$ as a linear combination of $B_1, B_2, \cdots, B_n$ and thereby provides the Lie algebra structure constants of $\mathfrak{g}$ relative to the chosen basis.

BC$_2$. *Given a basis $w_1, w_2, \cdots, w_d$ of $\mathbb{R}^d$ and the matrix entries $K(w_i, w_j)$ of a symmetric bilinear form K on $\mathbb{R}^d$, find a basis $v_1, v_2, \cdots, v_d$ such that $K(v_i, v_j) = 0$ for $i \neq j$.*

BC$_2$ may be performed by a variant of the standard Gram–Schmidt orthogonalization procedure. Thus suppose $k < d$ and $v_1, v_2, \cdots, v_k$ have been found such that $v_1, v_2, \cdots, v_k, w_{k+1}, \cdots, w_d$ is a basis of $\mathbb{R}^d$, $K(v_i, v_j) = 0$ for $i \neq j$ and $K(v_i, v_i) = 0$ if and only if $K(v_i, u) = 0$ for all $u \in \mathbb{R}^d$. Define the linear transformation Q_k on $\mathbb{R}^d$ by $Q_k u = u - \sum (K(u, v_i)/K(v_i, v_i))v_i$ where the sum is taken over all indices $i \leqq k$ for which $K(v_i, v_i) \neq 0$. Trivially $K(Q_k u, v_i) = 0$ for all $i \leqq k$. Put $v_{k+1} = Q_k w_{k+1}$ if either $K(Q_k w_{k+1}, Q_k w_{k+1}) \neq 0$ or $K(Q_k w_{k+1}, w_j) = 0$ for all $j \geqq k+1$. Otherwise select the first index $j > k+1$ such that $K(Q_k w_{k+1}, w_j) \neq 0$ and put $v_{k+1} = Q_k(w_{k+1} + c w_j)$ where c is any scalar for which $K(v_{k+1}, v_{k+1}) \neq 0$.

Recall that the Killing form on a Lie algebra $\mathfrak{h}$ is the symmetric bilinear form $K_\mathfrak{h}$ defined by $K_\mathfrak{h}(X, Y) = \text{trace}\,(\text{ad}\, X\ \text{ad}\, Y)$ where $\text{ad}\, X : \mathfrak{h} \to \mathfrak{h}$ is the linear transformation given by $\text{ad}\, X(Z) = [X, Z]$. Two of the standard results about the Killing form are that $\mathfrak{h}$ is semi-simple (respectively, compact and semi-simple) if and only if $K_\mathfrak{h}$ is nondegenerate (respectively, negative definite). Thus for $\mathfrak{g}$ as above with $B_1, \cdots, B_N$ the basis of $\mathfrak{g}$ found by BC$_1$, we use the BC$_1$ computations to compute the matrix of ad B_i for $i = 1, 2, \cdots, N$, then compute $K_\mathfrak{g}(B_i, B_j)$ for $1 \leqq i \leqq j \leqq N$, and finally use BC$_2$ to obtain a basis $C_1, C_2, \cdots, C_N$ relative to which $K_\mathfrak{g}$ is diagonal. It follows that $\mathfrak{g}$ is semi-simple (respectively, compact and semi-simple) if and only if $K_\mathfrak{g}(C_i, C_i) \neq 0$ (respectively, $K(C_i, C_i) < 0$) for $i = 1, 2, \cdots, N$.

BC$_3$. *Given a system of linear equations, find the general solution by rational operations on the coefficients.*

As usual, BC$_3$ is carried out most efficiently by performing Gaussian row reduction on the matrix of the system. To illustrate the use of BC$_3$, note that we may select any vector $x \in \mathbb{R}^n_0$, find a basis $B_1, B_2, \cdots, B_N$ of $\mathfrak{g}$ by BC$_1$, solve the system $(\sum c_j B_j)v = 0$ to find a basis for $\mathfrak{k}_x = \{X \in \mathfrak{g} : Xv = 0\}$, and thereby compute $\text{rank}_x \mathfrak{g} = N - \dim \mathfrak{k}_x$. From § 1, if $\text{rank}_x \mathfrak{g} < n$, then $\mathfrak{g}$ is not transitive. As another illustration, we can use BC$_3$ to find the *centralizer* $\mathfrak{z}$ of $\mathfrak{g}$, i.e. the collection of all matrices Z satisfying the system

[1] In fact it is easy to see that brackets of the special form mentioned in § 1 suffice (see, e.g. [3]).

$ZB_i - B_iZ = 0, i = 1, \cdots, N$ and thus commuting with every element of $\mathfrak{g}$. The centralizer is an associative subalgebra of $\mathscr{M}_n(\mathbb{R})$. When $\mathfrak{g}$ is transitive, then certainly there are no nontrivial $\mathfrak{g}$-invariant subspaces of $\mathbb{R}^n$ and this implies that every nonzero element of $\mathfrak{z}$ is invertible, i.e. $\mathfrak{z}$ is a division algebra. By a well known result, every finite dimensional division algebra over $\mathbb{R}$ is isomorphic to $\mathbb{R}$, $\mathbb{C}$, or $\mathbb{H}$ (the quaternions). Hence, if $\mathfrak{z}$ is not isomorphic to $\mathbb{R}$, $\mathbb{C}$, or $\mathbb{H}$, then $\mathfrak{g}$ is not transitive. This indicates the usefulness of our last basic computation.

BC$_4$. *Given a basis $\{Z_1, Z_2, \cdots, Z_k\}$ for an associative subalgebra $\mathscr{A}$ of $\mathscr{M}_n(\mathbb{R})$ which contains the identity matrix, determine whether $\mathscr{A}$ is isomorphic to $\mathbb{R}$, $\mathbb{C}$, or $\mathbb{H}$.*

Let E denote the $n \times n$ identity matrix. If $k \neq 1, 2,$ *or* 4, $\mathscr{A}$ cannot be isomorphic to $\mathbb{R}$, $\mathbb{C}$, or $\mathbb{H}$. If $k = 1$, $\mathscr{A} = \mathbb{R}E$, which is isomorphic to $\mathbb{R}$. For $k = 2$ or 4, an easy application of BC$_1$ allows us to convert to a basis $\{E, W_1, \cdots, W_{k-1}\}$ with trace $W_i = 0$ for $1 \leqq i \leqq k-1$. In the case when $k = 2$, it is trvial to check that $\mathscr{A} \approx \mathbb{C}$ if and only if W_1^2 is a negative multiple of E. When $k = 4$, we claim that $\mathscr{A}$ is isomorphic to $\mathbb{H}$ if and only if with suitable numbering of the W's, the symmetric matrix (α_{ij}), $1 \leqq i, j \leqq 2$, defined by $W_iW_j + W_jW_i = 2\alpha_{ij}E$ is real and negative definite. Indeed, given this condition, routine checking shows that

$$x_0 + x_1i + x_2j + x_3k \to x_0E + x_1I + x_2J + x_3K$$

is an isomorphism from $\mathbb{H}$ into $\mathscr{A}$ for $I = (-\alpha_{11})^{-1/2}W_1$, $J = [-\alpha_{11}/(\alpha_{11}\alpha_{22} - \alpha_{12}^2)]^{1/2}(W_2 - \alpha_{11}/\alpha_{11}W_1)$, and $K = IJ$. Conversely, given any isomorphism from $\mathbb{H}$ into $\mathscr{M}_n(\mathbb{R})$, the pure quaternions $x_1i + x_2j + x_3k$ must correspond to linear transformations with zero trace. But for q_1 and q_2 any two independent pure quaternions, the matrix $((q_iq_j + q_jq_i)/2)_{1 \leqq i,j \leqq 2}$ is easily seen to have real entries and to be negative definite. The claim follows.

3. The list of transitive Lie algebras. Recall that a Lie algebra $\mathfrak{g} \subset \mathfrak{gl}(n, \mathbb{R})$ is *transitive* by our definition if the corresponding connected subgroup $G \subset GL(n, \mathbb{R})$ acts transitively on $\mathbb{R}_0^n$. Define two subalgebras $\mathfrak{g}$ and $\tilde{\mathfrak{g}}$ of $\mathfrak{gl}(n, \mathbb{R})$ to be *equivalent* if there exists $A \in GL(n, \mathbb{R})$ such that $\tilde{\mathfrak{g}} = A\mathfrak{g}A^{-1}$. Clearly this defines an equivalence relation with the property that if $\mathfrak{g}$ is transitive, then any algebra equivalent to $\mathfrak{g}$ is also transitive. Hence we need list only one representative from each equivalence class of transitive algebras. It should be noted that although algebras which are equivalent are certainly isomorphic, it may happen that $\mathfrak{g}$ and $\tilde{\mathfrak{g}}$ are isomorphic with $\mathfrak{g}$ transitive and $\tilde{\mathfrak{g}}$ not transitive; for example, $\tilde{\mathfrak{g}}$ may be a faithful representation of $\mathfrak{g}$ on a space $\mathbb{R}^d$ with dimension $d > \dim \mathfrak{g}$. With this caveat, we reproduce the list of [1] together with some corrections, improvements in structural detail, and a more complete description of the matrices involved.

We begin by recalling some of the standard Lie algebra notations. For $m \geqq 1$ and $\mathbb{F} = \mathbb{C}$ (respectively, $\mathbb{H}$) restriction of scalar multiplication to $\mathbb{R}$ turns $\mathbb{F}^m$ into $\mathbb{R}^{2m}$ (respectively, $\mathbb{R}^{4m}$) and, through the usual identification of $m \times m$ matrices over $\mathbb{F}$ with transformations on $\mathbb{F}^m$, turns $m \times m$ complex (respectively, quaternionic) matrices into $2m \times 2m$ (respectively, $4m \times 4m$) real matrices. For $\mathbb{F} = \mathbb{R}$, $\mathbb{C}$, or $\mathbb{H}$, $\mathfrak{sl}(m, \mathbb{F})$ denotes the Lie algebra of all $m \times m$ matrices over F which have zero trace where the real trace is used for $F = \mathbb{R}$ or $\mathbb{H}$ and the complex trace for $\mathbb{F} = \mathbb{C}$. Hence the equivalence class of $\mathfrak{sl}(m, \mathbb{R})$ consists of $\mathfrak{sl}(m, \mathbb{R})$ alone while the equivalence class of $\mathfrak{sl}(m, \mathbb{C})$ consists of all subalgebras $\mathfrak{g}$ of $\mathfrak{gl}(2m, \mathbb{R})$ for which there exists a complex structure I (i.e. an element of $GL(2m, \mathbb{R})$ whose square is $-E$) such that $\mathfrak{g} = \{X \in \mathfrak{gl}(2m, \mathbb{R}) : [X, I] = 0$ and trace $X = 0 =$ trace $IX\}$. The equivalence class of $\mathfrak{sl}(m, \mathbb{H})$ consists of all subalgebras $\mathfrak{g}$ of $\mathfrak{sl}(4m, \mathbb{R})$ for which there exists an $\mathbb{H}$-structure on $\mathbb{R}^{4m}$ such that $\mathfrak{g}$ is the collection of all elements of $\mathfrak{sl}(4m, \mathbb{R})$ commuting with this $\mathbb{H}$-structure. The subalgebra of $\mathfrak{sl}(m, \mathbb{F})$ consisting of those elements X such that $X^* = -X$ is denoted by $\mathfrak{so}(m)$ for $\mathbb{F} = \mathbb{R}$, $\mathfrak{su}(m)$ for $\mathbb{F} = \mathbb{C}$, and $\mathfrak{sp}(m)$ for $\mathbb{F} = \mathbb{H}$. Here $X \to X^*$ may be viewed either as the conjugate transpose operation for $\mathbb{F}$-matrices or the restriction to $\mathbb{F}$-matrices of the transpose operation on real matrices. The associated equivalence classes consist of those subalgebras of $\mathfrak{gl}(n, \mathbb{R})$ ($n = m$, $2m$, or $4m$) maximal relative to the property that their elements commute with some $\mathbb{F}$-structure on $\mathbb{R}^n$, are skew relative to some positive definite $\mathbb{F}$-sesquilinear form on $\mathbb{R}^n \approx \mathbb{F}^m$, and, in the case $\mathbb{F} = \mathbb{C}$, have zero complex trace. For $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$, $\mathfrak{sp}(m, \mathbb{F})$ denotes the subalgebra of $\mathfrak{sl}(2m, \mathbb{F})$ consisting of those elements X such that JXJ is the $\mathbb{F}$-transpose of X where J is the $2m \times 2m$ F-matrix $\begin{bmatrix} 0 & E \\ -E & 0 \end{bmatrix}$; the associated equivalence classes are the subalgebras of $\mathfrak{gl}(n, \mathbb{R})$ ($n = 2m$ or $4m$) maximal relative to the property that their elements commute with some $\mathbb{F}$-structure and are skew relative to some non-singular, $\mathbb{F}$-bilinear, alternating form on $\mathbb{R}^n \approx \mathbb{F}^{2m}$. It is well known that the algebra $\mathfrak{so}(2m+1)$ has a representation on $\mathbb{R}^{2^m}$ for m congruent to 0 or 3 modulo 4. This representation is called the spin representation of $\mathfrak{so}(2m+1)$ and the associated algebra of $2^m \times 2^m$ real matrices is denoted by $\mathfrak{spin}(2m+1)$. Finally the algebra $\mathfrak{g}_2(-14)$ refers to the algebra of 7×7 matrices obtained from the representation on $\mathbb{R}^7$ of the compact simple Lie algebra of type G_2.

Every algebra in Table 1 is of the form $\mathfrak{g} = \mathfrak{g}_0 \oplus \mathfrak{c}$ (direct sum) where $\mathfrak{g}_0$ is semi-simple and $\mathfrak{c}$ is the center of $\mathfrak{g}$. In each case, $\mathfrak{g}_0$ acts irreducibly on the underlying Euclidean space $\mathbb{R}^n$, so the centralizer $\mathfrak{z}_0$ of $\tilde{\mathfrak{g}}_0$ is of the form $\mathbb{R}$, $\mathbb{C}$, or $\mathbb{H}$. Since $\mathfrak{c}$ is abelian and is contained in $\mathfrak{z}_0$, the dimension ε of $\mathfrak{c}$ is $\leqq 2$. The centralizer $\mathfrak{z}$ of $\mathfrak{g}$ is contained in $\mathfrak{z}_0$ and coincides with $\mathfrak{z}_0$ if $\mathfrak{z}_0$ is abelian. For $\mathfrak{z}_0 = \mathbb{H}$, however, $\mathfrak{z} = \mathfrak{z}_0$ if $\mathfrak{c} = \mathbb{R}E$ and otherwise $\mathfrak{z} = \mathbb{R}E$.

TABLE 1*

Transitive matrix algebras.

Type	n	N	Representative	$\mathfrak{z}_0$
I.1	m	$m(m-1)/2+1$	$\mathfrak{so}(m)\oplus\mathbb{R}$	$\mathbb{R}$
I.2	$2m$	$m^2-1+\varepsilon$ $(\varepsilon = 1, 2)$	$\mathfrak{su}(m)\oplus\mathfrak{c}$	$\mathbb{C}$
I.3	$4m$	$2m^2+m+\varepsilon$ $(\varepsilon = 1, 2)$	$\mathfrak{sp}(m)\oplus\mathfrak{c}$	$\mathbb{H}$
I.4	$4m$	$2m^2+m+4$	$\mathfrak{sp}(m)\oplus\mathbb{H}$	$\mathbb{R}$
I.5	8	22	$\mathfrak{spin}(7)\oplus\mathbb{R}$	$\mathbb{R}$
I.6	16	37	$\mathfrak{spin}(9)\oplus\mathbb{R}$	$\mathbb{R}$
I.7	7	15	$\mathfrak{g}_{2(-14)}\oplus\mathbb{R}$	$\mathbb{R}$
II.1	m	$m^2-1+\varepsilon$ $(\varepsilon = 0, 1)$	$\mathfrak{sl}(m, \mathbb{R})\oplus\mathfrak{c}$	$\mathbb{R}$
II.2	$2m$	$2(m^2-1)+\varepsilon$ $(\varepsilon = 0, 1, 2)$	$\mathfrak{sl}(m, \mathbb{C})\oplus\mathfrak{c}$	$\mathbb{C}$
II.3	$4m$	$4m^2-1+\varepsilon$ $(\varepsilon = 0, 1, 2)$	$\mathfrak{sl}(m, \mathbb{H})\oplus\mathfrak{c}$	$\mathbb{H}$
II.4	$2m$	$2m^2+m-\varepsilon$ $(\varepsilon = 0, 1)$	$\mathfrak{sp}(m, \mathbb{R})\oplus\mathfrak{c}$	$\mathbb{R}$
II.5	$4m$	$4m^2+2m+\varepsilon$ $(\varepsilon = 0, 1, 2)$	$\mathfrak{sp}(m, \mathbb{C})\oplus\mathfrak{c}$	$\mathbb{C}$
III	$4m$	$4m^2+2+\varepsilon$ $(\varepsilon = 0, 1)$ ·	$\mathfrak{sl}(m, \mathbb{H})\oplus\mathfrak{su}(2)\oplus\mathfrak{c}$	$\mathbb{R}$

* This table corrects minor errors in [1], to wit replacing $\mathfrak{spin}(5)$ and $\mathfrak{spin}(7)$ by $\mathfrak{spin}(7)$ and $\mathfrak{spin}(9)$, respectively, and changing the designation under III.

For each class, we list the dimension n of the underlying Euclidean space, the dimension N of the algebras in the class, one class representative, and the form of $\mathfrak{z}_0$. Whenever it appears, the parameter m is allowed to be any positive integer. The algebras of type II share the property that $\mathfrak{g}_0$ is a noncompact real or complex simple algebra which, by itself, is transitive. Hence $\mathfrak{g}_0 \oplus \mathfrak{c}$ is transitive if $\mathfrak{c}$ is any abelian subalgebra of $\mathfrak{z}_0$. The algebras of type I share the property that $\mathfrak{g}_0$ is compact and semi-simple (actually simple except for I.1 with $m=3$ and for I.4 where $\mathfrak{g}_0$ has the compact ideals $\mathfrak{sp}(m)$ and $\mathfrak{su}(2) \approx \{\text{pure quaternions}\}$). In each case the corresponding compact group G_0 acts transitively on the unit sphere. Here $\mathfrak{g}$ is transitive if $\mathfrak{c}$ is any abelian subalgebra of $\mathfrak{z}_0$ which is noncompact in the sense that some element of $\mathfrak{c}$ has eigenvalues with nonzero real part. Thus for $\mathfrak{z}_0 = \mathbb{R}E \approx \mathbb{R}$, $\mathfrak{c}$ must be equal to $\mathfrak{z}_0$, while for $\mathfrak{z}_0 \approx \mathbb{C}$ or $\mathbb{H}$, $\mathfrak{c}$ may be either of the form $\mathbb{R}F \approx \mathbb{R}$ for F any element of $\mathfrak{z}_0$ with nonzero real part (i.e. F^2 not a negative multiple of E) or of the form $\mathbb{R}E + \mathbb{R}I \approx \mathbb{C}$ for I a complex structure in $\mathfrak{z}_0$. As will be seen below, algebras of type III are the most difficult to recognize. The representative $\mathfrak{sl}(m, \mathbb{H}) \oplus \mathfrak{su}(2)$ for $\mathfrak{g}_0$ is to be regarded as the matrix subalgebra of $\mathfrak{gl}(4m, \mathbb{R})$ obtained by the natural left action of $\mathfrak{sl}(m, \mathbb{H})$ on $\mathbb{H}^m \approx \mathbb{R}^{4m}$ and the right action of $\mathfrak{su}(2)$ as multiplication by pure quaternions on $\mathbb{H}^m$. As with algebras of type II, $\mathfrak{g}_0$ is transitive, so the center $\mathfrak{c}$ may be either $\{0\}$ or $\mathbb{R}E = \mathfrak{z}_0$.

4. Algorithm for transitivity. Let $A_1, A_2, \cdots, A_r$ be given $n \times n$ real matrices generating a subalgebra $\mathfrak{g} \subset \mathfrak{gl}(n, \mathbb{R})$. We begin by attempting to match $\mathfrak{g}$ with an algebra $\tilde{\mathfrak{g}}$ from Table 1. Before doing anything else, a basis $B_1, B_2, \cdots, B_N$ for $\mathfrak{g}$ must be computed using BC$_1$. This determines the *dimension index* of $\mathfrak{g}$, i.e. the pair (N, n) where N is the dimension of $\mathfrak{g}$. Table 1 is then consulted to see whether there exists a transitive algebra $\tilde{\mathfrak{g}}$ with the same dimension index as $\mathfrak{g}$. If not, $\mathfrak{g}$ is certainly not transitive. If such an algebra $\tilde{\mathfrak{g}}$ exists, proceed to Step 1, 2 or 3 below depending on whether $\tilde{\mathfrak{g}}$ is of type I, II, or III. In the very rare situations where there is more than one such $\tilde{\mathfrak{g}}$, go first to Step 2, then to Step 1 if necessary, and, in final desperation, to Step 3. The computations in each step should be performed in the order indicated with the understanding that the algorithm terminates whenever a definite answer regarding transitivity is obtained. For the reader's convenience, along with each step, we indicate the relevance of each calculation and give a brief justification for the conclusions given.

Step 1. (i) *Select any nonzero vector v in $\mathbb{R}^n$ and use* BC$_1$ *to choose from $B_1 v, \cdots, B_N v$ a basis for the space $\mathfrak{k}_v = \{X \in \mathfrak{g} : Xv = 0\}$. If $N - \dim \mathfrak{k}_v \neq n$, $\mathfrak{g}$ is not transitive.*

(ii) *Use* BC$_3$ *and* BC$_4$ *to compute the centralizer $\mathfrak{z}$ of $\mathfrak{g}$ and determine whether $\mathfrak{z} \approx \mathbb{R}$, $\mathbb{C}$, or $\mathbb{H}$. If not, then $\mathfrak{g}$ is not transitive.*

(iii) *Using* BC$_1$ *and* BC$_2$, *obtain a basis $B_1', \cdots, B_N'$ for $\mathfrak{g}$ which diagonalizes the Killing form $K_{\mathfrak{g}}$. Let $a_i = K_{\mathfrak{g}}(B_i, B_i)$, ε the number of indices for which $a_i = 0$, and reorder if necessary so that $a_i \neq 0$ for $i > \varepsilon$. Then $\mathfrak{g}$ is not transitive if $\varepsilon > 2$. Also, $\mathfrak{g}$ is not transitive of type I if there exists $i > \varepsilon$ with $a_i > 0$. Finally, if $\varepsilon \leqq 2$ with $a_i < 0$ for all $i > \varepsilon$, $\mathfrak{g}$ is transitive of type* I.

In this step, we are trying to determine whether our $\mathfrak{g}$ is transitive and has the Lie algebra structure $\mathfrak{g} = \mathfrak{g}_0 \oplus \mathfrak{c}$ with $\mathfrak{g}_0$ compact semi-simple and $\mathfrak{c}$ abelian of dimension 1 or 2. It follows from the discussion of the Killing form under BC$_3$ and standard Lie algebra arguments that $\mathfrak{g}$ has the indicated form $\mathfrak{g}_0 \oplus \mathfrak{c}$ if and only if, in the notation given in (iii), $\varepsilon = 1$ or 2 and $a_i < 0$ for $i > \varepsilon$. Before tackling the fairly lengthy calculations in (iii), it is advisable to determine first whether the action on $\mathbb{R}_0^n$ of the corresponding group G has an open orbit (test (i)) and whether the centralizer of g is a division algebra (test (ii)); we know from the discussion in § 2 that these are necessary conditions for transitivity. Now suppose $\mathfrak{g}$ passes all three tests. Because G has an open orbit, $C = e^{\mathfrak{c}}$ cannot be compact. Since $\mathfrak{c}$ is contained in the devision algebra $\mathfrak{z}$ and $G = G_0 C$ for G_0 the subgroup of G with Lie algebra $\mathfrak{g}_0$, it follows easily that G_0 has an open orbit on the sphere of radius $|v|$. But compactness of G_0 implies that all G_0 orbits are closed so by connectivity, G_0 acts transitively on spheres and hence G is transitive on R_0^n.

An example. To illustrate this procedure, suppose $r=2$, $n=3$, and, for some number ε,

$$A_1=\begin{bmatrix}0 & 5 & 0\\ -13 & 16 & 0\\ 0 & 0 & 8\end{bmatrix},\qquad A_2=\begin{bmatrix}0 & 0 & 5\\ 0 & 0 & 8\\ \varepsilon & 0 & 0\end{bmatrix}.$$

Easy calculations show that

$$[A_1, A_2]=\begin{bmatrix}0 & 0 & 1\\ 0 & 0 & -1\\ 8\varepsilon & -5\varepsilon & 0\end{bmatrix}$$

with $[A_1,[A_1,A_2]]=-A_2$ and $[A_2,[A_1,A_2]]=-5\varepsilon A_1+40\varepsilon E$. For $\varepsilon=0$, $\mathfrak{g}$ is 3-dimensional and we can conclude that $\mathfrak{g}$ is not transitive either by observing that no algebra in type I has the dimension index (3, 3) or by using Step 1 (i) with $v=\begin{bmatrix}1\\0\\0\end{bmatrix}$ to obtain $\dim k_v=2\neq 0=3-3$. For $\varepsilon\neq 0$, $\mathfrak{g}$ is 4-dimensional and Table 1 tells us that the only transitive algebras with dimension index (4, 3) are those of type I.1. A convenient basis $\{B_1, B_2, B_3, B_4\}$ for $\mathfrak{g}$ is given by $B_1=E$, $B_2=A_1-8E$, $B_3=A_2$, and $B_4=[A_1, A_2]$. For a routine choice of v (e.g. a basis vector in $\mathbb{R}^3$), one sees at a glance that $\dim k_v=1$ so (i) is inconclusive. A few seconds work is enough to see that the centralizer $\mathfrak{z}$ of $\mathfrak{g}$ is $\mathbb{R}E$ so (ii) is inconclusive. Proceeding to (iii) and noting that $[B_1, B_j]=0$ for $j=2$, 3, 4, $[B_2, B_3]=B_4$, $[B_2, B_4]=-B_3$, $[B_3, B_4]=-5\varepsilon B_2$, one sees easily that the Killing form $K_{\mathfrak{g}}$ is diagonal relative to the given basis with the diagonal entries being $0, -2, 10\varepsilon, 10\varepsilon$. From (iii), we conclude that $\mathfrak{g}$ is transitive of type I.1 if and only if $\varepsilon<0$.

In the following, we use the familiar abbreviation iff for "if and only if". Recall that $\tilde{\mathfrak{g}}$ denotes an algebra of Table 1 whose dimension index is the same as that of $\mathfrak{g}$—if there is none, $\mathfrak{g}$ cannot be transitive.

Step 2. (i) *If* $\tilde{\mathfrak{g}}$ *is of type* II.1, $\mathfrak{g}$ *is equivalent to* $\tilde{\mathfrak{g}}$ *iff either* $(N, n)\neq(3, 2)$ *or trace* $B_i=0$, $i=1, 2, \cdots, N$.

(ii) *If* $\tilde{\mathfrak{g}}$ *is of type* II.2, *compute the centralizer* $\mathfrak{z}$ *of* $\mathfrak{g}$ *using* BC$_3$ *and* BC$_4$. *Then* $\mathfrak{g}$ *is equivalent to* $\tilde{\mathfrak{g}}$ *iff* $\mathfrak{z}\approx\mathbb{C}$ *and either* $(N, n)\neq(6, 4)$ *or trace* $B_i=0$, $i=1, 2, \cdots, N$.

(iii) *If* $\tilde{\mathfrak{g}}$ *is of type* II.3 *and if either* $\varepsilon=0$ *or if* $\varepsilon=1$ *and it is determined via* BC$_3$ *that* $E\in\mathfrak{g}$, *compute the centralizer of* $\mathfrak{g}$. *Then* $\mathfrak{g}$ *is equivalent to* $\tilde{\mathfrak{g}}$ *if* $\mathfrak{z}\approx\mathbb{H}$. *In all other cases when* $\tilde{\mathfrak{g}}$ *is of type* II.3, *use the structure constants of* $\mathfrak{g}$ *and* BC$_1$ *to select a maximal independent set* $C_1, C_2, \cdots, C_M$ *from the collection of elements* $[B_i, B_j]$, $1\leqq i<j\leqq N$. *If* $M\neq N-\varepsilon$, $\mathfrak{g}$ *is not equivalent to* $\tilde{\mathfrak{g}}$; *if* $M=N-\varepsilon$, $\mathfrak{g}$ *is equivalent to* $\tilde{\mathfrak{g}}$ *iff* BC$_3$ *and* BC$_4$ *show that the algebra* $\mathfrak{z}_0$ *of elements commuting with* $C_1, C_2, \cdots, C_M$ *is isomorphic to* $\mathbb{H}$.

(iv) *If* $\tilde{\mathfrak{g}}$ *is of type* II.4, *compute using* BC_3 *the set* $\mathfrak{s}$ *of solutions to the system* $J^t+J=0$, $JA_i+A_i^tJ=0$, $i=1, 2, \cdots, r$. *If* $\mathfrak{s}=\{0\}$, $\mathfrak{g}$ *is not equivalent to* $\tilde{\mathfrak{g}}$. *If* $\mathfrak{s}\neq\{0\}$, *select any nonzero element* J *in* $\mathfrak{s}$ *and compute* $\det J$; $\mathfrak{g}$ *is equivalent to* $\tilde{\mathfrak{g}}$ *iff* $\det J\neq 0$.

(v) *If* $\tilde{\mathfrak{g}}$ *is of type* II.5, *first compute the centralizer* $\mathfrak{z}$ *of* $\mathfrak{g}$; *if* $z\neq\mathbb{R}E+\mathbb{R}I\approx\mathbb{C}$, $\mathfrak{g}$ *is not equivalent to* $\tilde{\mathfrak{g}}$. *If* $\mathfrak{z}\approx\mathbb{C}$, *compute by* BC$_3$ *the set* $\mathfrak{s}$ *of solutions to the system* $J^tI=IJ$, $J^t+J=0$, $JA_i+A_i^tJ=0$, $i=1, 2, \cdots, r$. *If* $\mathfrak{s}=\{0\}$, $\mathfrak{g}$ *is not equivalent to* $\tilde{\mathfrak{g}}$. *Conversely,* $\mathfrak{g}$ *is equivalent to* $\tilde{\mathfrak{g}}$ *iff any nonzero element in* $\mathfrak{s}$ *is nonsingular.*

The simple dimension comparisons in (i)–(iii) rest on the observation that for $\mathbb{F}=\mathbb{R}$, $\mathbb{C}$, or $\mathbb{H}$ and $m>2$, any subalgebra $\mathfrak{g}$ of $\mathfrak{gl}(n, \mathbb{F})$ whose dimension is at least as large as that of $\mathfrak{sl}(m, \mathbb{F})$ must contain $\mathfrak{sl}(m, \mathbb{F})$. For $m=2$ and $F=\mathbb{R}$ or $\mathbb{C}$, it is necessary to check the zero trace condition in order to rule out the possibility that $\mathfrak{g}$ is isomorphic to the collection of upper triangular matrices. In (iii), the elements $C_1, \cdots, C_M$ are a basis for the derived subalgebra $\mathfrak{g}_0$ of $\mathfrak{g}$; the test given checks to see if $\mathfrak{g}_0$ is equivalent to $\mathfrak{sl}(m, \mathbb{H})$. For $\tilde{\mathfrak{g}}$ of type II.4 or II.5, a dimension comparison is not enough; we must also check to see that $\mathfrak{g}$ leaves invariant a nondegenerate alternating bilinear form. In (iv), this is trivially done by finding a nonsingular element in $\mathfrak{s}$. In (v), note that a real matrix J defines a complex $\mathfrak{g}$-invariant alternating form α by $\alpha(x, y) = \sum J_{ij}x_iy_j - \sqrt{-1}\sum J_{ij}(Ix)_iy_j$ precisely when $J \in \mathfrak{s}$.

5. The difficult case. We emphasize that Step 3 below is to be performed only under the following rare circumstances; $n = 4m$, $N = \dim \mathfrak{g}$ is either $N_0 = 4m^2+2$ or N_0+1, and Steps 1 and 2 indicate that $\mathfrak{g}$ is not of type I or II. Let $\mathfrak{g}_0$ be the subalgebra of $\mathfrak{g}$ consisting of those elements in $\mathfrak{g}$ with zero trace. From Table 1, $\mathfrak{g}$ is transitive precisely when $\mathfrak{g}_0$ is equivalent to $\tilde{\mathfrak{g}}_0 = \mathfrak{sl}(m, \mathbb{H}) + \mathfrak{su}(2)$ and then $\mathfrak{g} = \mathfrak{g}_0$ if $N = N_0$, $\mathfrak{g} = \mathfrak{g}_0 \oplus \mathbb{R}E$ if $N = N_0 + 1$. This explains the purpose for the simple dimension checks in test (i) below. Test (ii) is the familiar centralizer check while test (iii) checks to see if $\mathfrak{g}_0$ is not only semi-simple but has the same signature (or Cartan index) for its Killing form as does $\tilde{\mathfrak{g}}_0$. These tests have the same order of computational difficulty as those in Steps 1 and 2 and in most cases will allow us to conclude that $\mathfrak{g}$ is not transitive. However, if $\mathfrak{g}$ passes these tests and we have to move on to (iv)-(vi), the computational difficulty escalates considerably since we have to start dealing with the $N_0 \times N_0$ matrices ad X for $X \in \mathfrak{g}_0$. For example, if $m = 6$, step (iv) asks for the solution of a system of $(146)^3$ scalar equations in $(146)^2$ unknowns! We have included these remaining steps primarily for the sake of mathematical completeness. Test (iv) rests on the observation that the adjoint representation of any semi-simple algebra decomposes into the sum of irreducible, inequivalent representations corresponding to the algebra's simple ideals. It follows that the centralizer $\mathcal{A}$ of the adjoint representation is the direct sum of subalgebras corresponding to these simple ideals with each subalgebra isomorphic to $\mathbb{R}$ or $\mathbb{C}$ depending on whether the ideal is noncomplex or complex. Therefore, if our $\mathfrak{g}$ passes test (iv), we know that $\mathfrak{g}$ is either complex simple or has two noncomplex simple ideals. Successful passage of test (v) means that $\mathfrak{g} = \mathfrak{g}_1 \oplus \mathfrak{g}_2$ with $\mathfrak{g}_1$ a simple noncomplex ideal of dimension $N_0 - 3$ and $\mathfrak{g}_2$ a three-dimensional simple noncomplex ideal. Then $\mathfrak{g}_2$ is isomorphic to either $\mathfrak{sl}(2, \mathbb{R})$ or $\mathfrak{su}(2)$ and $\mathfrak{g}_1$ is a real form of a complex simple algebra with dimension $N_0 - 3$. Finally, in part (vi), we compute the rank r of $\mathfrak{g}_0$, i.e. the dimension of any Cartan subalgebra of $\mathfrak{g}_0$. If $r = 2m$, the classification of complex simple algebras forces $\mathfrak{g}_1$ to be a real form of $\mathfrak{sl}(2m, \mathbb{C})$ so $\mathfrak{g}_1$ is isomorphic to $\mathfrak{sl}(2m, \mathbb{R})$, $\mathfrak{sl}(m, \mathbb{H})$, or one of the pseudo-unitary algebras $\mathfrak{su}(p, 2m-p)$, $0 \leqq p \leqq m$. Out of the possible candidates for $\mathfrak{g}_0$, only $\mathfrak{sl}(m, \mathbb{H}) \oplus \mathfrak{su}(2)$, $\mathfrak{su}(p, 2m-p) \oplus \mathfrak{su}(2)$ for $p = m - \sqrt{(m-1)/2}$, and $\mathfrak{su}(p, 2m-p) \oplus \mathfrak{sl}(2, \mathbb{R})$ for $p = m - \sqrt{m/2}$ are compatible with the Cartan index computed in (iii). Since the last two do not have a faithful representation on $\mathbb{R}^{4m}$, $\mathfrak{g}_0$ must be equivalent to $\mathfrak{sl}(m, \mathbb{H}) \oplus \mathfrak{su}(2)$.

Step 3. (i) $\mathfrak{g}$ *is not transitive if either* $N = N_0$ *and* $\mathfrak{g} \neq \mathfrak{g}_0$ (*i.e. trace* $B_j \neq 0$ *for some* j) *or if* $N = N_0 + 1$ *and* $E \notin \mathfrak{g}$ (*determined via* BC$_3$). *If* $N = N_0 + 1$ *and* $E \in \mathfrak{g}$, *use* BC$_1$ *to select a basis* $B'_1, B'_2, \cdots, B'_{N_0}$ *for* $\mathfrak{g}_0$ *from the set of elements* $B_j - 1/(4m)(\text{trace } B_j)E$, $j = 1, 2, \cdots, N$.

(ii) *Compute the centralizer* $\mathfrak{z}$ *of* $\mathfrak{g}$ *using* BC$_3$. $\mathfrak{g}$ *is not transitive if* $\dim \mathfrak{z} > 1$.

(iii) *Starting from the known structure constants for* $\mathfrak{g}$ *and basis for* $\mathfrak{g}_0$ *found in* (i), *use* BC$_2$ *to obtain a new basis* $C_1, C_2, \cdots, C_{N_0}$ *such that* $K_{\mathfrak{g}_0}(C_i, C_j) = 0$ *for* $i \neq j$. *Then* $\mathfrak{g}$ *is not transitive if* $K_{\mathfrak{g}_0}(C_i, C_i) = 0$ *for some* i *or if the number of indices for which* $K_{\mathfrak{g}_0}(C_i, C_i) < 0$ *is unequal to* $2m^2 + m + 3$.

(iv) *Use the structure constants for* $\mathfrak{g}$ *and* BC$_3$ *to compute a basis for* $\mathcal{A} = \{A: \mathfrak{g}_0 \to \mathfrak{g}_0 | A \text{ ad } C_i - \text{ad } C_i\, A = 0,\ i = 1, 2, \cdots, N_0\}$. *If* $\dim \mathcal{A} \neq 2$, $\mathfrak{g}$ *is not transitive.*

(v) *Take any element* $F \in \mathcal{A}$ *which is not a scalar multiple of the identity and use* BC_2 *to compute a basis for* $B = \{B: \mathfrak{g}_0 \to \mathfrak{g}_0 | BF = FB\}$. *If* $\dim B \neq (N_0 - 3)^2 + 9$, $\mathfrak{g}$ *is not transitive.*

(vi) *For* $C_1, C_2, \cdots, C_{N_0}$ *as in* 5(iii), *use the structure constants of* $\mathfrak{g}_0$ *relative to these basis elements to compute the polynomial* $\det (\sum_{i=1}^{N_0} x_i \operatorname{ad} C_i - \lambda E)$ *in the indeterminants* $(x_1, x_2, \cdots, X_{N_0}, \lambda) \in \mathbb{R}^{N_0+1}$. *Express this polynomial in the form* $\sum_{k=1}^{N_0} \lambda^k \cdot f_k(x_1, x_2, \cdots, x_{N_0})$ *and find the largest integer r such that all of the polynomials* f_k *vanish identically for* $k < r$ *when written as linear combination of monomials* $x_1^{J_1} x_2^{J_2} \cdots x_{N_0}^{J_{N_0}}$ ($f_k \equiv 0$ *precisely when all coefficients are zero*). *Then* $\mathfrak{g}$ *is transitive if and only if* $r = 2m$.

6. Comments on the algorithm and easy cases. When n is odd, our algorithm is easy to apply since Table 1 yields only the transitive algebras of type I.1, II.1 and I.5 ($n = 7$) as candidates for $\mathfrak{g}$.

Another nice situation occurs when $n \leqq 8$. Since $\mathfrak{g}$ transitive implies $\mathfrak{g}$ irreducible, we can take advantage of a tabulation by Cartan [4] of all irreducible $n \times n$ matrix algebras for $n \leqq 8$. Now suppose $\mathfrak{g}$ is a semi-simple $n \times n$ matrix algebra acting irreducibly on R^n. If $\mathfrak{g}$ is compact (resp. noncompact) and $\dim \mathfrak{g}$ is the same as one of the list of compact (resp. noncompact) semi-simple irreducible algebras $\tilde{\mathfrak{g}}$ in Cartan's list, then for $n \leqq 7$ we know that $\mathfrak{g}$ is equivalent to $\tilde{\mathfrak{g}}$ and hence transitive if and only if $\tilde{\mathfrak{g}}$ is transitive—which is easy to check. Even in the case $n = 8$ there is only one exception to this, namely

$$\tilde{\mathfrak{g}}_1 = \mathfrak{sl}(2, \mathbb{H}) \oplus \mathfrak{su}(2) \quad \text{and} \quad \tilde{\mathfrak{g}}_2 = \mathfrak{sl}(4, \mathbb{R}) \oplus \mathfrak{sl}(2, \mathbb{R})$$

acting linearly on $\mathbb{R}^8$ (and hence an 8×8 matrix algebra). The first appears in our Table 1, but not the second. It acts on $\mathbb{R}^8$ identified with 4×2 real matrices as follows: if X is a 4×2 matrix and $(A, B) \in \mathfrak{sl}(4, \mathbb{R}) \oplus \mathfrak{sl}(2, \mathbb{R})$, then $(A, B)X = AX - XB$. This (linear) action is irreducible but not transitive. (It does have open orbits however.) Thus if the $n \times n$ matrices of our bilinear system generate an algebra $\mathfrak{g}$ which is found to have the same dimension as an algebra $\tilde{\mathfrak{g}} = \tilde{\mathfrak{g}}_0 + \tilde{\mathfrak{c}}$ in Table 1, then our procedure is as follows. We use BC_1 to compute a basis for the derived subalgebra $\mathfrak{g}_0$ of $\mathfrak{g}$, use BC_2 to diagonalize the Killing form on $\mathfrak{g}_0$ and use BC_3 and BC_4 to compute the centralizer $\mathfrak{z}_0$ of $\mathfrak{g}_0$. From these computations, we can immediately see whether $\dim \mathfrak{g}_0 = \dim \tilde{\mathfrak{g}}_0$, $\mathfrak{g}_0$ is semi-simple and irreducible, and $\mathfrak{g}_0$, $\tilde{\mathfrak{g}}_0$ both compact or both noncompact. If all of these conditions are satisfied and $n \leqq 7$ or $n = 8$ with $\tilde{\mathfrak{g}}_0 \neq \mathfrak{sl}(2, \mathbb{H}) \oplus \mathfrak{su}(2)$, then $\mathfrak{g}$ is equivalent to $\tilde{\mathfrak{g}}$. When $n = 8$ and $\tilde{\mathfrak{g}}_0 = \mathfrak{sl}(2, \mathbb{H}) \oplus \mathfrak{su}(2)$, $\mathfrak{g}$ is equivalent to $\tilde{\mathfrak{g}}$ if the above conditions are satisfied and our diagonalization of the Killing form for $\mathfrak{g}_0$ yields the same number of negative diagonal entries as for $\tilde{\mathfrak{g}}_0$, namely 13.

For any given even $n > 8$, one could in principle make use of the available wealth of information regarding the representations of semi-simple Lie algebras to extend Cartan's list to $\mathbb{R}^n$ and thereby determine the extent to which more computations need to be added. We have not tried to do this and see little point in doing so for the following reasons: the remaining calculations in Steps 1 and 2 are either easy or of the same order of difficulty as those given above while for the troublesome Step 3, we have already noted that all relevant computations get rapidly out of hand as n grows large.

REFERENCES

[1] W. BOOTHBY, *A transitivity problem from control theory*, J. Differential Equations, 17 (1975), pp. 296–307.

[2] R. W. BROCKETT, *On the algebraic structure of bilinear systems*, Theory and Applications of Variable Structure Systems, R. Mohler and A. Ruberti, eds., Academic Press, New York, 1972.

[3] E. B. DYNKIN, *Normed Lie Algebras and Analytic Groups*, A.M.S. Translations no. 97, American Mathematical Society, Providence, RI, 1953.

[4] E. CARTAN, *Les groupes projectifs continus réels qui ne laissent invariante aucune multiplicité plane*, J. Math. Pures Appl., 10 (1914), pp. 149–186.

[5] D. ELLIOTT, *A consequence of controllability*, J. Differential Equations, 10 (1971), pp. 364–370.

[6] D. ELLIOTT AND T. J. TARN, *Controllability and observability for bilinear systems*, unpublished.

[7] J. KUCĔRA, *Solution in the large of control problem* $\dot{x} = (Au + Bv)x$, Czechoslavak Math. J., 17 (1967), pp. 91–96.

[8] N. LEVITT AND H. J. SUSSMAN, *On controllability by means of two vector fields*, this Journal, 13 (1975), pp. 1271–1281.

[9] C. LOBRY, *Une proprieté générique des couples de champs de vecteurs*, Czechoslovak Math. J., 22 (97), (1972), pp. 230–237.

[10] R. R. MOHLER, *Bilinear Control Processes*, Academic Press, New York, 1973.

[11] J.-P. SERRE, *Lie Algebras and Lie Groups*, W. A. Benjamin, New York, 1965.

[12] H. J. SUSSMAN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 2, March 1979

0036-0363/79/1702-0005$01.00/0

OPTIMAL THINNING OF A POINT PROCESS*

P. BRÉMAUD†

Abstract. It is shown that a problem of control in which decisions occur at the jump times of an observed point process can be reformulated as an intensity control problem, and also that the methods of dynamic programming can accommodate the case of random controls. The particular example considered was addressed by Rishel [9; *Optimal control of a Poisson source*, Proceedings JACC Conference at Purdue, 1976] and is relative to the regulation of a point process to a given rate. The question of equivalence of various information patterns (whether input only, regulated input only, or both input and regulated input are observed) is answered by the affirmative.

1. Introduction: a special class of impulse control problems. In problems of impulsive control a decision d_n is to be taken at time τ_n for each $n \geqq 1$, where $(\tau_n, n \geqq 1)$ is a sequence of increasing times. The controller has the choice of both sequences $(\tau_n, n \geqq 1)$ and $(d_n, n \geqq 1)$ and this choice is adapted to the flow of information $(\mathcal{F}_t, t \geqq 0)$ (an increasing family of sub-σ-fields of $\mathcal{F}$, where $(\Omega, \mathcal{F}, P)$ is the subjacent probability space) in the following way: for each $n \geqq 1$, τ_n is an $\mathcal{F}_t$-stopping time and d_n is an $\mathcal{F}_{\tau_n}$-measurable random variable, taking its values in the decision space $(D, \mathcal{D})$.

Such problems have received attention by Bensoussan and Lions [1] when $\mathcal{F}_t$ is generated by a Wiener process or a diffusion. The corresponding dynamic programming conditions are not of the Hamilton–Jacobi type, but take the form of "quasi-variational inequations".

In the case of point process observations (when $\mathcal{F}_t$ is generated by a marked point process; see [3] for definitions) quasi-variational inequations also arise; however, when the sequence $(\tau_n, n \geqq 1)$ is constrained to be contained in the sequence of jumps of the observation process, it is possible to restate the problem as one of control of the intensity of a point process (intensity controls are considered by Boel and Varaiya [2] using the martingale approach to dynamic programming), for which the optimality conditions are of the Hamilton–Jacobi type.

Although the ideas in the present article can be developed in a general framework, we have preferred to present a case study relative to the cancellation problem of Rishel [9]: there, one takes decisions at the jump times $(T_n, n \geqq 1)$ of a point process, and for each n, the decision consists in cancelling or not cancelling the corresponding point T_n, in order to regulate the thinned process (the point process after cancellations) to a given rate. The line of arguments for this case can be followed in a large class of control problems where one has to take decisions at the jumps of the observed point process: dynamic file allocation as in Segall [10], nonpreemptive dynamic priority assignment in queuing systems, routing in communications networks, etc.

The choice of the cancellation problem to illustrate the method was motivated by a question of R. Rishel concerning random strategies (each cancellation is decided after tossing a coin; the bias $\tilde{u}_n$ of the coin depends upon the observation $\mathcal{F}_n$ at time T_n, and the sequence $(\tilde{u}_n, n \geqq 1)$ is the control): Is there a difference between the three cases: $\mathcal{F}_n$ is the past at time T_n of (1) the original point process only, (2) the thinned point process only, or (3) both processes? It turns out that, if one is interested in optimal controls, there is no difference. This follows from the completeness of the class of pure strategies

* Received by the editors May 5, 1977, and in revised form July 10, 1978.

† This article was written while the author was visiting the Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory of the University of California, Berkeley, California 94720. Present address: IRIA/LABORIA Domaine de Voluceau, 78150 Le Chesnay, France.

($\tilde{u}_n = 0$ or 1 for all $n \geqq 1$) and the fact that for pure strategies, the three information patterns are equivalent.

2. The original problem: Taking decisions at the jump times of a point process. Let $(T_n, n \geqq 0)$ be a sequence of nonnegative random variables defined on the measure space $(\Omega, \mathscr{F})$ and such that $T_0 \equiv 0$ and

$$T_n < \infty \Rightarrow T_n < T_{n+1}.$$

This sequence is called a *point process* over the positive half line without multiple points. With such a point process one associates the *counting process* N_t defined by

$$N_t = \begin{cases} n & \text{if } t \in [T_n, T_{n+1}),\ n \geqq 0, \\ \infty & \text{if } t \geqq T_\infty \overset{\mathrm{d}}{=} \lim T_n. \end{cases} \tag{2.1}$$

The T_n's can be interpreted as arrival times of tasks (resp. customers) in a processing unit (resp. service station).

Let now $(X_n, n \geqq 1)$ be a sequence of $\{0, 1\}$ valued random variables, to be interpreted in the following manner: if $X_n = 1$, the task (or customer) arriving at time T_n is admitted for processing (or service), otherwise it is dispatched somewhere else. Therefore, the flow of tasks (or customers) in the processing unit (or service station) is represented by the counting process Y_t defined by:

$$Y_t = \sum_{n \geqq 1} X_n 1(T_n \leqq t). \tag{2.2}$$

The nth cancellation decision X_n is taken after tossing a coin: if the coin falls "Heads up" then $X_n = 1$, 0 otherwise. The "coin" is biased and its bias varies with n and the available observation at time n. If $\mathscr{F}_n$ is for each $n \geqq 1$, a sub-σ-field of $\mathscr{F}$ representing the observation at time n, the probability that $X_n = 1$ conditioned by $\mathscr{F}_n$ is $\tilde{u}_n$. Therefore for each n, $\tilde{u}_n$ has to be $\mathscr{F}_n$-measurable. The choice of the sequence $\tilde{u} = (\tilde{u}_n, n \geqq 1)$ is at the discretion of the controller. The set $\tilde{U}$ consisting of all sequences $(\tilde{u}_n, n \geqq 1)$ of $[0, 1]$-valued random variables such that for each $n \geqq 1$, $\tilde{u}_n$ is $\mathscr{F}_n$-measurable is called the set of admissible controls.

A control $\tilde{u} \in \tilde{U}$ is sought that minimizes the expectation of

$$\int_0^{t_f} (Y_t - rt)^2 \, dt, \tag{2.3}$$

where t_f and r are positive real constants; t_f is the terminal time, and r is a rate at which the input Y_t is to be regulated by cancellation over N_t, or thinning of N_t (the average value of the quantity in (2.3) gives a measure of the quality of the regulation).

To the following three cases:

(1) $\mathscr{F}_n = \sigma(T_1, \cdots, T_n, X_0, X_1, \cdots, X_{n-1})$,

(2) $\mathscr{F}_n = \sigma(T_1, \cdots, T_n)$,

(3) $\mathscr{F}_n = \sigma(X_0 T_0, X_1 T_1, \cdots, X_{n-1} T_{n-1})$,

there correspond respectively the classes of admissible control $\tilde{U}_1$, $\tilde{U}_2$ and $\tilde{U}_3$. For each i, $i = 1, 2, 3$, define $\tilde{U}_{p,i}$ to be the set of pure strategies in $\tilde{U}_i$, that is to say the set of controls $\tilde{u}$ such that $\tilde{u} \in U_i$ and $\tilde{u}_n = 0$ or 1 for all $n \geqq 1$.

In this article we show that in the case where N_t is Poisson, there exists a control $\tilde{u}$ that belongs to the intersection $\bigcap_{i=1}^{3} \tilde{U}_{p,i}$ and is optimal for all classes of admissible controls $\tilde{U}_i$, $i = 1, 2, 3$. In other words the class of pure strategies is complete, and, as far as optimality is concerned, the three above observation patterns are equivalent.

The plan of proof is as follows:

(a) First we show that $\tilde{U}_{p,1} \equiv \tilde{U}_{p,2} \equiv \tilde{U}_{p,3}$.

(b) Then we show that an optimal control for the minimization problem corresponding to $\tilde{U}_1$ is in $\tilde{U}_{p,3}$.

Before proving the announced result, we need to state the problem in more precise mathematical terms. In particular the probability structure corresponding to each control has to be explicated. When this is done, we will replace the original problem with in fact one of control of the intensity of a point process, as studied in Boel and Varaiya [2]. We then apply the method of dynamic programming, via martingales.

3. The transformed problem: Control of intensity. Let Ω be the set of double sequences $\omega = (t_n, x_n, n \geqq 0)$ where

$$t_0 = 0; \quad t_n < \infty \Rightarrow t_n < t_{n+1}, \quad \forall n \geqq 0$$

and

$$x_0 = 0; \qquad x_n = 0 \text{ or } 1, \quad \forall n \geqq 0.$$

Let $\mathcal{F}$ be the σ-field generated by the mappings $T_n: \omega \to t_n$ and $X_n: \omega \to x_n$, for all $n \geqq 0$. Let N_t and Y_t be defined by (2.1) and (2.2). Let $\tilde{u} = (\tilde{u}_n, n \geqq 1)$ be a sequence of $[0, 1]$-valued random variables such that for each $n \geqq 1$, $\tilde{u}_n$ is $\sigma(X_0, \cdots, X_{n-1}, T_0, \cdots, T_n)$-measurable. The following requirements:

$$\text{(3.1)} \quad \tilde{P}_{\tilde{u}}[T_{n+1} - T_n \leqq t | \sigma(X_0, \cdots, X_n, T_0, \cdots, T_n)] = 1 - \exp\left\{-\int_{T_n}^{T_n+t} \lambda(s)\, ds\right\},$$

$$\text{(3.2)} \qquad \tilde{P}_{\tilde{u}}[X_n = 1 | \sigma(X_0, \cdots, X_{n-1}, T_0, \cdots, T_n)] = \tilde{u}_n,$$

where $t \to \lambda(t)$ is some nonnegative, measurable, locally integrable, real-valued function, completely characterize a probability measure $\tilde{P}_{\tilde{u}}$ on $(\Omega, \mathcal{F})$, up to completion (we assume in the sequel that $(\Omega, \mathcal{F}, \tilde{P}_{\tilde{u}})$ is indeed complete). Define Z_t to be the bivariate point process (N_t, Y_t). Then:

LEMMA 1. (a) *N_t is, with respect to $\tilde{P}_{\tilde{u}}$, a Poisson process with the intensity $\lambda(t)$ and for all $0 \leqq s \leqq t$, $N_t - N_s$ is $\tilde{P}_{\tilde{u}}$-independent of $\mathcal{F}_s^Z$.*

(b) *For some $\mathcal{F}_t^Z$-predictable* $[0, 1]$*-valued process u_t, Y_t has the $(\tilde{P}_{\tilde{u}}, \mathcal{F}_t^Z)$-intensity $\lambda(t)u_t$ and moreover*:

$$\text{(3.3)} \qquad u_{T_n} = \tilde{u}_n \quad \tilde{P}_{\tilde{u}}\text{-a.s.}$$

Proof. (a) From Lazaro [8, Lemma 3.3, pp. 286–287] it follows that:

$$\text{(3.4)} \qquad \mathcal{F}_{T_n}^Z = \sigma(X_0, X_1, \cdots, X_n, T_0, \cdots, T_n)$$

and from Boel, Varaiya and Wong [14, Cor. 2.4, p. 1002]:[1]

$$\text{(3.5)} \qquad \mathcal{F}_{T_n^-}^Z = \sigma(X_0, X_1, \cdots, X_{n-1}, T_0, \cdots, T_n).$$

Therefore (3.1) reads

$$\text{(3.6)} \qquad \tilde{P}_{\tilde{u}}[T_{n+1} - T_n \leqq t | \mathcal{F}_{T_n}^Z] = 1 - \exp\left\{-\int_{T_n}^{T_n+t} \lambda(s)\, ds\right\}$$

and by Jacod [7, Prop. 3.1, p. 241] it follows that $\lambda(s)$ is the $(\tilde{P}_{\tilde{u}}, \mathcal{F}_t^Z)$-intensity of N_t. From the version of Watanabe's characterization theorem given in [4], this suffices to ensure that conclusion (a) is true.

[1] Although the result of [14] is given in the special framework of Blackwell spaces, it is valid in our situation; without hypothesis on $(\Omega, \mathcal{F})$, as in [8]. The proof is an easy combination of the proofs in [8] and [14].

(b) For any nonnegative $\mathcal{F}_t^Z$-predictable process C_t, any $\tilde{u}$

$$0 \leqq \tilde{E}_{\tilde{u}}\left[\int_0^\infty C_s\,dY_s\right] \leqq \tilde{E}_{\tilde{u}}\left[\int_0^\infty C_s\,dN_s\right]. \tag{3.7}$$

This implies that on $P(\mathcal{F}_t^Z)$ the $\mathcal{F}_t$-predictable σ-field on $(0,\infty)\times\Omega$, the measure $\tilde{P}_{\tilde{u}}(d\omega)\,dY_t(\omega)$ is absolutely continuous with respect to the measure $\tilde{P}_{\tilde{u}}(d\omega)\,dN_t(\omega)$, and that a version of the corresponding Radon–Nikodym derivative, $u_t(\omega)$, is $[0,1]$-valued. Since, by definition of the intensity, $\tilde{P}_{\tilde{u}}(d\omega)\,dN_t(\omega)=\tilde{P}_{\tilde{u}}(d\omega)\lambda(t)\,dt$ on $\mathcal{P}(\mathcal{F}_t^Z)$, $\tilde{P}_{\tilde{u}}(d\omega)\,dY_t(\omega)=\tilde{P}_{\tilde{u}}(d\omega)u_t(\omega)\lambda(t)\,dt$ on $\mathcal{P}(\mathcal{F}_t^Z)$. In other words $u_t\lambda(t)$ is the $(\tilde{P}_{\tilde{u}},\mathcal{F}_t^Z)$-intensity of Y_t. Note that the process u_t is $[0,1]$-valued and $\mathcal{F}_t^Z$-predictable (as Radon–Nikodym derivative of measures on $\mathcal{P}(\mathcal{F}_t^Z)$, $(t,\omega)\to u_t(\omega)$ is $\mathcal{P}(\mathcal{F}_t^Z)$-measurable). In other words the marked point process $(T_n, X_n, n\geqq 1)$ admits the $(\tilde{P}_{\tilde{u}},\mathcal{F}_t^Z)$-predictable measure μ defined by:

$$\begin{aligned}\mu(\omega, dt\times\{1\}) &= \lambda(t)u_t(\omega)\,dt,\\ \mu(\omega, dt\times\{0\}) &= \lambda(t)(1-u_t(\omega))\,dt\end{aligned} \tag{3.8}$$

and therefore, as follows from Jacod [7, Prop. 3.1, p. 241]:

$$\tilde{P}_{\tilde{u}}[X_n=1|\mathcal{F}_{T_n-}^Z]=u_{T_n} \tag{3.9}$$

and (3.3) then follows from (3.9) and (3.5). □

With the help of Lemma 1, we can replace the original problem $\tilde{\mathcal{P}}$, which consists of minimizing $\tilde{J}(\tilde{u})=\tilde{E}_{\tilde{u}}[\int_0^{t_f}(Y_t-rt)^2\,dt]$ among all $\tilde{u}\in\tilde{U}$ (recall that $\tilde{U}$ can take 6 values: $\tilde{U}_i$, $\tilde{U}_{p,i}$, $i=1,2,3$, and therefore there are 6 problems $\tilde{\mathcal{P}}$: $\tilde{\mathcal{P}}_i$, $\tilde{\mathcal{P}}_{p,i}$, $i=1,2,3$), by an equivalent problem of control of the intensity of a point process:

Problem $\mathcal{P}$. Let U be the set of nonnegative $\mathcal{F}_t^Z$-predictable $[0,1]$-valued processes u and let for each $u\in U$, P_u be *the* probability measure on $(\Omega,\mathcal{F})$ that makes N_t a point process with the $(P_u,\mathcal{F}_t^Z)$-intensity $\lambda(t)$ and Y_t a point process with the $(P_u,\mathcal{F}_t^Z)$-intensity $\lambda(t)u_t$ (the existence of P_u is a result of [3]; the uniqueness is due to Jacod [7]). Problem $\mathcal{P}$ consists in minimizing $J(u)=E_u[\int_0^{t_f}(Y_t-rt)^2\,dt]$. Here again we can define six problems $\mathcal{P}_i$, $\mathcal{P}_{p,i}$, $i=1,2,3$ and six sets of admissible controls U_i, $U_{p,i}$, $i=1,2,3$, as we already did above for $\tilde{\mathcal{P}}$ and $\tilde{U}$.

Since our plan is to show that the optimal control $\tilde{u}^*$ for $\tilde{\mathcal{P}}_1$ is in $\tilde{U}_{p,3}$ we will—for notational convenience—denote $\tilde{U}_1$ and $\tilde{\mathcal{P}}_1$ by $\tilde{U}$ and $\tilde{\mathcal{P}}$. Similarly U_1 and $\mathcal{P}_1$ will be denoted by U and $\mathcal{P}$.

PROPOSITION 2. *If u^* is an optimal solution for $\mathcal{P}$, then $\tilde{u}^*$ defined by*

$$\tilde{u}_n^*=u_{T_n}^*,\quad \forall n\geqq 1, \tag{3.10}$$

is an optimal solution for $\tilde{\mathcal{P}}$.

Proof. To any $u\in U$, we can associate $\tilde{u}\in\tilde{U}$ by:

$$\tilde{u}_n=u_{T_n},\quad \forall n\geqq 1, \tag{3.11}$$

and moreover $P_u=\tilde{P}_{\tilde{u}}$: indeed, $\tilde{P}_{\tilde{u}}$ is the unique probability measure on $(\Omega,\mathcal{F})$ such that (3.1) and (3.2) are verified, and P_u also satisfies (3.1) and (3.2) (in the course of the proof of Lemma 1 we have seen that $P_u[X_n=1|\sigma(X_0,\cdots,X_n,T_0,\cdots,T_n)]=u_{T_n}$ which is equal to $\tilde{u}_n$ by definition of $\tilde{u}$).

Similarly, to $\tilde{u}\in\tilde{U}$, one can associate, according to Lemma 1, $u\in U$ such that (3.11) holds, and moreover $P_u=\tilde{P}_{\tilde{u}}$, by the same arguments as above.

The result is then clear since when $P_u\equiv\tilde{P}_{\tilde{u}}$, $\tilde{J}(\tilde{u})=J(u)$. □

We now proceed to the solution of $\mathcal{P}$.

4. The solution via dynamic programming. The following remark will be useful: if we let N be any fixed integer greater than rt_f, then clearly one can restrict one's attention to controls u such that $u_t(\omega)=0$ on $\{Y_t(\omega)\geqq N\}$. Indeed, if for some $t_0\in[0,t_f]$, $Y_{t_0}\geqq rt_f$, then for all $t\in[t_0,t_f]$ and all $k\geqq 0$, $(Y_t+k-rt)^2\geqq(Y_t-rt)^2$. Let us call $\bar{U}$ the class of controls $U\cap\{u/u_t(\omega)=0 \text{ on } \{Y_t(\omega)\geqq N\}\}$.

LEMMA 3. *Suppose there exists a family* $(t\to V(t,n), n\geqq 0)$ *of measurable mappings from* $\mathcal{R}$ *into* $\mathcal{R}$ *such that for all* $n\geqq 0$, $t\to V(t,n)$ *is differentiable and satisfies*

$$\frac{dV}{dt}(t,n)+\lambda(t)1(V(t,n+1)-V(t,n)\leqq 0)(V(t,n+1)-V(t,n))+(n-rt)^2=0 \tag{4.1}$$

and

$$V(t_f,n)=0. \tag{4.2}$$

Then u^* *defined by*:

$$u_t^*=1(V(t,Y_{t^-}+1)-V(t,Y_{t^-})\leqq 0) \tag{4.3}$$

is optimal for $\mathcal{P}$. *Moreover*:

$$V(0,0)=J(u^*)=\inf_{u\in U}J(u). \tag{4.4}$$

Proof. We first note that:

$$1(V(t,n+1)-V(t,n)\leqq 0)(V(t,n+1)-V(t,n))=\inf_{u\in[0,1]}u(V(t,n+1)-V(t,n)), \tag{4.4'}$$

so that (4.1) reads:

$$\frac{dV}{dt}(t,n)+\inf_{u\in[0,1]}\lambda(t)u[V(t,n+1)-V(t,n)]+(n-rt)^2=0. \tag{4.1'}$$

Let us now decompose $V(t,Y_t)$ as

$$V(t,Y_t)=V(0,0)+\sum_{0<T_n\leqq t}(V(T_n,Y_{T_n})-V(T_{n-1},Y_{T_{n-1}}))+V(t,Y_t)-V(\theta_t,Y_{\theta_t})$$

where

$$\theta_t=\sup(T_j/T_j\leqq t).$$

Equivalently:

$$\begin{aligned}V(t,Y_t)=V(0,0)&+\sum_{0<T_n\leqq t}(V(T_n,Y_{T_n})-V(T_n,Y_{T_{n-1}}))\\&+\sum_{0<T_n\leqq t}(V(T_n,Y_{T_{n-1}})-V(T_{n-1},Y_{T_{n-1}}))+V(t,Y_t)-V(\theta_t,Y_{\theta_t}).\end{aligned} \tag{4.5}$$

Now, observing that $Y_{T_{n-1}}=Y_{T_n^-}$ and $Y_t=Y_{\theta_t}$:

$$\sum_{0<T_n\leqq t}(V(T_n,Y_{T_{n-1}})-V(T_{n-1},Y_{T_{n-1}}))+V(t,Y_t)-V(\theta_t,Y_{\theta_t})=\int_0^t\frac{dV}{ds}(s,Y_s)\,ds. \tag{4.6}$$

Also:

$$\begin{aligned}V(T_n,Y_{T_n})-V(T_n,Y_{T_{n-1}})&=V(T_n,Y_{T_n^-}+X_n)-V(T_n,Y_{T_n^-})\\&=(V(T_n,Y_{T_n^-}+1)-V(T_n,Y_{T_n^-}))X_n\end{aligned}$$

and therefore:

$$E[(V(T_n, Y_{T_n}) - V(T_n, Y_{T_{n-1}}))1(T_n \leqq t)]$$
$$= E[(V(T_n, Y_{T_n^-}+1) - V(T_n, Y_{T_n^-}))u_{T_n}1(T_n \leqq t)]$$

(where we have made use of the fact that T_n is $\mathcal{F}^Z_{T_n^-}$-measurable and that $P_u[X_n = 1/\mathcal{F}^Z_{T_n^-}] = u_{T_n}$). Therefore, for any $u \in \tilde{U}$:

$$\text{(4.7)}\qquad E\left[\sum_{0<T_n\leqq t} (V(T_n, Y_{T_n}) - V(T_n, Y_{T_{n-1}}))\right] = E\left[\int_0^t (V(s, Y_{s^-}+1) - V(s, Y_{s^-}))u_s\, dN_s\right]$$
$$= E\left[\int_0^t (V(s, Y_{s^-}+1) - V(s, Y_{s^-}))u_s\lambda(s)\, ds\right]$$

by the integration theorem of [3].[2] Combining (4.5), (4.6) and (4.7) we obtain:

$$\text{(4.8)}\quad E_u[V(t, N_t)] = V(0,0) + E_u\left[\int_0^t \left\{\frac{dV}{ds}(s, Y_s) + \lambda(s)u_s(V(s, Y_s+1) - V(s, Y_s))\right\} ds\right].$$

Adding $\int_0^t (Y_s - rs)^2\, ds$ to both sides of (4.8) we obtain, after letting $t = t_f$ and using (4.2)

$$\text{(4.9)}\quad J(u) = V(0,0) + E_u\left[\int_0^{t_f} \left\{\frac{dV}{ds}(s, Y_s) + \lambda(s)u_s(V(s, Y_s+1) - V(s, Y_s)) + (Y_s - rs)^2\right\} ds\right].$$

By (4.1), $J(u) \geqq V(0,0)$ for all $u \in U$, and by definition of u^*, $J(u^*) = V(0,0)$. □

LEMMA 4. *There exists a solution to* (4.1)–(4.2).

Proof. Let N be as in the remark at the beginning of the present paragraph. Then clearly:

$$\text{(4.10)}\qquad V(t,n) = \int_t^{t_f} (n - rs)^2\, ds, \qquad n \geqq N,$$

satisfies

$$\frac{dV}{dt}(t,n) + \lambda(t)1(V(t, n+1) \leqq V(t,n))(V(t,n+1) - V(t,n)) + (n - rt)^2 = 0, \qquad n \geqq N,$$

and

$$V(t_f, n) = 0, \qquad n \geqq N,$$

since for such $V(t,n)$, $1(V(t,n+1) \leqq V(t,n)) = 0$ for all $n \geqq N$. Therefore what remains to be solved is the finite system of differential equations:

$$\text{(4.1}')\quad \frac{dV}{dt}(t,n) + \lambda(t)1(V(t,n+1) \leqq V(t,n))(V(t,n+1) - V(t,n)) + (n - rt)^2 = 0,$$
$$0 \leqq n \leqq N-1,$$

[2] At least for the controls $u \in \tilde{U}$ because then $(V(t, Y_{t^-}+1) - V(t, Y_{t^-}))u_t$ is a *bounded* $\mathcal{F}^Z_t$-predictable process (bounded by $2\sup_{0\leqq n\leqq N}\sup_{t\in[0,t_f]} V(t,n)$).

and

$$(4.2') \qquad V(t_f, n) = 0, \qquad 0 \leqq n \leqq N-1,$$

where $V(t, N) = \int_t^{t_f} (N - rs)^2 \, ds$. In view of the following definitions:

$$V(t) = \begin{pmatrix} V(t, 0) \\ V(t, 1) \\ \vdots \\ V(t, N-1) \end{pmatrix}, \qquad \phi(t) = \begin{pmatrix} (0 - rt)^2 \\ (1 - rt))^2 \\ \vdots \\ (N - 1 - rt)^2 \end{pmatrix},$$

$$A = \begin{pmatrix} -1 & +1 & 0 & \cdots & 0 \\ 0 & -1 & +1 & & \vdots \\ \vdots & & & -1 & +1 \\ 0 & \cdots & & 0 & -1 \end{pmatrix}$$

$$x = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{N-1} \end{pmatrix}, \qquad F(t, x) = \begin{pmatrix} F_0(t, x) \\ F_1(t, x) \\ \vdots \\ F_{N-1}(t, x) \end{pmatrix},$$

$$F_i(t, x) = \lambda(t) \inf_{u \in [0,1]} \{u x_i\}, \qquad i = 0, 1, \cdots, N-2,$$

$$F_{N-1}(t, x) = \lambda(t) \inf_{u \in [0,1]} \{u(V(t, N) - x_{N-1})\},$$

system (4.1′)–(4.2′) can be written:

$$(4.11) \qquad \frac{dV}{dt} + F(t, AV) + \phi.$$

Since the mappings $x \to F(t, x)$ is Lipschitz, as well as $x \to Ax$, (4.11) has a unique solution. □

We can now state the main result:

THEOREM 5 (*Completeness of Pure Strategies*). *There exists a common optimal solution to* $\tilde{\mathcal{P}}_i$, $i = 1, 2, 3$ (*resp.* $\mathcal{P}_i$, $i = 1, 2, 3$) *corresponding to a pure strategy.*

Proof. By Lemmas 3 and 4, there exists an optimal solution to $\mathcal{P}_1$ that belongs to $U_{p,3}$. Equivalently, by Proposition 2, there exists an optimal solution of $\tilde{\mathcal{P}}_1$ that belongs to $\tilde{U}_{p,3}$. The conclusion follows from:

LEMMA 6. $\tilde{U}_{p,1} = \tilde{U}_{p,2} = \tilde{U}_{p,3}$.

Proof. Clearly it suffices to show that $\tilde{U}_{p,1} \subset U_{p,2}$ and $\tilde{U}_{p,1} \subset \tilde{U}_{p,3}$. We only prove $\tilde{U}_{p,1} \subset \tilde{U}_{p,2}$, since $\tilde{U}_{p,1} \subset \tilde{U}_{p,3}$ [3] is proven with exactly the same arguments.

Let $\tilde{u} \in \tilde{U}_{p,1}$. There exists a sequence $(f^{(n)}, n \geqq 1)$ of measurable functions $f^{(n)}: R^{2n} \to \{0, 1\}$ such that:

$$(4.12) \qquad \tilde{u}_n = f^{(n)}(T_1, \cdots, T_n, X_0, \cdots, X_{n-1}), \quad \forall n \geqq 1.$$

[3] It has been pointed out by a referee that we do not need to show this in order to prove Theorem 5 since the optimal solution to ($\mathcal{P}_1$) is in $\tilde{U}_{p,3}$ and $\tilde{U}_{p,3} \subset \tilde{U}_{p,1} = \tilde{U}_{p,2}$. Also both referees have insisted, with good reasons, that we mention that the "shape" of $\tilde{u}$ be part of the information (by the shape, we mean the $f^{(n)}$'s such that $\tilde{u}_n = f^{(n)}(\mathcal{F}_n)$, see (4.12) for instance). This is equivalent to saying that the controller should not be so absent-minded that he forgets about his "strategy" (the "shape" of $\tilde{u}$).

Now since $\tilde{u}_n \in \{0, 1\}$ and $\tilde{u}_n = \tilde{P}_{\tilde{u}}[X_n = 1/\mathscr{F}_n]$, $X_n = \tilde{u}_n$, $\tilde{P}_{\tilde{u}}$ a.s., therefore:

$$\tilde{u}_1 = f^{(1)}(T_1, X_0) = f^{(1)}(T_1, 0) = g^{(1)}(T_1),$$

$$\tilde{u}_2 = f^{(2)}(T_1, T_2, X_1, X_0) = f^{(2)}(T_1, T_2, g^{(1)}(T_1), 0) = g^{(2)}(T_1, T_2), \quad \tilde{P}_{\tilde{u}} \text{ a.s.}$$

and so, to obtain the general representation

$$\tilde{u}_n = g^{(n)}(T_1, T_2, \cdots, T_n), \quad \tilde{P}_{\tilde{u}} \text{ a.s.}$$

where $g^{(n)}$ is a measurable mapping from R^n into $\{0, 1\}$. If we let $\bar{u}$ be defined by:

$$\tilde{u}'_n = g^{(n)}(T_1, T_2, \cdots, T_n)$$

then, firstly $\tilde{u}' \in \tilde{U}_{p,2}$, and secondly $\tilde{u}'_n = \tilde{u}_n$, $\tilde{P}_{\tilde{u}}$ a.s. And this clearly implies that $\tilde{P}_{\tilde{u}} = \tilde{P}_{\tilde{u}'}$, at least if both $(\Omega, \mathscr{F}, P_{\tilde{u}})$ and $(\Omega, \mathscr{F}, \tilde{P}_{\tilde{u}'})$ are completed, as was assumed.

5. Concluding Remarks.

(1) This article is a new version of [11]. Although the arguments in [11] are completely different and somewhat complicated, it shows the interaction between various methods of solution of stochastic control problems (dynamic programming, maximum principle along the lines of Bismut [12], control of Kolmogorov equations as in [13]). It also contains a partial treatment of the case where the incoming point process is Markovian.

(2) Although a particular problem was considered here, the method used for transforming a problem of impulsive control into one of intensity control and the proof of completeness of pure strategies are quite general.

(3) The fact that one could restrict attention to controls depending only upon the present state of the output has its analogues in other situations. For instance, in the dynamical file allocation problem considered by Segall [10], it could be proven that the class of controls considered there (depending only upon the present position of the file) is indeed complete. Note however, that this result strongly depends upon the Poissonian character of the inputs (in the present paper as well as in [10]). If the output were Markovian for instance, the optimal control would depend on the present value of the counting process N_t (see [11] for details).

REFERENCES

[1] A. Bensoussan and J. L. Lions, *Nouvelles méthodes en contrôle impulsionnel*, J. Appl. Math. Optimization, 1 (1975), pp. 289–312.

[2] R. Boel and P. Varaiya, *Optimal control of jump processes*, this Journal, 15 (1977), pp. 92–119.

[3] P. Brémaud, *The martingale theory of point processes with an intensity*, Proc. IRIA Coll. on Control Theory, Lecture Notes in Economics and Mathematical Systems, 107, Springer-Verlag, New York, 1974.

[4] ———, *An extension of Watanabe's characterization theorem for Poisson processes*, J. Appl. Probability, 8 (1974), pp. 396–399.

[5] P. Brémaud and J. Jacod, *Processus ponctuels et martingales: revue des résultats récents sur la modélisation et le filtrage*, Advances in Appl. Probability, 9 (1977), pp. 362–416.

[6] C. Dellacherie, *Capacités et processus stochastiques*, Springer-Verlag, Berlin, 1972.

[7] J. Jacod, *Multivariate point processes: predictable projection, Radon–Nikodym derivatives, representation of martingales*, Z. Wahrscheinlichkeits theorie und Verw. Gebiete, 31 (1975), pp. 235–253.

[8] J. de Sam Lazaro, *Sur les hélices du flot spécial sous une fonction*, Ibid., 30 (1974), pp. 279–302.

[9] R. Rishel, *Optimal control of a Poisson source*, Proc. JACC Conference at Purdue, 1976, pp. 531–535; also in IEEE Trans. Automatic Control to appear.

[10] A. Segall, *Dynamic file assignment in computer networks: Part I.* IEEE Trans. Automatic Control, AC-21 (1976), pp. 161–173.

[11] P. Brémaud, *Impulsive control of point processes: optimal cancellation of arrivals* (1977), preprint.

[12] J. M. BISMUT, *Control of jump processes and applications*, Bull. Soc. Math. France, 1977, to appear.

[13] P. BRÉMAUD, *Bang-bang controls of point processes*, Advances in Appl. Probability, 8 (1976), pp. 385–394.

[14] R. BOEL, P. VARAIYA AND E. WONG, *Martingales on jump processes, Part I: Representation results*, this Journal, 13 (1975), pp. 999–1021.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 2, March 1979

0036-0363/79/1702-0006$01.00/0

ON THE FAST SOLVING OF PARABOLIC BOUNDARY CONTROL PROBLEMS*

WOLFGANG HACKBUSCH†

Abstract. We present a multi-grid method for the solution of boundary control problems with quadratic cost functions, where the state is a solution of a parabolic initial-boundary value problem. The computational work is proportional to the work needed for the integration of the parabolic equation. The method can be extended also to nonlinear problems.

1. Introduction. Let $y(v)$ denote the solution of a parabolic initial-boundary value problem depending on the boundary value v (Neumann or Dirichlet boundary condition). If a cost function $J(v)$ is to be minimized, where J is a function of $y(v)$ and v, this is called a parabolic boundary control problem. The method presented in this paper is restricted to those problems, where the optimal control can be characterized by an equation of the form (2.4) or (2.8) involving the adjoint state $p(u)$. Also nonlinear equations like (5.2) can be treated by a slightly modified version of the algorithm.

We need no decoupling technique. Only a sequence of parabolic initial-boundary value problems is to be solved. The method can be applied to general problems. The coefficients of the parabolic system may depend on all variables. If the number of space variables exceeds one, also general regions are admissible. We do not fix the method used for discretizing the parabolic equations. One may choose a suitable difference scheme or a Galerkin method.

In the following section we present five examples of boundary control problems with different observations and different boundary conditions. Furthermore, the discretization is discussed in § 2. Section 3 investigates the equation that characterizes the optimal control. The fast solving of this equation is described in § 4. The foregoing sections are restricted to linear problems. A nonlinear problem is studied in § 5.

2. The boundary control problem and its discretization.

2.1. Notation. Following the notation of Lions [9] we denote by Ω a domain of $\mathbb{R}^n$ with the boundary Γ. $I=(0, T)$ is a finite time interval. The differential equation described below is to be satisfied in $Q=\Omega\times I$. Boundary conditions are prescribed on the lateral boundary $\Sigma=\Gamma\times I$. Σ_0 is defined by $\Gamma_0\times I$, where Γ_0 is a nonempty subset of Γ ($\Sigma_0=\Sigma$ possible). We summarize:

$$\Omega\subset\mathbb{R}^n,\quad \Gamma=\partial\Omega,\quad I=(0,T),\quad Q=\Omega\times I,\quad \Sigma=\Gamma\times I,\quad \Sigma_0\subset\Sigma.$$

Here we study *boundary* control problems. Therefore, the control functions are defined on Σ_0. The corresponding linear space is denoted by $\mathcal{U}$. In the main part of this paper we assume the case of no constrains, that means that the subset $\mathcal{U}_{ad}$ of *admissible* controls coincides with $\mathcal{U}$.

Let A be an elliptic differential operator of second order[1] with coefficients defined on Q. Let A^* be the formally adjoint operator. Let B and C be boundary operators of first order such that Green's formula

$$(2.1)\qquad (Ay, p)_{L^2(\Omega)}-(y, A^*p)_{L^2(\Omega)}=(y, Cp)_{L^2(\Gamma)}-(By, p)_{L^2(\Gamma)}$$

holds (cf. Lions and Magenes [10, pp. 2 and 157]).

* Received by the editors May 4, 1978.

† Mathematisches Institut, Universität zu Köln, Köln, West Germany.

[1] More general problems can be treated without difficulty.

2.2. Six examples of control problems. In § 3.2 we shall study the properties of the following control problems.

2.2.1. Observation of the total state (Neumann problem). We choose $\mathcal{U}_{ad} = L^2(\Sigma_0)$. For any $u \in \mathcal{U}_{ad}$, $y(u) = y(x, t; u)$ is defined as the solution of the parabolic initial-boundary value problem

$$\frac{\partial}{\partial t} y(u) + Ay(u) = f \qquad ((x, t) \in Q), \tag{2.2a}$$

$$By(u)|_\Sigma = \begin{cases} u & \text{on } \Sigma_0, \\ g & \text{on } \Sigma \backslash \Sigma_0, \end{cases} \tag{2.2b}$$

$$y(x, 0; u) = y_0(x) \qquad (x \in \Omega). \tag{2.2c}$$

We seek $u \in \mathcal{U}_{ad}$ such that $y(u) \approx z_d$ holds for a given function $z_d \in L^2(Q)$. If in addition the norm of u must be small, we obtain the following cost function

$$J(v) = \|y(v) - z_d\|^2_{L^2(Q)} + (Nv, v)_{L^2(\Sigma_0)} \qquad (v \in \mathcal{U}_{ad}),$$

where N is positive definite, e.g.,

$$N = \delta \times \text{identity}, \qquad \delta > 0. \tag{2.3}$$

The solution of $J(u) = \inf \{J(v) \colon v \in \mathcal{U}_{ad}\}$ is characterized by

$$u = -N^{-1}(p(u)|_{\Sigma_0}), \tag{2.4}$$

where $p(u)$ is the solution of

$$-\frac{\partial}{\partial t} p(u) + A^* p(u) = y(u) - z_d \qquad ((x, t) \in Q), \tag{2.5a}$$

$$Cp(u) = 0 \qquad ((x, t) \in \Sigma), \tag{2.5b}$$

$$p(x, T; u) = 0 \qquad (x \in \Omega). \tag{2.5c}$$

We repeat the proof given in [9]. $u \in \mathcal{U}_{ad}$ is optimal if and only if $J'(u) \cdot (v - u) \geqq 0$ holds for all $v \in \mathcal{U}_{ad}$, i.e.,

$$(y(u) - z_d, y(v) - y(u))_{L^2(Q)} + (Nu, v - u)_{L^2(\Sigma_0)} \geqq 0 \quad \text{for all } v \in \mathcal{U}_{ad}.$$

Using (2.5), (2.1) and (2.2) we obtain

$$\begin{aligned}
&(y(u) - z_d, y(v) - y(u))_{L^2(Q)} \\
&\quad = \left(-\frac{\partial}{\partial t} p(u) + A^* p(u), y(v) - y(u)\right)_{L^2(Q)} \\
&\quad = \left(p(u), \left(\frac{\partial}{\partial t} + A\right)(y(v) - y(u))\right)_{L^2(Q)} - (p(\cdot, t; u), y(\cdot, t; v) - y(\cdot, t; u))_{L^2(\Omega)}\Big|_{t=0}^{t=T} \\
&\qquad\quad - (Cp(u), y(v) - y(u))_{L^2(\Sigma)} + (p(u), B(y(v) - y(u)))_{L^2(\Sigma)} \\
&\quad = (p(u), v - u)_{L^2(\Sigma_0)}.
\end{aligned}$$

Equality (2.4) follows from $(Nu + p(u), v - u)_{L^2(\Sigma_0)} \geqq 0$ for all $v \in \mathcal{U}_{ad}$, since $\mathcal{U}_{ad}$ is a linear space.

By means of (2.4) the control u can be eliminated in (2.2b). The purpose of this paper is the numerical treatment of the arising coupled (y, p)-system. The application of the decoupling technique (cf. [9, p. 132]) is possible but too expensive.

2.2.2. Observation of the final state (Neumann problem). Take $\mathscr{U}_{ad} = L^2(\Sigma_0)$ and $z_d \in L^2(\Omega)$. We want to obtain $y(\cdot, T; u) \approx z_d$, where again $y(u)$ is the solution of (2.2). The cost function is

$$J(v) = \|y(\cdot, T; v) - z_d\|^2_{L^2(\Omega)} + (Nv, v)_{L^2(\Sigma_0)}.$$

The optimal control is determined by (2.4), where $p(u)$ is the solution of

$$-\frac{\partial}{\partial t} p(u) + A^* p(u) = 0 \qquad (\text{in } Q), \tag{2.6a}$$

$$Cp(u) = 0 \qquad (\text{on } \Sigma), \tag{2.6b}$$

$$p(x, T; u) = y(x, T; u) - z_d(x) \qquad (x \in \Omega) \tag{2.6c}$$

(cf. Lions [9, p. 124]).

2.2.3. Observation on the boundary (Neumann problem). Let $\mathscr{U}_{ad} = L^2(\Sigma_0)$ and $z_d \in L^2(\Sigma_0)$. We seek a solution $y(u)$ of (2.2) with $y(u)|_{\Sigma_0} \approx z_d$. The minimization of

$$J(v) = \|y(v) - z_d\|^2_{L^2(\Sigma_0)} + (Nv, v)_{L^2(\Sigma_0)}$$

leads us to (2.4), where $p(u)$ is the solution of (2.6a), (2.5c) and

$$Cp(u) = \begin{cases} y(u)|_{\Sigma_0} - z_d & \text{on } \Sigma_0, \\ 0 & \text{on } \Sigma \backslash \Sigma_0 \end{cases}$$

(cf. Lions [9, p. 187]).

2.2.4. Observation on the boundary (Dirichlet problem). Choose $\mathscr{U}_{ad} = L^2(\Sigma_0)$ and $z_d \in H^{-1}(\Sigma_0)$ (for the notation compare [9]). Let $y(u)$ be defined by (2.2a),

$$y(u)|_{\Sigma_0} = u, \qquad y(u)|_{\Sigma \backslash \Sigma_0} = g \tag{2.7}$$

and (2.2c). In order to obtain $By(u)|_{\Sigma_0} \approx z_d$, we define

$$J(v) = \|By(v)|_{\Sigma_0} - z_d\|^2_{H^{-1}(\Sigma_0)} + (Nv, v)_{L^2(\Sigma_0)},$$

where $\|f\|^2_{H^{-1}(\Sigma_0)} = (g, f)_{L^2(\Sigma_0)}$, $-\Delta_\Sigma g = f$, $g|_{\partial\Sigma_0} = 0$ (Δ_Σ: Laplace–Beltrami operator of Σ; $\Delta_\Sigma = \partial^2/\partial t^2$ for the one-dimensional case). The optimal control is characterized by

$$u = -N^{-1} Cp(u)|_{\Sigma_0}, \tag{2.8}$$

where u is the solution of (2.6a),

$$p(u)|_{\Sigma_0} = (-\Delta_\Sigma)^{-1}(By(u)|_{\Sigma_0} - z_d), \qquad p(u)|_{\Sigma \backslash \Sigma_0} = 0$$

and (2.5c) (cf. Lions [9, p. 198]).

2.2.5. Observation of the final state (Dirichlet problem). Let $\mathscr{U}_{ad} = L^2(\Sigma_0)$ and $z_d \in H^{-1}(\Omega)$. $y(u)$ be the solution of (2.2a), (2.7), (2.2c). If the cost function is

$$J(v) = \|y(\cdot, T; v) - z_d\|^2_{H^{-1}(\Omega)} + (Nv, v)_{L^2(\Sigma_0)}$$

the optimal control satisfies (2.8), where $p(u)$ is the solution of (2.6a) and

$$p(u)|_\Sigma = 0, \qquad -\Delta p(x, T; u) = y(x, T; u) - z_d \qquad (x \in \Omega),$$

$$p(x, T; u) = 0 \qquad (x \in \Gamma)$$

(cf. Lions [9, p. 202]).

2.2.6. Control by initial values and distributed control. This example is not a *boundary* control problem, but it can be treated similarly. Let $y(u) = y(u_1, u_2)$ be the solution of

$$\frac{\partial}{\partial t} y(u)+Ay(u)=f+\alpha u_1, \qquad By|_\Sigma = g,$$

$$y(\cdot, 0; u)=y_0+\beta u_2 \qquad (\alpha, \beta \in \mathbb{R}),$$

where $u=(u_1, u_2)\in \mathcal{U}_{ad}=L^2(Q)\times L^2(\Omega)$. Let $z_d \in L^2(\Omega)$. Optimizing

$$J(v)=\|y(\cdot, T; v)-z_d\|^2_{L^2(\Omega)}+(N_1v_1, v_1)_{L^2(Q)}+(N_2v_2, v_2)_{L^2(\Omega)}$$

we are led to $u_1=-\alpha N_1^{-1}p(u)$ and $u_2=-\beta N_2^{-1}p(\cdot, 0; u)$, where $p(u)$ satisfies

$$-\frac{\partial}{\partial t}p(u)+A^*p(u)=0, \qquad Cp(u)|_\Sigma=0, \qquad p(\cdot, T; u)=y(\cdot, T; u)-z_d.$$

For problems of this kind we proposed a fast numerical method in [2]. Nevertheless, the multi-grid methods of this paper can be applied. Only the considerations of § 4.2 must be changed.

2.3. Discretization. We have to solve the coupled system of $\{y, p\}$ (e.g. (2.2a, b, c) and (2.5a, b, c)), where the control u is eliminated (e.g. by (2.4)). Assume that the parabolic equations are discretized by a suitable difference scheme or a Galerkin method. The discretization parameters are the step sizes Δt and Δx of the interval $I=(0, T)$ and of the grid of Ω.

In the sequel we need a sequence of step sizes tending to zero. Let Δt_0 and Δx_0 be fixed and define

$$(2.9) \qquad \Delta t_\nu = 2^{-\nu}\Delta t_0, \qquad \Delta x_\nu = \begin{cases} 2^{-\nu/2}\Delta x_0 & (\nu \text{ even}) \\ \Delta x_{\nu-1} & (\nu \text{ odd}) \end{cases}$$

for $\nu \in \mathbb{N}_0 := \{0, 1, 2, \cdots\}$. In [4] we used the number $n_\nu = T/\Delta t_\nu + 1$ for indexing. Here it seems to be more appropriate to use the "level number" $\nu \in \mathbb{N}_0$ as index.

The discrete control function v_ν is defined for $t \in I$ with $t/\Delta t_\nu \in \mathbb{N}_0$. The discrete solutions corresponding to $y(v)$ and $p(v)$ are denoted by $y_\nu(v_\nu)$ and $p_\nu(v_\nu)$. Replacing the condition (2.4) or (2.8), respectively, by a discrete one, we are able to eliminate the control u_ν and obtain a discrete (y, p)-system. Its solution is $y_\nu = y_\nu(u_\nu)$ and $p_\nu = p_\nu(u_\nu)$, where u_ν denotes the discrete approximation to the optimal control u. In § 4.3 we shall give an example for the discretization.

We note that there is another possibility of discretization. Replace the parabolic equation for $y(v)$ by a difference method and discretize the integrals of the cost function $J(v)$ by sums taken over $y_\nu(v_\nu)$ and v_ν. Define $p_\nu(v_\nu)$ by the difference scheme, which is the adjoint one of the y_ν-scheme. Then the solution u_ν is not only an approximation to the optimal control u but also the true optimum with respect to the discrete cost function $J_\nu(u_\nu)$.

3. Interpretation by integral equations.

3.1. The integral equation. Consider the first example described in § 2.2.1. For a given control $v \in \mathcal{U}_{ad}$, $y(v)$ is defined by (2.2). In a second step $p(v)$ can be determined by (2.5) using $y(v)$. Now we define the function $w(v)=-N^{-1}(p(v)|_{\Sigma_0})$. By (2.4) the optimal control is the solution of $w(u)=u$. $u \mapsto w(u)$ is an affine mapping. Writing $w(u)=Ku+q$ (K linear, $q=w(0)$), we obtain the characterization of the optimal control u by

$$(3.1) \qquad u=Ku+q \qquad (u, q \in \mathcal{U}).$$

As illustrated in the following section, (3.1) may be called an integral equation (Fredholm's type, second kind).

In special cases, where K is explicitly known, it would be advantageous to solve the integral equation (3.1) instead of the coupled (y, p)-system. But even in the general case we shall use the representation (3.1).

Replacing the differential equations by difference schemes we obtain a discrete analogue of (3.1):

$$u_\nu = K_\nu u_\nu + q_\nu \qquad (\nu \in \mathbb{N}_0). \tag{3.2}$$

Theoretically, the matrix K_ν can be computed by calculating $w_\nu(v_\nu) = K_\nu v_\nu + q_\nu$ for $v_\nu = 0$ and all unit vectors v_ν. But only in the case of the coarsest grid size ($\nu = 0$) we shall put up with the corresponding amount of computational work. In the other cases we only have to compute $v_\nu \mapsto K_\nu v_\nu$ which requires the work of determining $y_\nu(v_\nu)$ and $p_\nu(v_\nu)$.

We shall regard K_ν as a discretization of the operator K and apply the fast algorithm for Fredholm's integral equations of second kind (cf. Hackbusch [4]). As emphasized in [4], the smoothing property of K or its powers is essential. In the next section we investigate for a model problem, how fine the topology of the range of K may be chosen so that K is still continuous.

3.2. Operator *K* for special cases. Our model problem is

$$\Omega = (0, \infty), \qquad A = A^* = -\partial^2/\partial x^2, \qquad \Sigma_0 = \Sigma = \{0\} \times I,$$
$$B = C = -\partial/\partial x, \qquad N = \delta \times \text{identity}.$$

Therefore, the equations (2.2) and (2.5) are

$$y_t = y_{xx} + f, \quad -p_t = p_{xx} + y - z_d, \quad y(x, 0) = y_0, \quad p(x, T) = 0,$$
$$-y_x(0, t) = u(t) = -p(0, t)/\delta, \qquad -p_x(0, t) = 0.$$

The term q of (3.1) vanishes if we choose $f = z_d = 0$ and $y_0 = 0$. Then, $y(u)$ has the solution

$$y(x, t) = \int_0^t \frac{u(\tau)}{\sqrt{\pi(t-\tau)}} e^{-x^2/[4(t-\tau)]}\, d\tau \qquad (x \geqq 0, 0 \leqq t \leqq T)$$

(cf. Ladyženskaja et al. [8, p. 261]). For $p(u)$ we obtain

$$p(x, t) = \frac{1}{2} \int_t^T \int_0^\infty \frac{y(\xi, \tau)}{\sqrt{\pi(t-\tau)}} \{e^{-(x-\xi)^2/[4(t-\tau)]} + e^{-(x+\xi)^2/[4(t-\tau)]}\}\, d\xi\, d\tau.$$

Substituting the representation of y and evaluating at $x = 0$, we obtain $p(0, t)$ and Ku:

$$(Ku)(t) = -\frac{1}{\delta\sqrt{\pi}} \int_0^T \{\sqrt{2T - t - \tau} - \sqrt{|t - \tau|}\} u(\tau)\, d\tau. \tag{3.3a}$$

One can verify that

$$\|K\|_{L^2(I) \to H^{3/2}(I)} \leqq C, \qquad \|K^2\|_{L^2(I) \to H^{2-\varepsilon}(I)} \leqq C(\varepsilon) \qquad (\varepsilon > 0) \tag{3.3b}$$

are the optimal results.

In order to give an idea of the operator K, we summarize the results for the remaining examples of § 2.2.

Second example (cf. § 2.2.2):

$$(3.4a)\qquad (Ku)(t) = -\frac{1}{\delta\sqrt{\pi}}\int_0^T \frac{u(\tau)}{\sqrt{2T-t-\tau}}\,d\tau \qquad (0 \leqq t \leqq T),$$

$$(3.4b)\qquad \|K\|_{L^2(I)\to H^{1/2}(I)} \leqq C, \qquad \|K^2\|_{L^2(I)\to H^{1-\varepsilon}(I)} \leqq C(\varepsilon) \qquad (\varepsilon > 0).$$

Third example (cf. § 2.2.3):

$$(3.5a)\qquad (Ku)(t) = -\frac{1}{\delta\pi}\int_0^T \log\left(\frac{2T-t-\tau+2\sqrt{(T-\tau)(T-t)}}{|t-\tau|}\right)u(\tau)\,d\tau,$$

$$(3.5b)\qquad \|K\|_{L^2(I)\to H^{1-\varepsilon}(I)} \leqq C(\varepsilon) \qquad (\varepsilon > 0).$$

Fourth example (cf. § 2.2.4): (3.5b) holds for

$$(3.6)\qquad \begin{aligned}(Ku)(t) = {}& -\frac{4\sqrt{T-t}}{T\pi\delta}\int_0^T \sqrt{T-\tau}\,u(\tau)\,d\tau \\ & -\frac{1}{\pi}\int_t^T \log\left(\frac{2T-t-\tau+2\sqrt{(T-\tau)(T-t)}}{\tau-t}\right)u(\tau)\,d\tau.\end{aligned}$$

Fifth example (cf. § 2.2.5): (3.4b) holds for

$$(3.7)\qquad (Ku)(t) = -\frac{1}{2\delta\sqrt{\pi}}\int_0^T \frac{u(\tau)}{\sqrt{2T-t-\tau}}\,d\tau.$$

Sixth example (cf. § 2.2.6): K is a continuous mapping from $L^2(Q)\times L^2(\Omega)$ into $H^{2,1}(Q)\times H^s(\Omega)$ with arbitrary $s \in \mathbb{R}$. K^2 maps $L^2(Q)\times L^2(\Omega)$ into $H^{3-\varepsilon,(3-\varepsilon)/2}(Q)\times H^s(\Omega)$ ($s\in\mathbb{R}$, $\varepsilon>0$).

4. Numerical method for solving the discrete system.

4.1. Description of the method. In [4] we described a multi-grid iteration basing on the following two-level method. Given a grid function v_ν, we apply the smoothing procedure m times

$$w_\nu \mapsto K_\nu w_\nu + q_\nu$$

(cf. (3.2)). The standard value of m is one. The result is denoted by $\tilde{v}_\nu$. The exact correction $(I_\nu - K_\nu)^{-1}\,d_\nu$ belonging to the defect

$$d_\nu = \tilde{v}_\nu - K_\nu\tilde{v}_\nu - q_\nu$$

is approximated by

$$\delta_\nu = p_{\nu,\nu-1}(I_{\nu-1} - K_{\nu-1})^{-1} r_{\nu-1,\nu}\,d_\nu.$$

I_ν is the identity mapping. $r_{\nu-1,\nu}$ denotes a restriction to the coarser grid with the step sizes $\Delta t_{\nu-1}$ and $\Delta x_{\nu-1}$ (level $\nu-1$), while $p_{\nu,\nu-1}$ is the prolongation from the coarser grid to the finer one (level ν). The simplest choice is

$$(4.1)\qquad (r_{\nu-1,\nu}w_\nu)(x,t) = \tfrac{1}{2}w_{\nu-1}(x,t) + \tfrac{1}{4}[w_{\nu-1}(x,t+\Delta t_\nu) + w_{\nu-1}(x,t-\Delta t_\nu)]$$

for $(x,t)\in\Sigma_{0,\nu-1}$,[2]

$$(4.2)\qquad (p_{\nu,\nu-1}w_{\nu-1})(x,t) = \begin{cases} w_{\nu-1}(x,t), & (x,t)\in\Sigma_{0\nu},\ t/\Delta t_{\nu-1}\in\mathbb{N}_0 \\ [w_{\nu-1}(x,t+\Delta t_\nu) + w_{\nu-1}(x,t-\Delta t_\nu)]/2 & (x,t)\in\Sigma_{0\nu},\ t/\Delta t_{\nu-1} - \tfrac{1}{2}\in\mathbb{N}_0 \end{cases}$$

[2] Define $w_{\nu-1}(x,-\Delta t_\nu) = w_{\nu=1}(x,0)$, $w_{\nu-1}(x,T+\Delta t_\nu) = w_{\nu-1}(x,T)$.

in the one-dimensional case. $\Sigma_{0\nu}$ denotes the grid points of Σ_0 with respect to the level ν. In the case of more space variables the definitions (4.1), (4.2) must be completed by a local summation or linear interpolation in the spatial directions of Σ_0, respectively. One iteration of the two-level method is given by the mapping

$$v_\nu \mapsto \tilde{v}_\nu - \delta_\nu.$$

The two-level method requires the solving of $(I_{\nu-1} - K_{\nu-1})^{-1} w_{\nu-1}$. The solution of this linear equation can be approximated by a two-level method corresponding to $\nu - 1$ and $\nu - 2$ etc. Applying the two-level iteration recursively, we obtain the multi-grid method. The following program similar to ALGOL describes the implementation of the multi-grid iteration. The procedure calls the vector-valued function *system*(μ, v_μ, d_μ) which is defined by $K_\mu v_\mu + d_\mu$. This result can be calculated as follows. Given the control v_μ, compute $y_\mu(v_\mu)$ and $p_\mu(v_\mu)$ by the corresponding *homogeneous* discrete systems. Then $-N_\mu^{-1}\ (p_\mu|_{\Sigma_0}) + d_\mu$ is the desired result. Note that for $\mu = \nu$ the argument d_ν always coincides with q. If $q = q_\nu$ is defined by (3.2), it is not necessary to know q_ν explicitly. Integrating the nonhomogeneous discrete systems, *system* $(\nu, v_\nu, q_\nu) = -N_\nu^{-1}\ (p_\nu(v_\nu)|_{\Sigma_0})$ holds. The matrix K_0 (better the LU-decomposition of $I_0 - K_0$) must be known.

```
**procedure** recursive ($\nu$, $i$, $m$, $v$, $q$);
**value** $\nu$, $i$; **integer** $\nu$, $i$, $m$; **array** $q$, $v$;
**comment** $\nu$: $\nu \in \mathbb{N}_0$, number of the actual level,
            $i$: number of the iteration steps on the level $\nu$,
            $m$: number of smoothing steps (usually $m = 1$),
            $v$: input value: starting value $v_\nu^{(0)}$, output value: $v_\nu^{(i)} \approx (I_\nu - K_\nu)^{-1} q$,
            $q$: $(I_\nu - K_\nu)v = q$ is to be solved;
**if** $\nu = 0$ **then** $v := (I_o - K_0)^{-1} q$ **else**
**for** $i := i$ **step** $-1$ **until** 1 **do**
**begin integer** $j$; **array** $d$, $w$;
    **for** $j := 1$ **step** 1 **until** $m$ **do** $v :=$ system($\nu$, $v$, $q$);
    $d := r_{\nu-1,\nu} * (v -$ system$(\nu, v, q))$; $w := 0$;
    recursive ($\nu - 1$, 2, $m$, $w$, $d$); **comment** $w \approx (I_{\nu-1} - K_{\nu-1})^{-1} d$;
    $v := v - p_{\nu,\nu-1} * w$
**end** $i$ iterations on the level $\nu$ by the multi-grid iteration;
```

In [4] we proved that the rate of convergence on the level ν is proportional to $2^{-\nu\gamma}$ (cf. § 4.3), where the positive number γ depends on the smoothness of the interpolation by $p_{\nu,\nu-1}$ and on the quality of the approximation of K_ν to K. The rate of convergence tends to zero for $\nu \to \infty$. But on the other way round this fact implies that the rate can exceed the value one if ν is small enough. Therefore, the step sizes Δt_0 and Δx_0 corresponding to $\nu = 0$ must be sufficiently small. The determination of the coarsest grid can be combined with the iterative process in the following way.

ALGORITHM.

1) choose preliminary values Δt_0, Δx_0;

2) set $\nu := 0$, compute K_0 and $u_0 := (I_0 - K_0)^{-1} q_0$;

3) set $\nu := \nu + 1$; start with $u_\nu^{(0)} := p_{\nu,\nu-1} u_{\nu-1}$ and compute $u_\nu^{(i)}$ (e.g. $i = 1$) by calling the procedure recursive (ν, i, 1, u_ν, q_ν); in case of divergence go to 5); set $u_\nu := u_\nu^{(i)}$;

4) estimate the discretization error e.g. by means of $u_\nu - p_{\nu,\nu-1} u_{\nu-1}$. If the error is not small enough, repeat 3);

5) set $\Delta t_0 := \Delta t_0/4$, $\Delta x_0 := \Delta x_0/2$; go to 2).

4.2. Amount of computational work. The following note shows that the work of one multi-grid iteration is of the same order as required for the integration of the (y, p)-systems.

NOTE 1. *Let N_ν be the number of operations needed for the integration of the discrete (y, p)-systems on the level ν, i.e., for the performance of the mapping $v_\nu \mapsto y_\nu(v_\nu) \mapsto p_\nu(v_\nu)$. We require*

$$(4.3) \qquad N_\nu \leqq C \cdot 2^{\nu(1+n/2)} \qquad (\nu \in \mathbb{N}_0;\ n\text{: dimension of } \Omega \subset \mathbb{R}^n).$$

If $n = 1$, (4.3) *follows from* (2.9). *But even if an implicit discretization is used for $n > 1$, the inequality* (4.3) *holds provided that a fast method* (*e.g., the multi-grid method of* [3]) *is applied to the discrete elliptic equations.*

We assume that the amount $\tilde{N}_\nu$ of the computational work of the prolongation be proportional to the number of grid points of $\Sigma_{0\nu}$:

$$\tilde{N}_\nu \leqq C \cdot 2^{\nu(n+1)/2}.$$

Here we exclude the example of § 2.2.6, *although similar results can be obtained in this case.*

Then one iteration of the multi-grid method, i.e., one call of the procedure recursive *with $i = 1$ requires less than*

$$(4.4) \qquad \left[\frac{1}{2}+\frac{m+\frac{1}{2}}{1-2^{-n/2}}\right]C2^{\nu(1+n/2)}+\tilde{C}\times\begin{Bmatrix}2^{\nu(n+1)/2}/[1-2^{(1-n)/2}]\\ \nu 2^\nu\end{Bmatrix}+2^\nu G_0 \quad \begin{matrix}(n>1)\\ (n=1)\end{matrix}$$

operations, where G_0 is the number of the grid points of Σ_{00} corresponding to the level $\nu = 0$. The determination of $I_0 - K_0$ and its LU-decomposition require

$$G_0N_0+\tfrac{2}{3}G_0^3$$

operations. For $m = n = 1$, the number (4.4) *is about* $5.6 \times N_\nu$.

Proof. Let M_ν be the work of one iteration on the level ν.

$$M_\nu \leqq (m+1)N_\nu + 2M_{\nu-1} + \tilde{N}_\nu \leqq (m+1)C2^{\nu(1+n/2)} + 2M_{\nu-1} + \tilde{C}2^{\nu(n+1)/2}$$

implies

$$M_\nu \leqq C2^{\nu(1+n/2)}(m+1)/(1-2^{-n/2})+\tilde{C}\times\begin{Bmatrix}(1-2^{(n-1)/2})^{-1}2^{\nu(n+1)/2}\\ \nu 2^\nu\end{Bmatrix}+2^\nu G_0.$$

The result (4.4) follows if we take into account that the first smoothing step on the levels $\mu < \nu$ is gratis, since $w = 0$ is used as starting value.

Usually, one iteration step per level is sufficient, i.e. $i = 1$ can be chosen in step 3) of the algorithm. Then the total amount is $\approx 8.7 \times N_\nu$ if $n = m = 1$ and $\approx 4.7 \times N_\nu$ if $m = 1$, $n = 2$.

Finally we remark that the storage requirement amounts to $O(2^\nu + 2^{\nu n/2}) = O(\Delta t_\nu^{-1} + \Delta x_\nu^{-n})$, since only one t-level of the functions y_ν and p_ν must be stored during the computation. Here again we have to exclude the case of the example of § 2.2.6.

4.3. Convergence.

4.3.1. Discussion of the necessary conditions. We repeat the essential assumptions that are needed for the proof of convergence (cf. [4]). There must be two Banach spaces B_0 and $B_1 \subset B_0$, where B_1 has a finer topology than B_0. We require that the mappings $K\colon B_0 \to B_0$, $K^m\colon B_0 \to B_1$ are continuous:

$$(4.5a) \qquad \|K\|_{B_0 \to B_0} \leqq C,$$

$$(4.5b) \qquad \|K^m\|_{B_0 \to B_1} \leqq C \qquad (m \geqq 1 \text{ fixed}),$$

where m is the number of smoothing steps (cf. § 4.1). If $m = 1$, (4.5b) implies (4.5a).

B_0^ν and B_1^ν are suitable vector spaces consisting of grid functions of the level $\nu \in \mathbb{N}_0$ defined on $\Sigma_{0\nu}$. Usually, $B_0^\nu = B_1^\nu$ holds, whereas the norms are different. $\|\cdot\|_{B_i^\nu}$ is a discrete analogue of $\|\cdot\|_{B_i}$ $(i = 0, 1)$. The connection is given by the prolongations P_ν: $B_0^\nu \to B_0$ and restrictions R_ν: $B_i \to B_i^\nu$ related to $p_{\nu,\nu-1}$: $B_i^{\nu-1} \to B_i^\nu$ and $r_{\nu-1,\nu}$: $B_i^\nu \to B_i^{\nu-1}$ $(i = 0, 1$; cf. [4]). The mapping K_ν defined in (3.2) must have properties analogous to (4.5a), (4.5b):

$$\text{(4.6a)} \qquad \|K_\nu\|_{B_0 \to B_0} \leqq C \qquad (\nu \in \mathbb{N}_0),$$

$$\text{(4.6b)} \qquad \|K_\nu^m\|_{B_0 \to B_1} \leqq C \qquad (\nu \in \mathbb{N}_0),$$

where the constant C does not depend on ν.

Finally we require the consistency, e.g., in the form[3]

$$\text{(4.7)} \qquad \|P_\nu K_\nu R_\nu - K\|_{B_1 \to B_0} \leqq C \cdot 2^{-\nu\beta} \qquad (\nu \in \mathbb{N}, \beta > 0).$$

The further conditions of [4] refer to $r_{\nu-1,\nu}$ and $p_{\nu,\nu-1}$. They are fulfilled for the restrictions and prolongations defined by (4.1), (4.2), if reasonable spaces B_i and B_i^ν are chosen.

Then Theorem 3.5 of [4] implies that the rate of convergence of the multi-grid method on the level ν is proportional to $2^{-\nu\gamma}$ with[4] $\gamma = \min(\beta, s, 2)$, where s is defined as follows. In our situation $B_0 = L^2(\Sigma_0)$ and the Sobolev space

$$B_1 = H^{2s,s}(\Sigma_0) = H^s(I, L^2(\Gamma_0)) \cap L^2(I, H^{2s}(\Gamma_0))$$

(cf. [10]) are the most suitable spaces. $s \leqq 2$ is the greatest number such that (4.5b) holds.

(4.5a) and (4.6a) are usual stability properties. (4.5b) is valid, since we choose the space B_1 according to (4.5b). But it seems not to be trivial that (4.6b) holds for the discretization proposed in § 2.3. Therefore, we shall give two examples which demonstrate that (4.6b) is a natural condition.

4.3.2. One-dimensional example. We return to the model problem of § 3.2. The spaces $L^2(\Sigma_0)$ and $H^{2s,s}(\Sigma_0)$ mentioned above degenerate to $B_0 = L^2(I)$ and $B_1 = H^s(I)$. From (3.3b), (3.4b), (3.5b) we see that s depends on the problem and on the choice of $m = 1$ or $m = 2$. s takes the values $\frac{1}{2}$, $1 - \varepsilon$, $\frac{3}{2}$, $2 - \varepsilon$. The following analysis is restricted to the second example (observation of the final state, Neumann problem, cf. § 2.2.2).

Let $B_0^\nu = B_1^\nu$ be the space of grid functions with step size $\Delta t_\nu = 2^{-\nu} \cdot \Delta t_0$ and define the prolongation P_ν: $B_0^\nu \to H^1(I)$ by piecewise linear interpolation. Then we may define the discrete norm

$$\|v_\nu\|_{H_\nu^s} := \|P_\nu v_\nu\|_{H^s(I)} \qquad (0 \leqq s \leqq 1,\ v_\nu \in B_0^\nu).$$

B_0^ν and B_1^ν are endowed with the norms $\|\cdot\|_{H_\nu^0}$ and $\|\cdot\|_{H_\nu^s}$ with $s = \frac{1}{2}$ or $s = 1 - \varepsilon$ according to $B_1 = H^s(I)$ and (3.4b). In the following we write H_ν^0 and H_ν^s instead of B_0 and B_1.

[3] (4.7) is sufficient for proving the consistency condition (2.12) of [4], provided that $\|I_\nu - R_\nu P_\nu\|_{B_1^\nu \to B_0^\nu} \leqq C 2^{-\nu\beta}$ holds. Note that the convergence property $\|[(I_\nu - K_\nu)^{-1} R_\nu - R_\nu (I - K)^{-1}]q\|_{B_0^\nu} \leqq C \cdot 2^{-\nu\beta} \|q\|_{B_1}$ does not coincide with $\|u_\nu - u\|_{B_0^\nu} \leqq C \cdot 2^{-\nu\beta} \|q\|_{B_1}$ (u, u_ν from (3.1), (3.2)), since $q_\nu \neq R_\nu q$. But if y_0, f, g, z_d are sufficiently smooth, $q_\nu - R_\nu q$ tends to zero, too.

[4] If $n > 1$ and $\gamma > 1$, the interpolation involved in (4.2) with respect to the spatial directions must be of higher order.

NOTE 2. *Discretize* (2.2a) *and* (2.2b) *by*

$$\frac{1}{\Delta t_\nu}[y_\nu(x,t)-y_\nu(x,t-\Delta t_\nu)] \qquad (t=\Delta t_\nu(\Delta t_\nu)T, \quad x=\Delta x_\nu(\Delta x_\nu)\infty),$$

$$\text{(4.8a)} \qquad =\frac{1}{\Delta x_\nu^2}[y_\nu(x+\Delta x_\nu,t)-2y_\nu(x,t)+y_\nu(x-\Delta x_\nu,t)]+f(x,t)$$

$$\text{(4.8b)} \qquad \frac{1}{\Delta x_\nu}[y_\nu(0,t)-y_\nu(\Delta x_\nu,t)]=u_\nu(t) \qquad (t=\Delta t_\nu(\Delta t_\nu)T)$$

and use analogous schemes for p_ν. *Furthermore,*

$$\text{(4.9)} \qquad 3\Delta x_\nu^2<\Delta t_\nu \leqq \text{const}\cdot \Delta x_\nu^2$$

must be fulfilled. Then the condition (4.6b) *holds. More precisely*:

$$\|K_\nu\|_{H_\nu^0\to H_\nu^{1/2}} \leqq C,$$
$$\|K_\nu^2\|_{H_\nu^0\to H_\nu^{1-\varepsilon}} \leqq C(\varepsilon), \qquad \|K_\nu^2\|_{H_\nu^0\to H_\nu^1} \leqq C|\ln(\Delta t_\nu)| \qquad (\varepsilon>0).$$

Proof. Use discrete Fourier transformations. The complete proof is given in the appendix of [7].

4.3.3. Discrete Galerkin method in the general case. Without loss of generality we again restrict our considerations to the problem of § 2.2.2 with N defined by (2.3). In order to determine K, the functions y_0, z_d, g, f are chosen to be zero. Lions and Magenes [10, p. 80] proved[5] that $v\in L^2(\Sigma_0)$ implies $y(u)$, $p(u)\in H^{3/2,/3/4}(Q)$. Therefore, the restriction $p(u)|_{\Sigma_0}$ belongs to $H^{1,1/2}(\Sigma_0)$ (cf. [10, p. 9]). Thus, (4.5b) holds for $B_1=H^{1,1/2}(\Sigma_0)$. In [5] we defined a discrete Galerkin solution corresponding to the step sizes Δx_ν, Δt_ν and proved[6] that

$$\text{(4.10)} \qquad \begin{aligned}\|P_\nu p_\nu(v)-p(v)\|_{H^{r,r/2}(Q)} &\leqq C(\Delta t_\nu)^{(s-r)/2}\|p(v)\|_{H^{s,s/2}(Q)}\\ &\leqq C'(\Delta t_\nu)^{(s-r)/2}\|v\|_{H^{s-3/2,s/2-3/4}(\Sigma_0)}\end{aligned}$$

holds for $0\leqq r\leqq s$, $1\leqq s\leqq 2$, where again P_ν denotes the prolongation by piece-wise linear interpolation. $p_\nu(v)$ and $y_\nu(v)$ are defined by means of the continuous control v. They coincide with $p_\nu(\tilde v_\nu)$ and $y_\nu(\tilde v_\nu)$, where $\tilde v_\nu$ is the discrete control obtained from v as

[5] We omit the conditions on the smoothness of the coefficients and of Γ required in [10] and [5]. Also the assumptions on the finite element spaces are to be seen from [5].

[6] If the extension of $A(t)$ and $B(t)$ for $t\in(T,2T)$ by $A^*(2T-t)$ and $C(2T-t)$ is sufficiently smooth, $p(t)$ coincides with $u(2T-t)$ for $t\in(0,T)$. Thus, (4.10) follows. In the general case (4.10) holds for $P_\nu y_\nu(v)-y(v)$. Denoting the solution of (2.6a, b) and $p(T)=y(T)$ by $p(y(T))$ we have

$$\begin{aligned}&\|p_\nu(v)-p(v)\|_{H^{r,r/2}(Q)}\\ &\leqq \|p_\nu(v)-p(y_\nu(T))\|_{H^{r,r/2}(Q)}+\|p(y_\nu(T))-p(y(T))\|_{H^{r,r/2}(Q)}\\ &\leqq C[(\Delta t_\nu)^\alpha\|p(y_\nu(T))\|_{H^{s,s/2}(Q)}+\|y_\nu(T)-y(T)\|_{H^{r-1}(\Omega)}]\\ &\leqq C'[(\Delta t_\nu)^\alpha\|y_\nu(T)\|_{H^{s-1}(\Omega)}+\|y_\nu(v)-y(v)\|_{H^{r,r/2}(Q)}]\\ &\leqq C''(\Delta t_\nu)^\alpha\|y(v)\|_{H^{s,s/2}(Q)}\leqq C'''(\Delta t_\nu)^\alpha\|v\|_{H^{s-3/2,s/2-3/4}(\Sigma_0)}\left(\alpha=\frac{s-r}{2}\right)\end{aligned}$$

for $1<r\leqq s\leqq 2$. This result is sufficient for demonstrating (4.11).

described in [5]. Note that $\tilde{v}_\nu \neq R_\nu v$. Choosing $r = s = \frac{3}{2}$ in (4.10), we obtain

$$\|[P_\nu p_\nu(v) - p(v)]|_{\Sigma_0}\|_{H^{1,1/2}(\Sigma_0)} \leqq C\|v\|_{L^2(\Sigma_0)}. \tag{4.11}$$

Together with (4.5b), $\|P_\nu p_\nu(v)\|_{H^{1,1/2}(\Sigma_0)} \leqq C\|v\|_{L^2(\Sigma_0)}$ follows. This implies also $\|K_\nu v_\nu\|_{H^{1,1/2}(\Sigma_0)} \leqq C\|v_\nu\|_{L^2(\Sigma_0)}$; thus, (4.6b) is proved.

Let $v \in H^{1,1/2}(\Sigma_0)$ be arbitrary and define $y(v)$ and $p(v)$ by the homogeneous equations (2.2), (2.6). As mentioned above $\tilde{v}_\nu$ is not equal to $R_\nu v$, but $\|\tilde{v}_\nu - R_\nu v\|_{B_0^\nu} \leqq C(\Delta t_\nu)^{1/2}\|v\|_{H^{1,1/2}(\Sigma_0)}$ holds. (4.10) yields the (nonoptimal) estimate

$$\|P_\nu K_\nu \tilde{v}_\nu - Kv\|_{L^2(\Sigma_0)} \leqq C\|P_\nu p_\nu - p\|_{H^{1,1/2}(Q)} \leqq C'\sqrt{\Delta t_\nu}\|v\|_{H^{1,1/2}(\Sigma_0)}$$

which proves (4.7) with $\beta = \frac{1}{2}$. Thus, the rate of convergence is $O(2^{-\nu\gamma})$ with $\gamma = \frac{1}{2}$.

4.4. Numerical Example. Consider again the model problem of § 3.2, but replace $\Omega = (0, \infty)$ by $\Omega = (0, 1)$. Choose $\Sigma_0 = \{0\} \times I$, $T = 1$, $\delta = 1$ and let

$$g = 0, \quad f = -\frac{\pi}{2}\cos\left(\frac{\pi}{2}(1-x)\right), \quad y_0 = -\frac{2}{\pi}\cos\left(\frac{\pi}{2}(1-x)\right), \quad z_d = 1 - \frac{2}{\pi}\cos\left(\frac{\pi}{2}(1-x)\right)$$

be the coefficients of the systems (2.2), (2.6). The optimal control is $u(t) = -1$.

We apply the multi-grid iteration with $\Delta t_0 = \frac{1}{2}$, $\Delta x_0 = \frac{1}{4}$ and different values of ν. The following table shows the observed rates of convergence of the multi-grid iteration on the level ν.

TABLE 1
Rate ρ_ν depending on the level number ν.

ν	1	2	3	4	5
Δt_ν	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
Δx_ν	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{32}$
ρ_ν	0.086	0.063	0.044	0.030	0.019

The theoretical rate is $\rho_\nu = O(\Delta t_\nu^\gamma)$ with $\gamma = \frac{1}{2}$ since $B_1 = H^{1/2}(I)$. This is confirmed by the values of the table for $(\rho_1/\rho_5)^{1/4} = 1.46$ approaches $\sqrt{2}$.

In order to give an idea of the results we list the errors

$$\varepsilon_\nu^{(i)} = \{\Delta t_\nu \Sigma |u_\nu^{(i)}(t) - u(t)|^2\}^{1/2} \qquad (i\text{: number of iteration})$$

(sum taken over $t = \Delta t_\nu(\Delta t_\nu)1$) and the discretization error

$$\varepsilon_\nu = \{\Delta t_\nu \Sigma |u_\nu(t) - u(t)|^2\}^{1/2}$$

that appear when the algorithm of § 4.1 is performed with $i = 1$ and the maximal level number $\nu = 5$.

level $\nu = 1$:	$\varepsilon_1^{(1)} = 7.8_{10}-3$	$\varepsilon_1 = 1.9_{10}-3$
level $\nu = 2$:	$\varepsilon_2^{(1)} = 1.7_{10}-3$	$\varepsilon_2 = 1.8_{10}-3$
level $\nu = 3$:	$\varepsilon_3^{(1)} = 1.8_{10}-3$	$\varepsilon_3 = 4.4_{10}-4$
level $\nu = 4$:	$\varepsilon_4^{(1)} = 4.1_{10}-4$	$\varepsilon_4 = 4.3_{10}-4$
level $\nu = 5$:	$\varepsilon_5^{(1)} = 4.3_{10}-4$	$\varepsilon_5 = 1.1_{10}-4$
	$\varepsilon_5^{(2)} = 1.1_{10}-4$	

The norms of the differences $\delta_5^{(i)} = u_5^{(i-1)} - u_5^{(i)}$ are

i	2	3	4	5	6
$\{\Delta t_5 \Sigma \delta_5^{(i)}\|^2\}^{1/2}$	$3.2_{10}-4$	$4.2_{10}-6$	$1.1_{10}-7$	$2.2_{10}-9$	$4.2_{10}-11$

resulting in the rate of convergence $\rho_5 = 0.019$.

5. Unilateral constraints. In the following we study a boundary control problem of unilateral type as one example of a nonlinear problem that can be treated by the nonlinear multi-grid method.

5.1. Characterization of the optimal control. Instead of $\mathcal{U}_{ad} = \mathcal{U} = L^2(\Sigma_0)$ we now choose

$$\mathcal{U}_{ad} = \{v \in \mathcal{U} = L^2(\Sigma_0): v \geqq 0 \text{ almost everywhere on } \Sigma_0\}$$

and assume (2.3): $N = \delta \times \text{identity}$. The optimal control u is defined by means of the adjoint state $p(u)$ (e.g. cf. Lions [9, p. 125]):

$$u = [-N^{-1}(p(u)|_{\Sigma_0})]_+, \qquad v_+ := \max(0, v). \tag{5.1}$$

Therefore, the linear equation (3.1) is to be replaced by

$$u = \mathcal{K}(u) := [Ku + q]_+. \tag{5.2}$$

The nonlinear equation (5.2) is discretized by

$$u_\nu = \mathcal{K}_\nu(u_\nu) := [K_\nu u_\nu + q_\nu]_+. \tag{5.3}$$

5.2. Nonlinear multi-grid iteration. Again we use the algorithm of § 4.1, but the procedure recursive must be replaced by the following procedure that is to be called with the parameter $f = 0$.

```
procedure nonlinear ($\nu$, $i$, $m$, $v$, $f$);
value $\nu$, $i$; integer $\nu$, $i$, $m$; array $v$, $f$;
comment $\nu$, $i$, $m$, $v$: compare procedure recursive,
        $v$, $f$: $v, f \in B_0^\nu$, the solution of $v = \mathcal{K}_\nu(v)+f$ is sought.
        Note that the solution $u_{\nu-1}$ of (5.3) is used. Since $u_{\nu-1}$ corresponds to the
        foregoing level, it is already computed;
if $\nu = 0$ then $v$ := (solution of $v = \mathcal{K}_0(v)+f$) else
for $i$ := $i$ step $-1$ until 1 do
begin integer $j$; array $d$, $w$;
        for $j$ := 1 step 1 until $m$ do $v := \mathcal{K}_\nu(v)+f$;
        $d := r_{\nu-1,\nu} * (v - \mathcal{K}_\nu(v) - f)$; $w := u_{\nu-1}$;
        nonlinear ($\nu-1$, 2, $m$, $w$, $d$); comment $w \approx \mathcal{K}_{\nu-1}(w)+d$;
        $v := v - p_{\nu,\nu-1} * (w - u_{\nu-1})$
end $i$ iterations on the level $\nu$ by the nonlinear multi-grid iteration;
```

This programme is equivalent to the procedure *recursive* if $\mathcal{K}$ is an affine mapping. Note that only on the level $\nu = 0$ a nonlinear equation is to be solved. All other computations are explicit. Useful modifications of the nonlinear multi-grid method are discussed in [11].

5.3. Convergence of the nonlinear multi-grid method. In [4] we proved that the rate of convergence of the nonlinear multi-grid iteration is asymptotically equal to the linear iteration with K_ν replaced by the Fréchet derivative of $\mathcal{K}_\nu(v_\nu)$ at $v_\nu = u_\nu$ (u_ν solution of (5.3)). Let U be a neighborhood of the solution u of (5.2) with respect to the topology of B_0 and define $U_\nu := \{v_\nu \in B_0^\nu : P_\nu v_\nu \in U\}$, where $P_\nu : B_0^\nu \to B_0$ is the prolongation mentioned in § 4.3.1. Assuming the convergence $P_\nu u_\nu \to u$, we require $u_\nu \in U_\nu$ for all $\nu \in \mathbb{N}_0$. It is sufficient that $\mathcal{K}_\nu$ fulfills the Lipschitz condition

$$\text{(5.4a)} \qquad \mathcal{K}_\nu(v_\nu) - \mathcal{K}_\nu(w_\nu) = L_\nu(v_\nu, w_\nu)(v_\nu - w_\nu)$$

for all v_ν, $w_\nu \in U_\nu$, where L_ν is a bounded linear operator:

$$\text{(5.4b)} \qquad \|L_\nu(v_\nu, w_\nu)\|_{B_0^\nu \to B_0^\nu} \leqq C \quad \text{for all } v_\nu, w_\nu \in U_\nu.$$

Moreover, we assume that $L_\nu(v_\nu, w_\nu)$ be continuous at (u_ν, u_ν), where u_ν is the solution of (5.3):

$$\text{(5.4c)} \qquad \|L_\nu(v_\nu, w_\nu) - L_\nu(u_\nu, u_\nu)\|_{B_0^\nu \to B_0^\nu} \to 0 \quad \text{if } v_\nu, w_\nu \to u_\nu.$$

The following note shows that the nonlinear operator $\mathcal{K}$ of (5.2) satisfies (5.4a, b, c) ($\mathcal{K}_\nu$, v_ν, w_ν to be replaced by $\mathcal{K}$, v, w). Analogous arguments can be used for the proof of (5.4a, b, c) with $\nu \in \mathbb{N}_0$.

NOTE 3. *Assume that*

$$\text{(5.5)} \qquad \textit{measure } \{(x, t) \in \Sigma_0 : |(Ku + q)(x, t)| \leqq \varepsilon\} \to 0 \quad \textit{as } \varepsilon \to 0.$$

Then (5.4) *holds with*

$$L(v, w) = \chi(v, w) \times K,$$

where χ *is the function*

$$\chi(v, w)(x, t) = \begin{cases} 0 & \textit{if } Kv + q \leqq 0, Kw + q \leqq 0, \\ (Kv + q)/(K(v - w)) & \textit{if } Kv + q > 0, Kw + q \leqq 0, \\ 1 & \textit{if } Kv + q \geqq 0, Kw + q > 0, \\ (Kw + q)/(K(w - v)) & \textit{if } Kv + q < 0, Kw + q > 0. \end{cases}$$

Proof. (5.4a) follows from the definition of $\mathcal{K}$ and L. Since the function χ takes only values of $[0, 1]$, the boundedness of K implies (5.4b). Let μ be the measure and denote by S the subset of Σ_0 consisting of all points, where the signs of $Kv + q$ or $Kw + q$ differ from the sign of $Ku + q$. Then

$$\|[L(v, w) - L(u, u)]z\|_{L^2(\Sigma_0)} \leqq \|Kz\|_{L^2(S)}$$
$$\leqq C\sqrt{\mu(S)}\|Kz\|_{H^{1/2}(\Sigma_0)} \leqq C'\sqrt{\mu(S)}\|z\|_{L^2(\Sigma_0)}$$

follows from $K : L^2(\Sigma_0) \to H^{1,1/2}(\Sigma_0) \subset H^{1/2}(\Sigma_0)$. It remains to show that $\mu(S) \to 0$ if v, $w \to u$. Let S_1 be the set of $(x, t) \in \Sigma_0$ with $Kv + q \geqq 0$ and $Ku + q \leqq 0$. Since S is the union of sets of this kind, it is sufficient to prove $\mu(S_1) \to 0$. Note that $S_1 \subset S_{11}(\varepsilon) \cup S_{12}(\varepsilon)$, where $S_{11}(\varepsilon) = \{(x, t) \in \Sigma_0 : Ku + q \leqq -\varepsilon,\ Kv + q \geqq 0\}$ and $S_{12}(\varepsilon) = \{(x, t) : |Ku + q| \leqq \varepsilon\}$. One concludes from $\mu(S_{11}(\varepsilon)) \cdot \varepsilon^2 \leqq \|K(v - u)\|^2_{L^2(\Sigma_0)} \leqq C\|v - u\|^2_{L^2(\Sigma_0)}$ that

$$\mu(S_1) \leqq C\varepsilon^{-2}\|v - u\|^2_{L^2(\Sigma_0)} + \mu(S_{12}(\varepsilon)).$$

Choosing $\varepsilon = \|v - u\|^{1/2}_{L^2(\Sigma_0)}$ we obtain from (5.5) that $\mu(S_1) \to 0$ as $v \to u$.

REFERENCES

[1] J. L. CASTI, *Dynamical Systems and Their Applications*, Academic Press, New York-San Francisco-London, 1977.

[2] W. HACKBUSCH, *A numerical method for solving parabolic equations with opposite orientations*, Computing, to appear.

[3] ———, *On the multi-grid method applied to difference equations*, Ibid., to appear.

[4] ———, *Die schnelle Auflösung der Fredholmschen Integralgleichung zweiter Art*, Beiträge Numer. Math., 9, to appear.

[5] ———, *Optimale $H^{2r,r}$-Fehlerabschätzungen der Galerkin-Lösung eines parabolischen Anfangsrandwertproblems*, Universität zu Köln, Mathematisches Institut, rep. 78-8, 1978.

[6] ———, *On the fast solving of elliptic control problems*, in preparation.

[7] ———, *On the fast solving of parabolic boundary control problems*, Universität zu Köln, Mathematisches Institut, report 78-9, 1978.

[8] O. A. LADYŽENSKAJA, J. A. SOLONNIKOV AND N. N. URAL'CEVA, *Linear and Quasi-linear Equations of Parabolic Type*, Translations of mathematical Monographs, 23, American Mathematical Society, Providence, RI, 1968.

[9] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin-Heidelberg-New York, 1971.

[10] J. L. LIONS AND E. MAGENES, *Non-homogeneous Boundary Value problems and Applications II*, Springer-Verlag, Berlin-Heidelberg-New York, 1972.

[11] W. HACKBUSCH, *An error analysis of the nonlinear multi-grid method of second kind*, Apl. Mat., to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 2, March 1979

NECESSARY AND SUFFICIENT CONDITIONS FOR A LOCAL MINIMUM. 1: A REDUCTION THEOREM AND FIRST ORDER CONDITIONS*

A. D. IOFFE†

Abstract. A new approach to the theory of necessary conditions is described. The core of the approach is a reduction theorem which replaces the initial constrained problem by a problem without constraints having the same solution. After this, the procedure for deriving first order necessary conditions becomes almost trivial, which is demonstrated by several examples.

1. Introduction. This work is concerned with the problem

$$\text{minimize } f_0(x) \tag{1.1}$$

subject to

$$F(x)=0; \qquad f_i(x)\leqq 0, \qquad i=1,\cdots,n; \tag{1.2}$$

$$x\in S, \tag{1.3}$$

where $f_0,\cdots,f_n$ are real-valued functions on a Banach space X, F is a mapping from X into another Banach space Y and $S\subset X$. We shall be interested in conditions for a local minimum in the above problem. Hence it suffices to regard f_i and F as defined only in a neighborhood of a prescribed point $z\in S$.

We shall assume throughout that f_i and F are *Lipschitz* in the neighborhood and that S is closed.

A usual theory of necessary conditions consists of two successive components. The first is *approximation*, that is, replacing the mappings and sets in question by those having simpler structure, usually linear or convex. The second is application of one or another known theory, usually connected with convex problems. Our approach diverges from this tradition only in one point: the approximation stage is no longer the first. It is preceded by a reduction of the initial problem to an unconstrained one. With this reduction made, the final stage becomes radically simplified; it might be compared with the calculation of the derivative of a composite function.

A close approach to necessary conditions has been recently developed by Clarke [1]. His method of building an unconstrained problem is more straightforward and requires less effort. But in applying this method to problems involving equality constraints, one will inevitably be dealing with an infinite sequence of unconstrained problems whose solutions converge to the solution of the initial problem. In contrast, we consider only one unconstrained problem with the same solution as (1.1)–(1.3). This is why it is possible to incorporate a wider spectrum of approximations in the framework of our approach. But for problems containing no equality constraint, both approaches coincide.

Unlike many abstract theories of necessary conditions, such as presented in works of Dubovitzkii and Miljutin [2], Gamkrelidze [3], Neustadt [9], Pshenichnyi [10] and others, this one is not essentially axiomatic. We consider only those approximations which are inherent in Lipschitz functions (generalized gradients of Clarke, approximations of Levitin–Miljutin–Osmolovskii type, etc.). Certain results for non-Lipschitz functions such as Halkin's theorem [4] do not thereby follow from ours. In all those

* Received by the editors January 16, 1976, and in revised form January 11, 1978.

† c/o R. T. Rockafellar, Department of Mathematics, University of Washington, Seattle, Washington 98195.

papers, however, as well as in Clarke's [1] and Warga's [12], only first-order conditions were considered, whereas we shall try to present here a much more advanced theory including conditions of Levitin–Miljutin–Osmolovskii type, higher order conditions etc.

This research has been strongly inspired by the remarkable work of Levitin, Miljutin and Osmolovskii [8]. To a certain extent, our theory is a reinterpretation of that developed by the three authors. Results similar to those established in [8], as well as differences between the two approaches, will be discussed in part 2, following this paper. This first part, which might be considered as introductory, contains the reduction theorem and a general (and very elementary, in fact!) description of first order necessary conditions. In further papers we hope to consider higher order conditions and optimal control problems.

We presume that the reader is familiar with elementary properties of subdifferentials of convex functions [7], [11]. Acquaintance with the generalized gradients of Clarke [1] is also desirable. Recall how they are defined. Let f be a Lipschitz function defined in a neighborhood of z. The function

$$h \to f^0(z; h) = \limsup_{\substack{u \to z \\ t \downarrow 0}} t^{-1}(f(u+th) - f(u))$$

is convex and continuous on X, and the set

$$\partial f(z) = \{x^* \in X^* \,|\, f^0(z; h) \geqq \langle x^*, h\rangle, \forall h \in X\} = \partial f^0(z; 0),$$

which is nonempty and weak* compact, is called the *generalized gradient* of f at z.

The related notions of tangent and normal cones will be also necessary for our purposes. Let $d_S(x)$ denote the distance from x to S, and let $z \in S$. The set

$$T_S(z) = \{h \in X \,|\, d_S^0(z; h) \leqq 0\}$$

is a closed convex cone called the *tangent cone* to S at z. The polar cone

$$N_S(z) = \{x^* \in X^* | \langle x^*, h\rangle \leqq 0, \forall h \in T_S(z)\}$$

is called the *normal* cone to S at z. Note that $N_S(z)$ contains $\partial d_S(z)$ and, moreover, the closure of the cone generated by $\partial d_S(z)$ coincides with $N_S(z)$.

2. The reduction theorem. Recall (see [6]) that z is said to be a *regular point* for F relative to S if there are $k > 0$ and a neighborhood U of z such that for all $x \in U \cap S$

$$d_Q(x) \leqq k \, \|F(x) - F(z)\|,$$

where $Q = \{x \in S \,|\, F(x) = F(z)\}$ and $d_Q(x)$ denotes the distance from x to Q.

In what follows, we assume that z satisfies not only (1.3) but also (1.2), in particular $F(z) = 0$, and denote

$$I = \{i \in \{1, \cdots, n\} \,|\, f_i(z) = 0\}.$$

THEOREM 1. *Let z be a regular point for F relative to S. If z is a local (isolated local) solution to* (1.1)–(1.3), *then for all sufficiently large $r > 0$, the function*

$$M_r(x) = \max\left\{f_0(x) - f_0(z), \max_{i \in I} f_i(x)\right\} + r(\|F(x)\| + d_S(x))$$

attains local (strict local) minimum at z.

Conversely, if $M_r(x)$ attains strict local minimum at z for some r, then z is an isolated local solution to (1.1)–(1.3).

Proof. The second part of the theorem is trivial. Similarly simple is the following assertion: if z is a local (isolated local) solution to (1.1)–(1.3), then z is a local (isolated local) solution to the problem:

$$\text{minimize } f(x) \tag{2.1}$$

subject to

$$F(x)=0, \qquad x \in S, \tag{2.2}$$

where

$$f(x)=\max\left\{f_0(x)-f_0(z), \max_{i\in I} f_i(x)\right\}.$$

Choose $q>0$ and a neighborhood V of z such that for any $x \in V \cap S$ there is $u \in S$ satisfying

$$f(u) \geqq f(z) \tag{2.3}$$

$$F(u)=0, \qquad \|x-u\| \leqq q\|F(x)\|. \tag{2.4}$$

Such choices are certainly possible since z is a local solution to (2.1), (2.2), z is a regular point for F relative to S and F is continuous in a neighborhood of z. Let $c>0$ be a Lipschitz constant for F and f on V. Take $r_1 \geqq qc$. If $x \in V \cap S$ and $u \in S$ is chosen in accordance with (2.3), (2.4), then

$$\begin{aligned} f(x) &\geqq f(x)-f(u)+f(z) \geqq -c\|x-u\|+f(z) \\ &\geqq -cq\|F(x)\|+f(z) \geqq -r_1\|F(x)\|+f(z). \end{aligned}$$

It follows that z is a local solution to the problem

$$\text{minimize } f(x)+r_1\|F(x)\|, \quad \text{subject to } x \in S.$$

In the case when z is an isolated local solution to (2.1), (2.2), we can take r_1 a bit greater to make z an isolated local solution to the latter problem.

Observe now that $x \in S$ is equivalent to $d_S(x)=0$, that z is obviously a regular point for $d_S(\cdot)$ (relative to X) and that $f(x)+r_1\|F(x)\|$ is Lipschitz. Using the same arguments as above, we find $r_2>0$ such that z is a local (isolated local) solution to the problem

$$\text{minimize } f(x)+r_1\|F(x)\|+r_2 d_S(x).$$

It remains to take $r=\max\{r_1, r_2\}$.

3. First order approximations. Let $f(x)$ be a real-valued function defined in a neighborhood of z. A real-valued function $\phi(x)$ will be called a *first order approximation* for f at z if

$$\phi(tx)=t\phi(x), \quad \forall t \geqq 0, \quad \forall x \in X$$

and

$$\limsup_{t\downarrow 0} t^{-1}(f(z+th)-f(z)-t\phi(h)) \leqq 0, \quad \forall h \in X. \tag{3.1}$$

For instance, if f is Lipschitz near z, then $\phi(x)=f^0(z;x)$ is a first order approximation for f at z, as immediately follows from the definition. This is probably the most important class of first order approximations: it contains usual linear approximations of continuously differentiable functions and directional derivatives of Lipschitz locally convex functions. We admit, however, that in particular situations other approximations might be useful.

In what follows, we shall consider only such first order approximations which are also *continuous*.

PROPOSITION 1. *Let ϕ be a first order approximation for f at z. If f attains local minimum at z, then ϕ attains absolute minimum at the origin. Hence if in addition ϕ is convex, then*

$$0 \in \partial\phi(0).$$

Proof. Fix $h \in X$. Then

$$f(z+th)-f(z)=t\phi(h)+r(t),$$

where

$$\limsup_{t\downarrow 0} t^{-1}r(t) \leqq 0.$$

If $\phi(h)<0$, then $t\phi(h)+r(t)<0$ for sufficiently small t. However, this quantity is equal to $f(z+th)-f(z)$ and must be nonnegative if t is small.

The proposition below contains the "calculus of approximations" and also is quite trivial.

PROPOSITION 2. *Let ϕ, $\phi_1, \cdots, \phi_n$ be first order approximations for f, $f_1, \cdots, f_n$ respectively at z. Then the following is true*:

(a) *if $k \geqq 0$, then $k\phi$ is a first order approximation for kf at z*;

(b) *$\phi_1+\cdots+\phi_n$ is a first order approximation for $f_1+\cdots+f_n$ at z*;

(c) $\max_{i\in I}\phi_i(x)$ *is a first order approximation for* $\max_{1\leqq i\leqq n} f_i(x)$ *at z, where*

$$I=\left\{i\in\{1,\cdots,n\}\middle|f_i(z)=\max_{1\leqq j\leqq n} f_j(z)\right\}.$$

4. A necessary condition for the initial problem. Let us return to the initial problem (1.1)–(1.3). Applying successively Theorem 1, Proposition 2, Proposition 1 and standard formulas for subdifferentials of convex functions, we come to the following result.

THEOREM 2. *Assume that*

(a) *there are convex functions $\phi_0, \cdots, \phi_n$, ψ, ρ which are first order approximations for $f_0, \cdots, f_n$, $\|F(\cdot)\|$, $d_S(\cdot)$ respectively at z and which are continuous but for at most one of them*;

(b) *z is a regular point for F relative to S.*

If z is a local solution to (1.1)–(1.3), *then there are numbers $\lambda_0 \geqq 0, \cdots, \lambda_n \geqq 0$, $r>0$ such that $\lambda_0+\cdots+\lambda_n=1$, $\lambda_i f_i(z)=0$ for $i=1,\cdots,n$* (*or equivalently, $\lambda_i=0$ if $i\neq 0$ and $i\notin I$*) *and*

$$0\in\sum_{i=0}^{n}\lambda_i\partial\phi_i(0)+r\partial\psi(0)+r\partial\rho(0).$$

To derive more specific results from here, one should apply one or another criterion for z to be a regular point for F and use certain particular kind of approximations.

5. The case dim $Y=m>\infty$. For this case, a workable condition sufficient for z to be a regular point was proved in [6]. We shall quote it in a slightly modified form.

Identify Y with R^m. Then

$$F(x)=(f_{n+1}(x),\cdots,f_{n+m}(x)), \tag{5.1}$$

where the functions f_{n+j} are defined and Lipschitz in a neighborhood of z. According to [6], z will be a regular point for F relative to S if the inclusions

$$\mu_1x_1^* +\cdots \mu_mx_m^* \in N_S(z), \qquad x_j^* \in \partial f_{n+j}(z), \qquad j=1,\cdots,m,$$

imply that $\mu_1=\cdots=\mu_m=0$.

If the opposite is true, then there are numbers $\lambda_{n+1},\cdots,\lambda_{n+m}$ not all equal to zero and such that

$$0\in\lambda_{n+1}\,\partial f_{n+1}(z)+\cdots+\lambda_{n+m}\,\partial f_{n+m}(z)+N_S(z). \tag{5.2}$$

Taking this into account, we get

THEOREM 3. *Let F be defined by* (5.1), *where the f_{n+j} are Lipschitz in a neighborhood of z, and let $\phi_0,\cdots,\phi_n$ be convex and continuous first order approximations for $f_0,\cdots,f_n$. If z is a local solution to* (1.1)–(1.3), *then there are numbers $\lambda_0,\cdots,\lambda_{n+m}$ not all equal to zero and such that*

$$\lambda_i\geqq 0 \quad \textit{for } i=0,\cdots,n; \qquad \lambda_if_i(z)=0 \quad \textit{for } i=1,\cdots,n; \tag{5.3}$$

$$0\in\sum_{i=0}^{n}\lambda_i\,\partial\phi_i(0)+\sum_{i=n+1}^{n+m}\lambda_i\,\partial f_i(z)+N_S(z). \tag{5.4}$$

We shall come to the result of Clarke [1] if we take $\phi_i(x)=f_i^0(z;x)$ and hence $\partial\phi_i(0)=\partial f_i(z)$. Observe that the inverse deduction is impossible since the method of Clarke demands that the approximations be defined at every point near z and satisfy certain semicontinuity conditions.

6. The case of differentiable *F*. Assume that F is strictly differentiable at z, i.e., F is Fréchet differentiable at z and $\|F(x+h)-F(x)-F'(z)h\|=r(x,h)\|h\|$, where $r(x,h)\to 0$ if $x\to z$, $h\to 0$. For this case too, there is a simple criterion for z to be a regular point [6, Thm. 2]. Here is one of many conceivable situations in which this criterion can be suitably applied.

THEOREM 4. *Assume that*

(a) *F is strictly differentiable at z;*

(b) *$R(F'(z))$ is a closed subspace in Y;*

(c) int $T_S(z)\neq\varnothing$ *and the cone-valued mapping $x\to T_S(x)$ is lower semicontinuous on S at z.*

Let $\phi_0,\cdots,\phi_n$ be first order convex approximations for $f_0,\cdots,f_n$ at z. If z is a local solution to (1.1)–(1.3) *then there are numbers $\lambda_i\geqq 0$, $i=0,\cdots,n$, and a vector $y^*\in Y^*$ not all equal to zero and such that*

$$-F'^*(z)y^*\in\sum_{i=0}^{n}\lambda_i\,\partial\phi_i(0)+N_S(z).$$

Here $R(F'(z))$ is the range of $F'(z)$.

Proof. If $R(F'(z))\neq Y$, then there is $y^*\neq 0$ which vanishes on $R(F'(z))$; in other words, $F'^*(z)y^*=0$ and it remains to take $\lambda_0=\cdots=\lambda_n=0$.

Assume that $R(F'(z))=Y$. If $(\operatorname{Ker} F'(z))\cap(\operatorname{int} T_S(z))=\varnothing$ we shall take a nonzero $x^*\in X^*$ separating $\operatorname{Ker} F'(z)$ and int $T_S(z)$ so that, say, $\langle x^*,x\rangle\geqq 0$ for $x\in T_S(z)$ and $\langle x^*,x\rangle=0$ for $x\in\operatorname{Ker} F'(z)$. Since $R(F'(z))=Y$, it follows that $x^*=F'^*(z)y^*$ for some $y^*\in Y^*$ which is obviously nonzero (see, for instance, [7, § 0.1]). Then

$\langle F'^*(z)y^*, x\rangle \geqq 0$ if $x \in T_S(z)$, and setting $\lambda_0 = \cdots = \lambda_n = 0$, we again get the desired result.

Assume finally that $R(F'(z)) = Y$ and that Ker $F'(z)$ meets the interior of $T_S(z)$. It follows from (c) that in this case there are $h \in \operatorname{Ker} F'(z)$, $\|h\| \leqq 1$ and $\alpha > 0$ such that any $u \in X$ with $\|u - h\| < \alpha$ belongs to $T_S(x)$ if $x \in S$ is sufficiently close to z (since $T_S(x)$ are closed convex and lower semicontinuous in x at z). For any such x,

$$C(F'(z), T_S(x)) = \sup_{\|y\| \leqq 1} \inf \{\|u\| \,|\, F'(z)u = y, u \in T_S(x)\}$$
$$\leqq \alpha^{-1} C(F'(z), X),$$

where

$$C(F'(z), X) = \sup_{\|y\| \leqq 1} \inf \{\|u\| \,|\, F'(z)u = y\}$$

because $F'(z)$ maps X onto Y.

It follows [6, Thm. 2] that z is a regular point for F relative to S. It remains to apply Theorem 2 by setting $\psi(h) = \|F'(z)h\|$, $\rho(h) = d_s^0(z; h)$.

REFERENCES

[1] F. H. CLARKE, *A new approach to Lagrange multipliers*, Math. of Operations Res., 1 (1976), pp. 165–174.

[2] A. YA. DUBOVITZKII AND A. A. MILJUTIN, *Extremum problems in presence of constraints*, Ž. Vyčisl. Mat. i Mat. Fiz., 5 (1965), pp. 395–453.

[3] R. V. GAMKRELIDZE, *First order necessary conditions and axiomatics in extremal problems*, Trudy Mat. Inst. Steklov. 112 (1971), 152–180.

[4] H. HALKIN, *Mathematical programming without differentiability*, Calculus of Variations and Control Theory, David L. Russel, ed., Academic Press, New York, 1976, pp. 279–288.

[5] ———, *Nonlinear nonconvex programming in infinite dimensional space*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, 1968, pp. 10–25.

[6] A. D. IOFFE, *Regular points of Lipschitz mapping*, Trans. Amer. Math. Soc., to appear.

[7] A. D. IOFFE AND V. M. TIKHOMIROV, *Theory of Extremal Problems*, "Nauka", Moscow, 1974; English translation, North-Holland, 1978.

[8] E. S. LEVITIN, A. A. MILJUTIN AND N. P. OSMOLOVSKII, *On conditions for a local minimum in a problem with constraints*, Mathematical Economics and Functional Analysis, B. S. Mitjagin, ed., "Nauka", Moscow, 1974, pp. 139–202. (In Russian.)

[9] L. NEUSTADT, *A general theory of extremals*, J. Comput. System Sci., 3 (1969), pp. 57–92.

[10] R. N. PSHENICHNYI, *Necessary Conditions for an Extremum*, Marcel Dekker, New York, 1971.

[11] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ., 1970.

[12] JACK WARGA, *Derivative containers, inverse functions and controllability*, Calculus of Variations and Control Theory, David L. Russel, ed., Academic Press, New York, 1976, pp. 13–46.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 2, March 1979

NECESSARY AND SUFFICIENT CONDITIONS FOR A LOCAL MINIMUM. 2: CONDITIONS OF LEVITIN-MILJUTIN-OSMOLOVSKII TYPE*

A. D. IOFFE†

Abstract. The general approach suggested in the preceding paper with the same title is applied to derive various necessary conditions similar to those which have been recently found by Levitin, Miljutin and Osmolovskii. The main technical tool is a new kind of approximation extending those used by the three authors. Normal problems are separately considered; conditions which are both necessary and sufficient are established for such problems.

1. Introduction. Here we continue to study the same problem as in [3]:

$$\text{minimize } f_0(x) \tag{1.1}$$

subject to

$$F(x)=0, \qquad f_i(x)\leqq 0, \qquad i=1,\cdots,n; \tag{1.2}$$

$$x\in S, \tag{1.3}$$

where $f_0,\cdots,f_n$ are real-valued functions defined in a neighborhood of a given point $z\in X$ satisfying (1.2) and (1.3), F is a mapping from the neighborhood into Y (X and Y being Banach spaces) and $S\subset X$. As in [3], we suppose that f_i and F are Lipschitz in a neighborhood of z.

The purpose of this paper is to present an alternative and somewhat more general version of the recent theory of Levitin, Miljutin and Osmolovskii [7], who have found a very powerful necessary condition of quite a new nature. The alternative character of our version is connected with the general approach described in [3]. Unlike the three authors, who used the traditional separation scheme enriched with a new and elegant idea of how to choose the sets to be separated, we establish the main result first for unconstrained problems and then, using the reduction theorem of [3] and ready formulas of convex analysis, pass to the general case. This makes the proofs more simple and the ideas more transparent. On the other hand, the problem considered here is somewhat more general since we incorporate nonfunctional constraints (1.3) which are absent in [7].

The principal technical equipment used here is a modified version of the approximation introduced by Levitin–Miljutin–Osmolovskii. These approximations are applicable to Lipschitz functions as well as certain other approximations suggested recently. To certain extent, the approximations of Levitin–Miljutin–Osmolovskii can be considered as a generalization of generalized gradients of Clarke [1] and derivative containers of Warga [10]. But they, as well as screens of Halkin [12] are oriented mainly to the first order conditions while the approximations of Levitin–Miljutin–Osmolovskii type were created to approach stronger necessary conditions. The crucial role of these approximations is revealed in § 3 in which an exact dual characterization for unconditional local minima of functions is presented. Using this characterization, we establish in § 4 our main result, which appears as a dual form of the reduction theorem of [3]. In examples which follow, a special emphasis is made on studying the cases of Y being finite dimensional or F being strictly differentiable. The concluding

* Received by the editors January 16, 1976, and in revised form January 11, 1978.

† c/o R. T. Rockafellar, Department of Mathematics, University of Washington, Seattle, Washington 98195.

section is devoted to normal problems in which case the main result assumes quite a perfect form presenting conditions both necessary and sufficient for a local minimum.

All notions and notations which are not explained here are the same as in [3]. We denote by

$$s(x, A) = \sup \{\langle x^*, x\rangle | x^* \in A\}$$

the support function of $A \subset X^*$. Conjugacy and subdifferentiation symbols near functions of two variables are referred to the second of them, so that $\phi^*(x, h^*)$ is the Fenchel conjugate to $\phi(x, \cdot)$ and $\partial\phi(x, h)$ is the subdifferential of $\phi(x, \cdot)$ at h, etc.

Here again the reader is assumed to be familiar with certain basic facts of convex analysis, especially those concerning dual operations ([5, § 3.4 and § 4.2]; [9, § 16]).

2. Approximations of Levitin–Miljutin–Osmolovskii type. Let $f(x)$ be a real-valued function in a neighborhood U of z. A real-valued function $\phi(x, h)$ defined on $U \times X$ will be called an *LMO-approximation* for f at z if

(i) $\phi(x, 0) = f(x)$;

(ii) for any x in a neighborhood of z, the function $h \to \phi(x, h)$ is convex continuous;

(iii)

$$\liminf_{\substack{x \to z \\ h \to 0}} \|h\|^{-1}(\phi(x, h) - f(x+h)) \geqq 0.$$

Remárk 1. This is a refinement of the notion of *thin convex approximation* introduced by Levitin, Miljutin and Osmolovskii in [7]. An equivalent modification was made simultaneously by the three authors under the name of *upper approximation.*

Consider several examples.

Example 1. Let f be strictly differentiable at z. Then

$$\phi(x, h) = f(x) + \langle f'(z), h\rangle$$

obviously satisfies (i)–(iii).

Example 2. Let $f(x) = g(G(x))$, where $G: X \to Y$ is strictly differentiable at z and g is convex and continuous in a neighborhood of $G(z)$. Then the function

$$\phi(x, h) = g(G(x) + G'(z)h)$$

is an LMO-approximation for f at z.

Example 3. Let f be Lipschitz in a neighborhood of z. Take sufficiently large $k > 0$ and set

$$\phi(x, h) = f(x) + k\|h\|.$$

More sophisticated examples for Lipschitz functions can be constructed with the help of the generalized gradients of Clarke.

Let $A \subset X^*$ be convex and weak* compact. We shall say that A is *locally effective* for f at z if for any $\varepsilon > 0$, there is $\delta = \delta(\varepsilon) > 0$ such that $\partial f(x) \subset A_\varepsilon$ whenever $\|x - z\| \leqq \delta(\varepsilon)$. Here A_ε is the (norm) ε-neighborhood of A in X^*.

To large extent, this notion is a generalization of the notion of ε-subdifferential of a convex function (see Example 5 below).

Example 4. Let f be the same as in Example 3, and let $A \subset X^*$ be locally effective for f at z. Then

$$\phi(x, h) = f(x) + s(h, A)$$

is an LMO-approximation for f at z.

Indeed, properties (i), (ii) are obviously satisfied. Take $\eta>0$ and $\delta>0$ such that $\partial f(x)\subset A_\eta$ if $\|x-z\|\leqq\delta$. If $\|x-z\|\leqq\delta/2$, $\|h\|\leqq\delta/2$ then $\|x+th-z\|\leqq\delta$ for any $t\in(0,1)$. By the mean value theorem of Lebourg [6] there are $t_0\in(0,1)$ and $x^*\in\partial f(x+t_0h)$ such that

$$\langle x^*, h\rangle=f(x+h)-f(x).$$

But $x^*\in A_\eta$ and hence

$$s(h, A_\eta)=s(h, A)+\eta\|h\|\geqq f(x+h)-f(x).$$

To conclude the example, observe that for a Lipschitz function, locally effective sets always exist. Indeed, every set

$$\overline{\text{conv}}\Big(\bigcup_{\|x-z\|\leqq\delta}\partial f(x)\Big)$$

is norm bounded and hence locally effective, provided δ is sufficiently small (because the multifunction $x\to\partial f(x)$ is upper semicontinuous). Moreover, if we are given a derivative container in the sense of Warga [10], then its convex closure is locally effective at the corresponding point.

In the next example we shall consider approximations extending directly those actually used in [4]. Let a convex function f and $\varepsilon>0$ be given. The set

$$\partial_\varepsilon f(x)=\{x^*\in X^*|f(x)+f^*(x^*)\leqq\langle x^*, x\rangle+\varepsilon\}$$

(nonempty if $x\in\operatorname{dom} f$) is called *ε-subdifferential* of f at x. This set (see [9]) is always convex, weak* closed and (if f is continuous at x) norm bounded. It is easy to verify furthermore that $\partial f(u)\subset\partial_\varepsilon f(x)$ if u lies sufficiently close to x. Thus $\partial_\varepsilon f(x)$ is locally effective for f at x. In particular

$$\sup\{\langle x^*, u\rangle-f^*(x^*)|x^*\in\partial_\varepsilon f(x)\}=f(u) \tag{2.1}$$

if u is not very far from x.

Example 5. Let $\phi(x, h)$ be an LMO-approximation for f, and set

$$\phi_\varepsilon(x, h)=\sup\{\langle h^*, h\rangle-\phi^*(x, h^*)|h^*\in\partial_\varepsilon\phi(x, 0)\}.$$

PROPOSITION 1. *If f is Lipschitz in a neighborhood of z and $\varepsilon>0$, then $\phi_\varepsilon(x, h)$ is an LMO-approximation for f at z.*

Proof. Conditions (i) and (ii) of the definition of LMO-approximations follow immediately from elementary properties of subdifferentials and the Fenchel conjugacy.

Let k be a Lipschitz constant for f. Given $\eta<k$, $\eta>0$, we choose a positive $\delta\leqq\varepsilon/(2k)$ such that

$$\phi(x, h)+\eta\|h\|\geqq f(x+h) \tag{2.2}$$

whenever $\|x-z\|\leqq\delta$, $\|h\|\leqq\delta$. It suffices to show that (2.2) would not be violated if ϕ is replaced by ϕ_ε.

Let x and h be such that $\|x-z\|\leqq\delta$, $\|h\|\leqq\delta$. By definition, $\phi_\varepsilon(x, h)\leqq\phi(x, h)$. If the equality holds for the x and h, nothing would remain to prove. Therefore we assume that $\phi_\varepsilon(x, h)<\phi(x, h)$ (which implies in particular that $h\neq0$). Let

$$t_0=\inf\{t>0|\phi_\varepsilon(x, th)<\phi(x, th)\}.$$

Then $t_0>0$ according to (2.1), and $t_0<1$ by what we have just assumed.

First we shall prove that there is an $h^*\in\partial\phi(x, t_0h)$ such that

$$\phi(x, 0)+\phi^*(x, h^*)=\varepsilon. \tag{2.3}$$

Indeed, since $\phi(x, t_0h)=\phi_\varepsilon(x, t_0h)$ and $\phi_\varepsilon \leqq \phi$, the inclusion $\partial\phi_\varepsilon(x, t_0h)\subset \partial\phi(x, t_0h)$ is true. But $\partial\phi_\varepsilon(x, t_0h)\neq \phi$ since $\phi(x, \cdot)$ is convex and continuous, and $\partial\phi_\varepsilon(x, t_0h)\subset \partial_\varepsilon\phi(x, 0)$ by definition. Hence there is $h_1^* \in \partial\phi(x, t_0h)$ such that $h_1^* \in \partial_\varepsilon\phi(x, 0)$ or in other words, $\phi(x, 0)+\phi^*(x, h_1^*)\leqq \varepsilon$.

On the other hand, if $t>t_0$, $\phi(x, th)>\phi_\varepsilon(x, th)$ and $h^*\in\partial\phi(x, th)$, then $h^*\notin\partial_\varepsilon\phi(x, 0)$ (otherwise we would have $\phi(x, th)=\phi_\varepsilon(x, th)$) and hence $\phi(x, 0)+\phi^*(x, h^*)\geqq\varepsilon$. Using routine convergence arguments (and taking into account that $\phi^*(x, h^*)=\langle h^*, th\rangle-\phi(x, th)$), we shall establish the existence of such an $h_2^*\in \partial_\varepsilon\phi(x, t_0h)$ that $\phi(x, 0)+\phi^*(x, h_2^*)\geqq\varepsilon$. Then a convex combination of h_1^* and h_2^* will satisfy (2.3).

It follows from (2.3) that

$$\varepsilon-\langle h^*, t_0h\rangle=\phi(x, 0)-\phi(x, t_0h),$$

which together with (2.2) implies that

$$\langle h^*, t_0h\rangle\geqq f(x+t_0h)-f(x)+\varepsilon-\eta t_0\|h\|\geqq\varepsilon-(k+\eta)t_0\|h\|.$$

As for $0<t_0<1$ and $\|h\|\leqq\delta\leqq\varepsilon/(2k)$, it follows that

$$\begin{aligned}\|h\|^{-1}\langle h^*, h\rangle &\geqq (t_0\|h\|)^{-1}\varepsilon-(k+\eta)\\ &\geqq \|h\|^{-1}\varepsilon-(k+\eta)\geqq k-\eta.\end{aligned}$$

Finally, since $h^*\in\partial\phi_\varepsilon(x, t_0h)$ (indeed, $h^*\in\partial\phi(x, t_0h)$ and $h^*\in\partial_\varepsilon\phi(x, 0)$),

$$\begin{aligned}\phi_\varepsilon(x, h)+\eta\|h\| &\geqq \phi_\varepsilon(x, t_0h)+\eta\|h\|+(1-t_0)\langle h^*, h\rangle\\ &\geqq \phi(x, t_0h)+\eta\|h\|+(1-t_0)(k-\eta)\|h\|\\ &\geqq \phi(x, t_0h)+\eta\|h\|\geqq f(x+t_0h),\end{aligned}$$

which completes the proof.

Here are two simple propositions reflecting elementary properties of the LMO-approximations.

PROPOSITION 2. *If $\phi(x, h)$ is an LMO-approximation for f at z, then*

$$\psi(h)=\phi'(z, 0; h)=\lim_{t\downarrow 0} t^{-1}(\phi(z, th)-\phi(z, 0))$$

is a first order approximation for f at z.

PROPOSITION 3. *Let $f_1, \cdots, f_n, f$ be Lipschitz in a neighborhood of z, and let $\phi_1, \cdots, \phi_n, \phi$ be LMO-approximations for $f_1, \cdots, f_n, f$ at z. Then the following are true:*

(a) *$k\phi$ is an LMO-approximation for kf if $k\geqq 0$;*

(b) *$\phi_1+\cdots+\phi_n$ is an LMO-approximation for $f_1+\cdots+f_n$;*

(c) $\max_{i\in I}\phi_i(x)$ *is an LMO-approximation for* $\max_{1\leqq i\leqq n} f_i(x)$, *where* $I=\{i\in\{1, \cdots, n\}\,|\, f_i(z)=\max_{1\leqq j\leqq n} f_j(z)\}$.

To conclude the section, we extend the notion of LMO-approximation to sets. Let, as before, $z\in S\subset X$, and let $x\to W(x)\subset X$ be a closed-convex-cone-valued multifunction defined in a neighborhood of z. We shall say that W is a *tangent LMO-approximation* for S at z if

$$\liminf_{\substack{x\to z\\ h\to 0,\, h\in W(x)}} \|h\|^{-1}(d_S(x)-d_S(x+h))\geqq 0.$$

PROPOSITION 4. *The following two conditions are equivalent*:

(a) *W is a tangent LMO-approximation for S at z*;

(b) $W(x)=\{h\in X\mid \phi(x,h)\leqq \phi(x,0)\}$, *where ϕ is a LMO-approximation for $d_S(\cdot)$ at z such that any function $h\to\phi(x,h)-\phi(x,0)$ is sublinear.*

Proof. Obviously, (b) implies (a). If (a) is true, we set

$$\phi(x,h)=d_S(x)+2d_{W(x)}(h).$$

Fix $\eta>0$. According to the definition, there is a $\delta>0$ such that $d_S(x+u)\leqq d_S(x)+\eta\|u\|$ whenever $\|x-z\|<\delta$, $\|u\|<\delta$, $u\in W(x)$.

Then, given $h\in X$, $x\in X$ such that $\|x-z\|<\delta$, $\|h\|<\delta/2$, we can find $u\in W(x)$ such that $\|u\|\leqq 2\|h\|<\delta$, $\|h-u\|\leqq 2d_{W(x)}(h)$. We have

$$\begin{aligned} d_S(x+h)&\leqq d_S(x+u)+\|h-u\|\\ &\leqq d_S(x)+2d_{W(x)}(h)+\eta\|u\|\leqq \phi(x,h)+\eta\|h\|. \end{aligned}$$

3. Dualization of the minimality condition. Let $\phi(x,h)$ be an LMO-approximation for $f(x)$ at z. Consider the function ($\eta>0$ fixed)

$$\varphi_\eta(x)=-\min\{\phi^*(x,h^*)\mid \|h^*\|\leqq\eta\}$$

(minimum is obviously attained.) This function has a very simple meaning. Let

$$p_\eta(x,h)=\phi(x,h)+\eta\|h\|.$$

Then (by elementary properties of Fenchel conjugacy)

$$\varphi_\eta(x)=\inf_{h\in X} p_\eta(x,h). \tag{3.1}$$

In this section, we prove the following crucial fact.

PROPOSITION 5. *Let f be Lipschitz in a neighborhood of z, and let $\phi(x,h)$ be an LMO-approximation for f at z. Then the following conditions are equivalent*:

(a) *f attains local minimum at z*;

(b) $0\in\partial\phi(z,0)$ *and for any $\eta>0$, φ_η attains local minimum at z*;

(c) $0\in\partial\phi(z,0)$ *and φ_η attains local minimum at z for some $\eta>0$.*

Proof. Note to begin with that

$$f(x)=\phi(x,0)\geqq\varphi_\eta(x) \tag{3.2}$$

since $\phi^*(x,h^*)+\phi(x,0)\geqq 0$ (the Young–Fenchel inequality).

(c)$\Rightarrow$(a): Since $0\in\partial\phi(z,0)$, we have

$$\varphi_\eta(z)\geqq-\phi^*(z,0)=\phi(z,0)=f(z),$$

which together with (3.2) implies (a).

(b)$\Rightarrow$(c) is trivial.

(a)$\Rightarrow$(b): The fact that $0\in\partial\phi(z,0)$ follows from Proposition 2 and from [3, Prop. 1].

Fix $\eta>0$. According to the definition, there is a $\delta_0>0$ such that

$$\phi(x,h)+(\eta/2)\|h\|\geqq f(x+h)\geqq f(z)$$

whenever $\|x-z\|\leqq\delta_0$, $\|h\|\leqq\delta_0$. For such x and h

$$p_\eta(x,h)\geqq f(z)+(\eta/2)\|h\|\geqq f(z). \tag{3.3}$$

Take $0<\delta\leqq\delta_0$ to ensure

$$f(x)\leqq f(z)+(\eta/2)\delta_0,\quad \text{if } \|x-z\|<\delta. \tag{3.4}$$

For such x,

$$p_\eta(x, 0) = f(x) \leqq f(z) + (\eta/2)\delta_0.$$

If in addition, $\|h\| = \delta_0$, then by (3.3)

$$p_\eta(x, h) \geqq f(z) + (\eta/2)\|h\| = f(z) + (\eta/2)\delta_0.$$

Since $p(x, \cdot)$ is convex, the latter two inequalities show that

$$\inf_{h \in X} p_\eta(x, h) = \inf_{\|h\| \leqq \delta_0} p_\eta(x, h),$$

whenever $\|x - z\| \leqq \delta$. Thus (3.1)–(3.3) imply for such x that

$$\varphi_\eta(x) \geqq f(z) \geqq \varphi_\eta(z).$$

Remark 2. As it follows from the proof, $\phi(x, h)$ need not be continuous in h; it suffices to assume that $\phi(x, \cdot)$ is *lower semicontinuous* and *proper* (nowhere equal to $-\infty$ and not everywhere equal to ∞).

4. The main theorem. Examples. We proved in [3] that under certain mild assumptions, z is a local minimum for

$$M_r(x) = \max\left\{f_0(x) - f_0(z), \max_{1 \leqq i \leqq n} f_i(x)\right\} + r(\|F(x)\| + d_S(x))$$

if it is a local solution to (1.1)–(1.3). We showed also that conversely, if M_r attains strict local minimum at z for some $r > 0$, then z is an isolated local solution to (1.1)–(1.3). In this section we make use of LMO-approximations to derive a dual equivalent to the just-quoted theorem. As a result, we describe a general way to built functions other than M_r but having the same properties.

From now on, we assume that

$$f_0(z) = 0 \tag{4.1}$$

and hence

$$M_r(x) = \max_{0 \leqq i \leqq n} f_i(x) + r(\|F(x)\| + d_S(x)).$$

If $\phi_0, \cdots, \phi_n, \psi, \rho$ are LMO-approximations for $f_0, \cdots, f_n, \|F(\cdot)\|, d_S(\cdot)$ respectively at z, then

$$g_r(x, h) = \max_{0 \leqq i \leqq \eta} \phi_i(x, h) + r(\psi(x, h) + \rho(x, h))$$

is an LMO-approximation for M_r at z by Proposition 3.

Fix $\eta > 0$, and let Γ_η denote the collection of all $(2n+4)$-tuples $(\lambda_0, \cdots, \lambda_n, h_0^*, \cdots, h_n^*, h^*, u^*)$ such that

$$\begin{aligned} &\lambda_i \geqq 0, \quad i = 0, \cdots, n; \qquad \lambda_i f_i(z) = 0, \quad i = 1, \cdots, n; \\ &\lambda_0 + \cdots + \lambda_n = 1; \end{aligned} \tag{4.2}$$

$$h_0^*, \cdots, h_n^*, h^*, u^* \in X^*; \qquad \left\| \sum_{i=0}^{n} \lambda_i h_i^* + h^* + u^* \right\| \leqq \eta. \tag{4.3}$$

If z is a local solution to (1.1)–(1.3), then Theorem 2 of [3] and Proposition 3 imply the existence of $\lambda_0, \cdots, \lambda_n$ satisfying (4.2) and $r > 0$ such that

$$0 \in \sum_{i=0}^{n} \lambda_i \partial\phi_i(z, 0) + r(\partial\psi(z, 0) + \partial\rho(z, 0)). \tag{4.4}$$

Let

$$M^*_{\eta r}(x) = -\min_{\Gamma_\eta}\left(\sum_{i=0}^{n}\lambda_i\phi_i^*(x, h_i^*) + r(\psi^*(x, h^*/r) + \rho^*(x, u^*/r))\right).$$

Remark 3. Strictly speaking, we should write $\Gamma_\eta(\phi_0, \cdots, \phi_n, \psi, \rho)$ and $M^*_{\eta r}(\phi, \cdots, \phi_n. \psi, \rho, x)$, which, however, would be too awkward. Hopefully, our simplified notations will lead to no confusion.

Applying Proposition 5 to M_r and g_r and using standard formulas for composite conjugate functions (which justify in particular writing "minimum", not "infimum" in the definition of $M^*_{\eta r}$), we come to the following characterization for solutions in (1.1)-(1.3).

THEOREM 1. *Let $\phi_0, \cdots, \phi_n, \psi, \rho$ be LMO-approximations for $f_0, \cdots, f_n$, $\|F(\cdot)\|$, $d_S(\cdot)$ respectively at z. Then the following two assertions are true.*

(a) *Let z be a regular point of F relative to S. If z is a local solution to* (1.1)–(1.3), *then for any $\eta > 0$ and any sufficiently large $r > 0$, the function $M^*_{\eta r}$ attains local minimum at z.*

(b) *If for some $\eta > 0$, $r > 0$, there are $\lambda_0, \cdots, \lambda_n$ such that* (4.2) *and* (4.4) *hold and $M^*_{\eta r}$ attains strict local minimum at z, then z is an isolated local solution to* (1.1)–(1.3).

This theorem is of course an exact dual equivalent to the reduction theorem of [3]. It allows us, however, to develop various conditions, say, by choosing various approximations, and should be considered more suitable for applications thereby.

Example 6. The function $M^*_{\eta r}$ assumes especially simple and natural form if we consider approximations as in Example 4.

Let $A_0, \cdots, A_n, C, D$ be locally effective sets for $f_0, \cdots, f_n, \|F(\cdot)\|, d_S(\cdot)$ respectively at z. Then

$$\begin{aligned} &\phi_i(x, h) = f_i(x, h) + s(h, A_i), \qquad i = 0, \cdots, n;\\ &\psi(x, h) = \|F(x)\| + s(h, C);\\ &\rho(x, h) = d_S(x) + s(h, D)\end{aligned} \tag{4.5}$$

are LMO-approximations for the corresponding functions. We have

$$\phi_i^*(x, h^*) = \begin{cases} -f_i(x), & \text{if } h^* \in A_i,\\ \infty, & \text{otherwise,}\end{cases} \tag{4.6}$$

etc. Therefore in this case

$$M^*_{\eta r}(x) = \max_{\Theta_{\eta r}}\left(\sum_{i=0}^{n}\lambda_i f_i(x)\right) + r(\|F(x)\| + d_s(x)),$$

where $\Theta_{\eta r}$ consists of all vectors $(\lambda_0, \cdots, \lambda_n)$ satisfying (4.2) and

$$0 \in \sum_{i=0}^{n}\lambda_i A_i + r(C + D) + B_\eta. \tag{4.7}$$

Here B_η denotes the ball of radius η around the origin.

Example 7. The result of the previous example can be further simplified in the case $\dim Y = m < \infty$. In this case, we can identify Y with R^m and set

$$F(x) = (f_{n+1}(x), \cdots, f_{n+m}(x)) \tag{4.8}$$

$$\|F(x)\| = |f_{n+1}(x)| + \cdots + |f_{n+m}(x)|. \tag{4.9}$$

If A_i is locally effective for f_i, then $-A_i$ is obviously locally effective for $-f_i$, hence

$$\psi_i(x, h)=\max\{f_i(x)+s(h, A_i), -f_i(x)+s(h, -A_i)\}$$

is an LMO-approximation for $|f_i|$ at z, hence

$$\psi(x, h)=\psi_{n+1}(x, h)+\cdots+\psi_{n+m}(x, h)$$

is an LMO-approximation for $\|F(\cdot)\|$ at z. We have

$$\psi_i^*(x, h^*)=\min\{-\alpha f_i(x)+(1-\alpha)f_i(x)|, 0\leqq\alpha\leqq 1, h^*\in\alpha A_i-(1-\alpha)A_i\}$$

(setting $\mu=2\alpha-1$)

$$=\min\{-\mu f_i(x)|\,|\mu|\leqq 1, h^*\in\mu A_i\}.$$

It follows that

$$\psi^*(x, h^*)=\min\left\{\sum_{i=n+1}^{n+m}\psi_i^*(x, h_i^*)\,\middle|\,\sum_{i=n+1}^{n+m}h_i^*=h^*\right\}$$

$$=\min\left\{-\sum_{i=n+1}^{n+m}\mu_i f_i(x)\,\middle|\,|\mu_i|\leqq 1, h^*\in\sum_{i=n+1}^{n+m}\mu_i A_i\right\}.$$

Thus in the case being considered,

$$M_{\eta r}^*=\max_{\Omega_{\eta r}}\mathscr{L}(\lambda_0, \cdots, \lambda_{n+m}, x)+rd_S(x), \tag{4.10}$$

where

$$\mathscr{L}(\lambda_0, \cdots, \lambda_{n+m}, x)=\lambda_0 f_0(x)+\cdots+\lambda_{n+m}f_{n+m}(x)$$

is the Lagrangian of the problem and $\Omega_{\eta r}$ contains those collections of multipliers $(\lambda_0, \cdots, \lambda_{n+m})$ which satisfy (4.2), $|\lambda_i|\leqq r$ for $i=n+1, \cdots, n+m$ and

$$0\in\sum_{i=0}^{n+m}\lambda_i A_i+rD+B_\eta \tag{4.11}$$

(D and B_η being the same as in Example 6).

Example 8. Another important particular situation arises in the case of F being strictly differentiable at z. In this case the most natural LMO-approximation for $F(\cdot)$ is (Example 3)

$$\psi(x, h)=\|F(x)+F'(z)h\|$$

so that

$$\psi^*(x, h^*)=-\max\{\langle y^*, F(x)\rangle\,|\,\|y^*\|\leqq 1, F'^*(z)y^*=h^*\}$$

and $M_{\eta r}^*$ assumes the form

$$M_{\eta r}^*(x)=\max_{\Delta_{\eta r}}\left(\langle y^*, F(x)\rangle-\sum_{i=0}^{n}\lambda_i\phi_i^*(x, h_i^*)-r\rho^*(x, u^*/r)\right),$$

where $\Delta_{\eta r}$ is the collection of $(2n+4)$-tuples $(\lambda_0, \cdots, \lambda_n, h_0^*, \cdots, h_n^*, y^*, u^*)$ such that $\|y^*\|\leqq r$ and $(\lambda_0, \cdots, \lambda_n, h_0^*, \cdots, h_n^*, F'^*(z)y^*, u^*)\in\Gamma_\eta$.

If in addition, ϕ_i and ρ are given by (4.5), then $M_{\eta r}^*$ is defined by formula (4.10), just the same as in the previous examples, but with

$$\mathscr{L}(\lambda_0, \cdots, \lambda_n, y^*, x)=\lambda_0 f(x)+\cdots+\lambda_n f_n(x)+\langle y^*, F(x)\rangle$$

and $\Omega_{\eta r}$ consisting of vectors $(\lambda_0, \cdots, \lambda_n, y^*)$ satisfying (4.2), $\|y^*\| \leqq r$ and

$$-F'^*(z)y^* \in \sum_{i=0}^{n} \lambda_1 A_i + rD + B_\eta.$$

Example 9. Consider again the foregoing situation assuming for simplicity that $S = X$, so that $d_S(x) \equiv 0$ and $\rho(x, h) \equiv 0$, and apply approximations $\phi_{i\varepsilon}$, etc., described in Example 5. We have

$$\phi^*_{i\varepsilon}(x, h^*) = \begin{cases} \phi^*_i(x, h^*), & \text{if } h^* \in \partial_\varepsilon \phi_i(x, 0), \\ \infty, & \text{otherwise.} \end{cases}$$

Therefore only those λ_i, h^*_i, y^* are needed to calculate $M^*_{\eta r}$ which belong to

$$\Lambda_{\eta\varepsilon r}(x) = \{(\lambda_0, \cdots, \lambda_n, h^*_0, \cdots, h^*_h, y^*, u^*) \in \Delta_{\eta r} \,|\, u^* = 0, h^*_i \in \partial_\varepsilon \phi_i(x, 0), i = 0, \cdots, n\}$$

($\rho^*(x, h^*)$ differs from infinity only if $h^* = 0$ and $\rho^*(x, 0) = 0$, hence we can take $u^* = 0$). Thus

$$M^*_{\eta r}(x) = \max_{\Lambda_{\eta\varepsilon r}(x)} \left(\langle y^*, F(x) \rangle - \sum_{i=0}^{n} \lambda_i \phi^*_i(x, h^*_i) \right).$$

Here $M^*_{\eta r}$ depends on ε too, so that $M^*_{\eta\varepsilon r}$ might be a better notation.

It remains to note that if $F'(z)$ maps X onto Y (which implies regularity according to the Lusternik theorem), any set $\Lambda_{\eta\varepsilon}(x) = \Lambda_{\eta\varepsilon\infty}(x)$ is compact in the weak* topology and we may replace $M^*_{\eta\varepsilon r}$ by

$$\varphi_{\eta\varepsilon}(x) = \max_{\Lambda_{\eta\varepsilon}(x)} \left(\langle y^*, F(x) \rangle - \sum_{i=0}^{n} \lambda_i \phi^*_i(x, h^*_i) \right).$$

This is just the function introduced by Levitin–Miljutin–Osmolovskii in [7].

We conclude this section with several remarks concerning the theorem and the examples.

Remark 4. As it follows from the Young–Fenchel inequality,

$$M^*_{\eta r}(x) \leqq M_r(x), \quad \forall x. \tag{4.12}$$

Thus any series of functions disposed between $M^*_{\eta r}$ and M_r will discriminate solutions as well. This suggests further possibilities to develop conditions for local minima. For instance under suitable assumptions ensuring regularity, functions

$$\max_{\Omega_{\eta\infty}} \mathscr{L}(\lambda_0, \cdots, x) + r d_S(x)$$

can be considered in the situations of Examples 7 and 8.

Remark 5. As it follows from the proof, the theorem will still hold if we assume that one of the approximations, say $\rho(x, h)$, is not everywhere finite and continuous in h. Indeed, the conjugate to the sum is still the infimal convolution of the conjugate functions (with the infimum being attained) if one of the functions, not more, is not continuous. The rest follows from Remark 2.

For instance, if we have a tangent LMO-approximation W for S at z, then we can take

$$\rho(x, h) = \begin{cases} d_S(x), & h \in W(x), \\ \infty, & \text{otherwise.} \end{cases}$$

Easy calculation shows that, in this case

$$M^*_{\eta r}(x) = -\min_{\bar{\Gamma}_\eta(x)} \left(\sum_{=0}^{n} \lambda_i \phi_i^*(x, h_i^*) + r\psi^*(x, h^*/r) \right) + rd_S(x)$$

where

$$\bar{\Gamma}_\eta(x) = \{(\lambda_0, \cdots, \lambda_n, h_0^*, \cdots, h_n^*, h^*) \mid (\lambda_0, \cdots, h^*, u^*) \in \Gamma_\eta \text{ for some } u^* \in W^0(x)\},$$

$W^0(x)$ being the polar cone to $W(x)$.

5. Normality. In both the reduction theorem of [3] and Theorem 1, one can easily notice a gap between the necessity and sufficiency parts: the latter includes the demand that z be an isolated local minimum for the corresponding function. To remove this gap, additional assumptions connected with the concept of normality are needed.

We shall say that the problem (1.1)–(1.3) is

(a) *normal* at z if $\lambda_0 > 0$ whenever $\lambda_0, \ldots, \lambda_n$ satisfy (4.2), $k > 0$ and

$$0 \in \sum_{i=0}^{n} \lambda_i \partial f_i(z) + k\partial \|F(z)\| + N_S(z); \tag{5.1}$$

(b) *strongly normal* at z if there are sets $A_0, \cdots, A_n, C, D$ which are locally effective for $f_0, \cdots, f_n, \|F(\cdot)\|, d_S(\cdot)$ respectively at z and such that $\lambda_0 > 0$ whenever $\lambda_0, \cdots, \lambda_n$ satisfy (4.2), $k > 0$ and

$$0 \in \sum_{i=0}^{n} \lambda_i A_i + k(C + D). \tag{5.2}$$

Let $\phi_0, \cdots, \phi_n, \psi, \rho$ be the same as in the preceding section. For any $\eta > 0$, $r > 0$, we consider the set N_η containing all $(2n+3)$ tuples $(\lambda_1, \cdots, \lambda_n, h_0^*, \cdots, h_n^*, h^*, u^*)$ which satisfy

$$\lambda_i \geqq 0, \qquad \lambda_i f_i(z) = 0, \qquad i = 1, \cdots, n; \tag{5.3}$$

$$\|h_0^* + \lambda_1 h_1^* + \cdots + \lambda_n h_n^* + h^* + u^*\| \leqq \eta, \tag{5.4}$$

and the function

$$D_{\eta r}(x) = -\inf_{N_\eta} (\phi_0^*(x, h_0^*) + \lambda_1 \phi_1(x, h_1^*) + \cdots + \lambda_n \phi_n^*(x, h_n^*) + r(\psi^*(x, h^*/r) + \rho^*(x, u^*/r))).$$

Consider also the function

$$P_r(x) = f_0(x) + r\left(\sum_{i=1}^{n} f_i^+(x) + \|F(x)\| + d_S(x) \right),$$

where $f^+ = \max(f, 0)$.

THEOREM 2. *Assume that z is a regular point for F relative to S and that the problem* (1.1)–(1.3) *is strongly normal at z. Then the following conditions are equivalent*:

(a) *z is a local solution to* (1.1)–(1.3);

(b) *P_r attains local minimum at z if r is sufficiently large*;

(c) *$D_{\eta r}$ attains local minimum at z for any $\eta > 0$ and sufficiently large $r > 0$, whenever $\phi_0, \cdots, \phi_n, \psi, \rho$ are LMO-approximations for $f_1, \cdots f_n, \|F(\cdot)\|, d_S(\cdot)$ at z and* (5.1) *holds for certain $\lambda_0, \cdots, \lambda_n$ satisfying* (4.2) *and $k > 0$.*

Remark 6. Results like equivalence of (a) and (b) were proved earlier (Howe [2], Pietrzykowski [8], Zangwill [11]) for finite dimensional smooth or convex problems in the presence of certain constraint qualifications.

Remark 7. Theorems 1 and 2 relate differently to the condition (4.1). Though it yields no theoretical restrictions, any result depending on this condition is rather difficult to use in practice because a priori knowledge of the minimal value of the function to be minimized is needed. Theorem 1 as well as the reduction theorem of [3] relies essentially on (4.1). On the other hand, Theorem 2 does not depend on the convention because both $D_{\eta r}$ and P_r are linear in f_0.

Proof of Theorem 2. (a)$\Rightarrow$(b): Let the sets $A_0, \cdots, A_n, C, D$ be locally effective for $f_0, \cdots, f_n, \|F(\cdot)\|, d_S(\cdot)$ respectively at z. Since our problem is strongly normal, there is another collection of locally effective sets, say $A_0', \cdots, A_n', C', D'$, such that $\lambda_0 > 0$ whenever (4.2) and (5.2) (the latter for primed sets) are satisfied. Taking, if necessary, $A_0 \cap A_0', \cdots$ instead of A_0, we can assume that $A_0 \subset A_0'$ etc.

Define ϕ_i, ψ and ρ by (4.5). Since (a) holds, the function

$$M_{\eta r}^*(x) = \max_{\Theta_{\eta r}} \sum_{i=0}^{n} \lambda_i f_i(x) + r(\|F(x)\| + d_S(x))$$

attains local minimum at z (Example 6) for any $\eta > 0$ and sufficiently large $r > 0$.

It is easy to see that $\lambda_0 > 0$ whenever (4.2), (4.7) are satisfied and η is sufficiently small. Moreover, we can find $\eta > 0$ and $\varepsilon > 0$ such that $\lambda_0 \geqq \varepsilon$ whenever $\lambda_0, \cdots, \lambda_n$ and r satisfy (4.2), (4.7).

Fix some x. If $f_0(x) \geqq 0$, then

$$\begin{aligned} M_{\eta r}^*(x) &\leqq f_0(x) + \max_{\Theta_{\eta r}} \sum_{i=1}^{n} \lambda_i f_i(x) + r(\|F(x)\| + d_S(x)) \\ &\leqq f_0(x) + \sum_{i=1}^{n} f_i^+(x) + r(\|F(x)\| + d_S(x)). \end{aligned}$$

If $f_0(x) < 0$, then

$$\begin{aligned} M_{\eta r}^*(x) &\leqq \varepsilon f_0(x) + \max_{\Theta_{\eta r}} \sum_{i=1}^{n} \lambda_i f_i(x) + r(\|F(x)\| + d_S(x)) \\ &\leqq \varepsilon f_0(x) + \sum_{i=1}^{n} f_i^+(x) + r(\|F(x)\| + d_S(x)). \end{aligned}$$

These two inequalities immediately imply (b) since $P_r(z) = M_{\eta r}^*(z) = 0$.

(b)$\Rightarrow$(c): Let $I = \{i \in \{1, \cdots, n\} \mid f_i(z) = 0\}$. Clearly,

$$Q_k(x) = f_0(x) + k\Big(\sum_{i \in I} f_i^+(x) + \|F(x)\| + d_S(x)\Big) = P_k(x)$$

in a neighborhood of z. Therefore Q_k attains local minimum at z if (b) holds.

By Proposition 3

$$g(x, h) = \phi_0(x, h) + k\Big(\sum_{i \in I} \phi_i^+(x, h) + \psi(x, h) + \rho(x, h)\Big)$$

is an LMO-approximation for Q_k at z if so are $\phi_0, \cdots, \rho$ for $f_0, \cdots, d_S(\cdot)$.

By Proposition 5

$$0 \in \partial g(z, 0) \tag{5.5}$$

and for every $\eta > 0$

$$\varphi_\eta(x) = -\min\{g^*(x, h^*) \mid \|h^*\| \leqq \eta\}$$

attains local minimum at z.

We have for $i \in I$

$$\partial\phi_i^+(z, 0) = \{\mu x^* \mid 0 \leqq \mu \leqq 1, x^* \in \partial\phi(z, 0)\};$$

hence (5.5) implies the existence of such $\mu_i \in [0, 1]$, $i \in I$, that

$$0 \in \partial\phi_0(z, 0) + k\Big(\sum_{i \in I} \mu_i \, \partial\phi_i(z, 0) + \partial\psi(z, 0) + \partial\rho(z, 0)\Big),$$

which implies the second part of (c).

Furthermore

$$(\phi_i^+)^*(x, u^*) = \min\{\mu\phi_i^*(x, h^*) \mid 0 \leqq \mu \leqq 1, \mu h^* = u^*\};$$

hence

$$g^*(x, w^*) = \min\Big\{\phi_0^*(x, h_0^*) + k \sum_{i \in I} \mu_i\phi_i^*(x, v_i^*/k) + k(\psi^*(x, h^*/k) + \rho^*(x, u^*/k))\Big\}$$

with the minimum being calculated over the set of all μ_i, v_i^* $(i \in I)$, h^*, u^* satisfying

$$0 \leqq \mu_i \leqq 1, \qquad h_0^* + \sum_{iI} \mu_i v_i^* + h^* + u^* = w^*.$$

Taking $\lambda_i = k\mu_i$, $h_i = v_i^*/k$ for $i \in I$, $\lambda_i = 0$ for $i \notin I$, we see that $\varphi_\eta(x)$ is just $D_{\eta k}(x)$. This proves (c).

(c)$\Rightarrow$(a): First we note that $D_{\eta r}(z) = f_0(z)$ (by virtue of (5.1)). On the other hand, if $f_i(x) \leqq 0$, $i = 1, \cdots, n$, $F(x) = 0$, $x \in S$, then

$$D_{\eta r}(x) \leqq f_0(x) + \sum_{i=1}^{n} \lambda_i f_i^+(x) + r(\|F(x)\| + d_S(x)) = f_0(x);$$

hence $f_0(x) \geqq f_0(z)$ for all admissible x lying in a neighborhood of z.

To conclude the section, we shall demonstrate that strong normality is not an exotic property: for the cases Y being finite dimensional or F being strictly differentiable it coincides with normality if suitable criteria for regularity are fulfilled. On the other hand, normality follows from simple constraint qualifications extending such a renowned condition as that of Slater.

PROPOSITION 8. *Let Y be finite dimensional so that F is defined by* (4.8). *Assume that*

$$(5.6) \qquad 0 \in \sum_{i=n+1}^{n+m} \lambda_i \partial f_i(z) + N_S(z) \quad \textit{implies} \quad \lambda_{n+1} = \cdots = \lambda_{n+m} = 0.$$

Then the problem (1.1)–(1.3) *is strongly normal at z if it is normal at z.*

Proof. Note to begin with, that the set conv $(A \cup (-A))$ is locally effective for $|f|$ if A is locally effective for f. Therefore

$$C = \sum_{i=n+1}^{n+m} \operatorname{conv}(A_i \cup (-A_i))$$

is locally effective for $\|F(\cdot)\|$ whenever the A_i are locally effective for f_i.

Take nonincreasing sequences $\{A_{0k}\}, \cdots, \{A_{n+m,k}\}, \{D_k\}$ of sets locally effective for $f_0, \cdots, f_{n+m}, d_S$ such that

$$\bigcap_{k=0}^{\infty} A_{ik} = \partial f_i(z), \qquad i = 1, \cdots, n+m, \qquad \bigcap_{k=0}^{\infty} D_k = \partial d_S(z), \tag{5.7}$$

and set

$$C_k = \sum_{i=n+1}^{n+m} \operatorname{conv}(A_{ik} \cup -A_{ik})). \tag{5.8}$$

Then the C_k are locally effective for $\|F(\cdot)\|$ at z.

Assume that the statement is false. Then for any $k = 1, 2, \cdots,$ there are numbers $\lambda_{1k} \geqq 0, \cdots, \lambda_{nk} \geqq 0, r_k > 0$ such that

$$0 \in \sum_{i=1}^{n} \lambda_{ik} A_{ik} + r_k(C_k + D_k), \qquad \sum_{i=1}^{n} \lambda_{ik} = 1.$$

By virtue of (5.8), this means that for any $k = 1, 2, \cdots,$ there are $x_{ik}^* \in A_{ik}, i = 1, \cdots, n+m, \mu_{ik}(|\mu_{ik}| \leqq 1), i = n+1, \cdots, n+m$ and $u^* \in D_k$ such that

$$0 = \sum_{i=1}^{n+m} \lambda_{ik} x_{ik}^* + r_k u_k^*$$

where $\lambda_{ik} = r_k \mu_{ik}$ for $i = n+1, \cdots, n+m$.

Any of the sequences $\{x_{ik}^*\}, \{u_k^*\}$ is bounded, hence weakly* precompact, and limit points of the sequences belong to $\partial f_i(z), \partial d_S(z)$ respectively, in view of (5.7).

By definition, the sequence $\{\lambda_{ik}\}$ is bounded if $i \leqq n$. If this is true for any $i = 1, \cdots, n+m$, then the sequence of

$$r_k u_k^* = -\sum_{i=1}^{n+m} \lambda_{ik} x_{ik}^*$$

is also bounded. In this case, no loss of generality will follow if we assume that for any $i = 1, \cdots, n+m$, λ_{ik} converges to some λ_i as $k \to \infty$. Let x_i^* be a limit point for the sequence $\{x_{ik}^*\}$, and let u^* be a limit point for $\{r_k u_k^*\}$. Then

$$0 = \sum_{i=1}^{n+m} \lambda_i x_i^* + u^*. \tag{5.9}$$

But $\lambda_i \geqq 0$ for $i = 1, \cdots, n$, $\lambda_1 + \cdots + \lambda_n = 1$, $x_i^* \in \partial f_i(z)$ and $u^* \in N_S(z)$ (because of the last equality in (5.7)). Thus (5.9) contradicts the fact that the problem is normal at z.

It follows that we must assume that

$$\gamma_k = \max_i |\lambda_{ik}| \to \infty \quad \text{if } k \to \infty.$$

Denote $\mu_{ik} = \lambda_{ik}/\gamma_k$, $t_k = r_k/\gamma_k$. Then

$$\begin{gathered} 0 = \sum_{i=1}^{n+m} \mu_{ik} x_{ik}^* + t_k u_k^*; \\ |\mu_{ik}| \leqq 1, \quad \forall i = 1, \cdots, n+m, \quad \forall k = 1, 2, \cdots, \\ \mu_{ik} \to 0 \quad \text{as } k \to \infty, \quad \text{for } i = 1, \cdots, n. \end{gathered} \tag{5.10}$$

As above, we may assume that for $i=n+1,\cdots,n+m$, μ_{ik} also converges to some μ_i. Obviously,

$$\max\{|\mu_i|\,|\,n+1\leqq i\leqq n+m\}=1. \tag{5.11}$$

Again we see that the sequence $\{t_k u_k^*\}$ is bounded. Let x_i^* be a limit point for $\{x_{ik}^*\}$, $i=1,\cdots,n+m$, and let u^* be a limit point for $\{t_k u_k^*\}$. Then $x_i^*\in\partial f_i(z)$, $u^*\in N_S(z)$ and (5.10) implies that

$$0=\sum_{i=n+1}^{n+m}\mu_i x_i^*+u^*,$$

which together with (5.11) contradicts (5.6). Q.E.D.

PROPOSITION 9. *Let F be strictly differentiable at z. Assume that there is $D\subset X$ which is locally effective for $d_S(\cdot)$ at z and such that*

$$C(F'(z),K_D)=\sup_{\|y\|\leqq 1}\inf\{\|h\|\,|\,h\in K_D,\ F'(z)h=y\}=c<\infty, \tag{5.12}$$

where

$$K_D=\{h\in X\,|\,\langle x^*,h\rangle\geqq 0,\ \forall x^*\in D\}.$$

Then the problem (1.1)–(1.3) *is strongly normal whenever it is normal.*

Proof. Since F is strictly differentiable at z, the set

$$C=\{x^*\in X^*\,|\,x^*=F'^*(z)y^* \text{ for some } \|y^*\|\leqq 1\}$$

is locally effective for $\|F(\cdot)\|$ at z.

Let A_{ik}, D_k satisfy (5.7) and, in addition, $D_k\subset D$.

If the statement is false, then we can find sequences $\{\lambda_{ik}\}$, $\{x_{ik}^*\}$, $\{r_k\}$, $\{y_k^*\}$, and u_k^* $(i=1,\cdots,n,\ k=1,2,\cdots)$ such that

$$\lambda_{ik}\geqq 0,\quad \sum_{i=1}^n\lambda_{ik}=1,\quad \|y_k^*\|\leqq r_k,\qquad k=1,2,\cdots,$$

$$x_{ik}^*\in A_{ik},\quad u_k^*\in D_k,\quad \forall i=1,\cdots,n,\quad \forall k=1,2,\cdots,$$

and

$$0=\sum_{i=1}^n\lambda_{ik}x_{ik}^*+F'^*(z)y_k^*+r_k u_k^*. \tag{5.13}$$

If the norms of y_k^* are bounded, such $\lambda_i\geqq 0, \sum_{i=1}^n\lambda_i=1$, $x_i^*\in\partial f_i(z)$, $y^*\in Y^*$ and $u^*\in N_S(z)$ exist that

$$0=\sum_{i=1}^n\lambda_i x_i^*+F'^*(z)y^*+u^*,$$

which is in contradiction with the normality assumption. (The proof is exactly the same as above.)

Hence this is not the case and we must suppose that $\|y_k^*\|\to\infty$. Denote $\mu_{ik}=\lambda_{ik}/\|y_k^*\|$, $v_k^*=y_k^*/\|y_k^*\|$, $t_k=r_k/\|y_k^*\|$. Then

$$h_k^*=\sum_{i=1}^n\mu_{ik}x_{ik}^*$$

norm converges to zero, $\|v_k^*\|=1$ and

$$0=F'^*(z)v_k^*+t_ku_k^*+h_k^*$$

or in other words,

$$\|F'^*(z)v_k^*+t_ku_k^*\|\to 0.$$

On the other hand, $u_k^*\in D_k\subset D$, because of which $\langle u_k^*, h\rangle \geqq 0$ for any $h\in K_D$. Therefore by (5.12)

$$\begin{aligned}\|F'^*(z)v_k^*+t_ku_k^*\| &\geqq \sup\{\langle F'^*(z)v_k^*, h\rangle \mid \|h\|\leqq 1, h\in K_D\}\\ &=\sup\{\langle v_k^*, F'(z)h\rangle \mid \|h\|\leqq 1, h\in K_D\}\\ &\geqq \frac{\|v_k^*\|}{2C(F'(z), K_D)}=\frac{1}{2c}\end{aligned}$$

and our hypothesis that $\|y_k^*\|\to\infty$ also proves to be wrong. This completes the proof.

Note (see [4]) that (5.12) is surely satisfied if the null-space of $F'(z)$ meets the interior of $T_S(z)$, the tangent cone to S at z, and the set-valued mapping $x\to T_S(x)$ is lower semicontinuous from the norm topology into the norm topology.

The concluding result, to follow, gives an elementary but, hopefully convenient, sufficient condition for the problem to be normal.

PROPOSITION 10. *Assume that z is a local solution to* (1.1)–(1.3) *and there is an $h\in X$ such that*

$$f_i^0(z, h)<0, \quad \textit{for } i=1,\cdots,n,$$
$$\|F(\cdot)\|^0(z, h)=0, \qquad d_S^0(z, h)=0.$$

Then the problem is normal at z.

REFERENCES

[1] F. H. CLARKE, *A new approach to Lagrange multipliers*, Math. of Operations Res., 1 (1976), 165–174.

[2] S. M. HOWE, *New conditions for exactness of a simple penalty function*, this Journal, 11 (1973), pp. 378–381.

[3] A. D. IOFFE, *Necessary and sufficient conditions for a local minimum*, 1: *A reduction theorem and first other conditions*, this Journal, 17 (1979), pp. 245–250.

[4] ———, *Regular points of Lipschitz mappings*, Trans. Amer. Math. Soc., to appear.

[5] A. D. IOFFE AND V. M. TIKHOMIROV, *Theory of Extremal Problems*, "Nauka", Moscow, 1974; English translation, North-Holland, 1978.

[6] G. LEBOURG, *Valeur moyenne pour gradient généralisé*, C. R. Acad. Sci. Paris Ser. A, 281(1975), pp. 795–797.

[7] E. S. LEVITIN, A. A. MILJUTIN AND N. P. OSMOLOVSKII, *On conditions for a local minimum in a problem with constraints*, Mathematical Economics and Functional Analysis, B. S. Mitjagin, ed., "Nauka", Moscow, 1974. (In Russian.)

[8] T. PIETRZYKOWSKI, *The potential method for conditional maxima in the locally compact metric spaces*, Numer. Math., 14 (1970), pp. 325–329.

[9] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.

[10] J. WARGA, *Controllability and necessary conditions* in *unilateral problems without differentiability assumptions*, this Journal, 14 (1976), pp. 546–573.

[11] W. I. ZANGWILL, *Non-linear programming via penalty functions*, Management Sci., 13 (1967), pp. 344–358.

[12] H. HALKIN, *Mathematical programming without differentiability*, Calculus of Variations and Control Theory, David L. Russel, ed., Academic Press, New York, 1976, pp. 279–288.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 2, March 1979

NECESSARY AND SUFFICIENT CONDITIONS FOR A LOCAL MINIMUM. 3: SECOND ORDER CONDITIONS AND AUGMENTED DUALITY*

A. D. IOFFE†

Abstract. The paper contains second order necessary and sufficient conditions for a minimum, exact smooth penalty theorems and augmented Lagrangian duality theorems which cover the case when more than one collection of Lagrange multipliers exist and the second order quadratic function can be lower estimated only by norms weaker than that in which the cost function and constrained mappings are differentiable.

1. Introduction. In what follows, X, Y, $Z, \cdots$ are Banach spaces, f, $g, \cdots$ are functions and F, $G, \cdots$ are mappings. By $\langle \cdot , \cdot \rangle$ we denote the canonical pairing between a Banach space X and its topological dual $X.^*$ If $F\colon X \to Y$ is a mapping, then $F'(x)$ will denote the Frechět derivative of F at x, $F''(x)$ will denote the second Frechět derivative at x and $F''(x)(u, h)$ will be the corresponding bilinear mapping from $X \times X$ into Y. If F depends on more than one variable, say $F = F(x, u)$, then the first and second derivatives of F with respect to x will be denoted by F_x and F_{xx} respectively, etc.

As it follows from the title, our purpose is to discuss second order conditions for a minimum and augmented duality theorems. This paper is almost independent of the first two parts of the work [5], [6]. The core of the approach developed here is, however, the same: variational analysis is applied first to an unconstrained problem which contains those arising after the reduction of the constrained problem we want to study.

The basic problem which will be studied here is

$$\text{minimize } f(x) = g(G(X)), \tag{1.1}$$

where g is a sublinear function on Y and $G\colon X \to Y$ will be assumed continuously differentiable in a neighborhood of a given point $z \in X$ as many trimes as necessary for each particular result.

We shall see in § 7 that this class of problems contains those unconstrained problems which appear after the reduction theorem of [5] has been applied to smooth problems with equality and finitely many inequality constraints.

To outline the nature of the results presented in the paper and to discuss briefly interrelations of the results with those already known, we consider for a while the problem

$$\text{minimize } f_0(x) \tag{1.2}$$

subject to

$$f_i(x) \leqq 0, \qquad i = 1, \cdots, n. \tag{1.3}$$

Let

$$\mathcal{L}(x, \lambda_0, \cdots, \lambda_n) = \lambda_0 f_0(x) + \lambda_1 f_1(x) + \cdots + \lambda_n f_n(x).$$

* Received by the editors January 16, 1976, and in revised form January 11, 1978.

† c/o R. T. Rockafellar, Department of Mathematics, University of Washington, Seattle, Washington 98195.

The "standard" second order sufficient condition for a point $z \in X$ to be a local solution for (1.2), (1.3) (see [3]) is that there are $k > 0$ and $\lambda_i \leqq 0$ such that

$$(1.4) \qquad \lambda_0 > 0, \qquad \lambda_i f_i(z) = 0, \qquad i = 1, \cdots, n;$$

$$(1.5) \qquad \lambda_0 f_0'(z) + \cdots + \lambda_n f_n'(z) = 0$$

(first order conditions) and

$$(1.6) \qquad \mathscr{L}_{xx}(z, \lambda_0, \cdots, \lambda_n)(h, h) \geqq k\|h\|^2$$

for all $h \in X$satisfying

$$\langle f_i'(z), h\rangle \leqq 0, \qquad i \in I_1,$$
$$\langle f_i'(z), h\rangle = 0, \qquad i \in I_2,$$

where

$$I_1 = \{i \in \{1, \cdots, n\} | f_i(z) = 0, \lambda_i = 0\},$$

and

$$I_2 = \{i \in \{1, \cdots, n\} | f_i(z) = 0, \lambda_i > 0\}.$$

(If X is finite dimensional, one can replace "$\geqq k\|h\|^2$" by "> 0" in the right-hand part of (1.6).)

This condition has two serious defects. The first is that it is far from being necessary. More precisely, after replacing $k\|h\|^2$ by 0 in the right hand part of (1.6), we shall not have, as a rule, a necessary condition, In fact, it will be necessary if only one (up to a multiplicative constant) collection of multipliers satisfying (1.4), (1.5) exists. But to guarantee the uniqueness of such a collection a priori, additional strong regularity assumptions are needed [3].

Another defect is that inequalities like (1.6) do not typically arise in many infinite dimensional problems. A typical situation is that the functions f_i are differentiable in X, but possible lower estimates for the second derivative of the Lagrangian function involve only a weaker norm in which the functions are not differentiable. (For instance, in calculus of variations, functionals are differentiable in the C-norm, but their second derivatives can be estimated from below only with the help of the L_2-norm.)

It was in the work of Levitin, Miljutin and Osmolovskii [7] that a sufficient second order condition free from the first defect was originally found. Situations similar to the two-norm discrepancy described above are often encountered in closely related areas such as variational inequalities, etc. (see [8]), and, no doubt, many specialists in extremal problems are aware of them. But we do not know a work containing a general second order result which would cover such situations.

Our first purpose is to give, for (1.1), a second order condition free from both defects. Note the role of the distance estimate (4.3), which alone makes such a condition possible. Observe also that the method of proof of the sufficient condition (Theorem 2) goes back to [7].

This condition is applied afterwards (§ 6) to derive nonconvex duality with nonlinear augmented Lagrangians. The form of the Lagrangians models the form suggested by Rockafellar [10]. We establish also an exact smooth penalty theorem generalizing an earlier result of Arrow and Solow [2].

In §§ 7, 8 we apply these results to the ordinary problem with equality and inequality constraints (with a possible infinite number of the first) thereby extending results of Arrow, Gould and Howe [1], Rockafellar [10], and Mangasarian [9]. The first two sections following the Introduction are devoted to necessary conditions, mainly to demonstrate that the sufficient conditions established thereafter are "almost necessary". Technically, the necessary conditions are direct corollaries of those given in [6].

In what follows, we denote

$$\mathscr{L}(x, y^*) = \langle y^*, G(x)\rangle$$

and call this function the *Lagrangian of the problem* (1.1). We shall see that the role of this function is quite the same as the role of ordinary Lagrangians in constrained problems.

2. A general necessary and sufficient condition. If $f(x)$ attains local minimum at z then $0 \in \partial f(z)$; in other words, there is $y^* \in \partial g(G(z))$ such that $G'^*(z)y^* = 0$. This is obviously the same as

$$(2.1) \qquad y^* \in \partial g(G(z)); \qquad \mathscr{L}_x(z, y^*) = 0.$$

The set of all y^* satisfying (2.1) will be denoted by Ω_0.

Let $\eta > 0$, $\varepsilon > 0$ and

$$\Omega_{\eta\varepsilon} = \{y^* \in Y^* \mid \|G'^*(z)y^*\| \leqq \eta;\ y^* \in \partial_\varepsilon g(G(z))\},$$

where $\partial_\varepsilon g(y)$ is the ε-subdifferential of g at y:

$$\partial_\varepsilon g(y) = \{y^* \mid y^* \in \partial g(0), \langle y^*, y\rangle - g(y) \geqq -\varepsilon\}.$$

In particular, we get Ω_0 if $\eta = \varepsilon = 0$.

Consider the function

$$\varphi_{\eta\varepsilon}(x) = \max\{\mathscr{L}(x, y^*) \mid y^* \in \Omega_{\eta\varepsilon}\}.$$

The maximum is obviously attained, since $\Omega_{\eta\varepsilon}$ is a closed subset of a weak* compact set $\partial g(0)$.

It is easy to see that

$$(2.2) \qquad \Omega_0 \neq \varnothing \quad \text{implies} \quad \varphi_{\eta\varepsilon}(z) = f(z), \quad \forall \eta\varepsilon.$$

Indeed, since $\Omega_{\eta\varepsilon} \subset \partial g(0)$,

$$\varphi_{\eta\varepsilon}(z) \leqq \max_{y^* \in \partial g(0)} \langle y^*, G(z)\rangle = g(G(z)).$$

On the other hand, if $y^* \in \Omega_0$ then $y^* \in \Omega_{\eta\varepsilon}$ and $y^* \in \partial g(G(z))$, that is,

$$g(G(z)) = \langle y^*, G(z)\rangle \leqq \varphi_{\eta\varepsilon}(z).$$

PROPOSITION 1. *Let G be strictly differentiable at z. Then the following conditions are equivalent*:

(a) *$f(x)$ attains a local* (*resp. strict local*) *minimum at z*;

(b) *$\Omega_0 \neq \varnothing$ and $\varphi_{\eta\varepsilon}$ attains a local* (*resp. strict local*) *minimum at z, for any $\eta > 0$, $\varepsilon > 0$*;

(c) *$\Omega_0 \neq \varnothing$ and $\varphi_{\eta\varepsilon}$ attains a local* (*resp. strict local*) *minimum at z for some $\eta > 0$, $\varepsilon > 0$.*

Proof. For any $\varepsilon > 0$, the function

$$\rho_\varepsilon(x, h) = g_\varepsilon(G'(z)h + G(x))$$

where $g_\varepsilon(y)=\sup\{\langle y^*, y\rangle | y^*\in\partial_\varepsilon g(G(z))\}$ is an LMO-approximation for f at z [6, Examples 2 and 5]. Apply Proposition 5 of [6].

3. A second order necessary condition. From now on, we assume that G is twice continuously differentiable near z.

THEOREM 1. *If f attains a local minimum at z, then*

$$\max\{\mathscr{L}_{xx}(z, y^*)(h, h)| y^*\in\Omega_0\}\geqq 0, \tag{3.1}$$

whenever $h\in X$ satisfies

$$g(G(z)+G'(z)h)\leqq g(G(z)) \tag{3.2}$$

Proof. By Proposition 1 and (2.2), $\Omega_0\neq\emptyset$ and $\varphi_{\eta\varepsilon}(x)\geqq\varphi_{\eta\varepsilon}(z)=f(z)$ in a neighborhood of z for any $\eta>0$, $\varepsilon>0$. Let $h\in X$ be given. We have (since $\Omega_{\eta\varepsilon}$ is compact)

$$\begin{aligned} f(z)&\leqq\varphi_{\eta\varepsilon}(z+th)=\max\{\langle y^*, G(z+th)\rangle | y^*\in\Omega_{\eta\varepsilon}\}\\ &=\max\left\{\left\langle y^*, G(z)+tG'(z)h+\frac{t^2}{2}\mathscr{L}_{xx}(z, y^*)(h, h)\right\rangle\middle| y^*\in\Omega_{\eta\varepsilon}\right\}+o(t^2).\end{aligned}$$

If $0\leqq t\leqq 1$ and h satisfies (3.2), then (since $\Omega_{\eta\varepsilon}\subset\partial g(0)$)

$$\langle y^*, G(z)+tG'(z)h\rangle\leqq g(G(z)+tG'(z)h)\leqq f(z).$$

These two relations yield

$$\max\{\mathscr{L}_{xx}(z, y^*)(h, h)| y^*\in\Omega_{\eta\varepsilon}\}\geqq 0 \tag{3.3}$$

for h satisfying (3.2).

By definition,

$$\bigcap_{\substack{\eta>0\\ \varepsilon>0}}\Omega_{\eta\varepsilon}=\Omega_0\neq\emptyset.$$

Since every $\Omega_{\eta\varepsilon}$ is weak* compact, it follows that for any $y\in Y$

$$\lim_{\substack{\eta\to 0\\ \varepsilon\to 0}}\max\{\langle y^*, y\rangle | y^*\in\Omega_{\eta\varepsilon}\}=\max\{\langle y^*, y\rangle | y^*\in\Omega_0\}.$$

It remains to combine this equality with (3.3).

The convex cone K_c generated by the set

$$\{h | g(G(z)+G'(z)h)\leqq g(G(z))\}$$

will be called the *critical cone* (for f at z) and its elements will be called *critical vectors*. Clearly, in Theorem 1, condition (3.2) can equivalently be replaced by $h\in K_c$. Observe also that

$$g(G(z)+G'(z)h)\geqq g(G(z))+\langle y^*, G'(z)h\rangle=g(G(z))$$

if $y^*\in\Omega_0$. Hence if $\Omega_0\neq\emptyset$, the critical cone is defined by

$$K_c=\{h\in X | g(G(z)+tG'(z)h)=g(G(z)) \text{ for some } t>0\}.$$

Remark 1. It is not difficult to verify that K_c cannot be replaced by $\{h | g'(G(z), G'(z)h)\leqq 0\}$.

Remark 2. The "standard" second order necessary condition for our case would be: there is a $y^*\in\Omega_0$ such that

$$\mathscr{L}_{xx}(z, y^*)(h, h)\geqq 0, \quad \forall h\in K_c.$$

For this condition to hold, additional regularity assumptions are needed which imply in particular that Ω_0 contains exactly one element. It is not difficult to find an example in which this condition fails to be valid whereas the conclusion of Theorem 1 holds.

Let $X=R^2$, $x=(\xi_1,\xi_2)$,

$$q_1(x)=(\xi_1)^2-(\xi_2)^2;\qquad q_2(x)=\xi_1(\xi_2-\xi_1);\qquad q_3(x)=-\xi_1(\xi_1+\xi_2)$$

and

$$f(x)=\max_i q_i(x).$$

It is easy to see that $f(x)\geqq f(0)=0$. We have

$$g(y_1,y_2,y_3)=\max_i y_i,\qquad G(x)=(q_1(x),q_2(x),q_3(x))$$

so that

$$\partial g(0)=\{(u_1,u_2,u_3)|u_i\geqq 0,\ u_1+u_2+u_3=1\}$$

and

$$G'(0)=0.$$

It follows that $\Omega_0=\partial g(0)$. On the other hand $G''(0)(h,h)=G(h)$ so that

$$\max_{\Omega_0}\mathscr{L}_{xx}(h,h)=f(h)\geqq 0$$

for all h. On the other hand, the quadratic form

$$\begin{aligned}&u_1q_1(x)+u_2q_2(x)+u_3q_3(x)\\&\quad=(u_1-u_2-u_3)(\xi_1)^2-u_1(\xi_2)^2+(u_2-u_3)\xi_1\,\xi_2\end{aligned}$$

assumes negative values either at (1, 0) or at (0, 1) whenever $u_i\geqq 0$, $u_1+u_2+u_3\neq 0$.

4. A second order sufficient condition. It is reasonable to expect after Theorem 1 that the condition: there is $k>0$ such that

$$\max\{\mathscr{L}_{xx}(z,y^*)(h,h)|y^*\in\Omega_0\}\geqq k\|h\|^2,\qquad \forall h\in K_c$$

is sufficient for z to be a local minimum for f (together with the first order condition $\Omega_0\neq\varnothing$). This is certainly the case, but as we have already mentioned in the Introduction, in many important problems, such conditions do not typically hold. Usually second order derivatives can be estimated by norms of certain other Banach spaces, not X.

Let W be another Banach space with the norm denoted by $|||\cdot|||$. We shall say that X is *densely imbedded* into W if there is a linear continuous one-to-one mapping $i: X\to W$ such that $i(X)$ is dense in W. For notational simplicity, we shall identify X and $i(X)$ and write $|||x|||$ instead of $|||i(x)|||$ so that

$$|||x|||\leqq\mu\|x\| \tag{4.1}$$

for some $\mu>0$ independent of x.

THEOREM 2. *Assume that X is densely imbedded in another Banach space W in such a way that*

(i) $$\|G''(z)(x,h)\|\leqq c|||x|||\,|||h|||,\quad\forall x\in X,\quad\forall h\in X$$

for some $c > 0$ independent of x and h, and

(ii)
$$\lim_{\substack{\|x-z\|\to 0\\ \|h\|\to 0}} |||h|||^{-2}\|G(z+h)-G(z)-G'(z)h-\tfrac{1}{2}G''(z)(h,h)\|=0.$$

Assume also that $\Omega_0 \neq \varnothing$ and there are $k>0$, $k_1>0$ such that

$$\max\{\mathcal{L}_{xx}(z,y^*)(h,h)|y^*\in\Omega_0\} \geqq k|||h|||^2, \quad \forall h\in K_c, \tag{4.2}$$

$$\rho(x,K_c)\leqq k_1(g(G(z)+G'(z)x)-g(G(z))), \quad \forall x\in X. \tag{4.3}$$

Then z is an isolated local minimum for $f(x)$.

Here $\rho(x,K_c)$ is the distance from x to K_c in the $|||\cdot|||$-norm, that is

$$\rho(x,K_c)=\inf\{|||x-u|||\,|u\in K_c\}.$$

Proof. We set, for notational simplicity,

$$A=G'(z), \qquad B=G''(z).$$

According to (ii)

$$f(z+h)=g(G(z)+Ah+\tfrac{1}{2}B(h,h))+o(|||h|||^2). \tag{4.4}$$

Let

$$a=\liminf_{\substack{\|h\|\to 0\\ h\neq 0}} |||h|||^{-2}(f(z+h)-f(z)).$$

We shall show that

$$2a\geqq \inf_{h\in K_c} |||h|||^{-2}\max\{\mathcal{L}_{xx}(z,y^*)(h,h)|y^*\in\Omega_0\}.$$

Along with (4.2), this will prove the theorem.

If $a=\infty$, there is nothing to prove. Suppose that $a<\infty$, and let the sequence $\{h_m\}$ be such that

$$\|h_m\|\to 0, \quad h_m\neq 0, \quad |||h_m|||^{-2}(f(z+h_m)-f(z))\to a.$$

First we shall verify that

$$\lim_{m\to\infty} |||h_m|||^{-1}(g(G(z)+Ah_m)-g(G(z)))=0. \tag{4.5}$$

Indeed, insofar as g is sublinear and in view of (4.4),

$$a\geqq \lim_{m\to\infty}\frac{g(G(z)+Ah_m)-g(G(z))-(1/2)g(B(h_m,h_m))}{|||h_m|||^2}.$$

But the quantity $|||h|||^{-2}B(h,h)$ is bounded on X by (i) and

$$g(G(z)+Ah)-g(G(z))\geqq 0, \quad \forall h\in X$$

since $\Omega_0\neq\varnothing$. Therefore

$$0\leqq \limsup_{m\to\infty} |||h_m|||^{-2}(g(G(z)+Ah_m)-g(G(z)))<\infty,$$

which immediately implies (4.5) because $|||h_m|||\to 0$.

By (4.3), for any m, there is $v_m\in K_c$ such that

$$|||v_m-h_m|||\leqq k_1(g(G(z)+Ah_m)-g(G(z))). \tag{4.6}$$

This yields (together with (4.5) and (4.1))

$$\lim_{m\to\infty} |||h_m|||^{-1}|||v_m - h_m||| = 0, \tag{4.7}$$

that is,

$$\lim_{m\to\infty} \frac{|||v_m|||}{|||h_m|||} = 1. \tag{4.8}$$

Denote $u_m = h_m - v_m$. Then

$$B(h_m, h_m) = b(v_m, v_m) + B(v_m, u_m) + B(u_m, v_m) + B(u_m, u_m).$$

Condition (4.8) combined with (i) and (4.7) implies the equality

$$\lim_{m\to\infty} |||h_m|||^{-2} B(v_m, u_m) = \lim_{m\to\infty} |||h_m|||^{-2} B(u_m, u_m) = 0.$$

Therefore (since $\langle y^*, G(z)\rangle = g(G(z))$ and $A^*y^* = 0$ for any $y^* \in \Omega_0$)

$$\begin{aligned} a &= \lim_{m\to\infty} |||h_m|||^{-2}[g(G(z) + Ah_m + \tfrac{1}{2}B(h_m, h_m)) - g(G(z))] \\ &\geqq \tfrac{1}{2} \lim_{m\to\infty} |||h_m|||^{-2} \max\{\langle y^*, B(h_m, h_m)\rangle | y^* \in \Omega_0\} \\ &= \tfrac{1}{2} \lim_{m\to\infty} |||h_m|||^{-2} \max\{\langle y^*, B(v_m, v_m)\rangle | y^* \in \Omega_0\} \geqq k/2. \end{aligned}$$

Remark 3. Conditions (i), (ii), do not mean, of course, that G is twice Fréchet differentiable at z with respect to the $|||\cdot|||$-norm. Moreover, G can be even discontinuous in this norm. Consider for instance the following simple example: $X = C(0, 1)$, $W = L_2(0, 1)$, $z(\cdot) = 0$ and $G: X \to R$ is defined by

$$G(x(\cdot)) = \int_0^1 x^4(t)\, dt.$$

Then G is twice Fréchet differentiable on X and (i), (ii) are obviously satisfied for $i: X \to W$ being the natural imbedding. Here $|||\cdot|||$ is the L_2-norm and, clearly, G is nowhere continuous on X with respect to this norm.

However, in the case when $\|\cdot\|$ and $|||\cdot|||$ are equivalent, conditions (i), (ii), follow automatically from the second order differentiability of G, and (4.2) coincides with the expected sufficient condition mentioned in the beginning of the section.

Remark 4. Condition (4.3) can be considered to a certain extent as a regularity condition. The nature of this condition will be clearer in § 7, where we consider a problem having a more customary form. We shall see also that, as a regularity assumption, (4.3) is very weak.

It is important to note that condition (4.3) holds for all vectors tx $(t > 1)$ if it holds for x. Indeed,

$$\rho(tx, K_c) = t\rho(x, K_c) \leqq k_1 t(g(G(z) + Ax) - g(G(z))).$$

If $t > 1$, then

$$\begin{aligned} t(g(G(z)+Ax)-g(G(z))) &= tg(G(z)+Ax)-(t-1)g(G(z))-g(G(z)) \\ &\leqq g(tG(z)+tAx-(t-1)G(z))-g(G(z)) \\ &= g(G(z)+A(tx))-g(G(z)). \end{aligned}$$

It follows that (4.3) need only be verified for those which lie in a neighborhood of the origin.

5. A theorem on exact smooth penalties. A nonnegative function $p(y)$ will be called a *penalty function* if it is equal to zero and strictly differentiable at $G(z)$ and there are $c > 0$, $\varepsilon > 0$ such that

$$(5.1) \qquad p(G(z)+G'(z)h) \geqq c\rho^2(h, K_c), \quad \text{if } \|h\| \leqq \varepsilon.$$

Let $p(y)$ be a penalty function. Consider the function

$$\mathscr{P}(x, y^*, m) = \mathscr{L}(x, y^*) + mp(G(x)),$$

which will be called the *penalty Lagrangian*. Formula

$$\mathscr{P}(x, \Omega, m) = \max_{y^* \in \Omega} \mathscr{P}(x, y^*, m)$$

extends the penalty Lagrangian to the set of all triples (x, Ω, m), where Ω is a weak*-compact subset of Y^*.

THEOREM 3. *Assume that X is densely imbedded in another Banach space W so that conditions* (i) *and* (ii) *of Theorem* 2 *are satisfied. Let $\Omega_0 \neq \varnothing$, and let Ω be a closed subset of Ω_0 such that for some $k > 0$*

$$(5.2) \qquad \max_{y^* \in \Omega} \mathscr{L}_{xx}(z, y^*)(h, h) \geqq k|||h|||^2, \quad \forall h \in K_c.$$

Then $\mathscr{P}(\cdot, \Omega, m)$ attains a local minimum at z whenever $p(y)$ is a penalty function and m is sufficiently large.

COROLLARY 3.1. *Assume the conditions of the theorem and* (4.3). *Then the function*

$$(5.3) \qquad \max_{y^* \in \Omega} \mathscr{L}(x, y^*) + m(f(x)-f(z))^2$$

attains local minimum at z if m is sufficiently large.

Proof. The function

$$p(y) = (g(y)-f(z))^2$$

is nonnegative and strictly differentiable at $y = G(z)$, and (5.1) follows from (4.3).

This corollary looks especially nice if Ω is a singleton. In this case the initial problem appears to be reducible to the minimization of a differentiable function. But also in general case, $\mathscr{P}(x, \Omega, m)$ is "smoother" than $f(x)$ if Ω is less than Ω_0.

The proof of Theorem 3 will follow from two lemmas.

LEMMA 1. *Let X be a normed space* (*not necessarily Banach*), *and let $B: X \times X \to Y$ be a bounded bilinear mapping. Assume that a weak*-compact set $\Omega \subset Y^*$ and a closed cone $K \subset X$ are given such that*

$$(5.4) \qquad \max_{y^* \in \Omega} \langle y^*, B(h, h)\rangle \geqq k\|h\|^2, \quad \forall h \in K \quad (k > 0).$$

Then

$$\max_{y^*\in\Omega}\langle y^*, B(h, h)\rangle + m\rho^2(h, K) \geqq \frac{3k}{4}\|h\|^2, \quad \forall h \in X \tag{5.5}$$

if m is sufficiently large.

Proof. Given $h\in X$, we can choose $x(h)\in K$ such that $\|h-x(h)\|\leqq 2\rho(h, K)$. Since Ω is norm bounded and B is a bounded bilinear mapping, there is $\gamma>0$ such that

$$\langle y^*, B(x, u)\rangle \leqq \gamma\|x\|\|u\|, \quad \forall x, u\in X, \quad \forall y^*\in\Omega.$$

On the other hand

$$\max_{y^*\in\Omega}\langle y^*, B(x(h), x(h))\rangle \geqq k\|x(h)\|^2$$

by (5.4). Therefore

$$\begin{aligned}\max_{\Omega}\langle y^*, B(h, h)\rangle + m\rho^2(h, K) &\geqq \max_{\Omega}\langle y^*, B(h, h)\rangle + (m/2)\|h-x(h)\|^2\\ &= \max_{\Omega}[\langle y^*, B(h-x(h), h-x(h))\rangle + \langle y^*, B(h-x(h), x(h))\rangle\\ &\qquad + \langle y^*, B(x(h), h-x(h))\rangle + \langle y^*, B(x(h), x(h))\rangle]\\ &\quad + (m/2)\|h-x(h)\|^2\\ &\geqq (m/2-\gamma)\|h-x(h)\|^2 - 2\gamma\|h-x(h)\|\,\|x(h)\| + k\|x(h)\|^2.\end{aligned}$$

If m is sufficiently large, this quantity is not less than

$$\frac{3k}{4}(\|h-x(h)\|+\|x(h)\|)^2 \geqq \frac{3k}{4}\|h\|^2. \quad \text{Q.E.D.}$$

LEMMA 2. *Under the assumptions of Theorem* 3, *the function*

$$a(x) = \max_{\Omega}\mathscr{L}(x, y^*) + m\rho^2(x-z, K_c) - (k/2)|||x-z|||^2$$

attains a local minimum at z if m is sufficiently large.

Proof. Let $y^*\in\Omega_0$. Then

$$\mathscr{L}(z, y^*) = \langle y^*, G(z)\rangle = g(G(z)); \qquad G'^*(z)y^* = 0.$$

Therefore

$$\mathscr{L}(z+h, y^*) = f(z) + \langle y^*, G''(z)(h, h)\rangle + \langle y^*, R(h)\rangle.$$

According to the condition (ii) of Theorem 2, and since Ω is bounded, we may be sure that

$$|\langle y^*, R(h)\rangle| \leqq (k/4)|||h|||^2$$

if h is sufficiently small.

On the other hand, applying Lemma 1 to X endowed with the $|||\cdot|||$-norm, we get from (5.1) that

$$\max_{\Omega}\mathscr{L}_{xx}(z, y^*)(h, h) + m\rho^2(h, K_c) \geqq \frac{3k}{4}|||h|||^2$$

if m is sufficiently large. Hence

$$a(z+h) - a(z) = \max_{\Omega}(\langle y^*, G''(z)(h, h)\rangle + \langle y^*, R(h)\rangle) + m\rho^2(h, K_c) - (k/4)|||h|||^2 \geqq 0$$

for all h sufficiently $\|\cdot\|$-close to zero if m is sufficiently large.

Theorem 3 follows immediately from Lemma 2. Indeed, $p'(G(z))=0$ since $p(\cdot)$ is a penalty function and hence

$$p(G(z+h))=p(G(z)+G'(z)h)+o(|||h|||^2).$$

Therefore

$$p(G(z+h))\geqq cp^2(h, K_c)-(k/(4m))|||h|||^2$$

if h is sufficiently small and hence

$$\mathscr{P}(z+h, \Omega, m)\geqq \mathscr{P}(z, \Omega, m)+(k/4)|||h|||^2$$

for all h close to the origin in the $\|\cdot\|$-norm if m is sufficiently large.

Remark 5. Observe that $f(x)$ does not necessarily attain local minimum at z if $\mathscr{P}(\cdot, \Omega, m)$ does. A sufficient condition for this is for instance that there is a $\gamma>0$ such that

$$p(y)\leqq \gamma(g(y)-f(z))^2.$$

Indeed, in this case

$$\begin{aligned}\mathscr{P}(x, \Omega, m)&\leqq f(x)-f(z)+m(f(x)-f(z))^2+f(z)\\ &=f(x)-f(z)+o(|f(x)-f(z)|)+f(z).\end{aligned}$$

6. Augmented Lagrangians and duality. In this section, we shall assume that $f(x)$ has the form

$$f(x)=f_0(x)+f_1(x), \tag{6.1}$$

where f_0 is twice continuously derivable,

$$f_1(x)=q(S(x)), \tag{6.2}$$

$S: X\to U$, $q: U\to R$ have the same properties as G and g and

$$f_1(z)=0. \tag{6.3}$$

This assumption involves no limitation in generality, since we can rewrite $g(G(x))$ in the form (6.1)–(6.3) by taking $f_0(x)=\langle y_0^*, G(x)\rangle$, where y_0^* is an arbitrary element of $\partial g(G(z))$, $S(x)=G(x)$ and

$$q(y)=\max\{\langle y^*-y_0^*, y\rangle | y^*\in\partial g(0)\}.$$

On the other hand, any function f satisfying (6.1)–(6.3) can be reformulated equivalently as $g(G(x))$: it suffices to take $G(x)=(f_0(x), S(x))$, $y=(\alpha, u)$ and $g(y)=\alpha+q(u)$. This reduction allows one also to reformulate all the preceding notations and results without any difficulties.

The Lagrangian function has the form

$$\mathscr{L}(x, u^*)=f_0(x)+\langle u^*, S(x))\rangle$$

and

$$\Omega_0=\{(1, u^*)|u^*\in N_0\},$$

where

$$N_0=\{u^*\in U^*\{u^*\in\partial q(S(z)), f_0'(z)+S'^*(z)u^*=0\}\},$$

so that

$$\max_{\Omega_0} \mathscr{L}_{xx}(z, u^*)(h, h) = f_0''(z)(h, h) + \max_{u^* \in N_0} \langle u^*, S''(z)(h, h)\rangle.$$

The critical cone is generated by the set

$$\{h \in X | \langle f_0'(z), h\rangle + q(S(z) + S'(z)h) \leqq 0\}.$$

Before stating the main result of this section, we shall introduce several notations and definitions.

For a given set $Q \subset U$ and a function $p(\cdot)$ on U, we shall define the *Q-conjugate* to $p(\cdot)$ by setting

$$p_Q^*(u, u^*) = \sup_{v \in Q} (\langle u^*, v\rangle - p(u - v)).$$

In what follows, Q will be a closed convex set containing 0 and $S(z)$ and such that $q(S(z)+u) \leqq 0$ whenever $u \in Q$. Such sets Q will be called *admissible*. We shall also say that $p(\cdot)$ is an *augmentation function* if $p(\cdot)$ is nonnegative, $p(0)=0$, $p(\cdot)$ is strictly differentiable at the origin and $\|u\| \to 0$ if $p(u) \to 0$. Such is for instance the function $p(u) = \phi(\|u\|)$, where ϕ is C_2, nonnegative and convex $\phi(0)=0$ and $\phi''(0) > 0$.

Consider also the function

$$\mathscr{L}(x, u^*, m) = \mathscr{L}(x, u^*) - mp_Q^*(S(x), u^*/m),$$

which we call, following Rockafellar, the *augmented Lagrangian*. (The notation $\mathscr{L}(x, u^*, m, Q, p)$ might be more precise but it is too awkward to be used.)

If $N \subset U^*$ is weak*-compact, we can also define

$$\mathscr{L}(x, N, m) = \max_{u^* \in N} \mathscr{L}(x, u^*,, m).$$

PROPOSITION 2. *Let Q be an admissible set, and let $p(\cdot)$ be an augmentation function. Then $\mathscr{L}(x, u^*, m)$ is nondecreasing in m and*

$$\mathscr{L}(z, u^*, m) = f_0(z) \quad \textit{if } u^* \in \partial q(S(z)).$$

Proof. To prove the first part, it suffices to note that

$$mp_Q^*(u, u^*/m) = \sup_{v \in Q} (\langle u^*, v\rangle - mp(u - v))$$

and that $p(\cdot)$ is nonnegative.

If $u^* \in \partial q(S(z))$ and $v \in Q$, then

$$\langle u^*, v\rangle \leqq q(S(z)+v) \leqq 0. \tag{6.4}$$

Therefore $p_Q^*(S(z), u^*/m) \leqq 0$ (again because $p(\cdot)$ is nonnegative). But taking $v = S(z)$, we have

$$\langle u^*, S(z)\rangle - mp(0) = q(S(z)) = 0,$$

hence $p_Q^*(S(z), u^*/m) = 0$ and hence by (6.3)

$$\mathscr{L}(z, u^*, m) = \mathscr{L}(z, u^*) = f_0(z).$$

In what follows, we shall usually consider those admissible sets and augmentation functions which satisfy the following *compatibility condition*: There is a real-valued function $\gamma(m)$ such that $\gamma(m)\to\infty$ if $m\to\infty$ and for any m starting with some m_0

$$mp_Q^*(S(z)+S'(z)h, u^*/m)+\gamma(m)\rho^2(h, K_c)\leqq 0 \tag{6.5}$$

for all h belonging to a neighborhood of the origin (which may depend on m), and for all $u^*\in N_0$.

THEOREM 4. *Assume that X is densely imbedded into another Banach space W so that the conditions* (i), (ii) *of Theorem* 2 *hold. Let $N_0\neq\varnothing$, and let $N\subset N_0$ be weak*-compact and such that*

$$\max_N \mathscr{L}_{xx}(z, u^*)(h, h)\geqq k|||h|||^2, \quad \forall h\in K_c \quad (k>0). \tag{6.6}$$

Suppose also that we are given an admissible set Q and an augmentation function $p(y)$ which satisfy the compatibility condition. Then the function $\mathscr{L}(\cdot, N, m)$ attains a local minimum at z if m is sufficiently large.

Proof. To prove the theorem, it suffices to show that for any $u^*\in\partial q(S(z))$,

$$\mathscr{L}(z+h, u^*, m)\geqq f_0(z)+\tfrac{1}{2}\mathscr{L}_{xx}(z, u^*)(h, h)+\gamma(m)\rho^2(h, K_c)+r(m, h)|||h|||^2, \tag{6.7}$$

where, for any m, $r(m, h)\to 0$ if $\|h\|\to 0$.

Indeed, applying Lemma 1 to the right-hand part of (6.7), we see that for any sufficiently large m, there is $\varepsilon=\varepsilon(m)$ such that $\mathscr{L}(z+h, N, m)\geqq f_0(z)+(k/2)|||h|||^2$ if $\|h\|\leqq\varepsilon(m)$. On the other hand, $\mathscr{L}(z, N, m)=f_0(z)$ by Proposition 2.

To prove (6.7), we shall show that, uniformly in $u^*\in N_0$,

$$mp_Q^*(S(z+h), u^*/m)-mp_Q^*(S(z)+S'(z)h, u^*/m)=r_1(m, h)|||h|||^2, \tag{6.8}$$

where $r_1(m, h)\to 0$ if $\|h\|\to 0$ for any fixed $m>0$. On the other hand,

$$\mathscr{L}(z+h, u^*)=f_0(z)+\tfrac{1}{2}\mathscr{L}_{xx}(z, u^*)(h, h)+r_2(h)|||h|||^2 \tag{6.9}$$

for any $u^*\in N_0$ which follows from the condition (ii) of Theorem 2. We can also rewrite (6.5) as follows:

$$-mp_Q^*(S(z)+S'(z)h, u^*/m)\geqq\gamma(m)\rho^2(h, K_c)+r_3(m, h), \tag{6.10}$$

where $r_3(m, h)=0$ if h is sufficiently small. Combining (6.8)–(6.10) and setting

$$r(m, h)=r_1(m, h)+r_2(m, h)+r_3(m, h),$$

we come to (6.7).

It remains to prove (6.8), which is equivalent to the fact that for all m and $u^*\in N_0$,

$$\lim_{\|h\|\to 0}\frac{mp_Q^*(S(z+h), u^*/m)-mp_Q^*(S(z)+S'(z)h, u^*/m)}{|||h|||^2}=0 \tag{6.11}$$

uniformly in $u^*\in N_0$.

Assume that (6.11) does not hold. Then there are $\varepsilon>0$ and sequences $\{h_n\}$ ($h_n\neq 0$) converging to zero and $\{u_n^*\}\subset N_0$ such that, say,

$$b_n=\frac{mp_Q^*(S(z+h_n), u_n^*/m)-mp_Q^*(S(z)+S'(z)h_n, u_n^*/m)}{|||h_n|||^2}\leqq-\varepsilon \tag{6.12}$$

(or $\geqq\varepsilon$, which case can be considered exactly in the same manner).

For any $n=1, 2, \cdots$ we can find $v_n\in Q$ such that

$$mp_Q^*(S(z)+S'(z)h_n, u_n^*/m)\leqq\langle u_n^*, v_n\rangle-mp(S(z)+S'(z)h_n-v_n)+\varepsilon\frac{|||h_n|||^2}{2}. \tag{6.13}$$

According to the definition,

$$mp_Q^*(S(z+h_n), u_n^*/m) \geqq \langle u_n^*, v_n\rangle - mp(S(z+h_n)-v_n), \tag{6.14}$$

and by (6.13), (6.14),

$$b_n \geqq m\frac{p(S(z)+S'(z)h_n-v_n)-p(S(z+h_n)-v_n)}{|||h_n|||^2}-\frac{\varepsilon}{2}. \tag{6.15}$$

We have, furthermore,

$$mp_Q^*(S(z)+S'(z)h_n, u_n^*/m) \geqq -mp(S(z)+S'(z)h_n) \to 0$$

since $0 \in Q$. Together with (6.4) and (6.13), this shows that

$$\langle u_n^*, v_n\rangle - mp(S(z)+S'(z)h_n - v_n) \to 0.$$

This in turn (again in view of (6.4) and the fact that $p(\cdot) \geqq 0$) implies that $p(S(z)+S'(z)h_n - v_n) \to 0$ and hence $v_n \to S(z)$, according to the definition of the augmentation function.

Insofar as $p(\cdot)$ is strictly differentiable at the origin, $\|h_n\| \to 0$ and $v_n \to S(z)$, we have, thanks to the condition (i) of Theorem 2, that

$$\begin{aligned} &p(S(z)+S'(z)h_n - v_n) - p(S(z+h_n)-v_n) \\ &\quad = o(S(z+h_n)-S(z)-S'(z)h_n) = o(|||h_n|||^2), \end{aligned}$$

which shows together with (6.15) that $\liminf b_n \geqq -\varepsilon/2$ in contradiction with (6.12). This completes the proof.

Remark 6. A sufficient condition for (6.5) to hold is that

$$\inf_{v \in Q} p(S(z)+S'(z)h - v) \geqq c\rho^2(h, K_c) \quad (c > 0)$$

(because $\langle u^*, v\rangle \leqq 0$ for $v \in Q$, $u^* \in \partial q(S(z))$). The latter inequality means that

$$\inf_{v \in Q} p(S(x)-v)$$

is a penalty function, hence by Theorem 3,

$$\mathscr{P}(x, N, m) = f_0(x) + \max_{u^* \in N} \langle u^*, S(x)\rangle + m \inf_{v \in Q} p(S(x)-v)$$

also attains local minimum at z.

Remark 7. As in the preceding section, the fact that $\mathscr{L}(x, N, m)$ attains a local minimum at z does not imply by itself that $f(x)$ also attains a local minimum at z. A sufficient condition for such a conclusion to be valid is that there are $\gamma > 0$ and $m_0 > 0$ such that

$$mp_Q^*(y, u^*/m) + \gamma(q(y) - q(S(z)))^2 \geqq 0$$

for all $u^* \in N$ and all y lying in a neighborhood of $mS(z)$. The demonstration is exactly the same as in Remark 5.

What really follows from z being a local minimum for $\mathscr{L}(x, N, m)$ is that $f_0(x)$ attains a local minimum subject to the condition $S(x) \in Q$. Indeed, if $S(x) \in Q$, then

$$mp_Q^*(S(x), u^*/m) = \sup_{v \in Q} (\langle u^*, v\rangle - mp(S(x)-v)) \geqq \langle u^*, S(x)\rangle;$$

hence

$$\mathcal{L}(x, u^*, m) = f_0(x) + \langle u^*, S(x)\rangle - mp_Q^*(S(x), u^*/m) \leqq f_0(x),$$

that is, $\mathcal{L}(x, N, m) \leqq f_0(x)$. On the other hand, $\mathcal{L}(z, N, m) = f_0(z)$ by Proposition 2.

Theorem 4 can be reformulated as a saddle-point result which might be useful for computational purpose.

THEOREM 5. *Let the assumptions of Theorem* 4 *hold. Then the inequalities*

$$\mathcal{L}(x, N, m) > \mathcal{L}(z, N, m) \geqq \mathcal{L}(z, N + u^*, m)$$

hold for all $x \neq z$ lying in a neighborhood of z and all $u^* \in U^*$ if m is sufficiently large.

Proof. Theorem 4 justifies the left-hand inequality. To prove the right, we note that for any $u^* \in U^*$,

$$\mathcal{L}(z, u^*, m) \leqq f_0(z),$$

due to the fact that $S(z) \in Q$ (see Remark 7 above). The result follows now from Proposition 2.

7. Smooth problems with equality and inequality constraints: second order conditions. In this section, we shall apply the foregoing results to the problem studied in the first two parts of our work:

$$\text{minimize } f_0(x) \tag{7.1}$$

subject to

$$f_i(x) \leqq 0, \qquad i = 1, \cdots, n; \qquad F(x) = 0, \tag{7.2}$$

where $f_i: X \to R$ and $F: X \to Y$. But in contrast with [5], [6], here we shall assume, f_i and F twice continuously differentiable. We shall also suppose that z satisfies (7.2) and

$$f_0(z) = 0. \tag{7.3}$$

This condition was discussed in [6]. Here we only note that it involves no theoretical restriction.

By the assumptions, the functions f_i and the mapping F are locally Lipschitz and hence the reduction theorem of [5] is applicable to the problem. This theorem says in particular that z is an isolated solution to (7.1), (7.2), if for some $r > 0$ the function

$$M_r(x) = \max_{0 \leqq i \leqq n} f_i(x) + r\|F(x)\|$$

attains a strict local minimum at z. But the problem of minimizing M_r is of the type studied here. To see this, it suffices to set

$$G(x) = (f_0(x), \cdots, f_n(x), F(x)): X \to R^{n+1} \times Y,$$

and

$$g_r(\alpha_0, \cdots, \alpha_n, y) = \max_{0 \leqq i \leqq n} \alpha_i + r\|y\|.$$

Then

$$M_r(x) = g_r(G(x)).$$

We see also that the Lagrangian function for M_r coincides with the traditional

Lagrangian for the problem (7.1), (7.2):

$$\mathcal{L}(x, \lambda_0, \cdots, \lambda_n, y^*) = \sum_{i=0}^{n} \lambda_i f_i(x) + \langle y^*, F(x) \rangle.$$

Let us denote by $\Omega_0(r)$ the collection of all multipliers $(\lambda_0, \cdots, \lambda_n, y^*)$ ($\lambda_i \in R$, $y^* \in Y^*$) such that

$$\lambda_i \geqq 0, \qquad i = 0, \cdots, n; \qquad \lambda_0 + \cdots + \lambda_n = 1, \qquad \|y^*\| \leqq r; \tag{7.4}$$

$$\lambda_i f_i(z) = 0, \qquad i = 1, \cdots, n; \tag{7.5}$$

$$\lambda_0 f_0'(z) + \cdots + \lambda_n f_n'(z) + F'^*(z) y^* = 0. \tag{7.6}$$

We denote also

$$\Omega_0 = \bigcup_{r>0} \Omega_0(r).$$

It follows from [5] that *under the additional assumption that $F'(z)$ maps X onto all of Y, the first order necessary condition for z to be a local solution to* (7.1), (7.2) *is that* $\Omega_0 \neq \varnothing$.

In what follows, we shall assume that this condition is fulfilled.

PROPOSITION 3. *The critical cone for M_r at z does not depend on r for large r and coincides with the set*

$$\{h \in X | \langle f_i'(z), h \rangle \leqq 0, i \in I; \; F'(z)h = 0\},$$

where

$$I = \{i \in \{0, \cdots, n\} | f_i(z) = 0\}.$$

(Note that $0 \in I$ by (7.3).)

Proof. Since $\Omega_0 \neq \varnothing$, there is an r_0 such that $\Omega_0(r_0) \neq \varnothing$. Let $(\lambda_0, \cdots, \lambda_n, y^*) \in \Omega_0(r_0)$. Then by (7.4)–(7.6),

$$\begin{aligned} &\max_{0 \leqq i \leqq n} (f_i(z) + \langle f_i'(z), h \rangle) + r_0 \|F'(z)h\| \\ &\geqq \sum_{i=0}^{n} \lambda_i (f_i(z) + \langle f_i'(z), h \rangle) + r_0 \|F'(z)h\| \\ &\geqq \sum_{i=0}^{n} \lambda_i f_i(z) + \sum_{i=0}^{n} \langle \lambda_i f_i'(z) + F'^*(z) y^*, h \rangle = 0 \end{aligned} \tag{7.7}$$

for any $h \in X$.

Let $r > r_0$. The critical cone for M_r is generated by the set

$$\{h \in X | \max_{0 \leqq i \leqq n} (f_i(z) + \langle f_i'(z), h \rangle) + r \|F'(z)h\| \leqq 0\}.$$

In view of (7.7), for any h belonging to this set, $(r - r_0)\|F'(z)h\| \leqq 0$, hence $\|F'(z)h\| = 0$ and hence $f_i(z) + \langle f_i'(z), h \rangle \leqq 0$ for all i. If h is sufficiently small, this condition is equivalent to $\langle f_i'(z), h \rangle \leqq 0$ for all $i \in I$. The rest is trivial.

From now on, we shall denote by K_c the critical cone for all M_r corresponding to large r.

Applying Theorem 1, we get:

THEOREM 6. *A second order necessary condition for z to be a local solution to* (7.1), (7.2) *is that* (*under the assumption that the range of $F'(z)$ is all of Y*)

$$\max_{\Omega_0} \mathscr{L}_{xx}(z, \lambda_0, \cdots, \lambda_n, y^*)(h, h) \geqq 0,$$

whenever $h \in K_c$.

This is the condition discovered by Levitin, Miljutin and Osmolovskii.

From now on we shall assume that X is densely inbedded in another Banach space W and conditions (i), (ii) of Theorem 2 are satisfied. We shall omit reformulating these conditions for the concrete situation being considered, which is quite trivial to do.

PROPOSITION 4. *Assume that $\Omega_0 \neq \varnothing$ and that the range of $F'(z)$ is a closed subspace of Y. Then there are $m > 0$, $r > 0$ such that*

$$\rho(h, K_c) \leqq m\Big(\max_{i \in I} \langle f_i'(z), h\rangle + r\|F'(z)h\|\Big)$$

for all $h \in X$.

Proof. As follows from the Hoffman theorem [4] and from Proposition 3 (see also (4.1)), there is a $\gamma > 0$ such that

$$\rho(h, K_c) \leqq \gamma\Big(\sum_{i \in I} \langle f_i'(z), h\rangle^+ + \|F'(z)h\|\Big), \quad \forall h \in X. \tag{7.8}$$

Let r_0 be so great that $\Omega_0(r_0) \neq \varnothing$. If

$$\max_{i \in I} \langle f_i'(z), h\rangle \geqq 0$$

then also

$$\max_{i \in I} \langle f_i'(z), h\rangle \geqq \frac{1}{n+1} \sum_{i \in I} \langle f_i'(z), h\rangle^+. \tag{7.9}$$

Combining (7.8) and (7.9), we get

$$\rho(h, K_c) \leqq \gamma(n+1)\Big(\max_{i \in I} \langle f_i'(z), h\rangle + \frac{1}{n+1}\|F'(z)h\|\Big).$$

If $\max_{i \in I}\langle f_i'(z), h\rangle < 0$, then by (7.8),

$$\rho(h, K_c) \leqq \gamma\|F'(z)h\|.$$

But, according to (7.7),

$$\max_{i \in I} \langle f_i'(z), h\rangle + r_0\|F'(z)h\| \geqq 0;$$

hence

$$\rho(h, K_c) \leqq \max_{i \in I} \langle f_i'(z), h\rangle + (\gamma + r_0)\|F'(z)h\|.$$

It remains to take

$$r = \max\{1/(n+1), \gamma + r_0\}, \qquad m = \max\{1, \gamma(n+1)\}.$$

Thus, under the assumptions of Proposition 4, the regularity condition (4.3) is satisfied for our problem.

Applying Theorem 2, we get:

THEOREM 7. *Under the assumptions that* $\Omega_0 \neq \emptyset$ *and the range of* $F'(z)$ *is closed, the second order sufficient condition for* z *to be an isolated local solution to* (7.1), (7.2) *is that there are* $r>0$, $k>0$ *such that*

$$\max_{\Omega_0(r)} \mathscr{L}_{xx}(z, \lambda_0, \cdots, \lambda_n, y^*)(h, h) \geqq k|||h|||^2 \tag{7.10}$$

for all $h \in K_c$.

This condition depends on the parameter r which is unknown beforehand and absent in the original statement of the problem. In certain situations, however, it is possible to replace $\Omega_0(r)$ by Ω_0, as follows from the proposition below.

PROPOSITION 5. *The statements*

(a) *there are* $k>0$, $r>0$ *such that* (7.10) *holds;*

(b) *there is* $k>0$ *such that*

$$\sup_{\Omega_0} \mathscr{L}_{xx}(z, \lambda_0, \cdots, \lambda_n, y^*) \geqq k|||h|||^2, \quad \forall h \in K_c \tag{7.11}$$

are equivalent under any of the following conditions:

(i) X *is finite dimensional;*

(ii) $F'(z)$ *maps* X *onto* Y;

(iii) *the range of* $F'(z)$ *is closed and the quantity in the left-hand part of* (7.11) *is finite for any* $h \in K_c$.

(Since Ω_0 can be unbounded, we must write supremum in (7.11). In fact, as we shall see, the second two of the list conditions permits replacing supremum by maximum. Note also that "k" in (a) can differ from "k" in (b).)

Proof. First we observe that (a) implies (b) since $\Omega_0(r) \subset \Omega_0$. Thus only the inverse implication is to be verified.

Assume that $\dim X < \infty$, (b) holds but (a) fails to hold. Since $\dim X < \infty$, K_c is closed by Proposition 3 and $\Omega_0(r)$ is compact, (7.10) is equivalent to

$$\max_{\Omega_0(r)} \mathscr{L}_{xx}(z, \lambda_0, \cdots, \lambda_n, y^*)(h, h) > 0, \quad \forall h \in K_c, \quad h \neq 0. \tag{7.12}$$

Insofar as (a) is not true, it follows that, for any $r>0$, there is an $h_r \in K_c$ such that $\|h_r\| = 1$ and

$$\max_{\Omega_0(r)} \mathscr{L}_{xx}(z, \lambda_0, \cdots, \lambda_n, y^*)(h_r, h_r) \leqq 0. \tag{7.13}$$

With no loss of generality, we can assume that h_r converges to some h. Obviously, $\|h\| = 1$ and $h \in K_c$. Since (b) holds, there is a $(\lambda_0, \cdots, \lambda_n, y^*) \in \Omega_0$ such that

$$\mathscr{L}_{xx}(z, \lambda_0, \cdots, \lambda_n, y^*)(h, h) > 0. \tag{7.14}$$

By definition, there is an $r_0 > 0$ such that $(\lambda_0, \cdots, \lambda_n, y^*) \in \Omega_0(r)$ for $r > r_0$. On the other hand, it follows from (7.14) that

$$\mathscr{L}_{xx}(z, \lambda_0, \cdots, \lambda_n, y^*)(h_r, h_r) > 0$$

if r is sufficiently large, which contradicts (7.13). Thus (b) actually implies (a) if (i) holds.

Assume now that $F'(z)$ maps X onto all of Y. Then $F'^*(z)$ in one-to-one and there is $\gamma>0$ such that $\|F'^*(z)y^*\|\geqq\gamma\|y^*\|$ (see, for instance, [6, Prop. 4]). The set

$$A=\left\{x^*\in X^*\,\middle|\,x^*=-\sum_{i\in I}\lambda_i f_i'(z),\,\lambda_i\geqq 0,\,\sum_{i\in I}\lambda_i=1\right\}$$

is a compact polyhedron, hence $a_0=\max\{\|x^*\|\,|x^*\in A\}<\infty$. Take r_0 so great that $\gamma r_0>a_0$. Then $F'^*(z)y^*$ cannot be equal to an element of A whenever $\|y^*\|\geqq r_0$. It follows that $\Omega_0=\Omega_0(r_0)$.

Assume finally that

$$(7.15)\qquad a(h)=\sup_{\Omega_0}\mathscr{L}_{xx}(z,\lambda_0,\cdots,\lambda_n,y^*)(h,h)<\infty$$

for all $h\in K_c$. It follows that $h\in K_c$ and $F'^*(z)u^*=0$ imply that $\langle u^*,F''(z)(h,h)\rangle=0$. Indeed, if $(\lambda_0,\cdots,\lambda_n,y^*)\in\Omega_0$ then $(\lambda_0,\cdots,\lambda_n,y^*+tu^*)\in\Omega_0$ for any $t\in R$. Therefore if $\langle u^*,F''(z),(h,h)\rangle\neq 0$ we would have

$$a(h)\geqq\mathscr{L}_{xx}(z,\lambda_0,\cdots,\lambda_n,y^*)\sup_t\langle tu^*,F''(z)(h,h)\rangle=\infty,$$

in contradiction with (7.15).

Since the range of $F'(z)$, denoted by V, has been assumed closed, it follows that $F''(z)(h,h)$ belongs to V when $h\in K_c$. This allows us to consider V instead of Y as the range space, whereby the problem appears to be reduced to the case when $F'(z)$ maps X onto the range space.

8. Smooth problems: penalty functions and augmented Lagrangians. Clearly, there are many penalty and augmentation functions and admissible sets for each particular problem. But we consider only one class of penalty functions, one class of augmentation functions and one class of admissible sets when studying the problem (7.1), (7.2). These classes are very natural as we shall see.

Everywhere in this section, $\phi(\alpha)$ will denote a function on R satisfying the condition

(C) *ϕ is convex nonnegative, twice continuously differentiable and* $\phi(0)=0$, $\phi''(0)>0$.

In particular, there is a $\gamma>0$ such that

$$(8.1)\qquad \phi(\alpha)\geqq\gamma\alpha^2\quad\text{in a neighborhood of zero.}$$

PROPOSITION 6. *Let ϕ satisfy condition* (C), *and let the range of $F'(z)$ be closed. Then*

$$p(\alpha_0,\cdots,\alpha_n,y)=\sum_{i=0}^n\phi(\alpha_i^+)+\phi(\|y\|)$$

is a penalty function for M_r if r is sufficiently large.

Proof. We have: $p(\cdots)\geqq 0$, $p(f_0(z),\cdots,f_n(z),F(z))=0$ (because of (7.3)) and p is continuously Fréchet differentiable of the origin. It remains to verify that p satisfies (5.1).

If r is sufficiently large, then the critical cone for M_r is defined by Proposition 3, and according to (7.8), there is $d>0$ such that

$$\rho^2(h, K_c)\leqq d\Big(\sum_{i\in I}((\langle f_i'(z), h\rangle^+)^2+\|F'(z)h\|^2\Big)$$
$$\leqq d\Big(\sum_{i=0}^{n}((f_i(z)+\langle f_i'(z), h\rangle)^+)^2+\|F'(z)h\|^2\Big),$$

which together with (8.1) shows that

$$\rho^2(h, K_c)\leqq\frac{d}{\gamma}\Big(\sum_{i=0}^{n}\phi((f_i(z)+\langle f_i'(z), h\rangle)^+)+\phi(\|F'(z)h\|)\Big) \tag{8.2}$$

for all h in a neighborhood of the origin. Q.E.D.

Applying Theorem 3, we get

THEOREM 8. *Assume that $\Omega_0\neq\varnothing$, the range of $F'(z)$ is closed and there is a weak*-compact set $\Omega\subset\Omega_0$ such that*

$$\max_{\Omega}\mathscr{L}_{xx}(z, \lambda_0, \cdots, \lambda_n, y^*)(h, h)\geqq k\|\|h\|\|^2, \quad \forall h\in K_c \tag{8.3}$$

$(k>0)$. *Assume also that the function $\phi(\alpha)$ on R satisfies* (C). *Then the function*

$$\mathscr{P}(x, \Omega, m)=\max_{\Omega}\mathscr{L}(x, \lambda_0, \cdots, \lambda_n, y^*)+m\Big(\sum_{i=0}^{n}\phi(f_i^+(x))+\phi(\|F(x)\|)\Big)$$

attains strict local minimum at z if m is sufficiently large.

Denote by R_- the nonpositive half-line.

PROPOSITION 7. *If ϕ satisfies* (C), *then*

(a)
$$\phi^*_{R_-}(\alpha, \lambda)=\sup_{\beta\leqq 0}(\lambda\beta-\phi(\alpha-\beta))$$
$$=\begin{cases}-\phi(\alpha), & \textit{if}\,\lambda+\phi'(\alpha)\geqq 0,\\ \lambda\alpha+\phi^*(-\lambda), & \textit{if}\,\lambda+\phi'(\alpha)<0;\end{cases}$$

(b)
$$\phi^*_{R_-}(\alpha, 0)=-\phi(\alpha^+).$$

(Here ϕ^* is the Fenchel conjugate to ϕ).

Proof. Fix α and λ and denote $s(\beta)=\lambda\beta-\phi(\alpha-\beta)$. This function is concave by (C). Therefore either $s(\beta)>s(0)$ for some $\beta<0$, in which case

$$\sup_{\beta\leqq 0}s(\beta)=\sup_{\beta}s(\beta)=\lambda\alpha+\phi^*(-\lambda),$$

or $s(\beta)\leqq s(0)$ for all $\beta<0$ and hence

$$\sup_{\beta\leqq 0}s(\beta)=s(0)=-\phi(\alpha).$$

Since $\phi'(\alpha)$ is nondecreasing, the necessary and sufficient condition for $s(\beta)\leqq s(0)$ for all $\beta<0$ is that $s'(0)\geqq 0$, which is the same as $\lambda+\phi'(\alpha)\geqq 0$. This proves (a); (b) can be verified by a direct calculation.

Let

$$a(\alpha_0, \cdots, \alpha_n, y)=\sum_{i=0}^{n}\phi(\alpha_i)+\phi(\|y\|)$$

and

$$Q=\{(\alpha_0,\cdots,\alpha_n,y)|\alpha_i\leqq 0,\ i=0,\cdots,n;\ y=0\}. \tag{8.4}$$

It is not difficult to see that $a(\cdots)$ is an augmentation function if ϕ satisfies (C) and that Q is an admissible cone.

We have (according to part (b) of Proposition 7);

$$\inf\{s(\alpha_0-\beta_0,\cdots,\alpha_n-\beta_n,y-v)|(\beta_0,\cdots,\beta_n,v)\in Q\}=p(\alpha_0,\cdots,\alpha_n,y).$$

By virtue of Remark 7 and Proposition 6, it follows that $a(\cdots)$ and Q satisfy the compatibility condition which allows one to apply Theorems 4 and 5.

To state the corresponding result, let us denote, for notational simplicity, the vectors of multipliers, $(\lambda_0,\cdots,\lambda_n,y^*)$ by ω, so that

$$\mathscr{L}(x,\omega)=\mathscr{L}(x,\lambda_0,\cdots,\lambda_n,y^*),$$

$$\mathscr{L}(x,\omega,m)=\mathscr{L}(x,\omega)-m\sum_{i=0}^{n}\phi^*_{R_-}(f_i(x),\lambda_i/m)+m\phi(\|F(x)\|),$$

$$\mathscr{L}(x,\Omega,m)=\max_{\Omega}\mathscr{L}(x,\omega,m).$$

THEOREM 9. *Let the assumptions of Theorem* 6 *hold. Then*

$$\mathscr{L}(x,\Omega,m)>\mathscr{L}(z,\Omega,m)\geqq\mathscr{L}(z,\Omega+\omega,m)$$

for any $x\neq z$ *lying in a neighborhood of* z *and all* $\omega\in R^{n+1}\times Y$, *provided* m *is sufficiently large.*

9. Smooth problems: augmented duality in presence of normality. Observe that Theorem 9 differs from the corresponding results of Rockafellar [10] and certain others even if Ω is a singleton. In all those results, the cost function f_0 enters the augmented Lagrangian in another way than do the inequality constraint functions, whereas our theorem relates equally to the cost function and the constraint functions.

Here we shall prove a result which will coincide with that of Rockafellar in the case where Ω is a singleton. This result, however, needs an additional assumption which is automatically satisfied if Ω contains exactly one element, but which is unnecessary for Theorem 9.

Consider the function

$$P_r(x)=f_0(x)+r\left(\sum_{i=1}^{n}f_i^+(x)+\|F(x)\|\right).$$

Clearly, z is a local solution to (7.1), (7.2) if P_r attains local minimum at z (even if (7.3) does not hold). Observe also that P_r has just the form (6.1)–(6.3), indeed, P_r will have been written in this form if we set

$$q(\alpha_1,\cdots,\alpha_n,y)=r(\alpha_1^++\cdots+\alpha_n^++\|y\|)$$

and

$$S(x)=(f_1(x),\cdots,f_n(x),F(x)). \tag{9.1}$$

We shall denote $(n+1)$-tuples $(\lambda_1,\cdots,\lambda_n,y^*)$ by ν and the Lagrangian function for P_r by

$$\hat{\mathscr{L}}(x,\nu)=\hat{\mathscr{L}}(x,\lambda_1,\cdots,\lambda_n,y^*)=f_0(x)+\textstyle\sum_{i=1}^{n}\lambda_if_i(x)+\langle y^*,F(x)\rangle.$$

Let

$$N_r=\{\nu=(\lambda_1,\cdots,\lambda_n,y^*)|0\leqq\lambda_i\leqq r,\lambda_if_i(z)=0,\|y^*\|\leqq r;\hat{\mathscr{L}}_x(z,\nu)=0\},$$

and

$$N_0=\bigcup_{r>0}N_r.$$

Just in the same way as in Proposition 3, we can verify that if $N_0\neq\varnothing$, the critical cone for P_r at z does not depend on r, for large r, and coincides with that for M_r:

$$K_c=\{h|\langle f_i'(z),h\rangle\leqq 0, i\in I; F'(z)h=0\},$$

where

$$I=\{0\}\cup I_0,\qquad I_0=\{i\in\{1,\cdots,n\}|f_i(z)=0\}.$$

Let a function ϕ on R be given. We set

$$\hat{\mathscr{L}}(x,\nu,m)=\hat{\mathscr{L}}(x,\nu)-m\sum_{i=1}^n\phi^*_{R_-}(f_i(x),\lambda_i/m)+m\phi(\|F(x)\|),$$

$$\hat{\mathscr{L}}(x,N,m)=\max_{\nu\in N}\hat{\mathscr{L}}(x,\nu,m).$$

THEOREM 10. *Assume that the range of $F'(z)$ is closed and there is a nonempty weak*-compact set $N\subset N_0$ such that*

(a) $\max_{\nu\in N}\hat{\mathscr{L}}_{xx}(z,\nu)(h,h)\geqq k|||h|||^2$, $\forall h\in K_c$ $(k>0)$;

(b) *there is an $\varepsilon>0$ such that $\lambda_i\geqq\varepsilon>0$ whenever $\nu=(\lambda_1,\cdots,\lambda_n,y^*)\in N$ and $\lambda_i>0$.*

Then z is a local solution to (7.1), (7.2). *Moreover, whenever $\phi(\alpha)$ satisfies* (C) *and m is sufficiently large, the inequalities*

$$\hat{\mathscr{L}}(x,N,m)>\hat{\mathscr{L}}(z,N,m)\geqq\hat{\mathscr{L}}(z,N+\nu,m)$$

hold for all $x\neq z$ lying in a neighborhood of z (*which may depend on ϕ and m*) *and all $\nu\in R^n\times y^*$.*

The additional assumption mentioned in the beginning of the section is just the condition (b). Clearly, it is automatically satisfied if N consists of a single element. There is one more case when this condition need not be verified.

Recall that the problem (7.1), (7.2) is called *normal* if $\lambda_0>0$ whenever $(\lambda_0,\cdots,\lambda_n,y^*)\in\Omega_0$. If $N_0\neq\varnothing$ and $F'(z)$ maps X onto Y; this is equivalent to N_0 being the convex closure of a finite set $\tilde{N}$. This set obviously satisfies condition (b). But any linear functional has the same maximum on $\tilde{N}$ as on N_0. Hence we have

COROLLARY 10.1. *Assume that $N_0\neq\varnothing$, $F'(z)$ maps X onto all of Y and the problem* (7.1), (7.2) *is normal. Assume also that there is $k>0$ such that*

$$\max_{N_0}\hat{\mathscr{L}}_{xx}(z,\nu)(h,h)\geqq k|||h|||^2,\quad\forall h\in K_c.$$

Then the conclusion of the theorem holds for $N=N_0$.

Proof of Theorem 10. Let us set

$$Q=\{(\alpha_1,\cdots,\alpha_n,y)\in R^n\times Y|\alpha_i\leqq 0, y=0\},$$

$$a(\alpha_1,\cdots,\alpha_n,y)=\sum_{i=1}^n\phi(\alpha_i)+\phi(\|y\|).$$

To prove the theorem, we have to verify that Q is an admissible set, $a(\cdots)$ is an augmentation function (both are trivial) which satisfy the compatibility condition. The second part of the theorem will follow then from Theorem 5, and the first part will follow from the second in view of Remark 7. (Indeed, conditions (7.2) are equivalent to $S(x)\in Q$, S being defined by (9.1).)

In our situation, the compatibility condition can be reformulated as follows: there exists a function $\gamma(m)$ such that $\gamma(m)\to\infty$ if $m\to\infty$ and

$$m\sum_{i=1}^{n}\phi^*_{R_-}(f_i(z)+\langle f_i'(z),h\rangle,\lambda_i/m)-m\phi(\|F'(z)h\|)+\gamma(m)\rho^2(h,K_c)\leqq 0 \tag{9.2}$$

for all sufficiently large m, for all $\nu\in N$ and all h in a neighborhood of the origin (maybe depending on m).

Fix some $\nu\in N$ and let $I_\nu=\{i\in I_0|\lambda_i>0\}$. Consider the set

$$K_\nu=\{h\in X\,|\,\langle f_i'(z),h\rangle\leqq 0,\ i\in I_0\backslash I_\nu;$$
$$\langle f_i'(z)h\rangle=0,\ i\in I_\nu;\ F'(z)h=0\}.$$

We claim that $K_\nu\subset K_c$. To see this, it suffices to verify that $\langle f_0'(z),h\rangle=0$ whenever $h\in K_\nu$. But since $\nu\in N_0$,

$$\langle f_0'(z),h\rangle=-\sum_{i\in I}\lambda_i\langle f_i'(z),h\rangle-\langle y^*,\ F'(z)h\rangle=0.$$

By the Hoffman theorem [4], there is a $c_1>0$ such that (see (8.1))

$$\begin{aligned}\rho^2(h,K_c)&\leqq\rho^2(h,K_\nu)\\ &\leqq c_1\Big[\sum_{i\in I_0\backslash I_\nu}(\langle f_i'(z),h\rangle^+)^2+\sum_{i\in I_\nu}\langle f_i'(z),h\rangle^2+\|F'(z)h\|^2\Big]\\ &\leqq\frac{c_1}{\gamma}\Big[\sum_{i\in I_0\backslash I_\nu}\phi(\langle f_i'(z),h\rangle^+)+\sum_{i\in I_\nu}\phi(\langle f_i'(z),h\rangle)+\phi(\|F'(z)h\|)\Big]\end{aligned} \tag{9.3}$$

for all h in a neighborhood of the origin.

If $i\in I_0\backslash I_\nu$, then $f_i(z)=0$, $\lambda_i=0$ and by the part (b) of Proposition 7

$$-\phi(\langle f_i'(z),h\rangle^+)=\phi^*_{R_-}(f_i(z)+\langle f_i'(z),h\rangle,\lambda_i/m). \tag{9.4}$$

If $i\notin I_0$, then $f_i(z)+\langle f_i'(z),h\rangle\leqq 0$ for all h in a neighborhood of the origin (which can be chosen the same for all $i\notin I_0$) and again $\lambda_i=0$. Hence for such i and h

$$0=-\phi([f_i(z)+\langle f_i'(z),h\rangle]^+)=\phi^*_{R_-}(f_i(z)+\langle f_i'(z),h\rangle,\lambda_i/m). \tag{9.5}$$

Finally, if $i\in I_\nu$, then, according to the condition (b) of the theorem, $\lambda_i\geqq\varepsilon$ and by Proposition 7

$$-\phi(\langle f_i'(z),h\rangle)=\phi^*_{R_-}(f_i(z)+\langle f_i'(z),h\rangle,\lambda_i/m),$$

whenever h is such that

$$\varepsilon\geqq m\phi'(\langle f_i'(z),h\rangle). \tag{9.6}$$

Since $\phi'(0)=0$, the latter inequality also defines a neighborhood of the origin depending on m but not depending on ν.

Substituting (9.4)–(9.6) in (9.3) and taking $\gamma(m)=m\gamma/c_1$, we get (9.2). This completes the proof.

REFERENCES

[1] K. J. Arrow, F. J. Gould and S. M. Howe, *A general saddle-point result for constrained minimization*, Math. Programming, 5 (1973), pp. 225–234.

[2] K. J. Arrow and R. M. Solow, *Gradient methods for constrained maxima with weakened assumptions*, Studies in Linear and Nonlinear Programming, K. J. Arrow, L. Hurwicz and H. Uzawa, eds., Stanford University Press, Stanford, CA, 1958, pp. 166–176.

[3] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming, Unconstrained Optimization Techniques*, John Wiley, New York, 1968.

[4] A. D. Ioffe, *Regular points of Lipschitz mappings*, Trans. Amer. Math. Soc. to appear.

[5] ———, *Necessary and sufficient conditions for a local minimum.* 1: *A reduction theorem and first order conditions*, this Journal.

[6] ———, *Necessary and sufficient conditions for local minimum.* 2: *Conditions of Levitin–Miljutin–Osmolovskii type*, this Journal.

[7] E. S. Levitin, A. A. Miljutin and N. P. Osmolovskii, *On conditions for a local minimum in a problem with constraints*, Mathematical Economics and Functional Analysis, B. S. Mitjagin, ed., "Nauka", Moscow, 1974, pp. 139–202. (In Russian.)

[8] J. L. Lions, *Controle Optimale de Systèmes Governés par des Equations aux Dérivées Partielles*, Dunod, Paris, 1968.

[9] O. L. Mangasarian, *Unconstrained Lagrangians in nonlinear programming*, this Journal, 13 (1975), pp. 772-791.

[10] R. T. Rockafellar, *Augmented Lagrangian multiplier rule and duality in nonconvex programming*, this Journal, 12 (1974), pp. 268–285.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 2, March 1979

0363-0129/79/1702-0010 $01.00/0

INVERTIBILITY OF NONLINEAR CONTROL SYSTEMS*

R. M. HIRSCHORN†

Abstract. This paper gives necessary and sufficient conditions for the invertibility of nonlinear control systems of the form $\dot{x}=A(x)+uB(x)$; $y=c(x)$, where the state space is a real analytic manifold. For invertible systems we construct nonlinear inverse systems. These results are used to study the question of functional controllability for nonlinear systems. The class of real analytic functions which can appear as outputs of a given nonlinear system is described, and a prefilter is constructed to generate the required control.

1. Introduction. A control system is invertible when the corresponding input-output map is injective. Thus given an output function one can, in theory, recover the control which was applied. There is a considerable amount of literature dealing with invertibility of linear control systems (cf. [2], [4], [5], [6], [7]). The purpose of this paper is to generalize some of these results to nonlinear systems.

We consider here systems of the form

$$(*)\qquad \begin{aligned} &\dot{x}(t)=A(x(t))+u(t)B(x(t)); \qquad x(0)=x_0\in M,\\ &y(t)=c(x(t)) \end{aligned}$$

where the state space M is a connected real analytic manifold, A, B are real analytic vector fields on M, $x\mapsto c(x)=(c_1(x), c_2(x), \cdots, c_l(x))$ is a real analytic mapping from M into R^l, and $u\in \mathcal{U}$, the class of real analytic functions from $[0,\infty)$ into R, the real numbers. One could let $c: M\to N$ be a real analytic mapping of manifolds, and our results remain valid (use coordinates in a nbhd (neighborhood) of $c(x_0)$ in N to reduce to the above case). If $x_0\in M$ and $u\in\mathcal{U}$, we denote the resulting solution of the differential equation $(*)$ by $x(t, u, x_0)$ and denote $c(x(t, u, x_0))$ by $y(t, u, x_0)$. We remark that $x(t, u, x_0)$ may not be defined for all t as A, B may not be complete.

Here $A, B\in V(M)$, the vector space over R of real analytic vector fields on M. If $(V', x_1, x_2, \cdots, x_n)$ is a coordinate system on M and $X\in V(M)$, then locally

$$X\restriction_{V'}=\sum_{i=1}^{n} a_i\frac{\partial}{\partial x_i}$$

where $a_1, \cdots, a_n$ are real analytic functions on V'. If $f\in C^\omega(M)$, the ring of real analytic functions on M, then $Xf\in C^\omega(M)$ is defined on each coordinate system $(V', x_1, \cdots, x_n)$ by $Xf(p)=\sum_{i=1}^n a_i(p)(\partial f/\partial x_i)(p)$ for $p\in V'$. Thus $V(M)$ acts on $C^\omega(M)$. We give $V(M)$ the structure of a real Lie algebra: if $X, Y\in V(M)$ the *Lie bracket* of X and Y is the vector field $[X, Y]$ and for each $f\in C^\omega(M)$, $p\in M$,

$$[X, Y]_p(f)=X_p(Yf)-Y_p(Xf) \qquad \text{(cf. [14])}$$

and we set $\mathrm{ad}_X^0Y=Y$, $\mathrm{ad}_X^nY=[X, \mathrm{ad}_X^{n-1}Y]$. Let $\mathscr{L}$ denote the Lie subalgebra of $V(M)$ generated by A and B (cf. [11], [16], [17]).

The class of systems we consider includes single-input linear, bilinear, and right-invariant systems, and has received considerable attention in the control literature (cf. [1], [3], [8], [9], [10], [11], [12], [13], [16], [17]). The main result on invertibility is Theorem 2.1. This theorem asserts that the above system is invertible if and only if $(\mathrm{ad}_A^kB)(c_j)\neq \mathbf{0}$ for some integer $k\geqq 0$ and some component c_j of the output map (here

* Received by the editors March 15, 1978, and in revised form July 7, 1978.

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

$c_j \in C^\omega(M)$, $\mathrm{ad}_A^k B \in V(M)$, and $\mathbf{0}$ refers to the zero function in $C^\omega(M)$). For single-input, single-output time-invariant linear systems this reduces to the standard criteria for invertibility (see Example 1). Theorem 3.1 considers the related problem of functional controllability, and the class of possible output functions is explicitly described.

2. Nonlinear invertibility. In this section we derive necessary and sufficient conditions for the invertibility of the nonlinear systems $(*)$. The standard linear result is worth noting. Consider the system

$$\dot{x} = Ax + bu; \qquad x(0) = x_0 \in R^n,$$
$$y = cx$$

where A is an $n \times n$ matrix over R, c is a $1 \times \mathrm{n}$ matrix, and b an $n \times 1$ matrix. This system fails to be invertible if $\exists u_1 \neq u_2$ such that the corresponding outputs are identical. Using the variation of constants formula, this condition becomes

$$c\,e^{tA}x_0 + \int_0^t c\,e^{(t-s)A}bu_1(s)\,ds \equiv c\,e^{tA}x_0 + \int_0^t c\,e^{(t-s)A}bu_2(s)\,ds,$$

or $c\,e^{tA}b \equiv 0$. Thus the system is invertible if one of cb, cAb, $cA^2b, \cdots, cA^{n-1}b$ is nonzero. Conversely, if $cA^kb = 0$ for $k = 0, 1, \cdots, n-1$, then $y(t) = c\,e^{tA}x_0$ for any choice of u; hence the system is not invertible. The *relative order* α is defined to be the least positive integer k such that $cA^{k-1}b \neq 0$, or $\alpha = \infty$ if $cA^kb = 0$ for all $k \geqq 0$. Thus in the linear case a system is invertible if and only if $\alpha < \infty$. We note that invertibility does not depend on the initial state x_0. For nonlinear systems this is not the case. Consider the linear system $\dot{x} = Ax + bu$; $x(0) = x_0$ with nonlinear output $y = c(x)$, where

$$x_0 = \begin{bmatrix} p \\ q \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$
$$c(x) = x_1^2 + x_2^2.$$

Solving the differential equation one finds that

$$y(t) = e^{2t}p^2 + v^2(t) + q^2 + 2qv(t)$$

where $v(t) = \int_0^t u(s)\,ds$. A straightforward computation shows that for different controls $u_1 \neq u_2$ one gets the same outputs if and only if $u_1 = -u_2$ and $p = 0$. Thus the system is invertible for all initial states in the state space R^2 which do not lie on the line $x_2 = 0$. Since the line $x_2 = 0$ is a subset of R^2 of Lebesgue measure zero, one could argue that, for practical purposes, this system is invertible. Further, if the system is invertible for x_0, then there exists an open nbhd of x_0 such that the system is invertible for all initial states in this nbhd. This motivates the following definitions:

DEFINITION. The nonlinear system $(*)$ is *invertible at* $x_0 \in M$ if whenever $u_1, u_2 \in \mathcal{U}$ are distinct controls,

$$y(t, u_1, x_0) \neq y(t, u_2, x_0).$$

DEFINITION. The system $(*)$ is *strongly invertible at* x_0 if there exists an open nbhd V of x_0 such that for all $x \in V$, the system is invertible at x.

DEFINITION. The system $(*)$ is *strongly invertible* if there exists an open and dense submanifold M_0 of M such that for all $x_0 \in M_0$, the system is strongly invertible at x_0.

We remark that for linear systems all of the above definitions are equivalent to the condition $\alpha < \infty$.

To define the relative order α for nonlinear systems, we note that the output map $c(x)=(c_1(x), \cdots, c_l(x))$ has the property that $c_i(\cdot)\in C^\omega(M)$ for $0\leqq i\leqq l$, and if A, B are the vector fields which describe the nonlinear system $(*)$, then $\mathrm{ad}_A^k B \in V(M)$ for all $k\geqq 0$. Thus $(\mathrm{ad}_A^k B)(c_i)\in C^\omega(M)$ for $k\geqq 0$, $0\leqq i\leqq l$, and we let $\mathbf{0}$ denote the zero function in $C^\omega(M)$.

DEFINITION. The *relative order* α of the nonlinear system $(*)$ is the least positive integer k such that $(\mathrm{ad}_A^{k-1}B)(c_i)\neq \mathbf{0}$ for some $0\leqq i\leqq l$, or $\alpha=\infty$ if $(\mathrm{ad}_A^k B)(c_i)=\mathbf{0}$ for all $k\geqq 0$, $0\leqq i\leqq l$. If $\alpha<\infty$ set i_α to be the least integer j such that $(\mathrm{ad}_A^{\alpha-1}B)(c_j)\neq \mathbf{0}$.

It turns out that $\alpha<\infty$ is a necessary and sufficient condition for strong invertibility in the nonlinear case. Since the proof of this result (Theorem 2.1) involves the construction of a left-inverse system we define inverse systems before proving Theorem 2.1.

Given an invertible system with output function $y(\cdot, u, x_0)$ one can, in theory, recover the control function $u(\cdot)$. A left-inverse system is a nonlinear system which, when driven by appropriate derivatives of y, produces $u(\cdot)$ as its output. Thus a left-inverse system provides a practical method for finding the input $u(\cdot)$ given $y(\cdot, u, x_0)$. In the time-invariant linear case an obvious choice for a left-inverse system is the system with transfer function $1/G(s)$ where $G(s)$ is the transfer function for the original system. Since $1/G(s)$ will not be a proper rational function it is necessary to input derivatives of y to add integrations to the inverse system (cf. [4], [5], [6]).

We now define a left-inverse system in the nonlinear case:

Consider the system

$$\dot z = F(z)+vG(z); \qquad z(0)=z_0\in N,$$
$$w=h(z)+vk(z)$$

where N is a real analytic manifold; $F, G\in V(N)$; $h, k\in C^\omega(N)$, and $v\in \mathcal{U}$. This system is called a *left-inverse* for the system $(*)$ if for some $0\leqq i\leqq l$,

$$w(\cdot, y_i^{(\alpha)}, z_0)=u(\cdot).$$

That is, if the input v to the inverse system is chosen to be the αth derivative of some component of the output $y(\cdot, u, x_0)$ of the original system, the left-inverse system produces $u(\cdot)$ for its output.

Note that the inverse system fails to be of the form $(*)$ because of the presence of the control in the output. The results of this section can easily be modified to allow systems with outputs

$$y(t)=c(x(t))+u(t)d(x(t))$$

where $d: M\to R^l$ is real analytic. The changes in the definition of relative order are analogous to those in the linear case, and as in the linear case, when d is nonzero the problem of invertibility becomes trivial.

If a nonlinear system is invertible then $\alpha<\infty$, $(\mathrm{ad}_A^{\alpha-1}B)(c_{i_\alpha})\neq \mathbf{0}$, and

$$M_\alpha=\{x\in M \mid (\mathrm{ad}_A^{\alpha-1}Bc_{i_\alpha})(x)\neq 0\}$$

is nonempty. Since $f=\mathrm{ad}_A^{\alpha-1}Bc_{i_\alpha}$ is a nonzero real analytic function it cannot vanish on any open subset of M and from the continuity of f it follows that M_α is an open dense subset of M, and hence a submanifold of M.

DEFINITION. If the system $(*)$ has $\alpha<\infty$ the open dense submanifold M_α of M described above will be called the *inverse submanifold* for the system.

The submanifold M_α will provide the state space for an inverse system.

THEOREM 2.1. *The nonlinear system* (*) *is strongly invertible if and only if* $\alpha < \infty$.

COROLLARY. *Suppose that the nonlinear system* (*) *is strongly invertible with relative order* α, *initial state* $x_0 \in M$, *and inverse submanifold* M_α. *If* $x_0 \in M_\alpha$ *then the system*

$$(**)\qquad \begin{aligned} \dot{z} &= F(z) + vG(z); \qquad z(0) = x_0, \\ w &= h(z) + vk(z) \end{aligned}$$

where $z \in M_\alpha$, $v \in \mathcal{U}$, $k(z) = 1/(BA^{\alpha-1}c_{i_\alpha})(z)$, $h(z) = -k(z)(A^\alpha c_{i_\alpha})(z)$, $F(z) = A(z) + h(z)B(z)$, *and* $G(z) = k(z)B(z)$, *acts as a left-inverse for the original system* (*).

Before proving Theorem 2.1 we establish the following result:

LEMMA 2.2. *Consider the system* (*) *with* $\alpha < \infty$, *and let* M_α *denote the inverse submanifold of* M, *where* $M_\alpha = \{x \in M | (\mathrm{ad}_A^{\alpha-1} Bc_{i_\alpha})(x) \neq 0\}$. *Then for all* $k \in \{1, \cdots, \alpha - 2\}$, $i \in \{1, \cdots, l\}$, $(BA^k)c_i = \mathbf{0}$, *and for all* $p \in M_\alpha$, $(BA^{\alpha-1}c_{i_\alpha})(p) \neq 0$.

Proof. Suppose $\alpha < \infty$. Then $(\mathrm{ad}_A^k B)c_i = \mathbf{0}$ for $i = 1, 2, \cdots, l$, $k = 0, 1, \cdots, \alpha - 2$, by definition of α. A simple argument shows that in general,

$$\mathrm{ad}_A^k Bc_i = (-1)^k \sum_{0 \leqq p \leqq k} (-1)^p \binom{k}{p} A^p BA^{k-p} c_i \qquad \text{(cf. p. 108 of [15])}.$$

If $\alpha = 2$, $Bc_i = \mathbf{0}$. If $\alpha > 2$, $Bc_i = \mathbf{0}$. If $\alpha > 2$ then a simple induction argument, using the above formula, shows that $(BA^k)c_i = \mathbf{0}$ for $k = 0, 1, \cdots, \alpha - 2$ and $i = 1, \cdots, l$. Using this fact, we have

$$\begin{aligned} (\mathrm{ad}_A^{\alpha-1} B)c_{i_\alpha} &= (-1)^{\alpha-1} \sum_{0 \leqq p \leqq \alpha-1} (-1)^p \binom{\alpha-1}{p} (A^p BA^{\alpha-1-p})c_{i_\alpha} \\ &= (-1)^{\alpha-1}(BA^{\alpha-1})c_{i_\alpha}. \end{aligned}$$

Since $((\mathrm{ad}_A^{\alpha-1}B)c_{i_\alpha})(q) \neq 0$ for all $q \in M_\alpha$, the proof is complete.

Proof (*Theorem* 2.1). *Necessity*: Suppose that $\alpha = \infty$. Let $u_1, u_2 \in \mathcal{U}$ be distinct controls and let $x_0 \in M$ be any initial state. Let $t \to A_t \cdot x_0$ be the integral curve for the vector field A passing through x_0. Then $x(t, u_i, x_0) \in I(\mathcal{L}_0, A_t \cdot x_0)$ for $i = 1, 2$, and those $t \in R$ for which the two trajectories are defined. Here $\mathcal{L}_0$ is the Lie subalgebra of $V(M)$ generated by $\{\mathrm{ad}_A^k B | k = 0, 1, \cdots\}$ and $I(\mathcal{L}_0, A_t \cdot x_0)$ is the unique maximal integral submanifold of M for the distribution determined by $\mathcal{L}_0$ which contains $A_t \cdot x_0$ (cf. [8], [11], [13]). It suffices to show that $\alpha = \infty$ implies that the output map c is constant on the submanifolds $I(\mathcal{L}_0, p)$ where $p \in M$, since this implies that $y(t, u_1, x_0) = c(x(t, u_1, x_0)) = c(x(t, u_2, x_0)) = y(t, u_2, x_0)$ for t sufficiently small. Since these curves are real analytic functions of t, $y(\cdot, u_1, x_0) = y(\cdot, u_2, x_0)$, and the system is not strongly invertible. Now $\alpha = \infty$ implies that $(\mathrm{ad}_A^k B)(c_i) = \mathbf{0}$ for $i = 1, 2, \cdots, l$ and $k = 0, 1, \cdots$, and it follows that $t \to c((\mathrm{ad}_A^k B)_t \cdot p)$ is a constant curve in R^l where $p \in M$, and $t \to (\mathrm{ad}_A^k B)_t \cdot p$ is an integral curve for $\mathrm{ad}_A^k B$. Since $S = \{\mathrm{ad}_A^k B | k = 0, 1, \cdots\}$ is a set of generators for the Lie algebra $\mathcal{L}_0$, it follows from Chow's theorem (cf. [8], [11], [13], [16]) that for each $q \in I(\mathcal{L}_0, p)$, $q = X_{t_1}^1 \circ X_{t_2}^2 \circ \cdots \circ X_{t_k}^k \cdot p$ where $X^i \in S$ and $t_i \in R$, and thus $c(q) = c(X_{t_1}^1 \circ X_{t_2}^2 \circ \cdots \circ X_{t_k}^k \cdot p) = c(X_{t_2}^2 \circ \cdots \circ X_{t_n}^k \cdot p) = \cdots = c(p)$.

Sufficiency. Suppose $\alpha < \infty$ and M_α is the inverse submanifold for the system (*). Since M_α is an open dense submanifold of M the system is strongly invertible if it is invertible for all $x_0 \in M_\alpha$.

Choose $x_0 \in M_\alpha$, $u \in \mathcal{U}$ and set $y(t) = y(t, u, x_0)$. Then

$$y_i^{(1)}(t) = \frac{dy_i}{dt}(t) = (Ac_i)(x(t)) + u(t)(Bc_i)(x(t))$$

and if $\alpha > 1$, $(Bc_i) = \mathbf{0}$ for $i = 1, \cdots, l$ and

$$y_i^{(2)}(t) = (A^2 c_i)(x(t)) + u(t)(BAc_i)(x(t)).$$

If $\alpha > 2$ then $BAc_i = \mathbf{0}$ for all $i = 1, \cdots, l$ by Lemma 2.2 and continuing this process we find that

$$y_{i_\alpha}^{(\alpha)}(t) = (A^\alpha c_{i_\alpha})(x(t)) + u(t)(BA^{\alpha-1}c_{i_\alpha})(x(t)),$$

where $(BA^{\alpha-1}c_{i_\alpha})(p) \neq 0$ for all $p \in M_\alpha$ from Lemma 2.2. Since $x_0 \in M_\alpha$ $\exists \varepsilon > 0$ s.t. (such that) $x(t, u, x_0) \in M_\alpha$ for $t \in [0, \varepsilon)$. We now construct a system which acts as a left-inverse for the original system. This new system, when driven by $y_{i_\alpha}^{(\alpha)}(t)$, will produce as its output $u(t)$ for $t \in [0, \varepsilon)$. Since u is real analytic, we will have a one-to-one correspondence between input functions and output functions, and the proof will be complete.

Consider the system (**) described in the Corollary to Theorem 2.1. Since $(BA^{\alpha-1}c_{i_\alpha})$ is a real analytic function on M which does not vanish on M_α, and since M_α is open in M, the system (**) is described by real analytic vector fields on M_α. To complete the proof we must show when $v(t) = y_{i_\alpha}^{(\alpha)}(t)$, the output $w(t)$ of (**) is $u(t)$. Suppose we choose $v(t) = y_{i_\alpha}^{(\alpha)}(t)$. The resulting differential equation is

$$(1) \qquad \dot{z}(t) = F(z(t)) + y_{i_\alpha}^{(\alpha)}(t)G(z(t)); \qquad z(0) = x_0.$$

In equation (1) we use the above definition for $y_{i_\alpha}^{(\alpha)}(t)$ and the definitions for the vector fields F and G, and observe that when $z(t) = x(t)$ equation (1) becomes

$$\dot{x}(t) = A(x(t)) + u(t)B(x(t)); \qquad x(0) = x_0.$$

Thus $z(t) = x(t)$ is the solution to (1) when $v(t) = y_{i_\alpha}^{(\alpha)}(t)$, and with this choice of $v(\cdot)$,

$$\begin{aligned} w(t) &= h(x(t)) + y_{i_\alpha}^{(\alpha)}(t)k(x(t)) \\ &= -k(x)(A^\alpha c_{i_\alpha})(x) + y_{i_\alpha}^{(\alpha)}(t)k(x), \end{aligned}$$

where $k(x) = 1/(BA^{\alpha-1}c_{i_\alpha})(x)$. Using the above expression for $y_{i_\alpha}^{(\alpha)}(t)$, we have $w(t) = u(t)$ and the proof is complete.

Proof (*Corollary*). Follows from the second half of the above proof.

3. Functional controllability. In this section we are concerned with determining the functions $f(t)$ which can be realized as the output of the nonlinear system (*) driven by a suitable input function. For linear systems this classification problem was solved by R. W. Brockett in 1965 [4]. He also showed that if $f(\cdot) = y(\cdot, u, x_0)$ for some control u, then the required control can be generated as the output of the left-inverse system driven by an appropriate derivative of $f(t)$. In this case we say that the left-inverse system acts as a *right-inverse* for the original system. We will show that the left-inverse system described in § 2 is also a right-inverse for nonlinear systems. We will restrict ourselves to single output systems ($l = 1$). The generalization to vector-valued output maps is straightforward.

THEOREM 3.1. *Consider the nonlinear system* (*) *with relative order* α.

If $\alpha = \infty$ *then for all initial states* $x_0 \in M$ *and* $\forall u \in \mathcal{U}$,

$$y(t, u, x_0) \equiv c(A_t \cdot x_0).$$

If $\alpha < \infty$, $x_0 \in M_\alpha$, *and* $f \in C^\omega(R)$ *then* $\exists u \in \mathcal{U}$ *such that* $y(\cdot, u, x_0) = f(\cdot)$ *if and only if*

$$f^{(k)}(0) = (A^k c)(x_0) \quad \textit{for } k = 0, 1, \cdots, \alpha - 1.$$

Proof. Suppose $\alpha = \infty$. In the proof of Theorem 2.1 we showed that $\alpha = \infty$ implies that $y(\cdot, u_1, x_0) = y(\cdot, u_2, x_0)$ for all $u_1, u_2 \in \mathcal{U}$, $\forall x_0 \in M$. Set $u_2 = \mathbf{0}$ so for all $u \in \mathcal{U}$, $y(t, u, x_0) = y(t, \mathbf{0}, x_0) = c(x(t, \mathbf{0}, x_0)) = c(A_t \cdot x_0)$, and this completes the first part of the proof.

Suppose that $\alpha < \infty$, $x_0 \in M_\alpha$, $f \in C^\omega(R)$ and $f(\cdot) = y(\cdot, u, x_0)$ for some $u \in \mathcal{U}$. Following the proof of Theorem 2.1 we differentiate $y(t, u, x_0)$ with respect to t and find that $f(0) = c(x_0)$, $f^{(1)}(0) = (Ac)(x_0) + u(0)(Bc)(x_0) = (Ac)(x_0)$ if $\alpha > 1, \cdots, f^{(\alpha-1)}(0) = (A^{\alpha-1}x)(x_0)$, as required.

Now suppose that $\alpha < \infty$, $f \in C^\omega(R)$ and $f^{(k)}(0) = (A^k c)(x_0)$ for $k = 0, 1, 2, \cdots, \alpha - 1$. If $x_0 \in M_\alpha$ we can find $u \in \mathcal{U}$ such that $y(\cdot, u, x_0) = f(\cdot)$ using the left-inverse system described in the Corollary to Theorem 2.1. Since $f^{(\alpha)}(\cdot) \in \mathcal{U}$ we can let $v(t) = f^{(\alpha)}(t)$ be the input to the inverse system (**), and set $z(t) = z(t, f^{(\alpha)}, x_0)$, and $w(t) = h(z(t)) + f^{(\alpha)}(t)k(z(t))$. If we can show that $y(\cdot, w, x_0) = f(\cdot)$ the proof will be complete, since $u = w$ is a control which produces $f(\cdot)$ as the output of the system (*).

When $u(t) = w(t)$ we know that $x(t) = x(t, w, x_0)$ satisfies the differential equation

$$\dot{x} = A(x) + wB(x); \qquad x(0) = x_0. \tag{2}$$

Claim. $z(t)$ satisfies (2): differentiating $z(t)$ we see that

$$\begin{aligned}\dot{z}(t) &= F(z(t)) + f^{(\alpha)}(t)G(z(t))\\ &= \{A(z(t)) + h(z(t))B(z(t))\} + f^{(\alpha)}(t)k(z(t))B(z(t))\\ &= A(z(t)) + \{h(z(t)) + f^{(\alpha)}(t)k(z(t))\}B(z(t))\\ &= A(z(t)) + w(t)B(z(t)),\end{aligned}$$

and $z(0) = x_0$.

Now set $y(t) = y(t, w, x_0)$. Since $y(t) = c(x(t))$ we have

$$\begin{aligned}&y(0) = c(x_0)\\ &y^{(1)}(0) = (Ac)(x_0)\\ &\quad\vdots\\ &y^{(\alpha-1)}(0) = (A^{\alpha-1}c)(x_0)\\ &y^{(\alpha)}(t) = (A^\alpha c)(x(t)) + w(t)(BA^{\alpha-1}c)(x(t))\end{aligned}$$

where $w(t) = h(z(t)) + f^{(\alpha)}(t)k(z(t))$. Since $x(\cdot) = z(\cdot)$,

$$y^{(\alpha)}(t) = (A^\alpha c)(x(t)) + \{h(x(t)) + f^{(\alpha)}(t)k(x(t))\}(BA^{\alpha-1}c)(x(t)).$$

Using the definitions for h and k,

$$\begin{aligned}&\{h(x) + f^{(\alpha)}(t)k(x)\}(BA^{\alpha-1}c)(x)\\ &\quad = \{-k(x)(A^\alpha c)(x) + f^{(\alpha)}(t)k(x)\}(1/k(x))\\ &\quad = -(A^\alpha c)(x) + f^{(\alpha)}(t), \quad \text{and} \quad y^{(\alpha)}(t) = f^{(\alpha)}(t).\end{aligned}$$

Thus $f^{(\alpha)}(t) = y^{(\alpha)}(t)$ and $y^{(k)}(0) = f^{(k)}(0)$ for $0 \leqq k \leqq \alpha - 1$. Integrating, we find that $y(\cdot, w, x_0) = f(\cdot)$, and thus $f(t)$ is the output of the original system if we set $u(t) = w(t)$. This completes the proof.

We remark that in proving this theorem we have shown that if $f(t)$ can be realized as the output of a nonlinear system for some control u, the required control can be generated as the output of the system (**) driven by $f^{(\alpha)}(t)$. We state this as a corollary.

COROLLARY. *Suppose that the nonlinear system* (*) *has* $\alpha < \infty$ *and initial state* $x_0 \in M_\alpha$. *Then the left-inverse system* (**) *acts as a right-inverse system. In particular if* $f(t) \in C^\omega(R)$ *can be realized as* $y(t, u, x_0)$ *for some control* $u \in \mathcal{U}$, *then* $f(t) = y(t, u_f, x_0)$ *where* $u_f(t) = w(t, f^{(\alpha)}, x_0)$.

4. Examples.

Example 1. Consider the linear system

$$\dot{x}(t) = Ax(t) + bu(t); \qquad x_0 \in R^n,$$
$$y(t) = cx(t)$$

where the state space is the real analytic manifold $M = R^n$, and $A, b \in V(M)$. For each $x \in M$, $A_x = Ax$ and $b_x = b$, where, as usual, $T_x(M)$ is identified with R^n for all $x \in M$. With this identification, $\forall f \in C^w(M)$, $(Af)(x) = df_x(Ax)$ where $df_x = ((\partial f/\partial x_1)(x), \cdots, (\partial f/\partial x_n)(x))$ and $(bf)(x) = df_x(b)$. Also, $\mathrm{ad}_A^k b = (-1)^k A^k b$, a constant vector field. Since $c: R^n \to R$ is linear, $dc_x = c$ and thus $(\mathrm{ad}_A^k b)(c)(x) = (-1)^k cA^k b$. In this case the relative order α is the least integer k such that $cA^{k-1}b \neq 0$, or $\alpha = \infty$ if $cA^k b = 0$ for $k = 0, 1, 2, \cdots, n-1$. This agrees with the definition of relative order in the linear case, and Theorem 2.1 generalizes the standard linear result. If $\alpha < \infty$ and the system is invertible, then $M_\alpha = M$, since $cA^{\alpha-1}b$ is independent of x. The left-inverse described in § 2 will be a linear system, and is the standard linear inverse system.

Example 2. Consider the bilinear system

$$\dot{x}(t) = Ax(t) + u(t)Bx(t); \qquad x \in R^3,$$
$$y(t) = cx(t)$$

where $M = R^3$, $u \in \mathcal{U}$, $c: R^3 \to R$ is linear, and

$$c = [1 \quad 0 \quad 1], \qquad A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

and

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

Here

$$[A, B] = BA - AB = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \mathrm{ad}_A^2 B = 0,$$

and $[B, [A, B]] = -[A, B]$. Thus $\mathcal{L}$ has a basis $\{A, B, \mathrm{ad}_A B\}$ and $\mathcal{L}_0$ has a basis $\{B, \mathrm{ad}_A B\}$, and is *not* Abelian. In this example, $(Bc)(x) = CBx = 0$ for all x, so $\alpha \neq 1$, and $(\mathrm{ad}_A B)(c)(x) = c\ \mathrm{ad}_A Bx = 0$ for all $x \in R^3$. Since $\mathrm{ad}_A^2 B = 0$, the relative order $\alpha = \infty$ and this system is not invertible. In fact, if x_0 has components (a_1, a_2, a_3), then the system equations become

$$\dot{x}_1 = x_1, \qquad x_1(0) = a_1,$$

$$\dot{x}_2 = x_3 + ux_2; \qquad x_2(0) = a_2,$$

$$\dot{x}_3 = 0; \qquad x_3(0) = a_3, \quad \text{and}$$

$$x_3(t) \equiv a_3, \quad x_1(t) \equiv a_1\, e^t, \quad x_2(t) = a_2\, e^{\int_0^t u(s)\,ds} + a_3 \int_0^t e^{\int_\tau^t u(s)\,ds}\, d\tau.$$

In this case, $y(t) = cx(t) = a_3 + a_1\, e^t$, and is independent of u for all x_0. This illustrates Theorem 2.1—A noninvertible system has an output which is unaffected by the control which is applied.

Example 3. Consider the nonlinear system

$$\text{(s)} \qquad \begin{aligned} \dot{x} &= A(x) + uB(x), \\ y &= c(x) \end{aligned}$$

where $M = R^3$, $A(x) = A(x_1, x_2, x_3) = (x_1x_2, x_1, x_2)$, $B(x) = (x_2^2, 0, 0)$ and $c(x) = e^{x_3}$. Here we identify the tangent space $T_x(R^3)$ with R^3, so for all $f \in C^\omega(R^3)$, and for all $X, Y \in V(R^3)$, $X_x = (a_1(x), a_2(x), a_3(x))$, $Y_x = (b_1(x), b_2(x), b_3(x))$ and $(Xf)(x) = \sum_{i=1}^3 a_i(x)\, \partial f/\partial x_i(x)$. Thus

$$([X, Y]f)(x) = [X, Y]_x f = \sum_{i=1}^{3} c_i(x) \frac{\partial f}{\partial x_i}(x)$$

where $[X, Y]_x = dY_xX_x - dX_xY_x = (c_1(x), c_2(x), c_3(x))$, dX_x is the Jacobian matrix $(\partial a_i/\partial x_j(x))$, and $dY_x = (\partial b_i/\partial x_j(x))$. By direct calculation we find that

$$(\mathrm{ad}_A B)_x = (2x_1x_2 - x_2^3, -x_2^2, 0)$$

and

$$(\mathrm{ad}_A^2 B)_x = (2x_1^2 - 2x_1x_2^2 - x_2^4, x_2^3 - 4x_1x_2, x_2^2).$$

Continuing we find $(\mathrm{ad}_A^x B)_x$ contains x_2^{n+2} in its first component, so that $\mathscr{L}_0$ is an infinite dimensional Lie algebra. Computing α we find that

$$(Bc)(x) = dc_xB_x = 0,$$

$$(\mathrm{ad}_A Bc)(x) = dc_x\ \mathrm{ad}_A B_x = 0 \quad \text{for all } x \in M,$$

and $(\mathrm{ad}_A^2 Bc)(x) = x_2^2\, e^{x_3} \neq \mathbf{0}$. Thus $\alpha = 3$ and this system is invertible. Since $\mathrm{ad}_A^2 Bc$ does not vanish on $\{x = (x_1, x_2, x_3) | x_2 \neq 0\}$, the inverse submanifold M_α is the complement of the plane $x_2 = 0$ in R^3, and is an open dense submanifold of R^3. If $x(0) = x_0 \in M_\alpha$, we can construct an inverse system of the form (**) and for this example $(BA^2c)(x) = x_2^2\, e^{x_3}$,

$$F(z) = (-z_2^3 - 3z_1z_2, z_1, z_2),$$

$$G(z) = (e^{-z_3}, 0, 0),$$

$$h(z) = -\left(z_2 + \frac{4z_1}{z_2}\right)$$

and

$$k(z) = \frac{e^{-z_3}}{z_2^2}.$$

From the proof of Theorem 2.1 we would expect that when

$$\begin{aligned} v(t) = y^{(\alpha)}(t) = y^{(3)}(t) &= (A^3c)(x(t)) + u(t)(BA^2c)(x(t)) \\ &= (4x_1x_2 + x_2^3)\, e^{x_3} + (x_2^2\, e^{x_3})u(t), \end{aligned}$$

we have $z(\cdot, y^{(\alpha)}, x_0)=x(\cdot, u, x_0)$ and $w(\cdot, y^{(\alpha)}, x_0)=u(\cdot)$. We leave the verification of the last equality to the reader and show that when $z(t)=x(t)$ the differential equation defining the inverse system is satisfied:

$$F(x)+y^{(3)}(t)G(x)=\begin{bmatrix}-x_2^3-3x_1x_2\\ x_1\\ x_2\end{bmatrix}+(4x_1x_2+x_2^3+ux_2^2)\,e^{x_3}\begin{bmatrix}e^{-x_3}\\ 0\\ 0\end{bmatrix}$$

$$=\begin{bmatrix}x_1x_2+ux_2^2\\ x_1\\ x_2\end{bmatrix}=A(x)+uB(x)=\dot{x},$$

as required, and since $x(0)=z(0)$, $x(\cdot, u, x_0)=z(\cdot, y^{(\alpha)}, x_0)$.

Finally, if we suppose that $x_0=(a_1, a_2, a_3)\in M_\alpha$ we can describe the range space of the input-output map defined by the system (s). From § 3 we know that $f\in C^\omega(R^3)$ is $y(t, u, x_0)$ for some $u\in\mathcal{U}$ if and only if $f(0)=c(x_0)$, $f^{(1)}(0)=(Ac)(x_0)$, and $f^{(2)}(0)=(A^2c)(x_0)$. For this example, we can conclude that $f(t)$ is an output of the system (s) for some $u\in\mathcal{U}$ if and only if $f(0)=e^{a_3}$, $f^{(1)}(0)=a_2\,e^{a_3}$ and $f^{(2)}(0)=(a_1+a_2^2)\,e^{a_3}$. In particular the output of this system can be made to track any polynomial function of the form

$$d_0+d_1d_0t+\tfrac{1}{2}(d_1^2+d_2)d_0t^2+d_3t^3+d_4t^4+\cdots+d_nt^n$$

where d_0, $d_1\neq 0$ and x_0 is chosen appropriately.

One can verify that the system (**) acts as a right-inverse for the system (s).

REFERENCES

[1] R. W. Brockett, *System theory on group manifolds and coset spaces*, this Journal, 10 (1972), pp. 265–284.

[2] ———, *Poles, zeros, and feedbacks: state space interpretation*, IEEE Trans. Automatic Control AC-10 (1965), pp. 129–135.

[3] ———, *Volterra series and geometric control theory*, Automatica, 12 (1976), pp. 167–176.

[4] R. W. Brockett and M. D. Mesarovic, *The reproducibility of multivariable systems*, J. Math. Anal. Appl., 11 (1965), pp. 548–563.

[5] M. K. Sain and J. L. Massey, *Invertibility of linear time-invariant dynamical systems*, IEEE Trans. Automatic Control, AC-14 (1969), pp. 141–149.

[6] L. M. Silverman, *Inversion of multivariable linear systems*, Ibid., AC-14 (1969), pp. 270–276.

[7] A. S. Willsky, *On the invertibility of linear systems*, Ibid., AC-19 (1974), pp. 272–274.

[8] R. M. Hirschorn, *Global controllability of nonlinear systems*, this Journal, 14 (1976), pp. 700–711.

[9] ———, *Invertibility of control system on Lie groups*, this Journal, 15 (1977), pp. 1034–1049.

[10] H. Sussmann and V. Jurdjevic, *Control systems on Lie groups*, J. Differential Equations, 16 (1972), pp. 313–329.

[11] ———, *Controllability of nonlinear systems*, Ibid., 12 (1972), pp. 95–116.

[12] G. Haynes and H. Hermes, *Nonlinear controllability via Lie theory*, this Journal, 8 (1970), pp. 450–460.

[13] C. Lobry, *Controlabilité des systèmes non linéaires*, this Journal, 8 (1970), pp. 573–605.

[14] F. Warner, *Foundations of Differentiable Manifolds and Lie Groups*, Scott, Foresman, Glenview, IL, 1970.

[15] V. S. Vardarajan, *Lie Groups, Lie Algebras, and their Representations*, Prentice-Hall, Englewood Cliffs, NJ, 1974.

[16] R. Hermann and A. Krener, *Nonlinear controllability and observability*, IEEE Trans. Automatic Control, to appear.

[17] H. Sussmann, *Existence and uniqueness of minimal realizations of nonlinear systems*, Math. Systems Theory, 10 (1977), pp. 263–284.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 2, March 1979

0363-0129/79/1702-0011 $01.00/0

COLLOCATION AT GAUSS POINTS AS A DISCRETIZATION IN OPTIMAL CONTROL*

G. W. REDDIEN†

Abstract. Collocation at Gauss points is shown to be a high order accurate discretization of certain unconstrained optimal control problems. Best possible convergence rates are established along with superconvergence results.

1. Introduction. The technique of replacing a continuous optimal control problem by a discretization in order to obtain a finite dimensional approximation problem for computational purposes is an old one. For some recent results on the convergence of such methods, see Budak, Berkovich and Solv'eva [3], Daniel [5], Mathis and Reddien [13] and Hager [8]. In a different direction, new high order and efficient methods for the numerical solution of two-point boundary value problems have been developed based on collocation using polynomial splines; see for example Russell and Shampine [16] and Lucas and Reddien [12] among many others. Particularly attractive among collocation methods are those involving collocation at Gauss points, e.g. DeBoor and Swartz [6], Cerrutti [4], Russell [17], Wittenbrink [20] and Reddien [14].

In this note we will study collocation at Gauss points as a discretization in optimal control. Such discretizations have apparently not been considered before, and, as we shall show, can be high order accurate methods. Discretizations using polynomial splines have been considered via the Ritz–Galerkin method, e.g. Bosarge et al. [1] and Falk [7], and the related Ritz–Trefftz principle, e.g. Bosarge et al. [2], Schultz [19] and Hager [9]. However, these methods generally require the use of a numerical quadrature in actual computations. The influence of such errors on high order estimates has apparently not been considered. However, see the important paper of Daniel [5] where convergence results are obtained. In a certain sense, the collocation methods given here can be viewed as quadrature approximations to the Ritz method.

In § 2, the problem and method to be studied will be defined and in § 3, basic convergence results will be given. We consider in § 3 essentially the same class of problems that was studied in [1], including the important linear-quadratic state regulator problem. In § 4, under additional assumptions, we improve the results given in § 3 and establish superconvergence results. Convergence rates given here are higher than those predicted using the same class of splines and the Ritz method [1], [7] or the Ritz–Trefftz method [19]. Our computational experience [15] indicates that the convergence rates established in those papers are not best possible. Indeed, we have been able to improve the bounds for the Ritz–Trefftz method. These results will appear elsewhere.

2. Problem and method. Define $J(u)=\int_0^1 g(x, u, t)\,dt$. We consider the problem

(P)

(a) $\min_u J(u)$, u *in* $L^2[0, 1]$,

subject to

(b) $\dot{x}(t)=f(x, u, t)$, $x(0)=x_0$, t *in* $[0, 1]$,

where $u(t)$ is an r-dimensional vector, $x(t)$ is an s-dimensional vector, $f(x, u, t)$ is an

* Received by the editors October 27, 1977, and in revised form August 8, 1978.

† Department of Mathematics, Vanderbilt University, Nashville, Tennessee 37235. This research was supported in part by the U.S. Army Research Office under Grant DAAG29-78-G-0126.

s-dimensional vector, and $g(x, u, t)$ is a scalar valued function. Let $W_2^\nu[0, 1]$ denote the usual Sobolev space of functions [1] with ν a positive integer. We will omit range designations for these functions as the range will be clear in the context used.

We will say that the problem (P) is in C^ν if and only if $f(x, u, t)$ and $g(x, u, t)$ are $\nu+1$ times continuously differentiable in x and u and ν times in t. The Lagrangian for (P) is

$$L(u, x, \lambda, \gamma) = J(u) + \int_0^1 (-\dot{x} + f(x, u, t))^T \lambda \, dt + (x(0) - x_0)^T \gamma$$

where λ is in W_2^1 and γ is a vector in R^s.

We now state three assumptions which will remain in effect throughout this paper: (A1) Problem (P) is in C^ν, $\nu \geqq 1$. (A2) The Lagrangian L is extremized at the four-tuple $(u^*, x^*, \lambda^*, \gamma^*)$ satisfying the conditions

$$\text{(2.1)} \qquad \begin{aligned} &\text{(a)} \quad \frac{\partial g^*}{\partial u} + \left(\frac{\partial f^*}{\partial u}\right)^T \lambda^* = 0, \\ &\text{(b)} \quad \dot{\lambda}^* + \left(\frac{\partial f^*}{\partial x}\right)^T \lambda^* + \frac{\partial g^*}{\partial x} = 0, \qquad \lambda^*(1) = 0, \\ &\text{(c)} \quad -\dot{x}^* + f(x^*, u^*, t) = 0, \qquad x^*(0) = x_0, \\ &\text{(d)} \quad \gamma^* = -\lambda^*(0), \end{aligned}$$

where the superscript * indicates the functions involved are evaluated at $u = u^*$ and $x = x^*$. Define the operator H by

$$H(u, x, \lambda) = \begin{bmatrix} g_{uu} + f_{uu}^T \lambda & g_{ux} + f_{ux}^T \lambda \\ g_{xu} + f_{xu}^T \lambda & g_{xx} + f_{fxx}^T \lambda \end{bmatrix}.$$

We assume (A3) that

$$\begin{bmatrix} \delta u \\ \delta x \end{bmatrix}^T H(\tilde{u}, \tilde{x}, \tilde{\lambda}) \begin{bmatrix} \delta u \\ \delta x \end{bmatrix} \geqq \sigma(|\delta u|^2 + |\delta x|^2)$$

where $\delta u = u - \tilde{u}$ and $\delta x = x - \tilde{x}$ with $(\tilde{u}, \tilde{x}, \tilde{\lambda})$ in some bounded convex neighborhood N of (u^*, x^*, λ^*) in the usual product norm on $W_2^1 \times W_2^1 \times W_2^1$ for some constant $\sigma > 0$ and uniformly in t, $0 \leqq t \leqq 1$. (A3) implies that the second variation of L with respect to (u, x) is strongly positive in N.

Remarks. Assumptions (A1), (A2) and (A3) constitute a set of local sufficiency conditions for the existence and uniqueness of a solution for problem (P). (See [1].) We will eliminate the multiplier γ from the Lagrangian by considering only variations which satisfy (2.1) (d).

The solution to (P) will be approximated using continuous polynomial splines. Let Δ_n: $0 = t_0 < t_1 < \cdots < t_n = 1$ be a partition of $[0, 1]$ with $|\Delta_n| \equiv \max_i (t_i - t_{i-1})$. Let $S(\Delta_n, p)$ denote the s-tuples of continuous polynomial splines of degree p over Δ_n (piecewise polynomials of degree $\leqq p$ on each subinterval). Let $C(\Delta_n, p)$ denote the r-tuples of continuous polynomial splines of degree p over Δ_n.

In the ith subinterval $[t_{i-1}, t_i]$ of Δ_n, let ξ_{ij}, $j = 1, \cdots, p$, denote the quadrature points for p-point Gaussian quadrature over $[t_{i-1}, t_i]$ and let w_{ij}, $j = 1, \cdots, p$, denote the associated weights. With these notations, the method we study is defined as follows:

$$
(2.2)\qquad
\begin{aligned}
&\text{(a) minimize}_{u_n} \sum_{i=1}^{n} \sum_{j=1}^{p} w_{ij} g(x_n(\xi_{ij}), u_n(\xi_{ij}), \xi_{ij})\\
&\text{(b) subject to } \dot{x}_n(\xi_{ij}) = f(x_n(\xi_{ij}), u_n(\xi_{ij}), \xi_{ij}), \qquad i = 1, \cdots, n;\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad j = 1, \cdots, p, \quad x_n(0) = x_0,
\end{aligned}
$$

where x_n is in $S(\Delta_n, p)$ and u_n is in $C(\Delta_n, p)$. In order to further simplify notation we let $\sum'_{i,j} h \equiv \sum_{i=1}^{n} \sum_{j=1}^{p} w_{ij} h(\xi_{ij})$.

The dimension of $S(\Delta_n, p)$ if $s = 1$ is $np + 1$. A Lagrange basis can be constructed as follows. Find np functions $\phi_{\alpha\beta}$, $\alpha = 1, \cdots, n$, $\beta = 1, \cdots, p$ so that $\phi_{\alpha\beta}(\xi_{ij}) = \delta_{\alpha\beta,ij}$ where $\delta_{\alpha\beta,ij} = 1$ if $\alpha = i$ and $\beta = j$ and is zero otherwise and $\phi_{\alpha\beta}(1) = 0$. Adjoin to this set a function ϕ_0 so that $\phi_0(1) = 1$ and $\phi_0(\xi_{ij}) = 0$. For $C(\Delta_n, p)$ with $r = 1$, we can use this same set as a basis. Now to form a basis for the s-tuples of splines in $S(\Delta_n, p)$, define $\phi_{ij}^k = \phi_{ij} e_k$ where $e_k = [0\ 0 \cdots 0\ 1\ 0 \cdots 0]$ and the one is in the kth position. In order to simplify notation, we will omit the superscript k and write, for example, if $\lambda_n \in S(\Delta_n, p)$ satisfies $\lambda_n(1) = 0$, then $\lambda_n = \sum_{i=1}^{n} \sum_{j=1}^{p} \alpha_{ij} \phi_{ij}$ for some set of real coefficients α_{ij}. Identical simplifications will be used for the basis for $C(\Delta_n, p)$.

3. Convergence results. The first convergence results of this paper will follow from the necessary conditions for (2.2) which we now derive.

THEOREM 3.1. *Let u_n^* and x_n^* be a solution to* (2.2). *Then there exists a function λ_n^* in $S(\Delta_n, p)$ so that*

$$
(3.1)\qquad
\begin{aligned}
&\text{(a)}\quad \dot{x}_n^*(\xi_{ij}) = f(x_n^*(\xi_{ij}), u_n^*(\xi_{ij}), \xi_{ij}),\\
&\text{(b)}\quad x_n^*(0) = x_0,\\
&\text{(c)}\quad \dot{\lambda}_n^*(\xi_{ij}) = -f_x^T(x_n^*, u_n^*, t)|_{\xi_{ij}} \lambda_n^*(\xi_{ij}) - g_x(x_n^*, u_n^*, t)|_{\xi_{ij}},\\
&\text{(d)}\quad \lambda_n^*(1) = 0,\\
&\text{(e)}\quad (g_u(x_n^*, u_n^*, t) + f_u(x_n^*, u_n^*, t)^T \lambda_n^*)|_{\xi_{ij}} = 0,\\
&\qquad\qquad\qquad\qquad\qquad\qquad i = 1, \cdots, n, \quad j = 1, \cdots, p.
\end{aligned}
$$

Proof. Equation (3.1) (a) can be written first as

$$
w_{ij} \phi_{ij}^T(\xi_{ij}) \dot{x}_n(\xi_{ij}) = w_{ij} \phi_{ij}^T(\xi_{ij}) f(x_n(\xi_{ij}), u_n(\xi_{ij}), \xi_{ij})
$$

for each i, j and then using the fact that $\phi_{ij}(\xi_{\alpha\beta}) = \delta_{ij,\alpha\beta}$ as

$$
\sum_{\alpha=1}^{n} \sum_{\beta=1}^{p} w_{\alpha\beta} \phi_{ij}^T(\xi_{\alpha\beta}) \dot{x}_n(\xi_{\alpha\beta}) = \sum_{\alpha=1}^{n} \sum_{\beta=1}^{p} w_{\alpha\beta} \phi_{ij}^T(\xi_{\alpha\beta}) f(x_n(\xi_{\alpha\beta}), u_n(\xi_{\alpha\beta}), \xi_{\alpha\beta}).
$$

Thus the Lagrangian for (2.2) can be written as

$$
L_n(u_n, x_n, \lambda_n) = \sum_{i,j}' g(x_n, u_n, t) + \sum_{i=1}^{n} \sum_{j=1}^{p} \lambda_{ij} \sum_{\alpha,\beta}' \phi_{ij}^T(-\dot{x}_n + f(x_n, u_n, t)).
$$

Define $\lambda_n = \sum_{i=1}^{n} \sum_{j=1}^{p} \lambda_{ij} \phi_{ij}$. Note this implies $\lambda(1) = 0$ from the way that the ϕ_{ij}'s were constructed. Then L_n may be written as

$$
L_n(u_n, x_n, \lambda_n) = \sum_{i,j}' g(x_n, u_n, t) + \sum_{i,j}' \lambda_n^T(-\dot{x}_n + f(x_n, u_n, t)).
$$

Note that $\lambda_n^T \dot{x}_n$ is a piecewise polynomial of degree $2p - 1$ on each subinterval of Δ_n.

Thus the quadrature formula will be exact so that $-\sum_{i,j}' \lambda_n^T \dot{x}_n = -\int_0^1 \lambda_n^T \dot{x}_n \, dt = \int_0^1 \dot{\lambda}_n^T x_n \, dt + \lambda_n(0) x_n(0)$. The discrete necessary conditions are given by

$$\text{(3.2)} \qquad \begin{aligned} &\text{(a)} \quad \frac{\partial L_n}{\partial \lambda_n} = 0, \\ &\text{(b)} \quad \frac{\partial L_n}{\partial x_n} = 0, \\ &\text{(c)} \quad \frac{\partial L_n}{\partial u_n} = 0. \end{aligned}$$

From (3.2) (a) one gets back (2.3) (b). Equation (3.2) (b) gives

$$\text{(3.3)} \qquad \sum_{r,s}' \phi_{ij}^T(g_x(x_n, u_n, t) + \dot{\lambda}_n + f_x^T(x_n, u_n, t)\lambda_n) = 0, \qquad i = 1, \cdots, n, \quad j = 1, \cdots, p,$$

and (3.2) (c) gives

$$\text{(3.4)} \qquad \sum_{r,s}' \phi_{ij}^T(g_u(x_n, u_n, t) + f_u^T(x_n, t)\lambda_n) = 0, \qquad i = 1, \cdots, n, \quad j = 1, \cdots, p.$$

Again using the fact that $\phi_{ij}(\xi_{rs}) = \delta_{ij,rs}$, equation (3.3) gives (3.1) (c) and (3.4) gives (3.1) (e), completing the proof.

We next want to establish the saddle point behavior of the discrete Lagrangian at a solution $(u_n^*, x_n^*, \lambda_n^*)$ in N. This result will imply the equivalence of dealing with the approximation problem through its formulation as a minimization problem in (2.2) or through the discrete necessary conditions in (3.1). However, this still leaves the question of existence of $(u_n^*, x_n^*, \lambda_n^*)$ in N. For the remainder of this section, we will assume (A4) that (2.2) has a solution in N. Sufficient conditions for existence are given in § 4.

We remark that the equivalent form generated in the proof of Theorem 3.1 for the Lagrangian is what would result if Gaussian quadrature were used to implement the Galerkin method.

THEOREM 3.2. *Let* (A1)–(A4) *hold. Then for all* λ_n *in* $S(\Delta_n, p)$ *satisfying* $\lambda(1) = 0$, *all* x_n *in* $S(\Delta_n, p)$ *satisfying* $x_n(0) = x_0$, *and all* u_n *in* $C(\Delta_n, p)$ *so that* (u_n, x_n, λ_n) *is in* N,

$$L_n(u_n^*, x_n^*, \lambda_n) = L_n(u_n^*, x_n^*, \lambda_n^*) \leqq L_n(u_n, x_n, \lambda_n^*).$$

Proof. The equality follows directly from (3.1) (a). Write $\Delta u_n = u_n - u_n^*$ and $\Delta x_n = x_n - x_n^*$ and expand $L_n(u, x_n, \lambda_n^*)$ to obtain

$$\text{(3.5)} \qquad \begin{aligned} L_n(u_n, x_n, \lambda_n^*) = {} & L_n(u_n^*, x_n^*, \lambda_n^*) + \sum_{i,j}' g_x(x_n^*, u_n^*, t)\Delta x_n + \sum_{i,j}' g_u(x_n^*, u_n^*, t)\Delta u_n \\ & + \sum_{i,j}' \int_0^1 \{g_{xx}(x_n^* + s\Delta x_n, u_n^*, t)\Delta x_n^2 + g_{uu}(x_n^*, u_n^* + s\Delta u_n, t)\Delta u_n^2\}(1-s)\, ds \\ & + \sum_{i,j}' \lambda_n^{*T}(-\dot{\Delta} x_n + f_x(x_n^*, u_n^*, t)\Delta x_n + f_u(x_n^*, u_n^*, t)\Delta u_n) \\ & + \sum_{i,j}' \lambda_n^{*T}(f_{xx}(x_n^* + s\Delta x_n, u_n^*, t)\Delta x_n^2 + f_{uu}(x_n^*, u_n^* + s\Delta u_n, t)\Delta u_n^2)(1-s)\, ds \\ & + \sum_{i,j}' (g_{xu}(x_n^* + s\Delta x_n, u_n^* + s\Delta u_n, t)\Delta x_n \Delta u_n \\ & \qquad + \lambda_n^{*T} f_{xu}(x_n^* + s\Delta x_n, u_n^* + s\Delta u_n, t)\Delta x_n \Delta u_n)(1-s)\, ds. \end{aligned}$$

Now using A3 and arguing as in the proof of Theorem 3.1, (3.5) becomes

$$
\begin{aligned}
L_n(u_n, x_n, \lambda_n^*) \geqq \sigma \sum_{i,j}{}' (|\Delta x_n|^2 + |\Delta u_n|^2) + L_n(u_n^*, x_n^*, \lambda_n^*)& \\
+ \sum_{i,j}{}' g_x(x_n^*, u_n^*, t)\Delta x_n + \sum_{i,j}{}' g_u(x_n^*, u_n^*, t)\Delta u_n& \\
+ \sum_{i,j}{}' \dot{\lambda}_n^{*T} \Delta x_n& \\
+ \sum_{i,j}{}' \lambda_n^{*T}(f_x(x_n^*, u_n^*, t)\Delta x_n + f_u(x_n^*, u_n^*, t)\Delta u_n).&
\end{aligned}
\tag{3.6}
$$

Then using (3.1) (c) and (3.1) (e) we obtain

$$
L_n(u_n, x_n, \lambda_n^*) \geqq \sigma \sum_{i,j}{}' (|\Delta x_n|^2 + |\Delta u_n|^2) + L_n(u_n^*, x_n^*, \lambda_n^*), \tag{3.7}
$$

completing the proof.

Now if u_n and x_n are candidate solutions to the discrete problem (2.2), i.e. (2.2) (b) is satisfied, then $L_n(u_n, x_n, \lambda_n^*) = \sum_{i,j}' g(x_n, u_n, t)$. If $\hat{u}_n$ and $\hat{x}_n$ (with some λ_n) satisfy the necessary conditions (3.1), then using Theorem 3.2 we have $\sum_{i,j}' g(\hat{x}_n, \hat{u}_n, t) \leqq \sum_{i,j}' g(x_n, u_n, t)$, i.e. $(\hat{x}_n, \hat{u}_n)$ solves (2.2). Thus (A3) implies the sufficiency of the necessary conditions. Moreover, from (3.7) it follows that the values of x_n and u_n are unique at the collocation points ξ_{ij}. Since $x_n(0) = x_0$, x_n is unique.

Now that we have shown the equivalence of dealing with the discrete problem either directly or through the discrete necessary conditions, convergence theorems can be established by an analysis of (3.1) as a discretization of (2.1). This is done in section 4 with the aid of an additional assumption on the problem. Before doing that, we first give a convergence theorem that shows optimal rates of convergence are obtainable without any more assumptions other than smoothness.

THEOREM 3.3. *Let* (P) *be in* C^{2p} *and let* (A2)–(A4) *hold. Then*

$$
\begin{aligned}
&\text{(a)} \quad \sum_{i,j}{}' (u_n^* - u^*)^T (u_n^* - u^*) = O(|\Delta_n|^{2(p+1)}), \\
&\text{(b)} \quad \sum_{i,j}{}' (x_n^* - x^*)^T (x_n^* - x^*) = O(|\Delta_n|^{2(p+1)}).
\end{aligned}
\tag{3.8}
$$

Proof. Using Lemma 4.1 to follow and standard differential equations arguments, it follows that x^* and λ^* have at least $2p+1$ continuous derivatives and u^* has at least $2p$ continuous derivatives.

From Theorem 3.2, it follows that for admissible u_n, x_n,

$$
L_n(u_n^*, x_n^*, \lambda_n^*) \leqq L_n(u_n, x_n, \lambda_n^*). \tag{3.9}
$$

As in the proof of Theorem 3.2, we have for admissible λ_n that

$$
\begin{aligned}
L_n(u_n^*, x_n^*, \lambda_n) \geqq L_n(u_n, x_n, \lambda_n) + \sigma \sum_{i,j}{}' (\Delta u_n^T \Delta u_n + \Delta x_n^T \Delta x_n)& \\
+ \sum_{i,j}{}' (g_x(x_n, u_n, t) + f_x^T(x_n, u_n, t)\lambda_n + \dot{\lambda}_n)^T \Delta x_n& \\
+ \sum_{i,j}{}' (g_u(x_n, u_n, t) + f_u^T(x_n, u_n, t)\lambda_n)\Delta u_n&
\end{aligned}
\tag{3.10}
$$

where $\Delta u_n = u_n^* - u_n$ and $\Delta x_n = x_n^* - x_n$. Combining (3.9) and (3.10) we obtain

$$\sum_{i,j}{}' \Delta\lambda_n^T(-\dot{x}_n + f(x_n, u_n, t))$$

$$\sigma \sum_{i,j}{}' (\Delta u_n^T \Delta u_n + \Delta x_n^T \Delta x_n) \leqq -\sum_{i,j}{}' (g_x(x_n, u_n, t) + f_n^T(x_n, u_n, t)\lambda_n + \dot{\lambda}_n)^T \Delta x_n \tag{3.11}$$

$$-\sum_{i,j}{}' (g_u(x_n, u_n, t) + f_u^T(x_n, u_n, t)\lambda_n)\Delta u_n$$

where $\Delta\lambda_n = \lambda_n^* - \lambda_n$. Now choose u_n to be the best L^∞-approximation of u^*. The estimate $\|u_n - u^*\|_\infty = O(|\Delta_n|^{p+1})$ follows from [18]. Now choose x_n to be the solution of the equations

$$-\dot{x}_n(\xi_{ij}) + f(x_n(\xi_{ij}), u_n(\xi_{ij}), \xi_{ij}) = 0 \tag{3.12}$$

for all i, j and with $x_n(0) = x_0$. Using (2.1) (c) and the smoothness of f, the problem

$$-\dot{x} + f(x, u_n, t) = 0, \qquad 0 < t < 1,$$

$x(0) = x_0$ has a unique solution $\hat{x}_n$ satisfying $\|\hat{x}_n - x^*\|_\infty = O(|\Delta_n|^{p+1})$. The functions $\hat{x}_n$ will be in C^1 because u_n is only continuous. However higher derivatives through order $2p+1$ will be piecewise smooth with jump discontinuities occurring only at the mesh points. It thus follows using the theory of collocation at Gauss points as given in Russell [17] that x_n is well-defined by (3.12) for all $|\Delta_n|$ sufficiently small and that $\|x_n - \hat{x}_n\|_\infty = O(|\Delta_n|^{p+1})$. Using the triangle inequality, it then follows that $\|x_n - x^*\|_\infty = O(|\Delta_n|^{p+1})$. Define λ_n to be the solution of the equations

$$\dot{\lambda}_n(\xi_{ij}) + f_x^T(x_n, u_n, t)\lambda_n|_{\xi_{ij}} + g_x(x_n, u_n, t)|_{\xi_{ij}} = 0$$

for all i, j and $\lambda_n(1) = 0$. Repeating the analysis above we have that λ_n is defined for $|\Delta_n|$ sufficiently small and that $\|\lambda_n - \lambda^*\|_\infty = O(|\Delta_n|^{p+1})$. Substituting u_n, x_n and λ_n into (3.11), the first two terms on the right hand side vanish and so

$$\sigma \sum_{i,j}{}' (\Delta u_n^T \Delta u_n + \Delta x_n^T \Delta x_n) \leqq -\sum_{i,j} (g_u(x_n, u_n, t) + f_u^T(x_n, u_n, t)\lambda_n)\Delta u_n. \tag{3.13}$$

Recall that $g_u(x^*, u^*, t) + f_u^T(x^*, u^*, t)\lambda^* = 0$. Thus

$$\sum_{i,j}{}'(g_u(x_n, u_n, t) + f_u^T(x_n, u_n, t)\lambda_n)\Delta u_n$$

$$= \sum_{i,j}{}' (g_u(x_n, u_n, t) - g_u(x^*, u^*, t))\Delta u_n$$

$$+ \sum_{i,j}{}' ((f_u^T(x_n, u_n, t) - f_u^T(x^*, u^*, t))\lambda_n + f_u^T(x^*, u^*, t)(\lambda_n - \lambda^*))\Delta u_n.$$

Using the estimates for $u_n - u^*$, $x_n - x^*$ and $\lambda_n - \lambda^*$ and the smoothness of f and g, we thus obtain via the Schwarz inequality that the right hand side of (3.13) is less than or equal to $O(|\Delta_n|^{p+1})(\sum_{i,j}' \Delta u_n^T \Delta u_n)^{1/2}$. Applying this to (3.11) gives

$$\sum_{i,j}{}' \Delta u_n^T \Delta u_n = O(|\Delta_n|^{2(p+1)}). \tag{3.14}$$

Since $\sum_{i,j}' (u_n - u^*)^T(u_n - u^*) = O(|\Delta_n|^{2(p+1)})$, an application of the triangle inequality with (3.14) gives (3.8) (a). Using (3.8) (a), (3.13) and the triangle inequality gives (3.8) (b), completing the proof.

Remarks. The bounds of Theorem 3.3 are given using a discrete L^2-norm and are one order higher than the analogous bounds of [1] for the Galerkin method. The bounds of [1] are given in the L^2-norm. Theorem 3.3 does not contain a bound for the dual variable. Using (3.1) (c) and collocation theory, convergence follows directly, but the order of convergence in the L^∞-norm will be less than the best bound because that is all the bounds (3.8) (a)–(b) imply for the L^∞-norm. It is possible with the added assumption that the matrix $-f_x^T(x^*, u^*, t)$ is uniformly positive definite in t to deduce an analogous bound to (3.8) (a)–(b) for $\lambda_n - \lambda^*$. The argument is similar to that given in Lasiecka [10]. (See also Lions [11].) We omit this here and in the next section give both existence and improved convergence under slightly different assumptions.

Also, Theorem 3.3 does not contain statements of error bounds in the case the assumed smoothness is lacking. The theory of collocation at Gauss points [17] requires extra smoothness to achieve optimal convergence rates. Here this means (P) must be in C^{2p}. Otherwise, for the problem studied, theoretical rates drop to order p until the smoothness of the solutions drops below C^{p+1}. Then convergence becomes of order $\nu-1$ if the solutions are in C^ν. These results are straightforward applications of the collocation theory [17] and statements of all the various cases are left to the reader.

4. High order error estimates. We will assume throughout this section that assumptions (A1)–(A3) hold. We first establish a lemma.

LEMMA 4.1. *The equation* $g_u(x, u, t)+f_u^T(x, u, t)\lambda = 0$ *can be solved uniquely for* $u = \phi(x, \lambda, t)$ *in a neighborhood* N' *of* (u^*, x^*, λ^*) *so that* $u^* = \phi(x^*, \lambda^*, t)$. *Moreover, if* P *is in* C^ν, ϕ *has* ν *continuous derivatives.*

Proof. This is a consequence of (A3) and the implicit function theorem.

Remark. In the case of quadratic-regulator problem

$$\min_u \frac{1}{2}\int_0^1 (x^TQx+u^TRu)\,dt \tag{4.1}$$

subject to

$$\dot{x} = Ax+Bu, \qquad x(0)=x_0, \tag{4.2}$$

with $R(t)$ positive definite, equation (2.1) (a) becomes

$$Ru+B^T\lambda = 0, \tag{4.3}$$

and so $u = -R^{-1}B^T\lambda$. The function $g_{uu}+f_{uu}^T\lambda$ in this case is simply R.

Using Lemma 4.1, the necessary conditions may be written as the two-point boundary value problem

$$\text{(4.4)}\qquad \begin{aligned} &\text{(a)} \quad \dot{\lambda}^*+f_x(x^*, \phi(x^*, \lambda^*, t), t)^T\lambda^*+g_x(x^*, \phi(x^*, \lambda^*, t), t)=0,\\ &\text{(b)} \quad -\dot{x}+f(x^*, \phi(x^*, \lambda^*, t), t)=0,\\ &\text{(c)} \quad x^*(0)=x_0, \qquad \lambda^*(1)=0. \end{aligned}$$

Also using Lemma 4.1, the necessary conditions derived in Theorem 3.1 can be written as

$$\text{(4.5)}\qquad \begin{aligned} &\text{(a)} \quad \dot{x}_n^*(\xi_{ij})=f(x_n^*(\xi_{ij}), \phi(x_n^*(\xi_{ij}), \lambda_n^*(\xi_{ij}), \xi_{ij}), \xi_{ij}),\\ &\text{(b)} \quad \dot{\lambda}_n^*(\xi_{ij})=-f_x^T(x_n^*(\xi_{ij}), \phi(x_n^*(\xi_{ij})\lambda_n^*(\xi_{ij}), \xi_{ij}), \xi_{ij})\lambda_n^*(\xi_{ij})\\ &\qquad\qquad -g_x(x_n^*(\xi_{ij}), \phi(x_n^*(\xi_{ij}), \lambda_n^*(\xi_{ij}), \xi_{ij}), \xi_{ij}),\\ &\qquad\qquad\qquad\qquad i=1,\cdots,n; \quad j=1,\cdots,p,\\ &\text{(c)} \quad x_n^*(0)=x_0, \qquad \lambda_n^*(1)=0. \end{aligned}$$

It follows that (4.5) represents the equations for collocation at Gauss points as an approximation scheme for the solution to the necessary conditions (4.4). With one additional assumption, we can now appeal to the theory of collocation methods for vector systems as given in, for example, Russell [17] and deduce convergence and convergence rates. We add the assumption (A5) that when (4.4) is linearized about (x^*, λ^*), the resulting variational problem is uniquely solvable. See [17] for details.

THEOREM 4.2. *Let hypotheses* (A1)–(A3) *and* (A5) *hold. Then solutions* x_n^*, λ_n^* *to* (4.5) *exist and are unique in a neighborhood of* (x^*, λ^*) *for all partitions* Δ_n *with* $|\Delta_n|$ *sufficiently small.*

If (P) *is in* C^{2p}, *then*

$$\|x_n^* - x^*\|_\infty = O(|\Delta_n|^{(p+1)}),$$

$$\|\lambda_n^* - \lambda^*\|_\infty = O(|\Delta_n|^{(p+1)}),$$

$$|x_n^*(t_i) - x^*(t_i)| = O(|\Delta_n|^{2p}),$$

$$|\lambda_n^*(t_i) - \lambda^*(t_i)| = O(|\Delta_n|^{2p}),$$

where $t_i \in \Delta_n$.

If we define $u_n = \phi(x_n^*, \lambda_n^*, t)$, it follows from Theorem 4.2 and Lemma 4.1 that $u_n - u^*$ will satisfy the bounds of Theorem 4.2. Of course u_n is not identically u_n^*, but $u_n(\xi_{ij}) = u_n^*(\xi_{ij})$, i.e., they agree at the collocation points.

Remark. Theorem 4.2 does not actually require the full strength of assumption (A3). What is actually required is a condition allowing the implicit function theorem to be used so that Lemma 4.1 holds. Then with (A5), existence, uniqueness, and convergence of the approximate solutions to (4.5) follows. The extra smoothness assumption of Theorem 4.2 leads to the high order rate. Note also that in contrast to Theorem 3.3, the bounds of Theorem 4.2 are in the L^∞-norm. Other convergence rates follow in the L^∞-norm if less smoothness is present.

REFERENCES

[1] W. E. BOSARGE, JR., O. G. JOHNSON, R. S. MCKNIGHT AND W. P. TIMLAKE, *The Ritz–Galerkin procedure for nonlinear control problems*, SIAM J. Numer. Anal., 10 (1973), 11, 94–111.

[2] W. E. BOSARGE, JR. AND O. G. JOHNSON, *Error bounds of high order accuracy for the state regulator problem via piecewise polynomial approximations*, this Journal, 9 (1971), pp. 15–28.

[3] B. M. BUDAK, E. M. BERKOVICH AND E. N. SOLV'EVA, *Difference approximations in optimal control problems*, this Journal, 7 (1969), pp. 18–31.

[4] JOHN H. CERUTTI, *Collocation for systems of ordinary differential equations*, Computer Sciences Technical Report #320, Computer Sciences Department, University of Wisconsin-Madison, 1974.

[5] J. W. DANIEL, *The Ritz–Galerkin method for abstract optimal control problems*, this Journal, 11 (1973), pp. 53–63.

[6] CARL DEBOOR AND BLAIR SWARTZ, *Collocation at Gaussian points*, SIAM J. Numer. Anal., 10 (1973), pp. 582–606.

[7] RICHARD S. FALK, *Approximation of a class of optimal control problems with order of convergence estimates*, J. Math. Anal. Appl., 44 (1973), pp. 28–47.

[8] WILLIAM W. HAGER, *Rates of convergence for discrete approximations to unconstrained control problems*, SIAM J. Numer. Anal., 13 (1976), pp. 449–471.

[9] ———, *The Ritz–Trefftz method for state and control constrained optimal control problems*, Ibid., 12 (1975), pp. 854–867.

[10] IRENA LASIECKA, *Finite difference approximation of optimal control for systems described by nonlinear differential equations with delay*, Control and Cybernetics, 5 (1976), pp. 35–67.

[11] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.

[12] T. R. LUCAS AND G. W. REDDIEN, *Some collocation methods for nonlinear boundary value problems*, SIAM J. Numer. Anal., 9 (1972), pp. 341–356.

[13] F. H. MATHIS AND G. W. REDDIEN, *Difference approximations to control problems with functional arguments*, this Journal, 16 (1978), pp. 436–449.

[14] G. W. REDDIEN, *Approximation methods for two-point boundary value problems with nonlinear boundary conditions*, SIAM J. Numer. Anal., 13 (1976), pp. 405–411.

[15] ———, *The Ritz–Galerkin method in optimal control*, in preparation.

[16] R. D. RUSSELL AND L. F. SHAMPINE, *A collocation method for boundary value problems*, Numer. Math., 19 (1972), pp. 1–28.

[17] R. D. RUSSELL, *Collocation for systems of boundary value problems*, Ibid., 23 (1974), pp. 119–133.

[18] M. H. SCHULTZ, *L^∞-multivariate approximation theory*, SIAM J. Numer. Anal., 6 (1969), pp. 161–183.

[19] MARTIN H. SCHULTZ, *Spline Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[20] K. A. WITTENBRINK, *High order projection methods of moment- and collocation-type for nonlinear boundary value problems*, Computing, 11 (1973), pp. 255–274.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 2, March 1979

0363-0129/79/1702-0012 $ 01.00/0

RITZ–TREFFTZ APPROXIMATIONS IN OPTIMAL CONTROL*

F. H. MATHIS† AND G. W. REDDIEN‡

Abstract. Known convergence rates for the control approximations generated by using the Ritz–Trefftz method on the state regulator problem are not optimal. Optimal rates are proven here using standard techniques.

1. Introduction. The Ritz–Trefftz method using polynomial splines for the state regulator problem has been analyzed by Bosarge and Johnson [1], Hager [3] and Schultz [5]. The purpose of this note is to show that the convergence rates derived in those papers for the control are not sharp, and that actually the method achieves the optimal rate of convergence. The convergence rate known for the state is seen to be sharp in the numerical examples presented here. Hager [3] considers state and control constrained problems which we do not treat here.

For $v \geqq 1$ let X^v denote the set of mappings x from $[0, 1]$ to R^v so that x is continuous and $\dot{x}$ exists and is continuous except for possibly finitely many points and $\|\dot{x}\|^2 = \int_0^1 |\dot{x}|^2\, dt < \infty$. Let $A(t)$ be a $v \times v$ matrix and $B(t)$ a $v \times r$ matrix for t in $[0, 1]$, both of which have entries which are continuous functions of t in $[0, 1]$. Let $Q(t)$ and $R(t)$ be respectively a $v \times v$ symmetric, positive definite and an $r \times r$ symmetric positive definite matrix, both of which are continuous functions of t in $[0, 1]$. Then the state regulator problem is to find a function u in X^r and a function x in X^v which minimize

$$(1.1) \qquad J(u) \equiv \frac{1}{2}\int_0^1 (x^TQx + u^TRu)\, dt$$

subject to

$$(1.2) \qquad \begin{aligned} &\text{(a)} \quad \dot{x} = Ax + Bu, \qquad 0 \leqq t \leqq 1, \\ &\text{(b)} \quad x(0) = x_0. \end{aligned}$$

Using known results of control theory [5], one can show that problem (1.1)–(1.2) is equivalent to the problem of finding a function λ in X^v which maximizes the Lagrangian

$$(1.3) \qquad L[u, x, \lambda] \equiv J(u) + \int_0^1 \lambda^T(-\dot{x} + Ax + Bu)\, dt + \lambda^T(0)[x_0 - x(0)]$$

subject to the constraint $\lambda(1) = 0$ where u and x are defined by

$$(1.4) \qquad u(t) = -R^{-1}(t)B^T(t)\lambda(t)$$

and

$$(1.5) \qquad x(t) = -Q^{-1}(t)(\dot{\lambda}(t) + A^T(t)\lambda(t))$$

both for t in $[0, 1]$. Substituting (1.4)–(1.5) into (1.3), we may express the Lagrangian in terms of λ alone. Define

$$(1.6) \quad a(w, v) = \int_0^1 ((Q^{-1}\dot{w} + Q^{-1}A^Tw)^T\dot{v} + (AQ^{-1}\dot{w} + AQ^{-1}A^Tw + BR^{-1}B^Tw)^Tv)\, dt.$$

* Received by the editors October 27, 1977, and in revised form August 8, 1978.

† Mathematics Department, Auburn University, Auburn, Alabama 36830.
‡ Mathematics Department, Vanderbilt University, Nashville, Tennessee 37235.

Then

$$-L[u, x, \lambda] = \tfrac{1}{2}a(\lambda, \lambda) - \lambda(0)^T x_0 \overset{\text{def.}}{=} F(\lambda).$$

The following characterization result is known [5].

THEOREM 1.1. *The optimal Lagrange multiplier exists and is the unique solution in* $X_0 \equiv \{\phi \in X^v\colon \phi(1) = 0\}$ *of the generalized Euler equation*

$$a(\lambda, y) = y^T(0)x_0, \quad \textit{for all } y \in X_0. \tag{1.7}$$

Let S be any finite dimensional subspace of X_0 with basis $\{\phi_i\}_{i=1}^n$. Then the Ritz–Trefftz method is to find λ_S which minimizes $F(\lambda)$ over S. It follows from [3] that the Ritz–Trefftz method is well-defined and solves (1.7) over S.

2. Convergence. Our convergence proof will be a modification of the argument given in Chapter 7 of [5] for the convergence of the Ritz method applied to elliptic boundary value problems. See also Chapter 1 of [6]. We need a few lemmas, the first of which can be found in Hager [2].

LEMMA 2.1. *There exist positive constants c_1 and c_2 so that for all w in X_0,*

$$c_1\|w\|_1^2 \leqq \|\dot{w} + A^T w\|^2 \leqq a(w, w) \leqq c_2\|w\|_1^2 \tag{2.1}$$

where

$$\|w\|_1^2 = \|\dot{w}\|^2 + \|w\|^2.$$

Proof. Recall that $w(1) = 0$. If $v \equiv \dot{w} + A^T w$, then using the Gronwall inequality it follows that $\|w\| \leqq c\|v\|$ for some constant $c > 0$. Now

$$\begin{aligned} a(w, w) &= \int_0^1 ((Q^{-1}\dot{w} + Q^{-1}A^T w)^T \dot{w} + (AQ^{-1}\dot{w} + AQ^{-1}A^T w + BR^{-1}B^T w)^T w)\, dt \\ &= \int_0^1 ((\dot{w} + A^T w)^T Q^{-1}(\dot{w} + A^T w) + (BR^{-1}B^T w)^T w)\, dt. \end{aligned} \tag{2.2}$$

Using the previous inequality, the positive definiteness of Q and R, and the assumed continuity for A, Q, R and B in (2.2), (2.1) follows.

We next give a result which establishes the strong coerciveness of $a(v, w)$ in the sense of [5].

LEMMA 2.2. *Let Q^{-1} and A be continuously differentiable. For g in X^v, there exists a unique function w in X_0 satisfying*

$$a(w, y) = \langle y, g\rangle \equiv \int_0^1 y^T g\, dt \tag{2.3}$$

for all y in X_0. Moreover, there exists a constant $c_3 > 0$ so that

$$\|\dot{w}\| \leqq c_3\|g\|, \tag{2.4}$$

where c_3 is independent of g.

Proof. The existence and uniqueness follows using Lemma 2.1 from Theorem 1.1 of Lions [4]. It is easy to see that the solution w is also the unique solution to the two-point boundary value problem

$$\frac{-d}{dt}(Q^{-1}\dot{w} + Q^{-1}A^T w) + AQ^{-1}\dot{w} + (AQ^{-1}A^T + BR^{-1}B^T)w = g, \qquad 0 \leqq t \leqq 1, \tag{2.5}$$

with boundary conditions

$$\dot{w}(0)+A^T(0)w(0)=0, \qquad w(1)=0. \tag{2.6}$$

The estimate (2.4) now follows by standard Green's function arguments.

We are now in a position to apply the analysis given for elliptic boundary value problems and the Ritz method using the technique of Nitsche.

We first note that by combining the Euler equations for λ and λ_S (see (1.7)), we obtain

$$a(\lambda_S-\lambda, \phi_i)=0, \quad \text{all } \phi_i \text{ in } S. \tag{2.7}$$

We next give a lemma that is a consequence of Theorem 1.1 in Strang and Fix [6]. Note from (1.6) that $a(\cdot, \cdot)$ is symmetric.

LEMMA 2.3. $a(\lambda-\lambda_s, \lambda-\lambda_s)=\inf_{s\in S} a(\lambda-s, \lambda-s)$.

We next choose for S spaces of polynomial splines. Let Δ_n: $0=t_0^n<t_1^n<\cdots<t_n^n=1$ be a partition of $[0, 1]$ and let $|\Delta_n|\equiv\max_i (t_i^n-t_{i-1}^n)$. We let $S(\Delta_n, p)$ denote the v-tuples of polynomial splines of degree $p\geqq 1$ over Δ_n. Each function u in $S(\Delta_n, p)$ will be a v-tuple of polynomials of degree p on each subinterval of Δ_n and u will have m continuous derivatives over $[0, 1]$ for some index $0\leqq m\leqq p-1$. We will omit the designation of m in our notation for $S(\Delta_n, p)$ since it is irrelevant here. Let $S_0(\Delta_n, p)$ denote the functions u in $S(\Delta_n, p)$ satisfying $u(1)=0$.

This next lemma follows from the results of Bosarge and Johnson [1].

LEMMA 2.4. *Let λ solve* (1.7), *let λ_n be the Ritz–Trefftz approximation to λ over $S_0(\Delta_n, p)$ and λ be in $C^k[0, 1]$, $k\geqq 1$. Then*

$$\|\dot{\lambda}-\dot{\lambda}_n+A^T(\lambda-\lambda_n)\|\leqq \text{const. } |\Delta_n|\gamma^{\min(k-1,p)}.$$

It is now straightforward to combine the preceeding lemmas as on pp. 48–49 in [6] and deduce the sharp bound.

THEOREM 2.5. *Let λ and λ_n be as in Lemma* 2.4 *and let the conditions of Lemma* 2.2 *hold. If λ is in $C^k[0, 1]$, $k\geqq 1$, then*

$$\|\lambda-\lambda_n\|\leqq \text{const. } |\Delta_n|^{\min(k,p+1)}.$$

Thus we have shown that the convergence rate for $\lambda-\lambda_n$ is best possible. Since $u-u_n=-R^{-1}B^T(\lambda-\lambda_n)$, this error also converges at the best rate. The formula (1.5) does not lead to an approximation to the state that achieves the best possible estimate because of the λ term. However, one could go back to (1.2)(a) given u_n and compute an optimal order approximation to the state if desired.

Example. We consider the simple scalar problem

$$\min_u \frac{1}{2}\int_0^1 (x^2+u^2)\,dt$$

subject to

$$\dot{x}=x+u, \qquad 0\leqq t\leqq 1, \qquad x(0)=1.$$

For the Ritz–Trefftz method using linear splines we obtained the results in the next table. The convergence rate, β, was computed based on the actual errors assuming the error behaved like const. $|\Delta_n|^{\beta}$. All meshes used were uniform. In this example, $\lambda=-u$ and $\lambda_n=-u_n$.

TABLE 1

Δ_n	$\|\lambda-\lambda_n\|$	β	$\|x-x_n\|$	β
$\frac{1}{5}$	$.264\cdot 10^{-2}$	—	$.997\cdot 10^{-1}$	—
$\frac{1}{10}$	$.657\cdot 10^{-3}$	2.006	$.501\cdot 10^{-1}$	.993
$\frac{1}{15}$	$.284\cdot 10^{-3}$	2.071	$.328\cdot 10^{-1}$	1.047
$\frac{1}{20}$	$.164\cdot 10^{-3}$	1.904	$.251\cdot 10^{-1}$	.930

In the next table, we present numerical results for the same problem, but using C^2-cubic splines.

TABLE 2

Δ_n	$\|\lambda-\lambda_n\|$	β	$\|x-x_n\|$	β
$\frac{1}{4}$	$.101\cdot 10^{-4}$	—	$.267\cdot 10^{-3}$	—
$\frac{1}{5}$	$.434\cdot 10^{-5}$	3.80	$.141\cdot 10^{-3}$	2.88
$\frac{1}{6}$	$.214\cdot 10^{-5}$	3.88	$.828\cdot 10^{-4}$	1.90
$\frac{1}{7}$	$.118\cdot 10^{-5}$	3.87	$.529\cdot 10^{-4}$	2.91

REFERENCES

[1] W. E. BOSARGE AND O. G. JOHNSON, *Error bounds of high-order accuracy for the state regulator problem via piecewise polynomial approximations*, this Journal, 9 (1971), pp. 15–28.

[2] W. W. HAGER, *Rates of convergence for discrete approximations to unconstrained control problems*, Ph.D. dissertation, M.I.T., 1974.

[3] ———, *The Ritz–Trefftz method for state and control constrained optimal control problems*, SIAM J. Numer. Anal., 6 (1975), pp. 854–867.

[4] J. L. LIONS, *Contrôle Optimal de Systèmes Gouvernés par des Equations aux Dérivées Partielles*, Dunod, Paris, 1968.

[5] M. H. SCHULTZ, *Spline Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[6] GILBERT STRANG AND GEORGE J. FIX, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 2, March 1979

0363–0129/79/1702–0013 $01.00/0

STRUCTURE MÉTRIQUE DES ORBITES DE FAMILLES SYMÉTRIQUES DE CHAMPS DE VECTEURS ET THÉORIE DU TEMPS MINIMUM*

ANDREA BACCIOTTI†

Sommaire. L'étude des propriétés et de la structure des ensembles des états atteignables ou orbites d'une famille de champs de vecteurs sur une variété différentiable a appelé récemment l'attention de plusieurs auteurs: il s'agit en effet d'une approche théorique très importante à la théorie des systèmes de contrôle non linéaires. Il est connu que une orbite S d'une famille symétrique peut être munie d'une structure de variété différentiable: dans cet article on montre que sous des conditions très raisonnables S peut être munie d'une structure d'espace métrique, qui induit sur S la même topologie de la structure différentiable. La fonction de distance qu'on va définir sur S est semblable à la fonction du temps minimum. En utilisant cette distance on peut alors poser les bases pour une théorie du problème du temps minimum.

Introduction. L'approche géométrique à la théorie du contrôle, devéloppée à partir des travaux de Hermann et Lobry [11], permet une étude très générale des propriétés de l'ensemble des états atteignables. Un des résultats les plus importants jusqu'ici connus est le théorème de Sussmann et Stefan (voir la proposition 1.2) qui montre comment l'ensemble des états atteignables par une famille symétrique de champs de vecteurs possède une structure naturelle de variété différentiable. Dans ce travail on introduisit sur l'ensemble des états atteignables une structure d'espace métrique: l'idée est de mesurer les distances au moyen du temps qu'on employe à les parcourir, le long des trajectoires de la famille donnée. Si nous n'avons pas à disposition des vitesses infinies, et si nous sommes en conditions convenables de autoaccessibilité (voir dans la suite) la topologie induite sur l'ensemble des états atteignables par sa structure métrique est équivalente à la topologie donnée sur l'espace des états.

La métrique introduite est strictement liée à la fonction du temps minimum: en utilisant cette métrique on peut alors déduire d'un façon très naturelle les résultats qui sont à la base de la théorie du temps minimum, tels que la continuité de l'ensemble des états atteignables par rapport au temps, la continuité de la fonction du temps minimum et une condition d'extrémalité pour les trajectoires optimales. Malheureusement, nous ne sommes pas en état d'achever notre exposition avec un théorème d'existence pour les contrôles optimaux, car en général, l'ensemble des points atteignables à un certain instant $T>0$ n'est pas fermé; et cela en tant que dans la théorie géométrique du contrôle on travaille structurellement avec des contrôles bang-bang continus par morceaux. Lorsque l'ensemble des points atteignables résulte fermé à tout instant $T \geqq 0$ (il est ainsi par exemple pour certains systèmes linéaires ou bilinéaires, voir [5]), la théorie du temps minimum peut être tout à fait développée de notre point de vue.

Voilà le plan du travail: la § 1 contient le liste des notations et les rappels nécessaires; la § 2 est consacré à la définition et à l'étude de la structure métrique sur les orbites d'une famille symétrique de champs de vecteurs; dans la § 3 on pose le problème du temps minimum et l'on montre la continuité de la fonction du temps minimum; enfin, dans la § 4 on achève l'étude de l'ensemble des états atteignables en démontrant, en particulier, qu'il dépend continûment du temps.

1. Notations, définition et rappels.

Dans ce travail nous utiliserons les notations suivantes:

k, h, n, μ sont des entiers positivs.

* Received by the editors August 20, 1977, and in final revised form March 13, 1978.

† Istituto di Matematica Applicata, Facoltà di Ingegneria, Firenze (Italia) Questo lavoro é stato svolto nell'ambito del Gruppo Nazionale per l'Analisi Funzionale e le sue Applicazioni del C.N.R.

$\mathbb{R}^n$ dénote l'espace des n-ples de nombres réels, $\mathbf{x}=(x_1, \cdots, x_n)$ et $\mathbb{R}^n_+$ dénote l'ensemble $\{\mathbf{x}=(x_1, \cdots, x_n)\in \mathbb{R}^n: x_i \geqq 0, i=1, \cdots, n\}$. Si $\mathbf{x}=(x_1, \cdots, x_n)\in\mathbb{R}^n$, on note $\|\mathbf{x}\|_1=\sum_{i=1}^n |x_i|$.

I est un ensemble de indices at I^k est l'ensemble des k-arrangements avec répétition d'éléments de I.

L'adhérence et l'interieur d'un ensemble A dans un espace topologique E sont dénotés respectivement par $\operatorname{adh}_E A$ et $\operatorname{int}_E A$ ou, s'il n'y a pas de possibilité de confusion, plus simplement par adh A et int A; la frontière de A est notée ∂A.

M est une variété différentiable paracompacte de dimension n et de classe $C^{\mu+1}$ $(1\leqq\mu\leqq\omega)$; les points de M sont dénotés par $m, p, q, \cdots$ et l'espace tangent à M en chaque point m est dénoté par TM_m. On suppose M munie d'une structure Riemanienne définie par un champ de tenseurs G (de classe C^μ) deux fois covariants (voir par exemple [3]): cela détermine, pour chaque $m\in M$, la donnée d'une forme bilinéaire $G_m(\cdot, \cdot)$ sur $TM_m\times TM_m$ symétrique, non dégenerée et définie positive.

$D=(X^i(\cdot))_{i\in I}$ est une famille de champs de vecteurs de classe C^μ sur M. Pour chaque $i\in I$ on note $(t, m)\mapsto X^i_t(m)$ le groupe à un paramètre engendré par $X^i(\cdot)$. On suppose que tous les champs de vecteurs de D soient complets: cela signifie que, pour chaque $i\in I$, le groupe $X^i_t(m)$ est défini quel que soit $(t, m)\in\mathbb{R}\times M$. Si $m\in M$ et $\xi=(i_1, \cdots, i_k)\in I^k$, on note $\rho_{\xi,m}(\cdot)$ l'application

$$(1)\qquad (t_1, \cdots, t_k)\mapsto X^{i_k}_{t_k}\circ\cdots\circ X^{i_1}_{t_1}(m): \mathbb{R}^k\to M$$

différentiable de classe C^μ. L'ensemble de toutes les applications $\rho_{\xi,m}$ définies par (1) est noté par $\mathscr{G}$.

Voilà maintenant quelques définitions et résultats classiques, que nous utiliserons dans la suite.

DÉFINITION. On dit qu'une famille D de champs de vecteurs est *symétrique* si pour tout $X(\cdot)\in D$, $-X(\cdot)\in D$.

DÉFINITION. Soient $m, m'\in M$. On dit que m' est *atteignable de m* si existent k, $\xi\in I^k$ et $\mathbf{t}=(t_1, \cdots, t_k)\in\mathbb{R}^k_+$ tels que $\rho_{\xi,m}(\mathbf{t})=m'$.

PROPOSITION 1.1. *Si la famille D est symétrique, la relation d'atteignabilité est une relation d'équivalence.*

Démonstration. La démonstration est triviale. □

DÉFINITION. Les classes d'équivalence de la relation d'atteignabilité en M sont appelées *orbites de la famille D*, ou bien ensembles des points ateignables.

PROPOSITION 1.2 (Sussmann–Stefan). *Soit S une orbite de la famille symétrique D de champs de vecteurs. Il existe sur S une structure différentiable σ telle que*:

(i) Pour chaque $m\in S$, $\rho_{\xi,m}(\cdot)$ définie par la (1) est une application continue de $\mathbb{R}^k$ en (S, σ), quels que soient k et $\xi\in I^k$.

(ii) (S, σ) est une sous-variété immerse en M.

Démonstration. Pour la démonstration voir [12], [13]. □

DÉFINITION. Soit $T\geqq 0$ et $m_0\in M$. L'ensemble

$$(2)\qquad R(T, m_0)=\{m\in M: \exists k, \xi\in I^k \text{ et } \mathbf{t}\in\mathbb{R}^k_+ \text{ t.q. } m=\rho_{\xi,m_0}(\mathbf{t}) \text{ et } \|\mathbf{t}\|_1=T\}$$

est appelé ensemble des points atteignables de m_0 à l'instant $T\geqq 0$.

PROPOSITION 1.3. *Soit D une famille symétrique de champs de vecteurs sur M et soit S une orbite. On a*

(i) *Si $m_0\in S$ alors $R(T, m_0)\subset S$ à tout instant $T\geqq 0$ et $S=\bigcup_{t\geqq 0} R(t, m_0)$.*

(ii) *Si $T_2>T_1>0$ alors $R(T_1, m_0)\subset R(T_2, m_0)$ pour chaque $m_0\in M$, et $R(T, m_0)=\bigcup_{0\leqq t\leqq T} R(t, m_0)$.*

(iii) *Si $m\in R(T, m')$ alors $m'\in R(T, m)$ quels que soient $m, m'\in S$ et $T>0$.*

Démonstration. Pour la démonstration voir [1]. □

DÉFINITION. On dit qu'un point $m_0 \in M$ est *autoaccessible à l'instant* $T > 0$ *par rapport à la famille* D si $m_0 \in \operatorname{int}_M R(T, m_0)$.

Des conditions d'autoaccessibilité sont données en [1]. Des conditions afin que $\operatorname{int} R(T, m_0) \neq \varnothing$ sont données en [14] dans le cas analitique.

2. Structure métrique des orbites. Soit S une orbite de la famille symétrique D. Soient $m, m' \in S$ et soit

$$d(m, m') = \inf\{T > 0 \colon m' \in R(T, m)\}. \tag{3}$$

La (3) définit une fonction de $S \times S$ en $\mathbb{R}_+$.

THÉORÈME 2.1. *Quelle que soit la famille symétrique D de champs de vecteurs complets sur M et quelle que soit l'orbite S, la fonction* (3) *satisfait les propriétés suivantes*:

(i) $m = m' \Rightarrow d(m, m') = 0$;

(ii) $d(m, m') = d(m', m)$;

(iii) $d(m, m') + d(m', m'') \geqq d(m, m'')$.

Selon la définition de [2, p. 1] la fonction (3) est donc un écart sur S.

Démonstration. La verification de (i) et (ii) est triviale; (iii) est une conséquence simple de (1), (2), (3). □

DÉFINITION. On dit qu'une famille D de champs de vecteurs est *localement bornée* quand, quel que soit $m \in M$, il existe un voisinage U de m et un nombre réel $L > 0$ tels que

$$G_q(X^i(q), X^i(q)) < L \tag{4}$$

quels que soient $q \in U$ et $i \in I$.

Soit (V, φ) une carte locale en m, et soient $f_j^i(q)$ les composantes du champ $X^i(\cdot)$ en cette carte. En écrivant (4) dans le repère (V, φ) on voit que D est localement bornée si et seulement si il existe un voisinage V' de m contenu en V et un nombre réel $L' > 0$ tels que

$$\left[\sum_{j=1}^{n} |f_j^i(q)|^2\right]^{1/2} < L', \tag{5}$$

quels que soient $q \in V'$ et $i \in I$. La définition de famille localement bornée ne dépend pas donc de la structure Riemanienne de M.

THÉORÈME 2.2. *Soit D une famille symétrique de champs de vecteurs complets sur M, et soit S une orbite de D en M. Si D est localement bornée, la fonction* $(m, m') \mapsto d(m, m')$ *définie par la* (3) *est une distance sur S.*

Démonstration. D'après le Théorème 2.1, il suffit de vérifier que

$$d(m, m') = 0 \Rightarrow m = m', \qquad m, m' \in S. \tag{6}$$

Soit $m \neq m'$. Puisque (S, σ) est un espace de Hausdorff, il existe un voisinage U de m où (4) vaut, tel que $m' \notin U$.

Soit $g(\cdot, \cdot)$ da distance géodésique induite sur M par la structure Riemanienne G (voir [3]), et soit $\varepsilon > 0$ tel que $\{p \in M \colon g(p, m) < \varepsilon\} = V \subset U$. Soient $\xi = (i_1, \cdots, i_k) \in I^k$ et $\mathbf{t} = (t_1, \cdots, t_k) \in \mathbb{R}_+^k$ tels que $m' = \rho_{\xi,m}(\mathbf{t})$ et posons

$$m_0 = m, \qquad m_h = X_{t_h}^{i_h}(m_{h-1}),$$

$$\tau_0 = 0, \qquad \tau_h = \sum_{j=1}^{h} t_j,$$

pour $h = 1, \cdots, k$. On a $\tau_k = \|\mathbf{t}\|_1$ et $m_k = m'$. Soit encore $\vartheta \mapsto \gamma(\vartheta) \colon [0, \tau_k] \to M$ la courbe définie par

$$(7) \qquad \gamma(0)=m_0, \qquad \gamma(\vartheta)=X^{i_h}_{\vartheta-\tau_{h-1}}(m_{h-1}) \quad \text{si } \tau_{h-1}<\vartheta\leqq\tau_h.$$

Telle courbe est continue, et ses extrêmes sont l'un en dedans, l'autre en dehors de V. Donc il existe $\bar{\vartheta}\leqq\tau$ tel que si $\vartheta<\bar{\vartheta}$, $\gamma(\vartheta)\in V$ et $\gamma(\bar{\vartheta})\in\partial V$; on peut supposer $\bar{\vartheta}=\tau_{\bar{h}}$, pour un certain indice $\bar{h}<k$. Soit l_h la longueur de l'arc décrit par la courbe $\vartheta\mapsto\gamma(\vartheta)$ lorsque $\tau_{h-1}\leqq\vartheta\leqq\tau_h$; il suit par la (4)

$$(8) \qquad l_h=\int_{\tau_{h-1}}^{\tau_h}[G_{\gamma(\vartheta)}(X^{i_h}(\gamma(\vartheta)), X^{i_h}(\gamma(\vartheta)))]^{1/2}\,d\vartheta\leqq {}_hL^{1/2}.$$

La longeur de l'arc décrit par la courbe $\vartheta\mapsto\gamma(\vartheta)$ lorsque $0\leqq\vartheta\leqq\tau_{\bar{h}}$ est $\sum_{j=1}^{\bar{h}} l_j\geqq\varepsilon$. Par (8) il suit alors $\varepsilon\leqq(\sum_{j=1}^{\bar{h}} t_j)\leqq\tau_k L$, c'est-à-dire $\|\mathbf{t}\|_1\geqq\varepsilon/L$, pour tous $k, \xi\in I^k$ et $\mathbf{t}\in\mathbb{R}^k_+$ tels que $m'=\rho_{\xi,m}(\mathbf{t})$. En prenant la borne inférieure la (6) est démonstrée. □

Si la famille D n'est pas localement bornée, la (3) ne donne pas en géneral une distance sur S: cela est montré par le simple exemple qui suit.

Exemple 2.1. Soit $M=\mathbb{R}$ avec l'structure usuelle euclidienne. La famille $D=(aX(\cdot))_{a\in\mathbb{R}}$, où $m\mapsto X(m)\equiv 1$, est symétrique, mais non localement bornée; en étant $(aX)_t(0)=at$, on voit aisément que $d(0,1)=0$.

Dans les hypothèse du Théorème 2.2 la distance (3) définit sur S une topologie métrique que nous notons δ. Le but de la partie finale de cette section est de comparer la topologie δ et la topologie naturelle de M. On commencera aussi à étudier les boules ouvertes ou fermées de δ par rapport aux ensembles des points atteignables.

PROPOSITION 2.1. *Soit $m_0\in S$. Dans les hypothèse du Théorème* 2.2 *si U_S est un voisinage quelconque de m_0 en (S,σ) alors il existe $T>0$ tel que $R(T, m_0)\subset U_S$.*

Démonstration. Soit D_S la famille des réstrictions des champs de D à la variété (S,σ): il est clair que D_S est encore localement bornée. Répétons la construction de la démonstration du Théorème 2.2 par rapport à un voisinage U_S de m_0 en (S,σ) et à tous $\rho_{\xi,m}\in\mathscr{G}$ et $\mathbf{t}\in\mathbb{R}^k$ avec $\|\mathbf{t}\|_1<\varepsilon/L$; on a nécessairement $\gamma(\vartheta)\in U_S$ pour tout $\vartheta\leqq\|\mathbf{t}\|_1$. Donc $R(\varepsilon/L, m_0)\subset U_S$. □

PROPOSITION 2.2. *Soit $m_0\in S$ et soit $\{m_\nu\}$ une suite en S telle que lim $d(m_\nu, m_0)=$ 0. Alors $\{m_\nu\}$ converge à m_0 dans la topologie de σ.*

Démonstration. Il est évident que, dans ces hypothèses, quel que soit $T>0$ il existe ν pour lequel $m_\nu\in R(T, m_0)$. L'énoncé suit alors par la Proposition 2.1 □

PROPOSITION 2.3. *Soit $T>0$ et soit $m_0\in S$; dans les hypothèses du Théorème* 2.2 *on a*:

(i) $\{m\in S\colon d(m, m_0)<T\}=\bigcup_{0\leqq t<T}R(t, m_0)$;

(ii) $R(T, m_0)\subset\{m\in S\colon d(m, m_0)\leqq T\}\subset \operatorname{adh}_{(S,\sigma)}R(T, m_0)$.

Démonstration. La (i) et la première inclusion de (ii) sont des conséquences immédiates des définitions. Pour démontrer la deuxiéme inclusion de (ii), soit $m\in\{m\in S\colon d(m, m_0)\leqq T\}$. Si $d(m, m_0)<T$, il est clair que $m\in\operatorname{adh} R(T, m_0)$; soit donc $d(m, m_0)=T$ et soit $\{T_\nu\}$, $T_\nu>0$, une suite de nombres réels qui converge à zéro. Par la définition même de la fonction $d(\cdot,\cdot)$ ils existent, pour chaque ν, $k_\nu\in\mathbb{N}$, $\xi_\nu\in I^{k_\nu}$ et $\mathbf{t}_\nu\in\mathbb{R}^{k_\nu}$ tels que $m=\rho_{\xi_\nu,m_0}(\mathbf{t}_\nu)$ et $T\leqq\|\mathbf{t}_\nu\|_1\leqq T+T_\nu$. Répétons la construction de la démonstration du Théorème 2.2, pour obtenir, de l'application ρ_{ξ_ν,m_0}, une courbe $\vartheta\mapsto\gamma_\nu(\vartheta)$, définie par $\vartheta\in[0,\|\mathbf{t}_\nu\|_1]$. Quel que soit ν, on a $m_\nu=\gamma_\nu(T)\in R(T, m_0)$ et $d(m_\nu, m)\leqq\|\mathbf{t}_\nu\|_1-T\leqq T_\nu$. La démonstration est complète par la Proposition 2.2. □

Dans la Proposition 2.3 paraissent les boules ouvertes et fermées de la topologie δ: puisque les orbites sont des sous-variétés immerses en M, on peut affirmer à ce point que la topologie δ est plus fine que la topologie naturelle de M.

THÉORÈME 2.3. *Soit D une famille symétrique et localement bornée. Si chaque point de M est autoaccessible à tout instant $T > 0$, alors les orbites de D en M coincident avec les composantes connexes de M et, sur chaque orbite, la topologie δ est équivalente à la topologie de M.*

Démonstration. L'hypothèse de autoaccessibilité à tout instant $T > 0$ entraîne que l'orbite que passe chaque point $m \in M$ a dimension n (voir [1]; la première affirmation du théorème est alors un corollaire de la Proposition 1.2 (voir [13]). On peut supposer maintenant pour semplicité que M soit formée par une composante connexe toute seule: par les Propositions 2.1 et 2.3, il suit alors que tout ouvert de M contient un ouvert de δ. Par contre, par la définition de autoaccessibilité, il est clair que, quel que soit $T > 0$, il un ouvert de M contenu en $\{m \in M: d(m, m_0) < T\} = \bigcup_{0 \leqq t < T} R(t, m_0)$. Le Théorème est donc montré. □

Si l'hypothèse de autoaccessibilité à tout instant $T > 0$ n'est pas vérifiée en tout point de M, la topologie δ peut être effectivement plus fine que la topologie de M, même si M est formée par une orbite toute seule: cela est montré par l'exemple qui suit.

Exemple 2.2. Considérons en $M = \mathbb{R}^2$ les champs

$$X^1(x, y) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad X^2(x, y) = \begin{pmatrix} 1 \\ f(x) \end{pmatrix}$$

où $f(x)$ est x^2 si $x \geqq 0$ et est zéro si $x < 0$. Soit $D = (\pm X^1, \pm X^2)$. Cet exemple est développé en [1]. La boule ouverte de rayon 1 et de centre $(-1, 0)$ dans la topologie δ est donnée par le segment $-2 < x < 0$, $y = 0$.

Les exemples suivants sont en relation avec la Proposition 2.3. On doit dire que l'Example 2.4 n'est pas beaucoup approprié dans ce contexte, car on y utilise un champ non complet; il me semble de toute façon interessant, parce que l'hypothèse de travailler avec des champs complets n'est pas considerée essentielle par presque tous les auteurs.

Exemple 2.3. Considérons en $M = \{(x, y) \in \mathbb{R}^2: y > 0\}$ les champs

$$X^1(x, y) = \begin{pmatrix} 1/y \\ 1/y \end{pmatrix}, \qquad X^2(x, y) = \begin{pmatrix} 1/y \\ -1/y \end{pmatrix},$$

et soit $D = (\pm X^1, \pm X^2)$. Soit encore $b > 0$, et soit n un entier positif. Du point $(0, b)$ on peut atteindre le point $(1, b)$ en suivant la courbe intégrale du champ X^1 jusqu'au point $(1/(2n), b + 1/(2n))$, puis la courbe intégrale du champ X^2 jusqu'au point $(2/(2n), b)$ et ainsi de suite. Le temps employé à la fin du chemin est égal à $b + 1/(4n)$. On a donc $d((0, b), (1, b)) = b$ mais $(1, b) \notin R(b, (0, b))$.

Exemple 2.4. Soit $M = \mathbb{R}^2 - \{0\}$ et considérons les champs

$$X^1(x, y) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad X^2(x, y) = \begin{pmatrix} 1 \\ f(x) \end{pmatrix},$$

où

$$f(x) = \begin{cases} (x+2)^2 & \text{si } x < -2, \\ 0 & \text{si } -2 \leqq x \leqq 2, \\ (x-2)^2 & \text{si } x > 2. \end{cases}$$

Soit $D = (\pm X^1, \pm X^2)$. En tout voisinage de $(-1, 0)$ il y a des points qu'on peut atteindre de $(1, 0)$ à un instant $T < 5$, mais le point $(-1, 0)$ n'est pas atteignable de $(1, 0)$ quel que soit $T < 5$. Il s'avère donc que l'inclusion $\{(x, y): d((x, y), (1, 0)) \leqq 5\} \subset \operatorname{adh} R(5, (1, 0))$ est propre.

3. La fonction du temps minimum. Soit $D=(X^i(\cdot))_{i\in I}$ une famille symétrique de champs de vecteurs complets C^ω sur M, et soit S une orbite de D en M. Il convient de penser M comme l'ensemble des états phisiques d'un système guidable, dont les champs de D représentent les possibles évolutions. Nous dénotons par m_0 un point fixé en S, qu'on assume comme "état désiré" par le système, c'est-à-dire l'état qu'on désire d'atteindre. Le problème du temps minimum consiste à déterminer (s'ils existent) pour chaque $m\in S$ donné, $k\in\mathbb{N}$, $\xi\in I^k$, $\mathbf{t}\in\mathbb{R}^k$ qui minimisent la valeur de $\|\mathbf{t}\|_1$, sous la condition que $\rho_{\xi,m}(\mathbf{t})=m_0$. Soit $T>0$; en étant D symétrique, l'ensemble des points desquels m_0 est atteignable à l'instant T coïncide avec $R(T,m_0)$ défini par (2). Le problème du temps minimum a donc solution pour un point m donné en S, si et seulement si il existe $T>0$ tel que

$$m\in R(T,m_0) \quad \text{et} \quad m\notin R(t,m_0), \quad \text{quel que soit } t<T. \tag{9}$$

Il y a une vaste littérature sur le problème du temps minimum. Ici nous rappelerons [4], [9], [10], où les résultats sont essentiellement fondés sur des théorèmes de compacité et de continuité des ensembles des points atteignables. On doit à Filippov [7] le théorème d'existence de solutions du problème du temps minimum dans le cas

$$\dot{y}=f(t,y,\boldsymbol{u}), \qquad t\geqq 0, \quad y\in\mathbb{R}^n, \quad \boldsymbol{u}\in\Omega\subset\mathbb{R}^h, \tag{10}$$

où $f(t,y,\boldsymbol{u})$ est continue et différentiable par-rapport à y, Ω est compact, $t\mapsto\boldsymbol{u}(t)$ est mesurable et l'ensemble $V(t,y)=f(t,y,\Omega)$ est convexe. Un exemple en [6] (voir aussi. [11]) montre qu'on ne peut pas se passer de cette dernière hypothèse.

La fonction

$$m\mapsto T_0(m)=d(m,m_0)\colon S\to\mathbb{R}_+ \tag{11}$$

où $(m,m_0)\mapsto d(m,m_0)$ est donnée par la (3), s'appelle la fonction du temps minimum. Les proprietés de la fonction du temps minimum sont étudiées en [8] par rapport à un système linéaire de la forme

$$\dot{y}=Ay+B\boldsymbol{u}, \qquad y\in\mathbb{R}^n, \quad \boldsymbol{u}\in\Omega\subset\mathbb{R}^h, \tag{12}$$

où A et B sont matrices constantes, $t\mapsto\boldsymbol{u}(t)$ est mesurables et

$$\Omega=\{\boldsymbol{u}=(u_1,\cdots,u_h)\in\mathbb{R}^h\colon |u_i|\leqq 1, i=1,\cdots,h\}, \tag{13}$$

et par rapport à l'état désiré $m_0=0\in\mathbb{R}^n$. Le résultat plus important que nous démontrerons à propos de la fonction (11) est le théorème suivant.

THÉORÈME 3.1. *Soit D une famille symétrique de champs de vecteurs complets, C^ω sur M. Si D est localement bornée et si chaque point de M est autoaccessible à tout instant $T>0$ par rapport à D, alors S est une composante connexe de M, et la fonction du temps minimum* (11) *est continue dans la topologie de M sur S.*

Démonstration. On déduit la première affirmation comme dans le Théorème 2.3. Soit alors $m_1\in S$, soit $\varepsilon>0$ et considérons l'ensemble $R(\varepsilon,m_1)$. Par l'hypothèse de autoaccessibilité il existe un voisinage U de m_1 en M contenu en $R(\varepsilon,m_1)$; donc la fonction $m\mapsto d(m,m_1)$ de S en $\mathbb{R}_+$, est continue dans le point m_1, quel que soit $m_1\in S$. La continuité de la fonction (11) est alors une conséquence de l'inégalité

$$|T_0(m)-T_0(m_1)|=|d(m,m_0)-d(m_1,m_0)|\leqq d(m,m_1). \qquad \square$$

Dans le cas (12), (13) la continuité de la fonction du temps minimum est montrée en [8] en utilisant le fait que, dans le cas (12), (13) l'ensemble $R(T,m_0)$ est fermé.

COROLLAIRE 3.2. *Suppose de nous trouver dans les hypothèses du Théorème* 3.1. *Supposons en plus que M soit connexe et que chaque sous-ensemble propre de M borné* *dans la métrique δ soit contenu dans un sous-ensemble compact de M (telle hypothèse est verifiée par exemple lorsque M est compacte, ou lorsque $M=\mathbb{R}^n$). Quel que soit $T_1>0$ tel que $R(T_1, m_0)$ soit un sous-ensemble propre de M, il existe $m\in R(T_1, m_0)$ tel que $T_0(m)=T_1$.*

Démonstration. On sait que, dans nos hypothèses, $S=M$. Soit $T_1>0$ tel que $R(T_1, m_0)$, et donc la boule $\{m\colon d(m, m_0)<T_1\}$, sont des sous-ensembles propres de M. Démontrons d'abord que, quel que soit $T>T_1$, la boule $\{m\colon d(m, m_0)<T_1\}$ est aussi un sous-ensemble propre de la boule $\{m\colon d(m, m_0)<T\}$. À cette fin, allons observer que, en étant M connexe, la boule ouverte $\{m\colon d(m, m_0)<T\}$ est un sous-ensemble propre de son adhérence; donc l'ensemble

$$\partial\{m\colon d(m, m_0)<T_1\}\subset\{m\colon d(m, m_0)=T_1\}$$

est non vide. Chaque point p tel que $d(p, m_0)=T_1$ est atteignable de m à tout instant $T>T_1$. En conclusion, $p\in\{m\colon d(m, m_0)<T\}$ si $T>T_1$, mais $p\notin\{m\colon d(m, m_0)<T_1\}$. Puisque la boule ouverte $\{m\colon d(m, m_0)<T_1\}$ est bornée en (M, δ), il existe un compact K qui la contient, et l'ensemble $K\backslash\{m\colon d(m, m_0)<T_1\}$ est encore un compact de M: puisque la fonction du temps minimum est continue, elle admet un minimum sur ce dernier ensemble (voir par exemple [6, p. 64]). Soit T^* la valeur de ce minimum; il est clair que $T^*\geqq T_1$, et l'on voit par l'absurde que $T^*=T_1$. En effect dans le cas contraire on aurait, à tout instant T tel que $T_1<T<T^*$,

$$\{m\colon d(m, m_0)<T_1\}=\{m\colon d(m, m_0)<T\}. \qquad \square$$

Le Corollaire 3.2 peut s'énoncer dans la manière suivante: dans les conditions posées sur D, M et T_1, il existe $m\in M$ tel que le problème du temps minimum avec les donnés m et m_0 a solution égale à T_1. On observe que l'hypothèse que M soit connexe n'est pas essentielle: il suffit de nous rapporter aux composantes connexes de M.

4. Quelques propriétés de l'ensemble des points atteignables. Le résultat le plus important de cette section concerne la continuité de la fonction multivoque $t\mapsto R(t, m_0)$. On dit qu'une fonction multivoque $t\mapsto F(t)$ de variable réelle qui prend ses valeurs dans un espace métrique E est continue au sens de Hausdorff si, quel que soit $\varepsilon>0$ il existe $\tau>0$ tel que $H(F(t), F(t_0))<\varepsilon$ toutes les fois que $|t-t_0|<\tau$. La "distance" $H(\cdot,\cdot)$ que nous employons ici est définie comme en [6, p. 61]: on doit observer que $H(\cdot,\cdot)$ est une distance au sense propre seulement sur la classe des sous-ensembles fermés de E.

THÉORÈME 4.1. *Soit D une famille symétrique et localement bornée sur la variété M et supposons que chaque point de M soit autoaccessible à tout instant $T>0$ par rapport à D. Quel que soit $m_0\in M$, la fonction multivoque $t\mapsto R(t, m_0)$ est continue au sens de Hausdorff dans l'espace (M, δ) à tout instant $t>0$.*

Démonstration. Supposons d'abord $t>t_0>0$, de façon que $R(t_0, m_0)\subset R(t, m_0)$ et $H(R(t, m_0), R(t_0, m_0))=\sup\{d(m, R(t_0, m_0))\colon m\in R(t, m_0))\}$. Soit $m_1\in R(t, m_0)$ et soient k, $\xi\in I^k$ et $\mathbf{t}\in\mathbb{R}^k_+$ tels que $m_1=\rho_{\xi,m_0}(\mathbf{t})$, avec $\|\mathbf{t}\|_1=t>t_0$. Construisons la courbe $\vartheta\mapsto\gamma(\vartheta)$ pour $\vartheta\in[0, t]$ comme dans la démonstration du Théorème 2.2, (7). Le point $m_2=\gamma(t_0)$ appartient à $R(t_0, m_0)$ et on a

$$d(m_1, m_2)\leqq t-t_0.$$

Donc $H(R(t, m_0), R(t_0, m_0))\leqq t-t_0$. On complète la démonstration en changeant les rôles de t_o et t. $\square$

La continuité de la fonction $t \mapsto R(t, m_0)$ est montrée en [4] dans le cas linéaire (12), (13) et en [9] dans les cas général (10), avec des hypothèses restrictives sur $f(t, \mathbf{y}, \mathbf{u})$.

Reprenons maintenant l'étude des ensembles des points atteignables par rapport aux boules ouvertes ou fermées de la topologie δ.

PROPOSITION 4.1. *Soit D une famille symétrique et localement bornée, avec la propriété que chaque point de M est autoaccessible à tout instant $T > 0$. On a alors, quel que soit $T > 0$ et quel que soit $m_0 \in M$,*

$$\mathrm{adh}_M R(T, m_0) = \{m \colon d(m, m_0) \leqq T\} = \{m \colon T_0(m) \leqq T\}.$$

Démonstration. Par la Proposition 2.3 (ii), nous avons déjà l'inclusion $\mathrm{adh}_M R(T, m_0) \supset \{m \colon T_0(m) \leqq T\}$. Il suffit donc de montrer que si $d(m, m_0) > T$ alors $m \notin \mathrm{adh}_M R(T, m_0)$. En étant m autoaccessible à tout instant, quel que soit $\varepsilon < d(m, m_0) - T$ il existe un ouvert U de M tel que $m \in U \subset R(\varepsilon, m)$. S'il y avait en U un point m' de $R(T, m_0)$ on pourrait atteindre m de m_0 à travers m' en temps $T + \varepsilon < d(m, m_0)$: le résultat est donc obtenu par l'absurde. □

PROPOSITION 4.2. *Dans les hypothèses de la Proposition* 4.1, *quels que soient $T > 0$ et $m_0 \in M$, on a*

$$\partial R(T, m_0) \subset \{m \colon d(m, m_0) = T\} = \{m \colon T_0(m) = T\}.$$

Démonstration. Il est clair par la Proposition 4.1 que $\partial R(T, m_0) \subset \{m \colon T_0(m) \leqq T\}$. Soit $m \in \partial R(T, m_0)$, et supposons $T_0(m) < T$. Quel que soit ε positif, par l'hypothèse d'autoaccessiblité il existe un ouvert U tel que $m \in U \subset R(\varepsilon, m)$: si l'on choisit $\varepsilon < T - T_0(m)$, on a $U \subset R(T, m_0)$, c'est-à-dire $m \notin \partial R(T, m_0)$. En conclusion $T_0(m) = T$. □

Supposons encore de nous trouver dans les conditions de la Proposition 4.1: l'inclusion

$$(14) \qquad \{m \colon d(m, m_0) < T\} = \{m \colon T_0(m) < T\} \subset \mathrm{int}_M R(T, m_0)$$

est triviale. Dans l'exemple suivant la (14) est une inclusion au sens propre: même dans cet exemple on utilise un champ non complet.

Exemple 4.1. Soit M la surface comprise entre deux sections planes orthogonales d'un cylindre de $\mathbb{R}^3$: nous dénotons les points de M avec des couples de nombres (x, y), $-1 < x \leqq 1$ et $-1 < y < 1$. Considérons les champs sur M

$$X^1(x, y) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad X^2(x, y) = \begin{pmatrix} \dfrac{1}{1-y} \\ 0 \end{pmatrix}, \qquad X^3(x, y) = \begin{pmatrix} \dfrac{1}{1+y} \\ 0 \end{pmatrix}$$

et soit $D = (\pm X^1, \pm X^2, \pm X^3)$. L'ensemble $R(1, (0, 0))$ est M tout entière, donc $\mathrm{int}\, R(1, (0, 0)) = M$, mais les points de la forme $(1, y)$ ne sont pas atteignables à $\mathrm{T} < 1$. On note que dans cet exemple, $\partial R(1, (0, 0)) = \varnothing$ tandis que $\{m \colon T_0(m) = 1\} = \{(x, y) \in M \colon x = 1\}$.

La proposition qui va suivre est une conséquence presque immédiate des définitions.

PROPOSITION 4.3. *Dans les hypothèses de la Proposition* 4.1, *on a*

$$(15) \qquad \mathrm{int}_M R(T, m_0) = \{m \colon T_0(m) < T\}$$

si et seulement si $\partial R(T, m_0) = \{m \colon T_0(m) = T\}$.

Les propositions 2.3, 4.1, 4.2 et 4.3 donnent une généralisation du Corollaire 3 de [8]. En particulier, la Proposition 4.3 donne une condition d'extrêmalité pour les trajectoires optimales (voir [4], [9], [10] pour l'idée d'extrêmalité). Supposons en effect que m_0 soit atteignable de m_1 en temps minimum T: si (15) est verifiée, on a que $m_1 \in R(T, m_0) \cap \partial R(T, m_0) = R(T, m_0) \backslash \{m: T_0(m) < T\}$. Dans le cas linéaire (12), (13) la (15) est toujours vérifiée (voir [8]): il suit alors le théorème classique d'extrêmalité.

On remarque enfin que, dans l'hypothèse (15) une trajectoire extrêmale à un certain instant $T > 0$, est extrêmale même à tout instant $t < T$.

Remerciement. Je remercie Prof. C. Lobry pour son attention envers mon travail. Je remercie aussi Prof. H. Hermes, doct. Mme G. Stefani et doct. J. B. Costal pour leurs suggestions utiles.

BIBLIOGRAPHIE

[1] A. BACCIOTTI, *Autoaccessibilité par familles symétriques de champs de vecteurs*, Ricerche di Automatica, 7 (1976), pp. 189–197.

[2] N. BOURBAKI, *Eléments de mathématique*, livre III, chap. 9, tome 8, Hermann, Paris, 1958.

[3] Y. CHOQUET-BRUHAT, *Géométrié différentielle et systèmes exterieurs*, Dunod, Paris, 1968.

[4] R. CONTI, *Problemi di controllo e di controllo ottimale*, UTET, Torino, 1974.

[5] ——— *Sul principio del bang-bang per i processi di controllo bilineari*, Matematiche, 30 (1975), pp. 363–371.

[6] J. DIEUDONNÉ, *Foundations of modern analysis*, Academic Press, New York, 1969.

[7] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, this Journal, 1 (1962), pp. 76–84.

[8] O. HAJEK, *Geometric theory of time-optimal control*, this Journal, 9 (1971), pp. 339–350.

[9] H. HERMES AND J. P. LASALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.

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

[11] C. LOBRY, *Contrôlabilité des systèmes non linéaires*, this Journal, 9 (1971), pp. 573–605.

[12] P. STEFAN, *Accessible sets, orbits, and foliations with singularities*, Proc. London Math. Soc. Ser. A, 29 (1974), pp. 699/713.

[13] H. J. SUSSMANN, *Orbits of families of vectors fields and integrability of distributions*, Trans. Amer. Math. Soc., 180 (1973), pp. 171–188.

[14] H. J. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 3, May 1979

1979 Society for Industrial and Applied Mathematics
0363-0129/79/1703-0001 $01.00/0

LIPSCHITZ CONTINUITY FOR CONSTRAINED PROCESSES*

WILLIAM W. HAGER†

Abstract. We study Lipschitz continuity properties for "constrained processes". As applications of our general theory, we consider mathematical programs and optimal control problems. We show that if the gradients of the binding constraints satisfy an independence condition, then the solution and the dual multipliers of a convex mathematical program are a Lipschitz continuous function of the data. Similarly, it is proved that the optimal control and the dual multipliers for strictly convex control problems with convex constraints on the state and the control are Lipschitz continuous in time. In both applications, estimates of the Lipschitz constant are given .

1. Introduction. We show that a constrained process is Lipschitz continuous on a convex domain if it is Lipschitz continuous on compatible pairs of elements in the domain. As applications of this result, we prove that if the gradients of the binding constraints satisfy an independence condition, the solution to a convex mathematical program is a Lipschitz continuous function of its data, and the solution to a strictly convex control problem with convex constraints on the state and the control is Lipschitz continuous in time. In both cases, the Lipschitz constant can be estimated.

The control regularity results provide the mathematical foundation for estimating the error in discrete approximations to control problems, and also give insight into the practicality of approximation schemes. As is well known to numerical analysts, the convergence rate for piecewise polynomial approximation is limited by regularity—if a problem's solution has only two derivatives in L^2, there is usually no advantage in using approximating piecewise polynomials of degree >1 (unless the structure of the solution is known). Also estimates of the error in approximation involve bounds on the derivatives of the function being approximated. Hence the bounds given below for the Lipschitz constants are a basis for a priori error estimates. Error estimates for the Ritz–Trefftz method are derived in [4].

2. The abstract problem. Let $\mathcal{S}$ be a Banach space, $\mathcal{D}$ be a convex subset of a Banach space, and $z: \mathcal{D} \to \mathcal{S}$ be continuous. Moreover, let $c: \mathcal{D} \to 2^{\{1, \cdots, n\}}$ = power set of $\{1, \cdots, n\}$ have the following property:

$$\text{If } \{d_k\} \subset \mathcal{D},\ d_k \to d \in \mathcal{D} \text{ as } k \to \infty, \text{ and} \\ I \subset c(d_k) \text{ for all } k, \text{ then } I \subset c(d). \tag{2.1}$$

In the applications, d represents the data of a program, $z(d)$ is the associated "solution," and $c(d)$ gives the "binding constraints" for $z(d)$.

We use the following notation: Given $d, e \in \mathcal{D}$, let (d, e) denote the ordered pair and define the segment:

$$[d, e] = \{(1-\lambda)d + \lambda e : 0 \leqq \lambda \leqq 1\}. \tag{2.2}$$

If $(d, e) \in \mathcal{D} \times \mathcal{D}$ and $\delta_k = [(1-\lambda_k)d + \lambda_k e] \in [d, e]$ for $k = 1, 2$, we say that $\delta_1 > \delta_2$ if $\lambda_1 > \lambda_2$. The data $d, e \in \mathcal{D}$ is called *compatible* if $c(d) = c(e)$ and $c(\delta) \subset c(d)$ for all $\delta \in [d, e]$. Finally let $\# I$ be the number of elements in the set I.

* Received by the editors July 17, 1975, and in final revised form June 18, 1978.

† Department of Mathematics, Carnegie-Mellon University, Pittsburgh, Pennsylvania, 15213. This work was supported in part by the National Science Foundation under Grant MCS 75-09457, and in part by the Office of Naval Research under Grant N00014-76-C-0369.

THEOREM 2.1. *If γ satisfies*

$$\|z(d)-z(e)\|_{\mathcal{S}} \leqq \gamma \|d-e\|_{\mathcal{D}} \tag{2.3}$$

for all compatible data $(d, e) \in \mathcal{D} \times \mathcal{D}$, then γ satisfies (2.3) *for all data $(d, e) \in \mathcal{D} \times \mathcal{D}$.*

The norm subscripts $\mathcal{S}$ and $\mathcal{D}$ are generally omitted since the choice of norm should be clear from context.

Proof of Theorem 2.1. Define the following set:

$$\begin{aligned} T_m = \{(d, e) \in \mathcal{D} \times \mathcal{D}\colon & \text{ There exists } I \subset \{1, \cdots, n\} \\ & \text{with } \# I \leqq m \text{ and } c(\delta) \subset I \text{ for all } \delta \in [d, e]\}. \end{aligned} \tag{2.4}$$

Since $T_0 \subset T_1 \cdots \subset T_n = \mathcal{D} \times \mathcal{D}$, our goal is to establish the validity of (2.3) on T_n. All pairs in T_0 are compatible, so (2.3) holds on T_0. Proceed by induction and suppose that (2.3) holds on T_{m-1}. Given $(d,e) \in T_m$, we shall construct a function $G\colon [d,e] \rightarrow [d,e]$ with the following properties: If $s \in [d,e]$ and $t = G(s)$, then

$$\begin{aligned} &\text{(i) either } t > s \text{ or } t = s = e, \\ &\text{(ii) either } \# c(t) = m \text{ or } t = e, \\ &\text{(iii) } \|z(t) - z(s)\| \leqq \gamma \|t - s\|. \end{aligned} \tag{2.5}$$

Assuming the existence of G, the inductive step is completed as follows: First construct sequences $\{s_k\}$ and $\{t_k\}$ starting with $s_0 = d$ and for $k \geqq 1$ setting $t_k = G(s_{k-1})$ and

$$s_k, = \sup \{t\colon c(t) = c(t_k) \text{ and } t_k \leqq t \leqq e\} \tag{2.6}$$

(conceivably $s_k = t_k$). Stop these sequences when s_k or t_k reaches e. We now show that these sequences have at most two elements.

Since $c(t) \subset I$ for $t \in [d, e]$ and $\# \mathrm{I} \leqq m$, we see by (ii) that for $s \in [d, e]$, $c(G(s)) = I$ if $G(s) < e$. Hence either $c(t_1) = c(G(s_0)) = I$ or $t_1 = e$ (and the sequence terminates). If $t_1 < e$, then by (2.1), $I \subset c(s_1)$; but $c(s_1) \subset I$, so we obtain $c(s_1) = I$. If $s_1 = e$, the sequence terminates. If $s_1 < e$, then by the definition of s_1, we have $c(\sigma) \neq I$ for all $s_1 < \sigma \leqq e$. Therefore, by (i) we have $t_2 > s_1$, and by (ii) we conclude that $t_2 = e$. Thus the sequences $\{s_k\}$ and $\{t_k\}$ have at most two elements as claimed.

By (iii) above, we have for all k:

$$\|z(t_k) - z(s_{k-1})\| \leqq \gamma \|t_k - s_{k-1}\|. \tag{2.7}$$

Now consider the interval $[t_k, s_k]$ when $s_k > t_k$. By the definition of s_k, there exists a sequence $\{\sigma_j\} \subset [d, e]$ such that $\sigma_j \rightarrow s_k$ as $j \rightarrow \infty$ and $c(\sigma_j) = c(t_k)$ for all j (conceivably $\sigma_j = s_k$ for all j). By (ii), we see that $\# c(t_k) = m = \# c(\sigma_j)$ and by the structure of T_m, we conclude that $c(t) \subset c(t_k) = c(\sigma_j)$ for all $t \in [t_k, \sigma_j]$. Thus σ_j, t_k are compatible and $\|z(t_k) - z(\sigma_j)\| \leqq \gamma \|t_k - \sigma_j\|$. Since z is continuous, we let $j \rightarrow \infty$ to obtain:

$$\|z(t_k) - z(s_k)\| \leqq \gamma \|t_k - s_k\|. \tag{2.8}$$

The triangle inequality, (2.7)–(2.8), and the ordering $s_0 \leqq t_1 \leqq s_1 \leqq t_2$ give us $\|z(d) - z(e)\| \leqq \gamma \|d - e\|$–the inductive step has been completed.

Now consider the construction of G. First set $G(e) = e$. Given $s \in [d, e]$ with $s < e$, we consider two cases:

Case 1. There exists $\{s_j\} \subset [d, e]$ such that $s_j \rightarrow s$ as $j \rightarrow \infty$, and for all j, both $\# c(s_j) = m$ and $s_{j+1} < s_j$.

In this case, define $G(s)=s_1$. By the structure of T_m, we conclude that s_1 and s_j are compatible for all j. Therefore, $\|z(s_j)-z(s_1)\| \leqq \gamma\|s_j-s_1\|$ for all j, and letting $j\to\infty$, we have $\|z(s)-z(s_1)\|\leqq\gamma\|s-s_1\|$. Hence (i)–(iii) are satisfied for $t=s_1=G(s)$.

Case 2. There exists an interval $[s,\rho]$ such that $s<\rho$ and $\# c(t)<m$ for all $s<t\leqq\rho$. In this case, define $G(s)=t$ where

$$t=\sup\{\tau: s<\tau<e,\ \# c(\sigma)<m \text{ for all } \sigma\in[s,\tau],\ \sigma\neq s\}.$$

Given $\tau\in[d,e]$, let $B(\Delta,\tau)$ be the intersection between $[d,e]$ and the ball with center τ and radius Δ. Property (2.1) implies the existence of $\Delta_\tau>0$ such that $c(\sigma)\subset c(\tau)$ for all $\sigma\in B(\Delta_\tau,\tau)$. Since $\# c(\tau)<m$ for $s<\tau<t$, we have $(\sigma_1,\sigma_2)\in T_{m-1}$ if $[\sigma_1,\sigma_2]\subset B(\Delta_\tau,\tau)$. Applying the induction hypothesis gives $\|z(\sigma_1)-z(\sigma_2)\|\leqq\gamma\|\sigma_1-\sigma_2\|$. Hence for all s_1, t_1 satisfying $s<s_1<t_1<t$, there exists a finite covering of $[s_1,t_1]$ using intervals $[\sigma_j,\sigma_{j+1}]$ where $s_1=\sigma_1<\sigma_2\cdots<\sigma_l=t_1$ and $(\sigma_j,\sigma_{j+1})\in T_{m-1}$ for all j. Applying the triangle inequality across these intervals, we get $\|z(s_1)-z(t_1)\|\leqq\gamma\|s_1-t_1\|$. Letting $s_1\to s$ and $t_1\to t$, we see that (i)–(iii) hold. □

Remark 2.2. Suppose that $\|d\|_{\mathscr{D}}=\|d\|_1+\|d\|_2$ where $\|\cdot\|_1$ and $\|\cdot\|_2$ are seminorms. It is easy to see that Theorem 2.1 remains valid if the right side of (2.3) is replaced by $\gamma_1\|d-e\|_1+\gamma_2\|d-e\|_2$.

The assumptions of Theorem 2.1 can be weakened to the following:

THEOREM 2.3. *Suppose that for some* $\varepsilon>0$, γ *satisfies*

$$\|z(d)-z(e)\|_{\mathscr{S}}\leqq\gamma\|d-e\|_{\mathscr{D}} \tag{2.9}$$

for all compatible data $(d,e)\in\mathscr{D}\times\mathscr{D}$ *for which* $\|d-e\|_{\mathscr{D}}\leqq\varepsilon$. *Then* γ *satisfies* (2.9) *for all data* $(d,e)\in\mathscr{D}\times\mathscr{D}$.

Proof. Let T_m be the set defined in (2.4). All pairs $(d,e)\in T_0$ are compatible; moreover, if $[\delta_1,\delta_2]\subset[d,e]$, then δ_1 and δ_2 are compatible. Thus $[d,e]$ can be expressed as the union of subintervals of length $\leqq\varepsilon$ and the endpoints of each subinterval are compatible. Applying (2.9) to each subinterval, we see that (2.9) holds for all $(d,e)\in T_0$.

Proceed by induction and suppose that (2.9) holds for all $(d,e)\in T_{m-1}$. Given $(d,e)\in T_m$, we observe that $(\delta_1,\delta_2)\in T_m$ if $[\delta_1,\delta_2]\subset[d,e]$. Exactly as in the proof of Theorem 2.1, $(\delta_1,\delta_2)\in T_m$ satisfies (2.9) if $\|\delta_1-\delta_2\|\leqq\varepsilon$. Expressing $[d,e]$ as the union of intervals with the length of each interval $\leqq\varepsilon$, we see that (2.9) holds for all $(d,e)\in T_m$—the inductive step has been completed. □

3. Quadratic programs. As an application of Theorem 2.1, consider the following quadratic program:

$$\text{(QP)}\qquad \begin{aligned}&\text{minimize} && \tfrac{1}{2}v^TRv+r^Tv\\ &\text{subject to} && Av+a\leqq 0,\quad Bv+b=0,\quad v\in R^n,\end{aligned}$$

where all matrices and vectors have compatible size (capital letters denote matrices and small letters denote vectors). If the gradients of the binding constraints for (QP) satisfy an independence condition, we prove that both the solution and the dual multiplier are Lipschitz continuous functions of the data $(R,r,\cdots,b)$. The extension of our results to more general convex programs is stated in Appendix D.

For related results, see papers [1], [2], [8] which differ from our results in the following aspects: Daniel [1] does not obtain Lipschitz continuity and does not consider the dual multipliers. Both Fiacco [2] and Robinson [7] require that strict complementary slackness and the second order sufficiency condition hold. We eliminate the former and relax the latter assumption. Finally our results are not local and the Lipschitz constant is estimated.

We use the following notation: Let $\|\cdot\|$ denote the Euclidean vector norm; this vector norm generates a matrix norm given by $\|E\| = \max\{\|Ev\| : \|v\| = 1\}$. If $S = (S_1, \cdots, S_n)$ and $T = (T_1, \cdots, T_n)$ are ordered n-tuples of matrices with each pair (S_j, T_j) having the same dimensions, define $S - T = (S_1 - T_1, \cdots, S_n - T_n)$, $cS = (cS_1, \cdots, cS_n)$, and $\|S\| = \|S_1\| + \cdots + \|S_n\|$. If J is a subset of the row indices of A, let A_J denote the submatrix consisting of those rows corresponding to elements of J. Similarly let v_J denote the vector with components v_j for $j \in J$.

Let $d = (R, r, A, a, B, b)$ be the data for (QP) and define

$$M(I, d) = \begin{bmatrix} A \\ B \end{bmatrix}_I,$$

$\delta(d)^T = (r^T, a^T, b^T)$, and $\Delta(d) = (R, A, B, A^T, B^T)$. Our analysis is restricted to programs for which there is a unique solution $u(d)$ and a unique dual multiplier $\lambda(d)$. Letting $J(d)$ denote the indices corresponding to binding constraints for $u(d)$ (including equality constraints), we define $M(d) = M(J(d), d)$.

Let D be any convex set of data satisfying the following conditions:

A1. For all $d \in D$, there exists a unique solution $u(d)$ to (QP).

A2. There exists $\Gamma_1, \Gamma_2 < \infty$ such that $\|R\| < \Gamma_1$ and $\|M(d)^T\| < \Gamma_2$ for all $d \in D$.

A3. There exist $\alpha, \beta > 0$ such that for all $d \in D$

$$v^T R v \geqq \alpha \|v\|^2 \quad \text{for all } v \text{ satisfying } M(d)v = 0, \tag{3.1}$$

$$\|M(d)^T \lambda\| \geqq \beta \|\lambda\| \quad \text{for all } \lambda. \tag{3.2}$$

Defining $\kappa_1 = \Gamma_1/\alpha$, $\kappa_2 = \Gamma_2/\beta$, and $\kappa_3 = \max\{\Gamma_1/\beta, 1\}$, we shall establish:

THEOREM 3.1. *There exists a constant $\rho < \infty$ such that for all $d_1, d_2 \in D$, we have*:

$$\begin{aligned} &\|u(d_1) - u(d_2)\| \text{ and } \|\lambda(d_1) - \lambda(d_2)\| \\ &\leqq \rho \|\delta(d_1) - \delta(d_2)\| + \rho^2 \|\Delta(d_1) - \Delta(d_2)\| (\|\delta(d_1)\| + \|\delta(d_2)\|). \end{aligned} \tag{3.3}$$

Moreover, $\rho \leqq \rho_m \equiv \alpha^{-1} + 2\kappa_1 \beta^{-1} + 4\kappa_1\kappa_2\kappa_3$.

As $\alpha \to 0$ or $\beta \to 0$, the bound ρ_m for ρ has the correct asymptotic behavior. For example, consider an unconstrained program with R positive definite. Let δr be a perturbation in r that is colinear with the eigenvector of R corresponding to the smallest eigenvalue α. The perturbation in the solution of (QP) satisfies $\|\delta u\| = \alpha^{-1}\|\delta r\|$; similarly $\rho_m \sim \alpha^{-1}$.

On the other hand, suppose that the binding constraints are nearly dependent as in the following program:

$$\begin{aligned} &\text{minimize} \quad \tfrac{1}{2}[v_1^2 + (v_2 - 1)^2] \\ &\text{subject to} \quad \begin{bmatrix} 0 & 1 \\ -\varepsilon & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \leqq \begin{bmatrix} 0 \\ a \end{bmatrix} \end{aligned} \tag{3.4}$$

where $-\varepsilon^2 \leqq a \leqq 0$. The solution to (3.4) lies at the corner of the feasible set. If λ_2 is the dual multiplier associated with the second constraint, the perturbation $\delta\lambda_2$ corresponding to a perturbation δa in the data satisfies $|\delta\lambda_2| = |\delta a|/\varepsilon^2$. Similarly $\rho_m \sim 1/\varepsilon^2$ since $\beta \sim \varepsilon$ and $\kappa_2\kappa_3 \sim 1/\varepsilon^2$.

We apply Theorem 2.1 using $z(d)^T=[u(d)^T,\lambda(d)^T]$ and $c(d)=J(d)$. Hence we must

(3.5) (i) establish the continuity of $u(\cdot)$ and $\lambda(\cdot)$ on D

and

(3.6) (ii) compute the constants γ_1 and γ_2 in Remark 2.2 where $\|d\|_1=\|\delta(d)\|$ and $\|d\|_2=\|\Delta(d)\|$.

(Since the constraints are continuous functions, (2.1) follows immediately from (i)). Let us begin with the computational question. First observe that the necessary conditions for (QP) can be expressed in the form:

$$N(d)\hat{z}(d)=f(d) \quad \text{where } \hat{z}(d)=\begin{bmatrix} u(d) \\ \lambda(d)_{J(d)} \end{bmatrix},$$
$$-f(d)=\begin{bmatrix} r \\ a_{J(d)} \\ b \end{bmatrix}, \quad \text{and} \quad N(d)=\begin{bmatrix} R & M(d)^T \\ M(d) & 0 \end{bmatrix}. \tag{3.7}$$

Lemma 3.2. *For all $d\in D$, $N(d)$ is nonsingular and $\|N(d)^{-1}\|\leqq\rho_m$.*

Proof. Given $d\in D$, let $M=M(d)$ and let e and z be related by:

$$e=\begin{bmatrix} e_1 \\ e_2 \end{bmatrix}=\begin{bmatrix} R & M^T \\ M & 0 \end{bmatrix}\begin{bmatrix} u \\ \lambda \end{bmatrix}, \quad \text{and} \quad z=\begin{bmatrix} u \\ \lambda \end{bmatrix}. \tag{3.8}$$

Since $\|N^{-1}\|=1/\min\{\|Nz\|:\|z\|=1\}$, we must show that $\|e\|\geqq\|z\|/\rho_m$ for all e and z satisfying (3.8).

Begin by expressing $u=u^p+u^\perp$ where $Mu^p=0$ and $u^\perp$ is perpendicular to the null space of M. Since $u^\perp$ lies in the range space of M^T, there exists y such that $u^\perp=M^Ty$, and

$$\|u^\perp\|\leqq\|M^T\|\,\|y\|\leqq\Gamma_2\|y\|. \tag{3.9}$$

Using A3, it is easily seen that $\beta\leqq$ smallest eigenvalue of MM^T; therefore, $\beta\|y\|\leqq\|MM^Ty\|$ and by (3.9), we have:

$$\|e_2\|=\|Mu\|=\|Mu^\perp\|=\|MM^Ty\|\geqq\beta\|y\|\geqq\frac{\beta}{\Gamma_2}\|u^\perp\| \tag{3.10}$$

or

$$\|u^\perp\|\leqq\kappa_2\|e\|. \tag{3.11}$$

Multiplying the first equation in (3.8) by u^p gives us

$$\begin{aligned}\|e_1\|\,\|u^p\|&\geqq(u^p)^TRu^p+(u^p)^TRu^\perp+(u^p)^TM^T\lambda\\ &\geqq\alpha\|u^p\|^2-\|u^p\|\,\|R\|\,\|u^\perp\|.\end{aligned} \tag{3.12}$$

Dividing by $\|u^p\|$ and applying (3.11), we get

$$\|u^p\|\leqq(\kappa_1\kappa_2+\alpha^{-1})\|e\|. \tag{3.13}$$

Again the first equation in (3.8) implies that

$$\|e_1\|\geqq\beta\|\lambda\|-\Gamma_1(\|u^p\|+\|u^\perp\|). \tag{3.14}$$

Inserting (3.11) and (3.13) into (3.14) and noting that κ_1, κ_2, $\kappa_3 \geqq 1$, we obtain

$$\|\lambda\| \leqq 2(\kappa_1\beta^{-1} + \kappa_1\kappa_2\kappa_3)\|e\|. \tag{3.15}$$

The proof is completed using the triangle inequality $\|z\| \leqq \|u^p\| + \|u^\perp\| + \|\lambda\|$ and the relations (3.11), (3.13), and (3.15). □

Observe that the mapping $d \to [u(d), \lambda(d)]$ is well defined. That is, by A1, (QP) has a unique solution $u(d)$ and consequently the binding constraint set is uniquely determined. If λ is any optimal dual multiplier, then $\hat{z}^T = \{u(d)^T, \lambda_{J(d)}^T]$ satisfies $N(d)\hat{z} = f(d)$. By Lemma 3.2, this equation has a unique solution, and by complementary slackness, the remaining components of λ are zero.

To estimate γ_1 and γ_2, let e_1, $e_2 \in D$ be compatible data; hence $J(e_1) = J(e_2)$. Defining $N_k = N(e_k)$, $\hat{z}_k = \hat{z}(e_k)$, and $f_k = f(e_k)$ for $k = 1, 2$, we obtain

$$\hat{z}_2 - \hat{z}_1 = N_1^{-1}[(f_2 - f_1) + (N_1 - N_2)N_2^{-1}f_2]. \tag{3.16}$$

Taking norms and applying Lemma 3.2 gives us

$$\begin{aligned} \|\hat{z}_2 - \hat{z}_1\| &\leqq \rho_m\|f_2 - f_1\| + \rho_m^2\|N_1 - N_2\|\,\|f_2\| \\ &\leqq \rho_m\|\delta(e_2) - \delta(e_1)\| + \rho_m^2\|\Delta(e_1) - \Delta(e_2)\|\,\|f_2\|. \end{aligned} \tag{3.17}$$

For all $e_2 \in [d_1, d_2]$, observe that $\|f_2\| \leqq \|\delta(d_1)\| + \|\delta(d_2)\|$. If $u(\cdot)$ and $\lambda(\cdot)$ are continuous, the proof of Theorem 3.1 is completed by choosing $\mathscr{D} = [d_1, d_2]$ and combining Theorem 2.1, Remark 2.2, and equation (3.17).

We conclude by proving (3.5). Defining $C(v, d) = \frac{1}{2}v^TRv + r^Tv$, $F(d) = \{v : Av + a \leqq 0,\ Bv + b = 0\} =$ feasible set, and $C(d) = \inf\{C(v, d) : v \in F(d)\} =$ optimal cost, we have

LEMMA 3.3. *$C(\cdot)$, $u(\cdot)$, and $\lambda(\cdot)$ are continuous functions on D.*

Proof. By Robinson's work [7] and (3.2), we know that:

(3.18) Given $d \in D$, there exist constants $c_1, c_2, c_3 > 0$ such that for all v_1, d_1, d_2 satisfying $\|d_1 - d\| \leqq c_1 \geqq \|d_2 - d\|$, $v_1 \in F(d_1)$, and $(1 + \|v_1\|)(\|d_1 - d\| + \|d_2 - d\|) \leqq c_2$, there exists $v_2 \in F(d_2)$ such that $\|v_1 - v_2\| \leqq c_3\|d_1 - d_2\|(1 + \|v_1\|)$.

Suppose that $\{d_k\} \subset D$ and $d_k \to d$ as $k \to \infty$. By (3.18), there exists $c > 0$ and $v_k \in F(d_k)$ such that $\|u(d) - v_k\| \leqq c\|d_k - d\|$ as $k \to \infty$. Thus Lim sup $C(d_k) \leqq C(d)$. Since $\|\delta(d_k)\|$ is uniformly bounded, $\|u(d_k)\| \leqq \rho_m\|\delta(d_k)\|$ is uniformly bounded. Applying (3.18) again, there exists $c > 0$ and $w_k \in F(d)$ such that $\|u(d_k) - w_k\| \leqq c\|d_k - d\|$ as $k \to \infty$. Hence $c(w_k, d) - C(u(d_k), d_k) \to 0$ as $k \to \infty$, Lim inf $C(d_k) \geqq C(d)$, and Lim $C(d_k) = C(d)$ as claimed.

Now suppose that $u(d_k)$ does not converge to $u(d)$. Since $\|u(d_k)\|$ is bounded uniformly, there exists a subsequence $\{d'_k\}$ and $v \neq u(d)$ such that $u(d'_k) \to v$ as $k \to \infty$. By (3.18), we see that $v \in F(d)$ and by the continuity of $C(\cdot)$, v is optimal in (QP). Hence A1 is violated and $u(d_k) \to u(d)$ as $k \to \infty$.

Finally consider $\lambda(\cdot)$. Since $u(\cdot)$ is continuous, $J(d_k) \subset J(d)$ for k sufficiently large. Letting $I = J(d)$, and (R_k, r_k) be the first two components of d_k, the following identity holds:

$$\begin{bmatrix} R_k & M(I,d_k)^T \\ M(I,d_k) & 0 \end{bmatrix}\begin{bmatrix} u(d_k) \\ \lambda(d_k)_I \end{bmatrix} = \begin{bmatrix} -r_k \\ M(I,d_k)u(d_k) \end{bmatrix}. \tag{3.19}$$

The right side of (3.19) converges to $f(d)$ as $k \to \infty$ while the matrix on the left side converges to the nonsingular matrix $N(d)$. Thus the solution to (3.19) converges to $\hat{z}(d)=N(d)^{-1}f(d)$; and in particular, $\lambda(d_k)_I\to\lambda(d)_I$ as $k\to\infty$. Since $J(d_k)\subset I$, the remaining components of $\lambda(d_k)$ and $\lambda(d)$ are zero by complementary slackness and we have $\lambda(d_k)\to\lambda(d)$ as $k\to\infty$. □

4. Optimal control regularity. As a second application of Theorem 2.1, we consider the solution regularity for a strictly convex control problem. The following notation is used for spaces of functions f: $[0,1]\to R^n$:

$C^p(R^n)$	Functions with continuous derivatives through order p.
$A(R^n)$	Absolutely continuous functions.
$BV(R^n)$	Functions of bounded variation that are left continuous on $[0, 1)$.
$L^\infty(R^n)$	Essentially bounded functions.
$L^p(R^n)$	Functions with $\int_0^1\|f(t)\|^p dt<\infty$.

The argument R^n above is omitted when the range is clear from context.

Consider the following problem:

$$\text{(CP)}\qquad \text{minimize}\quad \left\{C(x,u)=\int_0^1 f(x(t),u(t),t)dt\right\}$$

$$\text{(4.1)}\qquad \text{subject to}\quad \dot{x}(t)=A(t)x(t)+B(t)u(t)\quad\text{for almost every } t\in[0,1],$$

$$\left.\begin{array}{l}K_c(u(t),t)\leqq 0\\ K_s(x(t),t)\leqq 0\end{array}\right\}\text{ for all } t\in[0,1]$$

$$x(0)=x_0,\quad x\in A(R^n),\quad u\in L^2(R^m),$$

where K_c and K_s are vector valued with range in R^{m_c} and R^{m_s} respectively. (CP) is assumed to satisfy conditions (4.2)–(4.4) below, but first some notation is needed: Given two symmetric matrices M_1 and M_2, the statement $M_1>M_2$ means that M_1-M_2 is positive definite. If $g:R^{m_1}\times R^{m_2}\times\cdots\times R^{m_l}\to R$, we let $\nabla_j g$ and $\nabla_j^2 g$ denote the gradient and Hessian respectively of $g(y_1,\cdots,y_l)$ with respect to y_j where $y_k\in R^{m_k}$ for $k=1,\cdots,l$. We assume the following:

(4.2) A and B are Lipschitz continuous while f, K_c, K_s, and $\nabla_1 K_s(\cdot,\cdot)$ are C^2.

(4.3) Both $f(\cdot,\cdot,t)$ and the components of $K_s(\cdot,t)$ and $K_c(\cdot,t)$ are convex for all $t\in[0,1]$. Moreover, there exists $\alpha>0$ such that

$$\nabla_2^2 f(x,u,t)>\alpha I$$

for all $x\in R^n$, $u\in R^m$, and $t\in[0,1]$.

(4.4) There exists a continuous control $\bar{u}$, a corresponding trajectory $\bar{x}$, and a constant $\eta<0$ such that

$$K_c(\bar{u}(t),t)_j<\eta>K_s(\bar{x}(t),t)_i$$

for all $t\in[0,1]$, $j=1,\cdots,m_c$, and $i=1,\cdots,m_s$.

Using classical techniques in convex analysis, (4.2)–(4.4) imply for (CP) the existence of an optimal control $u^*\in L^2$ and a corresponding trajectory $x^*\in A$, and all optimal controls are equal almost everywhere. Furthermore, by Appendix A, $u^*\in L^\infty$.

The dual of (CP) is now introduced. Let $\langle \cdot, \cdot \rangle$ denote the L^2 inner product, and for $\nu \in BV(R^n)$, define the function $[\nu, \cdot]$ as follows:

$$[\nu, g] = \int_0^1 g(t)^T d\nu(t)$$

for all $g \in C^0(R^n)$. The *Lagrange dual function* associated with (CP) is given by

$$\begin{aligned}\mathscr{L}(p, \lambda, \nu) = \inf\{&C(x, u) + \langle p, \dot{x} - Ax - Bu\rangle \\ &+ \langle \lambda, K_c(u)\rangle + [\nu, K_s(x)] : x(0) = x_0, x \in A(R^n), u \in L^\infty(R^m)\}\end{aligned} \tag{4.5}$$

and the Lagrange dual to (CP) becomes:

(CD) $\quad \sup\{\mathscr{L}(p, \lambda, \nu) : (p, \nu) \in BV, \lambda \in L^1, \lambda \geqq 0, \nu(1) = 0, \nu \text{ nondecreasing}\}.$

Now recall our strong duality result [3] (which was actually proved under much weaker assumptions than those given above):

THEOREM 4.1. *If* (4.2)–(4.4) *hold, then there exist optimal solutions* (x^*, u^*) *to* (CP) *and* (p^*, λ^*, ν^*) *to* (CD). *Moreover,* (x^*, u^*) *achieve the minimum in* (4.5) *for* $(p, \lambda, \nu) = (p^*, \lambda^*, \nu^*)$ *and the following complementary slackness conditions hold*:

$$\langle \lambda^*, K_c(u^*)\rangle = 0 = [\nu^*, K_s(x^*)].$$

To obtain our regularity results, we strengthen assumption (4.4) and require uniform independence for the gradients of the binding constraints:

(4.6) There exists $\beta > 0$ such that for all $t \in [0, 1]$ and all z, we have:

$$\|[G_c(t)^T, B(t)^T G_s(t)^T]z\| \geqq \beta \|z\|$$

where $G_c(t)$ is the matrix whose rows are the gradients evaluated at $u^*(t)$ of components of $K_c(\cdot, t)$ corresponding to binding constraints for $u^*(t)$. The matrix $G_s(t)$ is defined similarly.

Defining the variable $q^*(\cdot) = \nabla_1 K_s(x^*(\cdot), \cdot)^T \nu^*(\cdot) - p^*(\cdot)$, our principal theorem is the following:

THEOREM 4.2. *If* (4.2)–(4.4) *and* (4.6) *hold, then there exist an optimal control* u^* *for* (CP), *a corresponding trajectory* x^*, *and optimal dual multipliers* (p^*, λ^*, ν^*) *such that* $(\dot{x}^*, \dot{q}^*, u^*, \lambda^*, \nu^*)$ *are Lipschitz continuous on* $[0, 1)$.

Remark 4.3. Using Lemma 4.4 and Remark 4.10, the Lipschitz constant can be estimated. The interval of Lipschitz continuity is $[0, 1)$ since ν may be discontinuous at $t = 1$. Theorem 4.2 is also valid when the system dynamics are only affine in the control (not necessarily the state). See Malanowski [6] for the modifications.

To prove Theorem 4.2, we apply Theorem 2.1 using $\mathscr{D} = (0, 1)$, $z(t)^T = (\lambda^*(t)^T, \nu^*(t)^T)$, and $c(t) =$ indices of binding constraints for $x^*(t)$ and $u^*(t)$. Hence we must

(i) establish that both ν^* and λ^* are continuous on $(0, 1)$ and

(ii) compute the constant γ in (2.3) for compatible data $d, e \in (0, 1)$.

The regularity of u^*, q^*, and x^* are obtained from the control minimum principle ((4.8) below), the adjoint equation ((4.7), below), and the system dynamics (4.1). Our analysis of (i)–(ii) is based on the necessary conditions for (CP) [3]:

$$\begin{aligned}&\dot{q}^*(t) = -A(t)^T(q^*(t) - G(t)^T\nu^*(t)) + \dot{G}(t)^T\nu^*(t)\\ &\qquad -\nabla_1 f(x^*(t), u^*(t), t) \quad \text{for almost every } t \in [0, 1], \quad \text{and}\\ &q^*(1) = 0 \quad \text{where } G(\cdot) = \nabla_1 K_s(x^*(\cdot), \cdot),\end{aligned} \tag{4.7}$$

$$\begin{aligned}&\nabla_2 f(x^*(t), u^*(t), t) + B(t)^T(q^*(t) - G(t)^T\nu^*(t))\\ &\qquad\qquad + \nabla_1 K_c(u^*(t), t)^T\lambda^*(t) = 0 \quad \text{for all } t \in [0, 1],\end{aligned} \tag{4.8}$$

$$\lambda^*(t)^T K_c(u^*(t), t) = 0 \quad \text{for all } t \in [0, 1], \tag{4.9}$$

and

$$[\nu^*, K_s(x^*)] = 0. \tag{4.10}$$

Although the control minimum principle (4.8) usually holds for almost every $t \in [0, 1]$, we show in Appendix B that by redefining (u^*, λ^*) on a set of measure zero, we get a new optimal control and dual multiplier satisfying relations (4.8)–(4.9) and the constraints $K_c(u^*(t), t) \leqq 0 \leqq \lambda^*(t)$ for all $t \in (0, 1)$.

We first consider the computational question (ii) above; henceforth, the superscript "*" on (x, q, u, λ, ν) is omitted and all variables are assumed optimal. To simplify the notation in the proofs, we assume that the cost functional is quadratic and the constraints on the state and the control are affine inequality constraints. That is, the following control problem is considered:

$$\begin{aligned}&\text{minimize} \quad \frac{1}{2}\int_0^1 [x(t)^T Q x(t) + u(t)^T R u(t)]\, dt\\ &\text{subject to} \quad \dot{x}(t) = Ax(t) + Bu(t) \quad \text{for almost every } t \in [0, 1],\\ &\qquad\qquad \left.\begin{array}{l} K_s x(t) + b_s \leqq 0 \\ K_c u(t) + b_c \leqq 0 \end{array}\right\} \text{for all } t \in [0, 1],\\ &\qquad\qquad x(0) = x_0, \qquad u \in L^2(R^m),\end{aligned} \tag{CP$'$}$$

where R is positive definite and Q is semidefinite. In Appendix C, we consider the more general problem (CP). Also note that in the context of (CP′), relations (4.7)–(4.8) become the following

$$\begin{aligned}&\dot{q}(t) = -A^T(q(t) - K_s^T\nu(t)) - Qx(t) \quad \text{for almost every } t \in [0, 1],\\ &q(1) = 0,\end{aligned} \tag{4.7$'$}$$

and

$$u(t) = R^{-1}[B^T(K_s^T\nu(t) - q(t)) - K_c^T\lambda(t)] \quad \text{for all } t \in [0, 1]. \tag{4.8$'$}$$

LEMMA 4.4. *If* (4.6) *holds,* $(\dot{x}, \dot{q})$ *are continuous and uniformly bounded on* $(0, 1)$ *and* u *is uniformly continuous on* $(0, 1)$, *then there exists a constant* γ *depending on* $\|\dot{x}\|_{L^\infty}$, $\|\dot{q}\|_{L^\infty}$, (R, A, B, K_c, K_s), *and* β *such that*

$$\|z(\sigma) - z(\tau)\| \leqq \gamma|\sigma - \tau| \tag{4.11}$$

for all compatible data $\sigma, \tau \in (0, 1)$.

Proof. Below we use "B" and "N" superscripts to denote binding and nonbinding components respectively. Since $K_s^B(t)x(t) + b_s^B(t) = 0$ and $\dot{x}$ is continuous, it is easy to see that

$$K_s^{\mathrm{B}}(t)\dot{x}(t)=0=K_s^{\mathrm{B}}(t)(Ax(t)+Bu(t)) \tag{4.12}$$

for all $t\in(0,1)$. Substituting in both the relations (4.12) and $K_c^{\mathrm{B}}(t)u(t)+b_c^{\mathrm{B}}(t)=0$ for $u(t)$ given by (4.8′), we obtain the system $N(t)z^{\mathrm{B}}(t)=f(t)$ where $z^T=(\lambda^T,\nu^T)$,

$$N(t)=M(t)R^{-1}M(t)^T, \quad M(t)^T=[K_c^{\mathrm{B}}(t)^T, -B^TK_s^{\mathrm{B}}(t)^T], \tag{4.13}$$

and

$$f(t)=\begin{bmatrix} -K_c^{\mathrm{B}}(t)R^{-1}B^T(q(t)-K_s^{\mathrm{N}}(t)^T\nu^{\mathrm{N}}(t))+b_c^{\mathrm{B}}(t) \\ -K_s^{\mathrm{B}}(t)(Ax(t)+BR^{-1}B^T[q(t)-K_s^{\mathrm{N}}(t)^T\nu^{\mathrm{N}}(t)]) \end{bmatrix}. \tag{4.14}$$

The matrix $N(t)$ is nonsingular by (4.6), and if $\sigma,\tau\in(0,1)$ are compatible, we have $N(\sigma)=N(\tau)$ and

$$\|z^{\mathrm{B}}(\sigma)-z^{\mathrm{B}}(\tau)\|\leqq\|N(\tau)^{-1}\|\,\|f(\sigma)-f(\tau)\|. \tag{4.15}$$

Moreover, by the definition of compatibility, state constraints nonbinding at $t=\sigma,\tau$ remain nonbinding on $[\sigma,\tau]$. Therefore the complementary slackness conditions (4.9)–(4.10) give us

$$\nu^{\mathrm{N}}(\sigma)=\nu^{\mathrm{N}}(\tau) \quad\text{and}\quad \lambda^{\mathrm{N}}(\sigma)=\lambda^{\mathrm{N}}(\tau)=0. \tag{4.16}$$

Since x and q are Lipschitz continuous, (4.14)–(4.16) combine to complete the proof. □

The continuity of z is based on the following lemmas.

LEMMA 4.5. *Let $E\subset(0,1)$ be the set of measure one on which both the system equation* (4.1) *and the adjoint equation* (4.7′) *are satisfied. Then $(\dot{q},\dot{x})$ have bounded variation on E while u has bounded variation on* (0, 1).

Proof. Combining relations (4.8′) and (4.9) and the condition $K_cu(t)+b_c\leqq0\leqq\lambda(t)$, we see that for all $t\in(0,1)$, $v=u(t)$ solves the following program:

$$\begin{aligned} &\text{minimize} \quad \Lambda(v,t)=\tfrac{1}{2}v^TRv+r(t)^Tv \\ &\text{subject to} \quad K_cv+b_c\leqq0 \end{aligned} \tag{4.17}$$

where $r(t)=B^T(q(t)-K_s^T\nu(t))$. Recall that the solution to (4.17) satisfies the following variational inequality:

$$\frac{\partial\Lambda}{\partial v}(v=u(t),t)(v-u(t))=(Ru(t)+r(t))^T(v-u(t))\geqq0 \tag{4.18}$$

for all v such that $K_cv+b_c\leqq0$. Substituting $t=\sigma,\tau$ and $v=u(\tau),u(\sigma)$ into (4.18) and adding the resulting inequalities, we obtain

$$\begin{aligned} \|u(\sigma)-u(\tau)\|&\leqq\|r(\sigma)-r(\tau)\|/\alpha \\ &\leqq c\|q(\sigma)-q(\tau)\|+c\|\nu(\sigma)-\nu(\tau)\| \end{aligned} \tag{4.19}$$

where $\alpha=$ the smallest eigenvalue of R, and c depends on α, K_s, and B. Since q and ν have bounded variation on [0, 1], u also has bounded variation on (0, 1) by (4.19). By the system and adjoint equations, $(\dot{q},\dot{x})$ have bounded variation on E. □

LEMMA 4.6. *Suppose that $(K_sx(\sigma)+b_s)_j=0$ for some j and some $\sigma\in(0,1)$. Then we have*

$$(K_s[Ax(\sigma)+Bu(\sigma)])_j\geqq0. \tag{4.20}$$

Moreover, given $\varepsilon > 0$, there exists $\delta > 0$ such that

$$(4.21)\qquad (K_s[Ax(t)+Bu(t)])_i \leqq \varepsilon \quad \text{for all } t \in I_\delta \equiv E \cap (\sigma, \sigma+\delta)$$

where E was defined in Lemma 4.5.

Proof. By Lemma 4.5, $(K_s\dot{x})_i$ has bounded variation (and hence right limits on E). If the right limit is $\leqq 0$, then (4.21) follows. Suppose that there exists $\Delta > 0$ with $(K_s\dot{x}(t))_i > 0$ for all $t \in I_\Delta$. Hence we have,

$$(4.22)\qquad 0 = (K_s x(\sigma)+b_s)_i = (K_s x(\sigma+\Delta)+b_s)_i - \int_\sigma^{\sigma+\Delta} (K_s\dot{x}(t))_i\,dt < 0$$

since $K_s x(\sigma+\Delta)+b_s)_i \leqq 0$. But (4.22) is impossible, so there exists a sequence $\{t_k\} \subset E$ with both $t_k \to \sigma^+$ as $k \to \infty$ and $(K_s\dot{x}(t_k))_i \leqq 0$ for all k.[1] Thus the right limit of $(K_s\dot{x})_i$ on E is $\leqq 0$ at $t = \sigma$, and (4.21) holds.

Similarly the left limit of $(K_s\dot{x})_i$ is $\geqq 0$ at $t = \sigma$. By Lemma 4.5, u has bounded variation and consequently $\dot{x}(t) = Ax(t)+Bu(t)$ is continuous from the left. Combining these relations, we have (4.20). □

LEMMA 4.7. *The optimal control u is uniformly continuous on* (0, 1).

Proof. By Lemma 4.5, u has bounded variation orr (0, 1); hence the right limit $u(\sigma^+)$ exists. Subtracting (4.20) from (4.21) and letting $\varepsilon \to 0$, we get

$$(4.23)\qquad (K_sB\,\delta u(\sigma))_i \leqq 0 \quad \text{if } (K_s x(\sigma)+b_s)_i = 0$$

where $\delta u(\sigma) = u(\sigma^+) - u(\sigma)$, the discontinuity at $t = \sigma$. Since R is positive definite, the program (4.17) has a unique solution; consequently $\Lambda(u(\sigma^+), \sigma^+) < \Lambda(u(\sigma), \sigma^+)$ if $u(\sigma) \neq u(\sigma^+)$. Below we show that $\Lambda(u(\sigma^+), \sigma^+) \geqq \Lambda(u(\sigma), \sigma^+)$. Therefore $u(\sigma^+) = u(\sigma)$ and u is continuous from the right. Since q and ν are left continuous, (4.19) implies that u is left continuous. Combining these results, u is continuous on (0, 1).

Using (4.18), observe that

$$(4.24)\qquad \begin{aligned} \frac{\partial\Lambda}{\partial v}(u(\sigma), \sigma^+)(\delta u(\sigma)) &= \frac{\partial\Lambda}{\partial v}(u(\sigma), \sigma)(\delta u(\sigma)) + (r(\sigma^+) - r(\sigma))^T\,\delta u(\sigma) \\ &\geqq (r(\sigma^+) - r(\sigma))^T \delta u(\sigma) \end{aligned}$$

where $r(\sigma^+) - r(\sigma) = -B^TK_s^T(\nu(\sigma^+) - \nu(\sigma))$. Combining (4.23) with the conditions $\nu(\sigma^+) \geqq \nu(\sigma)$ and $\nu_i(\sigma^+) > \nu_i(\sigma)$ only if $(K_s x(\sigma)+b_s)_i = 0$, we obtain

$$(4.25)\qquad (r(\sigma^+) - r(\sigma))^T \delta u(\sigma) \geqq 0.$$

Relations (4.24)–(4.25) and the convexity of $\Lambda(\cdot, t)$ imply that $\Lambda(u(\sigma^+), \sigma^+) \geqq \Lambda(u(\sigma), \sigma^+)$; thus $u(\sigma) = u(\sigma^+)$, the desired conclusion. Since u is both continuous and has bounded variation on (0, 1), we see that u is uniformly continuous on (0, 1). □

COROLLARY 4.8. *If* (4.6) *holds, then λ and ν are continuous and uniformly bounded on* (0, 1).

Proof. Combining (4.6) and (4.8)–(4.9), we see that $\lambda \in L^\infty$. Given $\sigma \in (0, 1)$, suppose that $\lambda(\cdot)$ is not continuous from the right at $t = \sigma$. Hence there exists a sequence $\{t_k\}$ and a limit μ such that both $t_k \to \sigma^+$ and $\lambda(t_k) \to \mu$ as $k \to \infty$ where $\mu \neq \lambda(\sigma)$. Since u and q are continuous on (0, 1), (4.8′) implies that

$$(4.26)\qquad B^TK_s^T(\nu(\sigma^+) - \nu(\sigma)) - K_c^T(\mu - \lambda(\sigma)) = 0.$$

By complementary slackness, $\nu_i(\sigma^+) \neq \nu_i(\sigma)$ only if $(K_s x(\sigma)+b_s)_i = 0$. Similarly, by the

[1] The notation $t_k \to \sigma^+$ (or σ^-) as $k \to \infty$ means that the sequence $\{t_k\}$ converges to σ from the right (or left).

continuity of u and complementary slackness, $\mu_j \neq \lambda_j(\sigma)$ only if $(K_c u(\sigma)+b_c)_j = 0$. Combining (4.6) and (4.26), we get $\nu(\sigma^+) = \nu(\sigma)$ and $\mu = \lambda(\sigma)$. Since ν is already left continuous and the above argument also applies to left limits of λ, the proof is complete. □

Proof of Theorem 4.2. By Lemma 4.7 and Corollary 4.8, (u, λ, ν) are continuous and uniformly bounded on $(0, 1)$. Thus $(\dot{x}, \dot{q})$ given by (4.1), (4.7′) are continuous and uniformly bounded on $(0, 1)$. Combining Lemma 4.4 and Theorem 2.1, λ and ν are Lipschitz continuous on $(0, 1)$ with the Lipschitz constant described in Lemma 4.4. From (4.8′), (4.1), and (4.7′), we see that $(u, \dot{x}, \dot{q})$ are Lipschitz continuous on $(0, 1)$.

Now consider the endpoints $t = 0, 1$. By changing the values of u or λ at a point, their optimality is not destroyed. Since u and λ are Lipschitz continuous on $(0, 1)$, their right and left limits at $t = 0$ and $t = 1$ both exist. Setting u and λ at $t = 0, 1$ to be their endpoint limits, we obtain optimal variables that are Lipschitz continuous on $[0, 1]$.

By the form of the dual functional, a jump in ν at $t = 0$ results in the following contribution to the dual cost:

$$(\nu(0^+) - \nu(0))^T (K_s x_0 + b_s).$$

Since $K_s x_0 + b_s < 0$ and $\nu(0^+) \geqq \nu(0)$, this term is maximized only if $\nu(0) = \nu(0^+)$. Consequently $(\dot{x}, \dot{q})$ are also continuous at $t = 0$.

Remark 4.9. The regularity stated in Theorem 4.1 is generally sharp. In [4], examples are given with either control or state constraints such that the derivative of $(\dot{x}, \dot{q}, u, \lambda, \nu)$ is discontinuous whenever a constraint changes from binding to nonbinding.

Although Theorem 4.1 is valid for general convex control problems, the constraints must be sufficiently smooth. For example, the optimal control for problem (4.27) below is Lipschitz continuous as long as $\ddot{\alpha}$ is essentially bounded on $[0, 1]$:

$$\begin{aligned} &\text{minimize} && \int_0^1 u(t)^2\, dt \\ &\text{subject to} && \dot{x}(t) = u(t) \quad \text{for almost every } t \in [0, 1], \\ &&& x(0) = x(1) = 0, \\ &&& x(t) \geqq \alpha(t) \quad \text{for all } t \in [0, 1]. \end{aligned} \tag{4.27}$$

If α is continuous, piecewise linear with $\dot{\alpha}$ $(t = .5)$ discontinuous, the optimal control is generally discontinuous at $t = .5$. The smoothness of α is needed in the development above at (4.12) where the state constraint is differentiated. □

Remark 4.10. In Lemma 4.4, we saw that the Lipschitz constant depended on the following: the parameter β, the matrices $\{R, A, B, K_c, K_s\}$, and the quantities $\|\dot{q}\|_{L^\infty}$ and $\|\dot{x}\|_{L^\infty}$. These last two quantities are now estimated. Let η be the constant given in (4.4), $\bar{c}$ be the cost associated with the control $\bar{u}$, and α be the smallest eigenvalue of R. By the structure of (CP′), we see that $(\alpha/2)\|u\|_{L^2}^2 \leqq \bar{c}$, and by (4.1), $\|x\|_{L^\infty}$ is bounded with an expression involving $\bar{c}$, α, $\|A\|$, and $\|B\|$. Since the optimal cost of (CP′) is nonnegative, we set $x = \bar{x}$ and $u = \bar{u}$ in (4.5) to obtain

$$0 \leqq \mathscr{L}(p, \lambda, \nu) \leqq \bar{c} + \eta[\nu, 1].$$

But $\eta < 0$, $\nu(1) = 0$, and ν is nondecreasing, so we have $\|\nu(t)\| \leqq$ (total variation of ν) $\leqq \bar{c}/|\eta|$. Using (4.7′), we estimate $\|q\|_{L^\infty}$ with an expression involving $\|A\|$, $\|Q\|$, $\|x\|_{L^\infty}$, $\|\nu\|_{L^\infty}$, and $\|K_S^T\|$. Hence (4.7′) also gives us a bound for $\|\dot{q}\|_{L^\infty}$. Inserting $v = \bar{u}(t)$ into

(4.18), we get an estimate for $\|u(t)\|$ in terms $\|\bar{u}(t)\|$, α, and $\|r(t)\|$. Finally we bound $\|\dot{x}\|_{L^\infty}$ using the relation (4.1) □

Appendix A. Essentially bounded controls.

THEOREM A.1. *If* (4.2)–(4.4) *hold, and* $(x^*, u^*) \in A \times L^2$ *are optimal in* (CP), *then* $u^* \in L^\infty$.

Proof. Defining $z^{*T} = (x^{*T}, u^{*T})$ and $\bar{z}^T = (x^{*T}, \bar{u}^T)$ where $(\bar{x}, \bar{u})$ were given in (4.4), (4.3) implies that

$$f(z^*(t), t) \geqq f(\bar{z}(t), t) + \nabla_1 f(\bar{z}(t), t)(z^*(t) - \bar{z}(t)) + \tfrac{1}{2}\alpha\|z^*(t) - \bar{z}(t)\|^2. \tag{A.1}$$

Integrating over $t \in [0, 1]$, we get:

$$C(z^*) \geqq C(\bar{z}) - \|\nabla_1 f(\bar{z}(\cdot), \cdot)\|_{L^2}\|z^* - \bar{z}\|_{L^2} + \tfrac{1}{2}\alpha\|z^* - \bar{z}\|_{L^2}^2. \tag{A.2}$$

Since $f \in C^1$, $\bar{z} \in C^0$, and $z^* \in L^2$, (A.2) shows that $C(z^*) > -\infty$. Therefore by [3, Thm. 2], there exists an optimal solution (p^*, λ^*, ν^*) to (CD), $\mathscr{L}(p^*, \lambda^*, \nu^*) = C(z^*)$, and $\langle \lambda^*, K_c(u^*)\rangle = 0 = [\nu^*, K_s(x^*)]$.

Now define the function $\varphi : R^m \times [0, 1] \to R$ as follows:

$$\begin{aligned} \varphi(u, t) = f(x^*(t), u, t) + p^*(t)^T(\dot{x}^*(t) - A(t)x^*(t) - B(t)u) \\ + \lambda^*(t)^T K_c(u, t). \end{aligned} \tag{A.3}$$

We shall establish that

$$\varphi(u^*(t), t) = \inf\{\varphi(u, t) : u \in R^m\} \tag{A.4}$$

for almost every $t \in [0, 1]$. Let E denote the set of measure 1 consisting of those $s \in [0, 1]$ such that $\lambda^*(s)^T K_c(u^*(s), s) = 0$, $\dot{x}^*(s) - A(s)x^*(s) - B(s)u^*(s) = 0$, and s is a Lebesgue point of the functions $\{p^*(\cdot)^T(\dot{x}^*(\cdot) - A(\cdot)x^*(\cdot)), f(z^*(\cdot), \cdot),\ p^*(\cdot)^T B(\cdot),\ \lambda^*(\cdot)\}$. Suppose that there exists $s \in E$ and $\hat{u} \in R^m$ with $\varphi(\hat{u}, s) < \varphi(u^*(s), s)$—we show that this is impossible, and hence (A.4) holds.

Define the interval

$$I_\delta = \{t \in [0, 1] : |t - s| \leqq \delta\}.$$

Since $s \in E$, we see that

$$\begin{aligned} \int_{I_\delta} \varphi(u^*(t), t)dt &= \int_{I_\delta} f(z^*(t), t)dt \\ &= f(z^*(s), s) \text{ meas } (I_\delta) + o(\text{meas } (I_\delta)) \\ &= \varphi(u^*(s), s) \text{ meas } (I_\delta) + o(\text{meas } (I_\delta)). \end{aligned} \tag{A.5}$$

Similarly, we have

$$\int_{I_\delta} \varphi(\hat{u}, t)dt = \varphi(\hat{u}, s) \text{ meas } (I_\delta) + o(\text{meas } (I_\delta)). \tag{A.6}$$

Since $\varphi(\hat{u}, s) < \varphi(u^*(s), s)$, (A.5) and (A.6) imply that

$$\int_{I_\delta} \varphi(\hat{u}, t)\, dt < \int_{I_\delta} \varphi(u^*(t), t)\, dt \tag{A.7}$$

for δ sufficiently small.

Define the sets

$$\begin{aligned} I_0^k &= \{t \in [0,1]: t \notin I_\delta, |u^*(t)| \leqq k\}, \\ I_\infty^k &= [0,1] - (I_\delta \cup I_0^k) \\ &= \{t \in [0,1]: t \notin I_\delta, |u^*(t)| > k\}, \end{aligned}$$

and the control

$$u_\delta^k = \begin{cases} \hat{u} & \text{for } t \in I_\delta, \\ u^*(t) & \text{for } t \in I_0^k, \\ 0 & \text{for } t \in I_\infty^k. \end{cases}$$

By the continuity properties of Lebesgue integrals, we know that

$$\begin{aligned} &\lim_{k\to\infty} \int_{I_0^k \cup I_\infty^k} [\varphi(u_\delta^k(t), t) - \varphi(u^*(t), t)]\, dt \\ &\quad = \lim_{k\to\infty} \int_{I_\infty^k} [\varphi(0, t) - \varphi(u^*(t), t)]\, dt = 0 \end{aligned} \tag{A.8}$$

since meas $(I_\infty^k) \to 0$ as $k \to \infty$ and $(\varphi(0, \cdot), \varphi(u^*(\cdot), \cdot)) \in L^1$.Combining (A.7) and (A.8), we see that for δ sufficiently small and k sufficiently large

$$\int_0^1 \varphi(u_\delta^k(t), t)\, dt < \int_0^1 \varphi(u^*(t), t)\, dt = C(z^*) \tag{A.9}$$

by complementary slackness. Hence (A.9) along with the definition of $\mathscr{L}$ implies that for k sufficiently large

$$\mathscr{L}(p^*, \lambda^*, \nu^*) \leqq \int_0^1 \varphi(u_\delta^k(t), t) dt + [\nu^*, K_s(x^*)] < C(z^*). \tag{A.10}$$

This contradicts the strong duality result that $\mathscr{L}(p^*, \lambda^*, \nu^*) = C(z^*)$, and (A.4) has been established.

Define the function

$$\Lambda(u, t) = f(x^*(t), u, t) - p^*(t)^T B(t) u. \tag{A.11}$$

Since $\lambda^*(t) \geqq 0 \geqq K_c(u^*(t), t)$ and $\lambda^*(t)^T K_c(u^*(t), t) = 0$ for almost every $t \in [0, 1]$, (A.4) and the convexity of $\varphi(\cdot, t)$ give us

$$\Lambda(u^*(t), t) = \inf \{\Lambda(u, t)\colon u \in R^m, K_c(u, t) \leqq 0\} \tag{A.12}$$

for almost every $t \in [0, 1]$. By the same reasoning used to derive (B.6) in Appendix B, we have

$$\tfrac{1}{2}\alpha |u^*(t) - \bar{u}(t)| \leqq |\nabla_1 \Lambda(\bar{u}(t), t)| \tag{A.13}$$

for almost every $t \in [0, 1]$. Therefore, $u^* \in L^\infty$ since $f \in C^1$, $(x^*, \bar{u}) \in C^0$, and $p^* \in BV$. □

Appendix B. Smoothing an optimal control. Suppose that both $f: R^m \times [0, 1] \to R$ and $K: R^m \times [0, 1] \to R^l$ are C^1, and the components of $K(\cdot, t)$ are convex for all $t \in [0, 1]$. Moreover, assume that

(i) there exists $\alpha > 0$ such that

$$f(y, t) \geqq f(z, t) + \nabla_1 f(z, t)(y - z) + \alpha \|y - z\|^2$$

for all $y, z \in R^m$ and $t \in [0, 1]$ satisfying $K(y, t) \leqq 0$ and $K(z, t) \leqq 0$,

(ii) there exists $\bar{u} \in C^0(R^m)$ such that

$$K(\bar{u}(t), t) \leqq 0 \quad \text{for all } t \in [0, 1].$$

LEMMA B.1. *Let $E \subset [0, 1]$ with measure $(E) = 1$, $r \in BV(R^m)$, u: $E \to R^m$, and λ: $E \to R^l$. Furthermore suppose that for all $t \in E$, we have*

$$\nabla_1 f(u(t), t) + r(t) + \nabla_1 K(u(t), t)^T \lambda(t) = 0, \tag{B.1}$$

$$K(u(t), t) \leqq 0 \leqq \lambda(t), \qquad \lambda(t)^T K(u(t), t) = 0, \tag{B.2}$$

and there exists a constant $\beta > 0$ such that for all $t \in E$ and all z

$$\|\nabla_1 K^B(u(t), t)^T z\| \geqq \beta \|z\| \tag{B.3}$$

where $K^B(u(t), t)$ denotes the components of $K(u(t), t)$ corresponding to binding constraints in (B.2). *Then there exists an extension of u and λ to* (0, 1) *such that* (B.1)–(B.2) *hold on* (0, 1).

Proof. Relations (B.1)–(B.2) imply that for all $t \in E$, $v = u(t)$ solves the following program:

$$\begin{aligned} &\text{minimize} && f(v, t) + r(t)^T v \\ &\text{subject to} && K(v, t) \leqq 0. \end{aligned} \tag{B.4}$$

By (i) and (ii) above, we have:

$$f(u(t), t) \geqq f(\bar{u}(t), t) + \nabla_1 f(\bar{u}(t), t)(u(t) - \bar{u}(t)) + \alpha \|u(t) - \bar{u}(t)\|^2, \tag{B.5}$$

for all $t \in E$. On the other hand, since $f(u(t), t) \leqq f(\bar{u}(t), t)$, (B.5) along with the Schwarz inequality give us

$$\|u(t) - \bar{u}(t)\| \leqq \|\nabla_1 f(\bar{u}(t), t)\|/\alpha \quad \text{for all } t \in E. \tag{B.6}$$

Therefore, $\|u(\cdot)\|$ is uniformly bounded on E.

Similarly, by (B.1)–(B.3) and the fact that both $\|\nabla_1 f(u(\cdot), \cdot)\|$ and $\|r(\cdot)\|$ are uniformly bounded on E, we conclude that $\|\lambda(\cdot)\|$ is uniformly bounded on E. Now given $\sigma \in (0, 1)/E$, let $\{\sigma_k\} \subset E$ be any sequence such that $\sigma_k \to \sigma^-$ as $k \to \infty$. Consequently, there exists a subsequence $\{\sigma'_k\}$ and limits w, μ such that $u(\sigma'_k) \to w$ and $\lambda(\sigma'_k) \to \mu$ as $k \to \infty$. Defining $u(\sigma) = w$ and $\lambda(\sigma) = \mu$, (B.2) holds trivially at $t = \sigma$ while (B.1) holds since r is left continuous. □

Appendix C. Control regularity for (CP). Under assumptions (4.2)–(4.4) and (4.6), we now consider the proof of Theorem 4.2 for (CP).

I. *Extension of Lemma* 4.4. We establish the existence of constants $\varepsilon > 0$ and $\gamma < \infty$ such that (4.11) holds for all compatible data σ, $\tau \in (0, 1)$ satisfying $|\sigma - \tau| \leqq \varepsilon$. Hence the proof of Theorem 4.2 can be completed using Theorem 2.3 instead of Theorem 2.1. Suppose that σ, $\tau \in [0, 1]$ are compatible. Using the notation of Lemma 4.4 and assumption (4.6), we differentiate the state constraint $K_s^{\mathrm{B}}(x(t), t) = 0$ to get the following analogue of (4.12):

$$G_s(t)(A(t)x(t) + B(t)u(t)) + \frac{\partial}{\partial t} K_s^{\mathrm{B}}(x(t), t) = 0. \tag{C.1}$$

Substituting $t=\sigma, \tau$ and subtracting the resulting relations, we get

$$G_s(\sigma)B(\sigma)(u(\tau)-u(\sigma))=O(|\tau-\sigma|) \tag{C.2}$$

since x is Lipschitz continuous and (4.2) holds.

Similarly substituting $t=\sigma, \tau$ in the relation $K_c^{\mathrm{B}}(u(t), t)=0$ and subtracting the resulting relations, we obtain

$$S(u(\tau)-u(\sigma))=O(|\tau-\sigma|) \tag{C.3}$$

where

$$S=\int_0^1 \nabla_1 K_c^{\mathrm{B}}((1-s)u(\sigma)+su(\tau), \sigma)\,ds. \tag{C.4}$$

Combining (4.6) and (4.8)–(4.9), we see that $\lambda \in L^\infty$. Hence substituting $t=\sigma, \tau$ in (4.8) and subtracting the resulting equalities, we find that

$$T(u(\tau)-u(\sigma))+M(\sigma)^T(z^{\mathrm{B}}(\tau)-z^{\mathrm{B}}(\sigma))=O(|\tau-\sigma|) \tag{C.5}$$

where

$$\begin{aligned} T=\int_0^1 &[\nabla_2^2 f(x(\sigma), (1-s)u(\sigma)+su(\tau), \sigma)\\ &+\nabla_1^2 K_c((1-s)u(\sigma)+su(\tau), \sigma)^T\lambda(\tau)]\,ds \quad \text{and}\\ M(\sigma)^T&=[G_c(\sigma)^T, -B(\sigma)^T G_s(\sigma)^T]. \end{aligned} \tag{C.6}$$

From (4.3), the convexity of K_c, and the fact that $\lambda(\tau) \geqq 0$, we conclude that the smallest eigenvalue of T is bounded from below by α. Substituting into (C.2) and (C.3) for $(u(\tau)-u(\sigma))$ given by (C.5), we obtain a relation of the form

$$N(z^{\mathrm{B}}(\tau)-z^{\mathrm{B}}(\sigma))=O(|\tau-\sigma|) \tag{C.7}$$

where N is nonsingular if $|\tau-\sigma|$ is sufficiently small. Combining (4.6), the observation that $T>\alpha I$, and the assumption that u is uniformly continuous, we conclude that there exists $\varepsilon>0$ such that $\|N^{-1}\|$ is bounded by a constant that is independent of σ, τ if $|\sigma-\tau| \leqq \varepsilon$. The proof is completed as in Lemma 4.4.

II. *Extension of Lemma* 4.5. Program (4.17) is replaced by (A.12) and the bound (4.19) for the change in solution in terms of the data is replaced by Theorem D.1.

III. *Extension of Lemma* 4.6. Throughout the proof, $K_s\dot{x}(t)$ should be replaced by

$$\begin{aligned} \frac{d}{dt}K_s(x(t), t)&=G(t)\dot{x}(t)+\frac{\partial}{\partial t}K_s(x(t),t)\\ &=G(t)(A(t)x(t)+B(t)u(t))+\frac{\partial}{\partial t}K_s(x(t), t) \end{aligned} \tag{C.8}$$

Hence (4.20) and (4.21) become respectively:

$$\left[G(\sigma)(A(\sigma)x(\sigma)+B(\sigma)u(\sigma))+\frac{\partial}{\partial t}K_s(x(\sigma), \sigma)\right]_j \geqq 0, \tag{C.9}$$

$$\left[G(t)(A(t)x(t)+B(t)u(t))+\frac{\partial}{\partial t}K_s(x(t), t)\right]_j \leqq \varepsilon \tag{C.10}$$

for all $t \in I_\delta = E \cap (\sigma, \sigma+\delta)$.

IV. *Extension of Lemma* 4.7. Relation (4.23) is replaced by

$$[G(\sigma)B(\sigma)\,\delta u(\sigma)]_j \leqq 0 \quad \text{if } K_s(x(\sigma),\sigma)_j = 0. \tag{C.11}$$

Relation (4.24) is unchanged; however, we now have

$$r(\sigma^+)-r(\sigma) = -B(\sigma)^T G_s(\sigma)^T(\nu(\sigma^+)-\nu(\sigma)). \tag{C.12}$$

V. *Extension of Corollary* 4.8. Relation (4.26) is replaced by

$$B(\sigma)^T G(\sigma)^T(\nu(\sigma^+)-\nu(\sigma)) - \nabla_1 K_c(u(\sigma),\sigma)^T(\mu-\lambda(\sigma)) = 0.$$

VI. *Proof of Theorem* 4.2. This proof is completely unchanged except that the reference to Theorem 2.1 is replaced by a reference to Theorem 2.3.

Appendix D. Convex programs. Let $D \subset R^m$ be a convex set of data and for given $d \in D$, consider the program:

$$\text{(P)}\qquad \begin{array}{ll} \text{minimize} & f(v,d) \\ \text{subject to} & A(v,d) \leqq 0, \quad B(v,d)=0, \quad v \in R^n, \end{array}$$

where $f: R^n \times R^m \to R$, $A: R^n \times R^m \to R^{l_i}$, and $B: R^n \times R^m \to R^{l_e}$. Our analysis is restricted to programs for which there is a unique solution $u(d)$ and a unique dual multiplier $\lambda(d)$. Defining

$$M(d) = \begin{bmatrix} \nabla_1 A(u(d),d) \\ \nabla_1 B(u(d),d) \end{bmatrix}_{J(d)} \tag{D.1}$$

where $J(d)$ denotes the indices corresponding to binding constraints for $u(d)$, we make the following assumptions:

(D.2) f, A, and B are C^2, $f(\cdot, d)$ and the components of $A(\cdot, d)$ are convex for all $d \in D$, and $B(\cdot, d)$ is affine for all $d \in D$,

(D.3) for all $d \in D$, there exists a unique solution $u(d)$ to (P),

(D.4) there exists a constant Γ such that $\|d\|$ and $\|u(d)\| \leqq \Gamma$ for all $d \in D$,

(D.5) there exists $\alpha, \beta > 0$ such that for all $d \in D$,

$$v^T \nabla_1^2 f(u(d),d) v \geqq \alpha \|v\|^2 \quad \text{for all } v \text{ satisfying } M(d)v = 0,$$

and

$$\|M(d)^T \lambda\| \geqq \beta \|\lambda\| \quad \text{for all } \lambda.$$

Using techniques similar to those of § 3, the following result can be established:

THEOREM D.1. *If* (D.2)–(D.5) *hold, then there exists a constant* $\rho < \infty$ *such that for all* $d_1, d_2 \in D$, *we have*:

$$\|u(d_1)-u(d_2)\| \quad \text{and} \quad \|\lambda(d_1)-\lambda(d_2)\| \leqq \rho \|d_1-d_2\|.$$

Acknowledgments. We wish to thank Professor Juan Schäffer and Greg Knowles for carefully reading and critiquing Theorem 2.1; also the many suggestions in writing style given by Professor Gilbert Strang were appreciated. Finally much credit should be attributed to Professor R. T. Rockafellar whose letter of encouragement was a primary motivation for completing this work.

REFERENCES

[1] J. W. DANIEL, *Stability of the solution of definite quadratic programs*, Math. Programming, 5 (1973), pp. 41–53.

[2] A. V. FIACCO, *Sensitivity analysis for nonlinear programming using penalty methods*, Tech. Rep. T 275, Institute for Management Science and Engineering, George Washington Univ., Washington, D.C., March, 1973.

[3] W. W. HAGER AND S. K. MITTER, *Lagrange duality theory for convex control problems*, this Journal, 14 (1976), pp. 843–856.

[4] W. W. HAGER, *The Ritz–Trefftz method for state and control constrained optimal control problems*, SIAM J. Numer. Anal., 12 (1975), pp. 854–867.

[5] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, translated by S. K. Mitter, Springer-Verlag, New York, 1971.

[6] K. MALANOWSKI, *On the regularity of solutions to optimal control problems for systems linear with respect to control variable*, to appear.

[7] S. M. ROBINSON, *Perturbation in finite dimensional systems of linear inequalities and equations*, Rep. 1357, Mathematics Research Center, Univ. of Wisconsin, Madison, June, 1973.

[8] ———, *Perturbed Kuhn–Tucker points and rates of convergence for a class of nonlinear programming algorithms*, Math. Programming, 7 (1974), pp. 1–16.

[9] ———, *Generalized equations and their solutions, part I: basic theory*, Mathematics Research Center, Rep. 1812, Univ. of Wisconsin, Madison, November, 1977.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 3, May 1979

FEEDBACK SYSTEMS DESCRIBED BY MONOTONE OPERATORS*

VACLAV DOLEZAL†

Abstract. In this paper feedback systems described by nonlinear (possibly multivalued) monotone operators are considered. We establish conditions for the existence and uniqueness of outputs corresponding to a given pair of inputs, conditions for causality, Lipschitz continuity and stability. Also, feedback systems over an extended Hilbert space are discussed. Finally, linear approximations to a nonlinear feedback system are studied.

Introduction. In the last two decades a considerable amount of research has been devoted to nonlinear feedback systems. Starting with the pioneering paper [19] by G. Zames, the functional analysis proved to be a useful tool in the analysis and synthesis of nonlinear systems. The early results concerning stability, Lipschitz continuity and possible solvability ([19] through [23]) were mostly based on the assumption that a certain quantity is less than one. Later, refinements of these results led to various circle criteria [16], [17] for systems, whose underlying space is the space $L_2[0, \infty)$.

On the other hand, it was soon recognized that the concept of causality is quite natural and of extreme importance in the feedback system theory, [4], [5], [9], [11], [13], [18]. In particular, a theorem of J. C. Willems [18] permits us to avoid a natural setting of extended spaces for stability considerations and thus simplify the mathematical framework.

More recently, new results were obtained for feedback systems described by causal C_0-contractions or operators from a special class [10], [6], [7], and for systems over particular spaces, [1] through [3]. Also, the circle criteria have been generalized, [14], [15].

In this paper we are concerned primarily with a feedback system $[A_1, A_2]$ over a Hilbert space H (see Fig. 1) such that the operators $A_1+\alpha_1 I$ and $A_2+\alpha_2 I$ are monotone for some $\alpha_1 \geqq 0$ and $\alpha_2 < 0$, or $\alpha_1 < 0$ and $\alpha_2 \geqq 0$. Most of the results given below are based on Rockafellar's surjectivity theorem for maximal monotone, coercive operators [12].

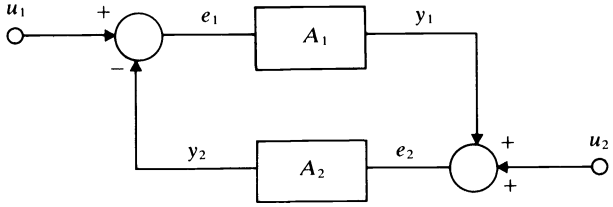

FIG. 1

In the first part of the paper we give fairly general theorems on the existence and uniqueness of the errors e_1, e_2 (or the outputs y_1, y_2) corresponding to a given pair of inputs (u_1, u_2). We assume that A_1 and A_2 are nonlinear, possibly multivalued operators on a linear space H, which actually arise in practical engineering problems. It turns out that the behavior of a feedback system is completely determined by a certain mapping M_a (defined by equation (2)).

* Received by the editors November 28, 1977, and in revised form August 4, 1978.

† Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, New York 11794. This research was supported by The National Science Foundation under Grant MPS7505268.

Based on this, the second part deals with systems, whose underlying space is a Hilbert space, and whose operators satisfy the requirements mentioned above. We give simple yet effective conditions for normality (i.e., existence and uniqueness of errors e_1, e_2), causality, Lipschitz continuity and stability. Crudely speaking, these conditions are given in terms of the "gain" and "minimal slope" of operators A_1, A_2.

In the third part it is indicated, how the results obtained can be generalized for systems over extended Hilbert spaces.

If the operators A_1, A_2 are not linear, then expressing explicitly the errors e_1, e_2 via inputs u_1, u_2 amounts to inverting a nonlinear operator M_{u_2}, (see formula (13)). However, if an approximate solution suffices, inverting M_{u_2} can be avoided by linearizing the operators A_1 and A_2 in a vicinity of an operating point (assumed to be zero). This problem is briefly discussed in the fourth part, where estimates for the quality of an approximation are given.

1. General theorems. Let us begin with several definitions.

Let H be a linear (not necessarily normed) space, and let 2^H be the collection of all subsets of H.

Given a mapping $A: H \to 2^H$, we let

$$D(A) = \{x : x \in H, Ax \neq \phi\},$$

and

$$AK = \bigcup_{x \in K} Ax$$

for every $K \subset H$.

If, in particular, $D(A) = H$ and Ax is a singleton for each $x \in H$, A will be called an operator.

Moreover, a mapping $A: H \to 2^H$ will be called simple, if

$$x_1, x_2 \in H, x_1 \neq x_2 \Rightarrow (Ax_1) \cap (Ax_2) = \phi.$$

(Clearly, if A is an operator and A is simple, then A is 1-to-1).

Now, if $A_1, A_2: H \to 2^H$, then the ordered pair $[A_1, A_2]$ will be called a feedback system (further F.S.) over H.

Having the physical interpretation in mind (see Fig. 1), we introduce the following definition:

DEFINITION 1. Let $[A_1, A_2]$ be a F.S. over H. If $(u_1, u_2) \in H^2 = H \times H$, then a pair $(e_1, e_2) \in H^2$ will be called *a solution of* $[A_1, A_2]$ *corresponding to* (u_1, u_2), if there exist $y_1 \in A_1 e_1$ and $y_2 \in A_2 e_2$ such that

$$e_1 = u_1 - y_2, \qquad e_2 = u_2 + y_1. \tag{1}$$

This fact will be symbolized by writing $(u_1, u_2) \mapsto (e_1, e_2)$. Furthermore, the F.S. $[A_1, A_2]$ will be called

(a) solvable, if for any $(u_1, u_2) \in H^2$ there exists at least one solution $(e_1, e_2) \in H^2$ corresponding to (u_1, u_2),

(b) unambiguous, if each solution is determined uniquely,

(c) normal, if $[A_1, A_2]$ is solvable and unambiguous.

To state an assertion on solvability and unambiguousness, we introduce the following mapping: For each $a \in H$, let $M_a: H \to 2^H$ be defined by

$$M_a x = x + A_2(a + A_1 x). \tag{2}$$

THEOREM 1. *Let* $[A_1, A_2]$ *be a F.S. over* H; *then* $[A_1, A_2]$ *is solvable* $\Leftrightarrow M_a H = H$ *for each* $a \in H$.

Proof. Let $M_aH = H$ for every $a \in H$. Choose $(u_1, u_2) \in H^2$. Since $M_{u_2}H = H$, there exists at least one $e_1 \in H$ such that $u_1 \in M_{u_2}e_1$, i.e., $u_1 \in e_1 + A_2(u_2 + A_1e_1)$. Thus, there exists $y_2 \in A_2(u_2 + A_1e_1)$ such that

$$u_1 = e_1 + y_2. \tag{3}$$

Moreover, there exists $e_2 \in u_2 + A_1e_1$ such that $y_2 \in A_2e_2$; also, there exists $y_1 \in A_1e_1$ such that

$$e_2 = u_2 + y_1. \tag{4}$$

Thus, there exists $(e_1, e_2) \in H^2$ and $y_1 \in A_1e_1$, $y_2 \in A_2e_2$ such that (3), (4) hold, i.e., (e_1, e_2) is a solution of $[A_1, A_2]$ corresponding to (u_1, u_2). Hence, $[A_1, A_2]$ is solvable.

(ii) Let $[A_1, A_2]$ be solvable. Choose $a \in H$ and show that $M_aH = H$. By solvability, for any $u_1 \in H$ there exists $(e_1, e_2) \in H^2$ and $y_1 \in A_1e_1$, $y_2 \in A_2e_2$ such that

$$e_1 + y_2 = u_1, \qquad e_2 - y_1 = a. \tag{5}$$

Thus, $u_1 \in e_1 + A_2e_2$, and $e_2 = a + y_1 \in a + A_1e_1$; hence, $u_1 \in e_1 + A_2(a + A_1e_1) = M_ae_1$. Consequently, $H \subset \bigcup_{z \in H} M_az = M_aH$, i.e., $H = M_aH$.

THEOREM 2. *Let* $[A_1, A_2]$ *be a F.S. over* H.

(i) *If* $[A_1, A_2]$ *is unambiguous, then* M_a *is simple for each* $a \in H$.

(ii) *If* A_1 *is an operator and* M_a *is simple for each* $a \in H$, *then* $[A_1, A_2]$ *is unambiguous.*

Proof. (i) Choose $a \in H$ and assume that, for some $e_1, e_1' \in H$, $(M_ae_1) \cap (M_ae_1') \neq \phi$, i.e., there exists $u_1 \in H$ such that

$$u_1 \in M_ae_1 = e_1 + A_2(a + A_1e_1) \tag{6}$$

and

$$u_1 \in M_ae_1' = e_1' + A_2(a + A_1e_1'). \tag{7}$$

By (6), there exists $y_2 \in A_2(a + A_1e_1)$ such that

$$u_1 = e_1 + y_2. \tag{8}$$

Also, there exists $e_2 \in a + A_1e_1$ such that $y_2 \in A_2e_2$, and there exists $y_1 \in A_1e_1$ such that

$$e_2 = a + y_1. \tag{9}$$

Hence, (8), (9) show that $(e_1, e_2) \in H^2$ is a solution corresponding to (u_1, a). Similarly, (7) implies that $(e_1', e_2') \in H^2$ is a solution corresponding to (u_1, a). Consequently, $e_1 = e_1'$, i.e., M_a is simple.

(ii) Suppose that, for some $(u_1, u_2) \in H^2$, there exist solutions $(e_1, e_2) \in H^2$ and $(e_1', e_2') \in H^2$ corresponding to (u_1, u_2). Thus, there exists $y_2 \in A_2e_2$ and $y_2' \in A_2e_2'$ such that

$$\begin{aligned} e_1 + y_2 &= u_1, & e_1' + y_2' &= u_1, \\ e_2 - A_1e_1 &= u_2, & e_2' - A_1e_1' &= u_2. \end{aligned} \tag{10}$$

From the first pair of (10) it follows that $u_1 \in e_1 + A_2e_2$ and $e_2 = u_2 + A_1e_1$, i.e.,

$$u_1 \in e_1 + A_2(u_2 + A_1e_1) = M_{u_2}e_1. \tag{11}$$

Similarly, the second pair implies that

$$u_1 \in e_1' + A_2(u_2 + A_1e_1') = M_{u_2}e_1'. \tag{12}$$

Since M_{u_2} is simple, we have by (11), (12), $e_1=e_1'$. Then (10) yields $e_2=e_2'$, i.e., $(e_1,e_2)=(e_1',e_2')$.

Hence, $[A_1, A_2]$ is unambiguous.

COROLLARY 1. *Let $[A_1, A_2]$ be a F.S. over H and let A_1, A_2 be operators; then $[A_1, A_2]$ is*

(i) *solvable $\Leftrightarrow$ the operator M_a is onto H for each $a\in H$,*

(ii) *unambiguous $\Leftrightarrow$ the operator M_a is 1-to-1 for each $a\in H$,*

(iii) *normal $\Leftrightarrow M_a$ is invertible for each $a\in H$.*

In this case, for any $(u_1,u_2)\in H^2$, the solution $(e_1,e_2)\in H^2$ is given by

$$(e_1,e_2)=(M_{u_2}^{-1}u_1,\ u_2+A_1M_{u_2}^{-1}u_1). \tag{13}$$

Proof. Obvious.

COROLLARY 2. *Let $[A_1, A_2]$ be a F.S. over H, let A_1, A_2 be operators, and let A_2 be linear; then $[A_1, A_2]$ is*

(i) *solvable $\Leftrightarrow$ the operator $I+A_2A_1$ is onto H,*

(ii) *unambiguous $\Leftrightarrow$ the operator $I+A_2A_1$ is 1-to-1,*

(iii) *normal $\Leftrightarrow I+A_2A_1$ is invertible.*

In this case, for any $(u_1,u_2)\in H^2$, the solution $(e_1,e_2)\in H^2$ is given by

$$(e_1,e_2)=((I+A_2A_1)^{-1}(u_1-A_2u_2),\ u_2+A_1(I+A_2A_1)^{-1}(u_1-A_2u_2)). \tag{14}$$

Proof. It suffices to realize that M_a is onto H (1-to-1) for each $a\in H\Leftrightarrow I+A_2A_1$ is onto H (1-to-1). As far as (14) is concerned, for the solution (e_1,e_2) we have by the definition,

$$e_1+A_2e_2=u_1, \qquad e_2-A_1e_1=u_2. \tag{15}$$

Thus, $e_1+A_2(u_2+A_1e_1)=u_1$, i.e., $(I+A_2A_1)e_1=u_1-A_2u_2$. Hence, $e_1=(I+A_2A_1)^{-1}(u_1-A_2u_2)$. The rest follows from (15).

2. Feedback systems over a Hilbert space. In this part of the paper we will assume that H is a real Hilbert space with inner product $\langle\cdot,\cdot\rangle$.

As known [12], a mapping $A\colon H\to 2^H$ is called monotone [strictly monotone], if

$$\langle y_1-y_2, x_1-x_2\rangle\geqq 0$$

whenever $x_1,x_2\in H$, $y_1\in Ax_1$, $y_2\in Ax_2$, [>0 whenever $x_1,x_2\in H$, $x_1\neq x_2$, $y_1\in Ax_1$, $y_2\in Ax_2$].

LEMMA 1. *Let A, $B\colon H\to 2^H$ be monotone; if either*

(i) *B is strictly monotone, or*

(ii) *A is a strictly monotone operator, then the mapping $N=I+AB$ is simple.*

Proof. Assume that for some x_1, $x_2\in H$ we have $(Nx_1)\cap(Nx_2)\neq\phi$, i.e., there exists $y\in H$ such that $y\in x_1+ABx_1$ and $y\in x_2+ABx_2$. Then $y=x_1+z_1$ for some $z_1\in ABx_1$, and $y=x_2+z_2$ for some $z_2\in ABx_2$. Also, there exists $y_1\in Bx_1$ such that $z_1\in Ay_1$, and there exists $y_2\in Bx_2$ such that $z_2\in Ay_2$. From this we have

$$x_1-x_2+z_1-z_2=0, \tag{16}$$

and consequently,

$$\langle y_1-y_2, x_1-x_2\rangle+\langle y_1-y_2, z_1-z_2\rangle=0. \tag{17}$$

Since both inner products in (17) are nonnegative by monotonicity of A and B, it follows that

$$(18)\qquad \langle y_1-y_2, x_1-x_2\rangle = 0,$$

$$(19)\qquad \langle z_1-z_2, y_1-y_2\rangle = 0.$$

Now, if (i) holds, we have $x_1=x_2$ by (18). Otherwise, if (ii) holds, (19) yields $y_1=y_2$. Thus, $Ay_1=Ay_2$, and since both Ay_1 and Ay_2 are singletons, we have $z_1=z_2$. Hence, by (16), $x_1=x_2$.

THEOREM 3. *Let A_1, A_2: $H\to 2^H$ be monotone, and let A_1 be an operator. If either*

(i) *A_1 is strictly monotone, or*

(ii) *A_2 is a strictly monotone operator,*

then the F.S. $[A_1, A_2]$ over H is unambiguous.

COROLLARY 3. *If A_1, A_2 are monotone operators and one of them is strictly monotone, then $[A_1, A_2]$ over H is unambiguous.*

Proof. Referring to Theorem 2, (ii), it suffices to show that M_a is simple on H for each $a\in H$. To do this, note that the following is true:

If K: $H\to 2^H$, $a\in H$ and K_a: $H\to 2^H$ is defined by $K_ax = a+Kx$, then K is [strictly] monotone $\Leftrightarrow K_a$ is [strictly] monotone. Recalling (2) and Lemma 1, the proof follows. The corollary is obvious.

At this point, let us introduce concepts that are needed for discussing causality.

For every $T\in R^1$, let S_T be an orthogonal projection from H into itself, and let $\{S_T: T\in R^1\}$ be a resolution of identity in H.

Let A: $H\to 2^H$; then A will be called causal [13], if

$$(20)\qquad x_1, x_2\in D(A),\ S_Tx_1=S_Tx_2 \Rightarrow S_TAx_1=S_TAx_2.$$

It is easy to see that the following assertion is true:

Let A: $H\to 2^H$ and assume that $S_TD(A)\subset D(A)$ for all $T\in R^1$. Then A is causal $\Leftrightarrow S_TAx = S_TAS_Tx$ for each $T\in R^1$ and $x\in D(A)$.

Also, we have:

Let A, B: $H\to 2^H$ be causal; then

(i) $A+B$ is causal,

(ii) BA is causal provided $AH\subset D(B)$.

DEFINITION 2. Let $[A_1, A_2]$ be a F.S. over H.

(a) If $(u_1, u^*)\mapsto(e_1, e_2)$, $(u_1', u^*)\mapsto(e_1', e_2')$ and $S_Tu_1=S_Tu_1'$ implies that

$$(21)\qquad S_Te_1=S_Te_1',\qquad S_Te_2=S_Te_2',$$

then $[A_1, A_2]$ will be called *causal by the first input.*

(b) If $(u_1, u_2)\mapsto(e_1, e_2)$, $(u_1', u_2')\mapsto(e_1', e_2')$ and $S_Tu_1=S_Tu_1'$, $S_Tu_2=S_Tu_2'$ implies (21), then $[A_1, A_2]$ will be called causal.

We will need the following proposition:

LEMMA 2. *Let A: $H\to 2^H$ be causal and assume that $S_TD(A)\subset D(A)$ for each $T\in R^1$; if*

(a) *A is strictly monotone, then*

$$\langle S_Ty_1-S_Ty_2, x_1-x_2\rangle > 0$$

whenever $x_1, x_2\in D(A)$, $S_Tx_1\neq S_Tx_2$, $y_1\in Ax_1$, $y_2\in Ax_2$;

(b) *A is monotone, then*

$$\langle S_Ty_1-S_Ty_2, x_1-x_2\rangle \geqq 0$$

whenever $x_1, x_2\in D(A)$, $y_1\in Ax_1$, $y_2\in Ax_2$.

Proof. (a) If $u_1, u_2 \in D(A)$, $u_1 \neq u_2$, and $v_1 \in Au_1$, $v_2 \in Au_2$, then

$$\langle v_1 - v_2, u_1 - u_2\rangle > 0. \tag{22}$$

Let $x_1, x_2 \in D(A)$ be such that, for some $T \in R^1$, $S_T x_1 \neq S_T x_2$; then (22) yields

$$\langle z_1 - z_2, S_T x_1 - S_T x_2\rangle > 0$$

for any $z_1 \in AS_T x_1$, $z_2 \in AS_T x_2$, i.e.,

$$\langle S_T z_1 - S_T z_2, x_1 - x_2\rangle > 0.$$

Thus,

$$\langle w_1 - w_2, x_1 - x_2\rangle > 0 \tag{23}$$

for $S_T x_1 \neq S_T x_2$ and $w_1 \in S_T AS_T x_1$, $w_2 \in S_T AS_T x_2$. (Indeed, if $w_i \in S_T AS_T x_i$, then $w_i = S_T g_i$ for some $g_i \in AS_T x_i$.) However, by the above assertion, $S_T AS_T u = S_T Au$ for each $u \in D(A)$. Hence,

$$\langle S_T y_1 - S_T y_2, x_1 - x_2\rangle > 0 \tag{24}$$

whenever $S_T x_1 \neq S_T x_2$, $y_i \in Ax_i$, because $S_T y_i \in S_T Ax_i = S_T AS_T x_i$.

(b) For any $x_1, x_2 \in D(A)$ and $T \in R^1$ we have

$$\langle z_1 - z_2, S_T x_1 - S_T x_2\rangle \geqq 0 \tag{25}$$

whenever $z_i \in AS_T x_i$. Thus,

$$\langle S_T z_1 - S_T z_2, x_1 - x_2\rangle \geqq 0, \tag{26}$$

so that

$$\langle w_1 - w_2, x_1 - x_2\rangle \geqq 0 \tag{27}$$

whenever $w_i \in S_T AS_T x_i$. Hence,

$$\langle S_T y_1 - S_T y_2, x_1 - x_2\rangle \geqq 0 \tag{28}$$

whenever $y_i \in Ax_i$, because $S_T y_i \in S_T Ax_i = S_T AS_T x_i$.

THEOREM 4. *Let A_1, A_2: $H \to 2^H$ be causal and strictly monotone, and let A_1 be an operator. Assume that $S_T D(A_i) \subset D(A_i)$ for each $T \in R^1$ and $i = 1, 2$. Then the F.S. $[A_1, A_2]$ over H is unambiguous and causal.*

Proof. The unambiguity of $[A_1, A_2]$ follows from Theorem 3. Next, let $(u_1, u_2) \mapsto (e_1, e_2)$, $(u_1', u_2') \mapsto (e_1', e_2')$, and assume that, for some $T \in R^1$,

$$S_T u_1 = S_T u_1', \qquad S_T u_2 = S_T u_2'. \tag{29}$$

Then there exists $y_i \in A_i e_i$ and $y_i' \in A_i e_i'$, $i = 1, 2$ such that

$$\begin{aligned} e_1 + y_2 &= u_1, & e_1' + y_2' &= u_1', \\ e_2 - y_1 &= u_2, & e_2' - y_1' &= u_2'. \end{aligned} \tag{30}$$

However, equations (30) and (29) yield

$$\begin{aligned} S_T(e_1 - e_1') + S_T(y_2 - y_2') &= 0, \\ S_T(e_2 - e_2') - S_T(y_1 - y_1') &= 0. \end{aligned} \tag{31}$$

Thus

$$\begin{aligned} \langle S_T(e_1 - e_1'), e_2 - e_2'\rangle + \langle S_T(y_2 - y_2'), e_2 - e_2' / &= 0, \\ \langle S_T(e_2 - e_2'), e_1 - e_1'\rangle - \langle S_T(y_1 - y_1'), e_1 - e_1'\rangle &= 0. \end{aligned} \tag{32}$$

Since S_T is selfadjoint, it follows that

$$(33) \qquad \langle S_T(y_2-y_2'), e_2-e_2'\rangle + \langle S_T(y_1-y_1'), e_1-e_1'\rangle = 0.$$

Since both A_1 and A_2 are monotone, it follows by Lemma 2 (b) that both inner products in (33) are nonnegative, and consequently,

$$(34) \qquad \langle S_T(y_2-y_2'), e_2-e_2'\rangle = 0,$$

$$(35) \qquad \langle S_T(y_1-y_1'), e_1-e_1'\rangle = 0.$$

Thus, by Lemma 2 (a), $S_Te_1 = S_Te_1'$ and $S_Te_2 = S_Te_2'$.

Remark 1. If a F.S. satisfies the assumptions of Theorem 4, then $S_Tu_i = S_Tu_i'$, $i = 1, 2 \Rightarrow S_Te_i = S_Te_i'$, $i = 1, 2$. This in turn implies by (31) that $S_Ty_i = S_Ty_i'$, $i = 1, 2$, i.e., the F.S. is causal in regard to outputs y_1 and y_2.

In order to simplify the formulation and proofs of further theorems, let us introduce the following notation.

Let $\mathcal{M}$ be the set of all operators $N\colon H \to H$ such that

$$(36) \qquad \mu_N = \inf_{\substack{x_1,x_2\in H\\ x_1\neq x_2}} \langle Nx_1-Nx_2, x_1-x_2\rangle \|x_1-x_2\|^{-2} > -\infty.$$

Observe that if $N, M \in \mathcal{M}$ and $\alpha \geqq 0$, then $N+M$, $\alpha N \in \mathcal{M}$ and $\mu_{N+M} \geqq \mu_N + \mu_M$, $\mu_{\alpha N} = \alpha\mu_N$. Also, it is clear that N is monotone [strongly monotone]$\Leftrightarrow \mu_N \geqq 0$ [$\mu_N > 0$].

Furthermore, let Lip be the set of all operators $N\colon H \to H$ such that

$$(37) \qquad \|N\|^* = \sup_{\substack{x_1,x_2\in H\\ x_1\neq x_2}} \|Nx_1-Nx_2\|\cdot\|x_1-x_2\|^{-1} < \infty.$$

It is clear that $\|N\|^* \geqq 0$, and $\|N\|^* = 0 \Leftrightarrow N$ is a constant operator. Also, if $N_1, N_2 \in \mathrm{Lip}$ and α_1, α_2 are real numbers, then $\alpha_1N_1+\alpha_2N_2$, $N_1N_2 \in \mathrm{Lip}$ and

$$\|\alpha_1N_1\|^* = |\alpha_1|\cdot\|N_1\|^*, \qquad \|N_1+N_2\|^* \leqq \|N_1\|^* + \|N_2\|^*,$$
$$\|N_1N_2\|^* \leqq \|N_1\|^*\cdot\|N_2\|^*.$$

If, in particular, N is linear, then N is bounded $\Leftrightarrow N \in \mathrm{Lip}$. In this case $\|N\| = \|N\|^*$.

Finally, note that by virtue of Schwarz inequality,

(i) $\mathrm{Lip} \subset \mathcal{M}$,

(ii) $\|N\|^* \geqq |\mu_N|$ for every $N \in \mathrm{Lip}$.

The numbers $\|N\|^*$ and μ_N can be interpreted crudely as a "gain" and "minimal slope" of the operator N, respectively.

LEMMA 3. *Let $N \in \mathcal{M}$ and let $\mu_N > 0$; if N is hemicontinuous, then N is invertible, $N^{-1} \in Lip$, $\mu_{N^{-1}} \geqq 0$ and*

$$(38) \qquad \|N^{-1}\|^* \leqq \mu_N^{-1}.$$

If, in addition, $N \in \mathrm{Lip}$, *then*

$$(39) \qquad \mu_{N^{-1}} \geqq \mu_N\|N\|^{*-2}.$$

Proof. For all $x_1, x_2 \in H$ we have

$$(40) \qquad \langle Nx_1-Nx_2, x_1-x_2\rangle \geqq \mu_N\|x_1-x_2\|^2.$$

Thus, N is monotone, and because N is hemicontinuous, it is maximal monotone. Also, (40) shows that N is coercive. Hence, $NH=H$ (see [12]).

Moreover, (40) implies that N is 1-to-1, and consequently, N^{-1} exists. By Schwarz inequality we have from (40)

$$\|Nx_1-Nx_2\|\geqq \mu_N\|x_1-x_2\|.$$

Hence, for any $y_1, y_2\in H$, $\|N^{-1}y_1-N^{-1}y_2\|\leqq \mu_N^{-1}\|y_1-y_2\|$, which gives (38).

Also, by (40), $\langle y_1-y_2, N^{-1}y_1-N^{-1}y_2\rangle\geqq 0$, so that $\mu_{N^{-1}}\geqq 0$.

If $N\in\mathrm{Lip}$, then

$$\|Nx_1-Nx_2\|\leqq\|N\|^*\|x_1-x_2\| \tag{41}$$

for all $x_1, x_2\in H$. However, (40) and (41) yield:

$$\langle Nx_1-Nx_2, x_1-x_2\rangle\geqq\mu_N\|N\|^{*-2}\|Nx_1-Nx_2\|^2.$$

Choosing $y_1, y_2\in H$ and setting $x_1=N^{-1}y_1$, $x_2=N^{-1}y_2$, we get

$$\langle y_1-y_2, N^{-1}y_1-N^{-1}y_2\rangle\geqq\mu_N\|N\|^{*-2}\|y_1-y_2\|^2.$$

This, however, proves (39).

LEMMA 4. *Let $N\in\mathcal{M}$ be hemicontinuous and let $\mu_N>0$; if N is causal, then N^{-1} is also causal.*

Proof. By $N\in\mathcal{M}$ we have for any $x_1, x_2\in H$ and $T\in R^1$, $\langle NS_Tx_1-NS_Tx_2, S_Tx_1-S_Tx_2\rangle\geqq\mu_N\|S_Tx_1-S_Tx_2\|^2$; thus, by causality of N,

$$\langle S_TNx_1-S_TNx_2, x_1-x_1\rangle\geqq\mu_N\|S_Tx_1-S_Tx_2\|^2. \tag{42}$$

Since N is invertible by Lemma 3, choose $y_1, y_2\in H$ and put $x_i=N^{-1}y_i$, $i=1,2$, in (42). We get:

$$\langle S_Ty_1-S_Ty_2, N^{-1}y_1-N^{-1}y_2\rangle\geqq\mu_N\|S_TN^{-1}y_1-S_TN^{-1}y_2\|^2.$$

Thus, if $S_Ty_1=S_Ty_2$, we have $S_TN^{-1}y_1=S_TN^{-1}y_2$, and causality of N^{-1} is proven.

LEMMA 5. *Let $A\in\mathcal{M}$ be hemicontinuous, and let $B\in Lip$, $\mu_B>0$. If $\mu_A+\mu_B\|B\|^{*-2}>0$, then $I+AB$ is invertible, $(I+AB)^{-1}\in Lip$ and*

$$\|(I+AB)^{-1}\|^*\leqq\mu_B^{-1}(\mu_A+\mu_B\|B\|^{*-2})^{-1}. \tag{43}$$

If, in addition, both A and B are causal, then $(I+AB)^{-1}$ is also causal.

Proof. The assumptions $B\in\mathrm{Lip}$, $\mu_B>0$ imply by Lemma 3 that B is invertible, $B^{-1}\in\mathcal{M}$, $\mu_{B^{-1}}\geqq\mu_B\|B\|^{*-2}$, $B^{-1}\in\mathrm{Lip}$ and $\|B^{-1}\|^*\leqq\mu_B^{-1}$. Thus, $B^{-1}+A$ is hemicontinuous, $B^{-1}+A\in\mathcal{M}$ and

$$\mu_{B^{-1}+A}\geqq\mu_A+\mu_B\|B\|^{*-2}>0. \tag{44}$$

Hence, again by Lemma 3, $B^{-1}+A$ is invertible, $(B^{-1}+A)^{-1}\in\mathrm{Lip}$ and

$$\|(B^{-1}+A)^{-1}\|^*\leqq(\mu_A+\mu_B\|B\|^{*-2})^{-1}.$$

Next, $I+AB=(B^{-1}+A)B$, and consequently, $I+AB$ is invertible. Since

$$(I+AB)^{-1}=B^{-1}(B^{-1}+A)^{-1} \tag{45}$$

we have by the above

$$\|(I+AB)^{-1}\|^*\leqq\|B^{-1}\|^*\|(B^{-1}+A)^{-1}\|^*\leqq\mu_B^{-1}(\mu_A+\mu_B\|B\|^{*-2})^{-1}.$$

Now, let A and B be causal. Then Lemma 4 and the above relations show that B^{-1} is causal, and consequently, $B^{-1}+A$ is causal. Moreover, (44) implies that $(B^{-1}+A)^{-1}$ is causal. Hence, by (45), $(I+AB)^{-1}$ is causal.

THEOREM 5. *Let $A_1 \in Lip$, $\mu_{A_1} > 0$, and let $A_2 \in \mathcal{M}$ be hemicontinuous. If $\mu_{A_2} + \mu_{A_1}\|A_1\|^{*-2} > 0$, then the F.S. $[A_1, A_2]$ over H is normal.*

Moreover, if $(u_1, u^*) \mapsto (e_1, e_2)$ and $(u_1', u^*) \mapsto (e_1', e_2')$, then

$$\|e_1 - e_1'\| \leqq \mu_{A_1}^{-1}(\mu_{A_2} + \mu_{A_1}\|A_1\|^{*-2})^{-1}\|u_1 - u_1'\|, \tag{46}$$

and

$$\|e_2 - e_2'\| \leqq \|A_1\|^* \mu_{A_1}^{-1}(\mu_{A_2} + \mu_{A_1}\|A_1\|^{*-2})^{-1}\|u_1 - u_1'\|. \tag{47}$$

If, in addition, both A_1 and A_2 are causal, then $[A_1, A_2]$ is causal by the first input.

Proof. Let $A = A_2$, and, for a chosen $u^* \in H$, let $B: H \to H$ be defined by $Bx = u^* + A_1x$. Then clearly $B \in \text{Lip}$, $\mu_B = \mu_{A_1} > 0$ and $\|B\|^* = \|A_1\|^*$. Thus, by Lemma 5, $I + AB$ is invertible and (43) holds. Hence, by Corollary 1, (iii), $[A_1, A_2]$ is normal (see (2)).

Moreover, from (13) it follows that

$$e_1 - e_1' = M_{u^*}^{-1}u_1 - M_{u^*}^{-1}u_1', \tag{48}$$

and

$$e_2 - e_2' = A_1M_{u^*}^{-1}u_1 - A_1M_{u^*}^{-1}u_1'. \tag{49}$$

Since $M_{u^*} = I + AB$, (46) follows from (43). Also, (49) yields $\|e_2 - e_2'\| \leqq \|A_1\|^*\|M_{u^*}^{-1}u_1 - M_{u^*}^{-1}u_1'\|$. The rest is obvious.

As for the causality of $[A_1, A_2]$ by the first input, assume that $(u_1, u^*) \mapsto (e_1, e_2)$, $(u_1', u^*) \mapsto (e_1', e_2')$ and $S_Tu_1 = S_Tu_2$ for some $T \in R^1$. Clearly, with a fixed u^*, the operator B is causal. Thus, by Lemma 5, $M_{u^*}^{-1}$ is causal and (48) yields

$$S_T(e_1 - e_1') = S_TM_{u^*}^{-1}u_1 - S_TM_{u^*}^{-1}u_1' = S_TM_{u^*}^{-1}S_Tu_1 - S_TM_{u^*}^{-1}S_Tu_1' = 0.$$

Finally, (49) and causality of A_1 imply that $S_T(e_2 - e_2') = 0$. Theorem 5 can be improved, if either A_1 or A_2 is linear. To this end, we need the following:

LEMMA 6. *Let $A \in \mathcal{M}$ be a linear operator with $\mu_A > 0$, and let $B \in \mathcal{M}$ be a hemicontinuous operator with $\mu_B \leqq 0$. If $\mu_A + \mu_B\|A\|^2 > 0$, then $I + AB$ is invertible, $(I + AB)^{-1} \in Lip$ and*

$$\|(I + AB)^{-1}\|^* \leqq \|A\|(\mu_A + \mu_B\|A\|^2)^{-1}. \tag{50}$$

If, in addition, both A and B are causal, then $(I + AB)^{-1}$ is also causal.

Proof. First observe that A is bounded. Indeed, since A is linear, it is hemicontinuous. Thus, by Lemma 3, A is invertible, $A^{-1} \in \text{Lip}$ and $\|A^{-1}\| \leqq \mu_A^{-1}$. Consequently, by the open mapping theorem, A itself is bounded.

Next, consider the operator $C = (I + AB)A^* = A^* + ABA^*$. Clearly, C is hemicontinuous. Indeed, A^* is continuous and for any $x_0, w \in H$ and number sequence $t_n \to 0$ we have

$$BA^*(x_0 + t_nw) = B(A^*x_0 + t_nA^*w) \overset{w}{\to} BA^*x_0.$$

Since A is bounded, it follows that $ABA^*(x_0 + t_nw) \overset{w}{\to} ABA^*x_0$.

Moreover, for any $x_1, x_2 \in H$,

$$\begin{aligned} &\langle Cx_1 - Cx_2, x_1 - x_2\rangle \\ &= \langle A^*(x_1 - x_2), x_1 - x_2\rangle + \langle ABA^*x_1 - ABA^*x_2, x_1 - x_2\rangle \\ &= \langle A(x_1 - x_2), x_1 - x_2\rangle + \langle BA^*x_1 - BA^*x_2, A^*x_1 - A^*x_2\rangle. \end{aligned} \tag{51}$$

However,

$$\langle BA^*x_1 - BA^*x_2, A^*x_1 - A^*x_2\rangle \geqq \mu_B\|A^*(x_1-x_2)\|^2,$$

and

$$\|A^*(x_1-x_2)\| \leqq \|A^*\|\,\|x_1-x_2\| = \|A\|\,\|x_1-x_2\|.$$

Thus, by (51)

$$\langle Cx_1 - Cx_2, x_1 - x_2\rangle \geqq (\mu_A + \mu_B\|A\|^2)\|x_1-x_2\|^2,$$

i.e.,

$$\mu_C \geqq \mu_A + \mu_B\|A\|^2 > 0.$$

Hence, by Lemma 3, C is invertible, $C^{-1} \in \mathrm{Lip}$ and

$$\|C^{-1}\|^* \leqq \mu_C^{-1} \leqq (\mu_A + \mu_B\|A\|^2)^{-1}. \tag{52}$$

Now, since A^* is invertible and $\|A^{*-1}\| \leqq \mu_A^{-1}$, we have $I + AB = CA^{*-1}$. Consequently, $I + AB$ is also invertible, and the equation $(I+AB)^{-1} = A^*C^{-1}$ yields immediately (50) by (52).

To prove causality, note first that, by Lemma 4, A^{-1} is causal. Also note that, by Lemma 3, $\mu_{A^{-1}} \geqq \mu_A\|A\|^{-2}$. Thus, $A^{-1}+B \in \mathcal{M}$, $\mu_{A^{-1}+B} \geqq \mu_B + \mu_A\|A\|^{-2} > 0$ and $A^{-1}+B$ is hemicontinuous. Hence, by Lemma 3, $A^{-1}+B$ is invertible, and $(A^{-1}+B)^{-1}$ is causal by Lemma 4. Since $I+AB = A(A^{-1}+B)$, we have $(I+AB)^{-1} = (A^{-1}+B)^{-1}A^{-1}$, and consequently, $(I+AB)^{-1}$ is causal.

Remark 2. The invertibility of $I+AB$ can also be proved via invertibility of $A^{-1}+B$, but then we obtain a worse estimate than (50).

THEOREM 6. *Let $A_1 \in \mathcal{M}$ with $\mu_{A_1} \leqq 0$ be hemicontinuous, and let $A_2 \in \mathcal{M}$ with $\mu_{A_2} > 0$ be linear. If $\mu_{A_2} + \mu_{A_1}\|A_2\|^2 > 0$, then the F.S. $[A_1, A_2]$ over H is normal.*

Moreover,

(i) *if both A_1 and A_2 are causal, then $[A_1, A_2]$ is causal.*

(ii) *if $(u_1, u_2) \mapsto (e_1, e_2)$ and $(u_1', u_2') \mapsto (e_1', e_2')$, then*

$$\|e_1 - e_1'\| \leqq \lambda\|u_1 - u_1'\| + \lambda\|A_2\| \cdot \|u_2 - u_2'\|, \tag{53}$$

where

$$\lambda = \|A_2\|(\mu_{A_2} + \mu_{A_1}\|A_2\|^2)^{-1}.$$

If, in addition, $A_1 \in Lip$, then

$$\|e_2 - e_2'\| \leqq \|A_1\|^*\lambda\|u_1 - u_1'\| + (1 + \|A_1\|^*\lambda\|A_2\|)\|u_2 - u_2'\|. \tag{54}$$

Proof. By Lemma 6 it follows that the operator $I + A_2A_1$ is invertible and $\|(I+A_2A_1)^{-1}\|^* \leqq \lambda$. Thus, by Corollary 2, $[A_1, A_2]$ is normal.

Moreover, by (14) it follows that

$$\begin{aligned}\|e_1 - e_1'\| &= \|(I+A_2A_1)^{-1}(u_1 - A_2u_2) - (I+A_2A_1)^{-1}(u_1' - A_2u_2')\| \\ &\leqq \lambda\|u_1 - u_1'\| + \lambda\|A_2\| \cdot \|u_2 - u_2'\|,\end{aligned}$$

which is (53).

Similarly, if $A_1 \in \mathrm{Lip}$, then (14) yields

$$\|e_2 - e_2'\| \leqq \|u_2 - u_2'\| + \|A_1\|^*\lambda\|u_1 - A_2u_2 - u_1' + A_2u_2'\|,$$

which readily gives (54).

Finally, let A_1, A_2 be causal: If $(u_1, u_2)\mapsto(e_1, e_2)$, $(u_1', u_2')\mapsto(e_1', e_2')$ and $S_Tu_1 = S_Tu_1'$, $S_Tu_2 = S_Tu_2'$, then (14) and causality of $K = (I + A_2A_1)^{-1}$ yield

$$\begin{aligned} S_Te_1 - S_Te_1' &= S_TK(u_1 - A_2u_2) - S_TK(u_1' - A_2u_2') \\ &= S_TK(S_Tu_1 - S_TA_2S_Tu_2) - S_TK(S_Tu_1' - S_TA_2S_Tu_2') = 0. \end{aligned}$$

In the same manner it follows that $S_Te_2 - S_Te_2' = 0$ which completes the proof.

Remark 3. If a F.S. $[A_1, A_2]$ over H satisfies the assumptions of Theorem 6 (including $A_1 \in \text{Lip}$), and if $A_10 = 0$, then $(0, 0) \in H^2$ is the unique solution of $[A_1, A_2]$ corresponding to $(0, 0) \in H^2$. Thus, if $(u_1, u_2)\mapsto(e_1, e_2)$, we get from (53) and (54)

$$\begin{aligned} &\|e_1\| \leqq \lambda\|u_1\| + \lambda\|A_2\| \cdot \|u_2\|, \\ &\|e_2\| \leqq \|A_1\|^*\lambda\|u_1\| + (1 + \|A_1\|^*\lambda\|A_2\|)\|u_2\|. \end{aligned}$$

Consequently, $[A_1, A_2]$ is stable in the standard sense (see [8], p. 52).

Let us now prove a "linear counterpart" of Lemma 5.

LEMMA 7. *Let $A \in \mathcal{M}$ be hemicontinuous with $\mu_A \leqq 0$, and let $B \in \mathcal{M}$ be linear with $\mu_B > 0$. If $\mu_B + \mu_A\|B\|^2 > 0$, then $I + AB$ is invertible, $(I + AB)^{-1} \in Lip$ and*

$$\|(I + AB)^{-1}\|^* \leqq \|B\|(\mu_B + \mu_A\|B\|^2)^{-1}. \tag{55}$$

If, in addition, both A and B are causal, then $(I + AB)^{-1}$ is also causal.

Proof. Since linearity of B and $B \in \mathcal{M}$, $\mu_B > 0$ imply that B is bounded, i.e., $B \in \text{Lip}$, our assertions concerning invertibility of $I + AB$ and possible causality of $(I + AB)^{-1}$ follow immediately from Lemma 5. However, the bound for $\|(I + AB)^{-1}\|^*$ given by (55) is smaller than that given by (43), since the latter is a $\mu_B^{-1}\|B\| \geqq 1$ multiple of the former.

To confirm (55), we use an argument similar to the proof of Lemma 6. First, note that B^* is invertible and $\|B^{*-1}\| \leqq \mu_B^{-1}$. Letting $C = B^*(I + AB)$, we verify as before that C is hemicontinuous and satisfies the condition

$$\langle Cx_1 - Cx_2, x_1 - x_2\rangle \geqq (\mu_B + \mu_A\|B\|^2)\|x_1 - x_2\|^2$$

for all $x_1, x_2 \in H$. Thus, by Lemma 3, C is invertible, $C^{-1} \in \text{Lip}$ and

$$\|C^{-1}\|^* \leqq (\mu_B + \mu_A\|B\|^2)^{-1}. \tag{56}$$

Now, since B^* is invertible and $\|B^{*-1}\| \leqq \mu_B^{-1}$, we have $I + AB = B^{*-1}C$. Consequently, $I + AB$ is invertible and $(I + AB)^{-1} = C^{-1}B^*$. From this and (56) we conclude that (55) holds.

LEMMA 8. *Let $A, B \colon H \to H$ be operators, let B be linear, and let B and $N = I + AB$ be invertible. If $a \in H$, let $M_a \colon H \to H$ be defined by*

$$M_ax = x + A(a + Bx). \tag{57}$$

Then M_a is invertible, and

$$M_a^{-1}x = N^{-1}(x + B^{-1}a) - B^{-1}a \tag{58}$$

for each $x \in H$.

Proof. Let Q denote the operator standing on the right-hand side of (58). Then, for any $x \in H$, we have

$$\begin{aligned} QM_ax &= N^{-1}(x + A(a + Bx) + B^{-1}a) - B^{-1}a \\ &= N^{-1}(x + B^{-1}a + AB(x + B^{-1}a)) - B^{-1}a = x. \end{aligned}$$

Also,

$$\begin{aligned}M_aQx&=N^{-1}(x+B^{-1}a)-B^{-1}a+A(a+B[N^{-1}(x+B^{-1}a)-B^{-1}a])\\&=N^{-1}(x+B^{-1}a)-B^{-1}a+ABN^{-1}(x+B^{-1}a)\\&=(I+AB)N^{-1}(x+B^{-1}a)-B^{-1}a=x.\end{aligned}$$

Hence, $Q=M_a^{-1}$.

THEOREM 7. *Let $A_1\in\mathcal{M}$ with $\mu_{A_1}>0$ be linear, and let $A_2\in\mathcal{M}$ with $\mu_{A_2}\leqq 0$ be hemicontinuous. If $\mu_{A_1}+\mu_{A_2}\|A_1\|^2>0$, then the F.S. $[A_1, A_2]$ over H is normal. Moreover, if $(u_1, u_2)\mapsto(e_1, e_2)$ and $(u_1', u_2')\mapsto(e_1', e_2')$, then*

$$\|e_1-e_1'\|\leqq\lambda\|u_1-u_1'\|+(1+\lambda)\mu_{A_1}^{-1}\|u_2-u_2'\| \tag{59}$$

and

$$\|e_2-e_2'\|\leqq\|A_1\|\lambda\|u_1-u_1'\|+[1+(1+\lambda)\mu_{A_1}^{-1}\|A_1\|]\|u_2-u_2'\|, \tag{60}$$

where

$$\lambda=\|A_1\|(\mu_{A_1}+\mu_{A_2}\|A_1\|^2)^{-1}.$$

If, in addition, both A_1 and A_2 are causal, then $[A_1, A_2]$ is causal.

Proof. By Lemma 3, A_1 is invertible, by Lemma 7, $N=I+A_2A_1$ is invertible, their inverses are in Lip and $\|A_1^{-1}\|\leqq\mu_{A_1}^{-1}$, $\|N^{-1}\|^*\leqq\lambda$. By Lemma 8, M_a defined by (2) is invertible for any $a\in H$ and (58) holds. Thus, by Corollary 1, $[A_1, A_2]$ is normal.

Moreover, by (13) and (58) we have

$$\begin{aligned}\|e_1-e_1'\|&=\|M_{u_2}^{-1}u_1-M_{u_2'}^{-1}u_1'\|\\&=\|N^{-1}(u_1+A_1^{-1}u_2)-A_1^{-1}u_2-N^{-1}(u_1'+A_1^{-1}u_2')-A_1^{-1}u_2'\|\\&\leqq\lambda\|u_1-u_1'+A_1^{-1}(u_2-u_2')\|+\|A_1^{-1}(u_2-u_2')\|\\&\leqq\lambda\|u_1-u_1'\|+(1+\lambda)u_{A_1}^{-1}\|u_2-u_2'\|.\end{aligned} \tag{61}$$

Also by (13),

$$\|e_2-e_2'\|\leqq\|u_2-u_2'\|+\|A_1\|\,\|M_{u_2}^{-1}u_1-M_{u_2'}^{-1}u_1'\|$$

which gives readily (60) by invoking (61). Finally, let A_1 and A_2 be causal. Then A_1^{-1} is causal by Lemma 4, and N^{-1} is causal by Lemma 7. If $(u_1, u_2)\mapsto(e_1, e_2)$, $(u_1', u_2')\mapsto(e_1', e_2')$ and $S_Tu_1=S_Tu_1'$, $S_Tu_2=S_Tu_2'$, then (13) and (58) yield

$$\begin{aligned}S_Te_1-S_Te_1'&=S_TN^{-1}(u_1+A_1^{-1}u_2)-S_TA_1^{-1}u_2-S_TN^{-1}(u_1'+A_1^{-1}u_2')-S_TA_1^{-1}u_2'\\&=S_TN^{-1}(S_Tu_1+S_TA_1^{-1}S_Tu_2)-S_TA_1^{-1}S_Tu_2\\&\qquad-S_TN^{-1}(S_Tu_1'+S_TA_1^{-1}S_Tu_2')+S_TA_1^{-1}S_Tu_2'=0.\end{aligned}$$

A similar argument shows that $S_Te_2-S_Te_2'=0$ which concludes the proof.

Remark 4. If a F.S. $[A_1, A_2]$ over H satisfies the assumptions of Theorem 7 and if $A_20=0$, then $(0, 0)\in H^2$ is the unique solution corresponding to $(0, 0)\in H^2$, and (59), (60) yield

$$\begin{aligned}\|e_1\|&\leqq\lambda\|u_1\|+(1+\lambda)\mu_{A_1}^{-1}\|u_2\|,\\\|e_2\|&\leqq\|A_1\|\lambda\|u_1\|+[1+(1+\lambda)\mu_{A_1}^{-1}\|A_1\|]\|u_2\|.\end{aligned}$$

Hence, $[A_1, A_2]$ is stable in the standard sense.

In concluding this part of the paper, let us present an example of a specific F.S.

Example 1. Let $n \geqq 1$ be an integer, and let $H = L_2^n$, where L_2^n is the n-fold Cartesian product of $L_2[0, \infty)$ with itself equipped with usual inner product.

Let D be a real $n \times n$ matrix, and denote

$$d = \inf_{\substack{|\xi|=1 \\ \xi \in R^n}} \xi^T D \xi \tag{62}$$

(here, $|\cdot|$ denotes the Euclidean norm).

Also, let $K(t)$ be an $n \times n$ matrix whose entries $K_{ij}(t) \in L_1[0, \infty) \cap L_2[0, \infty)$, and let $\hat{K}(iw)$ be the Fourier transform of $K(t)$, (defined as zero for $t < 0$).

Denote

$$k = \tfrac{1}{2} \inf_{w \in R^1} \inf_{|\xi|=1} \xi^T (\hat{K}(iw) + \overline{\hat{K}(iw)^T}) \xi, \tag{63}$$

and

$$\kappa = \sup_{w \in R^1} \Lambda(\hat{K}(iw)), \tag{64}$$

where $\Lambda(M)$ denotes the square root of the largest eigenvalue of the matrix $\bar{M}^T M$. (Note that $-\infty < k \leqq 0$ and $0 \leqq \kappa < \infty$).

Furthermore, let ϕ: $R^n \to R^n$ be such that

$$\phi(0) = 0, \tag{65}$$

$$|\phi(\xi_1) - \phi(\xi_2)| \leqq \beta |\xi_1 - \xi_2| \tag{66}$$

for some $\beta > 0$ and all $\xi_1, \xi_2 \in R^n$, and

$$\inf_{\substack{\xi_1, \xi_2 \in R^n \\ \xi_1 \neq \xi_2}} (\phi(\xi_1) - \phi(\xi_2))^T (\xi_1 - \xi_2) \cdot |\xi_1 - \xi_2|^{-2} = a_2 \leqq 0. \tag{67}$$

A typical example of such function ϕ in one dimension is given in Fig. 2.

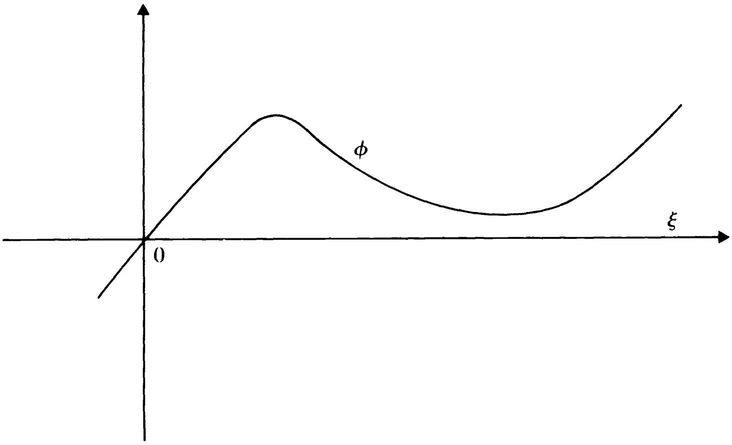

FIG. 2

Now, define the operators A_1, A_2 on H by

$$(A_1 x)(t) = Dx(t) + \int_0^t K(t-\tau) x(\tau)\, d\tau, \qquad t \geqq 0, \tag{68}$$

and

(69) $$(A_2x)(t)=\phi(x(t)).$$

We are going to show that, if

(70) $$d+k>0$$

and

(71) $$d+k+a_2(|D|+\kappa)^2>0,$$

($|D|$ denotes the norm of D associated with the Euclidean vector norm), then the F.S. $[A_1, A_2]$ is normal.

Indeed, it is clear that A_1 is linear, maps H into itself and is bounded. A straightforward argument using Parseval's equality and (63) shows that $A_1 \in \mathcal{M}$ and $\mu_{A_1} \geqq d+k>0$ by (70). Also, it is known [16] that $\|A_1\| \leqq |D|+\kappa$.

On the other hand (65) and (66) show that A_2 maps H into itself and is a continuous mapping. A little thought will persuade us that, due to (67), $A_2 \in \mathcal{M}$ and $\mu_{A_2}=a_2 \leqq 0$.

Thus, we have by (71), $\mu_{A_1}+\mu_{A_2}\|A_2\|^2 \geqq d+k+a_2(|D|+\kappa)^2>0$. Hence, by Theorem 7, our F.S. $[A_1, A_2]$ is normal. Also, estimates (59), (60) hold with

$$\lambda \leqq (|D|+\kappa)(d+k+a_2(|D|+\kappa)^2)^{-1},$$

i.e., our F.S. is Lipschitz continuous in both inputs.

Finally, it is well known that, if $S_T: L_2^n \to L_2^n$ is defined by $(S_Tx)(t)=x(t)$ for $t \leqq T$ and $(S_Tx)(t)=0$ for $t>T$, then $\{S_T: T \in R^1\}$ is a resolution of identity on L_2^n. Also, it is clear that A_1 and A_2 are causal with respect to $\{S_T: T \in R^1\}$. Hence, by Theorem 7, our F.S. is causal, i.e., $u_1(t)=u_1'(t)$ and $u_2(t)=u_2'(t)$ on $[0, T]$ implies that $e_1(t)=e_1'(t)$ and $e_2(t)=e_2'(t)$ on $[0, T]$.

3. Feedback systems over an extended Hilbert space. The results obtained in the second part can easily be generalized to the case in which the underlying space is an extended Hilbert space. As it will be apparent later, the Lemma 10 given below is a crucial tool for deriving results concerning extended spaces. However, since otherwise the proofs of generalized results are quite straightforward, we will discuss only a counterpart of Theorem 5.

We introduce the following framework:

Let H_e be a real linear space and let $H \subset H_e$ be a Hilbert space. Moreover, let $\Pi=\{P_\alpha: \alpha \in I\}$ be a nonempty collection of linear operators $P_\alpha: H_e \to H_e$ which satisfies the following axioms:

(i) $P^2=P$ for each $P \in \Pi$.

(ii) $P_1P_2=P_2P_1$ and $P_1P_2 \in \Pi$ for all $P_1, P_2 \in \Pi$.

(iii) If $x \in H_e$, then $Px \in H$ for each $P \in \Pi$.

(iv) If, for every $P \in \Pi$, the element $x^{(P)}$ is in $H_P=PH$ and $P_0x^{(P_1)}=P_0x^{(P_2)}$ for all $P_1, P_2 \in \Pi$ with $P_0=P_1P_2$, then there exists $x \in H_e$ such that $x^{(P)}=Px$ for every $P \in \Pi$.

(v) If $x \in H$, then $\|Px\| \leqq \|x\|$ for each $P \in \Pi$.

(vi) If $x \in H_e$ and $\|Px\| \leqq a$ for every $P \in \Pi$ and some fixed $a \geqq 0$, then $x \in H$ and $\|x\| \leqq a$.

Then H_e will be called an extended Hilbert space, or an extension of H.

LEMMA 9. (a) *$P_0P_1=P_0P_2=P_0$ for every $P_1, P_2 \in \Pi$ and $P_0=P_1P_2$.*

(b) *$PH_e=PH=H_P \subset H$ for each $P \in \Pi$.*

(c) *If $x \in H_e$ and $Px=0$ for each $P \in \Pi$, then $x=0$.*

(d) *The element $x \in H_e$ in axiom* (iv) *is unique.*

(e) *$\langle Px, y\rangle=\langle x, Py\rangle$ for every $x, y \in H$ and $P \in \Pi$.*

Proof. (a) By (i), (ii) we have $P_0P_1 = P_1P_2P_1 = P_1^2P_2 = P_1P_2 = P_0$. Also, $P_0P_2 = P_1P_2^2 = P_0$.

(b) If $z \in PH_e$, then $z = Py$ for some $y \in H_e$. However, by (i), $z = P(Py)$ and $Py \in H$ by (iii), so that $z \in PH$. Hence, $PH_e \subset PH$. Conversely, since $H \subset H_e$, we have $PH \subset PH_e$. The inclusion $PH_e \subset H$ is in fact the axiom (iii).

The assertion (c) is a trivial consequence of (vi). Similarly, (d) follows immediately from (c).

As for (e), choose $x, y \in H$ and $P \in \Pi$. Since $Px, Py \in H$ by (b), we have for every real λ by (i) and (v),

$$\|Px + \lambda y\| \geqq \|P(Px + \lambda y)\| = \|Px + \lambda Py\|,$$

i.e.

$$2\lambda\{\langle Px, y\rangle - \langle Px, Py\rangle\} + \lambda^2\{\langle y, y\rangle - \langle Py, Py\rangle\} \geqq 0.$$

Thus, necessarily

$$\langle Px, y\rangle = \langle Px, Py\rangle.$$

Interchanging the role of x and y, we get $\langle Px, Py\rangle = \langle x, Py\rangle$ and (e) follows.

Let $A: H_e \to H_e$; then A will be called causal, if

$$PA = PAP \tag{72}$$

for every $P \in \Pi$.

Before proceeding further, let us point out the following facts: The axioms (i), (v) and assertion (e) show that each $P \in \Pi$ restricted to H is an orthogonal projection on H. Moreover, if in particular the collection $\Pi = \{P_\alpha : \alpha \in I\}$ can be indexed by reals, i.e., $I = R^1$, if $P_{\alpha_1}P_{\alpha_1} = P_{\min[\alpha_1,\alpha_2]}$ and if we impose the additional requirement: (vii) P_α is right-continuous at every $\alpha \in R^1$ and $P_\alpha x \to x$, $P_{-\alpha}x \to 0$ as $\alpha \to \infty$ for every $x \in H$, then we can easily verify that Π is a resolution of identity on H. Thus, if an operator $A: H_e \to H_e$ is causal by definition (72) and $AH \subset H$, then its restriction $A_0: H \to H$ to H is causal in the sense defined in the second part.

LEMMA 10. *Let A: $H_e \to H_e$ be causal. For each $P \in \Pi$, let A_P: $H_P \to H_P = PH$ be the restriction of PA to H_P. Then A is invertible and the inverse A^{-1}: $H_e \to H_e$ is causal $\Leftrightarrow A_P$ is invertible for every $P \in \Pi$. In this case,*

$$PA^{-1} = A_P^{-1}P \tag{73}$$

for every $P \in \Pi$, and A_P^{-1} is the restriction of PA^{-1} to H_P.

Proof. First note that, due to Lemma 9(b), A_P truly maps H_P into itself.

(1) Assume that A_P is invertible for each $P \in \Pi$. Choose $y \in H_e$, and for each $P \in \Pi$ let

$$x^{(P)} = A_P^{-1}Py \in H_P. \tag{74}$$

We are going to show that

$$P_0x^{(P_1)} = P_0x^{(P_2)} \tag{75}$$

whenever $P_1, P_2 \in \Pi$ and $P_0 = P_1P_2$. Indeed, by (74) we have $A_Px^{(P)} = Py$, i.e.

$$P_iAx^{(P_i)} = P_iy, \tag{76}$$

$i = 1, 2$. Thus, by Lemma 9(a), $P_0Ax^{(P_i)} = P_0y$, and by causality of A, $P_0A(P_0x^{(P_i)}) =$

P_0y. Since $P_0y, P_0x^{(P_i)} \in P_0H = H_{P_0}$, it follows that

$$A_{P_0}(P_0x^{(P_i)}) = P_0y. \tag{77}$$

Thus, (75) follows by invertibility of A_{P_0}.

Next, by virtue of (75) and (iv) there exists a unique $x \in H_e$ such that

$$x^{(P)} = Px \tag{78}$$

for every $P \in \Pi$.

The element x is a solution of the equation $Ax = y$. Indeed, choose $P \in \Pi$. Then we have by causality of A and (78), (74), $P(Ax - y) = PAx - Py = PAPx - Py = PAx^{(P)} - Py = A_Px^{(P)} - Py = 0$. Hence, by Lemma 9(c), $Ax - y = 0$.

As for uniqueness of x, let $\tilde{x} \in H_e$ be such that $A\tilde{x} = y$. Then we have for every $P \in \Pi$, $PAP\tilde{x} = Py$ and also $PAPx = Py$. Hence, $A_PP\tilde{x} = Py$, $A_PPx = Py$ which implies that $P\tilde{x} = Px$. Thus, by Lemma 9(c), $x = \tilde{x}$. Consequently, A is invertible.

Moreover, since $x = A^{-1}y$, we have for any $P \in \Pi$, $Px = PA^{-1}y$. Using (78) and (74) it follows that $PA^{-1}y = x^{(P)} = A_P^{-1}Py$, which confirms (73). Also, since $A_P^{-1}P$ restricted to H_P is A_P^{-1}, (73) proves that A_P^{-1} is the restriction of PA^{-1} to H_P.

Finally, by (73) and (i), $PA^{-1}P = A_P^{-1}P^2 = A_P^{-1}P = PA^{-1}$, i.e., $\tilde{A}^{-1}$ is causal.

(2) Conversely, let A be invertible, and let A^{-1} be causal. Choose $P \in \Pi$ and denote $B: H_P \to H_P$ the restriction of PA^{-1} to H_P. If $x \in H_P$, we have $BA_Px = BPAx = PA^{-1}PAx = PA^{-1}Ax = Px = x$. Also, $A_PBx = A_PPA^{-1}x = PAPA^{-1}x = PAA^{-1}x = Px = x$. Hence, $B = A_P^{-1}$, i.e., A_P is invertible.

Now we can state the counterpart of Theorem 5.

THEOREM 8. *Let A_1, A_2: $H_e \to H_e$ be causal operators. For every $P \in \Pi$, let A_{1P} and A_{2P} be the restriction of PA_1 and PA_2 to $H_P = PH$, respectively. Assume that, for each $P \in \Pi$,*

(i) *$A_{1P} \in Lip$ and $\mu_{A_{1P}} > 0$,*

(ii) *$A_{2P} \in \mathcal{M}$ and is hemicontinuous,*

(iii) *$\mu_{A_{2P}} + \mu_{A_{1P}}\|A_{1P}\|^{*-2} > 0$.*

Then the F.S. $[A_1, A_2]$ over H_e is normal and causal, i.e., if $(u_1, u_2) \in H_e^2$, $(u_1, u_2) \to (e_1, e_2) \in H_e^2$, $(u_1', u_2') \in H_e^2$, $(u_1', u_2') \to (e_1', e_2') \in H^2$ and $Pu_1 = Pu_1'$, $Pu_2 = Pu_2'$, then $Pe_1 = Pe_1'$ and $Pe_2 = Pe_2'$.

Moreover, assume that, in addition, there exists $\lambda > 0$ such that

$$\mu_{A_{1P}}^{-1}(\mu_{A_{2P}} + \mu_{A_{1P}}\|A_{1P}\|^{*-2})^{-1} \leqq \lambda \tag{79}$$

for all $P \in \Pi$. If $(u_1, u^*) \to (e_1, e_2) \in H_e^2$, $(u_1', u^*) \to (e_1', e_2') \in H_e^2$ and $u_1 - u_1' \in H$, then $e_1 - e_1' \in H$ and

$$\|e_1 - e_1'\| \leqq \lambda \|u_1 - u_1'\|. \tag{80}$$

If, in addition, $\|A_{1P}\|^* \leqq \kappa$ for all $P \in \Pi$ and some $\kappa > 0$, then also $e_2 - e_2' \in H$ and we have

$$\|e_2 - e_2'\| < \kappa\lambda \|u_1 - u_1'\|. \tag{81}$$

Proof. Choose a fixed $z \in H_e$ and consider the operator M_z: $H_e \to H_e$ defined by $M_zx = x + A_2(z + A_1x)$. Defining the operator B_z: $H_e \to H_e$ by $B_zx = z + A_1x$, we see that B_z is causal and $M_z = I + A_2B_z$.

Next, choose $P \in \Pi$. Since P is an orthogonal projection on H, it follows that H_P is a Hilbert space of its own right. Also, let B_{zP}: $H_P \to H_P$ be the restriction of PB_z to H_P. By assumption (i), $B_{zP} \in \text{Lip}$ and $\|B_{zP}\|^* = \|A_{1P}\|^*$, $\mu_{B_{zP}} = \mu_{A_{1P}} > 0$. Hence, by (ii), (iii) and

Lemma 5, the operator $N_P = I + A_{2P}B_{zP}\colon H_P \to H_P$ is invertible, $N_P^{-1} \in \mathrm{Lip}$ and

$$\|N_P^{-1}\|^* \leqq \mu_{A_{1P}}^{-1}(\mu_{A_{2P}} + \mu_{A_{1P}}\|A_{1P}\|^{*-2})^{-1} = \lambda_P. \tag{82}$$

However, N_P is the restriction of PM_z to H_P, since for any $x \in H_P$ we have $PM_zx = P(I + A_2B_z)x = x + PA_2PB_zx = (I + A_{2P}B_{zP})x$. Also, M_z is causal, since $PM_z = P + PA_2B_z = P + PA_2PB_zP = P(I + A_2B_z)P = PM_zP$. Hence, by Lemma 10, M_z is invertible, M_z^{-1} is causal and we have by (73),

$$PM_z^{-1} = N_P^{-1}P. \tag{83}$$

Consequently, $[A_1, A_2]$ is normal by Corollary 1.

To prove causality of $[A_1, A_2]$, observe first that for any $z \in H_e$ and $P \in \Pi$ we have $PM_z = PM_{Pz}$. Now, assume that $(u_1, u_2) \to (e_1, e_2)$, $(u_1', u_2') \to (e_1', e_2')$ and $Pu_1 = Pu_1' = p$, $Pu_2 = Pu_2' = q$. Then we have by (13), $M_{u_2}e_1 = u_1$, $M_{u_2'}e_1' = u_1'$. Consequently, $PM_{u_2}Pe_1 = Pu_1$, $PM_{u_2'}Pe_1' = Pu_1'$, i.e.,

$$PM_qPe_1 = p, \qquad PM_qPe_1' = p. \tag{84}$$

Since $p, Pe_1, Pe_1' \in H_P$ and PM_q is invertible (it is the above operator N_P for $z = q$), (84) shows that $Pe_1 = Pe_1'$.

Moreover, since $e_2 = A_1e_1 + u_2$ and $e_2' = A_1e_1' + u_2'$ by (1), we conclude readily that $Pe_2 = Pe_2'$.

Finally, assume that (79) is satisfied, and let $(u_1, u^*) \to (e_1, e_2)$, $(u_1', u^*) \to (e_1', e_2')$ and $u_1 - u_1' \in H$. Invoking (13) and (83) it follows that $Pe_1 = PM_{u^*}^{-1}u_1 = PM_{u^*}^{-1}Pu_1 = N_P^{-1}Pu_1$ and similarly for Pe_1'. Hence, by (82), (79) and axiom (v),

$$\begin{aligned}\|P(e_1 - e_1')\| = \|N_P^{-1}Pu_1 - N_P^{-1}Pu_1'\| &\leqq \|N_P^{-1}\|^*\|P(u_1 - u_1')\| \\ &\leqq \lambda\|P(u_1 - u_1')\| \leqq \lambda\|u_1 - u_1'\|.\end{aligned}$$

Thus, by axiom (vi), $e_1 - e_1' \in H$ and (80) holds.

Moreover, since $e_2 = A_1e_1 + u^*$, we have for every $P \in \Pi$, $P(e_2 - e_2') = PA_1Pe_1 - PA_1Pe_1'$. Hence, if $\|A_{1P}\|^* \leqq \kappa$ for all $P \in \Pi$, we get

$$\|P(e_2 - e_2')\| \leqq \|A_{1P}\|^*\|P(e_1 - e_1')\| \leqq \kappa\|e_1 - e_1'\| \leqq \kappa\lambda\|u_1 - u_1'\|.$$

Then axiom (vi) concludes the proof.

Using Lemmas 10 and 6, 7, we can prove analogues of Theorems 6, 7 which concern feedback systems over the extended space H_e, but we omit the details.

On the other hand, let us emphasize the following aspect of results established in the second part as seen from the viewpoint of extended spaces. Assume that the operators $A_1, A_2\colon H_e \to H_e$ have the properties $A_1H \subset H$, $A_2H \subset H$ and $A_10 = A_20 = 0$. Then by a theorem of Willems [18] the F.S. $[A_1, A_2]$ over H_e is well-posed, precisely if the F.S. $[A_1^0, A_2^0]$ over H is normal and Lipschitz continuous, where A_1^0 and A_2^0 is the restriction of A_1 and A_2 to H, respectively. Thus, if a given F.S. $[A_1, A_2]$ satisfies the above requirements and $[A_1^0, A_2^0]$ meets conditions of any Theorem 5–7 then $(A_1, A_2]$ is well-posed.

4. Linearization. Consider a F.S. $[A_1, A_2]$ over a real Hilbert space H, whose operators A_1, A_2 satisfy the condition $A_10 = A_20 = 0$. As it is apparent from (13), to describe the behavior of $[A_1, A_2]$ amounts to establishing the inverse M_a^{-1} for each $a \in H$. If A_1 and A_2 are not linear, this is usually a difficult task. If we wish to avoid inverting nonlinear operators, and are satisfied with an approximate solution we can linearize $[A_1, A_2]$ in a vicinity of the origin $(0, 0) \in H^2$. More specifically, we can choose an appropriate approximating F.S. $[{}^0A_1, {}^0A_2]$ with ${}^0A_1, {}^0A_2$ being linear and expect that the errors ${}^0e_1$, ${}^0e_2$ of $[{}^0A_1, {}^0A_2]$, corresponding to a pair of inputs $(u_1, u_2) \in H^2$ with $\|u_1\|, \|u_2\| \leqq r$, will be sufficiently close to the errors e_1, e_2 of $[A_1, A_2]$ that correspond to the same pair (u_1, u_2).

It turns out that if the operators $A_1+\alpha_1 I$ and $A_2+\alpha_2 I$ are monotone for some $\alpha_1 \geqq 0, \alpha_2 < 0$ (or $\alpha_1 < 0, \alpha_2 \geqq 0$), and if ${}^0A_i$ is sufficiently close to A_i, $i = 1, 2$, on a certain ball centered at the origin, then we can derive inequalities of the form

$$\|e_1 - {}^0e_1\| \leqq K_{11}\|u_1\| + K_{12}\|u_2\|,$$

$$\|e_2 = {}^0e_2\| \leqq K_{21}\|u_1\| + K_{22}\|u_2\|.$$

These estimates are furnished by Theorems 9–11 below, which also supply conditions for Lipschitz continuity of $[A_1, A_2]$ and $[{}^0A_1, {}^0A_2]$.

Definition 3. Let $[A_1, A_2]$ be a normal F.S. over a real Hilbert space H. Then $[A_1, A_2]$ will be called

(a) *Lipschitz continuous in the first input*, if there exist numbers $\lambda_{11}, \lambda_{21} > 0$ such that

$$\|e_1 - e_1'\| \leqq \lambda_{11}\|u_1 - u_1'\|, \qquad \|e_2 - e_2'\| \leqq \lambda_{21}\|u_1 - u_1'\|$$

whenever $(u_1, u^*) \to (e_1, e_2)$ and $(u_1', u^*) \to (e_1', e_2')$;

(b) *Lipschitz continuous in both inputs*, if there exist numbers $\lambda_{11}, \lambda_{12}, \lambda_{21}, \lambda_{22} > 0$ such that

$$\|e_1 - e_1'\| \leqq \lambda_{11}\|u_1 - u_1'\| + \lambda_{12}\|u_2 - u_2'\|,$$

$$\|e_2 - e_2'\| \leqq \lambda_{21}\|u_1 - u_1'\| + \lambda_{22}\|u_2 - u_2'\|$$

whenever $(u_1, u_2) \to (e_1, e_2)$ and $(u_1', u_2') \to (e_1', e_2')$.

Now, we can state the first approximation result. If $r > 0$, we denote $B_r = \{x : x \in H, \|x\| \leqq r\}$.

Theorem 9. *Let $A_1, A_2 \in Lip$, and let $r > 0$. Assume that*

(a) *there exists a linear operator ${}^0A_1$: $H \to H$ and number a_1 with $0 \leqq a_1 < u_{A_1}$ such that*

$$\|(A_1 - {}^0A_1)x\| \leqq a_1\|x\| \tag{85}$$

for all $x \in B_{\nu r}$, where

$$\nu = \mu_{A_1}^{-1}(\mu_{A_2} + \mu_{A_1}\|A_1\|^{*-2})^{-1}, \tag{86}$$

(b) *there exists a linear operator ${}^0A_2$: $H \to H$ and a number $a_2 \geqq 0$ such that*

$$\|(A_2 - {}^0A_2)x\| \leqq a_2\|x\| \tag{87}$$

for all $x \in B_{(1+\|A_1\|^*\nu)r}$,

(c) $\mu_{A_2} - a_2 + (\mu_{A_1} - a_1)(a_1 + \|A_1\|^*)^{-2} > 0$.

Then

(i) *both F.S.'s $[A_1, A_2]$ and $[{}^0A_1, {}^0A_2]$ are normal and Lipschitz continuous in the first input,*

(ii) *if $u_1, u_2 \in B_r$ and $(u_1, u_2) \to (e_1, e_2)$ for $[A_1, A_2]$, $(u_1, u_2) \to ({}^oe_1, {}^oe_2)$ for $[{}^0A_1, {}^0A_2]$, we have*

$$\|e_1 - {}^0e_1\| \leqq K_{11}\|u_1\| + K_{12}\|u_2\|, \tag{88}$$

$$\|e_2 - {}^0e_2\| \leqq K_{21}\|u_1\| + K_{22}\|u_2\|, \tag{89}$$

where

$$
\begin{aligned}
&K_{11} = \kappa\nu(a_2\|A_1\|^* + a_1\|{}^0A_2\|), \qquad K_{12} = \kappa a_2,\\
&K_{21} = a_1\nu + \|{}^0A_1\|\kappa\nu(a_2\|A_1\|^* + a_1\|{}^0A_2\|),\\
&K_{22} = \kappa\|{}^0A_1\|a_2,\\
&\kappa = (\mu_{A_1} - a_1)^{-1}(\mu_{A_2} - a_2 + (\mu_{A_1} - a_1)\|{}^0A_1\|^{-2})^{-1}.
\end{aligned}
\tag{90}
$$

Proof. First (85) implies that $A_1 0 = 0$. Also, if $x \in B_{\nu r}$, we have

$$\|{}^0A_1x\| \leqq \|{}^0A_1x - A_1x\| + \|A_1x\| \leqq a_1\|x\| + \|A_1\|^*\|x\|.$$

Since ${}^0A_1$ is linear, this inequality holds for all $x \in H$, and consequently, ${}^0A_1$ is bounded with $\|{}^0A_1\| \leqq \|a_1 + \|A_1\|^*$.

On the other hand, if $x \in B_{\nu r}$, we have

$$\langle {}^0A_1x, x\rangle = \langle A_1x, x\rangle - \langle (A_1 - {}^0A_1)x, x\rangle, \tag{91}$$

and, by (85)

$$\langle (A_1 - {}^0A_1)x, x\rangle \leqq a_1\|x\|^2.$$

Hence, by (91) and $\langle A_1x, x\rangle \geqq \mu_{A_1}\|x\|^2$,

$$\langle {}^0A_1x, x\rangle \geqq (\mu_{A_1} - a_1)\|x\|^2. \tag{92}$$

Due to linearity of ${}^0A_1$, (92) holds for all $x \in H$. Hence, $\mu_{{}^0A_1} \geqq \mu_{A_1} - a_1 > 0$ by our hypothesis.

Moreover, since ${}^0A_1$ is bounded, Lemma 3 shows that ${}^0A_1$ is invertible.

Using the same argument and (87) it follows that $A_2 0 = 0$, ${}^0A_2$ is bounded with $\|{}^0A_2\| \leqq a_2 + \|A_2\|^*$, and that $\mu_{{}^0A_2} \geqq \mu_{A_2} - a_2$.

Choose now $z \in H$ and define the operators M_z, ${}^0M_z$: $H \to H$ by

$$M_zx = x + A_2(z + A_1x), \qquad {}^0M_zx = x + {}^0A_2(z + {}^0A_1x). \tag{93}$$

Also, let B_z, ${}^0B_z$: $H \to H$ be defined by

$$B_zx = z + A_1x, \qquad {}^0B_zx = z + {}^0A_1x. \tag{94}$$

It is obvious that

$$M_z = I + A_2B_z, \qquad {}^0M_z = I + {}^0A_2{}^0B_z. \tag{95}$$

Moreover, from (94) it follows that B_z, ${}^0B_z \in \mathrm{Lip}$,

$$\mu_{B_z} = \mu_{A_1}, \qquad \mu_{{}^0B_z} = \mu_{{}^0A_1} \geqq \mu_{A_1} - a_1 > 0 \tag{96}$$

and

$$\|B_z\|^* = \|A_1\|^*, \qquad \|{}^0B_z\|^* = \|{}^0A_1\|. \tag{97}$$

Next, from (96), (97) and (c) it follows that $\mu_{A_2} + \mu_{B_z}\|B_z\|^{*-2} = \mu_{A_2} + \mu_{A_1}\|A_1\|^{*-2} > \mu_{A_2} - a_2 + (\mu_{A_1} - a_1)(a_1 + \|A_1\|^{*-2}) > 0$. Since $\mu_{B_z} = \mu_{A_1} > 0$, Lemma 5 shows that the operator M_z is invertible, $M_z^{-1} \in \mathrm{Lip}$ and

$$\|M_z^{-1}\|^* \leqq \mu_{A_1}^{-1}(\mu_{A_2} + \mu_{A_1}\|A_1\|^{*-2})^{-1} = \nu. \tag{98}$$

Hence, by Corollary 1, $[A_1, A_2]$ is normal. Moreover, (98) and (13) show that $[A_1, A_2]$ is Lipschitz continuous in the first input.

Similarly, we get by (96), (97) and (c),

$$\begin{aligned}\mu_{{}^0A_2}+\mu_{{}^0B_z}\|{}^0B_z\|^{*-2}&=\mu_{{}^0A_2}+\mu_{{}^0A_1}\|{}^0A_1\|^{-2}\\&\geqq \mu_{A_2}-a_2+(\mu_{A_1}-a_1)(a_1+\|A_1\|^*)^{-2}>0,\end{aligned}\tag{99}$$

and because $\mu_{{}^0B_z}=\mu_{{}^0A_1}>0$, it follows by Lemma 5 that ${}^0M_z$ is invertible, ${}^0M_z^{-1}\in \mathrm{Lip}$ and

$$\begin{aligned}\|{}^0M_z^{-1}\|&\leqq \mu^{-1}_{{}^0B_z}(\mu_{{}^0A_2}+\mu_{{}^0B_z}\|{}^0B_z\|^{*-2})^{-1}\\&\leqq(\mu_{A_1}-a_1)^{-1}(\mu_{A_2}-a_2+(\mu_{A_1}-a_1)\|{}^0A_1\|^{-2})^{-1}=\kappa.\end{aligned}\tag{100}$$

Thus, by Corollary 1, $[{}^0A_1, {}^0A_2]$ is normal, and (100), (13) show that $[{}^0A_1, {}^0A_2]$ is Lipschitz continuous in the first input. Hence, our assertion (i) is proven.

To prove (ii), define the linear operator $N\colon H\to H$ by

$$N=I+{}^0A_2{}^0A_1.\tag{101}$$

Since ${}^0A_1, {}^0A_2\in \mathrm{Lip}$ and $\mu_{{}^0A_1}=\mu_{{}^0B_z}\geqq \mu_{A_1}-a_1>0$, inequality (99) shows by Lemma 5 that N is invertible, $N^{-1}\in \mathrm{Lip}$ and

$$\|N^{-1}\|\leqq \kappa.\tag{102}$$

Choose now $x, z\in B_r$, and denote $M_z^{-1}x=w$. Since ${}^0A_1$ and N are invertible, we have by Lemma 8,

$$\begin{aligned}M_z^{-1}x-{}^0M_z^{-1}x&=M_z^{-1}x-N^{-1}(x+{}^0A_1^{-1}z)+{}^0A_1^{-1}z\\&=(M_z^{-1}x-N^{-1}x)+(I-N^{-1}){}^0A_1^{-1}z\\&=-N^{-1}(M_z-N)M_z^{-1}x+N^{-1}(N-I){}^0A_1^{-1}z\\&=-N^{-1}(M_zw-Nw)+N^{-1}\,{}^0A_2{}^0A_1{}^0A_1^{-1}z\\&=-N^{-1}(w+A_2(z+A_1w)-w-{}^0A_2^0A_1w)+N^{-1}\,{}^0A_2z\\&=-N^{-1}\{[A_2(z+A_1w)-{}^0A_2(z+A_1w)]\\&\qquad+[{}^0A_2(z+A_1w)-{}^0A_2{}^0A_1w]\}+N^{-1}\,{}^0A_2z\\&=-N^{-1}(A_2-{}^0A_2)(z+A_1w)-N^{-1}\,{}^0A_2(A_1-{}^0A_1)w.\end{aligned}\tag{103}$$

However, by our assumption,

$$\|w\|\leqq\|M_z^{-1}\|^*_{\cdot}\|x\|\leqq \nu\|x\|\leqq \nu r,\tag{104}$$

i.e., $w\in B_{\nu r}$. Also,

$$\|z+A_1w\|\leqq\|z\|+\|A_1\|^*_{\cdot}\|w\|\leqq r+\|A_1\|^*\nu r,$$

i.e., $z+A_1w\in B_{(1+\|A_1\|^*\nu)r}$. Hence, (103) yields, by (85) and (87),

$$\begin{aligned}\|M_z^{-1}x-{}^0M_z^{-1}x\|&\leqq\|N^{-1}\|a_2(\|z\|+\|A_1\|^*_{\cdot}\|w\|)+\|N^{-1}\|\cdot\|{}^0A_2\|a_1\|w\|\\&=\|N^{-1}\|(a_2\|A_1\|^*+a_1\|{}^0A_2\|)\|w\|+\|N^{-1}\|a_2\|z\|\\&\leqq\kappa\nu(a_2\|A_1\|^*+a_1\|{}^0A_2\|)\|x\|+\kappa a_2\|z\|\\&=K_{11}\|x\|+K_{12}\|z\|.\end{aligned}\tag{105}$$

To conclude the proof, choose $(u_1, u_2)\in H^2$ with $u_1, u_2\in B_r$, and let $(u_1, u_2)\mapsto(e_1, e_2)$ for $[A_1, A_2]$, and $(u_1, u_2)\mapsto({}^0e_1, {}^0e_2)$ for $[{}^0A_1, {}^0A_2]$. Referring to (13) in Corollary 1, we have

$$e_1 - {}^0e_1 = M_{u_2}^{-1}u_1 - {}^0M_{u_2}^{-1}u_1, \tag{106}$$

$$e_2 - {}^0e_2 = A_1M_{u_2}^{-1}u_1 - {}^0A_1{}^0M_{u_2}^{-1}u_1. \tag{107}$$

Using (105), we obtain readily from (106),

$$\|e_1 - {}^0e_1\| \leqq K_{11}\|u_1\| + K_{12}\|u_2\|.$$

Hence, (88) is confirmed.

On the other hand, (107) yields

$$\|e_2 - {}^0e_2\| \leqq \|A_1M_{u_2}^{-1}u_1 - {}^0A_1M_{u_2}^{-1}u_1\| + \|{}^0A_1M_{u_2}^{-1}u_1 - {}^0A_1{}^0M_{u_2}^{-1}u_1\|.$$

Since $\|M_{u_2}^{-1}u_1\| \leqq \nu\|u_1\| \leqq \nu r$, we have by (85) and (105),

$$\begin{aligned}\|e_2 - {}^0e_2\| &\leqq a_1\nu\|u_1\| + \|{}^0A_1\|(K_{11}\|u_1\| + K_{12}\|u_2\|)\\ &= (a_1\nu + \|{}^0A_1\|K_{11})\|u_1\| + \|{}^0A_1\|K_{12}\|u_2\|\\ &= K_{21}\|u_1\| + K_{22}\|u_2\|.\end{aligned}$$

Thus, (89) holds and the proof is complete.

The estimates given in Theorem 9 can be improved, if either the operator A_1 or A_2 is linear. Let us first discuss the case of A_2 being linear.

THEOREM 10. *Let $A_1 \in Lip$ with $u_{A_1} \leqq 0$, let $A_2 \in \mathcal{M}$ with $\mu_{A_2} > 0$ be linear, and let $r > 0$. Assume that*

(a) *there exists a linear operator ${}^0A_1$: $H \to H$ with $\mu_{{}^0A_1} \leqq 0$ and $a_1 \geqq 0$ such that*

$$\|(A_1 - {}^0A_1)x\| \leqq a_1\|x\| \tag{108}$$

for all $x \in B_{\omega(1+\|A_2\|)r}$, where

$$\omega = \|A_2\|(\mu_{A_2} + \mu_{A_1}\|A_2\|^2)^{-1}, \tag{109}$$

(b) $\mu_{A_2} + (\mu_{A_1} - a_1)\|A_2\|^2 > 0$.

Then

(i) *both F.S.'s $[A_1, A_2]$ and $[{}^0A_1, A_2]$ are normal and Lipschitz continuous in both inputs,*

(ii) *if $u_1, u_2 \in B_r$ and $(u_1, u_2) \to (e_1, e_2)$ for $[A_1, A_2]$, $(u_1, u_2) \to ({}^0e_1, {}^0e_2)$ for $[{}^0A_1, A_2]$, we have*

$$\|e_1 - {}^0e_1\| \leqq \lambda\|u_1 - A_2u_2\| \leqq \lambda\|u_1\| + \lambda\|A_2\|\cdot\|u_2\|, \tag{110}$$

$$\begin{aligned}\|e_2 - {}^0e_2\| &\leqq (a_1\omega + \|{}^0A_1\|\lambda)\|u_1 - A_2u_2\|\\ &\leqq (a_1\omega + \|{}^0A_1\|\lambda)(\|u_1\| + \|A_2\|\cdot\|u_2\|),\end{aligned} \tag{111}$$

where

$$\lambda = a_1\|A_2\|^3(\mu_{A_2} + \mu_{A_1}\|A_2\|^2)^{-1}(\mu_{A_2} + \mu_{{}^0A_1}\|A_2\|^2)^{-1}. \tag{112}$$

Proof. Let the operators N, ${}^0N$: $H \to H$ be defined by

$$N = I + A_2A_1, \qquad {}^0N = I + A_2{}^0A_1. \tag{113}$$

As in the proof of Theorem 9 it follows from (108) that $A_10 = 0$, and consequently, $N0 = 0$. Also, it follows that ${}^0A_1$ is bounded with $\|{}^0A_1\| \leqq a_1 + \|A_1\|^*$, and that $\langle {}^0A_1x, x\rangle \geqq (\mu_{A_1} - a_1)\|x\|^2$ for all $x \in H$. Thus, ${}^0A_1 \in \mathcal{M}$ and $0 \geqq \mu_{{}^0A_1} \geqq \mu_{A_1} - a_1$.

Moreover, by our hypothesis (b),

$$\mu_{A_2}+\mu_{A_1}\|A_2\|^2>a_1\|A_2\|^2\geqq 0.$$

Hence, by Lemma 6, N is invertible, $N^{-1}\in\mathrm{Lip}$ and $\|N^{-1}\|^*\leqq\omega$. Since A_2 is linear, we can invoke the Corollary 2 and conclude that $[A_1, A_2]$ is normal. Also, the fact that $A_1, A_2, N^{-1}\in\mathrm{Lip}$ and (14) show immediately that $[A_1, A_2]$ is Lipschitz continuous in both inputs.

Next, since $\mu_{{}^0A_1}\leqq 0$ and $\mu_{A_2}+\mu_{{}^0A_1}\|A_2\|^2\geqq\mu_{A_2}+(\mu_{A_1}-a_1)\|A_2\|^2>0$ by (b), Lemma 6 implies that the operator ${}^0N$ is invertible, ${}^0N^{-1}\in\mathrm{Lip}$ and $\|{}^0N^{-1}\|\leqq\|A_2\|(\mu_{A_2}+\mu_{{}^0A_1}\|A_2\|^2)^{-1}=\kappa$. Hence, again by Corollary 2, $[{}^0A_1, A_2]$ is normal and Lipschitz continuous in both inputs, which proves our claim (i).

Now, let $x\in B_{(1+\|A_2\|)r}$; then

$$\|N^{-1}x\|\leqq\|N^{-1}\|^*\|x\|\leqq\omega\|x\|\leqq\omega(1+\|A_2\|)r, \tag{114}$$

so that $N^{-1}x\in B_{\omega(1+\|A_2\|)r}$. Thus, we have by (108) and linearity of ${}^0N^{-1}$,

$$\begin{aligned}\|N^{-1}x-{}^0N^{-1}x\|&=\|{}^0N^{-1}({}^0N-N)N^{-1}x\|\\&=\|{}^0N^{-1}A_2({}^0A_1-A_1)N^{-1}x\|\\&\leqq\|{}^0N^{-1}\|\cdot\|A_2\|\cdot a_1\|N^{-1}\|^*\|x\|\\&\leqq\kappa\omega\|A_2\|a_1\|x\|=\lambda\|x\|.\end{aligned}\tag{115}$$

To conclude the proof, choose $u_1, u_2\in B_r$ and let $(u_1, u_2)\mapsto(e_1, e_2)$ for $[A_1, A_2]$ and $(u_1, u_2)\mapsto({}^0e_1, {}^0e_2)$ for $[{}^0\mathrm{A}_1, \mathrm{A}_2]$. Then we have

$$\|u_1-A_2u_2\|\leqq\|u_1\|+\|A_2\|\cdot\|u_2\|\leqq(1+\|A_2\|)r, \tag{116}$$

i.e., $w=u_1-A_2u_2\in B_{(1+\|A_2\|)r}$. Also, by (114), $N^{-1}w\in B_{\omega(1+\|A_2\|)r}$. Thus, invoking (14) in Corollary 2, we get by (115),

$$\|e_1-{}^0e_1\|=\|(N^{-1}-{}^0N^{-1})w\|\leqq\lambda\|w\|\leqq\lambda\|u_1\|+\lambda\|A_2\|\cdot\|u_2\|.$$

Moreover, by (108) and (115),

$$\begin{aligned}\|e_2-{}^0e_2\|&=\|A_1N^{-1}w-{}^0A_1{}^0N^{-1}w\|\\&\leqq\|A_1N^{-1}w={}^0A_1N^{-1}w\|+\|{}^0A_1N^{-1}w-{}^0A_1{}^0N^{-1}w\|\\&\leqq a_1\|N^{-1}w\|+\|{}^0A_1\|\lambda\|w\|\\&\leqq(a_1\omega+\|{}^0A_1\|\lambda)\|w\|\\&\leqq(a_1\omega+\|{}^0A_1\|\lambda)(\|u_1\|+\|A_2\|\cdot\|u_2\|).\end{aligned}$$

This, however, proves (110) and (111).

Remark 5. The assumptions $\mu_{A_1}\leqq 0$, $\mu_{{}^0A_1}\leqq 0$ made in Theorem 10 can sometimes be inconvenient. Fortunately, they can be dropped provided we are satisfied with estimates that are worse than (110) and (111). The proof of this uses an assertion like Lemma 6 which does not impose the condition $\mu_B\leqq 0$, and which can be proved via a decomposition $I+AB=A(A^{-1}+B)$. The details are omitted.

If the operator A_1 is linear, we have the following result:

THEOREM 11. *Let $A_1\in\mathcal{M}$ with $\mu_{A_1}>0$ be linear, let $A_2\in$ Lip with $\mu_{A_2}\leqq 0$, and let $r>0$. Assume that*

(a) *there exists a linear operator ${}^0A_2$: $H\to H$ with $\mu_{{}^0A_2}\leqq 0$ and $a_2\geqq 0$ such that*

$$\|(A_2-{}^0A_2)x\|\leqq a_2\|x\| \tag{117}$$

for all $x \in B_{\rho\|A_1\|(1+\mu_{A_1}^{1})r}$, *where*

$$\rho = \|A_1\|(\mu_{A_1} + \mu_{A_2}\|A_1\|^2)^{-1}, \tag{118}$$

(b) $\mu_{A_1} + (\mu_{A_2} - a_2)\|A_1\|^2 > 0.$

Then

(i) *both F.S.'s* $[A_1, A_2]$ *and* $[A_1, {}^0A_2]$ *are normal and Lipschitz continuous in both inputs,*

(ii) *if* $u_1, u_2 \in B_r$ *and* $(u_1, u_2) \mapsto (e_1, e_2)$ *for* $[A_1, A_2]$, $(u_1, u_2) \mapsto ({}^0e_1, {}^0e_2)$ *for* $[A_1, {}^0A_2]$, *we have*

$$\|e_1 - {}^0e_1\| \leqq \lambda\,\|u_1 + A_1^{-1}u_2\| \leqq \lambda\,\|u_1\| + \lambda\mu_{A_1}^{-1}\,\|u_2\|, \tag{119}$$

$$\|e_2 - {}^0e_2\| \leqq \lambda\,\|A_1\| \cdot \|u_1 + A_1^{-1}u_2\| \leqq \lambda\,\|A_1\| \cdot \|u_1\| + \lambda\,\|A_1\|\mu_{A_1}^{-1}\,\|u_2\|, \tag{120}$$

where

$$\lambda = a_2\|A_1\|^3(\mu_{A_1} + \mu_{A_2}\|A_1\|^2)^{-1}(\mu_{A_1} + \mu_{{}^0A_2}\|A_1\|^2)^{-1}.$$

Proof. Using (117) we conclude as before that $A_20 = 0$, A_2 is bounded with $\|{}^0A_2\| \leqq a_2 + \|A_2\|^*$, and $0 \geqq \mu_{{}^0A_2} \geqq \mu_{A_2} - a_2$. Moreover, due to Lemma 3, A_1 is bounded, invertible and $\|A_1^{-1}\|\mu_{A_1}^{-1}$.

Consider now the operators $N = I + A_2A_1$ and ${}^0N = I + {}^0A_2A_1$. Since $\mu_{A_1} > 0$, $\mu_{A_2} \leqq 0$ and $\mu_{A_1} + \mu_{A_2}\|A_1\|^2 \geqq \mu_{A_1} + (\mu_{A_2} - a_2)\|A_1\|^2 > 0$ by (b), it follows by Lemma 7 that N is invertible, $N^{-1} \in \mathrm{Lip}$ and

$$\|N^{-1}\|^* \leqq \|A_1\|(\mu_{A_1} + \mu_{A_2}\|A_1\|^2)^{-1} = \rho. \tag{121}$$

Similarly, since $\mu_{A_1} + \mu_{{}^0A_2}\|A_1\|^2 \geqq \mu_{A_1} + (\mu_{A_2} - a_2)\|A_1\|^2 > 0$ and $\mu_{{}^0A_2} \leqq 0$, Lemma 7 shows that ${}^0N$ is also invertible, ${}^0N^{-1} \in \mathrm{Lip}$ and

$$\|{}^0N^{-1}\| \leqq \|A_1\|(\mu_{A_1} + \mu_{{}^0A_2}\|A_1\|^2)^{-1} = \eta. \tag{122}$$

Next, choose a $z \in H$ and define the operators M_z, ${}^0M_z$: $H \to H$ by

$$M_zx = x + A_2(z + A_1x), \qquad {}^0M_zx = x + {}^0A_2(z + A_1x). \tag{123}$$

Since N, ${}^0N$ are invertible, and A_1 is invertible and linear, Lemma 8 shows that M_z and ${}^0M_z$ are invertible. Also, by (58),

$$\begin{aligned} M_z^{-1}x &= N^{-1}(x + A_1^{-1}z) - A_1^{-1}z, \\ {}^0M_z^{-1}x &= {}^0N^{-1}(x + A_1^{-1}z) - A_1^{-1}z. \end{aligned} \tag{124}$$

However, from (124) it follows readily that M_z^{-1}, ${}^0M_z^{-1} \in \mathrm{Lip}$ and

$$\|M_z^{-1}\|^* = \|N^{-1}\|^* \leqq \rho, \qquad \|{}^0M_z^{-1}\|^* = \|{}^0N^{-1}\| \leqq \eta. \tag{125}$$

Thus, invoking Corollary 1, we see that $[A_1, A_2]$ and $[A_1, {}^0A_2]$ are normal. Moreover, if, for $[A_1, A_2]$, $(u_1, u_2) \mapsto (e_1, e_2)$ and $(u_1', u_2') \mapsto (e_1', e_2')$, it follows by (13) and (124) that

$$\begin{aligned} \|e_1 - e_1'\| &= \|M_{u_2}^{-1}u_1 - M_{u_2'}^{-1}u_1'\| \\ &= \|N^{-1}(u_1 + A_1^{-1}u_2) - A_1^{-1}u_2 - N^{-1}(u_1' + A_1^{-1}u_2') + A_1^{-1}u_2'\| \\ &\leqq \|N^{-1}\|^*\|u_1 - u_1' + A_1^{-1}(u_2 - u_2')\| + \|A_1^{-1}\|\cdot\|u_2 - u_2'\| \\ &\leqq \rho\|u_1 - u_1'\| + \|A_1^{-1}\|(1 + \rho)\|u_2 - u_2'\|. \end{aligned} \tag{126}$$

Similarly, by (13) and (124),

$$\begin{aligned}e_2-e_2' &= u_2-u_2'+A_1(M_{u_2}^{-1}u_1-M_{u_2}^{-1}u_1')\\ &= u_2-u_2'+A_1\{N^{-1}(u_1+A_1^{-1}u_2)-A_1^{-1}u_2-N^{-1}(u_1'+A_1^{-1}u_2')+A_1^{-1}u_2'\}\\ &= A_1\{N^{-1}(u_1+A_1^{-1}u_2)-N^{-1}(u_1'+A_1^{-1}u_2')\}.\end{aligned}$$

Hence, by (121),

$$\begin{aligned}\|u_2-u_2'\| &\leqq \|A_1\|\,\|N^{-1}\|^*\|u_1-u_1'+A_1^{-1}(u_2-u_2')\|\\ &\leqq \rho\|A_1\|\cdot\|u_1-u_1'\|+\rho\|A_1\|\cdot\|A_1^{-1}\|\cdot\|u_2-u_2'\|.\end{aligned}\tag{127}$$

Thus, due to (126), (127), the F.S. $[A_1, A_2]$ is Lipschitz continuous in both inputs.

The same argument applies to $[A_1, {}^0A_2]$; hence, our claim (i) holds.

Finally, let $u_1, u_2 \in B_r$ and let $(u_1, u_2)\mapsto(e_1, e_2)$ for $[A_1, A_2]$, and $(u_1, u_2)\mapsto({}^0e_1, {}^0e_2)$ for $[A_1, {}^0A_2]$. Denote $w=u_1+A_1^{-1}u_2$; then

$$\|A_1N^{-1}w\|\leqq\|A_1\|\rho(\|u_1\|+u_{A_1}^{-1}\|u_2\|)\leqq\|A_1\|\rho(1+u_{A_1}^{-1})r,\tag{128}$$

i.e., $A_1N^{-1}w\in B_{\|A_1\|\rho(1+u_{A_1}^{-1})r}$. Thus, we have by (13), (124) and (117),

$$\begin{aligned}\|e_1-{}^0e_1\| &= \|M_{u_1}^{-1}u_2-{}^0M_{u_1}^{-1}u_2\|\\ &= \|(N^{-1}-{}^0N^{-1})w\|\\ &= \|{}^0N^{-1}({}^0N-N)N^{-1}w\|\\ &= \|{}^0N^{-1}({}^0A_2A_1-A_2A_1)N^{-1}w\|\\ &= \|{}^0N^{-1}({}^0A_2-A_2)A_1N^{-1}w\|\\ &\leqq \|{}^0N^{-1}\|a_2\|A_1\|\cdot\|N^{-1}\|^*\|w\|\\ &\leqq \eta a_2\|A_1\|\rho\|w\|\\ &= \lambda\|w\|\leqq\lambda(\|u_1\|+u_{A_1}^{-1}\|u_2\|).\end{aligned}\tag{129}$$

Hence, (119) is confirmed.

Moreover, by (13) and (129),

$$\begin{aligned}\|e_2-{}^0e_2\| &= \|A_1M_{u_2}^{-1}u_1-A_1{}^0M_{u_2}^{-1}u_1\|\\ &\leqq \|A_1\|\lambda\|w\|\\ &\leqq \|A_1\|\lambda(\|u_1\|+\mu_{A_1}^{-1}\|u_2\|).\end{aligned}\tag{130}$$

Hence, (120) holds and the proof is complete.

Remark 6. A similar comment like Remark 5 applies to Theorem 11. Note also that if the assumptions $\mu_{A_2}\leqq 0$, $\mu_{{}^0A_2}\leqq 0$ are dropped, the corresponding claim follows from Theorem 9.

Example 2. Let A_1, A_2 be the same operators as in Example 1. Moreover, assume that there exists a constant $n\times n$ matrix F with

$$a_0=\inf_{\substack{|\xi|=1\\ \xi\in R^n}}\xi^TF\xi\leqq 0$$

such that

$$|\phi(\xi)-F\xi|\leqq a|\xi|\tag{131}$$

for all $\xi\in R^n$.

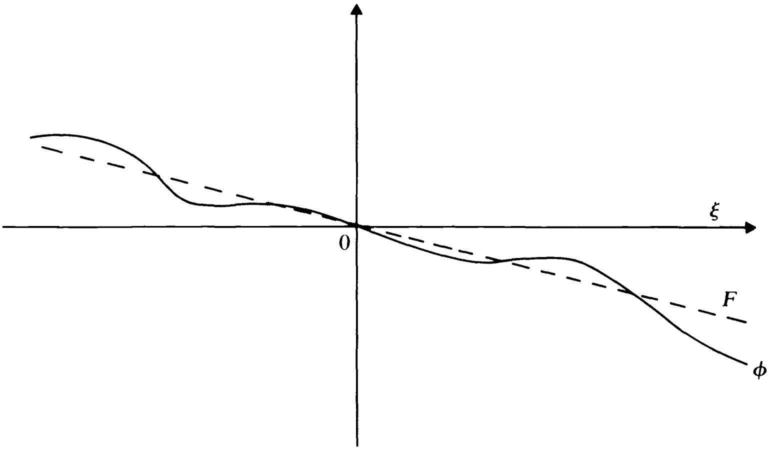

FIG. 3

A typical example of such ϕ and F in one dimension is given in Fig. 3.

Define the operator ${}^0A_2$ on H by

$$({}^0A_2x)(t) = Fx(t), \qquad t \geqq 0. \tag{132}$$

We are going to show that if (70) and the inequality

$$d + k + (a_2 - a)(|D| + \kappa)^2 > 0 \tag{133}$$

hold, then the F.S.'s $[A_1, A_2]$ and $[A_1, {}^0A_2]$ are normal, Lipschitz continuous in both inputs and we have

$$\begin{aligned} \|e_1 - {}^0e_1\| &\leqq \lambda^*\|u_1\| + \lambda^*(d+k)^{-1}\|u_2\|, \\ \|e_2 - {}^0e_2\| &\leqq \lambda^*(|D| + \kappa)(\|u_1\| + (d+k)^{-1}\|u_2\|) \end{aligned} \tag{134}$$

whenever $(u_1, u_2) \to (e_1, e_2)$ for $[A_1, A_2]$ and $(u_1, u_2) \to ({}^0e_1, {}^0e_2)$ for $[A_1, {}^0A_2]$, where

$$\lambda^* = a(|D| + \kappa)^3[d + k + a_2(|D| + \kappa)^2]^{-1}[d + k + a_0(|D| + \kappa)^2]^{-1}. \tag{135}$$

Indeed, it is easy to see that $\mu_{{}^0A_2} = a_0 \leqq 0$, and that $\|(A_2 - {}^0A_2)x\| \leqq a\|x\|$ for all $x \in H$ by virtue of (131). Moreover, using the facts established in Example 1 and (133), we confirm that $\mu_{A_1} + (\mu_{A_2} - a)\|A_1\|^2 > 0$. Hence, all assumptions of Theorem 11 are met with any $r > 0$, and consequently, (119) and (120) hold. Introducing our bounds for $\|A_1\|$, μ_{A_1}, etc. into (119) and (120), we get readily (134) and (135).

REFERENCES

[1] P. E. CAINES, *Causality, stability and inverse systems*, Internat. J. Systems Sci., (1973), pp. 825–832.

[2] F. M. CALLIER AND C. A. DESOER, *L^p-stability $(1 \leqq p \leqq \infty)$ of multivariable time-varying feedback systems that are open-loop unstable*, Proc. of Colloquium on Point Mappings, LASS, Toulouse, France.

[3] ———, *L_p-stability $(1 \leqq p \leqq \infty)$ of multivariable nonlinear time-varying feedback systems that are open-loop unstable*, Internat. J. Control, 20 (1974), pp. 65–72.

[4] M. J. DAMBORG, *Suitability of the basic nonlinear operator feedback system*, Tech. Rpt. 37, Syst. Engrg. Lab., Univ. of Michigan, 1967.

[5] M. J. DAMBORG AND A. NAYLOR, *Fundamental structure of input-output stability for feedback systems*, IEEE Trans., SSC-6, (1970) pp. 92–96.

[6] R. M. DESANTIS AND W. A. PORTER, *On the analysis of feedback systems with a multipower open loop chain*, Systems Engrg. Lab., Tech. Rpt. No. 75, Univ. of Michigan, 1973.

[7] ———, *On the analysis of feedback systems with a polynomic plant*, Internat. J. Control, 21 (1975), pp. 159–175.

[8] C. A. DESOER AND M. VIDYASAGAR, *Feedback Systems: Input-Output Properties*, Academic Press, New York, 1975.

[9] P. L. FALB, M. I. FREEDMAN AND G. ZAMES, *Input-output stability: A general point of view*, Rpt. PM-54, NASA Elec. Res. Center, 1968.

[10] A. FEINTUCH, *Causal C_0 operators and feedback stability*, Math. Systems Theory, to appear.

[11] M. I. FREEDMAN, P. L. FALB AND G. ZAMES, *A Hilbert space stability theory over locally compact Abelian groups*, this Journal, 7 (1969), pp. 479–493.

[12] R. T. ROCKAFELLAR, *On the maximality of sums of nonlinear monotone operators*, Trans. Amer. Math. Soc., May, 1970, pp. 75–88.

[13] R. SAEKS, *Causality in Hilbert space*, SIAM Rev., 12 (1970), pp. 357–383.

[14] ———, *On the encirclement condition and its generalization*, IEEE Trans. Circuits and Systems, CAS-22, (1975), pp. 780–785.

[15] R. SAEKS AND R. A. DECARLO, *Stability and Homotopy*, Alternatives for Linear Multivariable Control, NEC, Chicago, 1978, pp. 247–252.

[16] I. W. SANDBERG, *On the L_2-boundedness of solutions of nonlinear functional equations*, Bell systems Tech. J., 43 (1968), pp. 1601–1608.

[17] ———, *Some results on the theory of physical systems governed by nonlinear functional equations*. Ibid., 44 (1965), p. 871.

[18] J. C. WILLEMS, *Stability, instability, invertibility and causality*, this Journal, 7 (1969), pp. 645–671.

[19] G. ZAMES, *Functional analysis applied to nonlinear feedback systems*, IEEE Trans. Comm. Tech., CT-10 (1963), pp. 392–404.

[20] ———, *On input-output stability of time-varying nonlinear feedback systems*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 228–238 and 465–476.

[21] ———, *On the stability of nonlinear, time-varying systems*, Proc. Natl. Electronics Conf., vol. 20, 1964, pp. 725–730.

[22] ———, *Realizability conditions for nonlinear feedback systems*, IEEE Trans. Comm. Tech. CT-11 (1964), p. 186.

[23] ———, *Realizability of nonlinear filters and feedback systems*, Quarterly Progress Rpt., No. 56, Res. Lab. of Electronics, Mass. Inst. of Tech., Cambridge, 1960, pp. 137–143.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 3, May 1979

ON THE STOCHASTIC REALIZATION PROBLEM*

ANDERS LINDQUIST† AND GIORGIO PICCI‡

Abstract. Given a mean square continuous stochastic vector process y with stationary increments and a rational spectral density Φ such that $\Phi(\infty)$ is finite and nonsingular, consider the problem of finding all minimal (wide sense) Markov representations (*stochastic realizations*) of y. All such realizations are characterized and classified with respect to deterministic as well as probabilistic properties. It is shown that only certain realizations (*internal stochastic realizations*) can be determined from the given output process y. All others (*external stochastic realizations*) require that the probability space be extended with an exogeneous random component. A complete characterization of the sets of internal and external stochastic realizations is provided. It is shown that the state process of any internal stochastic realization can be expressed in terms of two steady-state Kalman–Bucy filters, one evolving forward in time over the infinite past and one backward over the infinite future. An algorithm is presented which generates families of external realizations defined on the same probability space and totally ordered with respect to state covariances.

1. Introduction. One of the most common models of random phenomena in control theory is provided by the linear stochastic system

$$dx = Ax\,dx + B\,dw, \tag{1.1a}$$

$$dz = Cx\,dt + D\,dw, \tag{1.1b}$$

where A, B, C and D are constant matrices of dimensions $n \times n$, $n \times k$, $m \times n$ and $m \times k$ respectively, and w is a k-dimensional mean-square continuous stochastic process with zero mean, stationary orthogonal increments, and $w(0) = 0$. Here we shall assume that w is defined on the whole real line R, that is

$$E\{w(t)\} = 0 \quad \text{for all } t \in R, \qquad E\{w(t)w(s)'\} = \tfrac{1}{2}\{|t| + |s| - |t-s|\}I \tag{1.2}$$

[35; p. 51], where $E\{\cdot\}$ denotes mathematical expectation and prime (′) transposition. (All vectors without prime are column vectors.) For later reference, let $\mathscr{W}_k$ denote the class of all such orthogonal increment processes, the index referring to the dimension; more generally we shall say that the process is of class $\mathscr{W}$. Moreover, we assume that A is a stability matrix, i.e. all the eigenvalues of A are situated in the left complex half-plane; we shall write $\operatorname{Re}\{\lambda(A)\} < 0$ for short. This assumption will insure that (1.1a) has the unique solution

$$x(t) = \int_{-\infty}^{t} e^{A(t-\tau)}B\,dw(\tau) \tag{1.3}$$

on the real line, where the integral is defined in quadratic mean. This is an n-dimensional vector process. If, in addition, we assume that $z(0) = 0$, the m-dimensional process z can be determined uniquely by integrating (1.1b). We shall call x the *state process*, w the *input process* and z the *output process*. Clearly the state process x is (wide sense) stationary, i.e. the *state covariance matrix*

$$P = E\{x(t)x(t)'\} \tag{1.4}$$

* Received by the editors June 14, 1977 and in revised form March 15, 1978.

† Department of Mathematics, University of Kentucky, Lexington, Kentucky 40506, U.S.A. This work was supported by the National Science Foundation under Grant MPS 75-07028.

‡ Laboratorio per Ricerche di Dinamica dei Sistemi e di Elettronica Biomedica, Consiglio Nazionale delle Ricerche, Casella Postale 1075, 35100 Padova, Italy. This work was supported partially by the National Science Foundation under Grant MPS 75-07028 and partially by the Air Force Office of Scientific Research under Grant AFOSR-78-3519.

does not depend on t, and it satisfies the Lyapunov equation

$$AP+PA'+BB'=0. \tag{1.5}$$

(See e.g. [35].) The output process z has stationary increments.

Each $w \in \mathcal{W}_k$ has a unique spectral representation

$$w(t)=\int_{-\infty}^{\infty} \frac{e^{i\omega t}-1}{i\omega}\, d\hat{w}(\omega) \tag{1.6}$$

[12; p. 205], where $d\hat{w}$ is an orthogonal stochastic measure such that $E\{d\hat{w}(\omega)\, d\hat{w}(\omega)\dagger\}=I\, d\omega$. (Here † denotes the complex conjugation and transposition.) Then (1.3) may be written

$$x(t)=\int_{-\infty}^{\infty} e^{i\omega t}(i\omega I-A)^{-1}B\, d\hat{w}(\omega). \tag{1.7a}$$

(Indeed, making the substitution $(sI-A)^{-1}=(1/s)[I+A(sI-A)^{-1}]$, (1.7a) is seen to satisfy (1.1a.) Inserting (1.7a) into (1.1b) and integrating yields

$$z(t)=\int_{-\infty}^{\infty} \frac{e^{i\omega t}-1}{i\omega} W(i\omega)\, d\hat{w}(\omega), \tag{1.7b}$$

where

$$W(s)=C(sI-A)^{-1}B+D. \tag{1.8}$$

We shall call W the *transfer function* of (1.1). Relation (1.7b) is a spectral representation of z; $d\hat{z}(\omega):=W(i\omega)\, d\hat{\omega}(\omega)$ being an orthogonal stochastic measure such that

$$E\{d\hat{z}(\omega)\, d\hat{z}(\omega)\dagger\}=\Phi(i\omega)\, d\omega, \tag{1.9}$$

where Φ is the *spectral density* given by

$$\Phi(s)=W(s)W(-s)'. \tag{1.10}$$

This is an $m \times m$-matrix of rational functions such that (i) each element of Φ is analytic on the imaginary axis, (ii) Φ is parahermitian, i.e. $\Phi(-s)=\Phi(s)'$, (iii) $\Phi(i\omega)$ is nonnegative definite Hermitian for all real ω, and (iv) $\Phi(\infty)<\infty$. Such a Φ is called a *spectral function* [3], [4].

In this paper we consider the following inverse problem. Let $\{y(t);\, t \in R\}$ be a given mean-square continuous and purely nondeterministic m-dimensional stochastic process with zero mean, stationary increments and $y(0)=0$. Then there is a spectral representation

$$y(t)=\int_{-\infty}^{\infty} \frac{e^{i\omega t}-1}{i\omega}\, d\hat{y}(\omega) \tag{1.11}$$

[12; p. 205], where $d\hat{y}$ is an orthogonal stochastic measure such that [9]

$$E\{d\hat{y}(\omega)\, d\hat{y}(\omega)\dagger\}=\Phi(i\omega)\, d\omega. \tag{1.12}$$

Here Φ is an $m \times m$-matrix of real rational functions satisfying conditions (i)–(iv) above. Setting $R:=\Phi(\infty)$, we also assume that (v) R^{-1} exists and that (vi) $\Phi(i\omega)$ is positive definite for all real ω. The problem is to find representations (1.1) such that the output process z is equivalent to the given process y in some sense to be specified below. Such a representation will be called a *stochastic realization.*

More precisely, the system (1.1) will be called a *wide sense stochastic realization* of y if z has the same spectral density Φ as y and a *proper stochastic realization* if, for each

$t \in (-\infty, \infty)$, $z(t) = y(t)$, a.s. (In the sequel we shall leave out the "a.s.", hence regarding such equivalent processes as equal.) Clearly each proper stochastic realization is also a wide sense stochastic realization, but the converse is not true.

The stochastic realization problem is related to the *spectral factorization problem*: Given a rational spectral function Φ, find all matrices $W(s)$ of real rational functions with all its poles in Re $(s) < 0$ and satisfying (1.10). Such a function will be called a *stable spectral factor.* Let $\delta\{\cdot\}$ denote McMillan degree [8]. Then $\delta\{W\} \geqq \frac{1}{2}\delta\{\Phi\}$; if there is equality we shall say that W is *minimal.* We have seen that the transfer function (1.8) of any wide sense stochastic realization of y is a stable spectral factor of the spectral density of y. Conversely any such spectral factor W is the transfer function of an equivalence class of wide sense stochastic realizations. In fact, for *any* orthogonal stochastic measure $d\hat{w}$ such that $E\{d\hat{w}(\omega)\ d\hat{w}(\omega)^\dagger\} = I\, d\omega$, the process

$$z(t) = \int_{-\infty}^{\infty} \frac{e^{i\omega t} - 1}{i\omega} W(i\omega)\, d\hat{w}(\omega) \tag{1.13}$$

has the same spectral density as y. Since W is a real rational matrix function analytic in Re $(s) \geqq 0$, there is a quadruplet $[A, B, C, D]$ of matrices such that (1.8) holds [8], with A a stability matrix. Now let x be defined by (1.7a) and w by (1.6). Then w is of class $\mathscr{W}$ and (x, z) satisfy (1.1) as asserted. Note that $[A, B, C, D]$ defines one wide sense stochastic realization for *each* $w \in \mathscr{W}_k$. Since these realizations are equivalent up to second-order properties of z, in the sequel we shall say that $[A, B, C, D]$ is a wide sense stochastic realization, thereby referring to the whole equivalence class. To avoid trivialities we shall assume that the representation (1.8) is chosen so that the dimension of the matrix A equals $\delta(W)$, i.e. we shall only consider quadruplets $[A, B, C, D]$ for which (A, B) is controllable and (A, C) is observable [8]. We shall call a stochastic realization *minimal* if it corresponds to a minimal spectral factor. Hence, the minimal stochastic realizations are precisely those representations (1.1) which have a state process of smallest possible dimension, i.e. $n = \frac{1}{2}\delta(\Phi)$. In this paper we shall restrict our attention to such realizations, the basic problem being to find all of them.

Determining all wide sense minimal stochastic realizations $[A, B, C, D]$ is a deterministic problem which has been studied extensively by, among others, B. D. O. Anderson [5], Faurre [11] and J. C. Willems [32], the first of whom has named it the *inverse problem of covariance generation.* To facilitate its solution we note that the spectral density of y can be written

$$\Phi(s) = Z(s) + Z(-s)', \tag{1.14}$$

where Z is positive real[1] and rational, and $\delta(Z) = n$ [3], [4], [11], [32]. Let

$$Z(s) = H(sI - F)^{-1}G + \tfrac{1}{2}R \tag{1.15}$$

be a minimal realization [8] of Z, i.e. F, G and H are constant matrices of dimensions $n \times n$, $n \times m$ and $m \times n$ respectively. Hence F is a stability matrix, (F, G) is controllable and (H, F) is observable [8]. There are computational procedures for determining (F, G, H, R) from Φ [8], [13], [31], [38], so in the sequel we shall assume that such a quadruplet is given.

It can be shown [5] that all wide sense minimal stochastic realizations are given by

$$[A, B, C, D] = [TFT^{-1}, T(B_1, B_2)S, HT^{-1}, (R^{1/2}, 0)S] \tag{1.16}$$

[1] A real rational function Z without poles on the imaginary axis is said to be *positive real* if it has no poles in Re $[s] > 0$ and $Z(i\omega) + Z(-i\omega)'$ is nonnegative definite Hermitian for all real ω.

where the nonsingular matrix T and the orthogonal matrix S are arbitrary, $R^{1/2}$ is the symmetric square-root of R, and (B_1, B_2) are two matrices, $n \times m$ and $n \times p$ respectively (p is arbitrary), such that (P, B_1, B_2) satisfy the conditions

(1.17a) $$FP + PF' + B_1B_1' + B_2B_2' = 0$$

(1.17b) $$PH' + B_1R^{1/2} = G,$$

(1.17c) P is a symmetric, positive definite $n \times n$-matrix.

Conversely, any $[A, B, C, D]$ constructed in this fashion is a wide sense minimal realization. It is no restriction to set $T = I$ and $S = I$ in (1.16), i.e. to consider only realizations of the form

(1.18a) $$dx = Fx\,dt + B_1\,du + B_2\,dv,$$

(1.18b) $$dz = Hx\,dt + R^{1/2}\,du$$

where $w = \binom{u}{v} \in \mathcal{W}_{m+p}$. In fact, all other stochastic realizations can be obtained from (1.18) by multiplying (1.18a) by an arbitrary T and transforming w by an orthogonal transformation. Consequently we shall be working in a fixed coordinate system, thereby identifying each transfer function (spectral factor) W with one quadruplet $[F, B, H, (R^{1/2}, 0)]$. Hence the wide sense problem is reduced to determining $B = (B_1, B_2)$.

The main topic of this paper is the characterization of all *proper* minimal stochastic realizations. This is a probabilistic problem. In addition to the input-output map of (1.1) we need to determine the input process w, which is no longer arbitrary; hence we shall be looking for quintuplets $[A, B, C, D, w]$. For an arbitrary representation (1.1), let $(\Omega, \mathcal{F}, P)$ be a probability space on which both y and w are defined, and define $H(y)$ and $H(w)$ to be the closed linear hulls in $L_2(\Omega, \mathcal{F}, P)$ of $\{y_i(t); t \in (-\infty, \infty), i = 1, 2, \cdots, m\}$ and $\{w_i(t); t \in (-\infty, \infty), i = 1, 2, \cdots, k\}$ respectively. Since y is given, $H(y)$ is fixed, whereas $H(w)$ varies with different choices of representation (1.1). For a proper stochastic realization we will always have $H(y) \subset H(w)$. We shall say that $[A, B, C, D, w]$ is an *internal stochastic realization* if $H(y) = H(w)$ and an *external stochastic realization* if $H(y) \neq H(w)$, adding the attribute *minimal* as appropriate. Hence the internal realizations are precisely those proper stochastic realizations which can be constructed in terms of the given process y, whereas the external realizations require extending our probabilistic setting with an exogeneous noise generator unrelated to y. Various aspects of the proper stochastic realization problem have been studied by Akaike [1], [2], Picci [23], [24] and Rozanov [26], but here we shall give a *complete* characterization of all such realizations. (In [21] the internal realizations are constructed from basic principles without first assuming that they are defined by models of type (1.1).) After submitting this paper we have learned about a series of as yet unpublished papers by Ruckebusch [27]–[29] containing discrete-time counterparts of some of the results presented here; these papers provide an alternative approach to the problem.

The outline of the paper goes as follows. Section 2 is devoted to preliminaries and definitions. In § 3 we show that to each proper stochastic realization there is a representation (1.1) with $\mathrm{Re}\{\lambda(A)\} > 0$ and $z = y$, the dynamic relations of which evolve backward in time. These representations, which are an important tool in our subsequent analysis, are called *proper backward stochastic realizations.* In §§ 4 and 5 all internal stochastic realizations are characterized, and it is shown that these are precisely the proper stochastic realizations for which $B_2 = 0$. Each internal state process can be expressed in terms of two steady-state Kalman–Bucy estimates, one filter evolving in the forward direction from time $t=-\infty$ and the other in the backward direction from $t=\infty$. Sections 6 and 7 are devoted to external stochastic realizations. First, in § 6, we construct a system of differential equations in B_1 and B_2 which generates families of wide sense stochastic realizations, totally ordered with respect to state covariances. In § 7 this result is interpreted in terms of proper stochastic realizations and a complete characterization of all such realizations is provided.

This paper extends the results reported (without proofs) in our short note [20].

2. Preliminaries and definitions. Let the function Λ: $R^{n\times n}\to R^{n\times n}$ be given by

$$\Lambda(P)=FP+PF'+(G-PH')R^{-1}(G-PH')', \tag{2.1}$$

and define the set $\mathcal{P}=\{P|P'=P;\ \Lambda(P)\leqq 0\}$ of symmetric $n\times n$-matrices, where $Q\geqq 0$ ($Q>0$) means that Q is nonnegative (positive) definite. Also introduce the subset $\mathcal{P}_0=\{P\in\mathcal{P}|\Lambda(P)=0\}$.

In the following theorem we collect some facts from Anderson [5], Faurre [11] and Willems [32].

THEOREM 2.1. *The set $\mathcal{P}$ is closed, bounded and convex, and there are two elements P_* and P^* in $\mathcal{P}_0$ such that*

$$P_*\leqq P\leqq P^* \quad \textit{for all } P\in\mathcal{P}. \tag{2.2}$$

Moreover, $\mathcal{P}$ is the set of all solutions P of (1.17), *and $\mathcal{P}_0$ is the set of all such solutions for which $B_2=0$.*

Each $P\in\mathcal{P}$ can be interpreted as the covariance matrix (1.4) of the corresponding stochastic realization (1.18). Consequently, there is a minimum-variance (P_*) and a maximum-variance (P^*) wide sense stochastic realization, and for these realizations we have $B_2=0$.

For each $P\in\mathcal{P}$, define the *feedback matrix*

$$\Gamma=F-(G-PH')R^{-1}H, \tag{2.3}$$

the significance of which will be made clear below. Let the feedback matrices corresponding to P_* and P^* be denoted Γ_* and Γ^* respectively. It can be shown that $\text{Re}\{\lambda(\Gamma_*)\}<0$ and $\text{Re}\{\lambda(\Gamma^*)\}>0$ [32, p. 260], [11, p. 53]. Consequently, for each matrix N, the Lyapunov equation

$$\Gamma_*'M+M\Gamma_*+H'R^{-1}H+N=0 \tag{2.4}$$

has a unique solution $M_*(N)$, which is positive definite whenever N is nonnegative definite. In fact, since (F,H) is controllable, so is (Γ,H). (See e.g. [36].) Likewise

$$-\Gamma^{*\prime}M-M\Gamma^*+H'R^{-1}H+N=0 \tag{2.5}$$

has a unique positive definite solution $M^*(N)$ for each $N\geqq 0$. Furthermore, define $\mathcal{P}_+=\{P\in\mathcal{P}|P>P_*\}$ and $\mathcal{P}_-=\{P\in\mathcal{P}|P<P^*\}$. Since $\Phi(i\omega)>0$ for all real ω, $P_*<P^*$ [32, p. 360], and consequently $\mathcal{P}_+$ and $\mathcal{P}_-$ are nonempty.

THEOREM 2.2. *Let Π and $\bar{\Pi}$ be the unique solutions of the $n\times n$-matrix differential equations*

$$\dot{\Pi}(t)=\Lambda(\Pi(t)); \qquad \Pi(0)=0 \tag{2.6}$$

and

$$\dot{\bar{\Pi}}(t)=\bar{\Lambda}(\bar{\Pi}(t)); \qquad \bar{\Pi}(0)=0 \tag{2.7}$$

respectively, where Λ is given by (2.1) *and $\bar{\Lambda}$ by*

$$\bar{\Lambda}(P)=F'P+PF+(H'-PG)R^{-1}(H'-PG)'. \tag{2.8}$$

Then $\Pi(t)\to P_*$ *and* $\bar{\Pi}(t)\to(P^*)^{-1}$ *as* $t\to\infty$. *Moreover, the matrix* $P=P_*+[M_*(N)]^{-1}$ *belongs to* $\mathcal{P}_+$ *if and only if* $N\geqq 0$. *Likewise,* $P=P^*-[M^*(N)]^{-1}$ *belongs to* $\mathcal{P}_-$ *if and only if* $N\geqq 0$. *Finally,* $P^*-P_*=[M_*(0)]^{-1}=[M^*(0)]^{-1}$.

Various versions of this theorem can be found in [7] and [11]. It provides us with a procedure to determine all elements in $\mathcal{P}_+\cup\mathcal{P}_-$: First compute P_* and P^*. Then varying N over the nonnegative cone will generate the other elements in $\mathcal{P}_+\cup\mathcal{P}_-$. The corresponding wide sense stochastic realizations $[F, B, H, (R^{1/2}, 0)]$ can then be obtained by determining $B=(B_1, B_2)$ from

$$B_1=(G-PH')R^{-1/2}, \tag{2.9a}$$

$$B_2B_2'=-\Lambda(P), \tag{2.9b}$$

which is merely (1.17) reformulated.

In § 6 another method for generating wide sense stochastic realizations is presented, which is formulated directly in terms of B, the unknown quantity in $[F, B, H', (R^{1/2}, 0)]$. Hence the intermediate step of determining P will be eliminated. Define $\mathcal{B}$ to be the set of all $B=(B_1, B_2)$ given by (2.9) as P ranges over $\mathcal{P}$. Let $\mathcal{B}_0$, $\mathcal{B}_+$ and $\mathcal{B}_-$ be defined analogously in terms of $\mathcal{P}_0$, $\mathcal{P}_+$ and $\mathcal{P}_-$. The set $\mathcal{B}_0$ consists of all $B\in\mathcal{B}$ with $B_2=0$ (Theorem 2.1). In particular, let B_* and B^* be the unique elements in $\mathcal{B}_0$ corresponding to P_* and P^* respectively.

All stochastic processes in this paper will have finite second order moments. Given a k-dimensional vector process η of this type, defined on some probability space $(\Omega, \mathcal{F}, P)$, and a subset I of $(-\infty,\infty)$, let $H_I(\eta)$ be the closed linear hull in $L_2(\Omega, \mathcal{F}, P)$ of the stochastic variables $\{\eta_i(t);\ t\in I, i=1, 2, \cdots, k\}$. (We write $H_t(\eta)$ if the set I contains only the point t.) If ξ is an l-dimensional stochastic vector such that $\xi_i\in H_I(\eta)$, $i=1, 2, \cdots, l$, we shall misuse notations slightly by writing $\xi\in H_I(\eta)$. For $\zeta\in L_2(\Omega, \mathcal{F}, P)$, let $\hat{E}\{\zeta|H_I(\eta)\}$ be the projection of ζ onto $H_I(\eta)$, i.e. the wide sense conditional mean in the terminology of Doob [10]. (We shall sometimes write $\hat{E}\{\zeta|\eta(t)\}$ instead of $\hat{E}\{\zeta|H_t(\eta)\}$.) For simplicity let $H(\eta)$, $H_t^-(\eta)$ and $H_t^+(\eta)$ denote $H_{(-\infty,\infty)}(\eta)$, $H_{(-\infty,t)}(\eta)$ and $H_{[t,\infty)}(\eta)$ respectively. Moreover, set $\eta_t(\tau)=\eta(t+\tau)-\eta(t)$, and define $H_t^-(d\eta)$ and $H_t^+(d\eta)$ to be respectively $H_0^-(\eta_t)$ and $H_0^+(\eta_t)$. Note that if $\eta(0)=0$ (which is often the case with the processes studied in this paper), we have $H_\infty^-(d\eta)=H(\eta)$.

As mentioned in § 1, any mean-square continuous stochastic vector process $\{\eta(t);\ t\in R\}$ with stationary increments and $\eta(0)=0$ has a representation of the form

$$\eta(t)=\int_{-\infty}^{\infty}\frac{e^{i\omega t}-1}{i\omega}\,d\hat{\eta}(\omega) \tag{2.10}$$

[12; p. 205], where $d\hat{\eta}$ is an orthogonal stochastic measure, called the *stochastic spectral measure* of η. If, in addition, η is purely nondeterministic, it has an absolutely continuous spectral distribution [9], i.e.

$$E\{d\hat{\eta}(\omega)\,d\hat{\eta}(\omega)\dagger\}=S(i\omega)\,d\omega \tag{2.11}$$

where S is the *spectral density* of η. If $E\{\eta(t)\}=0$ for all t and $S=I$ (identity), η is said to be of class $\mathcal{W}$. The spectral decomposition (2.10) defines an isometric correspondence between $H(\eta)$ and $L_2(R, S(i\omega)\,d\omega)$ under which $\eta(t)$ corresponds to $(e^{i\omega t}-1)/i\omega$; hence to any real random variable $\xi\in H(\eta)$ there corresponds an (essentially) unique $g\in L_2(R, S(i\omega)\,d\omega)$ such that

$$\xi=\int_{-\infty}^{\infty}g(\omega)\,d\hat{\eta}(\omega).$$

In fact, the system of functions $\{(e^{i\omega t}-1)/i\omega;\ t\in R\}$ is complete in $L_2(R, S(i\omega)\,d\omega)$[12; p. 204]. Hence we have the following lemma which we shall need below.

LEMMA 2.3. *Let ξ and η be mean-square continuous and purely nondeterministic stochastic vector processes, defined on the whole real line R, with (jointly) stationary increments and such that $\xi(t)\in H(\eta)$ for all $t\in R$. Let $S(i\omega)$ be the spectral density of η, and assume that $\xi(0)=0$. Then there is a matrix-valued function K such that $((e^{i\omega t}-1)/i\omega)\cdot K(i\omega)\in L_2(R, S(i\omega)\,d\omega)$ for all $t\in R$ and such that*

$$\xi(t)=\int_{-\infty}^{\infty}\frac{e^{i\omega t}-1}{i\omega}K(i\omega)\,d\hat{\eta}\,(\omega). \tag{2.12}$$

If, in addition, ξ and η are both of class $\mathcal{W}$,

$$K(s)K(-s)'=I. \tag{2.13}$$

The last statement follows from $d\hat{\xi}=K(i\omega)\,d\hat{\eta}$ and the fact that both ξ and η have identity spectral densities.

3. Forward and backward stochastic realizations. Let $\{x(t);\ t\in R\}$ *be an* n-dimensional wide sense Markov process, i.e.

$$\hat{E}\{x(s)|H_t^-(x)\}=\hat{E}\{x(s)|x(t)\}\quad \text{for } s\geqq t, \tag{3.1}$$

or equivalently

$$\hat{E}\{x(s)|H_t^+(x)\}=\hat{E}\{x(s)|x(t)\}\quad \text{for } s\leqq t. \tag{3.2}$$

In addition, assume that x is purely nondeterministic and (wide sense) stationary. It is well-known [11] that such a process can be described as the solution of a system of linear stochastic differential equations of the type

$$dx=Ax\,dt+B\,dw, \tag{3.3}$$

where A and B are constant matrices, $\operatorname{Re}\{\lambda(A)\}<0$, and w is a vector process of class $\mathcal{W}$ such that[2] $H_t^+(dw)\perp H_t^-(x)$ for all $t\in R$. [In fact, A being a stability matrix implies that (3.3) has the solution (1.3), and consequently $H_t^-(x)\subset H_t^-(dw)\perp H_t^+(dw)$.] Moreover, the covariance matrix $P:=E\{x(t)x(t)'\}$ satisfies (1.5). The model (3.3) is clearly unsymmetric with respect to time, $x(t)$ being orthogonal to future increments of w, but not to past ones. Hence we shall call (3.3) the *forward representation* of x.

We shall now show that x has a *backward representation* also, i.e. a model (3.3) with $\operatorname{Re}\{\lambda(A)\}>0$ and $H_t^-(dw)\perp H_t^+(x)$ for all $t\in R$. To this end first observe that the forward representation (3.3) can be integrated between t and s to yield

$$x(s)=e^{A(s-t)}x(t)+\int_t^s e^{A(s-\tau)}B\,dw(\tau), \tag{3.4}$$

where the two terms are orthogonal if and only if $s\geqq t$; in this case it can be seen that (3.4) is precisely the orthogonal decomposition

$$x(s)=\hat{E}\{x(s)|H_t^-(x)\}+[x(s)-\hat{E}\{x(s)|H_t^-(x)\}]. \tag{3.5}$$

We shall use a symmetric argument to determine the backward representation. More precisely, for $s\leqq t$ we shall derive a backward version of (3.4) from the decomposition

$$x(s)=\hat{E}\{x(s)|H_t^+(x)\}+[x(s)-\hat{E}\{x(s)|H_t^+(x)\}]. \tag{3.6}$$

[2] "$H_1\perp H_2$" means "H_1 and H_2 are orthogonal".

In view of the Markov property (4.2) and the standard projection formula [11] the first term in (3.6) can be written

$$\begin{aligned}\hat{E}\{x(s)|H_t^+(x)\} &= E\{x(s)x(t)'\}E\{x(t)x(t)'\}^{-1}x(t)\\ &= Pe^{A'(t-s)}P^{-1}x(t) = e^{-PA'P^{-1}(s-t)}x(t),\end{aligned} \tag{3.7}$$

where we have used (3.4) to evaluate $E\{x(s)x(t)'\}$. From (3.7) it is clear that

$$\xi(t) = e^{PA'P^{-1}t}x(t) \tag{3.8}$$

is a wide sense *backward martingale* with respect to the family $\{H_t^+(x)\}$, i.e.

$$\hat{E}\{\xi(s)|H_t^+(x)\} = \xi(t) \quad \text{for } s \leqq t, \tag{3.9}$$

and using (3.3) we obtain

$$d\xi = e^{PA'P^{-1}t}[(AP + PA')P^{-1}x\,dt + B\,dw],$$

which, because of (1.5), may be written

$$d\xi = e^{PA'P^{-1}t}B(dw - B'P^{-1}x\,dt). \tag{3.10}$$

LEMMA 3.1. *Let $\{x(t);\ t \in R\}$ be the solution on $(-\infty, \infty)$ of (3.3), and let P be the covariance matrix of x. Then the vector process $\bar{w}$, defined by*

$$d\bar{w} = dw - B'P^{-1}x\,dt; \qquad \bar{w}(0) = 0, \tag{3.11}$$

belongs to class $\mathcal{W}$, and $H_t^-(d\bar{w})$ is orthogonal to $H_t^+(x)$ for all $t \in R$.

Proof. Inserting (1.6) and (1.7a) for w and x in (3.11) yields

$$\bar{w}(t) = \int_{-\infty}^{\infty} \frac{e^{i\omega t} - 1}{i\omega} T(i\omega)\,d\hat{w}(\omega). \tag{3.12}$$

where

$$T(s) = I - B'P^{-1}(sI - A)^{-1}B. \tag{3.13}$$

Consequently $\bar{w}$ is a zero-mean, mean-square continuous vector process with stationary increments and spectral density $T(s)T(-s)'$ and such that $\bar{w}(0) = 0$. Then, to see that $\bar{w}$ is of class $\mathcal{W}$, it just remains to show that

$$T(s)T(-s)' = I. \tag{3.14}$$

To this end first note that

$$T(s)T(-s)' = I - B'P^{-1}(sI - A)^{-1}B - N(s)P^{-1}B, \tag{3.15}$$

where

$$N(s) = T(s)B'(-sI - A')^{-1} \tag{3.16a}$$

$$= B'(-sI - A')^{-1} - B'P^{-1}(sI - A)^{-1}BB'(-sI - A')^{-1}. \tag{3.16b}$$

In view of (1.5) we may write

$$BB' = (sI - A)P + P(-sI - A')$$

which inserted into (3.16b) yields

$$N(s) = -B'P^{-1}(sI - A)^{-1}P. \tag{3.17}$$

Now (3.15) and (3.17) together yield (3.14). To show that $H_t^-(d\bar{w}) \perp H_t^+(x)$, take

$t_1 \leqq t_2 \leqq t_3$ and form

$$E\{[\bar{w}(t_1)-\bar{w}(t_2)]x(t_3)'\}=\int_{-\infty}^{\infty}\frac{e^{i\omega t_1}-e^{i\omega t_2}}{i\omega}\,e^{-i\omega t_3}N(i\omega)\,d\omega. \tag{3.18}$$

Here we have used (3.12), (1.7a) and (3.16a) to obtain (3.18). But $(e^{-i\omega\alpha}-e^{-i\omega\beta})/i\omega$ is the Fourier transform of the indicator function $\chi_{(\alpha,\beta)}$ of the interval (α, β) and, in view of (3.17), $N(i\omega)$ is the Fourier transform of $-B'P^{-1}\,e^{At}P\chi_{(0,\infty)}$. Hence Parseval's Theorem yields

$$E\{[\bar{w}(t_1)-\bar{w}(t_2)]x(t_3)'\}=B'P^{-1}\int_{-\infty}^{\infty}\chi_{(t_1-t_3,t_2-t_3)}(t)\chi_{(0,\infty)}(t)\,e^{At}\,dt\,P,$$

which is zero whenever t_1, $t_2 \leqq t_3$. □

Consequently, in view of (3.7)–(3.11), (3.6) can be written

$$\begin{aligned} x(s)&=e^{-PA'P^{-1}(s-t)}x(t)+e^{-PA'P^{-1}s}[\xi(s)-\xi(t)] \\ &=e^{-PA'P^{-1}(s-t)}x(t)+\int_t^s e^{-PA'P^{-1}(s-\tau)}B\,d\bar{w}\,(\tau), \end{aligned} \tag{3.19}$$

which is the backward counterpart of (3.4). Since $\operatorname{Re}\{\lambda(-PA'P^{-1})\}>0$ and $H_t^-(d\bar{w})\perp H_t^+(x)$ for all $t\in R$,

$$dx=-PA'P^{-1}x\,dt+B\,d\bar{w}, \tag{3.20}$$

obtained by differentiating (3.19), is a backward representation of x. In [22], [30] it was shown that, for arbitrary w and $\bar{w}$ of class $\mathscr{W}$, the solutions on $(-\infty, \infty)$ of (3.3) and (3.20) have the same second-order properties. Here we have demonstrated that, for the particular choice (3.11) of $\bar{w}$, these systems actually represent the same wide sense Markov process. We record this observation in the following theorem.

THEOREM 3.2. *Let $\{x(t);\, t\in R\}$ be a vector-valued, wide sense stationary, purely nondeterministic, wide sense Markov process with covariance matrix P. Then x has a forward representation* (3.3) *with* $\operatorname{Re}\{\lambda(A)\}<0$ *and $H_t^+(dw)\perp H_t^-(x)$ for all $t\in R$, and a corresponding backward representation* (3.20) *with $H_t^-(d\bar{w})\perp H_t^+(x)$ for all $t\in R$. The processes x, w and $\bar{w}$ are related as in* (3.11).

In § 1 we only considered stochastic realizations for which $\operatorname{Re}\{\lambda(A)\}<0$, i.e. with the state process x written in the forward form. From what has been said above, it is clear that we will get an isomorphic theory by reversing time. In particular, let us consider representations of the type

$$d\bar{x}=\bar{A}\bar{x}\,dt+\bar{B}\,d\bar{w}, \tag{3.21a}$$

$$d\bar{z}=\bar{C}\bar{x}\,dt+\bar{D}\,d\bar{w}, \tag{3.21b}$$

where $\operatorname{Re}\{\lambda(\bar{A})\}>0$ and $H_t^-(dw)\perp H_t^+(x)$ for all $t\in R$. We shall call (3.21) a proper or a wide sense *backward stochastic realization* of y, depending on whether the solution $\bar{z}$ of (3.21) on $(-\infty, \infty)$ equals y or has the same spectral density as y. Equation (3.21a) has the unique solution

$$\bar{x}(t)=-\int_t^{\infty}e^{\bar{A}(t-\tau)}\bar{B}\,d\bar{w}(\tau) \tag{3.22}$$

on $(-\infty, \infty)$, and by the procedure used in § 1 we obtain

$$\bar{z}(t)=\int_{-\infty}^{\infty}\frac{e^{i\omega t}-1}{i\omega}\,\bar{W}(i\omega)\,d\hat{\bar{w}}(\omega) \tag{3.23}$$

where

$$\bar{W}(s)=\bar{C}(sI-\bar{A})^{-1}\bar{B}+\bar{D}. \tag{3.24}$$

If (3.21) is a backward stochastic realization of y, we must have

$$\bar{W}(s)\bar{W}(-s)'=\Phi(s), \tag{3.25}$$

i.e. $\bar{W}$ is a *strictly unstable spectral factor* of Φ. Conversely, each such spectral factor $\bar{W}$ is the transfer function of an equivalence class of wide sense backward stochastic realizations; to see this proceed as in § 1. If $\bar{W}$ is minimal, we shall say that the realization (3.21) is *minimal*; only such representations will be considered in the sequel.

Consider the problem of determining all strictly unstable minimal spectral factors (3.24) of Φ. Since $\bar{W}(-s)\bar{W}(s)'=\Phi(s)'$, this problem is equivalent to finding all stable minimal factors $\bar{W}(-s)$ of $\Phi(s)'$. Given the representation (1.14)–(1.15), we have

$$\Phi(s)'=\bar{Z}(s)+\bar{Z}(-s)', \tag{3.26}$$

where $\bar{Z}$ is the positive real matrix function Z', i.e.

$$\bar{Z}(s)=G'(sI-F')^{-1}H'+\tfrac{1}{2}R. \tag{3.27}$$

Consequently we have reduced the problem to the one considered in § 1. In fact, all stable factors

$$\bar{W}(-s)=\bar{C}(sI+\bar{A})^{-1}(-\bar{B})+\bar{D} \tag{3.28}$$

of $\Phi(s)'$ are given by

$$[-\bar{A},-\bar{B},\bar{C},\bar{D}]=[TF'T^{-1},\ T(-\bar{B}_1,-\bar{B}_2)S,\ G'T^{-1},\ (R^{1/2},0)S] \tag{3.29}$$

where T is any nonsingular $n\times n$-matrix, S is any orthogonal matrix of appropriate dimension and $(\bar{B}_1,\bar{B}_2)$ satisfy

$$F'\bar{P}+\bar{P}F+\bar{B}_1\bar{B}_1'+\bar{B}_2\bar{B}_2'=0, \tag{3.30a}$$

$$\bar{P}G-\bar{B}_1R^{1/2}=H', \tag{3.30b}$$

$$\bar{P}\text{ is a symmetric, positive definite } n\times n\text{-matrix.} \tag{3.30c}$$

This the *dual spectral factorization problem* considered by Anderson [6] and Faurre [11]. As in the forward setting it is no restriction to take $T=I$ and $S=I$, i.e. to consider backward stochastic realizations of the form $[-F',(\bar{B}_1,\bar{B}_2),G',(R^{1/2},0)]$ only; then $\bar{P}$ in (3.30) is the state covariance matrix.

Let $\bar{\Lambda}$ be given by (2.8) and define $\bar{\mathcal{P}}=\{P=P'|\bar{\Lambda}(P)\leqq 0\}$ and $\bar{\mathcal{P}}_0=\{P\in\bar{\mathcal{P}}|\bar{\Lambda}(P)=0\}$. By Theorem 2.1, the set $\bar{\mathcal{P}}$ is closed, bounded and convex, and there are two elements $\bar{P}_*$ and $\bar{P}^*$ in $\bar{\mathcal{P}}_0$ such that $\bar{P}_*\leqq P\leqq\bar{P}^*$ for all $P\in\bar{\mathcal{P}}$. Moreover, $\bar{\mathcal{P}}$ is the set of all solutions $\bar{P}$ of (3.30), and $\bar{\mathcal{P}}_0$ is the set of all such solutions for which $\bar{B}_2=0$. Let $\bar{\mathcal{B}}$ be the set of all solutions $\bar{B}=(\bar{B}_1,\bar{B}_2)$ of (3.30a)–(3.30b) as $\bar{P}$ varies over $\bar{\mathcal{P}}$, and let $\bar{B}_*$ and $\bar{B}^*$ be the elements in $\bar{B}$ corresponding to $\bar{P}_*$ and $\bar{P}^*$ respectively. As expressed by the following lemma (which is essentially the same as one found in [11]) there is a one-one correspondence between $\mathcal{P}$ and $\bar{\mathcal{P}}$ as well as between $\mathcal{B}$ and $\bar{\mathcal{B}}$.

LEMMA 3.3. *The set of matrices* $(\bar{P},\bar{B}_1,\bar{B}_2)$ *given by*

$$\bar{P}=P^{-1}, \tag{3.31a}$$

$$(\bar{B}_1,\bar{B}_2)=P^{-1}(B_1,B_2) \tag{3.31b}$$

is a solution of (3.30) *if and only if* (P, B_1, B_2) *is a solution of* (1.17). *In particular,* $\bar{P}_* = (P^*)^{-1}$, $\bar{P}^* = (P_*)^{-1}$, $\bar{B}_* = (P^*)^{-1}B^*$ *and* $\bar{B}^* = (P_*)^{-1}B_*$.

Proof. Pre- and postmultiplying (1.17a) by P^{-1} and premultiplying (1.17b) by P^{-1}, it is seen that P is a solution of (1.17) if and only if (3.31a) is a solution of (3.30) with $(\bar{B}_1, \bar{B}_2)$ given by (3.31b). The rest of the statement then follows trivially from (3.31). □

Lemma 3.3 defines a bijective mapping between the sets $\mathscr{P}$ and $\bar{\mathscr{P}}$. This raises the question whether to each *proper* minimal stochastic realization with transfer function W there is a *unique* proper backward minimal stochastic realization whose transfer function is the dual spectral factor $\bar{W}$, and vice versa. In general this is not true, for a spectral factor may correspond to many proper minimal stochastic realizations (Theorem 7.1). However, we shall see that if, in addition, we require that the two realizations have the same *state space*, i.e. $H_t(\bar{x}) = H_t(x)$, for all $t \in R$, there is such a one-one correspondence under mild conditions on B, and that the input processes are related as in Lemma 3.1. Of course, taking (3.31) *and* (3.11) as the starting point, the families of forward and backward proper minimal stochastic realizations are seen to be bijectively related regardless of any condition on B.

THEOREM 3.4. *Let* (F, G, H, R) *be defined as in* § 1. *To each proper minimal stochastic realization of* y *of the form*

$$dx = Fx\,dt + B_1\,du + B_2\,dv, \tag{3.32a}$$

$$dy = Hx\,dt + R^{1/2}\,du, \tag{3.32b}$$

with state covariance matrix P, *there is one and, if* B_2 *has linearly independent columns, only one proper backward minimal stochastic realization of the form*

$$d\bar{x} = -F'\bar{x}\,dt + \bar{B}_1\,d\bar{u} + \vec{B}_2\,d\bar{v}, \tag{3.33a}$$

$$dy = G'\bar{x}\,dt + R^{1/2}\,d\bar{u}, \tag{3.33b}$$

with state covariance $\bar{P}$, *such that* (3.31) *holds and* $H_t(\bar{x}) = H_t(x)$ *for all* $t \in R$. *Conversely, to each realization* (3.33) *there is one and, if* $\bar{B}_2$ *has linearly independent columns, only one realization* (3.32) *such that* (3.31) *holds and* $H_t(x) = H_t(\bar{x})$ *for all* $t \in R$. *The stochastic processes in the two realizations are related in the following way*

$$\bar{x}(t) = P^{-1}x(t), \tag{3.34}$$

$$d\bar{u} = du - B_1'P^{-1}x\,dt; \qquad \bar{u}(0) = 0, \tag{3.35a}$$

$$d\bar{v} = dv - B_2'P^{-1}x\,dt; \qquad \bar{v}(0) = 0. \tag{3.35b}$$

The relations (3.31), (3.34) *and* (3.35) *define a bijective mapping between the families* (3.32) *and* (3.33) *of forward and backward stochastic realizations.*

Proof. The backward representation (3.20) corresponding to (3.32a) is

$$dx = -PF'P^{-1}x\,dt + B_1\,d\bar{u} + B_2\,d\bar{v}, \tag{3.36a}$$

where, according to Theorem 3.2, $\bar{u}$ and $\bar{v}$ are given by (3.35). Then (3.32b) and (3.35a) together yield

$$dy = (HP + R^{1/2}B_1')P^{-1}x\,dt + R^{1/2}\,d\bar{u},$$

which, in view of (1.17b), is the same as

$$dy = G'P^{-1}x\,dt + R^{1/2}\,d\bar{u}. \tag{3.36b}$$

Now let $\bar{x}$ be defined by (3.34). Then $H_t(\bar{x}) = H_t(x)$ for all $t \in R$ and $\bar{x}$ has the covariance matrix (3.31a). Moreover, (3.34) applied to (3.36) yields (3.33) with $\bar{B}$ given by (3.31b). Secondly, consider an arbitrary proper backward minimal realization

$$d\tilde{x} = -F'\tilde{x}\,dt + \bar{B}_1\,d\tilde{u} + \bar{B}_2\,d\tilde{v}, \tag{3.37a}$$

$$dy = G'\tilde{x}\,dt + R^{1/2}\,d\tilde{u} \tag{3.37b}$$

with $\bar{B}$ given by (3.31b) and $H_t(\tilde{x}) = H_t(x)$ for all $t \in R$. Due to the last condition, there is a nonsingular matrix S such that $x(t) = S\tilde{x}(t)$; since x and $\tilde{x}$ are stationary, S is constant. Set $T = P^{-1}S$. Then in view of (3.34), $\bar{x}(t) = T\tilde{x}(t)$. Hence (3.37) can be written

$$d\bar{x} = -TF'T^{-1}\bar{x}\,dt + T\bar{B}_1\,d\tilde{u} + T\bar{B}_2\,d\tilde{v}, \tag{3.38a}$$

$$dy = G'T^{-1}\bar{x}\,dt + R^{1/2}\,d\tilde{u}. \tag{3.38b}$$

Since $\bar{x}$ and $\tilde{x}$ have the same covariance matrix $\bar{P}$, we must have $T\bar{P}T' = \bar{P}$. Hence, in view of (3.38), (3.30) holds also with $(\bar{P}, F', \bar{B}, G')$ exchanged for $(T\bar{P}T', TF'T^{-1}, T\bar{B}, G'T^{-1})$; in particular, (3.30b) yields $T(\bar{P}G' - \bar{B}_1R^{1/2}) = H'$, which together with the original (3.30b) gives us $TH' = H'$. We also have $TF'T^{-1} = F'$. To see this, form $\hat{E}\{\bar{x}(s)|H_t^+(\bar{x})\}$ for all $s \leqq t$ by using first (3.33) and then (3.38); we get $e^{-F'(s-t)}\bar{x}(t)$ and $e^{-TF'T^{-1}(s-t)}\bar{x}(t)$ respectively. Hence $(F')^iH' = T(F')^iT^{-1}H' = T(F')^iH'$ for $i = 1, 2, \cdots, n$, and since (H, F) is observable we must have $T = I$. Therefore $\tilde{x} = \bar{x}$. Then comparing (3.33b) and (3.37b), we see that $\tilde{u} = \bar{u}$, and hence (3.33a) and (3.37a) yield $\tilde{v} = \bar{v}$, for the columns of $\bar{B}$ are linearly independent. Hence (3.33) and (3.37) are identical. Finally, the converse statement is obtained in the same way starting out with the backward realization (3.33). □

4. The minimum- and maximum-variance realizations. The proper stochastic realizations corresponding to P_* and P^*, the minimum and maximum elements of the set $\mathcal{P}$, will play an important role in what follows. Therefore we shall begin by providing an interpretation of these.

Consider an arbitrary proper minimal stochastic realization of the form (3.32) and with state covariance P. It is not hard to see that such a realization exists; we postpone the proof of this to § 7 (Theorem 7.1). It is well-known [35] that, for each fixed $T \in R$, the estimate

$$\hat{x}(t;\,T) = \hat{E}\{x(t)|H_{[T,t]}(dy)\} \qquad (t \geqq T) \tag{4.1}$$

is generated by the *Kalman–Bucy filter*

$$d\hat{x} = F\hat{x}\,dt + K(t-T)\,d\nu_T; \qquad \hat{x}(T;\,T) = 0 \qquad (T \leqq t < \infty), \tag{4.2a}$$

where $\{\nu_T(t);\, t \in [T, \infty)\}$ is the transient *innovation process*, defined by[3]

$$d\nu_T = R^{-1/2}(dy - H\hat{x}\,dt); \qquad \nu_T(\max\{0, T\}) = 0. \tag{4.2b}$$

The matrix function K, called the *Kalman–Bucy gain*, can be determined from the matrix Riccati equation

$$\dot{\Sigma} = F\Sigma + \Sigma F' - KK' + BB'; \qquad \Sigma(0) = P, \tag{4.3a}$$

$$K = \Sigma H'R^{-1/2} + B_1. \tag{4.3b}$$

[3] Our choice of initial conditions in (4.2b) and (4.5b), which are otherwise arbitrary, is to insure that $\nu_T(0) = 0$ ($\bar{\nu}_T(0) = 0$) for negative (positive) T.

In the same manner, given an arbitrary proper backward minimal stochastic realization of the form (3.33), it can be seen that

$$\hat{x}_b(t, T)=\hat{E}\{\bar{x}(t)|H_{[t,T]}(dy)\} \qquad (t \leqq T) \tag{4.4}$$

is given by the *backward Kalman–Bucy filter*

$$d\hat{x}_b=-F'\hat{x}_b\, dt+\bar{K}(T-t)\, d\bar{\nu}_T; \qquad \hat{x}_b(T, T)=0 \qquad (-\infty<t\leqq T), \tag{4.5a}$$

where $\{\bar{\nu}_T(t);\ t\in(-\infty, T]\}$, defined by

$$d\bar{\nu}_T=R^{-1/2}(dy-G'\hat{x}_b\, dt); \qquad \nu_T(\min\{0, T\})=0, \tag{4.5b}$$

is the transient *backward innovation process*, introduced in [17]. Here $\bar{K}$ is given by the dual matrix Riccati equation

$$\dot{\bar{\Sigma}}=F'\bar{\Sigma}+\bar{\Sigma}F-\bar{K}\bar{K}'+\bar{B}\bar{B}'; \qquad \bar{\Sigma}(0)=\bar{P}, \tag{4.6a}$$

$$\bar{K}=\bar{\Sigma}GR^{-1/2}-\bar{B}_1. \tag{4.6b}$$

Note that both ν_T and $\bar{\nu}_T$ are normalized orthogonal increment processes [17], so (4.2) and (4.5) can be regarded as a pair of "nonstationary stochastic realizations" of y. We shall now demonstrate that the steady-state versions of these representations are indeed proper stochastic realizations in the sense of this paper.

THEOREM 4.1. *There is one and only one proper stochastic realization* (3.32) *with state covariance matrix* P_*, *namely*

$$dx_*=Fx_*\, dt+B_*\, du_*, \qquad dy=Hx_*\, dt+R^{1/2}\, du_*, \tag{4.7}$$

and it is the steady-state Kalman–Bucy filter in the sense that, for each $t\in R$, $x_*(t)$, $u_*(t)$ *and* B_* *are the limits in mean square of* $\hat{x}(t, T)$, $\nu_T(t)$ *and* $K(t-T)$ *respectively as* $T\to-\infty$. *The* innovation process u_* *satisfies*

$$H_t^-(du_*)=H_t^-(dy) \tag{4.8}$$

for all $t\in R$, *and the projection of the state* $x(t)$ *of any stochastic realization* (3.32) *onto* $H_t^-(dy)$, *being given by*

$$\hat{E}\{x(t)|H_t^-(dy)\}=x_*(t), \tag{4.9}$$

is invariant with respect to the particular realization.

THEOREM 4.2. *There is one and only one proper stochastic realization* (3.32) *with state covariance* P^*, *namely*

$$dx^*=Fx^*\, dt+B^*\, du^*, \qquad dy=Hx^*\, dt+R^{1/2}\, du^*, \tag{4.10}$$

and it is the forward counterpart (*in the sense of Theorem* 3.4) *of the backward stochastic realization*

$$d\bar{x}_*=-F'\bar{x}_*\, dt+\bar{B}_*\, d\bar{u}_*, \qquad dy=G'\bar{x}_*\, dt+R^{1/2}\, d\bar{u}_* \tag{4.11}$$

where $\bar{x}_*(t)$, $\bar{u}_*(t)$ *and* $\bar{B}_*$ *are the limits in mean square of* $\hat{x}_b(t; T)$, $\bar{\nu}_T(t)$ *and* $\bar{K}(T-t)$ *respectively as* $T\to\infty$. *Then* x^* *and* u^* *are given by*

$$x^*(t)=P^*\bar{x}_*(t), \tag{4.12}$$

$$du^*=d\bar{u}_*-\bar{B}_*'P^*\bar{x}_*\, dt; \qquad u^*(0)=0 \tag{4.13}$$

and B_* *by Lemma* 3.3. *The* backward innovation process $\bar{u}_*$ *has the property*

$$H_t^+(d\bar{u}_*)=H_t^+(dy) \tag{4.14}$$

for all $t\in R$, *and*

$$\hat{E}\{\bar{x}(t)|H_t^+(dy)\}=\bar{x}_*(t) \tag{4.15}$$

for the state process $\bar{x}$ *of any backward stochastic realization* (3.33).

Before proving these theorems a few remarks are in order:

(i) It is well-known that

$$E\{[x(t)-\hat{x}(t;\ T)][x(t)-\hat{x}(t;\ T)]'\}=\Sigma(t-T), \tag{4.16}$$

where Σ is given by (4.3); the stationarity of x insures that (4.16) depends on the difference $t-T$ only. Likewise, set $E\{\hat{x}(t;\ T)\hat{x}(t;\ T)'\}=\Pi(t-T)$. Then

$$\Sigma(t)=P-\Pi(t). \tag{4.17}$$

Inserting (4.17) into (4.3) and applying (1.17) it is seen that Π satisfies (2.6) and that

$$K=(G-\Pi H')R^{-1/2}. \tag{4.18}$$

Hence $K(t)\to B_*$ as $t\to\infty$ by Theorem 2.2. The corresponding dual results are analogous. Consequently one could base the proofs of Theorems 4.1 and 4.2 on Theorem 2.2, but instead we shall offer a self-contained proof which is more direct. Note that (4.18) together with (2.6), and its dual counterparts, imply that the filters (4.2) and (4.5) are in fact invariant with respect to the particular realization which provides the process $x(\bar{x})$.

(ii) The choice of (3.33) as the standard form for the backward stochastic realizations rather than (3.36) is motivated by the dual spectral factorization problem. Relation (4.15) provides an additional justification for this choice. As in (4.9), the left member of (4.15) is invariant with respect to variations in the state process $\bar{x}$. On the other hand, were we to project the state process x of (3.36) onto the future space $H_t^+(dy)$, we would have

$$\hat{E}\{x(t)|H_t^+(dy)\}=P(P^*)^{-1}x^*(t), \tag{4.19}$$

which does not enjoy the same invariance properties. Indeed the natural setting for the process x is the forward, and not the backward, realization problem.

Proof of Theorem 4.1. For each fixed $t\in R$ the process $\{\xi(\tau);\ \tau\geqq -t\}$, where $\xi(\tau)=\hat{x}(t;\ -\tau)$, is a uniformly integrable wide sense martingale [10], and therefore $\hat{x}(t;\ T)$ tends to a limit $x_*(t)$ in mean square as $T\to-\infty$. Moreover,

$$\hat{x}(t,\ T)=\hat{E}\{x(t)|H_{[T,t]}(dy)\}\to\hat{E}\{x(t)|\textstyle\bigvee_{T\leqq t}H_{[T,t]}(dy)\} \tag{4.20}$$

in mean square [10], and hence (4.9) holds (a.s. for each t), for $\bigvee_{T\leqq t}H_{[T,t]}(dy)=H_t^-(dy)$. Then ν_T tends to a limit process u_*. Since ν_T has normalized orthogonal increments, the same must hold for u_*; hence u_* is of class $\mathscr{W}$. In view of (4.20), $\Pi(t)$ and $K(t)$, as given by (4.17) and (4.18), tend to limits; let us call these Π_∞ and K_∞ respectively. Consequently, x_* and u_* must satisfy

$$dx_*=Fx_*\,dt+K_\infty\,du_*,\qquad dy=Hx_*\,dt+R^{1/2}\,du_*,$$

which is a proper minimal stochastic realization of y with state covariance Π_∞. Thus $\Pi_\infty\in\mathscr{P}$. But since (4.16) is nonnegative definite for all $t\in R$, (4.17) implies that $P\geqq\Pi_\infty$, and this holds for all $P\in\mathscr{P}$, for the realization (3.32) is arbitrary. (By Theorem 7.1 there is a proper stochastic realization for each $P\in\mathscr{P}$.) Therefore $\Pi_\infty=P_*$, and consequently $K_\infty=B_*$. Given P_*, the matrix B_* is uniquely determined by (2.9a). Moreover, as we shall see in § 5, u_* is uniquely determined as a *causal* function of y through relations (5.10b) and (5.12). Hence there is only one proper stochastic realization (3.32) with $P=P_*$, and moreover $H_t^-(du_*)\subset H_t^-(dy)$. Since, in addition, $H_t^-(du_*)\supset H_t^-(dy)$, (4.8) holds. Also, since x_* is uniquely determined, the limit (4.20) is independent of the choice of state process x. □

Proof of Theorem 4.2. The statements concerning (4.11), (4.14) and (4.15) follow along the same lines as in the proof of Theorem 4.1, just reversing time. Then the statements concerning (4.10), (4.12) and (4.13) are a consequence of Theorem 3.4. □

5. Internal stochastic realizations. Consider an arbitrary proper stochastic realization (3.32) and its backward counterpart (3.33). The following lemma describes the relationship between the two input processes w and $\bar{w}$ and the output process y.

LEMMA 5.1. *Let* $(w, \bar{w})$ *be the pair of input processes defined above. Then the following relations hold for all* $t \in R$.

(i) $H_t^-(dy) \subset H_t^-(dw)$ *and* $H_t^+(dy) \subset H_t^+(d\bar{w})$,

(ii) $H(y) \subset H(w)$,

(iii) $H_t^-(d\bar{w}) \subset H_t^-(dw)$ *and* $H_t^+(dw) \subset H_t^+(d\bar{w})$,

(iv) $H(\bar{w}) = H(w)$.

Proof. Relations (i) and (ii) are trivial consequences of (1.1b) and (1.3) and (3.21b) and (3.22), recalling that $z = \bar{z} = y$. To obtain (iii), insert first (1.3) and then $\bar{x} = P^{-1}x$, as given by (3.22), into (3.11). Then (iv) is proven by letting $t \to \infty$ in the first of relations (iii) and $t \to -\infty$ in the second. □

Since the input process w is of class $\mathcal{W}$, (i) implies that the future increments of w are orthogonal to the past increments of y, i.e. $H_t^+(dw) \perp H_t^-(dy)$ for all $t \in R$. In the same manner it can be seen that $H_t^-(d\bar{w}) \perp H_t^+(dy)$ for all t. It follows from Theorem 5.5 below that the innovation process u_* and the backward innovation process $\bar{u}_*$ are the only input processes to satisfy relations (i) with equality; they satisfy (4.8) and (4.14) respectively. The only thing we can say about the future space of u^* is that $H_t^+(du^*) \subset H_t^+(dy)$, which follows from Theorem 5.5. Hence we have again detected a certain lack of symmetry between the minimum- and maximum-variance realizations.

We shall now consider those realizations for which the converse of relation (ii) holds.

DEFINITIONS. The proper forward or backward stochastic realization $[A, B, C, D; w]$ of y is said to be *internal* if $H(w) = H(y)$. If $H(w) \neq H(y)$, the realization is said to be *external.*

For an internal stochastic realization, the input process w can be expressed in terms of the output y. Therefore, if x is the state process, $x(t) \in H(y)$ for all $t \in (-\infty, \infty)$. In view of Lemma 5.1 (iv), the backward counterpart of any internal (forward) realization is also internal. Hence, in the sequel, we shall restrict our attention to forward realizations, and only consider backward ones when there is an interplay between the forward and backward settings. We now turn to the characterization of the set of internal realizations.

THEOREM 5.2. *A proper stochastic realization of y is internal if and only if it has a square transfer function W, i.e.* $W(s)$ *is* $m \times m$.

Proof. The proof consists of two parts. First we show that $H(w) = H(y)$ if and only if W has a left inverse. Secondly we show that W has a left inverse if and only if it is $m \times m$.

(i) Assume that $w(t) \in H(y)$ for all $t \in R$. Then there is a representation

$$w(t) = \int_{-\infty}^{\infty} \frac{e^{i\omega t} - 1}{i\omega} K(i\omega)\, d\hat{y}(\omega) \tag{5.1}$$

satisfying the conditions of Lemma 2.3. Therefore, since the stochastic spectral measure is unique, $d\hat{w}=K(i\omega)\,d\hat{y}$. But

$$d\hat{y}=W(i\omega)\,d\hat{w}, \tag{5.2}$$

for $y=z$ satisfies (1.7b), and consequently

$$d\hat{w}=K(i\omega)W(i\omega)\,d\hat{w}. \tag{5.3}$$

Postmultiply (5.3) by $d\hat{w}\dagger$, take expectation, and note that $E\{d\hat{w}\,d\hat{w}\dagger\}=I\,d\omega$ to see that

$$K(s)W(s)=I \tag{5.4}$$

by analytic continuation. Hence W has a left inverse. Conversely, assume that W has a left inverse K. Then (5.3) holds, and, in view of (5.2), we have (5.1). Hence $w(t)\in H(y)$ for all $t\in R$, and therefore $H(w)=H(y)$ (Lemma 5.1 (ii)).

(ii) An $m\times k$ rational transfer matrix $W(s)$ has a left inverse if and only if $\rho\{W\}=k$, where ρ stands for rank, defined with respect to the field of rational functions [34; p. 162, Thm. 5.5.3]. Therefore it remains to show that $\rho\{W\}=k$ if and only if $k=m$. To this end, apply Sylvester's inequality [34; p. 40] to (1.10) to obtain

$$\rho\{W(s)\}+\rho\{W(-s)'\}-k\leqq\rho\{\Phi\}\leqq\min\,[\rho\{W(s)\},\rho\{W(-s)'\}],$$

which can be written

$$2\rho\{W\}-k\leqq m\leqq\rho\{W\}, \tag{5.5}$$

for $\rho\{\Phi\}=m$. Consequently, if $\rho\{W\}=k$, we have $k=m$. Conversely, if $k=m$, (5.5) implies that $\rho\{W\}=k$. □

COROLLARY 5.3. *A proper minimal stochastic realization in the standard form* (3.32) *is internal if and only if* $B_2=0$.

Proof. The transfer function of (3.32) is

$$W(s)=H(sI-F)^{-1}(B_1,B_2)+(R^{1/2},0), \tag{5.6}$$

which is square if and only if $B_2=0$. □

Consequently the internal stochastic realizations in standard form are precisely the representations of the type

$$dx=Fx\,dt+B\,du, \tag{5.7a}$$

$$dy=Hx\,dt+R^{1/2}\,du \tag{5.7b}$$

among which we have the minimum-variance realization (4.7) and the maximum-variance realization (4.10).

THEOREM 5.4. *There is a one-one correspondence between the family of internal realizations* (5.7) *and the set* $\mathcal{P}_0$ *of solutions of the algebraic Riccati equation* $\Lambda(P)=0$. *The input process u of* (5.7) *is given by*

$$u(t)=\int_{-\infty}^{\infty}\frac{e^{i\omega t}-1}{i\omega}W^{-1}(i\omega)\,d\hat{y}, \tag{5.8}$$

where W is the transfer function of (5.7).

Proof. Each stochastic realization (5.7) has a state covariance matrix P which belongs to $\mathcal{P}_0$, since $B_2=0$. (Theorem 2.1). Hence it remains to show that to each $P\in\mathcal{P}_0$ there is one and only one proper stochastic realization (5.7) and that u is given by (5.8): To each $P\in\mathcal{P}_0$ there is one and only one spectral factor of the form (5.6), namely the square factor

$$W(s)=H(sI-F)^{-1}B+R^{1/2}, \tag{5.9}$$

for B is uniquely determined by (2.9a). Since R is nonsingular, (5.9) has an inverse W^{-1}. First define u by (5.8). Then $d\hat{y} = W(i\omega)\, d\hat{u}$, which transformed to the time domain yields (5.7). Secondly, let u be the input process of a proper stochastic realization with transfer function (5.9). Then $d\hat{y} = W(i\omega)\, d\hat{u}$, and hence u is given by (5.8). □

The internal realization (5.7) can be inverted in the time domain also by rewriting it in the form

$$dx = \Gamma x\, dt + BR^{-1/2}\, dy, \tag{5.10a}$$

$$du = R^{-1/2}(dy - Hx\, dt) \tag{5.10b}$$

where, in view of (2.9a),

$$\Gamma = F - BR^{-1/2}H \tag{5.11}$$

is the feedback matrix (2.3). Once there is a solution of (5.10a), u is given by (5.10b). For the two extreme realizations, corresponding to P_* and P^*, such solutions are immediate, namely

$$x_*(t) = \int_{-\infty}^{t} e^{\Gamma_*(t-\tau)} B_* R^{-1/2}\, dy(\tau) \tag{5.12}$$

and

$$x^*(t) = -\int_{t}^{\infty} e^{\Gamma^*(t-\tau)} B^* R^{-1/2}\, dy(\tau) \tag{5.13}$$

respectively. In fact, all eigenvalues of Γ_* (Γ^*) have negative (positive) real parts. (See § 2.) Then u_* and u^* can be determined from (5.10b).

Other internal stochastic realizations can now be handled by integrating stable modes over the past and unstable over the future, provided that the matrix Γ has no eigenvalues on the imaginary axis. However, since $P_* < P^*$ [32, p. 260], no such eigenvalues occur for $\mathcal{P}_0$-realizations [33, p. 630; Remark 19]. In fact the solution is surprisingly simple.

THEOREM 5.5. *Consider an internal stochastic realization* (5.7). *Let* $\Pi^+(\Pi^-)$ *be the projection operator onto the invariant subspace spanned by the eigenvectors corresponding to eigenvalues of the feedback matrix* (5.11) *with positive* (*negative*) *real parts. Then*

$$x(t) = \Pi^- x_*(t) + \Pi^+ x^*(t), \tag{5.14}$$

where x_* *and* x^* *are given by* (5.12) *and* (5.13). *The input process u is given by*

$$du = R^{-1/2}[dy - H\Pi^- x_*(t)\, dt - H\Pi^+ x^*(t)\, dt]. \tag{5.15}$$

The proof of Theorem 5.5 is based on the following lemma.

LEMMA 5.6 (J. C. Willems). *Let* $P \in \mathcal{P}_0$, *and let* Π^+ *and* Π^- *be defined as in Theorem* 5.5. *Then* $\Pi^+ + \Pi^- = I$ *and*

$$P = \Pi^- P_* + \Pi^+ P^*. \tag{5.16}$$

Moreover, with Γ_* *and* Γ^* *defined as above,*

$$\Pi^-\Gamma_*\Pi^- = \Pi^-\Gamma_* \quad \textit{and} \quad \Pi^+\Gamma^*\Pi^+ = \Pi^+\Gamma^*. \tag{5.17}$$

In view of the fact that $P^* - P_* > 0$ and (H, F) is observable (see § 1), this result is an immediate consequence of Theorem 6 and Lemma 8 in [33].

Proof of Theorem 5.5. Let P be the state covariance matrix of the stochastic realization (5.7). Hence $P \in \mathcal{P}_0$ (Corollary 5.3). Since $(\Pi^-)^2 = \Pi^-$ and $\Pi^-\Pi^+ = 0$, we have $\Pi^- P = \Pi^- P_*$ from (5.16). Consequently, in view of (2.9a) and (5.11),

$$\Pi^- B = \Pi^- B_*, \tag{5.18a}$$

$$\Pi^- \Gamma = \Pi^- \Gamma_* = \Pi^- \Gamma_* \Pi^-, \tag{5.18b}$$

where in the last relation we have also used (5.17). Hence, premultiplying (5.10a) by Π^- and using (5.18), it is seen that $\Pi^- x(t)$ satisfies the differential equation

$$d\xi = \Pi^- \Gamma_* \xi \, dt + \Pi^- B_* R^{-1/2} \, dy \tag{5.19}$$

on $(-\infty, \infty)$. But $\Pi^- x_*(t)$, too, satisfies (5.19) on $(-\infty, \infty)$. To see this, use (5.17). Therefore, since (5.19) has a unique solution on $(-\infty, \infty)$, we must have $\Pi^- x(t) = \Pi^- x_*(t)$ for all $t \in R$. In the same way we show that $\Pi^+ x(t) = \Pi^+ x^*(t)$. Hence, (5.14) follows from $\Pi^+ + \Pi^- = I$ (Lemma 5.6). Then insert (5.14) into (5.10b) to obtain (5.15). □

It follows from (5.12) and (5.13) that $x_*(t) \in H_t^-(dy)$ and $x^*(t) \in H_t^+(dy)$ for each $t \in R$. Therefore, (5.14) decomposes $x(t) \in H(y)$ into two components, one in $H_t^-(dy)$ and one in $H_t^+(dy)$. In view of (4.8) and (4.14), we can acquire symmetry between past and future by using (4.12) to rewrite (5.14) in the form

$$x(t) = \Pi^- x_*(t) + \Pi^+ P^* \bar{x}_*(t). \tag{5.20}$$

Consequently, the state process of any internal stochastic realization can be expressed in terms of the steady-state forward and backward Kalman–Bucy estimates, x_* and $\bar{x}_*$, and therefore it can be constructed from a linear combination of the filters (4.2) and (4.5), by taking the limit in quadratic mean.

6. Families of totally ordered stochastic realizations. Considering minimal stochastic realizations in the standard form (3.32) leaves only the matrix $B = (B_1, B_2)$ and the input process $w = \binom{u}{v}$ to be determined, the parameters (F, G, H, R) being given. This section will be devoted to studying the set $\mathcal{B}$ of feasible matrices B, defined in § 2; finding w will be the topic of § 7.

It was shown in § 4 (Theorem 4.1) that

$$B_* = \lim_{t \to \infty} K(t), \tag{6.1}$$

where K is the Kalman–Bucy gain function. This fact together with the following theorem provide us with a means to determine B_* directly without first having to obtain P_*.

THEOREM 6.1 (Kailath–Lindquist). *Let (K, Q) be the unique solution on $[0, \infty)$ of the system of matrix differential equations*

$$\dot{K} = -QQ'H'R^{-1/2}; \qquad K(0) = GR^{-1/2} \tag{6.2a}$$

$$\dot{Q} = (F - KR^{-1/2}H)Q; \qquad Q(0) = GR^{-1/2}. \tag{6.2b}$$

Then K is the Kalman–Bucy gain function. The filter covariance function Π, defined in § 4 (Remark (i)), satisfies

$$\dot{\Pi} = QQ'; \qquad \Pi(0) = 0. \tag{6.3}$$

Note that, although different realizations (3.32) yield different Riccati equations (4.3) [but the same filter (4.2)], the *non-Riccati algorithm* (6.2) is invariant over $\mathcal{P}$, depending only on the known quantities (F, G, H, R). If needed, P_* can be determined as the limit of $\Pi(t)$ as $t \to \infty$ (Theorem 2.2), where Π is generated by either (2.6) or (6.3). The system (6.2)–(6.3) is precisely the algorithm derived in [17] by using the transient backward innovation process (4.5b) and in [16] by factoring the matrix differential equation (4.3). A dual non-Riccati algorithm generating the backward Kalman–Bucy gain $\bar{K}$ and the backward filter covariance $\bar{\Pi}$ can be derived analogously by using the *forward* innovation (4.2b) or alternatively from (4.7) by applying the technique of [16]; formally it can be obtained by merely exchanging (F, G, H, R) for (F', G', H', R) in (6.2).

It can be seen that $K(t)$ approaches B_* from outside of $\mathscr{B}$. In fact, as one can see by comparing (2.9a) and (4.18), $K(t)$ is related to $\Pi(t)$ as B_* to P_*, and, in view of (6.3), Π is monotonely nondecreasing starting out with $0 \notin \mathscr{P}$ at $t=0$; hence $\Pi(t) \leqq P_*$ for all t. Here we shall show that there are equations similar to (6.2) whose trajectories, with the proper initial conditions, lie entirely inside $\mathscr{B}$. These equations will consequently generate families of wide sense stochastic realizations. Again the basic idea is to eliminate the need of going via the auxilliary quantity P.

THEOREM 6.2. *Let $[F, B_0, H, (R^{1/2}, 0)]$ be an arbitrary wide sense minimal stochastic realization of y in standard form, and let $\theta \to B(\theta) = [B_1(\theta), B_2(\theta)]$ be the unique solution on $(-\infty, \infty)$ of the system of matrix differential equations*

$$\frac{dB_1}{d\theta} = B_2 B_2' H' R^{-1/2}, \tag{6.4a}$$

$$\frac{dB_2}{d\theta} = (F - B_1 R^{-1/2} H) B_2 \tag{6.4b}$$

with initial condition $B(0) = B_0$. For each $\theta \in (-\infty, \infty)$, let $P(\theta)$ be the unique solution of the Lyapunov equation

$$FP + PF' + B(\theta)B(\theta)' = 0. \tag{6.5}$$

Then, for each $\theta \in (-\infty, \infty)$, $[F, B(\theta), H, (R^{1/2}, 0)]$ is a wide sense minimal stochastic realization of y with state covariance matrix $P(\theta)$. This family of realizations is totally ordered in the sense that $P(\theta_2) \leqq P(\theta_1)$ for $\theta_1 \leqq \theta_2$. If $B_0 \in \mathscr{B}_-$, $B(\theta) \to (B_*, 0)$ as $\theta \to \infty$, and if $B_0 \in \mathscr{B}_+$, $B(\theta) \to (B^*, 0)$ as $\theta \to -\infty$. The function $\theta \to P(\theta)$ satisfies the differential equations (6.7) *and*

$$\frac{dP}{d\theta} = -B_2 B_2', \tag{6.6}$$

and also conditions (iii) *and* (iv) *of Lemma* 6.3 *where here P_0 may be any point on the trajectory $\{P(\theta); -\infty < \theta < \infty\}$.*

The proof of this theorem is based on the following lemma.

LEMMA 6.3. *Let Λ be defined by* (2.1). *Then, for each $P_0 \in \mathscr{P}$, the matrix differential equation*

$$\frac{dP}{d\theta} = \Lambda(P(\theta)); \qquad P(0) = P_0 \tag{6.7}$$

has a unique solution on $(-\infty, \infty)$, such that (i) *$P(\theta) \in \mathscr{P}$ for all $\theta \in (-\infty, \infty)$,* (ii) *$P(\theta_2) \leqq P(\theta_1)$ for $\theta_1 \leqq \theta_2$,* (iii) *if $P_0 \in \mathscr{P}_-$, $P(\theta) \to P_*$ as $\theta \to \infty$, and* (iv) *if $P_0 \in \mathscr{P}_+$, $P(\theta) \to P^*$ as $\theta \to -\infty$.*

Proof. First note that (6.7) can be replaced by the system

$$\frac{dP}{d\theta} = U(\theta)\Lambda(P_0)U(\theta)'; \qquad P(0) = P_0, \tag{6.8a}$$

$$\frac{dU}{d\theta} = \Gamma(\theta)U(\theta); \qquad U(0) = I, \tag{6.8b}$$

where $\Gamma(\theta)$ is the feedback matrix (2.3) corresponding to $P(\theta)$. To see this, reformulate (6.7) to read

$$\frac{dP}{d\theta}=(F-GR^{-1}H)P+P(F-GR^{-1}H)'+PH'R^{-1}HP+GR^{-1}G,$$

and use the differentiation technique employed by Kailath in [15], i.e. observe that

$$\frac{d^2P}{d\theta^2}=\Gamma(\theta)\frac{dP}{d\theta}+\frac{dP}{d\theta}\Gamma(\theta)'; \qquad \frac{dP}{d\theta}(0)=\Lambda(P_0),$$

and integrate to obtain (6.8).

Clearly the Riccati equation (6.7) has a unique solution locally in the neighborhood of $\theta=0$. In fact, at least for small θ, $P(\theta)=Y(\theta)X(\theta)^{-1}$, where the $n\times n$-matrix valued functions X and Y satisfy a system of linear differential equations such that $X(\theta)^{-1}$ exists for sufficiently small θ [8, p. 156]. Since $P_0\in\mathcal{P}$, $\Lambda(P_0)\leqq 0$, and hence, in view of (6.8a), the condition

$$\text{(6.9)}\qquad \frac{dP}{d\theta}\leqq 0$$

holds along this trajectory. Consequently, (6.7) implies $\Lambda(P(\theta))\leqq 0$, i.e. the trajectory is contained in the bounded (Theorem 2.1) set $\mathcal{P}$. Hence the solution can be extended to the whole real line, for $P(\theta)$ will never leave $\mathcal{P}$. Since Λ is locally Lipschitz, this solution is unique. This also proves (i), and (ii) is a consequence of (6.9).

To prove (iv) we use an argument similar to that in Willems [33, p. 631]. In view of the fact that $\Lambda(P_*)\leqq 0$, $S(\theta):=P(\theta)-P_*$ is the solution of

$$\frac{dS}{d\theta}=\Gamma_*S+S\Gamma_*'+SH'R^{-1}HS; \qquad S(0)=P_0-P_*.$$

Since $S(0)>0$ (for $P_0\in\mathcal{P}_+$) and $dS/d\theta\leqq 0$ (by (6.9)), $S(\theta)>0$ for $\theta\leqq 0$. Consequently S^{-1} exists on $(-\infty,0]$. Let M_* be defined as in Theorem 2.2, and define $V:=S^{-1}-M_*(0)$. It is easy to see that V satisfies

$$\frac{dV}{d\theta}=-\Gamma_*'V-V\Gamma_*$$

on $(-\infty,0]$. Since $\operatorname{Re}\{\lambda(-\Gamma_*)\}>0$, $V(\theta)\to 0$ as $\theta\to-\infty$, and hence $S(\theta)\to[M_*(0)]^{-1}=P^*-P_*$ (Theorem 2.2). Therefore $P(\theta)\to P^*$ as $\theta\to-\infty$. This proves (iv). The proof of (iii) is analogous; just exchange substar ($_*$) by superstar (*) everywhere and $(-\infty,0]$ for $[0,\infty)$. (Now $S(0)<0$ for $\theta\geqq 0$.) □

Hence, given any P_0 in $\mathcal{P}_+\cap\mathcal{P}_-$, we may construct a trajectory $\mathcal{T}\subset\mathcal{P}$ extending from P^* through P_0 to P_* so that $\mathcal{T}$ is a totally ordered set of matrices P satisfying (1.17). The only difference between (2.6) and (6.7) is the initial conditions ($0\notin\mathcal{P}$); the differential equation is the same. Its critical points are precisely the elements of $\mathcal{P}_0$, one of which (P_*) is locally stable in the forward direction and another of which (P^*) is stable in the backward direction (cf. [33]). Note, however, that (6.2) and (6.4) are not exactly the same, although they are derived from the same differential equation. A dual (backward) version of (6.1) can be obtained by factoring (2.7), with $\Pi(0)\in\bar{\mathcal{P}}$, as above.

Proof of Theorem 6.2. Let P_0 be the state covariance of the initial realization $[F, B_0, H, (R^{1/2}, 0)]$, and let $\{P(\theta);\ -\infty<\theta<\infty\}$ be the trajectory through P_0 defined by

Lemma 6.3. Define $B(\theta)$ as

$$B_1(\theta)=[G-P(\theta)H']R^{-1/2}, \tag{6.10a}$$

$$B_2(\theta)=U(\theta)(B_0)_2, \tag{6.10b}$$

where U is given by (6.8b). Then (6.6) and (6.4a) follow from (6.8a) (for $\Lambda(P_0)=-(B_0)_2(B_0)_2'$) and (6.4b) is a consequence of (6.8b) and (6.10). A local Lipschitz condition insures uniqueness. In view of (6.6) and (6.7), we have $B_2(\theta)B_2(\theta)'=-\Lambda(P(\theta))$, which together with (6.10a) yields (6.5). Since Re $\{\lambda(F)\}<0$ and $(F, B(\theta))$ is controllable (for (F, B_0) is), (6.5) has a unique positive definite, symmetric solution [8]. This fact together with (6.5) and (6.10a) insures that $(P(\theta), B(\theta))$ satisfies (1.17), and consequently $[F, B(\theta), H, (R^{1/2}, 0)]$ is a wide sense stochastic realization with state covariance $P(\theta)$. By Lemma 6.2, $P(\theta)$ satisfies conditions (ii)–(iv), and obviously the last two conditions hold for any P_0 on the trajectory $\{P(\theta); -\infty<\theta<\infty\}$. Finally, the fact that $B_1(\theta)$ tends to $B_*(B^*)$ as $\theta\to\infty(\theta\to-\infty)$ under the stated conditions, follows from conditions (iii) and (iv) and (6.10a). Since $dP/d\theta\to 0$, (6.6) implies that $B_2(\theta)\to 0$ as $\theta\to\pm\infty$. □

In the next section we shall interpret Theorem 6.2 in terms of proper stochastic realizations.

7. External stochastic realizations. The following theorem gives a complete characterization of all proper minimal stochastic realizations.

THEOREM 7.1. *Let*

$$dx=Fx\,dt+B_1\,du+B_2\,dv, \tag{7.1a}$$

$$dy=Hx\,dt+R^{1/2}\,du \tag{7.1b}$$

be a proper minimal stochastic realization of y, *and let* $W_1(s)$ *and* $W_2(s)$ *be defined by*

$$W_1(s)=H(sI-F)^{-1}B_1+R^{1/2}, \tag{7.2a}$$

$$W_2(s)=H(sI-F)^{-1}B_2. \tag{7.2b}$$

Then

$$W(s)=[W_1(s), W_2(s)] \tag{7.3}$$

is a minimal stable spectral factor of the spectral density Φ *of* y, *and the input processes are given by*

$$v(t)=\int_{-\infty}^{\infty}\frac{e^{i\omega t}-1}{i\omega}W_2(-i\omega)'\Phi^{-1}(i\omega)\,d\hat{y}(\omega)+z(t) \tag{7.4a}$$

$$u(t)=\int_{-\infty}^{\infty}\frac{e^{i\omega t}-1}{i\omega}W_1(-i\omega)'\Phi^{-1}(i\omega)\,d\hat{y}(\omega) \tag{7.4b}$$

$$-\int_{-\infty}^{\infty}\frac{e^{i\omega t}-1}{i\omega}W_1^{-1}(i\omega)W_2(i\omega)\,d\hat{z}(\omega)$$

where z *is a mean-square continuous, purely nondeterministic stochastic vector process with stationary increments, zero mean, spectral density*

$$\Psi(s)=I-W_2(-s)'\Phi^{-1}(s)W_2(s), \tag{7.5}$$

and $z(0)=0$. *Moreover,* $\Psi(i\omega)>0$ *for all real* ω *and* $H(z)\perp H(y)$; *we shall call* z *the* exogeneous input component. *Conversely, for each minimal stable spectral factor* (7.3) *of* Φ, *there is a minimal proper stochastic realization* (7.1) *with* u *and* v *given by* (7.4), z *being an arbitrary stochastic vector process with all the properties prescribed above.*

Proof. It was shown in § 1 that, with (7.1) given, (7.3) is a minimal stable spectral factor of Φ; this result is restated here for completeness only. To see that u and v are given by (7.4), first decompose v as

$$v(t)=\hat{E}\{v(t)|H(y)\}+z(t). \tag{7.6}$$

Then $H(z)\perp H(y)$. Given the properties of v and y described in § 1, it is easy to see that the first term in this decomposition is a mean-square continuous, purely nondeterministic vector process with stationary increments, so the same must hold for z; in addition, z has zero mean and $z(0)=0$. Hence, since

$$d\hat{y}(\omega)=W_*(i\omega)\,d\hat{u}_*(\omega), \tag{7.7}$$

where $d\hat{u}_*$ is the stochastic spectral measure of the innovation process u_* and W_* is the transfer function of (4.7), and in view of Lemma 2.3, (7.6) can be written

$$v(t)=\int_{-\infty}^{\infty}\frac{e^{i\omega t}-1}{i\omega}Z(i\omega)\,d\hat{u}_*(\omega)+\int_{-\infty}^{\infty}\frac{e^{i\omega t}-1}{i\omega}\,d\hat{z}(\omega), \tag{7.8}$$

for some Z to be determined. Let Ψ denote the spectral density of the process z. Clearly there is a representation

$$d\hat{z}(\omega)=T(i\omega)\,d\hat{\mu}(\omega), \tag{7.9}$$

where $d\hat{\mu}$ is the stochastic spectral measure of a process μ of classs $\mathcal{W}$ such that $H(\mu)\perp H(y)$, and $T(s)$ is a spectral factor of $\Psi(s)$. Then (7.8) can be written

$$d\hat{v}=Z(i\omega)\,d\hat{u}_*+T(i\omega)\,d\hat{\mu}. \tag{7.10a}$$

Therefore, inserting (7.7) and (7.10a) into

$$d\hat{y}=W_1(i\omega)\,d\hat{u}+W_2(i\omega)\,d\hat{v}, \tag{7.11}$$

which is (7.1) rewritten in terms of spectral measures, and solving for $d\hat{u}$, we obtain

$$d\hat{u}=X(i\omega)\,d\hat{u}_*+Y(i\omega)T(i\omega)\,d\hat{\mu}, \tag{7.10b}$$

where

$$X(s)=W_1^{-1}(s)W_*(s)-W_1^{-1}(s)W_2(s)Z(s) \tag{7.12}$$

and

$$Y(s)=-W_1^{-1}(s)W_2(s), \tag{7.13a}$$

for the matrix R being nonsingular insures that W_1 has an inverse. Since both $\binom{u}{v}$ and $\binom{u_*}{\mu}$ are vector processes of class $\mathcal{W}$, the coefficient matrix function of (7.10), i.e.

$$K(s)=\begin{bmatrix}X(s) & Y(s)T(s)\\ Z(s) & T(s)\end{bmatrix},$$

satisfies relation (2.13) of Lemma 2.3, i.e.

$$X(s)X(-s)'+Y(s)T(s)T(-s)'Y(-s)'=I, \tag{7.14a}$$

$$X(s)Z(-s)'+Y(s)T(s)T(-s)'=0, \tag{7.14b}$$

$$Z(s)Z(-s)'+T(s)T(-s)'=I. \tag{7.14c}$$

Then inserting (7.12) into (7.14b) and applying (7.14c), we have

$$Z(s) = W_2(-s)' W_*^{-1}(-s)', \tag{7.13b}$$

which inserted into (7.12) yields

$$X(s) = W_1(-s)' W_*^{-1}(-s)'. \tag{7.13c}$$

To obtain this, we have used the fact that

$$\Phi(s) = W_1(s) W_1(-s)' + W_2(s) W_2(-s)'. \tag{7.15}$$

Now (7.10) together with (7.7) and (7.13) yield (7.4), and (7.13b) and (7.14c) give us (7.5), for $T(s)T(-s)' = \Psi(s)$. By using the matrix inversion lemma [14, p. 124], we can see that

$$\Psi(s) = [I + W_2(-s)' W_1^{-1}(-s) W_1^{-1}(s) W_2(s)]^{-1}. \tag{7.16}$$

Hence $\Psi(i\omega) > 0$ for all real ω.

Secondly, assume that a minimal stable spectral factor (7.3) is given; from it we can determine a quadruplet $[F, (B_1, B_2), H, (R^{1/2}, 0)]$. Let z be an arbitrary mean-square continuous process with stationary increments, zero mean, and spectral density (7.5), and such that $z(0) = 0$ and $H(z) \perp H(y)$. Since z has a rational spectral density, it is purely nondeterministic [9]. Define u and v by (7.4). Then the corresponding stochastic spectral measures $d\hat{u}$ and $d\hat{v}$ are given by (7.10) with X, Y, Z and T defined by (7.13) and (7.9). Straightforward calculations using (7.15) show that X, Y, Z and T satisfy (7.14), and consequently $\binom{u}{v}$ is a process of class $\mathscr{W}$. Finally, with the help of (7.15), we can see that $d\hat{u}$ and $d\hat{v}$ thus defined satisfy (7.11) (the z-components cancel), and therefore (7.1) is a proper stochastic realization of y. □

Theorem 7.1 provides us with an alternative proof of the "only if" part of Corollary 5.3. (Theorem 5.5 gives an alternative proof of the "if" part.) In fact, since $\Psi(i\omega) > 0$ for all real ω, the exogeneous input component z is never identically zero. Therefore, unless $B_2 = 0$, the output of (7.1) contains a component orthogonal to $H(y)$.

We are now in a position to interpret Theorem 6.2 in terms of proper minimal stochastic realizations. Consider an arbitrary such realization

$$dx = Fx\,dt + (B_0)_1\,du_0 + (B_0)_2\,dv_0, \qquad dy = Hx\,dt + R^{1/2}\,du_0 \tag{7.17}$$

with exogeneous input component z_0 having spectral density $\Psi_0(s)$. Let $T_0(s)$ be a square spectral factor of $\Psi_0(s)$ and define

$$\mu(t) = \int_{-\infty}^{\infty} \frac{e^{i\omega t} - 1}{i\omega} T_0^{-1}(i\omega)\,d\hat{z}_0(\omega). \tag{7.18}$$

(Since $\Psi_0(i\omega) > 0$ for all ω, $T_0(s)$ has an inverse.) Then, $\mu \in \mathscr{W}_k$, where k is the number of columns of $(B_0)_2$. Let $\mathscr{F}$ be the sigma-algebra generated by $\{y(t), \mu(t);\ t \in R\}$ and form the probability space $(\Omega, \mathscr{F}, P)$ on which (7.17) is defined. Then (7.17) gives rise to a family of proper minimal stochastic realizations

$$dx_\theta = Fx_\theta\,dt + B_1(\theta)\,du_\theta + B_2(\theta)\,dv_\theta, \qquad dy = Hx_\theta\,dt + R^{1/2}\,du_\theta, \tag{7.19}$$

which are defined on the same probability space $(\Omega, \mathscr{F}, P)$ and which are totally ordered in the sense that the state covariance function $P(\theta) = E\{x_\theta(t)x_\theta(t)'\}$ is monotonely nonincreasing in θ. In fact, for each $\theta \in [-\infty, \infty]$, define $W_1(s;\theta)$ and $W_2(s;\theta)$ by inserting $[B_1(\theta), B_2(\theta)]$, generated by (6.4), into (7.2), and let

$$z_\theta(t) = \int_{-\infty}^{\infty} \frac{e^{i\omega t} - 1}{i\omega} T_\theta(i\omega)\,d\hat{\mu}(\omega), \tag{7.20}$$

where $T_\theta(s)$ is a square spectral factor of

$$\Psi_\theta(s)=I-W_2(-s;\theta)'\Phi^{-1}(s)W_2(s;\theta). \tag{7.21}$$

(We may for example take all T_θ to be minimum phase.) Then define u_θ and v_θ by inserting $W_1(s;\theta)$, $W_2(s;\theta)$ and z_θ into (7.4). Hence $x_\theta(t)$, $u_\theta(t)$ and $v_\theta(t)$ belong to $H(y,\mu)$ for all t and all θ. If $B_0\in\mathscr{B}_-$, the family (7.19) will contain the steady-state Kalman–Bucy filter (4.7); if $B_0\in\mathscr{B}_+$, it will contain the maximum-variance model (4.10). Finally, if $B_0\in\mathscr{B}_0$, (7.19) will only contain one realization, (7.17) itself.

REFERENCES

[1] H. Akaike, *Markovian representation of stochastic processes by canonical variables*, this Journal, 13 (1975), pp. 162–173.

[2] ———, *Stochastic theory of minimal realization*, IEEE Trans. Automatic Control, AC-19 (1974), pp. 667–674.

[3] B. D. O. Anderson, *An algebraic solution to the spectral factorization problem*, Ibid., Automatic Control AC-12 (1967), pp. 410–414.

[4] ———, *A system theory criterion for positive real matrices*, this Journal, 5 (1967), pp. 171–182.

[5] ———, *The inverse problem of stationary covariance generation*, J. Statis. Phys., 1 (1969), pp. 133–147.

[6] ———, *Dual form of a positive real lemma*, Proc. IEEE, 55 (1967), pp. 1749–1750.

[7] B. D. O. Anderson and S. Vongpanitlerd, *Network Analysis and Synthesis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[8] R. W. Brockett, *Finite Dimensional Linear Systems*, Wiley, New York, 1970.

[9] H. Cramer and M. R. Leadbetter, *Stationary and Related Stochastic Processes*, Wiley, New York, 1967.

[10] J. L. Doob, *Stochastic Processes*, Wiley, New York, 1953.

[11] P. Faurre, *Realisations markoviennes de processus stationnaires*, Research Report no. 13, March 1973, IRIA (LABORIA), Le Chesnay, France.

[12] I. I. Gikhman and A. V. Skorokhod, *Introduction to the Theory of Random Processes, W. B.* Saunders, Philadelphia, 1969.

[13] B. L. Ho and R. E. Kalman, *Effective construction of linear state-variable models from input/output functions*, Proc. 3rd Allerton Conf. Circuits and Systems Theory, 1965, pp. 449–459.

[14] A. S. Householder, *The Theory of Matrices in Numerical Analysis*, Blaisdell, New York, 1964.

[15] T. Kailath, *Some Chandrasekhar-type algorithms for quadratic regulators*, Proc. IEEE Decision and Control Conference (New Orleans, Dec. 1972).

[16] ———, *Some new algorithms for recursive estimation in constant linear systems*, IEEE Trans. Information Theory IT-19 (1973), pp. 750–760.

[17] A. Lindquist, *Optimal filtering of continuous-time stationary processes by means of the backward innovation process*, this Journal, 12 (1974), pp. 747–754.

[18] ———, *On Fredholm integral equations, Toeplitz equations and Kalman–Bucy filtering*, Appl. Math. Optimization, 1 (1975), pp. 355–373.

[19] ———, *Linear least-squares prediction based on covariance data from stationary processes with finite-dimensional realizations*, Proc. Second European Congress on Operations Research (Stockholm, Sweden), North-Holland, Amsterdam, 1976, pp. 281–286.

[20] A. Lindquist and G. Picci, *A note on the stochastic realization problem*, Proc. Intern. Conf. Information Sciences and Systems (Patras, Greece, Aug. 1976), Hemisphere Publishing Corporation, 1977, pp. 1–5.

[21] ———, *On the structure of minimal splitting subspaces in stochastic realization theory*, Proc. 1977 Decision and Control Conference (New Orleans), pp. 42–48.

[22] L. Ljung and T. Kailath, *Backwards markovian models for second-order stochastic processes*, IEEE Trans. Information Theory, IT-22 (1976), pp. 488–491.

[23] G. Picci, *Stochastic realization of Gaussian processes*, Proc. IEEE, 64 (1976), pp. 112–122.

[24] ———, *Some connections between the theory of sufficient statistics and the identifiability problem*, SIAM J. Appl. Math., 33 (1977), pp. 383–398.

[25] V. M. Popov, *Hyperstability of Control Systems*, Springer-Verlag, New York, 1973.

[26] Yu. A. Rozanov, *On two selected topics connected with stochastic systems theory*, Appl. Math. Optimization, 3 (1976), pp. 73–80.

[27] G. RUCKEBUSCH, *Representations markoviennes de processus gaussiens stationnaires*, Thèse 3ème cycle, Paris VI, 1975.

[28] ———, *Representations markoviennes de processus gaussiens stationnaires et applications statistiques*, Centre de Mathematique Appliquées, internal report no. 18, Ecole Polytechnique, Palaiseau, France.

[29] ———, *On the theory of markovian representation*, preprint.

[30] G. S. SIDHU AND U. A. DESAI, *New smoothing algorithms based on reversed-time lumped models*, IEEE Trans. Automatic Control, AC-21 (1976), pp. 538–541.

[31] L. M. SILVERMAN AND H. E. MEADOWS, *Equivalence and synthesis of time variable linear systems*, Proc. 4th Allerton Conf. Circuit and Systems Theory, 1966, pp. 776–784.

[32] J. C. WILLEMS, *Dissipative dynamical systems, Part II: Linear systems with quadratic supply rates*, Arch. Rational Mech. Anal., 45 (1972), pp. 352–393.

[33] ———, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 621–634.

[34] W. A. WOLOVICH, *Linear Multivariable Systems*, Springer-Verlag, New York, 1974.

[35] E. WONG, *Stochastic Processes in Information and Dynamical Systems*, McGraw-Hill, New York, 1971.

[36] W. M. WONHAM, *On a matrix Riccati equation of stochastic control*, this Journal, 6 (1968), pp. 681–697.

[37] D. C. YOULA, *On the factorization of rational matrices*, IRE Trans. Information Theory, IT-7 (1961), pp. 172–189.

[38] D. C. YOULA AND P. TISSI, *n-port synthesis via reactance extraction—Part I*, IEEE Intern. Convention Record, 14 (1966), pp. 183–205.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 3, May 1979

STABILITY FOR A MULTI-RATE SAMPLED-DATA SYSTEM*

DAVID P. STANFORD†

Abstract. A sampled-data system with sampling interval lengths selected from a finite set is considered. Stabilizability of the system via feedbacks associated with sampling interval lengths is studied, and conditions for stabilizability involving "pre-contractiveness", "contractiveness", and "positive definiteness" of a finite set of matrices are given. Included in these results is a generalization of a theorem by P. Stein stating that for a real square matrix H, $\lim_{n\to\infty} H^n = 0$ if and only if there is a symmetric matrix Q such that $Q - H^TQH$ is positive definite. Finally, some results concerning a choice of feedbacks which will produce stability are presented.

Introduction. In this paper we study stabilizability of a sampled-data system in which the sampling interval length varies over a fixed finite set of positive numbers. The discretization of this system, as described in § 1, leads to a finite set of square matrices with which we hope, by successive multiplication, to be able to steer each vector to the origin. The matrices arrived at are a function of the selection of certain feedback matrices.

Section 2 determines a necessary and sufficient condition ("pre-contractiveness") for stabilizability of the system, and a sufficient condition ("contractiveness") for stabilizability. These conditions are stated in terms of arbitrary vector norms, and the norm-dependence of the conditions is discussed.

In § 3, contractiveness of a set of matrices is related to a notion of positive definiteness for a set of symmetric matrices, and a geometric description of such sets is discussed.

In § 4, we examine the choice of feedback matrices used to produce the system to be stabilized. "Best possible" choices are given for making the system contractive relative to a norm $(x^TQx)^{1/2}$ with Q positive definite. Finally, contractibility of the system relative to such a norm is related to positive definiteness of a set of symmetric matrices.

1. Formulation of the problem. We wish to stabilize a linear control system of the form

$$\dot{x} = Ax + Bu \tag{1}$$

where x is a real n-dimensional "state" vector, u is a real m-dimensional "control" vector, A is a real constant $n \times n$ matrix, and B a real constant $n \times m$ matrix. It is well known that on the interval $[t_1, t]$, a continuous control u will steer the state $x(t_1) = z$ to the state

$$x(t) = e^{TA}z + \int_0^T e^{vA}Bu(t-v)\,dv, \tag{2}$$

where T is the interval length $t - t_1$.

Using this, we define the multi-rate sampled-data system as follows. A finite set of positive numbers $\{S_1, S_2, \cdots, S_N\}$ is selected. These are the sampling interval lengths. At a sampling instant the state x_k is used to determine the next sampling interval length S_i and a *constant* control u_k to be applied in (1) to produce the state x_{k+1} at the next sampling instant, S_i time units later.

* Received by the editors May 24, 1978. This research was supported by NASA-Langley Research Center under Grant NAS1-11958.

† Department of Mathematics and Computer Science, College of William and Mary, Williamsburg, Virginia 23185.

Using (2),

$$x_{k+1}=e^{S_iA}x_k+\int_0^{S_i} e^{vA}\,dv\,Bu_k. \tag{3}$$

The constant value of u_k is to be determined by a feedback matrix depending only on the sampling interval length S_i selected. Thus for each i, $1 \leqq i \leqq N$, we must define an $m \times n$ real constant matrix F_i, and then when interval length S_i is chosen, we take

$$u_k=F_ix_k.$$

Then (3) becomes,

$$x_{k+1}=\left(e^{S_iA}+\int_0^{S_i} e^{vA}\,dv\,BF_i\right)x_k. \tag{4}$$

For $1 \leqq i \leqq N$, we define

$$C_i=e^{S_iA}, \quad \text{and} \quad D_i=\int_0^{S_i} e^{vA}\,dv\,B, \tag{5}$$

so that (4) can be written:

$$x_{k+1}=(C_i+D_iF_i)x_k. \tag{6}$$

Since our goal is to stabilize the solution to (1), we may view the problem as follows: We begin by taking the N interval lengths $S_1, S_2, \cdots, S_N$ as given. From the matrices A and B we determine C_i and D_i, $1 \leqq i \leqq N$, using (5). We then wish to select $F_1, F_2, \cdots, F_N$ so that the system (6) is stable in the sense that for any initial x_0, a sequence $\{i_k\}_{k=0}^{\infty}$ exists with $1 \leqq i_k \leqq N$ so that

$$x_{k+1}=(C_{i_k}+D_{i_k}F_{i_k})x_k$$

defines a sequence converging to 0.

DEFINITION 1. A set $\{H_1, H_2, \cdots, H_N\}$ of $n \times n$ matrices is *convergent* provided that, for each x in R^n, there is a sequence $\{p(x)_k\}_{k=1}^{\infty}$ such that $1 \leqq p(x)_k \leqq N$ and the sequence

$$x, \quad H_{p(x)_1}x, \quad H_{p(x)_2}H_{p(x)_1}x, \cdots$$

converges to 0. We assert this convergence by writing

$$\lim_{k\to\infty}\left(\prod_{i=k}^{1} H_{p(x)_i}\right)x=0.$$

Clearly the system (6) is stable for a selection $F_1, F_2, \cdots, F_N$ of feedback matrices if and only if the set $\{H_1, H_2, \cdots, H_N\}$ is convergent, where $H_i=C_i+D_iF_i$, $1 \leqq i \leqq N$.

2. Contractive and pre-contractive sets. In this section we relate convergence of a set of matrices to other properties of the set which we now introduce.

It is well known that if $N=1$, the set $\{H_1\}$ is convergent if and only if the spectral radius of H_1 is less than one; i.e., H_1 satisfies $\|H_1x\|<\|x\|$ for all nonzero $x \in R^n$ (Euclidean norm). Consider, then, the following example:

Example 1.

$$H_1=\begin{bmatrix} .5 & 0 \\ 0 & 2 \end{bmatrix}, \quad H_2=\begin{bmatrix} 1.25 & -.75 \\ -.75 & 1.25 \end{bmatrix}, \quad H_3=\begin{bmatrix} 2 & 0 \\ 0 & .5 \end{bmatrix}, \quad H_4=\begin{bmatrix} 1.25 & .75 \\ .75 & 1.25 \end{bmatrix}.$$

It can be routinely verified that each H_i has spectral radius 2. But $\{H_1, H_2, H_3, H_4\}$ is convergent, as seen from the following facts:

1. H_1 contracts each nonzero vector in the closed cone C_1, co-axial with the x-axis, having vertex angle at the origin, and measuring 45°. H_2, H_3, and H_4 act similarly on vectors in cones C_2, C_3, and C_4 which are counter-clockwise rotations of C_1 through 45°, 90°, and 135° respectively.

2. There is a number β, $0<\beta<1$, such that if $x \in R^2$, $x \neq 0$, then one of the vectors H_1x, H_2x, H_3x, H_4x satisfies:

$$\|H_ix\| < \beta\|x\|.$$

Thus the set $\{H_1, H_2, H_3, H_4\}$ is (exponentially) convergent, and for some sequence from $\{H_1, H_2, H_3, H_4\}$

$$\|x_k\| < \beta^k\|x_0\|.$$

The example suggests the following definition:

DEFINITION 2. Let $\|\cdot\|_v$ be any norm on R^n. The set $\{H_1, H_2, \cdots, H_N\}$ of $n \times n$ matrices is *contractive relative to* $\|\cdot\|_v$ provided that, if $x \in R^n$, $x \neq 0$, there is an i, $1 \leqq i \leqq N$, such that $\|H_ix\|_v < \|x\|_v$.

The fact that contractiveness of a set of matrices relative to some norm implies (exponential) convergence will appear as a corollary to Theorem 1.

Consider another example with $N=2$.

Example 2.

$$H_1 = \begin{bmatrix} 1/2 & 0 \\ 0 & 2 \end{bmatrix}, \quad H_2 = \begin{bmatrix} \sqrt{3}/2 & 1/2 \\ -1/2 & \sqrt{3}/2 \end{bmatrix}.$$

H_1 is the H_1 of Example 1, while H_2 effects clockwise rotation through 30°. The set $\{H_1, H_2\}$ is not contractive relative to the Euclidean norm, for neither matrix contracts (0, 1). The set is convergent, however, for if x is in the cone C_1 described in the earlier example, we may apply H_1 to contract x by a factor of β. If x is not in C_1, then the application of H_2 no more than 11 times results in a vector of the same length as x and lying in C_1, and so a subsequent application of H_1 contracts by a factor of β, and the set $\{H_1, H_2\}$ is convergent. We are thus led to the following definition and theorem:

DEFINITION 3. Let $\|\cdot\|_v$ be any norm on R^n. The set $\{H_1, H_2, \cdots, H_N\}$ of $n \times n$ matrices is *pre-contractive relative to* $\|\cdot\|_v$ provided that, if $x \in R^n$, $x \neq 0$, there is a finite sequence $\{q(x)_i\}_{i=1}^{n(x)}$, $1 \leqq q(x)_i \leqq N$, such that

$$\left\|\left(\prod_{i=n(x)}^{1} H_{q(x)_i}\right)x\right\|_v < \|x\|_v.$$

THEOREM 1. *Let $\|\cdot\|_v$ be any norm on R^n, and let $K = \{H_1, H_2, \cdots, H_N\}$ be a set of $n \times n$ matrices. Then K is pre-contractive relative to $\|\cdot\|_v$ if and only if K is convergent.*

The proof of Theorem 1 depends on the following two lemmas, each of which can be verified by standard arguments based on the compactness of the unit sphere in R^n. These arguments are omitted here.

LEMMA 1.1. *If K is pre-contractive relative to $\|\cdot\|_v$, then there is a positive integer M such that, if $x \in R^n$ and $x \neq 0$, there is a finite sequence $\{q(x)_i\}_{i=1}^{n(x)}$, $1 \leqq q(x)_i \leqq N$, such that*

$$\left\|\left(\prod_{i=n(x)}^{1} H_{q(x)_i}\right)x\right\|_v < \|x\|_v \quad \textit{and} \quad n(x) \leqq M.$$

LEMMA 1.2. *If K is contractive relative to $\|\cdot\|_v$, then there is a $\beta<1$ such that, if $x \in R^n$ and $x \neq 0$, there is an i, $1 \leqq i \leqq N$, such that*

$$\|H_i x\|_v \leqq \beta \|x\|_v.$$

Proof of Theorem 1. It is immediate that convergence of K implies pre-contractiveness of K relative to $\|\cdot\|_v$.

Assume K pre-contractive relative to $\|\cdot\|_v$. Let M be the positive integer given in Lemma 1.1. Then the set

$$\hat{K} = \left\{ \prod_{i=k}^{1} H_{p_i} \middle| k \leqq M, 1 \leqq p_i \leqq N \right\}$$

is contractive relative to $\|\cdot\|_v$. Choose $\beta<1$ for $\hat{K}$ using Lemma 1.2. Then we have, for each x in R^n, a sequence $\{q(x)_i\}_{i=1}^{n(x)}$ with

$$\left\| \left(\prod_{i=n(x)}^{1} H_{q(x)_i} \right) x \right\|_v \leqq \beta \|x\|_v \quad \text{and} \quad n(x) \leqq M.$$

To show K convergent, select $x \in R^n$ and define

$$p(x)_i = q(x)_i \quad \text{for } 1 \leqq i \leqq n(x).$$

Let $y_0 = x$ and $y_1 = (\prod_{i=n(y_0)}^{1} H_{q(y_0)_i}) y_0$.

Define

$$p(x)_i = q(y_1)_{i-n(y_0)}, \qquad n(y_0)+1 \leqq i \leqq n(y_0)+n(y_1).$$

Let $y_2 = (\prod_{i=n(y_1)}^{1} H_{q(y_1)_i}) y_1$, and continue in this way to form the sequence $\{p(x)_k\}_{k=1}^{\infty}$. It is clear from this construction that if $k = n(y_0)+n(y_1)+\cdots+n(y_l)$ for some l, then

$$\left\| \left(\prod_{i=k}^{1} H_{p(x)_i} \right) x \right\|_v \leqq \beta^l \|x\|_v. \tag{7}$$

To conclude the proof, we must bound the left side of (7) for

$$n(y_0)+n(y_1)+\cdots+n(y_l) < k < n(y_0)+n(y_1)+\cdots+n(y_l)+n(y_{l+1}).$$

Suppose k satisfies this inequality. Let $s = n(y_0)+n(y_1)+\cdots+n(y_l)$, and let $t = k-s$. Since $n(y_{l+1}) \leqq M$, we have $t<M$. Let $\|\cdot\|_0$ be the operator norm on $n \times n$ matrices generated by $\|\cdot\|_v$, and let

$$B = \max \{\|H_1\|_0, \|H_2\|_0, \cdots, \|H_N\|_0\}.$$

Then

$$\begin{aligned} \left\| \left(\prod_{i=k}^{1} H_{p(x)_i} \right) x \right\|_v &\leqq \prod_{i=k}^{s+1} \|H_{p(x)_i}\|_0 \left\| \left(\prod_{i=s}^{1} H_{p(x)_i} \right) x \right\|_v \\ &\leqq B^t \beta^l \|x\|_v < B^M \beta^l \|x\|_v. \end{aligned}$$

As $k \to \infty$, we have $l \to \infty$ so K is convergent and Theorem 1 is proved.

COROLLARY 1.1. *Pre-contractiveness is norm-independent.*

Proof. The proof is immediate.

COROLLARY 1.2. *If K is contractive relative to $\|\cdot\|_v$, then K is exponentially convergent relative to $\|\cdot\|_v$, in the sense that, for some number Γ and some β, $0<\beta<1$, we have, for each $x \in R^n$, a sequence $\{p(x)_i\}_{i=1}^{\infty}$ so that*

$$\left\| \left(\prod_{i=k}^{1} H_{p(x)_i} \right) x \right\|_v \leqq \Gamma \beta^k \|x\|_v \quad \textit{for each } k.$$

Proof. The M in the proof of Theorem 1 can be taken as 1 when K is contractive relative to $\|\cdot\|_v$, so each $n(y_i)$ is 1 and (7) holds, with $l=k$, for each k. Thus K is exponentially convergent with $\Gamma=1$.

We conclude this section with two examples which answer the questions: (a) Is contractiveness norm-dependent? (b) Does pre-contractiveness imply contractiveness relative to some norm?

(a) Contractiveness, unlike pre-contractiveness, is norm-dependent, as is seen by the following example.

Example 3. Let

$$Q=\begin{bmatrix}1/2 & 0\\ 0 & 1\end{bmatrix} \quad\text{and}\quad \|x\|_Q=(x^TQx)^{1/2}=(\tfrac{1}{2}x_1^2+x_2^2)^{1/2}$$

for all $x\in R^2$. If $H_1=\begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}$ and $K=\{H_1\}$, it can easily be seen that K is contractive relative to $\|\cdot\|_Q$, but not contractive relative to the Euclidean norm on R^2.

(b) There is a set of matrices which is pre-contractive, but not contractive relative to any norm.

To justify this statement we need the following theorem, whose proof is straightforward and is omitted.

THEOREM 2. *Suppose* $\{\alpha_1,\alpha_2,\cdots,\alpha_n\}$ *is a basis for* R^n, $K=\{H_1,H_2,\cdots,H_n\}$ *is a set of* $n\times n$ *matrixes, and* $\lambda_1,\lambda_2,\cdots,\lambda_n$ *are numbers such that*

$$H_j\alpha_i=\begin{cases}\lambda_j\alpha_j, & i=j,\\ 0\ , & i\neq j.\end{cases}$$

Suppose $\|\cdot\|$ *is a norm on* R^n *and* K *is contractive relative to* $\|\cdot\|$. *For* $x\in R^n$, $x=x_1\alpha_1+x_2\alpha_2+\cdots+x_n\alpha_n$, *let*

$$\|x\|^*=|x_1|\,\|\alpha_1\|+|x_2|\,\|\alpha_2\|+\cdots+|x_n|\,\|\alpha_n\|.$$

Then 1. $\|\cdot\|^*$ *is a norm on* R^n

2. $\|\alpha_j\|^*=\|\alpha_j\|$, $1\leqq j\leqq n$

3. $\|x\|\leqq\|x\|^*$, $x\in R^n$.

4. K *is contractive relative to* $\|\cdot\|^*$.

Geometrically, the unit sphere of $\|\cdot\|^*$ is the convex hull of vectors $\pm(1/\|\alpha_i\|)\alpha_i$, $1\leqq i\leqq n$.

Example 4. Let

$$H_1=\begin{bmatrix}3 & 0\\ 0 & 0\end{bmatrix},\qquad H_2=\begin{bmatrix}0 & 0\\ 0 & 4\end{bmatrix},\qquad K=\{H_1,H_2\}.$$

K is pre-contractive, since $H_1H_2=0$. Suppose there is a norm $\|\cdot\|$ on R^2 relative to which K is contractive. Let

$$\alpha_1=\begin{bmatrix}1\\ 0\end{bmatrix},\qquad \alpha_2=\begin{bmatrix}0\\ 1\end{bmatrix}.$$

Then $H_1\alpha_1=3\alpha_1$, $H_2\alpha_2=4\alpha_2$, and $H_1\alpha_2=H_2\alpha_1=0$. Thus, by Theorem 4,

$$\left\|\begin{bmatrix}x\\ y\end{bmatrix}\right\|^*=|x|\,\|\alpha_1\|+|y|\,\|\alpha_2\|$$

defines a norm on R^2 agreeing with $\|\cdot\|$ for α_1 and α_2, and relative to which K is contractive. Thus, for any $x>0$, either

$$\left\|H_1\begin{bmatrix}x\\1\end{bmatrix}\right\|^* < \left\|\begin{bmatrix}x\\1\end{bmatrix}\right\|^* \quad\text{or}\quad \left\|H_2\begin{bmatrix}x\\1\end{bmatrix}\right\|^* < \left\|\begin{bmatrix}x\\1\end{bmatrix}\right\|^*,$$

which simplifies to:

either $$x<\frac{\|\alpha_2\|}{2\|\alpha_1\|} \quad\text{or}\quad x>\frac{3\|\alpha_2\|}{\|\alpha_1\|}.$$

Since this holds for all $x>0$, we must have $\|\alpha_2\|/(2\|\alpha_1\|)>3\|\alpha_2\|/\|\alpha_1\|$, which implies $\frac{1}{2}>3$. This contradiction establishes that there is no norm on R^2 relative to which K is contractive.

3. A criterion for contractiveness—positive definite sets of matrices. In [2], P. Stein proves the following theorem.

THEOREM. *If H is a real or complex square matrix, a necessary and sufficient condition that* $\lim_{n\to\infty} H^n=0$ *is that there exists a positive definite Hermitian matrix Q for which $Q-H^*QH$ is positive definite. If H is real, Q may be taken real and symmetric.*

Now $\lim_{n\to\infty} H^n=0$ if and only if $\{H\}$ is a contractive set relative to some norm. In Stein's theorem, the statement that $Q-H^*QH$ is positive definite implies that

$$x^*(Q-H^*QH)x>0, \quad\text{or}\quad x^*H^*QHx<x^*Qx, \quad \text{for } x\neq 0.$$

This suggests the following criterion for contractiveness.

DEFINITION 4. If Q is a positive definite symmetric matrix, $\|x\|_Q$ denotes the norm $(x^TQx)^{1/2}$.

DEFINITION 5. The set $W=\{A_1, A_2, \cdots, A_N\}$ of symmetric $n\times n$ matrices is a *positive definite set* provided that, if $x\in R^n$, $x\neq 0$, then there is an i, $1\leqq i\leqq N$, such that $x^TA_ix>0$.

THEOREM 3. *Let $K=\{H_1, H_2, \cdots, H_N\}$ be a set of $n\times n$ matrices, and let Q be an $n\times n$ positive definite symmetric matrix. Form the set $W=\{Q-H_1^TQH_1, Q-H_2^TQH_2, \cdots, Q-H_N^TQH_N\}$. Then K is contractive relative to $\|\cdot\|_Q$ if and only if W is a positive definite set.*

Proof. Clearly $\|H_ix\|_Q<\|x\|_Q$ if and only if $x^TH_i^TQH_ix<x^TQx$; that is, $x^T(Q-H_i^TQH_i)x>0$. The theorem follows.

We now present a geometric test for positive definiteness of a set of symmetric matrices. For $0\leqq p$, $0\leqq m$, and $p+m\leqq n$, let

$$S(p,m)=\left\{y\in R^n \,\middle|\, \sum_{j=1}^{p} y_j^2>\sum_{j=p+1}^{p+m} y_j^2\right\}.$$

$S(p,m)$ is a cone in R^n, in the sense that it is closed under multiplication by nonzero scalars. Notice that if $m=0$, $S(p,m)$ is the complement in R^n of an $(n-p)$-dimensional subspace of R^n.

THEOREM 4. *Suppose that A is an $n\times n$ symmetric matrix, and that $D=P^TAP$, where P is nonsingular and D is the unique matrix conjugate to A having the canonical form*

$$D=\operatorname{Diag}\{1, 1, \cdots, 1, -1, -1, \cdots, -1, 0, 0, \cdots, 0\}.$$

If D contains p 1's and m -1's on its diagonal (that is the index *of A is p and the* co-index *of A is m), then*

$$\{x\in R^m \mid x^TAx>0\}=P(S(p,m)),$$

where $P(S(p,m))$ denotes $\{Py \mid y\in S(p,m)\}$.

Proof. Let $y=P^{-1}x$. It is easily checked that $x^TAx>0$ if and only if $y^TDy>0$, and that this occurs if and only if $y\in S(p,m)$; that is, $x\in P(S(p,m))$.

COROLLARY 4.1. *If $W=\{A_1,A_2,\cdots,A_N\}$ is a set of $n\times n$ symmetric matrices and, for $1\leqq i\leqq N$, $D_i=P_i^TA_iP_i$ is the canonical form of A_i, and if the index and co-index of A_i are p_i and m_i respectively, then W is a positive definite set if and only if*

$$\bigcup_{i=1}^{N} P_i(S(p_i,m_i))=R^n-\{0\}.$$

An $n\times n$ positive definite matrix has all eigenvalues positive (its index is n). A generalization of this to positive definite sets is that a positive definite set of $n\times n$ symmetric matrices must have the property that the number of positive eigenvalues of all the matrices in the set (counting multiplicities) must be greater than or equal to n. This is the implication of Theorem 5, the proof of which uses Corollary 4.1 and the following dimensional inequality:

LEMMA 5.1. *If $V_1,V_2,\cdots,V_N$ are subspaces of R^n, then*

$$\dim\left(\bigcap_{i=1}^{N}V_i\right)\geqq\left(\sum_{i=1}^{N}\dim(V_i)\right)-(N-1)n.$$

Proof. Let $d_i=\dim(V_1)$ and $d_0=\dim(\bigcap_{i=1}^{N}V_i)$. For $1\leqq i\leqq N$, let A_i be an $(n-d_i)\times n$ matrix having null-space V_i. Let A be the block matrix

$$A=\begin{bmatrix}A_1\\A_2\\\vdots\\A_N\end{bmatrix}.$$

A is $(\sum_{i=1}^{N}(n-d_i))\times n$; that is, A is $(Nn-\sum_{i=1}^{N}d_i)\times n$. Clearly, for $x\in R^n$, $Ax=0$ if and only if $A_ix=0$ for $i=1,2,\cdots,N$, so $d_0=\text{nullity}(A)$. Now $Nn-\sum_{i=1}^{N}d_i\geqq\text{rank}(A)=n-d_0$, so $d_0\geqq\sum_{i=1}^{N}d_i-(N-1)n$, and this proves the lemma.

THEOREM 5. *If $W=\{A_1,A_2,\cdots,A_N\}$ is a positive definite set of symmetric matrices and the index of A_i is p_i for $1\leqq i\leqq N$, then*

$$\sum_{i=1}^{N}p_i\geqq n.$$

Proof. Let $D_i=P_i^TA_iP_i$ be the canonical form of A_i, and let m_i be the co-index of A_i. Corollary 4.1 gives

$$\bigcup_{i=1}^{N} P_i(S(p_i,m_i))=R^n-\{0\}.$$

Thus $\bigcap_{i=1}^{N}P_i(S(p_i,m_i))^c=\{0\}$, where "$c$" denotes complement in R^n. Now for each i, $S(p_i,m_i)\subset S(p_i,0)$, so $P_i(S(p_i,m_i))\subset P_i(S(p_i,0))$ and $P_i(S(p_i,0))^c\subset P_i(S(p_i,m_i))^c$.

Thus

$$\bigcap_{i=1}^{N}P_i(S(p_i,0))^c\subset\bigcap_{i=1}^{N}P_i(S(p_i,m_i))^c=\{0\}. \tag{8}$$

But $S(p_i,0)^c$ is a subspace of R^n of dimension $n-p_i$. Since P_i is nonsingular, $P_i(S(p_i,0))^c=P_i(S(p_i,0)^c)$ is a subspace of R^n of dimension $n-p_i$. Thus, using (8) and

Lemma 5.1,

$$0=\dim\left(\bigcap_{i=1}^{N} P_i(S(p_i,0))^c\right) \geqq \left(\sum_{i=1}^{N}(n-p_i)\right)-(N-1)n=n-\sum_{i=1}^{N} p_i,$$

so $\sum_{i=1}^{N} p_i \geqq n$.

4. Selection of feedback matrices. We consider in this section the following problem: suppose that the matrices A and B for the system $\dot{x}=Ax+Bu$ have been given, and that time-interval lengths $S_1, S_2, \cdots, S_N$ have been determined for a corresponding multi-rate sampled-data system. Then the matrices $C_i=e^{S_iA}$ and $D_i=\int_0^{S_i} e^{vA}\,dv\,B$ are determined. How can feedback matrices $F_1, F_2, \cdots, F_N$ be selected so that the set

$$K=\{H_i=C_i+D_iF_i \mid i=1,2,\cdots,N\}$$

is convergent? It would be useful to know when K can, in fact, be made contractive relative to some norm, since this implies exponential convergence (Corollary 1.2). Theorem 6 describes how the choice of the feedback matrix F_i can control distances between the column space of H_i and any other given subspace of R^n. By choosing this subspace to be $\{0\}$, we can make K contractive if that is possible.

We will employ the following notation: throughout this section Q is an $n\times n$ symmetric positive definite matrix. For x in R^n,

$$\|x\|_Q \quad \text{denotes the norm} \quad (x^TQx)^{1/2}.$$

If A is an $n\times m$ matrix, $\mathrm{Row}_i(A)$ and $\mathrm{Col}_i(A)$ denote the ith row and ith column of A, $NS(A)$ and $CS(A)$ denote the nullspace and column space of A, and for y in R^n, $\mathrm{dist}_Q(y, A)$ denotes the distance in $\|\cdot\|_Q$ from y to $CS(A)$. A^+ will denote the Moore–Penrose pseudoinverse of A. Finally, if W is a subspace of R^n,

$$W^{\perp Q}=\{y\in R^n \mid y^TQx=0 \text{ for all } x\in W\}.$$

LEMMA 6.1. *If A is $n\times m$ and $W=CS(A)$, then*

$$W^{\perp Q}=NS(A^TQ).$$

Proof. If $y\in W^{\perp Q}$ then $\mathrm{Row}_i(A^T)Qy=\mathrm{Col}_i(A)^TQy=0$ for $i=1,2,\cdots,n$. Thus $A^TQy=0$. Conversely, if $y\in NS(A^TQ)$, we have $(Az)^TQy=z^TA^TQy=0$ for all $z\in R^n$, so $y^TQx=x^TQy=0$ for all $x\in W$.

LEMMA 6.2. *If A is $n\times m$, $W=CS(A)$, and $z\in R^n$, then $z=z_W+z_{W^{\perp Q}}$, where*

$$z_W=A(A^TQA)^+A^TQz \quad \textit{is in } W,$$

and

$$z_{W^{\perp Q}}=(I-A(A^TQA)^+A^TQ)z \quad \textit{is in } W^{\perp Q}.$$

Proof. Since $R^n=W\oplus W^{\perp Q}$ there are (unique) $\alpha\in W$ and $\beta\in W^{\perp Q}$ such that $z=\alpha+\beta$. Let $v\in R^n$ such that $\alpha=Av$. Since $\beta\in W^{\perp Q}$

$$A^TQ(z-Av)=A^TQ\beta=0.$$

Thus the equation $A^TQAx=A^TQz$ has the solution $x=v$, and so

$$u=(A^TQA)^+A^TQz$$

is a solution of the same equation. This implies that $z-Au\in NS(A^TQ)=W^{\perp Q}$.

Clearly, $Au \in W$, so we have

$$z = Au + (z - Au),$$

with $Au = A(A^TQA)^+A^TQz = z_W$ in W and $z - Au = (I - (A(A^TQA)^+A^TQ)z) = z_{W^{\perp Q}}$ in $W^{\perp Q}$.

THEOREM 6. *Suppose C is $n \times n$ and nonsingular, D is $n \times m$, and V is $n \times k$. Let $L = I - V(V^TQV)^+V^TQ$, $E = LD$, and $F = -(E^TQE)^+E^TQLC$, and let G be any $n \times m$ matrix. Then, for each x in R^n,*

$$\operatorname{dist}_Q((C+DF)x, V) \leqq \operatorname{dist}_Q((C+DG)x, V).$$

Proof. For any y in R^n and any $n \times l$ matrix A, $\operatorname{dist}_Q(y, A) = Q-$length of the Q-orthogonal projection of y onto $CS(A)^{\perp Q}$. By Lemma 6.2, then

$$\operatorname{dist}_Q(y, A) = \|y_{CS(A)^{\perp Q}}\|_Q = \|(I - A(A^TQA)^+A^TQ)y\|_Q.$$

Thus $\operatorname{dist}_Q((C+DG)x, V) = \|L(C+DG)x\|_Q$. Let $T = GC^{-1}$, $y = Cx$, and $W = CS(E)$. Then $L(C+DG)x = L(I+DT)y = [Ly]_{W^{\perp Q}} + [Ly]_W + ETy$. Since the first term is in $W^{\perp Q}$ and the sum of the second and third terms is in W, we obtain

$$\operatorname{dist}_Q((C+DG)x, V)^2 = \|L(C+DG)x\|_G^2 = \|[Ly]_{W^{\perp Q}}\|_Q^2 + \|[Ly]_W + ETy\|_Q^2.$$

The first of these terms is independent of G, and the second becomes 0 when $G = F$, as can be easily checked using Lemma 6.2. Thus

$$\operatorname{dist}_Q((C+DG)x, V) \quad \text{is minimized by} \quad G = F.$$

By choosing V to be the $n \times 1$ matrix of zeros, we obtain the following corollary concerning feedback selection in the multi-rate sampled-data system. Notice that in that system, each $C_i = e^{S_iA}$ is nonsingular.

COROLLARY 6.1. *If C_i is $n \times n$ and nonsingular, and D_i is $n \times m$ for $i = 1, 2, \cdots, N$, and if there exist feedback matrices $F_1, F_2, \cdots, F_N$ such that the set*

$$K = \{H_i = C_i + D_iF_i \mid i = 1, 2, \cdots, N\}$$

is contractive relative to $\|\cdot\|_Q$, then the choice

$$F_i = -(D_i^TQD_i)^+D_iQC_i$$

makes K contractive relative to $\|\cdot\|_Q$. In addition, letting $Q = I$, we obtain

COROLLARY 6.2. *In the notation of Corollary 6.1, if any choice of feedbacks makes K contractive relative to the Euclidean norm, then*

$$F_i = -D_i^+C_i \quad \textit{does so.}$$

We conclude this section with a theorem combining Corollary 6.1 with Theorem 3 to relate "contractibility" of a set of C_i's and D_i's to positive definiteness of a set of symmetric matrices. Recall that, since Q is symmetric positive definite, $Q = U^TU$ for some nonsingular matrix U.

LEMMA 7.1. *If C is $n \times n$, D is $n \times m$, $F = -(D^TQD)^+D^TQC$, and $H = C + DF$, then*

$$Q - H^TQH = U^TU - (UC)^T[I - (UD)(UD)^+](UC).$$

Proof. Let $J = Q - H^TQH = Q - (C+DF)^TQ(C+DF)$. Then $J = Q - [(I - D \cdot (D^TQD)^+D^TQ)C]^TQ[(I - D(D^TQD)^+D^TQ)C]$. Algebraic manipulation produces $J = Q - C^TQC + C^TQD(D^TQD)^+D^TQC + C^TQD(D^TQD)^+D^TQC - C^TQD \cdot \{(D^TQD)^+D^TQD(D^TQD)^+\}D^TQC$. Since the quantity in $\{\cdot\}$ reduces to $(D^TQD)^+$, we obtain

$$J = Q + C^T[QD(D^TQD)^+D^TQ - Q]C. \tag{9}$$

Now

$$QD(D^TQD)^+D^TQ = U^T(UD)[(UD)^T(UD)]^+(UD)^TU. \tag{10}$$

The properties of the Moore–Penrose pseudo-inverse (see [1]) give

$$((UD)^T)^+(UD)^T(UD) = UD, \quad \text{and} \quad (UD)^T(UD)(UD)^+ = (UD)^T.$$

Thus $((UD)^T)^+(UD)^T(UD)(UD)^T(UD)(UD)^T(UD)(UD)^+ = (UD)(UD)^T(UD)\cdot(UD)^T$. By [1, p. 18], this is sufficient for

$$[(UD)^T(UD)]^+ = (UD)^+((UD)^T)^+.$$

Then by (10),

$$\begin{aligned} QD(D^TQD)^+D^TQ &= U^T(UD)(UD)^+((UD)^T)^+(UD)^TU \\ &= U^T[(UD)(UD)^+][(UD)(UD)^+]^TU = U^T[(UD)(UD)^+]U. \end{aligned}$$

Thus by (9),

$$\begin{aligned} J &= U^TU + C^TU^T[(UD)(UD)^+]UC - C^TU^TUC \\ &= U^TU - (UC)^T[I - (UD)(UD)^+](UC), \end{aligned}$$

which was to be shown.

THEOREM 7. *Suppose C_i is $n \times n$ nonsingular and D_i is $n \times m$ for $i = 1, 2, \cdots, N$. Let*

$$J_i = U^TU - (UC_i)^T[I - (UD_i)(UD_i)^+](UC_i), \qquad i = 1, 2, \cdots, N,$$

and let $W = \{J_1, J_2, \cdots, J_N\}$. Then these statements are equivalent:

(a) *There exist $m \times n$ matrices $F_1, F_2, \cdots, F_N$ such that*

$$K = \{H_i = C_i + D_iF_i \mid i = 1, 2, \cdots, N\}$$

is contractive relative to $\|\cdot\|_Q$.

(b) *W is a positive definite set.*

Proof. Suppose (a) holds. By Corollary 6.1, K is contractive relative to $\|\cdot\|_Q$ for the choice $F_i = -(D_i^TQD_i)^+D_i^TQC_i$, $i = 1, 2, \cdots, N$. By Lemma 7.1, $Q - H_i^TQH_i = J_i$, and so by Theorem 3, W is a positive definite set. Clearly, if (b) is assumed true, the choice $F_i = -(D_i^TQD)^+D_i^TQC_i$ produces a set K contractive relative to $\|\cdot\|_Q$.

The "cone of contraction" $P(S(p, m))$ described in Theorem 4 is centered about the subspace of R^n containing $(x_1, x_2, \cdots, x_p, 0, \cdots, 0)$. It is to be hoped that, when contractiveness cannot be achieved, Theorem 6 will be useful in achieving pre-contractiveness by enabling us to aim the column space of H_i into a cone of contractiveness for H_j.

Acknowledgment. The author expresses his thanks to Dr. Alper Caglayan and to Dr. Douglas Price of NASA-Langley for bringing the problem to his attention and for many helpful discussions.

REFERENCES

[1] T. L. BOULLION AND P. L. ODELL, *Generalized Inverse Matrices*, John Wiley, New York, 1971.

[2] P. STEIN, *Some general theorems on iterants*, J. Res. Nat. Bur. Standards, 48 (1952), pp. 82–83.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 3, May 1979

ON THE CONSTRUCTION OF NONAUTONOMOUS STABLE SYSTEMS WITH APPLICATIONS TO ADAPTIVE IDENTIFICATION AND CONTROL*

Dedicated to Lt. Col. Samuel Lindsay, upon the occasion of his retirement from the U.S. Army.

A. P. MORGAN†

Abstract. In this paper, the uniform asymptotic stability of nonautonomous ordinary differential equations with Lyapunov functions whose derivatives are negative semidefinite is studied. A general framework for constructing and analyzing such systems is established, and applications to adaptive schemes for identification and control are described. Specific rates of convergence and robustness estimates are also given.

1. Introduction. In some recent papers on adaptive identification and control, the following method of developing asymptotically stable linear nonautonomous systems has been used. First, system elements which can be chosen at the discretion of the controller are specified so that the overall system has a Lyapunov function with a negative semidefinite derivative. Then asymptotic stability is implied by an "excitedness" property. A fundamental related mathematical idea is periodicity, and Lyapunov theorems for the asymptotic stability of periodic systems are fundamental to a number of these papers. (See P. M. Lion [15], Narendra and Kudva [22], and Narendra and McBride [23].)

The purpose of this paper is to provide mathematical results that will allow the same basic method of constructing asymptotically stable systems for nonlinear and nonperiodic cases. Further, rate of convergence and robustness estimates are also provided. The proofs given here are elementary and constructive, and hopefully this will expedite the development of computer algorithms for the control applications.

Previous mathematical work relating most directly to the results in this paper is in LaSalle [12], [13], Burton [4], Haddock [6], [7], [8], Morgan and Narendra [19], [20], and Morgan [18]. The control theory inspiration comes from such papers as Lion [15]; Narendra and McBride [23], Yuan and Wonham [26], and Narendra and Kudva [22]. The recent papers by Artstein [2], Anderson [1], Kreisselmeier [11], and Nuyan and Carroll [24] are also relevant. The material developed here logically follows Morgan and Narendra [19] and [20], providing a unified and general framework in which the results in those two papers can be developed.

In § 2 necessary preliminary definitions are stated. Section 3 contains the main results of the paper, the key result being Theorem 3. Sections 4 and 5 develop special cases, while § 6 is devoted to a discussion of generalizations and variations on the § 3 material. An outline of the specific control theory ideas which motivated the previous material is given in § 7. The last section contains a technical result.

2. Definitions and a theorem. In this section a number of definitions and a theorem are presented.

2.1. Notation and conventions.

DEFINITION 1 (Notation). (a) R^n denotes n-dimensional Euclidean space. R^+ denotes the nonnegative real numbers.

* Received by the editors December 15, 1976, and in final revised form July 17, 1978.

† Division of Systems and Computer Services, Medical College of Georgia, Augusta, Georgia. Now at Mathematics Department, General Motors Research Laboratories, Warren, Michigan 48090.

(b) The open ball of radius ε about $0 \in R^n$ is denoted S_ε. The closure of a subset S of R^n is denoted $\bar{S}$. The annulus about 0 defined by $\varepsilon_1 > \varepsilon_2 > 0$ is $A_{\varepsilon_1,\varepsilon_2} = S_{\varepsilon_1} - \bar{S}_{\varepsilon_2}$. If C is a closed subset of R^n, then $S_\varepsilon(C) = \{x \in R^n | d(x, C) < \varepsilon\}$ where $d(x, C) = \inf\{|x - y| \,|\, y \in C\}$.

(c) K denotes the collection of all strictly increasing continuous functions $k: R^+ \to R^+$ with $k(0) = 0$. (See Hahn [9, p. 7].) KK_0 denotes the collection of all continuous functions $k: R^+ \times R^+ \to R^+$ with $k(t_1, 0) = 0$ for all t_1 and strictly increasing in the second variable with t_1 fixed.

K' and KK_0' denote the collections of functions as above but with "strictly increasing" replaced by "monotonically increasing".

(d) If $x \in R^1$, then $[x]$ denotes the smallest integer n such that $n > x$. ($[\cdot]$ is the "greater integer" function.) For example, $[3/2] = 2$ and $[1] = 2$.

DEFINITION 2. The following notation and basic assumptions will be fixed from now on. Let G be an open bounded subset of R^n containing 0. Assume f: $R^+ \times G \to R^n$ is measurable in t, continuous in x, and $f(t, 0) = 0$ for all $t \in R^+$. Consider the equation

$$\dot{x} = f(t, x). \tag{1}$$

Assume that G_0 is an open subset of G containing 0 such that, if $x(t)$ is a solution of (1) with initial condition in G_0, then $x(t)$ is continuable to $+\infty$. In other words, if $x(t_0) \in G_0$ for some $t_0 \in R^+$, then $x(t) \in G$ can be defined for all $t > t_0$.

DEFINITION 3. The smooth function V: $R^+ \times G \to R^+$ is a Lyapunov function for (1) at 0 if

1. $V(t, 0) = 0$ for all $t \in R^+$,
2. given annulus A, there is a positive constant b_0 such that $b_0 \geqq V(t, x) \geqq 0$ for all $(t, x) \in R^+ \times (A \cap G)$.
3. $\dot{V}(t, x) \leqq 0$ where $\dot{V}(t, x) = (\partial V/\partial t)(t, x) + (\partial V/\partial x)(t, x) f(t, x)$.

It is not assumed that V is positive definite. Also, note that if V is a Lyapunov function for (1) at 0, it does not necessarily follow that 0 is uniformly stable. (See Hahn [9] or Hale [10] for definitions of (uniform) stability and (uniform) asymptotic stability.) The abbreviation u.a.s. is used for uniform asymptotic stability.

If 0 is asymptotically stable and $G' \subseteq G$, then "G' is in the basin" means that solutions with initial conditions in G' go to 0 as $t \to \infty$.

2.2. Rates of convergence and persistence.

DEFINITION 4. The "rate of convergence" of (1) across annulus $A = A_{\varepsilon_1,\varepsilon_2}$ is the smallest positive number r such that, if $x(t)$ is a solution and $x(t_0) \in A$, then $x(t) \in S_{\varepsilon_2}$ for all $t \geqq t_0 + r$.

The *rate of persistence* of (1) in annulus A is the smallest positive number r such that, if $x(t)$ is a solution and $x(t_0) \in A$, then there is a $t_1 \in [t_0, t_0 + r]$ such that $x(t_1) \notin A$.

In other words, if r is the rate of persistence of (1) in A, then no solution can remain in A for r consecutive units of time. If 0 is uniformly stable, then the existence of finite rates of persistence for annuli about 0 implies that 0 is u.a.s. Further, in this case, rates of persistence can be used to derive rates of convergence.

For example, let $\varepsilon_2 > \varepsilon_2$ and suppose that solutions with initial points in S_{ε_1} never leave $S_{\varepsilon_1'}$ for some $\varepsilon_1' \geqq \varepsilon_1$ and solutions with initial points in $S_{\varepsilon_2'}$ never leave S_{ε_2} for some $\varepsilon_2' \leqq \varepsilon_2$. Suppose that r is an upper bound for the rate of persistence of (1) in annulus $A' = A_{\varepsilon_1',\varepsilon_2'}$. Let $x(t)$ be a solution with $|x(t_0)| < \varepsilon_1$. Since $x(t)$ must leave A' for some $t_1 \in [t_0, t_0 + r]$ and since $x(t)$ cannot go outside A', it must go inside. We conclude that if $|x(t_0)| < \varepsilon_1$, then $|x(t)| \leqq \varepsilon_2$ for all $t \geqq t_0 + r$. Thus an upper bound for the rate of persistence in A' is an upper bound for the rate of convergence across $A = A_{\varepsilon_1,\varepsilon_2}$.

If $V(x)=|x|^2$ is a Lyapunov function for (1) at 0, then $|x(t)|$ can never increase for any solution $x(t)$ and the rate of persistence in A coincides with the rate of convergence across A.

The linear case is particularly simple. Let $X(t, t_0)$ denote the fundamental solution, and suppose that $|X(t, t_0)| \leqq c_0$ for any t_0 and all $t \geqq t_0$. Let $\varepsilon_1 > \varepsilon_2 > 0$ and suppose that r is an upper bound for the rate of persistence of the equation in annulus $A_{c_0\varepsilon_1, \varepsilon_2/c_0}$. Then

$$|X(t, t_0)| \leqq (\varepsilon_1/\varepsilon_2)c_0\, e^{-L(t-t_0)}$$

for any t_0 and all $t \geqq t_0$ where $L=(\ln(\varepsilon_1/\varepsilon_2))/r$.

2.3. Uniform boundedness and continuity conditions. The following assumptions are used in several contexts. They will be cited as necessary.

Let $\beta: R^+ \times G \to R^n$. Then "$\dot\beta(t_0, x_0)$ exists" means β is smooth in an open subset of $R^+ \times G$ containing (t_0, x_0). In this case, we have $\dot\beta(t_0, x_0) = (\partial\beta/\partial t)(t_0, x_0) + (\partial\beta/\partial x)(t_0, x_0) f(t_0, x_0)$. (However, this definition of $\dot\beta$ can be extended to continuous β in the usual way. See Hale [10, p. 293].)

Assumption A. There is a $k \in K$ such that $|\beta(t, x)| \leqq k(|x|)$ for all $x \in G$, a.e. $t \in R^+$.

Assumption B. There is a $k \in K$ such that $|\beta(t, x) - \beta(t, y)| \leqq k(|x-y|)$ for $x, y \in G$, a.e. $t \in R^+$.

Assumption B$_0$. Assumption B with "for all $t \in R^+$" replacing "a.e. $t \in R^+$".

Assumption C. The function $\dot\beta(t, x)$ exists for all $(t, x) \in R^+ \times G$, and $\dot\beta$ obeys Assumption B.

Assumption D. Given annulus A, there is $\eta \in K$ such that, if $x(t)$ is a solution in A for $t \in [a, b]$, then $|\beta(b, x(b)) - \beta(a, x(a))| \leqq \eta(|b-a|)$.

Note that "β obeys a uniform Lipschitz condition" is equivalent to "β obeys Assumption B$_0$ with $k(s)=c_0 s$ for some constant c_0".

If $\dot\beta(t, x)$ exists for all (t, x), and, for each annulus A, there is a constant k_0 such that $|\dot\beta(t, x)| \leqq k_0$ for $(t, x) \in R^+ \times (A \cap G)$, then β obeys Assumption D. In this case,

$$|\beta(b, x(b)) - \beta(a, x(a))| = \left| \int_a^b \dot\beta(\tau, x(\tau))\, d\tau \right| \leqq \int_a^b |\dot\beta(\tau, x(\tau))|\, d\tau$$

$$\leqq \int_a^b k_0\, d\tau = k_0(b-a) \equiv \eta(b-a).$$

The following lemma will be used in the proof of Theorem 2.

LEMMA. *Assume W obeys Assumption* D *with $\eta \in K$ for annulus A. Suppose there is a solution $x(t) \in A$ for $t \in [a, b]$, and suppose there is a $t_* \in [a, b]$ and $\gamma > 0$ such that $W(t_*, x(t_*)) \geqq \gamma$. Then there is an interval $I \subseteq [a, b]$ such that $W(t, x(t)) \geqq \gamma/2$ for $t \in I$ and the length of I is at least* $\min\{(b-a)/2, \eta^{-1}(\gamma/2)\}$.

Proof. For $t, t_* \in [a, b]$, we have $|W(t, x(t)) - W(t_*, x(t_*))| \leqq \eta(|t-t_*|)$, implying $W(t, x(t)) \geqq \gamma - \eta(|t-t_*|)$. Thus $W(t, x(t)) \geqq \gamma/2$ as long as $\eta(|t-t_*|) \leqq \gamma/2$. Now, either $t_* - a \geqq (b-a)/2$ or $b - t_* \geqq (b-a)/2$. If $t_* - a \geqq (b-a)/2$, let

$$I = (a, t_*) \cap (t_* - \eta^{-1}(\gamma/2), t_*).$$

Otherwise, define

$$I = (t_*, b) \cap (t_*, t_* + \eta^{-1}(\gamma/2)).$$

2.4. Excitedness conditions. Let $\alpha: R^+ \times G \to R^n$, and suppose that Ω is a collection of measurable functions with domains closed intervals in R^+ and range G.

DEFINITION 5. Let T be a positive constant, and let ϕ_i and s_i be sequences of positive numbers with $s_i \to \infty$. Then "α is pointwise uniformly exciting (PTUE) with respect to Ω with ϕ_i, T, and s_i" means

1. $s_{i+1} - s_i \leqq T$ for all i,
2. given index i and $\omega \in \Omega$ with $\omega(t)$ defined for $t \in [s_i, s_{i+1}]$, there is a $t_0 \in [s_i, s_{i+1}]$ such that $|\alpha(t_0, \omega(t_0))| > \phi_i$.

DEFINITION 6. Let T be a positive constant, and let ϕ_i and s_i be sequences of positive numbers with $s_i \to \infty$. Then "α is uniformly exciting (UE) with respect to Ω with ϕ_i, T, and s_i" means

1. $s_{i+1} - s_i \leqq T$ for all i,
2. given index i and $\omega \in \Omega$ with $\omega(t)$ defined for $t \in [s_i, s_{i+1}]$, there is an interval $[a, b] \subseteq [s_i, s_{i+1}]$ such that

$$\left| \int_a^b \alpha(\tau, \omega(\tau))\, d\tau \right| > \phi_i.$$

If Ω is the collection of all constant functions on annulus A, then we usually say "with respect to constants in A" in place of "with respect to Ω" in using the above definitions.

If $\alpha(t, x) = B(t)x$ is linear, then it is convenient to use the following abbreviated statement: "$B(t)$ is PTUE (UE respectively)" means "$\alpha(t, x)$ is PTUE (UE respectively) with respect to constants of unit length".

The term "persistently exciting" is used by Yuan and Wonham in [26] to describe a property which is similar to the above but somewhat less general. Excitedness conditions are central to the results in Morgan and Narendra [19] and [20].

It is interesting to note that, if 0 is u.a.s., then $f(t, x)$ must be UE as follows.

THEOREM 1. *Assume that $f(t, x)$ obeys a uniform Lipschitz condition on $R^+ \times G$. If* 0 *is u.a.s. for* (1) *with basin containing G, then, for any annulus A, there is a $\phi_0 > 0$ such that $f(t, x)$ is UE with respect to constants in $A \cap G$ with $\phi_i = \phi_0$ for all i.*

Proof. See § 8.

COROLLARY 1. *In the above result, "UE" can be replaced by "PTUE".*

Proof. With the uniform Lipschitz condition, UE is equivalent to PTUE.

In Morgan and Narendra [19] and [20], it is shown in several uniformly bounded linear contexts that conditions equivalent to UE are necessary for u.a.s. Artstein, in a preprint [2] that I have just received, has established a result similar to Theorem 1.

3. Main results. Assume from now on that 0 is uniformly stable, V is a Lyapunov function for (1) at 0, and there is a $W: R^+ \times G \to R^+$ with $-\dot{V}(t, x) \geqq W(t, x) \geqq 0$ for all $(t, x) \in R^+ \times G$. Let $H_t = \{x \in G \mid W(t, x) = 0\}$.

The theorems to follow have the same basic structure: a (generalized) "distance from H_t" function $\alpha(t, x)$ is identified so that, given annulus A, there is a collection Ω of functions $\omega: [t_1, t_2] \to G$ and

(a) if $x(t)$ is a solution to (1) and $\alpha(t, x(t))$ is "small" for $t \in [t_1, t_2]$, then there is an $\omega \in \Omega$ that is "near" $x(t)$ for $t \in [t_1, t_2]$;

(b) α is PTUE with respect to Ω, or

(b′) $\dot{\alpha}$ is defined and UE with respect to Ω.

The u.a.s. of 0 for (1) follows.

We shall assume for the remainder of this section that W oveys Assumption D. In Theorem 2, it is shown that, under this assumption, if W is PTUE with respect to solutions to (1) in $A \cap G$ for all annuli A, then u.a.s. follows. Theorems 3 and 4 are corollaries to Theorem 2, and they are the main results of this paper. Theorems and Corollaries 5 and 6 are further consequences.

In § 6, we will see that Assumption D for W can be weakened or omitted if other conditions are strengthened. In particular, piecewise continuous W are allowable. In general, the results in this section can be extended in a number of ways. Section 6 sketches some of these.

THEOREM 2. *The equilibrium* 0 *for* (1) *is u.a.s. with* G_0 *in the basin if, for each annulus* A, *there is a* $\gamma>0$ *such that* W *is PTUE with respect to* Ω_A *with* $\phi_i=\gamma$ *for all* i, *where* Ω_A *denotes the solutions to* (1) *in* A.

Further, if $b_0 \geqq V(t,x)$ *for all* $(t,x)\in R^+\times(A\cap G)$, T *and* $\phi_i=\gamma$ *are as in the definition of PTUE, and* η *is as in the definition of Assumption* D, *then* $[2b_0/(\gamma d_0)]2T$ *is an upper bound for the rate of persistence of* (1) *in* A *where* $d_0=\min\{T,\eta^{-1}(\gamma/2)\}$. (Recall $[\cdot]$ denotes the greater integer function. See Definition 1(d).)

Note that this is an extension of the result of LaSalle for periodic systems by which, if $f(t,x)$ and $V(t,x)$ are periodic in t and $\dot V(t,x(t))\equiv 0$ for solution $x(t)$ implies $x(t)\equiv 0$, then u.a.s. follows. See LaSalle [12].

We may generalize Theorem 2 by supposing that

$$-\dot V(t,x)\geqq W(t,x)-\alpha(t)$$

where W obeys the hypothesis of the theorem and $\alpha\colon R^+\to R^+$ with constant $\alpha_0>0$ such that

$$\int_{t_0}^{t_0+T}\alpha(\tau)\,d\tau\leqq\alpha_0$$

for any $t_0\in R^+$ and $\gamma d_0>2\alpha_0$. In this case, the rate of persistence becomes $[2b_0/(\gamma d_0-2\alpha_0)]2T$. This result can be proven by a minor modification of the proof of Theorem 2 given below. It provides an approach for estimating the "robustness" or "structural stability" of the uniform asymptotic stability. It gives an estimate of the amount of degradation of the rate of persistence caused by disturbances bounded as above by α_0.

Proof. Let A be an annulus. Then we have T, $\phi_i=\gamma$, and $s_i\to\infty$ with $s_{i+1}-s_i\leqq T$ as in the definition of PTUE. Suppose $x(t)$ is a solution and $x(t)\in A$ for $t\in[t_0,t_0+2mT]=I$ for some $t_0\in R^+$ and some positive integer m. Let $I_k=[t_0+2(k-1)T,t_0+2kT]$ for $k=1,2,\cdots,m$. For each k, there is an index i_k such that $[s_{i_k},s_{i_k+1}]\subseteq I_k$. Therefore

$$\int_{I_k}W(\tau,x(\tau))\,d\tau\geqq\frac{\gamma d_0}{2}$$

for each k, by PTUE and the lemma in § 2. It follows that

$$\int_I W(\tau,x(\tau))\,d\tau\geqq\frac{m\gamma d_0}{2}.$$

Thus $b_0\geqq V(t_0,x(t_0))\geqq m\gamma d_0/2$, and therefore $2b_0/(\gamma d_0)\geqq m$. This puts the stated upper bound on the length of time that $x(t)$ can stay in A.

To apply Theorem 2, we need two definitions. These are central to the method of constructing stable systems being discussed.

DEFINITION 7. Suppose that, for each $t\in R^+$, there is a neighborhood of H_t, $N(H_t)$, and a continuous $\alpha_t\colon N(H_t)\to R^n$ such that $\alpha_t(x)=0$ if $x\in H_t$. Let $\alpha(t,x)$ be defined by $\alpha(t,x)=\alpha_t(x)$. Then the pair (W,α) is *admissible* if there is a $\delta\in KK_0$ and constant γ_0^* with $\infty\geqq\gamma_0^*>0$ such that

(a) if $0\leqq\gamma\leqq\gamma_0^*$, $x\in N(H_t)$, and $|\alpha(t,x)|>\delta(t,\gamma)$, then $W(t,x)>\gamma$;

(b) if $x\in G$ and $x\notin N(H_t)$, then $W(t,x)>\gamma_0^*$.

Example 1. Suppose $W(t,x)=w(Q(t,x))$ where $w\in K'$ and $Q\colon R^+\times G\to R^+$.

Then we can take $\alpha(t,x)=Q(t,x)$, $\delta(t,\gamma)=w^{-1}(\gamma)$, $\gamma_0^*=\infty$. One case of interest is $Q(t,x)=|f(t,x)|$. This occurs, for example, when $f(t,x)=-Q(t)x$ where $Q(t)$ is a positive semidefinite matrix. (See Theorem 5 and Corollary 5 below.)

Example 2. Consider the two dimensional linear system

$$\dot{x}_1=-ax_1-b(t)x_2,$$
$$\dot{x}_2=b(t)x_1$$

where a is a positive constant and $b\colon R^+\to R^1$. With $V(x)=\frac{1}{2}(x_1^2+x_2^2)$, we have $-\dot{V}(t,x)=ax_1^2$. Thus $W(t,x)=ax_1^2$, $H_t=H_0=\{(0,x_2)\in R^2\}$, $N(H_t)=R^2$, $\alpha(t,x)=x_1$, $\delta(t,\gamma)=\sqrt{\gamma/a}$, $\gamma_0^*=\infty$. (See Corollary 6.)

Example 3. Suppose there is a retraction $\pi_t\colon N(H_t)\to H_t$ of a neighborhood of H_t onto H_t. (In other words, π_t is continuous and $\pi_t(x)=x$ if $x\in H_t$.) Then we let $\alpha_t(x)=x-\pi_t(x)$, and $N(H_t)$, $\delta(t,\gamma)$, and γ_0^* are defined depending on the geometry of H_t. (See §§ 4 and 5.)

From now on, let Ω denote a collection of measurable functions with domains intervals in R^+ and range G. Suppose (W,α) is admissible with $N(H_t)$, δ, and γ_0^* as above.

DEFINITION 8. Let annulus A and Ω be given. Then "Ω approximates solutions to (1) in A near H_t" means there are positive constants T and γ_1^*, a sequence of constants $s_i\to\infty$, and a sequence of functions $\theta_i\in KK_0$ such that

(a) $s_{i+1}-s_i\leqq T$ for all i, and

(b) if $x(t)$ is a solution to (1) with $x(t)\in N(H_t)\cap A$ for all $t\in[s_i,s_{i+1}]$ for some index i and $|\alpha(t,x(t))|\leqq\delta(t,\gamma)$ for all $t\in[s_i,s_{i+1}]$, then there is an $\omega\in\Omega$ with $|x(t)-\omega(t)|\leqq\theta(t,\gamma)$ for all $t\in[s_i,s_{i+1}]$.

Continuing the previous set of examples we have the following:

Example 1 (continued). Let $Q(t,x)=|f(t,x)|$, Ω be the constants in annulus A, and

$$\theta_i(t,\gamma)=\int_{s_i}^{t}\delta(\tau,\gamma)\,d\tau=w^{-1}(\gamma)(t-s_i).$$

Suppose $x(t)$ is a solution in A with $|\alpha(t,x(t))|\leqq\delta(t,x(t))$ for $t\in[s_i,t]$. Then

$$|x(t)-x(s_i)|\leqq\int_{s_i}^{t}|\dot{x}(\tau)|\,d\tau=\int_{s_i}^{t}|f(\tau,x(\tau))|\,d\tau$$
$$\leqq\int_{s_i}^{t}\delta(\tau,\gamma)\,d\tau=\theta_i(t,\gamma).$$

The $\omega\in\Omega$ for $x(t)$ is thus $\omega(t)\equiv x(s_i)$. (We may take $\gamma_1^*=\infty$.)

Example 2 (continued). Let Ω be the constants in $H_0\cap A$ and

$$\theta_i(t,\gamma)=\sqrt{\gamma/a}\left(3+\int_{s_i}^{t}|b(\tau)|\,d\tau\right).$$

Suppose $x(t)$ is a solution in A with $|\alpha(t,x(t))|\leqq\delta(t,\gamma)$ for $t\in[s_i,t]$. Thus $|x_1(t)|\leqq\sqrt{\gamma/a}$. Then

$$|x_2(t)-x_2(s_i)|\leqq\int_{s_i}^{t}|\dot{x}_2(\tau)|\,d\tau=\int_{s_i}^{t}|b(\tau)x_1(\tau)|\,d\tau$$
$$\leqq\sqrt{\gamma/a}\int_{s_i}^{t}|b(\tau)|\,d\tau.$$

Thus

$$\begin{aligned}|x(t)-x(s_i)| &\leqq |x_1(t)-x_1(s_i)|+|x_2(t)-x_2(s_i)|\\ &\leqq |x_1(t)|+|x_1(s_i)|+|x_2(t)-x_2(s_i)|\\ &\leqq \sqrt{\gamma/a}\left(1+1+\int_{s_i}^{t}|b(\tau)|\,d\tau\right).\end{aligned}$$

Now $|x(s_i)-(0,x_2(s_i))|=|x_1(s_i)|\leqq\sqrt{\gamma/a}$. Therefore $|x(t)-(0,x_2(s_i))|\leqq|x(t)-x(s_i)|+|x(s_i)-(0,x_2(s_i))|\leqq\theta_i(t,\gamma)$. The $\omega\in\Omega$ for $x(t)$ is thus $\omega(t)\equiv(0,x_2(s_i))$. (Again, take $\gamma_1^*=\infty$.)

Example 3 (continued). See §§ 4 and 5 for examples in which α is defined from a retraction $\pi_t\colon N(H_t)\to H$ (although Example 2 above is a case in point). These sections include examples in which Ω does not consist of constants and $N(H_t)$ is not all of R^n.

Recall that we are assuming W obeys Assumption D with $\eta\in K$.

THEOREM 3. *The equilibrium* 0 *for* (1) *is u.a.s. with* G_0 *in the basin if the following conditions hold.*

1. (W,α) *is admissible with* $N(H_t)$, δ, *and* γ_0^*.

2. α *obeys assumption* B_0 *with* $k_0\in K$.

3. *For each annulus* A, *there are* Ω, *positive constants* T, γ_0, γ_1 *with* $\gamma_0^*\geqq\gamma_0$ *and* $\gamma_1^*\geqq\gamma_1$, *a sequence of constants* $s_i\to\infty$, *and a sequence of functions* $\theta_i\in KK_0$ *such that*

(a) Ω *approximates solutions to* (1) *in* A *near* H_t *with* T, γ_1^*, s_i, *and* θ_i,

(b) α *is PTUE with respect to* Ω *with* $\phi_i=\delta_i(\gamma_0)+k_0(\theta_i(\gamma_1))$, T, s_i *where*

$$\delta_i(\gamma_0)=\max\{\delta(t,\gamma_0)|t\in[s_i,s_{i+1}]\},$$

$$\theta_i(\gamma_1)=\max\{\theta_i(t,\gamma_1)|t\in[s_i,s_{i+1}]\}.$$

Further, with $b_0\geqq V(t,x)$ *for* $(t,x)\in R^+\times(A\cap G)$, $[2b_0/(\gamma d_0)]2T$ *is an upper bound for the rate of persistence of* (1) *in* A *where* $\gamma=\min\{\gamma_0,\gamma_1\}$ *and* $d_0=\min\{T,\eta^{-1}(\gamma/2)\}$.

Proof. This is a corollary to Theorem 2.

1. Let A be an annulus. Let Ω, γ_0, γ_1, T, s_i, and θ_i be as in the hypothesis. Fix the index i. Suppose $x(t)$ is a solution to (1) with $x(t)\in A$ for $t\in[s_i,s_{i+1}]=I_i$. Suppose $|\alpha(t,x(t))|\leqq\delta(t,\gamma_1)$ for $t\in I_i$. Then, by condition 3(a) in the hypothesis, there is an $\omega\in\Omega$ with $|x(t)-\omega(t)|\leqq\theta_i(t,\gamma_1)$ for $t\in I_i$.

2. Now there is a $t_0\in I_i$ such that $|\alpha(t_0,\omega(t_0))|>\delta_i(\gamma_0)+k_0(\theta_i(\gamma_1))$. Also $|\alpha(t,x(t))-\alpha(t,\omega(t))|\leqq k_0(|x(t)-\omega(t)|)\leqq k_0(\theta_i(t,\gamma_1))$ for $t\in I_i$. Thus $|\alpha(t_0,x(t_0))|\geqq|\alpha(t_0,\omega(t_0))|-k_0(\theta_i(t_0,\gamma_1))>\delta_i(\gamma_0)+k_0(\theta_i(\gamma_1))-k_0(\theta_i(t_0,\gamma_1))\geqq\delta_i(\gamma_0)\geqq\delta(t_0,\gamma_0)$.

3. Therefore, $W(t_0,x(t_0))>\gamma_0$. We conclude that W is PTUE with respect to Ω_A with $\phi_i=\gamma$ where $\gamma=\min\{\gamma_0,\gamma_1\}$, T, and s_i. Now apply Theorem 2.

THEOREM 4. *The conclusion of Theorem* 3 *remains true exactly as written if we replace condition* 2 *of the hypothesis by*

2′. α *obeys Assumption* C *with* $k\in K$,

and replace 3(b) *by*

3(b)′. $\dot\alpha(t,x)$ *is UE with respect to* Ω *with* $\phi_i=2\delta_i(\gamma_0)+k(\theta_i(\gamma_1))T$, T, s_i.

Proof. This is a corollary to the proof of Theorem 3. Parts 1 and 3 of that proof carry over exactly as written. We shall see now that the conclusion of part 2 follows by a simple argument.

There is an interval $I=[a,b]\subseteq[s_i,s_{i+1}]$ such that

$$\left|\int_a^b\dot\alpha(\tau,\omega(\tau))\,d\tau\right|>\phi_i.$$

Now $|\dot{\alpha}(t, \omega(t))-\dot{\alpha}(t, x(t))| \leqq k(|\omega(t)-x(t)|) \leqq k(\theta_i(t, \gamma_1)) \leqq k(\theta_i(\gamma_1))$ for a.e. $t \in I$. Therefore,

$$\left|\int_a^b \dot{\alpha}(\tau, \omega(\tau))\, d\tau-\int_a^b \dot{\alpha}(\tau, x(\tau))\, d\tau\right| \leqq \int_a^b |\dot{\alpha}(\tau, \omega(\tau))-\dot{\alpha}(\tau, x(\tau))|\, d\tau$$
$$\leqq k(\theta_i(\gamma_1))T.$$

Thus

$$\left|\int_a^b \dot{\alpha}(\tau, x(\tau))\, d\tau\right|>\phi_i-k(\theta_i(\gamma_1))T=2\delta_i(\gamma_0).$$

Therefore, $|\alpha(b, x(b))-\alpha(a, x(a))|>2\delta_i(\gamma_0)$ implying there is a $t_0 \in I$ such that $|\alpha(t_0, x(t_0))|>\delta_i(\gamma_0)$. This completes the proof of Theorem 4.

The remark after Theorem 2 on "robustness" carries over to Theorems 3 and 4.

Theorems and Corollaries 5 and 6 below provide some simple concrete applications of these results. This includes versions of the main theorems for Morgan and Narendra [19] and [20]. In §§ 4 and 5, more complicated consequences are given. There are many possible variations and generalizations of this material. For example, Theorem 3 (Theorem 4 respectively) requires Lipschitz-like conditions on W and α (W and $\dot{\alpha}$ respectively), and these can be weakened. Thus W can be piecewise continuous under fairly general circumstances. In § 6 some such extensions are outlined.

THEOREM 5. *Assume that* $W(t, x)=w(|f(t, x)|)$ *for some* $w \in K$. *Suppose, for every* $\varepsilon>0$, *there is a* $w_\varepsilon \in K$ *such that* $|w(a)-w(b)| \leqq w_\varepsilon(|a-b|)$ *for* $a, b \in[0, \varepsilon]$. *Suppose that there is a constant c such that*

$$|f(t, x)-f(s, y)| \leqq c(|t-s|+|x-y|)$$

for $(t, x),(s, y) \in R^{+} \times G$.

Then 0 *for* (1) *is u.a.s. with* G_0 *in the basin if and only if, for each annulus A, there is a* $\phi_0>0$ *such that* $f(t, x)$ *is PTUE with respect to constants in* $A \cap G$ *with* $\phi_i=\phi_0$ *for all i.*

Further, if $A=A_{\varepsilon_1, \varepsilon_2}$, $b_0 \geqq V(t, x)$ *for all* $(t, x) \in R^{+} \times(A \cap G)$, *and* $|f(t, x)|$ *is PTUE with respect to constants in* $A \cap G$ *with* ϕ_0, T, *then* $[2 b_0 /(\gamma d_0)] 2 T$ *is an upper bound for the rate of persistence for* (1) *in A where*

$$\gamma=w(\phi_0 /(1+c T)),$$
$$d_0=\min \{T,(w_{\varepsilon_1}^{-1}(\gamma / 2)) /(c(1+c \varepsilon_1))\}$$

A version of this result was proven in Morgan and Narendra [19]. As in that paper, the $|f(t, x)-f(s, y)| \leqq c(|t-s|-|x-y|)$ requirement may be replaced by a uniform Lipschitz condition if a different excitedness definition is substituted for PTUE. See § 6.3.

Proof of Theorem 5. The necessity of PTUE for u.a.s. is immediate from Corollary 1. The sufficiency follows from Theorem 3, as shown below.

1. (W, α) is admissible as in Example 1:

$$W(t, x)=w(|f(t, x)|),$$
$$\alpha(t, x)=|f(t, x)|,$$
$$\delta(t, \gamma)=w^{-1}(\gamma),$$
$$N(H_t)=G,$$
$$\gamma_0^*=\infty.$$

2. W obeys Assumption D with $\eta \in K$: Define $\eta(s) = w_{\varepsilon_1}((c+c^2\varepsilon_1)s)$. Let $x(t)$ be a solution in $A = A_{\varepsilon_1,\varepsilon_2}$ for $t \in [a, b]$. Then

$$\begin{aligned}
|W(b, x(b)) - W(a, x(a))| &= |w(|f(b, x(b))|) - w(|f(a, x(a))|)| \\
&\leqq w_{\varepsilon_1}(|f(b, x(b)) - f(a, x(a))|) \\
&\leqq w_{\varepsilon_1}(c((b-a) + |x(b) - x(a)|)) \\
&\leqq w_{\varepsilon_1}\left(c\left((b-a) + \int_a^b |\dot{x}(\tau)|\, d\tau\right)\right) \\
&\leqq w_{\varepsilon_1}\left(c\left((b-a) + c\int_a^b |x(\tau)|\, d\tau\right)\right) \\
&\leqq w_{\varepsilon_1}(c((b-a) + c^2\varepsilon_1(b-a))) \\
&= w_{\varepsilon_1}((c+c^2\varepsilon_1)(b-a)) = \eta(b-a).
\end{aligned}$$

3. $|f(t, x)|$ obeys Assumption B_0 with $k_0(s) = cs$:

$$\big||f(t, x)| - |f(t, y)|\big| \leqq |f(t, x) - f(t, y)| \leqq c|x-y|.$$

4. Ω = constants in $A \cap G$ approximates solutions to (1) in A near H_t with $\theta_i(t, \gamma) = w^{-1}(\gamma)(t - s_i)$: This was shown in Example 1 (continued).

5. Now, the hypothesis of Theorem 3 is satisfied if there are γ_0, γ_1, T, $s_i \to \infty$ such that $\alpha(t, x) = |f(t, x)|$ is PTUE with respect to Ω with $\phi_i = w^{-1}(\gamma_0) + cw^{-1}(\gamma_1)T$, T, and s_i. This follows for $\gamma_0 = \gamma_1 = w(\phi_0/(1+cT))$, because then $\phi_i = \phi_0$ for all i.

COROLLARY 5. *Let $Q(t)$ be a symmetric positive semidefinite nonautonomous $n \times n$ matrix. Suppose that there is a constant c with $c \geqq |Q(t)|$ and $|Q(t) - Q(s)| \leqq c|t-s|$ for all $t, s \in R^+$. Suppose there is a $\phi_0 > 0$ such that $Q(t)$ is PTUE with $\phi_i = \phi_0$ for all i and with some T and $s_i \to \infty$. Then, with $X(t, t_0)$ denoting the fundamental solution of $\dot{x} = -Q(t)x$,*

$$|X(t, t_0)| \leqq 2\, e^{-L(t-t_0)}$$

for any t_0 and $t \geqq t_0$ with $L = \ln(2)r$, $r = [8/(\gamma d_0)]2T$, and

$$\begin{aligned}
\gamma &= (2/c)(\phi_0/(1+cT))^2 \\
d_0 &= \min\{T, \gamma/(16(1+2c))\}
\end{aligned}$$

Conversely, if 0 is u.a.s., then there is a $\phi_0 > 0$ such that $Q(t)$ is PTUE with $\phi_i = \phi_0$ for all i.

This is similar to the main theorem in Morgan and Narendra [19]. As noted after Theorem 5, the $|Q(t) - Q(s)| \leqq c|t-s|$ condition may be omitted if a modified PTUE condition is used.

Proof. This is a simple consequence of Theorem 5. We have

$$\begin{aligned}
&f(t, x) = -Q(t)x, \\
&V(t, x) = |x|^2, \\
&-\dot{V}(t, x) = 2x^TQ(t)x \geqq (2/c)x^TQ(t)Q(t)x = (2/c)|Q(t)x|^2 \\
&\qquad = w(|f(t, x)|), \quad \text{where } w(s) = (2/c)s^2.
\end{aligned}$$

Now $|Q(t)x|$ is PTUE with respect to unit constants with ϕ_0, T. Therefore, $|Q(t)x|$ is PTUE with respect to constants in A with ϕ_0, T where $A=A_{\varepsilon_1,\varepsilon_2}$ with $\varepsilon_1=2$ and $\varepsilon_2=1$. We also have $G=R^n$, $b_0=4$, and $w_\varepsilon(s)=(4\varepsilon/c)s$ for any $\varepsilon>0$.

By Theorem 5, r is an upper bound for the rate of persistence of (1) in A. The result follows from this and the material in § 2.2.

It will be worthwhile to note here some general conditions under which constants approximate solutions. Suppose

1. there is a $\xi\in KK_0$ such that if $|\alpha(t,x)|\leqq\delta(t,\gamma)$, then $|f(t,x)-\dot{\alpha}(t,x)|\leqq\xi(t,\gamma)$,
2. there is a $\hat{\delta}\in KK_0$ such that if $|\alpha(t,x)|\leqq\delta(t,\gamma)$, then $d(x,H_t)\leqq\hat{\delta}(t,\gamma)$.

Let $s_i\to\infty$ with $s_{i+1}-s_i\leqq T$ and $\Omega=\bigcup_i(H_{s_i}\cup A)$, a collection of constant functions. Then Ω approximates solutions to (1) in A near H_t with θ_i, T, and s_i where

$$\theta_i(t,\gamma)=\hat{\delta}(s_i,\gamma)+\delta(s_i,\gamma)+\delta(t,\gamma)+\int_{s_i}^{t}\xi(\tau,\gamma)\,d\tau.$$

This can be seen as follows. Let $x(t)$ be a solution in A for $t\in[s_i,s_{i+1}]$. Then

$$\begin{aligned}&||x(t)-x(s_i)|-|\alpha(t,x(t))-\alpha(s_i,x(s_i))||\\&\quad=\left|\left|\int_{s_i}^{t}\dot{x}(\tau)\,d\tau\right|-\left|\int_{s_i}^{t}\dot{\alpha}(\tau,x(\tau))\,d\tau\right|\right|\leqq\left|\int_{s_i}^{t}(f(\tau,x(\tau))-\dot{\alpha}(\tau,x(\tau)))\,d\tau\right|.\end{aligned}$$

Therefore, if $|\alpha(\tau,x(\tau))|\leqq\delta(\tau,\gamma)$ for $\tau\in[s_i,t]$, then

$$\begin{aligned}|x(t)-x(s_i)|&\leqq|\alpha(t,x(t))|+|\alpha(s_i,x(s_i))|+\int_{s_i}^{t}\xi(\tau,\gamma)\,d\tau\\&\leqq\delta(t,\gamma)+\delta(s_i,\gamma)+\int_{s_i}^{t}\xi(\tau,\gamma)\,d\tau.\end{aligned}$$

Also $d(x(s_i),H_{s_i})\leqq\hat{\delta}(s_i,\gamma)$. Thus there is an $x_i\in H_{s_i}$ with $|x(s_i)-x_i|\leqq\hat{\delta}(s_i,\gamma)$. Therefore, $|x(t)-x_i|\leqq|x(t)-x(s_i)|+|x(s_i)-x_i|\leqq\theta_i(t,\gamma)$.

Note also that the UE of $\dot{\alpha}$ with respect to Ω is equivalent to the UE of f with respect to Ω, since $f(t,\omega)=\dot{\alpha}(t,\omega)$ for $\omega\in\Omega$.

Some consequences of these remarks and Theorem 4 follow.

Let m be an integer with $0<m<n$. Then $R^n=R^{n-m}\times R^m$, and we may write $x=(x_1,x_2)\in R^{n-m}\times R^m$, $f(t,x)=(f_1(t,x),f_2(t,x))\in R^{n-m}\times R^m$.

THEOREM 6. *Assume that* $W(t,x)=w(|x_1|)$ *for some* $w\in K$. *Suppose, for every* $\varepsilon>0$, *there is a* $w_\varepsilon\in K$ *such that* $|w(a)-w(b)|\leqq w_\varepsilon(|a-b|)$ *for* $a,b\in[0,\varepsilon]$. *Suppose there is a* $\xi_0\in K$ *such that* $|f_2(t,x)|\leqq\xi_0(|x_1|)$ *for* $(t,x)\in R^+\times G$. *Assume f obeys a uniform Lipschitz condition with constant c.*

Define $\beta(t,x_2)=f_1(t,0,x_2)$. *Then* 0 *for* (1) *is u.a.s. with* G_0 *in the basin if and only if, for each annulus A, there is a* $\phi_0>0$ *such that* β *is UE with respect to constants in* $A_0=A\cap H_0$ *with* $\phi_i=\phi_0$ *for all i.*

Further, if $A=A_{\varepsilon_1,\varepsilon_2}$ *and* $b_0\geqq V(t,x)$ *for* $(t,x)\in R^+\times(A\cap G)$, *and* $\beta(t,x_2)$ *is UE with respect to constants in* A_0 *with* ϕ_0, T, *then* $[2b_0/(\gamma d_0)]2T$ *is an upper bound for the rate of persistence of* (1) *in A where* $\gamma=\min\{\gamma_0,\gamma_1\}$, γ_0 *and* γ_1 *obey* $\phi_0\geqq 2w^{-1}(\gamma_0)+c(3w^{-1}(\gamma_1)+\xi_0(w^{-1}(\gamma_1))T)T$, *and* $d_0=\min\{T,(w_{\varepsilon_1}^{-1}(\gamma/2))/(c\varepsilon_1)\}$.

Proof. The necessity of UE for u.a.s. is immediate from Theorem 1. The sufficiency follows from Theorem 4, as shown below.

1. (W, α) is admissible:

$$\begin{aligned} &W(t, x) = w(|x_1|), \\ &\alpha(t, x) = x_1, \qquad \dot\alpha(t, x) = f_1(t, x), \\ &\delta(t, \gamma) = w^{-1}(\gamma), \\ &N(H_t) = G, \\ &\gamma_0^* = \infty, \\ &H_0 = \{(0, x_2) \in R^{n-m} \times R^m\} = H_t \quad \text{for all } t \in R^+, \\ &\dot\alpha(t, x) = \beta(t, x_2) \quad \text{for } (0, x_2) \in H_0. \end{aligned}$$

2. W obeys Assumption D with $\eta \in K$: Define $\eta(s) = w_{\varepsilon_1}(c\varepsilon_1 s)$. Let $x(t)$ be a solution in $A = A_{\varepsilon_1, \varepsilon_2}$ for $t \in [a, b]$. Then

$$\begin{aligned} |W(b, x(b)) - W(a, x(a))| &= |w(|x_1(b)|) - w(|x_1(a)|)| \\ &\leqq w_{\varepsilon_1}(||x_1(b)| - |x_1(a)||) \leqq w_{\varepsilon_1}(|\dot x_1(b) - x_1(a)|) \\ &\leqq w_{\varepsilon_1}\left(\int_a^b |\dot x_1(\tau)|\, d\tau\right) \leqq w_{\varepsilon_1}(c\varepsilon_1(b-a)) = \eta(b-a). \end{aligned}$$

3. $\dot\alpha$ obeys Assumption B with $k(s) = cs$:

$$|f_1(t, x) - f_1(t, y)| \leqq |f(t, x) - f(t, y)| \leqq c|x - y|.$$

4. $\Omega = \{(0, x_2) \in A_0\}$ approximates solutions in A near H_0 where $A_0 = A \cap H_0$ with $\theta_i(t, \gamma) = 3w^{-1}(\gamma) + \xi_0(w^{-1}(\gamma))(t - s_i)$: This follows from the comments preceding the statement of the theorem. Define $\xi(t, \gamma) = \xi_0(w^{-1}(\gamma))$. Then if $|\alpha(t, x)| \leqq \delta(t, \gamma)$, then $|f(t, x) - \dot\alpha(t, x)| \leqq \xi(t, \gamma)$. Also $d(x, H_0) = |x_1|$, so $\hat\delta = \delta$.

5. Now, the hypothesis of Theorem 4 is satisfied if there are $\gamma_0, \gamma_1, T, s_i \to \infty$ such that $\dot\alpha(t, x)$ is UE with respect to Ω with $\phi_i = 2\delta(\gamma_0) + k(\theta_i(\gamma_1))T$. This follows for any choice of γ_0 and γ_1 obeying the inequality in the hypothesis.

The linear result below appeared in Morgan and Narendra [20] without the rate of convergence estimate.

COROLLARY 6. *Consider the linear system*

$$\begin{pmatrix} \dot x_1 \\ \dot x_2 \end{pmatrix} = \begin{pmatrix} -Q_0(t) & -P_1(t)^T \\ P_2(t) & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{2}$$

where $Q_0(t)$ is $(n-m) \times (n-m)$ and $P_1(t)$, $P_2(t)$ are $m \times (n-m)$ continuous bounded matrices.

(a) *If* 0 *for* (2) *is u.a.s., then there is a $\phi_0 > 0$ such that $P_1(t)$ is UE with $\phi_i = \phi_0$ for all i.*

(b) *If there is a symmetric positive definite differentiable matrix $J(t)$ such that $\dot J + JQ_0^T + Q_0 J$ is positive definite, $P_2(t) = J(t)P_1(t)$, and there is a $\phi_0 > 0$ such that $P_1(t)$ is UE with $\phi_i = \phi_0$ for all i, then* 0 *for* (2) *is u.a.s.*

Further, if $X(t, t_0)$ denotes the fundamental solution of (2), *then*

$$|X(t, t_0)| \leqq (2j_1/j_2)\, e^{-L(t-t_0)}$$

for any t_0 and $t \geqq t_0$ with $L = \ln(2)r$, $r = [8j_1^2/(\gamma d_0)]2T$, $\gamma = \min\{\gamma_0, \gamma_1\}$ where

$$\phi_0 \geqq 2\sqrt{\gamma_0/w_0} + c(3\sqrt{\gamma_1/w_0} + j_1^2 c\sqrt{\gamma_1/w_0}T)T, \qquad d_0 = \min\{T, \gamma/(16cw_0)\},$$

and j_1, j_2, c, w_0 are positive constants with $j_1^2 \geqq |J(t)| \geqq j_2^2$, $j_1 \geqq 1$, $j_2 \leqq 1$, $|B(t)| \leqq c$ where B *is the matrix on the right hand side of* (2), $|\dot{J}(t)+J(t)Q_0(t)^T+Q_0(t)J(t)|\geqq w_0$ *for all* $t\in R^+$, *and* $P_1(t)$ *is UE with* ϕ_0 *and* T.

Proof. This is a simple consequence of Theorem 6 and the material in § 2.2. We have

$$j_2^2|x|^2\leqq V(t,x)\leqq j_1^2|x|^2,$$

and

$$-\dot{V}(t,x)=x_1^T(\dot{J}(t)+J(t)Q_0(t)^T+Q_0(t)J(t))x_1\geqq w_0|x_1|^2\geqq 0.$$

It follows that $|X(t,t_0)|\leqq j_1/j_2$ for $t\geqq t_0$.

In the notation of the theorem:

$$\begin{aligned}&w(s)=w_0s^2,\\&\xi_0(s)=j_1^2cs,\\&b_0=4j_1^2\quad\text{taking } \varepsilon_1=2 \text{ and } \varepsilon_2=1,\\&w_\varepsilon(s)=2w_0\varepsilon s,\\&\beta(t,x_2)=P_1(t)^Tx_2.\end{aligned}$$

4. Admissible (W, α) and Ω. In this section and the next, we will focus attention on identifying systems that possess admissible (W, α) and on identifying convenient classes of Ω which approximate solutions. In § 6 the emphasis will be on such considerations as weakening the Assumption D for W and Assumptions B_0 or C for α requirements of Theorems 3 and 4.

Suppose, for the rest of this section, that $W(t,x)=k(x^TQ(t)x)$ where $k\in K$ and $Q(t)$ is a symmetric positive semidefinite matrix. If $f(t,x)$ is linear, we can take annulus A always to be $A_{2,1}$ (or any other fixed annulus). However, $f(t,x)$ need not be linear.

4.1. Admissibility. Three schemes for admissibility, labeled (a), (b), and (c), are presented below. For simple systems they are essentially equivalent, but even in simple cases there are conceptual advantages of one over the other. In particular, the third scheme emphasizes the geometry of admissibility, although it is computationally more difficult to realize than the others.

(a)
$$\begin{aligned}&\alpha(t,x)=x^TQ(t)x,\\&\delta(t,\gamma)=k^{-1}(\gamma),\\&N(H_t)=R^n,\\&\gamma_0^*=\infty.\end{aligned}$$

(b)
$$\begin{aligned}&\alpha(t,x)=Q(t)x,\\&\delta(t,\gamma)=\sqrt{|Q(t)|k^{-1}(\gamma)},\\&N(H_t)=R^n,\\&\gamma_0^*=\infty.\end{aligned}$$

Compare Example 1, Theorem 5, and Corollary 5.

(c) To outline the third scheme, we need some preliminaries. For any t, $Q(t)$ can be diagonalized:

$$Q(t)=L(t)^T\Lambda(t)L(t)$$

where $L(t)$ is orthogonal and $\Lambda(t) = \operatorname{diag}(\lambda_1(t), \lambda_2(t), \cdots, \lambda_n(t))$ with $\lambda_i(t)$ the ith eigenvalue of $Q(t)$. (See Bellman [3].) If $Q(t)$ is piecewise continuous (measurable, respectively) in t, then $L(t)$ and $\Lambda(t)$ are piecewise continuous (measurable) in t also.

Define

$$\Lambda_\nu(t) = \operatorname{diag}(\sigma(\lambda_1(t)), \cdots, \sigma(\lambda_n(t))),$$

$$\Lambda_\mu(t) = \operatorname{diag}(\bar{\sigma}(\lambda_1(t)), \cdots, \bar{\sigma}(\lambda_n(t)))$$

where

$$\sigma(s) = \begin{cases} 1 & \text{if } s \neq 0, \\ 0 & \text{if } s = 0, \end{cases}$$

$$\bar{\sigma}(s) = \begin{cases} 0 & \text{if } s \neq 0, \\ 1 & \text{if } s = 0. \end{cases}$$

Define $Q_\nu(t) = L(t)^T \Lambda_\nu(t) L(t)$, $Q_\mu(t) = L(t)^T \Lambda_\mu(t) L(t)$, $x_{\mu(t)} = Q_\mu(t)x$, and $x_{\nu(t)} = Q_\nu(t)x$ for any $x \in R^n$. Then $x = x_{\mu(t)} + x_{\nu(t)}$ for any $(t, x) \in R^+ \times R^n$, and $H_t = \{x_{\mu(t)} | x \in R^n\}$ which is thus a hyperplane through 0. Moreover, $x^T Q(t) x = x^T_{\nu(t)} Q(t) x_{\nu(t)}$ and $d(x, H_t) = |x_{\nu(t)}|$ for any $(t, x) \in R^+ \times R^n$.

Note that $x_{\mu(t)}$ is the orthogonal projection of x onto H_t, and $x_{\nu(t)}$ is the orthogonal projection of x onto the normal complement of H_t. (The orthogonality is with respect to the dot product.)

Let $q(t)^2$ denote the smallest nonzero eigenvalue of $Q(t)$. Define

$$\alpha(t, x) = x_{\nu(t)},$$

$$\delta(t, \gamma) = \begin{cases} \sqrt{k^{-1}(\gamma)}/q(t) & \text{if } |Q(t)| \neq 0, \\ \infty & \text{if } |Q(t)| = 0, \end{cases}$$

$$\gamma_0^* \text{ is any value,} \quad 0 < \gamma_0^* \leqq \infty,$$

$$N(H_t) = \begin{cases} S_{\delta(t, \gamma_0^*})(H_t) & \text{if } \gamma_0^* < \infty, \\ R^n & \text{if } \gamma_0^* = \infty. \end{cases}$$

Compare Example 2 and Corollary 6. Also see Example 4 below.

4.2. Approximating Ω. It is difficult to be specific about the choice of Ω in general. Intuitively, Ω should consist of solutions to the "simplified" system $\dot{y} = f(t, y)$ when y is "near" H_t. For Theorem 5, $f(t, y) = 0$ when $y \in H_t$ which indicated Ω = constants as the natural choice. For Theorem 6, $f(t, y)$ was not 0 when $y \in H_t$, but solutions to $\dot{y} = f(t, y)$ for y near H_t could stay near H_t only by being nearly constant of the form $(0, y_2)$. Thus, these constants were the natural choice for Ω in this case.

In order to provide more specific information on the choice of Ω, let us consider the following type of quasi-linear system:

$$(3) \qquad \dot{x} = P(t, x)x - Q(t)x$$

where $P(t, x)$ is a skew symmetric $n \times n$ matrix for each $(t, x) \in R^+ \times G$ and $Q(t)$ is a symmetric positive semidefinite matrix for each $t \in R^+$. (Note that $\dot{x} = -Q(t)x$ from Corollary 5 and (2) from Corollary 6, when Q_0 is symmetric positive definite, are of this form.) The analysis of this case will illustrate some general ideas.

With $V(t, x) = \frac{1}{2}|x|^2$, we get $-\dot{V}(t, x) = x^T Q(t) x = W(t, x)$. Let us choose admissibility scheme (b), so that $\alpha(t, x) = Q(t)x$ and $\delta(t, \gamma) = \sqrt{\gamma |Q(t)|}$. Now, when $W = 0$, $\dot{x} = f(t, x)$ simplifies to $\dot{y} = P(t, y)y$ which suggests defining Ω to be solutions to this equation. A variant on this idea is presented below.

Let A be an annulus, and let $T > 0$ and $s_i \to \infty$ with $s_{i+1} - s_i \leqq T$ for each i. Let γ_0^* be a fixed positive number. Define

$$\gamma_0(t) = \delta(t, \gamma_0^*) = \sqrt{|Q(t)|\gamma_0^*},$$

$$\theta_i(t, \gamma) = \int_{s_i}^{t} \delta(\tau, \gamma)\, d\tau = \sqrt{\gamma} \int_{s_i}^{t} \sqrt{|Q(\tau)|}\, d\tau,$$

$$\gamma_i(t) = \gamma_0(t) + |Q(t)|\theta_i(t, \gamma_0^*),$$

$$N(H_t) = \{x \,|\, |\alpha(t, x)| < \gamma_0(t)\}.$$

For every smooth $u\colon [s_i, s_{i+1}] \to A \cap N(H_t)$, define

$$\Omega_u = \{\omega(t) \in A \,|\, |\alpha(t, \omega(t))| < \gamma_i(t),\ \dot\omega(t) = P(t, u(t))\omega(t) \text{ for } t \in [s_i, s_{i+1}]\}.$$

Then define $\Omega = \bigcup \Omega_u$ where the union is over all such u, for all i.

Now, by comparing (3) with $\dot\omega = P(t, u(t))\omega$ via variation of constants and noting that solutions to (3) are allowable choices for $u(t)$, we see that Ω approximates solutions to (1) in A near H_t with s_i, T, γ_0^*, $N(H_t)$, and $\theta_i(t, \gamma)$ as defined above. Given solution $x(t)$ with $x(t_0) \in N(H_t)$, the $\omega \in \Omega$ associated with $x(t)$ is defined by $\dot\omega = P(t, x(t))\omega$ with $\omega(t_0) = x(t_0)$.

Sometimes there is a collection Ω' of functions that are "close to" those in Ω and have desirable additional properties (e.g. computable or simple form). Then Ω' can be used in place of Ω, and this can result in a simplification of the stability criterion. Theorems and Corollaries 5 and 6 illustrate this. Example 4 below provides an example for which Ω does not consist of constants.

Example 4. Consider the three dimensional system $\dot x = P(t)x - Q(t)x$ where

$$P(t) = \begin{pmatrix} 0 & -a(t) & 0 \\ a(t) & 0 & -b(t) \\ 0 & b(t) & 0 \end{pmatrix},$$

and $Q(t) = \operatorname{diag}(q(t)^2, 0, 0)$ where $a, b, q\colon R^+ \to R^1$ are measurable functions. Further assume $q(t) > 0$ for all t.

Let us choose scheme (b) for admissibility. Thus we can use the material developed above: $V(t, x) = \frac{1}{2}|x|^2$ and $-\dot V(t, x) = x^T Q(t) x = q(t)^2 x_1^2$, $k(s) = s$, $\alpha(t, x) = q(t)^2 x_1$, $\delta(t, \gamma) = \sqrt{\gamma} q(t)$. Let $s_i \to \infty$ and T be given so that $s_{i+1} - s_i \leqq T$. Then $\gamma_0(t) = \sqrt{\gamma_0^*} q(t)$,

$$\theta_i(t, \gamma) = \int_{s_i}^{t} \delta(\tau, \gamma)\, d\tau = \sqrt{\gamma} \int_{s_i}^{t} q(\tau)\, d\tau,$$

$$\gamma_i(t) = q(t)\sqrt{\gamma_0^*} + q(t)^2 \sqrt{\gamma_0^*} \int_{s_i}^{t} q(\tau)\, d\tau,$$

$$H_t = \{x \in R^3 \,|\, x_1 = 0\},$$

$$N(H_t) = \{x \in R^3 \,|\, |x_1| < \sqrt{\gamma_0^*}/q(t)\},$$

$$A = A_{\varepsilon_1, \varepsilon_2},$$

$$\Omega = \{\omega(t) \,|\, \varepsilon_1 > |\omega(t)| > \varepsilon_2,\ |\omega_1(t)| < \gamma_i(t)/q(t)^2,$$
$$\dot\omega(t) = P(t)\omega(t) \text{ for } t \in [s_i, s_{i+1}] \text{ for some } i\}.$$

Now, Ω approximates solutions to $\dot x = P(t)x - Q(t)x$ in A near H_t with s_i, T, γ_0^*, $N(H_t)$, $\theta_i(t, \gamma)$ as above. However, we can define an Ω' with elements "near" those in Ω

that also approximates solutions. Further, this Ω' has certain advantages over Ω as noted below.

Define

$$P_0(t)=\begin{pmatrix} 0 & -b(t) \\ b(t) & 0 \end{pmatrix}.$$

Let $\Omega'=\{(0, v_0(t))\,|\,v_0(t)$ is a solution to $\dot{v}_0=P_0(t)v_0\}$.

Claim. Ω' approximates solutions to $\dot{x}=P(t)x-Q(t)x$ in A near H_t with s_i, T, γ_0^*, $N(H_t)$ as above, and

$$\theta_1'(t,\gamma)=2(\theta_i(s_i;\gamma)+\theta_i(t,\gamma))+2(\sqrt{\gamma}/q(s_i))$$
$$+(\sqrt{\gamma}/q(t))+\int_{s_i}^{t}|a(\tau)|(\theta_i(\tau,\gamma)+(\sqrt{\gamma}/q(\tau)))\,d\tau.$$

This Ω' has the advantage of being lower dimensional than Ω and of being explicitly computable. (We can solve $\dot{v}_0=P_0(t)v_0$.)

The lemma below is the key fact in the proof of the claim. Let

$$P'(t)=\left(\begin{array}{c|cc} 0 & 0 & 0 \\ \hline 0 & \multicolumn{2}{c}{\multirow{2}{*}{$P_0(t)$}} \\ 0 & & \end{array}\right),$$

and let $P''(t)=P(t)-P'(t)$.

LEMMA. *Let ω and v be solutions to $\dot{\omega}=P(t)\omega$ and $\dot{v}=P'(t)v$ respectively with $\omega(s_i)=v(s_i)$. Suppose $\varepsilon(t)$ exists with $|\omega_1(t)|\leqq\varepsilon(t)$ for $t\in[s_i,s_{i+1}]$. Then*

$$|v(t)-\omega(t)|\leqq\varepsilon(t)+\varepsilon(s_i)+\int_{s_i}^{t}|a(\tau)|\varepsilon(\tau)\,d\tau$$

for $t\in[s_i,s_{i+1}]$.

Let us first establish that the claim follows from the lemma. Then the lemma will be proven.

Let $x(t)$ be a solution in $A\cap N(H_t)$ with $|\alpha(t,x(t))|\leqq\delta(t,\gamma)$ for $t\in[s_i,s_{i+1}]$. Thus $|x_1(t)|\leqq\sqrt{\gamma}/q(t)$ for $t\in[s_i,s_{i+1}]$. Then there is an $\omega(t)\in\Omega$ such that $|x(t)-\omega(t)|\leqq\theta_i(t,\gamma)$ for $t\in[s_i,s_{i+1}]$. Thus $|x_1(t)-\omega_1(t)|\leqq\theta_i(t,\gamma)$ implying $|\omega_1(t)|\leqq|x_1(t)|+\theta_i(t,\gamma)\leqq(\sqrt{\gamma}/q(t))+\theta_i(t,\gamma)$. Let $\varepsilon(t)=(\sqrt{\gamma}/q(t))+\theta_i(t,\gamma)$ and apply the lemma. Then there is a $v(t)$ with $\dot{v}(t)=P(t)v(t)$ and $\omega(s_i)=v(s_i)$ such that

$$|v(t)-\omega(t)|\leqq\varepsilon(t)+\varepsilon(s_i)+\int_{s_i}^{t}|a(\tau)|\varepsilon(\tau)\,d\tau.$$

Because of the form of P', $v_1(t)=v_1(s_i)=\omega_1(s_i)$ is constant. Then

$$|(0,v_2(t),v_3(t))-\omega(t)|\leqq|v_1(t)|+\varepsilon(t)+\varepsilon(s_i)+\int_{s_i}^{t}|a(\tau)|\varepsilon(\tau)\,d\tau$$
$$\leqq\varepsilon(s_i)+\varepsilon(t)+\varepsilon(s_i)+\int_{s_i}^{t}|a(\tau)|\varepsilon(\tau)\,d\tau$$

and $v_0(t)=(v_2(t),v_3(t))$ is a solution to $\dot{v}_0=P_0(t)v_0$. Thus

$$|x(t)-(0,v_0(t))|\leqq|x(t)-\omega(t)|+|\omega(t)-(0,v_0(t))|$$
$$\leqq\theta_i(t,\gamma)+2\varepsilon(s_i)+\varepsilon(t)+\int_{s_i}^{t}|a(\tau)|\varepsilon(\tau)\,d\tau=\theta_i'(t,\gamma).$$

Proof of lemma. Let $V(t)$ be the fundamental solution for $\dot{v}=P(t)v$. Then by variation of parameters

$$|\omega(t)-v(t)|=\left|\int_{s_i}^{t} V(t-\tau)P''(\tau)\omega(\tau)\,d\tau\right|.$$

Now

$$V(t)=\left(\begin{array}{c|cc} 1 & 0 & 0 \\ \hline 0 & \multicolumn{2}{c}{\multirow{2}{*}{$E(t)$}} \\ 0 & & \end{array}\right) \quad \text{where } E(t) \text{ is orthogonal,}$$

and

$$V(t-\tau)=\left(\begin{array}{c|cc} 1 & 0 & 0 \\ \hline 0 & \multicolumn{2}{c}{\multirow{2}{*}{$E(t)E(\tau)^{-1}$}} \\ 0 & & \end{array}\right).$$

Also

$$P''(\tau)\omega(\tau)=\begin{pmatrix} -a(\tau)\omega_2(\tau) \\ a(\tau)\omega_1(\tau) \\ 0 \end{pmatrix}.$$

Thus

$$\begin{aligned} |\omega(t)-v(t)| &\leqq \left|\int_{s_i}^{t} a(\tau)\omega_2(\tau)\,d\tau\right| + \int_{s_i}^{t} |a(\tau)|\,|\omega_1(\tau)|\,d\tau \\ &\leqq \left|\int_{s_i}^{t} \dot{\omega}_1(\tau)\,d\tau\right| + \int_{s_i}^{t} |a(\tau)|\varepsilon(\tau)\,d\tau \\ &\leqq |\omega_1(t)-\omega_1(s_i)| + \int_{s_i}^{t} |a(\tau)|\varepsilon(\tau)\,d\tau \\ &\leqq \varepsilon(t)+\varepsilon(s_i)+\int_{s_i}^{t} |a(\tau)|\varepsilon(\tau)\,d\tau. \end{aligned}$$

This completes the proof of the lemma.

5. The case $\alpha(t,x)=x-\pi_t(x)$ where $\pi_t\colon N(H_t)\to H_t$ is a retraction. In this section, we consider the case that $\alpha(t,x)=x-\pi_t(x)$ where $\pi_t\colon N(H_t)\to H_t$ is a retraction of a neighborhood of H_t onto H_t. (In other words, π_t is a continuous function with $\pi_t(x)=x$ if $x\in H_t$.) The α for admissibility scheme (c) from the previous section was of this form: $\alpha(t,x)=x_{\nu(t)}=x-\pi_t(x)$ where $\pi_t(x)=x_{\mu(t)}$.

In § 5.1, we consider some general comments on admissibility and Ω. In §§ 5.2 and 5.3, sufficient conditions for existence of π_t are given.

5.1. We have $H_t=\{x\in G \mid W(t,x)=0\}$, a closed subset of G. Also $N(H_t)$ is an open subset of G containing H_t, and $\pi_t\colon N(H_t)\to H_t$ is continuous with $\pi_t(x)=x$ if $x\in H_t$. Letting $\alpha(t,x)=x-\pi_t(x)$, we can define $\delta(t,\gamma)$ by the relation

$$|\alpha(t,x)|>\delta(t,\gamma)\Rightarrow W(t,x)>\gamma,$$

because W_t and π_t are continuous. This holds for $0<\gamma\leqq\gamma_t^*$ for some γ_t^* which depends on properties of W_t and π_t. The existence of γ_0^* which is independent of t does not follow necessarily. It must be established as a separate condition.

The specification of Ω generally requires some insight. Two Ω which are useful to keep in mind but rarely directly computable are described below. First

$$\Omega_1 = \{y(t) \mid y(t) \text{ is a solution to } \dot{y} = f(t, y) \text{ and } y(t) \in N_0(H_t) \cap A \text{ for } t \in [s_i, s_{i+1}] \text{ for some } i\}$$

where $N_0(H_t) \subseteq N(H_t)$ is some neighborhood of H_t, perhaps depending on A or other parameters, and $\theta_i(t, \gamma) = 0$. Now define Ω_2 from Ω_1 as follows:

$$\Omega_2 = \{z(t) = \pi_t(y(t)) \mid y(t) \in \Omega_1\}$$

with

$$\theta_i(t, \gamma) = \delta(t, \gamma).$$

Since we cannot specify Ω_1 or Ω_2 unless we can solve $\dot{y} = f(t, y)$ for y near H_t, the use of these Ω is generally nonroutine. In some cases, however, we can show that α (or $\dot{\alpha}$, respectively) is PTUE (or UE) with respect to all smooth $\omega(t) \in H_t$ obeying certain general properties (e.g. uniformly bounded derivatives). Then we conclude u.a.s. from Theorem 3 (or Theorem 4) by reference to Ω_2. Part of the usefulness of having test functions $\omega(t)$ in H_t is that $\alpha(t, x)$ (or $\dot{\alpha}(t, x)$) may have a simpler form for $x \in H_t$. (See Corollary 6 or the proposition below.)

As noted in § 4, the form of $f(t, x)$ when $x \in H_t$ will sometimes suggest an approximating Ω. The proof that such an Ω is approximating can often be formulated by comparison with Ω_1 or Ω_2. (See the proofs of Theorems 5 and 6 and Example 4.)

5.2. In this subsection, a special case is described. Suppose there is a continuous $\sigma_t \colon R^m \to R^{n-m}$ such that $W(t, x) = 0$ if and only if $x_1 = \sigma_t(x_2)$ where $x = (x_1, x_2) \in R^{n-m} \times R^m$. (For example, $W(t, x) = k(|x_1 - \sigma_t(x_2)|)$ for some $k \in K$.) Then $p_t \colon R^m \to H_t$ parametrizes H_t where $p_t(x_2) = (\sigma_t(x_2), x_2)$. Thus

$$H_t = \{(\sigma_t(x_2), x_2) \mid x_2 \in R^m\},$$

and $\pi_t \colon R^n \to H_t$ can be defined by $\pi_t(x_1, x_2) = (\sigma_t(x_2), x_2)$.

For example, in Corollary 6 W is of this form with $\sigma_t \equiv 0$. As a further illustration, consider the following proposition.

PROPOSITION. *Let $u \colon R^+ \to R^1$ be continuous and bounded. Consider the two dimensional system*

$$\left.\begin{aligned} \dot{x}_1 &= -(x_1 - x_2^2) + u(t)x_2 \\ \dot{x}_2 &= -u(t)x_1 + x_2(x_1 - x_2^2) \end{aligned}\right\} f(t, x). \tag{4}$$

Then 0 *is u.a.s. if and only if there is a $\phi_0 > 0$ such that $u(t)$ is UE with $\phi_i = \phi_0$ for all i.*

Proof. Let $V(x) = \frac{1}{2}|x|^2$. Then $-\dot{V}(x) = (x_1 - x_2^2)^2 = W(x)$. Therefore, 0 is uniformly stable. Now $\sigma_t(x_2) = \sigma_0(x_2) = x_2^2$, $W(x) = (x_1 - \sigma_0(x_2))^2$, and $H_t = H_0$, the parabola $x_1 = x_2^2$, for all $t \in R^+$. Also $\pi_0 \colon R^2 \to H_0$ is defined by $\pi_0(x_1, x_2) = (x_2^2, x_2)$. (See Fig. 1.) Thus $\alpha(t, x) = x - \pi_0(x) = (x_1 - x_2^2, 0)$, and (W, α) is admissible with $\delta(t, \gamma) = \sqrt{\gamma}$.

Let $A = A_{\varepsilon_1, \varepsilon_2}$ be an annulus, and $\gamma_A^* = \varepsilon_2/2$. Let

$$c_1 = \max\{|x_2| \mid (x_1, x_2) \in A \cap S\gamma_A^*(H_0)\},$$

$$c_2 = \min\{|x_2| \mid (x_1, x_2) \in A \cap S\gamma_A^*(H_0)\},$$

$$c = \max\{|\dot{x}_2| \mid x \in A\}.$$

(See Fig. 2.) Let

$$\Omega = \{\omega(t) = (y(t)^2, y(t)) \mid y(t) \text{ smooth}, |y(t)| \geqq c_2\sqrt{1 + c_2^2}, |\dot{y}(t)| \leqq c\sqrt{1 + 4c_1^2}\}$$

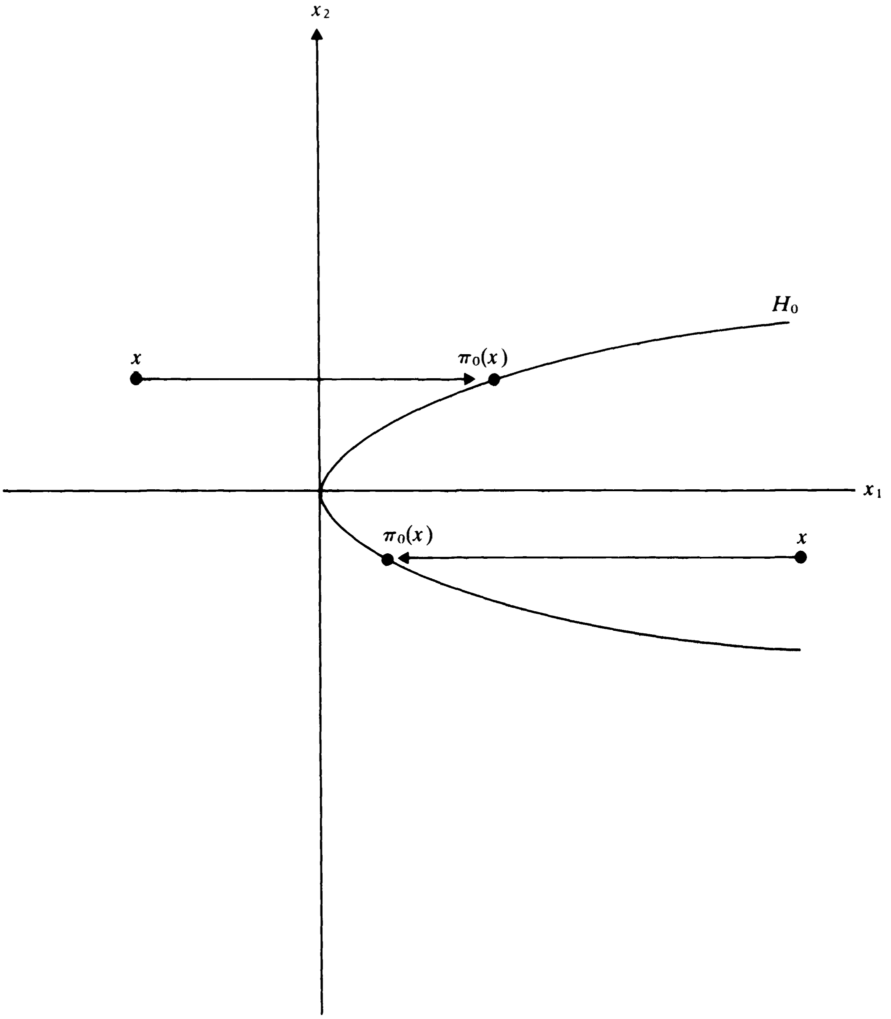

FIG. 1. *Parabola H_0 with retraction π_0: $R^2 \to H_0$ defined by $\pi_0(x_1, x_2) = (x_2^2, x_2)$.*

and

$$\theta_i(t, \gamma) = \sqrt{\gamma}.$$

Claim 1. Let $x(t)$ be a solution in $A \cap S_{\gamma \bar{A}}^{*}(H_0)$ and let $w(t) = (x_2(t)^2, x_2(t))$. Then

1. if $|\alpha(t, x(t))| \leqq \delta(t, \gamma)$, then $|x(t) - w(t)| \leqq \theta_i(t, \gamma)$,
2. $w(t) \in \Omega$.

The proof of the claim, which is very simple, is omitted.

It follows from Claim 1 that Ω approximates solutions.

Claim 2. If $u(t)$ is UE with ϕ_0, T, $s_i \to \infty$, then $\dot{\alpha}(t, x)$ is UE with respect to Ω with $\phi_i = \phi_0'$, a constant for all i, T, $s_i \to \infty$.

By Claims 1 and 2, u.a.s. follows from Theorem 4.

Proof of Claim 2. We have $\dot{\alpha}(t, x) = (f_1(t, x) - 2x_2 f_2(t, x), 0)$. Suppose $\omega(t) = (y(t)^2, y(t)) \in \Omega$. Then $\dot{\alpha}(t, \omega(t)) = (u(t)y(t)(1 + 2y(t)^2), 0)$.

Now, given index i, there is an interval $(a, b) \subseteq (s_i, s_{i+1})$ such that

$$\left| \int_a^b u(\tau)\, d\tau \right| \geqq \phi_0.$$

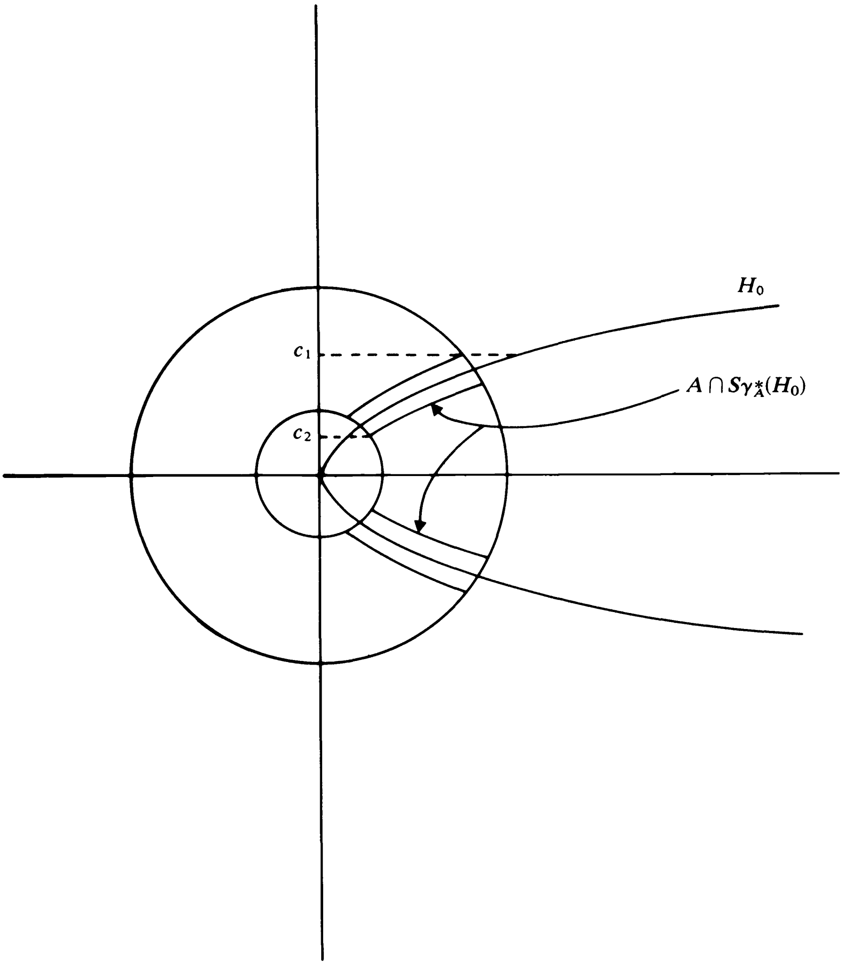

FIG. 2. *Parabola H_0 and annulus A with constants c_1 and c_2 indicated.*

Then, for each positive integer N, there is an interval $(a_N, b_N) \subseteq (a, b)$ such that $b_N - a_N = (b-a)/N$ and

$$\left| \int_{a_N}^{b_N} u(\tau)\, d\tau \right| \geqq \phi_0/N.$$

Since $\dot{y}(t)$ is uniformly bounded above by $c\sqrt{1+4c_1^2}$, as N gets large $y(t)(1+2y(t)^2)$ becomes essentially constant over the interval (a_N, b_N) and can be factored out of the integral

$$\int_{a_N}^{b_N} u(\tau)y(\tau)(1+2y(\tau)^2)\, d\tau.$$

The result follows.

This completes the proof of Claim 2 and the sufficiency part of the proposition. The necessity that $u(t)$ be UE follows from Theorem 1.

5.3. If H_t is a smooth submanifold of G, then the existence of π_t follows from elementary differential topology. The basic ideas are outlined below.

If M is an m-dimensional smooth (embedded) submanifold of R^n, then each $x_0 \in M$ has a "regular neighborhood" $N(x_0)$. In other words, $N(x_0)$ is open in R^n and contains x_0, and there is a diffeomorphism $\phi: N(x_0) \to S_1$ with $\phi(x_0) = 0$ and $\phi \,|\, M \cap N(x_0)$ a diffeomorphism of $M \cap N(x_0)$ onto $S_1 \cap (\{0\} \times R^m)$. (See Guillemin and Pollack [5, Chap. 1].)

Now S_1 retracts to $S_1 \cap (\{0\} \times R^m)$, and any such retraction induces a retraction of $N(x_0)$ to $M \cap N(x_0)$ via ϕ. If the retractions are chosen to be orthogonal projections relative to some fixed inner product in R^n, then these $N(x_0)$ can be "patched together" to cover M and provide a smooth $\pi: N(M) \to M$ with $N(M) = \bigcup N(x_0)$ where the union is over all $x_0 \in M$. This $N(M)$ is called a "tubular neighborhood" of M in R^n. (See Guillemin and Pollack [5, Chap. 2].)

Now, if $H_t \subseteq G$ is a smooth manifold for each t, then $N(H_t)$ exists with smooth $\pi_t: N(H_t) \to H_t$. (The dimension of H_t may vary with t.) Note that $N(H_t)$ and π_t depend on the choice of an inner product and on the choice of a covering of H_t by regular neighborhoods. Then $|x - \pi_t(x)|$ is the distance of x from H_t (relative to the induced metric), and the existence of $\delta(t, \gamma)$ follows from the continuity of W_t and π_t as noted in § 5.1.

For computations, the principal difficulty is in explicitly identifying π_t. Often, $N(H_t)$ can be taken to be W_t^{-1} ([0, ε)) for some ε. However, π_t must be determined by

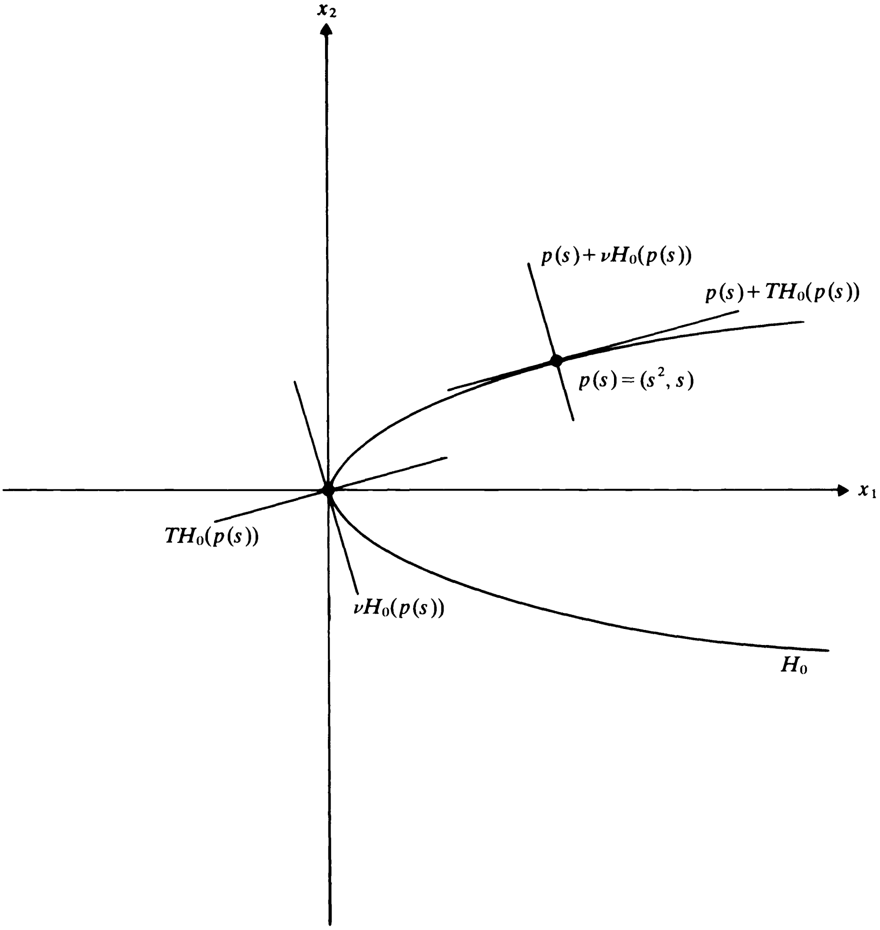

FIG. 3. *Parabola H_0 with tangent and normal lines for point $p(s)$.*

solving orthogonality equations. For example, if H_t is parametrized by a local diffeomorphism $p: R^m \to H_t$, then $TH_t(p(s)) = Dp(s)(R^m)$ is the tangent hyperplane to H_t at $p(s)$ (translated to the origin). Then $\nu H_t(p(s))$, the normal hyperplane at $p(s)$, is given by $\nu H_t(p(s)) = (Dp(s)(R^m))^{\perp}$ which can be computed explicitly as the solution set to $(n-m)\times m$ linear equations. Then $R^n = TH_t(p(s)) + \nu H_t(p(s))$ and $\pi_t(x) = p(s)$ if $x \in p(s) + \nu H_t(p(s))$ and $x \in N(H_t)$.

To illustrate these comments, let us take a second look at the proposition proven above and redefine π_0 to be the tubular neighborhood projection.

A parametrization $p: R^1 \to R^2$ for H_0 is given by $p(s) = (s^2, s)$. Then $Dp(s) = (2s, 1)$, $TH_0(p(s)) = \{c(2s, 1) \mid c \in R^1\}$, and $\nu H_0(p(s)) = \{c(-1, 2s) \mid c \in R^1\}$. (See Fig. 3.) Therefore, the normal line through $p(s) = (s^2, s)$ has the parametrized form $l_s(c) = (s^2, s) + c(-1, 2s)$, and it is easy to confirm that

$$N(H_0) = \{l_s(c) \mid s \in R^1, c > -\tfrac{1}{4}\}$$

is a well defined tubular neighborhood of H_0 with $\pi_0: N(H_0) \to H_0$ defined by $\pi_0(l_s(c)) = (s^2, s)$. (See Fig. 4.) If $(x_1, x_2) \in R^2$, then, by solving $(x_1, x_2) = l_s(c)$ for s and c,

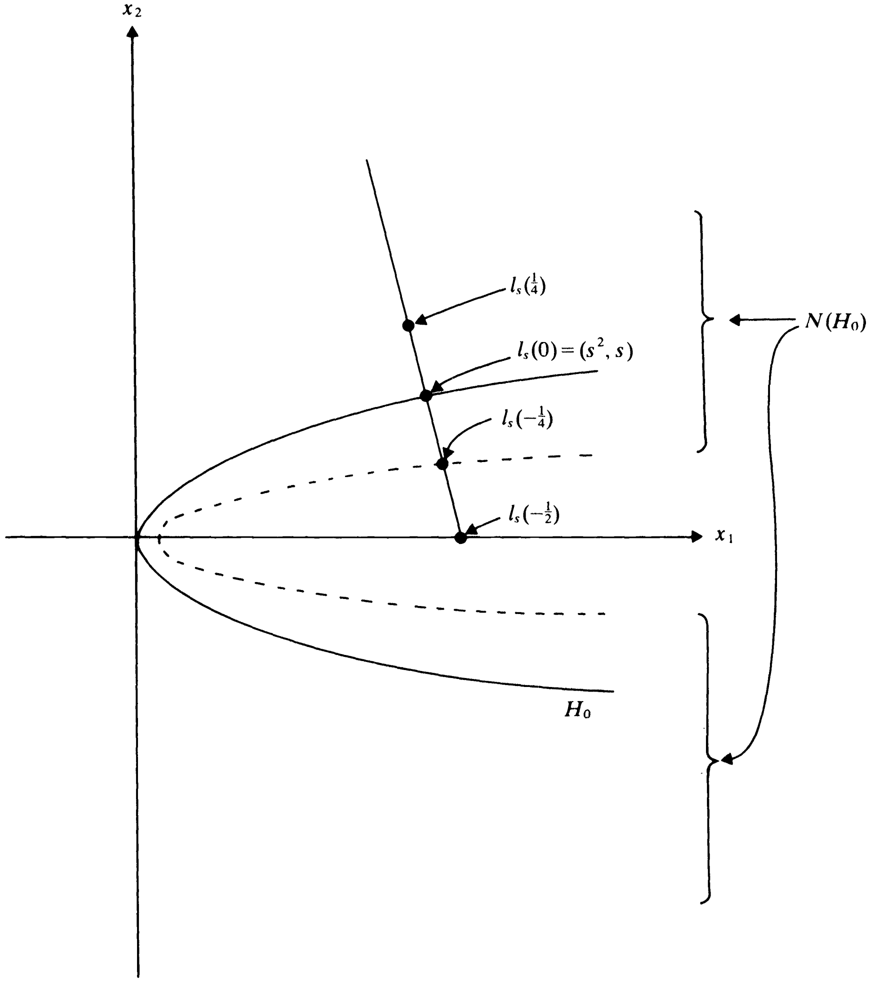

FIG. 4. *Parabola H_0 with parametrized normal line, $l_s(c) = (s^2, s) + c(-1, 2s)$. $N(H_0)$ is the region to the left of the dotted lines.*

we can write π_0 in (x_1, x_2)-coordinates. Then $\alpha(t, x_1, x_2) = (x_1, x_2) - \pi_0(x_1, x_2)$, $\dot{\alpha}(t, x_1, x_2) = (I - D\pi_0(x_1, x_2))f(x_1, x_2)$, and we can proceed to apply Theorem 4.

Now $(x_1, x_2) = l_s(c)$ reduces to the cubic equation

$$s^3 + (\tfrac{1}{2} - x_1)s - \tfrac{1}{2}x_2 = 0$$

which can be solved with the cubic formula. There are three cases, depending on the sign of

$$\left(\frac{1}{6} - \frac{x_1}{3}\right)^2 - \frac{x_2}{64}.$$

We will not continue with this analysis. Clearly the previous discussion is computationally much simpler.

6. Generalizations of the main theorems. In this section, some modifications of the material in § 3 are outlined. In particular, we will see that the theorems can be extended to cases in which W and α are piecewise continuous. The interest in such extensions is motivated by the use of piecewise constant and other piecewise smooth system elements in control applications. There is no attempt at completeness here. The results below are only illustrative of some types of possible extensions.

6.1. In this subsection, we see that Theorem 4 holds for cases in which $\dot{\alpha}(t, x)$ does not exist for all t or does not obey assumption B for all t. First we need the following.

Assumption C'. There is a positive constant L, sequence $t_i \to \infty$, and $k \in K$ such that

1. $t_{i+1} - t_i \geqq L$ for all i, and
2. for each index i, $\dot{\alpha}(t, x)$ exists for $(t, x) \in (t_i, t_{i+1}) \times G$ and $|\dot{\alpha}(t, x) - \dot{\alpha}(t, y)| \leqq k(|x - y|)$ for $t \in (t_i, t_{i+1})$, $x, y \in G$.

THEOREM 4 (version 2). *The conclusion of Theorem* 3 *remains true exactly as written if we replace condition* 2 *of the hypothesis by*

2′. *α obeys Assumption C' with L, t_i, and $k \in K$*

and replace 3(b) *by*

3(b)′. *$\dot{\alpha}(t, x)$ is UE with respect to Ω with $\phi_i = m(2\delta_i(\gamma_0) + k(\theta_i(\gamma_1))T)$, T, s_i where* $m = [T/L] + 1$.

Proof. This is a simple modification of the proof of Theorem 4. There is an interval $(a, b) \subseteq (s_i, s_{i+1})$ such that

$$\left| \int_a^b \dot{\alpha}(\tau, \omega(\tau))\, d\tau \right| \geqq \phi_i.$$

Define $(a_j, b_j) = (a, b) \cap (t_j, t_{j+1})$. At most m of these (a_j, b_j) are nonempty. Then there is some j_0 such that

$$\left| \int_{a_{j_0}}^{b_{j_0}} \dot{\alpha}(\tau, \omega(\tau))\, d\tau \right| \geqq \phi_i/m.$$

By Assumption C', $|\dot{\alpha}(t, x) - \dot{\alpha}(t, y)| \leqq k(|x - y|)$ for $t \in (a_{j_0}, b_{j_0})$ and $x, y \in G$. Now continue as in the proof of Theorem 4.

6.2. In this subsection, we see that Theorem 2 holds for W which do not obey Assumption D but rather obey a more general Assumption D′. In particular, W may be piecewise continuous. It follows that Theorems 3, 4, 5, and 6 extend to this more general case. First we need the following.

Assumption D′. Given annulus A, there are constant L, sequence $t_i \to \infty$, and $\eta \in K$ such that

1. $t_{i+1} - t_i \geqq L$ for all i,

2. if $x(t)$ is a solution and there is an index i such that $x(t) \in A$ for $t \in [a, b] \subseteq (t_i, t_{i+1})$, then

$$|W(b, x(b)) - W(a, x(a))| \leqq \eta(|b-a|),$$

and

3. if $x(t)$ is a solution, $W(t_i, x(t_i)) > \gamma$ for some index i and constant $\gamma > 0$, and there is an $\varepsilon > 0$ with $x(t) \in A$ for $t \in (t_i - \varepsilon, t_i + \varepsilon)$, then there is a δ with $0 < \delta \leqq \varepsilon$ such that either $W(t_i - \delta, x(t_i - \delta)) \geqq \gamma$ or $W(t_i + \delta, x(t_i + \delta)) \geqq \gamma$.

Compare this restriction on W with the class $P[0, \infty)$ defined by Yuan and Wonham [26].

LEMMA (version 2). *Assume W obeys Assumption* D′ *with η, L, and $t_i \to \infty$ for annulus A. Suppose there is an index i and solution $x(t) \in A$ for $t \in (t_i, t_{i+1})$, and suppose there is a $t_* \in (t_i, t_{i+1})$ and $\gamma > 0$ such that $W(t_*, x(t_*)) \geqq \gamma$.*

Then there is an interval $I \subseteq (t_i, t_{i+1})$ such that $W(t, x(t)) \geqq \gamma/2$ for $t \in I$ and the length of I is at least $\min\{L/2, \eta^{-1}(\gamma/2)\}$.

The proof is essentially the same as that of the lemma in § 3.

THEOREM 2 (version 2). *If W obeys Assumption* D′ *instead of Assumption* D, *then Theorem* 2 *remains true as stated except that the given upper bound for the rate of persistence must be replaced by*

$$\left[\frac{2b_0}{\gamma d_0} + 1\right](2L + 3T).$$

Proof. Define a sequence of positive integers $k_1, k_2, \cdots, k_m, \cdots$ and a sequence of nonnegative real numbers $\sigma_1, \sigma_2, \cdots, \sigma_m, \cdots$ as follows:

$$\sigma_0 = 0,$$

$$\sigma_r = \sigma_{r-1} + (2k_r + 1)$$

where

$$L \leqq s_{\sigma_{r-1}} + k_r - s_{\sigma_{r-1}} \leqq L + T$$

and

$$L \leqq s_{\sigma_r} - s_{\sigma_{r-1}} + k_r + 1 \leqq L + T \quad \text{for } r > 0.$$

Note that $\sigma_r - \sigma_{r-1} \leqq 2L + 3T$ for all r.

Let r be fixed. Suppose $x(t) \in A$ for $t \in [\sigma_{r-1}, \sigma_r]$. Then there is an interval $(a, b) \subseteq (\sigma_{r-1}, \sigma_r)$ such that $b - a \geqq L$, $W(t_*, x(t_*)) \geqq \gamma$ for some $t_* \in (a, b)$, and $|W(t, x(t)) - W(s, x(s))| \leqq \eta(|t-s|)$ for $t, s \in (a, b)$.

Applying the lemma, we get

$$\int_a^b W(\tau, x(\tau))\, d\tau \geqq \frac{\gamma}{2} d_0.$$

Thus

$$\int_{\sigma_{r-1}}^{\sigma_r} W(\tau, x(\tau))\, d\tau \geqq \frac{\gamma}{2} d_0 \quad \text{for } r > 0.$$

Let $M=2L+3T$. Let $m>0$ be an integer. Suppose $x(t)$ is a solution in A for $t\in[t_0, t_0+mM]=I$ for some $t_0\in R^+$. Then there are $m-1$ intervals (σ_{j-1}, σ_j) contained in I. Thus

$$\int_{t_0}^{t_0+mM} W(\tau, x(\tau))\,d\tau \geqq (m-1)\frac{\gamma}{2}d_0.$$

But then $b_0\geqq(m-1)(\gamma/2)\,d_0$, implying $(2b_0/(\gamma d_0))+1\geqq m$.

6.3. In this subsection, excitedness conditions PTUE and UE are modified to PTUE′ and UE′ and Definitions 5 and 6 are replaced by 5′ and 6′. Then Theorems 2, 3, and 4 become Theorems 2′, 3′, and 4′. No version of Assumption D for W is needed; α is required to obey a version of Assumption B (Assumption C, respectively) for Theorem 3′ (Theorem 4′).

The main modification of the § 3 material outlined below is to replace pointwise equations with integral equations. In some contexts, the resulting excitedness conditions are more natural. For example, the main theorem in Morgan and Narendra [19] is essentially the version of Corollary 5 that would follow from Theorem 3′ below.

First we need the new excitedness conditions.

DEFINITION 5′. Let $\alpha\colon R^+\times G\to R^n$ be measurable. Let Ω be given. Let T be a positive constant and s_i and ϕ_i sequences of positive numbers with $s_i\to\infty$.

Then "α is PTUE′ with respect to Ω with T, ϕ_i, and s_i" means

1. $s_{i+1}-s_i\leqq T$ for all i,
2. given index i and $\omega\in\Omega$ with $\omega(t)$ defined for $t\in[s_i, s_{i+1}]$, then

$$\int_{s_i}^{s_{i+1}} |\alpha(\tau, \omega(\tau))|\,d\tau>\phi_i.$$

DEFINITION 6′. Let $\alpha\colon R^+\times G\to R^n$ be measurable. Let Ω be given. Let T be a positive constant, s_i, a_i, and ϕ_i sequences of positive numbers with $s_i\to\infty$.

Then "α is UE′ with respect to Ω with T, ϕ_i, a_i, s_i" means

1. $s_{i+1}-s_i\leqq T$ for all i,
2. given index i and $\omega\in\Omega$ with $\omega(t)$ defined for $t\in[s_i, s_{i+1}]$, then

$$\int_{s_i}^{s_{i+1}} \left|\int_{a_i}^{\tau} \alpha(\sigma, \omega(\sigma))\,d\sigma\right| d\tau>\phi_i.$$

THEOREM 2′. *Theorem 2 holds if PTUE is replaced by PTUE′ and the rate of persistence bound is replaced by* $[b_0/\gamma]2T$.

The proof is a simple modification of the proof of Theorem 2.

Now we need some more definitions.

Assumption B′. There is a $k\in K$ such that $|\alpha(t, x)-\alpha(t, y)|\leqq k(|x-y|)$ for x, $y\in G$, a.e. $t\in R^+$. Further, given $T>0$, there is a $\tilde{k}\in K$ such that

$$\int_0^t k(s(\tau))\,d\tau\leqq\tilde{k}\left(\int_0^t s(\tau)\,d\tau\right)$$

for any continuous $s\colon[0, t]\to G$ with $t\leqq T$.

For example, if $k(s)=s^2$, then $k(s)=Ns$ where $G\subseteq S_N$. Also, if $k(s)=\sqrt{s}$, then $k(s)=\sqrt{T}\sqrt{s}$.

DEFINITION 7′. Replace (a) and (b) in Definition 7 by

(a′) if $0\leqq\gamma\leqq\gamma_0^*$, $x(t)$ is a solution and $x(t)\in N(H_t)$ for $t\in[t_1, t_2]$, and

$$\int_{t_1}^{t_2} |\alpha(\tau, x(\tau))|\,d\tau>\int_{t_1}^{t_2} \delta(\tau, \gamma)\,d\tau,$$

then

$$\int_{t_1}^{t_2} W(\tau, x(\tau))\, d\tau > \gamma,$$

(b′) if $x(t)$ is a solution for $t \in [t_1, t_2]$ and there is a $t_0 \in [t_1, t_2]$ such that $x(t_0) \notin N(H_t)$, then

$$\int_{t_1}^{t_2} W(\tau, x(\tau))\, d\tau > \gamma_0^*.$$

Definition 8′. Replace (b) in Definition 8 by

(b′). If $x(t)$ is a solution to (1) with $x(t) \in N(H_t) \cap A$ for all $t \in [s_i, s_{i+1}]$ and γ is a number with $0 < \gamma \leqq \gamma_1^*$ and

$$\int_{s_i}^{s_{i+1}} |\alpha(\tau, x(\tau))|\, d\tau \leqq \int_{s_i}^{s_{i+1}} \delta(\tau, \gamma)\, d\tau,$$

then there is an $\omega \in \Omega$ with

$$\int_{s_i}^{s_{i+1}} |x(\tau) - \omega(\tau)|\, d\tau \leqq \int_{s_i}^{s_{i+1}} \theta_i(\tau, \gamma)\, d\tau.$$

Theorem 3′. *Theorem* 3 *holds with the above Definitions* 7′ *and* 8′ *for "admissible" and "Ω approximates solution", Assumption* B′ *replacing Assumption* B_0, *PTUE′ replacing PTUE, and with*

$$\phi_i = \int_{s_i}^{s_{i+1}} \delta(\tau, \gamma_0)\, d\tau + \tilde{k}\left(\int_{s_i}^{s_{i+1}} \theta_i(\tau, \gamma_1)\, d\tau\right)$$

and rate of persistence bound $[b_0/\gamma]2T$.

Note that we need the $\tilde{k}$ from Assumption B′ to define ϕ_i.

Theorem 4′. *Theorem* 4 *holds with Assumption* B′ *replacing Assumption* B *and*

$$\phi_i = \int_{s_i}^{s_{i+1}} \delta(\tau, \gamma_0)\, d\tau + \int_{s_i}^{s_{i+1}} \tilde{k}\left(\int_{s_i}^{\tau} \theta_i(\sigma, \gamma_1)\, d\sigma\right) d\tau + (\max\{|\alpha(a_i, x)| \mid x \in A\})T.$$

The proofs are simple modifications of the proofs of Theorems 3 and 4.

6.4. The following fact may be useful for certain examples. All parameters and functions not now explicitly defined to depend on annulus A can be allowed to do so. In particular, this is true of $N(H_t)$, as well as V and W. The proofs will be unchanged. The choice of what to define as depending on A in § 3 was based on "naturalness" relative to the key examples.

One delicate point here, however, is that if $k \in K$ for Assumption B is to depend on annulus A, then the $\omega \in \Omega$ must be in A, which is sometimes an inconvenience. (See the proof of Theorem 3.) However, given A, there is generally an $A' \supseteq A$ with $\omega(t) \in A'$, and k can be chosen to depend on this A'.

6.5. In this last subsection, let us note that there is an asymptotic stability version of each of the results above (except Theorem 1). By simply omitting mention of the "T" which consistently occurs and replacing "u.a.s." by "asymptotically stable", a series of true theorems follow with proofs essentially the same as before.

7. Control motivation. In this section the control motivation for the material in previous sections is presented. In particular, necessary and sufficient conditions for the convergence of adaptive identification and control schemes from Lion [15] and Narendra and Kudva [22] are given.

7.1. The identification problem. Let us consider the following model reference identification problem. Let

$$f_i(t,x)\colon R^+\times R^n\to R^n \quad \text{for } i=1,2,\cdots,r,$$
$$g_j(t,u)\colon R^+\times R^m\to R^n \quad \text{for } j=1,2,\cdots,s,$$

be measurable functions.

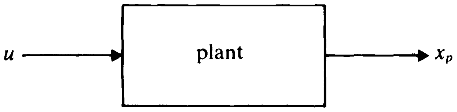

FIG. 5

The plant whose parameters we wish to identify with input $u(t)\in R^m$ and output $x_p(t)\in R^n$ is given by the (nonlinear) equation

$$\dot{x}_p=\sum_{i=1}^{r} a_i f_i(t,x_p)+\sum_{j=1}^{s} b_j g_j(t,u)$$

where $a_1,\cdots,a_r$ and $b_1,\cdots,b_s$ are constants. (See Fig. 5.) The problem is to determine $a_1,\cdots,a_r$ and $b_1,\cdots,b_s$ from the input-output pair $(u(t),x_p(t))$. The method is to compare the plant with a model given by

$$\dot{x}_m=\sum_{i=1}^{r} \alpha_i f_i(t,x_m)+\sum_{j=1}^{s} \beta_j g_j(t,u)$$

or some similar equation. We compare $(u(t),x_p(t))$ with $(u(t),x_m(t))$ adjusting the parameters α_i, β_j via "adaptive identification laws" on $\dot{\alpha}_i$ and $\dot{\beta}_j$ until $\alpha_i\to a_i$ and $\beta_j\to b_j$. (See Fig. 6.) Let us examine two different identification schemes.

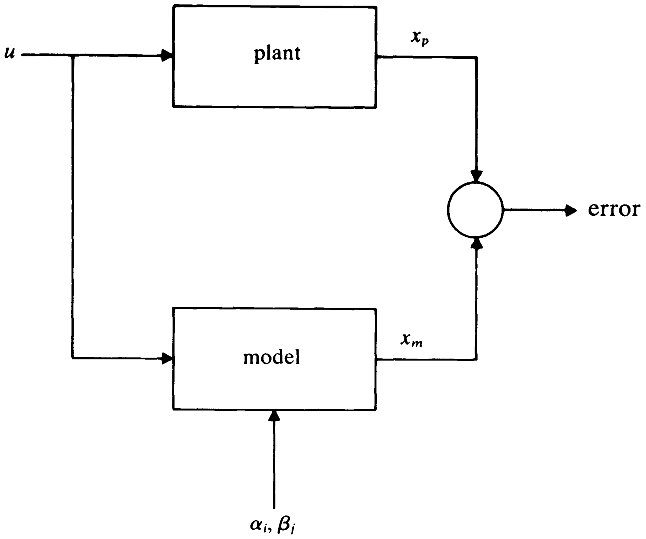

FIG. 6. *Model reference identification scheme.*

7.2. Lion's adaptive observer. The following approach appeared in Lion [15]. Define an error function using the model equation and the plant output:

$$\begin{aligned} e &= \sum \alpha_i f_i(t, x_p) + \sum \beta_j g_j(t, u) - \dot{x}_p \\ &= \sum (\alpha_i - a_i) f_i(t, x_p) + \sum (\beta_j - b_j) g_j(t, u) \\ &= \sum (\Delta\alpha_i) f_i(t, x_p) + \sum (\Delta\beta_j) g_j(t, u). \end{aligned}$$

Define the adaptive laws by

$$\Delta\dot{\alpha}_i \frac{-1}{2} \frac{\partial e^T e}{\partial \alpha_i} = -e^T f_i(t, x_p),$$

$$\Delta\dot{\beta}_j = \frac{-1}{2} \frac{\partial e^T e}{\partial \beta_j} = -e^T g_j(t, u).$$

(Lion describes this as a "steepest descent law" on the surface $F = e^T e$.)

Letting $\Delta\alpha = (\Delta\alpha_1, \cdots . \Delta\alpha_r)^T$, $\Delta\beta = (\Delta\beta_1, \cdots, \Delta\beta_s)^T$, $f = (f_1^T, \cdots, f_r^T)^T$, and $g = (g_1^T, \cdots, g_s^T)^T$, we have

$$\begin{pmatrix} \Delta\dot{\alpha} \\ \Delta\dot{\beta} \end{pmatrix} = -\begin{pmatrix} f \\ g \end{pmatrix} (f^T \quad g^T) \begin{pmatrix} \Delta\alpha \\ \Delta\beta \end{pmatrix}. \tag{5}$$

The stability of this type of system is characterized by Corollary 5. We conclude that 0 is u.a.s. for (5) if and only if the time varying matrix $(f(t, x_p(t))^T \quad g(t, u(t))^T)$ is PTUE. (This assumes that the various uniform continuity conditions required by Corollary 5 hold. But note also the comments in §§ 3 and 4 on weakening these conditions.)

Thus the scheme converges uniformly if and only if the above PTUE condition holds. Lion observes that if f_i, g_j, x_p, and u are all periodic, then the result of LaSalle noted after Theorem 2 implies the sufficiency of the above condition for u.a.s. See also Morgan and Narendra [19].

7.3. Narendra and Kudva's identification scheme. Narendra and Kudva [22] use a slightly different model equation and error function, as follows:

$$\begin{aligned} \dot{x}_m &= -(x_m - x_p) + \sum \alpha_i f_i(t, x_p) + \sum \beta_j g_j(t, u), \\ e &= x_m - x_p, \\ \dot{e} &= \dot{x}_m - \dot{x}_p = -e + \sum (\alpha_i - a_i) f_i(t, x_p) + \sum (\beta_j - b_j) g_j(t, u) \\ &= -e + \sum \Delta\alpha_i f_i(t, x_p) + \sum \Delta\beta_j g_j(t, u). \end{aligned}$$

The adaptive identification laws are defined by

$$\Delta\dot{\alpha}_i = -f_i^T e,$$

$$\Delta\dot{\beta}_j = -g_j^T e.$$

Letting $f = (f_1, \cdots, f_r)$, $g = (g_1, \cdots, g_s)$, $\Delta\alpha = (\alpha_1, \cdots, \alpha_r)^T$, and $\Delta\beta = (\beta_1, \cdots, \beta_s)^T$, we have

$$\begin{pmatrix} \dot{e} \\ \Delta\dot{\alpha} \\ \Delta\dot{\beta} \end{pmatrix} = \begin{pmatrix} -I & f & g \\ -f^T & 0 & 0 \\ -g^T & 0 & 0 \end{pmatrix} \begin{pmatrix} e \\ \Delta\alpha \\ \Delta\beta \end{pmatrix}. \tag{6}$$

The stability of this type of system is characterized by Corollary 6. It follows that 0 is u.a.s. for (6) if and only if the time varying matrix $(f(t, x_p(t)) \quad g(t, u(t)))$ is UE. (This assumes that f and g are uniformly bounded in t.)

Narendra and Kudva also note the relevance of the result of LaSalle cited above if (6) is periodic. Yuan and Wonham in [26] give some nonperiodic sufficient conditions for the asymptotic stability of (6). See also Morgan and Narendra [20].

7.4. The adaptive control problem. The following description of the adaptive controller is derived from material in Narendra and Kudva [22]. Although not the most general formulation, it does serve to illustrate how the theorems from previous sections relate to this type of problem.

Consider

$$\dot{x}_p = (A_p(t) + B_p(t)Q(t)F(t))x_p + B_p(t)Q(t)u \qquad \text{(plant)},$$
$$\dot{x}_m = A_m(t)x_m + B_m(t)u \qquad \text{(model)},$$

where F and Q represent feedback and feedforward respectively. (See Fig. 7.) Q, F, A_m, B_m, x_m, u, and x_p are accessible, but A_p and B_p are not. We want to specify a scheme that continuously adjusts $Q(t)$ and $F(t)$ so that $|x_p(t) - x_m(t)|$ converges to 0. Thus A_m, B_m, A_p, B_p, and u are taken as given, and Q and F can be chosen. The adaptive control laws will be choices of $\dot{Q}$ and $\dot{F}$ which may involve A_m, B_m, x_m, u, and x_p but which should not (directly) involve A_p or B_p. Once these laws are defined, we want to find necessary and sufficient conditions on the plant and model for the uniform convergence of the scheme.

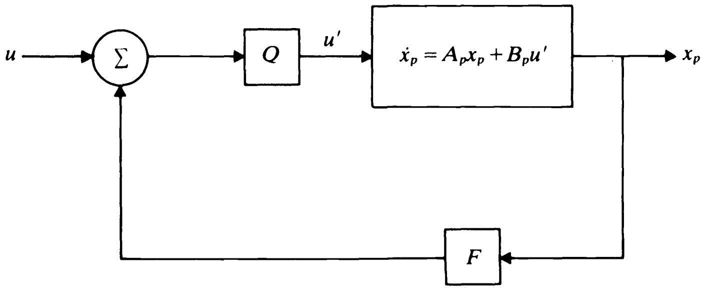

FIG. 7. *Configuration of plant with feedforward and feedback gain matrices (from Narendra and Kudva* [22]).

It is assumed that $\dot{x} = A_m(t)x$ is u.a.s. at 0 and that there are constant matrices Q^* and F^* such that $B_p(t)Q^* = B_m(t)$ and $B_m(t)F^* = A_m(t) - A_p(t)$ for all $t \in R^+$. (See Narendra and Kudva [22] for a discussion of the control significance of these matching assumptions.) Let $P(t)$ and $R(t)$ be positive definite matrices such that

$$\dot{P}(t) + P(t)A_m(t) + A_m(t)^T P(t) = -R(t).$$

We further assume Q^* is symmetric positive definite. Although there are many important examples that satisfy this last requirement, we would rather Q^* be a general invertible square matrix. However, this seems to lead to various technical problems in finding a globally defined Lyapunov function. It is convenient to take $Q = Q^*(I - Q_0)$ and $F = F^* - F_0$.

7.5. A Narendra and Kudva type controller. Define the error function $e = x_m - x_p$. Then

$$\begin{aligned}\dot{e} = \dot{x}_m - \dot{x}_p &= A_m e + A_m x_p - A_p x_p - B_p Q F x_p + B_m u - B_p Q u \\ &= A_m e + B_m F_0 x_p + B_m Q_0 (F x_p + u).\end{aligned}$$

Choosing

$$\dot{F}_0 = -B_m^T P e x_p^T,$$
$$\dot{Q}_0 = -Q^{*-1} B_m^T P e (F x_p + u)^T,$$

we find that the resulting (quadratic) system $(\dot{e}, \dot{F}_0, \dot{Q}_0)$ has a Lyapunov function

$$V(t, e, F_0, Q_0) = \tfrac{1}{2}(e^T P(t) e + \operatorname{tr}(F_0^T F_0 + Q_0^T Q^* Q_0))$$

with

$$\dot{V}(t, e, F_0, Q_0) = -e^T R(t) e.$$

Therefore, 0 is uniformly stable. Now, the uniform convergence of the control scheme is equivalent to 0 being u.a.s. The u.a.s. of 0, however, is completely characterized by Theorems 1 and 6. Thus, 0 is u.a.s. for $(\dot{e}, \dot{F}_0, \dot{Q}_0)$ on G if and only if the function

$$\alpha(t, F_0, Q_0) \equiv B_m(t) F_0 x_p(t) + B_m(t) Q_0((F^* - F_0) x_p(t) + u(t))$$

is UE with respect to constants in $G \cap A$ for all annular regions A about 0.

The resulting adaptive control laws are

$$\dot{F} = B_m^T P e x_p^T,$$
$$\dot{Q} = B_m^T P e (F x_p + u)^T.$$

7.6. Another controller. With the adaptive control problem formulated as in § 7.4 above, define the error equation

$$e = A_m x_p + B_m u - \dot{x}_p = B_m F_0 x_p + B_m Q_0 (F x_p + u).$$

This suggests

$$\dot{F}_0 = -B_m^T e x_p^T,$$
$$\dot{Q}_0 = -Q^{*-1} B_m^T (F x_p + u)^T$$

as in the Lion observer, yielding the adaptive control laws

$$\dot{F} = B_m^T e x_p^T,$$
$$\dot{Q} = B_m^T e (F x_p + u)^T.$$

By definition of Lyapunov function $V(t, F_0, G_0) = \operatorname{tr}(F_0^T F_0 + Q_0^T Q^* Q_0)$, it is routine to confirm that, given annulus A, there is a constant c such that $-\dot{V}(t, F_0, Q_0) \geqq c \operatorname{tr}(\dot{F}_0^T \dot{F}_0 + \dot{Q}_0^T \dot{Q}_0)$ for all $(F_0, Q_0) \in A$, where c depends on the annulus and the upper bounds of B_m, Q^*, F^*, x_p, and u. Therefore the scheme converges uniformly for $(F_0, Q_0) \in G$ if and only if the function

$$\alpha(t, F_0, Q_0) = \operatorname{tr}(\dot{F}_0^T \dot{F}_0 + \dot{Q}_0^T \dot{Q}_0) = |x_p(t)|^2 |B_m^T(t) B_m(t)(F_0 x_p(t) + Q_0 y(t))|^2$$
$$+ |y(t)|^2 |Q^{*-1} B_m^T(t) B_m(t)(F_0 x_p(t) + Q_0 y(t))|^2,$$

where $y(t) = F x_p(t) + u(t) = (F^* - F_0) x_p(t) + u(t)$, is PTUE with respect to constants in $G \cap A$ for all annular regions A about 0. See Theorem 5.

8. Proof of Theorem 1. This section contains the proof of Theorem 1. (Compare the proof of Theorem 3 in Morgan and Narendra [20].)

8.1. We need the following strong u.a.s. converse theorem of Massera [16], cited on pp. 244–245 of Hahn [9].

Let the differential equation $\dot{x} = f(t, x)$ have a u.a.s. equilibrium and suppose that in a domain $R^+ \times G$ the right side satisfies a Lipschitz condition. Then there exists in $R^+ \times G$ a positive definite decrescent Lyapunov function with a negative definite derivative, which has partial derivatives of any order desired with respect to all of its variables. If there exists a uniform Lipschitz constant, then V can be so determined that all the partial derivatives are uniformly bounded and that, in fact, the same bound can be used everywhere.

8.2. The following technical lemma is also needed.

LEMMA. *Let $\omega: R^+ \to R^1$ be measurable and bounded. Assume there are constants $a > 0$ and $b > 0$ such that*

$$\left| \int_{t_0}^{t} \omega(\tau)\, d\tau \right| \geqq a(t - t_0) - b$$

for all $t \geqq t_0 \geqq 0$.

Then there are positive constants δ_1, ϕ, and T such that if $t_0 \geqq 0$, then there is $t_1 \in [t_0, t_0 + T]$ such that

$$\left| \int_{t_1}^{t_1+\delta} \omega(\tau)\, d\tau \right| \geqq \phi\delta$$

for all δ with $0 \leqq \delta \leqq \delta_1$.

The (easy) proof is in Morgan and Narendra [20].

8.3. Since $x = 0$ is u.a.s. and $f(t, x)$ obeys a uniform Lipschitz condition, Massera's theorem guarantees the existence of Lyapunov function V with the listed properties. Let $A = A_{\varepsilon_1, \varepsilon_2}$ be an annular region about 0, and define $S = G \cap A$. Thus, since S is bounded, we have constant $k_0 > 0$ such that $|x| \leqq k_0$,

$$\left| \frac{\partial V}{\partial x}(t, x) \right| \leqq k_0, \quad \text{and} \quad \left| \frac{\partial^2 V}{\partial t\, \partial x}(t, x) \right| \leqq k_0 \quad \text{for all } x \in S \text{ and } t \in R^+.$$

We may also assume $|f(t, x)| \leqq k_0 |x|$ for all $x \in S$ and $t \in R^+$. Since V is positive definite decrescent, there are positive constants c_2 and c_1 with $c_2|x|^2 \leqq V(t, x) \leqq c_1|x|^2$ for all $x \in S$ and $t \in R^+$. Also there is a positive constant c_3 such that $-\dot{V}(t, x) \geqq c_3|x|^2$ for all $x \in S$ and $t \in R^+$.

8.4. Choose $x_0 \in S$. Consider the perturbed system

$$(1') \qquad \dot{x} = f(t, x) - f(t, x_0).$$

Then $x_0(t) \equiv x_0$ is a (constant) solution to (1'), and

$$\dot{V}_{1'}(t, x) = \dot{V}_1(t, x) - \frac{\partial V}{\partial x}(t, x) f(t, x_0).$$

Thus

$$\frac{\partial V}{\partial t}(t, x_0) = \dot{V}_1(t, x_0) - \frac{\partial V}{\partial x}(t, x_0) f(t, x_0)$$

from which the inequality

$$\left| \int_{t_0}^{t} \frac{\partial V}{\partial x}(\tau, x_0) f(\tau, x_0)\, d\tau \right| \geqq c_3 \varepsilon_2^2 (t - t_0) - c_1 \varepsilon_1^2$$

follows easily.

8.5. Now we apply the lemma from § 8.2 above to conclude that there are constants δ_1, ϕ, and T such that if $t_0 \in R^+$, then there is some $t_1 \in [t_0, t_0 + T]$ with

$$\left| \int_{t_1}^{t_1+\delta} \frac{\partial V}{\partial x}(\tau, x_0) f(\tau, x_0)\, d\tau \right| \geqq \phi\delta$$

for all δ with $0 < \delta \leqq \delta_1$.

8.6. Now $|(\partial^2 V/\partial t\, \partial x)(t, x)| \leqq k_0$ for all $x \in S$ and $t \in R^+$ implies that

$$\left| \frac{\partial V}{\partial x}(\tau, x_0) - \frac{\partial V}{\partial x}(t_1, x_0) \right| \leqq k_0(\tau - t_1) \quad \text{for any } \tau \geqq t_1 \geqq 0.$$

Thus it follows that

$$\left| \int_{t_1}^{t_1+\delta} \left(\frac{\partial V}{\partial x}(\tau, x_0) f(\tau, x_0) - \frac{\partial V}{\partial x}(t_1, x_0) f(\tau, x_0) \right) d\tau \right| \leqq \frac{1}{2} k_0^3 \delta^2.$$

Therefore,

$$\left| \int_{t_1}^{t_1+\delta} \frac{\partial V}{\partial x}(\tau, x_0) f(\tau, x_0)\, d\tau \right| - \frac{1}{2} k_0^3 \delta^2 \leqq k_0 \left| \int_{t_1}^{t_1+\delta} f(\tau, x_0)\, d\tau \right|.$$

8.7. Now, choosing $t_1 \in [t_0, t_0 + T]$ as in § 8.5, we have

$$\left| \int_{t_1}^{t_1+\delta} f(\tau, x_0)\, d\tau \right| \geqq \frac{\phi\delta}{k_0} - \frac{1}{2} k_0^2 \delta^2$$

for all δ with $0 < \delta \leqq \delta_1$. Thus it is clear that there is a δ_0 with $0 < \delta_0 \leqq \delta_1$ such that the right-hand side of the above expression is positive. This δ_0 is clearly independent of the choice of $x_0 \in S$.

Acknowledgment. I would like to thank the referee for his helpful suggestions to improve the readability of this paper. Also I would like to express my appreciation to my wife, Ann S. Morgan, for drawing the figures.

REFERENCES

[1] B. D. O. ANDERSON, *Exponential stability of linear equations arising in adaptive identification*, IEEE Trans. Automatic Control, AC-22 (1977), pp. 83–87.

[2] Z. ARTSTEIN, *Uniform asymptotic stability via the limiting equations*, preprint, Dept. of Mathematics, The Weizmann Institute of Science, Rehovot, Israel, 1976.

[3] R. BELLMAN, *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1960.

[4] T. A. BURTON, *An extension of Lyapunov's direct method*, J. Math. Anal. Appl., 28 (1969), pp. 545–552.

[5] V. GUILLEMIN AND A. POLLACK, *Differential Topology*, Prentice-Hall, Englewood Cliffs, NJ, 1974.

[6] J. R. HADDOCK, *Some refinements of asymptotic stability theory*, Ann. Mat. Pura Appl., LXXXIX (1971), pp. 393–402.

[7] ———, *On Lyapunov functions for nonautonomous systems*, J. Math. Anal. Appl., 47 (1974), pp. 599–603.

[8] ———, *Stability theory for nonautonomous systems*, Proc. Internat. Conf. Diff. Eqs., Brown Univ., Providence, RI, August 1974.

[9] W. HAHN, *The Stability of Motion*, Springer-Verlag, Berlin, 1969.

[10] J. K. HALE, *Ordinary Differential Equations*, Wiley-Interscience, New York, 1969.

[11] G. KREISSELMEIER, *Adaptive observers with exponential rate of convergence*, IEEE Trans. Automatic Control, AC-22 (1977), pp. 2–8.

[12] J. P. LASALLE, *Asymptotic stability criteria*, Proceedings of the Symposia in Appl. Math., Hydrodynamic Stability, vol. 13, Amer. Math. Soc., Providence, RI, 1962, pp. 299–307.

[13] ———. *Stability theory for ordinary differential equations*, J. Differential Equations, 4 (1968), pp. 57–65.

[14] ———, *Stability of nonautonomous systems*, Nonlinear Analysis, 1 (1976), pp. 83–91.

[15] P. M. LION, *Rapid identification of linear and nonlinear systems*, AIAA J., 5 (1967), pp. 1835–1842.

[16] J. L. MASSERA, *Contribution to stability theory*, Ann. of Math., 64 (1956), pp. 182–206; *Erratum*, Ann. of Math., 68 (1958), p. 202.

[17] J. MILNOR, *Topology from the Differentiable Viewpoint*, University of Virginia Press, Charlottesville, VA, 1965.

[18] A. P. MORGAN, *Uniform asymptotic stability and Lyapunov functions with negative semidefinite derivatives*, Conference on Dynamical Systems, Gainesville, Florida, March 1976.

[19] A. P. MORGAN AND K. S. NARENDRA, *On the uniform asymptotic stability of certain nonautonomous linear differential equations*, this Journal, 15 (1977), pp. 5–24.

[20] ———, *On the stability of nonautonomous differential equations* $\dot{x} = [A + B(t)]x$, *with skew symmetric matrix* $B(t)$, this Journal, 15 (1977), pp. 163–176.

[21] J. R. MUNKRES, *Elementary Differential Topology*, Annals Studies 54, Princeton University Press, Princeton, NJ, 1963.

[22] K. S. NARENDRA AND P. KUDVA, *Stable adaptive schemes for system identification and control, Parts I and II*, IEEE Trans. Systems, Man, and Cybernetics, SMC-4 (1974), pp. 542–560.

[23] K. S. NARENDRA AND L. E. MCBRIDE, *Multiparameter self-optimizing systems using correlation techniques*, IEEE Trans. Automatic Control, AC-9 (1964), pp. 31–38.

[24] S. NUYAN AND R. L. CARROLL, *An adaptive observer with an arbitrarily fast rate of convergence*, Proc. 1976 IEEE Conf. on Decision and Control, Clearwater, FL, December 1976.

[25] M. SPIVAK, *A Comprehensive Introduction to Differential Geometry*, Vol. 1, Publish or Perish, Boston, MA, 1974.

[26] J. S. C. YUAN AND W. M. WONHAM, *Probing signals for model reference identification*, IEEE Trans. Automatic Control, AC-22 (1977), pp. 530–538.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 3, May 1979

0363-0129/79/1703-0006$01.00/0

CONTINUOUS DEPENDENCE OF SOLUTIONS OF A DIFFERENTIAL INCLUSION ON THE RIGHT HAND SIDE WITH APPLICATIONS TO STABILITY OF OPTIMAL CONTROL PROBLEMS*

G. I. STASSINOPOULOS† AND R. B. VINTER†

Abstract. Continuous dependence of the solution set of a differential inclusion on the right hand side is investigated. A mode of convergence for right hand sides, involving the notion of weak convergence of set-valued functions, is determined which is necessary and sufficient for convergence of the solution sets. In the case that the right hand sides depend affinely on the control, perturbations of the control constraint set with respect to which the solution set is stable are more explicitly characterized. Our results are applied to give necessary and sufficient conditions for stability of a general class of control problems under perturbations of the dynamics.

1. Introduction. We investigate continuous dependence of the solution set of the differential inclusion

$$\dot{x} \in F(t, x), \qquad x(0) = a, \tag{1.1}$$

on the right hand side, $F(\cdot, \cdot)$.

Motivation is provided by dynamical systems modeled as

$$\dot{x} = f(t, x, u), \qquad u(t) \in U(t, x), \qquad x(0) = a. \tag{1.2}$$

It is important to establish what tolerances are permitted in specification of $f(\cdot, \cdot, \cdot)$ and the control constraint set $U(\cdot, \cdot)$ in order that the set of responses $x(\cdot)$ (the solution set) achievable by the model adequately approximates that of the system. Fundamental to such questions is the sense in which the solution set depends continuously on $f(\cdot, \cdot, \cdot)$ and $U(\cdot, \cdot)$.

We find it convenient to treat for the most part differential inclusions (1.1). Perturbations in $f(\cdot, \cdot, \cdot)$ and $U(\cdot, \cdot)$ affect the solution set only in so far as they modify the velocity set $f(t, x, U(t, x))$, which defines the right hand side of a differential inclusion $F(t, x)$, and conditions for continuous dependence are most simply expressed through $F(\cdot, \cdot)$ directly. We point out also that differential inclusions provided a natural setting for study of state dependent control constraint sets $U(\cdot, \cdot)$. Some generality is provided too by giving the boundedness and Lipschitz continuity hypotheses needed here in terms of the set valued function $F(\cdot, \cdot)$. Indeed such hypotheses apply even in some special situations where $F(\cdot, \cdot)$ arises from (1.2) with $f(\cdot, \cdot, \cdot)$, $U(\cdot, \cdot)$ not continuous in their x-dependence;[1] a trivial instance occurs when $f(\cdot, \cdot, \cdot)$ is identically zero and $U(\cdot, \cdot)$ is any discontinuous set-valued function.

Conditions are first given for convergence of solution sets to the solution set of a nominal differential inclusion, in terms of convergence in some weak sense of the right hand sides evaluated on trajectories in the nominal solution set. Convergence of the solution set is understood in the sense of Hausdorff convergence with respect to the supremum norm on $C(I, \mathbb{R}^n)$. The conditions are the most general possible in that they are necessary and sufficient for convergence.

One immediate application is to models of the form (1.2) where $f(\cdot, \cdot, \cdot)$ is affine in its u-dependence, and $U(\cdot, \cdot)$ is independent of x; necessary and sufficient conditions are given for continuous dependence on $U(\cdot)$, when $f(\cdot, \cdot, \cdot)$ does not vary. This

* Received by the editors July 6, 1977, and in final revised form July 27, 1978.

† Imperial College of Science and Technology, London SW7 2BZ, England.

[1] When $f(\cdot, \cdot, \cdot)$ and $U(\cdot, \cdot)$ are not continuous in their x-dependence however, the sense in which the original equations and the corresponding differential inclusion are equivalent needs clarification (Fillipov's selection lemma does not apply in such situations). We thank a reviewer for this observation.

application generalizes results of Artstein [1] relating to models (1.2) linear in (x, u). The difficulties in achieving this generalization arise from the fact that in our wider setting the solution set is only implicitly defined; in [1] the solution set has a convenient representation through the variation of constants formula, to which results on convergence of integrals of set-valued functions may be applied directly.

A sufficient condition is provided by requiring weak convergence of the velocity sets along all constant trajectories $x(\cdot)$. Making an additional assumption, we show that this stronger, but simpler, condition is necessary if we require convergence of the solution sets for arbitrary initial conditions. This generalizes to differential inclusions a theorem of Artstein [2] for differential equations.

Finally we examine the implications of our results for stability of control problems. Necessary and sufficient conditions are given for continuous dependence of the minimal cost on parameters for arbitrary cost functions in a certain class. The same conditions imply upper semi-continuity of the set of optimal trajectories.

The debt to Artstein's work in the present paper will be evident: the main results here interpolate between those in [1] which concern reachable sets, but restrict attention to linear control systems, and those in [2] which concern nonlinear differential equations but not differential inclusions. The properties of set-valued functions and, particularly, the notion of weak convergence, introduced in [1], are extensively used. We stress however that our results require some essentially new developments. The technical core of this paper is an analogue for differential inclusions of the 'equivalent approximations' lemma for ordinary differential equations; while the lemma for ordinary differential equations follow simply from the assumptions, this is no longer the case here and some delicate constructions are involved. Thus a trivial step in [2] becomes the most serious problem in developing the broader results of this paper. Of course the need for studying equivalent approximations does not arise in [1] because of the assumption of linearity.

Convexity of the velocity sets is assumed. When the differential inclusion arises from a model (1.2) this assumption amounts to admissibility of relaxed controls. We could equally have given our results in terms of strong closures (in $C(I, \mathbb{R}^n)$) of the solution sets, and of convex hulls of the velocity sets.

There is an extensive literature dealing with stability of solution sets to differential inclusions, though much of this is concerned with dependence on initial conditions (see [12], [13] for some typical results). Hermes [9]. Granger [14] and Bridgland [4] give conditions for continuous dependence on right hand sides, and to this extent provide an overlap with results in this paper, but the perturbations considered are very restrictive compared with ours and are far from providing "necessary conditions" for continuous dependence.

It is hoped that the equivalent approximations results will find applications in the study of algorithm convergence.

2. Notation and preliminaries. The interval $[0, 1]$ is written I.

We denote by $|\cdot|$ the Euclidean norm, and by $|\cdot|_T$ the supremum norm on $C(T, \mathbb{R}^n)$ for $T \subset I$, a nontrivial closed subinterval.

The space of nonempty, convex, compact subsets of $\mathbb{R}^n$ is written $\mathcal{R}$, and the space of nonempty, compact subsets of $C(I, \mathbb{R}^n)$ is written $\mathcal{C}$.

Both $\mathcal{R}$ and $\mathcal{C}$ are endowed with their Hausdorff metrics:

$$\operatorname{dist}(A, B) = \max\{\max_{a\in A}\{\operatorname{dist}(a, B)\}, \max_{b\in B}\{\operatorname{dist}(b, A)\}\}.$$
$$(\operatorname{dist}(x, A) = \min_{a\in A} |x - a|).$$

We adopt also the notation

$$\overline{\text{dist}}\,(A,B)=\max_{a\in A}\{\text{dist}\,(a,B)\},\qquad \|A\|=\text{dist}\,(\{0\},A)$$

for asymetric distance and 'norm' on $\mathcal{R}$, $\mathcal{C}$. Functions, their values and points in their range space will be distinguished typically as $x(\cdot)$, $x(t)$, x. Upper case letters, e.g. $F(\cdot)$, denote $\mathcal{R}$-valued functions, while lower case letters, e.g. $f(\cdot)$, are reserved as symbols for $\mathbb{R}^n$-valued functions.

We term $\mathcal{R}$-valued functions $F(\cdot)$ measurable (Borel measurable) when $\{t\in I\,|\,F(t)\cap C\neq\emptyset\}$ is Lebesgue measurable (respectively Borel measurable) for $C\subset\mathbb{R}^n$, closed. (In view of the σ-compactness of $(\mathbb{R}^n,|\cdot|)$, we could equivalently have defined measurability in terms of open sets or, indeed, open balls.) If the $\mathcal{R}$-valued function $F(\cdot)$ on I is measurable, then $F(\cdot)$ may be taken Borel measurable by adjustment on a null set (this follows from [7, Thm. 1]), and the real-valued function $\|F(\cdot)\|$ is (Lebesgue) measurable [1, Cor. 2.3]. $L_1(I,\mathcal{R})$ is now introduced as the space of measurable $\mathcal{R}$-valued functions $F(\cdot)$ on I (modulo null functions) such that $\|F(\cdot)\|$ is integrable.

The $\mathbb{R}^n$-valued function $f(\cdot)$ on I is a selector of the $\mathcal{R}$-valued function $F(\cdot)$ on I when $f(\cdot)$ is measurable and $f(t)\in F(t)$, a.e. $t\in I$.

Given $F(\cdot)\in L_1(I,\mathcal{R})$ and $E\subset I$, a measurable subset, we write

$$\int_E F(t)\,dt=\left\{x\in\mathbb{R}^n\,|\,x=\int_E f(t)\,dt;\ f(\cdot)\text{ is a selector of }F(\cdot)\right\}.$$

The $\mathcal{R}$-valued function on I, defined through

$$H(t)=\int_{[0,t]}F(\tau)\,d\tau,\qquad \text{all } t\in I,$$

with $F(\cdot)\in L_1(I,\mathcal{R})$, is termed the Aumann integral.

Following [3], we introduce also the trajectory integral $\mathcal{I}F(\cdot)$ of $F(\cdot)\in L_1(I,\mathcal{R})$:

$$\mathcal{I}F(\cdot)=\{x(\cdot)\in C(I,\mathbb{R}^n)|x(\cdot)\text{ is absolutely continuous},$$
$$x(0)=0\text{ and }\dot{x}(\cdot)\text{ is a selector of }F(\cdot)\}.$$

It is important to distinguish the Aumann integral, which is an $\mathcal{R}$-valued function, from the trajectory integral, which is a subset of $C(I,\mathbb{R}^n)$.

Let Λ be an index set. We say that the subset $\{F_\lambda(\cdot)|\lambda\in\Lambda\}$ of $L_1(I,\mathcal{R})$ is

(i) bounded in $L_1(I,\mathcal{R})$ (or strongly bounded) when $\{\|F_\lambda(\cdot)\|\,|\,\lambda\in\Lambda\}$ is a bounded subset of $L_1(I,\mathcal{R})$,

(ii) uniformly integrably bounded when there exists $m_\Lambda(\cdot)\in L_1(I,\mathbb{R})$ such that $\|F_\lambda(t)\|\leqq m_\Lambda(t)$ a.e. $t\in I$, for all $\lambda\in\Lambda$,

(iii) uniformly integrable if, given $\varepsilon>0$, there exists $\eta>0$ such that if $E\subset I$ is measurable with Lebesgue measure smaller than η, then $\|\int_E F_\lambda(t)\,dt\|\leqq\varepsilon$, for all $\lambda\in\Lambda$.

We remark that since we are dealing with a finite measure space (iii) implies (i).

Following Artstein [1], we define weak convergence of sequences in $L_1(I,\mathcal{R})$:

The sequence $\{F_i(\cdot)\}_{i=1}^\infty$ converges weakly to $F_0(\cdot)$ in $L_1(I,\mathcal{R})$ when, for every measurable subset $E\subset I$, $\int_E F_i(t)\,dt$ converges to $\int_E F_0(t)\,dt$ in $\mathcal{R}$.

Since we have limited consideration to *convex* valued functions in defining $L_1(I,\mathcal{R})$, weak limits in $L_1(I,\mathcal{R})$ are unique [1, Remark 4.3]. Also the trajectory integral is a compact subset of $C(I,\mathbb{R}^n)$.

3. A class of differential inclusions. Let there be given nonnegative integrable functions $h(\cdot)$ and $k(\cdot)$ on I, and $a\in\mathbb{R}^n$. $h(\cdot)$ and $k(\cdot)$ will remain fixed throughout the paper. The point a will remain fixed except in § 9.

We set

$$P=\left\{(t,x)\in I\times\mathbb{R}^n \,\middle|\, |x-a|\leqq\int_0^t h(\tau)\,d\tau\right\}.$$

We take $\mathcal{M}$ to be the set of functions $D(\cdot,\cdot)$ defined on some relatively open subset $\mathcal{O}$ of $I\times\mathbb{R}^n$ containing P ($\mathcal{O}$ may possibly depend on $D(\cdot,\cdot)$), which satisfy:

A1: If x lies in the projection of $\mathcal{O}$ on $\mathbb{R}^n$, then $D(\cdot,x)\in L_1(T,\mathcal{R})$, for every compact subinterval $T\subset I$ such that $T\times\{x\}\subset\mathcal{O}$.

A2: For a.e. $t\in I$, $D(t,\cdot)$ satisfies

$$\operatorname{dist}(D(t,x),D(t,x'))\leqq k(t)|x-x'|$$

for all $x, x'\in\mathcal{O}_t$.

A3: $\sup_{x\in\mathcal{O}_t}\|D(t,x)\|\leqq h(t)$ a.e. $t\in I$. (In A2, A3, $\mathcal{O}_t=\{x\in\mathbb{R}^n|(t,x)\in\mathcal{O}\}$.)

The set $\mathcal{M}$ comprises the right hand sides $D(\cdot,\cdot)$ of differential inclusions of interest here. The domains of the $D(\cdot,\cdot)$'s are taken to be subsets $\mathcal{O}$ of $I\times\mathbb{R}^n$ rather than the whole of $I\times\mathbb{R}^n$ to permit consideration of certain $D(\cdot,\cdot)$'s which are, loosely speaking, unbounded in their x-dependence. For example, $D(\cdot,\cdot)$ defined through the $n\times n$ matrix A and $U\in\mathcal{R}$:

$$D(t,x)=Ax+U$$

satisfies the assumptions with

$$\mathcal{O}=\left\{(t,x)\,\middle|\, |x-a|<\int_0^t h(\tau)\,d\tau+\varepsilon\right\}\qquad(\varepsilon>0)$$

when we take

$$h(t)=(|A|\cdot(|a|+\varepsilon)+\|U\|)\,e^{|A|t}.$$

(In the definition of $h(\cdot)$, $|A|=\max_{|x|=1}|Ax|$.)

Let $D(\cdot,\cdot)\in\mathcal{M}$ be given. It is known [3, Lemma 2.8] that $D(\cdot,x(\cdot))\in L_1(I,\mathcal{R})$ for every $x(\cdot)\in C(I,\mathbb{R}^n)$ with graph in $\mathcal{O}$. Moreover for every $(\bar{t},\bar{x})\in P$, there exists an absolutely continuous $\mathbb{R}^n$-valued function $d(\cdot)$ on $[\bar{t},1]$ such that

$$(3.1)\qquad \begin{aligned}&\dot{d}(t)\in D(t,d(t))\quad \text{a.e. } t\in[\bar{t},1],\\ &d(\bar{t})=\bar{x},\end{aligned}$$

and $d(\cdot)$ has graph in P. Existence of a local solution $d(\cdot)$ with graph in $\mathcal{O}$ is established in [3, Thm. 4.1]. Assumption A3 however assures that the graph of $d(\cdot)$ lies in P, and that $d(\cdot)$ may be developed over $[\bar{t},1]$.

For $D(\cdot,\cdot)\in\mathcal{M}$, we define the solution set of

$$(3.2)\qquad \begin{aligned}&\dot{x}(t)\in D(t,x(t))\quad \text{a.e. } t\in I,\\ &x(0)=a,\end{aligned}$$

as $\mathcal{D}=\{d(\cdot)\in C(I,\mathbb{R}^n)|d(\cdot)$ is absolutely continuous and satisfies (3.2)$\}$. $\mathcal{D}$, defined in this way, will be referred to as the solution set of $D(\cdot,\cdot)\in\mathcal{M}$. We have already observed that elements in $\mathcal{D}$ have graph in P. A standard argument, using the property that $D(\cdot,\cdot)$ is convex-valued, gives that $\mathcal{D}$ is a compact subset of $C(I,\mathbb{R}^n)$, i.e. an element in $\mathcal{C}$.

4. Convergence of integrals. We present here results which relate convergence of integrals of set-valued functions and convergence of their integrands:

Artstein [1, Thm. 6.4] has proved the following property of the Aumann integral:

THEOREM 4.1. *Let $F_0(\cdot)$, $\{F_i(\cdot)\}_{i=1}^{\infty}$ belong to $L_1(I, \mathbb{R}^n)$. If $\{F_i(\cdot)\}_{i=1}^{\infty}$ converges weakly to $F_0(\cdot)$, then $\int_{[0,t]} F_i(\tau)\,d\tau$ converges to $\int_{[0,t]} F_0(\tau)\,d\tau$ uniformly in t. The converse is true provided that the family $\{F_i(\cdot)\}_{i=1}^{\infty}$ is uniformly integrable.*

We shall need an analogous property of the trajectory integral:

THEOREM 4.2. *Let $F_0(\cdot)$, $\{F_i(\cdot)\}_{i=1}^{\infty}$ belong to $L_1(I, \mathscr{R})$. If $\{F_i(\cdot)\}_{i=1}^{\infty}$ converges weakly to $F_0(\cdot)$, then $\mathscr{I}F_i(\cdot)$ converges to $\mathscr{I}F_0(\cdot)$ in $\mathscr{C}$. The converse is true provided that the family $\{F_i(\cdot)\}_{i=1}^{\infty}$ is uniformly integrable.*

Proof. Suppose that $F_i(\cdot)\to F_0(\cdot)$ weakly. To establish convergence of the trajectory integrals we need only show that, given an arbitrary subsequence, there exists a further subsequence such that $\overline{\text{dist}}\,(\mathscr{I}F_i(\cdot), \mathscr{I}F_0(\cdot))\to 0$ and $\overline{\text{dist}}\,(\mathscr{I}F_0(\cdot), \mathscr{I}F_i(\cdot))\to 0$.

Limit attention to an arbitrary subsequence. Consider first $\{\overline{\text{dist}}\,(\mathscr{I}F_i(\cdot), \mathscr{I}F_0(\cdot))\}_{i=1}^{\infty}$. For each i there exists, by compactness of $\mathscr{I}F_i(\cdot)$, some $h_i(\cdot)\in \mathscr{I}F_i(\cdot)$ such that $\text{dist}\,(h_i(\cdot), \mathscr{I}F_0(\cdot)) = \overline{\text{dist}}\,(\mathscr{I}F_i(\cdot), \mathscr{I}F_0(\cdot))$. Let $f_i(\cdot)$ be the selector of $F_i(\cdot)$ for which $h_i(t) = \int_0^t f_i(\tau)\,d\tau$, $t\in I$. By assumption $F_i(\cdot)\to F_0(\cdot)$ weakly in $L_1(I, \mathscr{R})$ and $F_0(\cdot)$ is convex-valued; by [1, Props. 4.10 and 4.11] then, $\{f_i(\cdot)\}_{i=1}^{\infty}$ converges weakly to some selector $f_0(\cdot)$ of $F_0(\cdot)$ for some subsequence. It follows that $h_i(\cdot)\to h_0(\cdot)$ uniformly, where $h_0(t) = \int_0^t f_0(\tau)\,d\tau$, $t\in I$, and therefore $\overline{\text{dist}}\,(\mathscr{I}F_i(\cdot), \mathscr{I}F_0(\cdot))\to 0$.

Now consider $\{\overline{\text{dist}}\,(\mathscr{I}F_0(\cdot), \mathscr{I}F_i(\cdot))\}_{i=1}^{\infty}$. For each i we may choose $h_{0i}(\cdot)\in \mathscr{I}F_0(\cdot)$ such that $\text{dist}\,(h_{0i}(\cdot), \mathscr{I}F_i(\cdot)) = \text{dist}\,(\mathscr{I}F_0(\cdot), \mathscr{I}F_i(\cdot))$. Let $f_{0i}(\cdot)$ be the selector of $F_0(\cdot)$ for which $h_{0i}(t) = \int_0^t f_{0i}(\tau)\,d\tau$, $t\in I$. The selectors of $F_0(\cdot)$ are weakly sequentially compact whence (for a further subsequence) $\{f_{0i}(\cdot)\}_{i=1}^{\infty}$ converges weakly in $L_1(I, \mathscr{R})$ to some $f_0(\cdot)$ which is a selector of $F_0(\cdot)$. But by [1, Prop. 4.12], $f_0(\cdot)$ is the weak limit of some sequence $\{f_i(\cdot)\}_{i=1}^{\infty}$ with $f_i(\cdot)$ a selector of $F_i(\cdot)$ for each i. Set $h_i(t) = \int_0^t f_i(\tau)\,d\tau$, $t\in I$. Since $f_i(\cdot) - f_{0i}(\cdot)\to 0$ weakly, $h_i(\cdot) - h_{0i}(\cdot)\to 0$ uniformly and therefore $\overline{\text{dist}}\,(\mathscr{I}F_0(\cdot), \mathscr{I}F_i(\cdot))\to 0$. The first assertion is proved.

The second assertion follows from Theorem 4.1, since convergence of trajectory integrals implies pointwise convergence of the Aumann integrals. □

Bridgland [3, Thm. 3.2] gives a dominated convergence theorem for trajectory integrals, which is a rather special case of Theorem 4.2.

We conclude immediately from Theorems 4.2 and 4.1 a useful relationship between convergence of Aumann integrals and trajectory integrals.

COROLLARY 4.3. *Let $F_0(\cdot)$, $\{F_i(\cdot)\}_{i=1}^{\infty}$ belong to $L_1(I, \mathscr{R})$. Suppose that $\{F_i(\cdot)\}_{i=1}^{\infty}$ is uniformly integrable. Then $\{\mathscr{I}F_i(\cdot)\}_{i=1}^{\infty}$ converges to $\mathscr{I}F_0(\cdot)$ in $\mathscr{C}$ if and only if $\int_{[0,t]} F_i(\tau)\,d\tau$ converges to $\int_{[0,t]} F_0(\tau)\,d\tau$ for each $t\in I$.*

5. Equivalent approximations. The main results, to follow in § 7, concern the stability of the solution set under perturbations of the right hand side of a differential inclusion. Conditions for convergence of a sequence of solution sets are given in terms of weak convergence of the corresponding right hand sides evaluated along limiting trajectories. Such results obviously require developing in the present differential inclusions setting an analogue of the equivalent approximations lemma, of importance in ordinary differential equations, which asserts that uniform convergence of trajectories corresponding to the perturbed ordinary differential equations is equivalent to the convergence of the integrals of the perturbed right hand sides evaluated along the uniform limit of the trajectories [8, Lemma 35.3].

The equivalent approximations lemma for singleton-valued functions follows rather trivially from the assumptions on the family of right hand sides (in fact the

Lipschitz condition A2), but the corresponding result in the present setting involves some rather complicated constructions. We shall repeatedly use the following theorem due to Filippov [6, Thm. 1]:

THEOREM 5.1. *Given $D(\cdot,\cdot)\in\mathcal{M}$, an interval $T=[t_0,t_1]\subset I$, and an absolutely continuous function $y(\cdot)\in C(T,\mathbb{R}^n)$ with graph in P, we write*

$$\rho(t)=\operatorname{dist}(\dot{y}(t),D(t,y(t))),\qquad t\in T.$$

Then $\rho(\cdot)$ is integrable, and there exists a solution $x(\cdot)$ of

$$\dot{x}(t)\in D(t,x(t)),\qquad x(t_0)=y(t_0),$$

such that

$$|x(\cdot)-y(\cdot)|_T\leqq C\int_{t_0}^{t_1}\rho(\tau)\,d\tau$$

where the constant C is independent of $D(\cdot,\cdot)$ and $y(\cdot)$.

The theorem in [6] is stated for $D(\cdot,\cdot)$'s continuous in their t-dependence, but the proof adapts in an obvious manner to apply under the more general hypotheses of Theorem 5.1.

In the following results, it is understood that $D(\cdot,\cdot)$, $G(\cdot,\cdot)$ are elements in $\mathcal{M}$. $\mathcal{D}$, $\mathcal{G}$ denote the corresponding solution sets (for common initial value $x(0)=a$). $\mathcal{D}_i$ is the solution set of $D_i(\cdot,\cdot)$ etc.

LEMMA 5.2. *Given $\varepsilon>0$, there exists $\eta>0$ such that*

$$\operatorname{dist}(\mathcal{J}D(\cdot,d(\cdot)),\mathcal{J}G(\cdot,g(\cdot)))\leqq\varepsilon$$

whenever $\operatorname{dist}(\mathcal{D},\mathcal{G})\leqq\eta$ *and $d(\cdot)\in\mathcal{D}$, $g(\cdot)\in\mathcal{G}$ with $|d(\cdot)-g(\cdot)|_I\leqq\eta$.*

LEMMA 5.3. *Given $\eta>0$, there exists $\varepsilon>0$ such that* $\operatorname{dist}(\mathcal{D},\mathcal{G})\leqq\eta$ *whenever* $\operatorname{dist}(\mathcal{J}G(\cdot,g(\cdot)),\mathcal{J}D(\cdot,g(\cdot)))\leqq\varepsilon$ *for all $g(\cdot)\in\mathcal{G}$.*

These two lemmas are proved in § 6.

The lemmas combine to give the new equivalent approximations result:

PROPOSITION 5.4. *Let $D_0(\cdot,\cdot)$ and $\{D_i(\cdot,\cdot)\}_{i=1}^{\infty}$ belong to $\mathcal{M}$. Then, if $\mathcal{D}_i\to\mathcal{D}_0$ in $\mathcal{C}$ and $d_i(\cdot)\to d(\cdot)$ in $C(I,\mathbb{R}^n)$ with $d_i(\cdot)\in\mathcal{D}_i$, $i=1,2,\cdots$, we have that*

$$\operatorname{dist}(\mathcal{J}D_i(\cdot,d_i(\cdot)),\mathcal{J}D_0(\cdot,d(\cdot)))\to 0.$$

On the other hand, if

$$\operatorname{dist}(\mathcal{J}D_i(\cdot,d(\cdot)),\mathcal{J}D_0(\cdot,d(\cdot)))\to 0$$

for every $d(\cdot)\in\mathcal{D}_0$, then $\mathcal{D}_i\to\mathcal{D}_0$.

Proof. Suppose that $\mathcal{D}_i\to\mathcal{D}_0$, $d_i(\cdot)\to d(\cdot)$ and that, for each i, $d_i(\cdot)\in\mathcal{D}_i$. Then $d(\cdot)\in\mathcal{D}_0$. The first assertion is now seen as a restatement of Lemma 5.2.

Consider now the second assertion. This will follow from Lemma 5.3 if, under the stated hypotheses, $\operatorname{dist}(\mathcal{J}D_i(\cdot,d(\cdot)),\mathcal{J}D_0(\cdot,d(\cdot)))\to 0$ uniformly over $d(\cdot)\in\mathcal{D}_0$. Define $\varepsilon_i(d)=\operatorname{dist}(\mathcal{J}D_i(\cdot,d(\cdot)),\mathcal{J}D_0(\cdot,d(\cdot)))$. The sequence $\{\varepsilon_i(\cdot)\}_{i=1}^{\infty}$ is an equicontinuous family of functions on the compact set $\mathcal{D}_0$. Indeed for $d(\cdot)$, $d'(\cdot)\in\mathcal{D}_0$

$$\varepsilon_i(d')\leqq\operatorname{dist}(\mathcal{J}D_i(\cdot,d'(\cdot)),\mathcal{J}D_i(\cdot,d(\cdot)))+\varepsilon_i(d)+\operatorname{dist}(\mathcal{J}D_0(\cdot,d(\cdot)),\mathcal{J}D_0(\cdot,d'(\cdot)))$$

$$\leqq\varepsilon_i(d)+2|d(\cdot)-d'(\cdot)|_I\cdot\int_0^1 k(t)\,dt,$$

whence, by symmetry, we have that

$$|\varepsilon_i(d')-\varepsilon_i(d)|_I\leqq 2|d(\cdot)-d'(\cdot)|_I\cdot\int_0^1 k(t)\,dt.$$

The sequence converges uniformly to the zero function then since it converges pointwise. □

6. Equivalent approximations: Proof of the lemmas. We let $\{a\}$ denote the element $x(\cdot)\in C(I, \mathbb{R}^n)$ for which $x(t)=a$, for all $t\in I$.

Proof of Lemma 5.2. Take $d(\cdot)\in \mathscr{D}$, $\delta(\cdot)\in\{a\}+\mathscr{I}D(\cdot, d(\cdot))$. Let $\{t_0=0, t_1, \cdots, t_N=1\}$ partition I uniformly into N intervals. Define $\delta_k(\cdot)$ on I_k, $I_k=[t_k, t_{k+1}]$ as

$$\delta_k(t)=\delta(t)-\delta(t_k)+d(t_k), \qquad t\in I_k. \tag{6.1}$$

Then $\delta_k(\cdot)\in C(I_k, \mathbb{R}^n)$ is absolutely continuous and

$$\dot{\delta}_k(t)\in D(t, d(t)) \quad \text{a.e. } t\in I_k: \qquad \delta(t_k)=d(t_k).$$

Furthermore

$$\begin{aligned}\int_{t_k}^{t_{k+1}} \operatorname{dist}(\dot{\delta}_k(t), D(t, \delta_k(t)))\,dt &\leqq \int_{t_k}^{t_{k+1}} \operatorname{dist}(D(t, d(t)), D(t, \delta_k(t)))\,dt\\ &\leqq \int_{t_k}^{t_{k+1}} k(t)|d(t)-\delta_k(t)|\,dt\\ &\leqq 2\int_{t_k}^{t_{k+1}} k(t)\int_{t_k}^{t} h(\tau)\,d\tau\,dt.\end{aligned}$$

Let us apply Theorem 5.1 to the differential inclusion:

$$\dot{x}(t)\in D(t, x(t)), \quad \text{a.e. } t\in I_k, \quad x(t_k)=d(t_k). \tag{6.2}$$

In view of the last estimate there exists a function $d_k(\cdot)$ on I_k being a solution to (6.2) and such that

$$|d_k(\cdot)-\delta_k(\cdot)|_{I_k}\leqq 2C\int_{t_k}^{t_{k+1}} k(t)\int_{t_k}^{t} h(\tau)\,d\tau\,dt. \tag{6.3}$$

Notice that, by construction, $d_k(\cdot)$ must be the restriction to I_k of some element of $\mathscr{D}$. We now define a function $\tilde{\delta}(\cdot)$ on I, by piecing together the $d_k(\cdot)$'s to give a continuous function:

$$\begin{aligned}&\tilde{\delta}(0)=a,\\ &\tilde{\delta}(t)=d_k(t)-d_k(t_k)+\tilde{\delta}(t_k), \qquad t\in I_k, \text{for } k=0, 1, \cdots, N-1.\end{aligned} \tag{6.4}$$

In view of (6.3)

$$|\tilde{\delta}(\cdot)-\delta(\cdot)|_I\leqq 2C\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} k(t)\int_{t_k}^{t} h(\tau)\,d\tau\,dt=2C\int_0^1 k(t)\tilde{h}_N(t)\,dt \tag{6.5}$$

where $\tilde{h}_N(\cdot)$ is defined as

$$\tilde{h}_N(t)=\int_{t_k}^{t} h(\tau)\,d\tau, \qquad t\in[t_k, t_{k+1}), \quad k=0, 1, \cdots, N-1. \tag{6.6}$$

Now suppose that $\operatorname{dist}(\mathscr{D}, \mathscr{G})\leqq\eta$ and let $g(\cdot)\in\mathscr{G}$ be such that $|g(\cdot)-d(\cdot)|_I\leqq\eta$. Since each $d_k(\cdot)$ is the restriction to I_k of an element in $\mathscr{D}$, by hypothesis we may choose, for each k, functions $g_k(\cdot)\in C(I_k, \mathbb{R}^n)$, being restrictions to I_k of elements in $\mathscr{G}$, which satisfy

$$|g_k(\cdot)-d_k(\cdot)|_{I_k}\leqq\eta. \tag{6.7}$$

We now take $\tilde{\gamma}(\cdot)$ to be the function on I obtained by piecing together the $g_k(\cdot)$'s to give a continuous function:

$$\begin{aligned} &\tilde{\gamma}(0)=a, \\ (6.8) \qquad &\tilde{\gamma}(t)=g_k(t)-g_k(t_k)+\tilde{\gamma}(t_k), \qquad t\in I_k \quad \text{for } k=0,1,\cdots,N-1. \end{aligned}$$

In view of (6.7)

$$|\tilde{\gamma}(\cdot)-\tilde{\delta}(\cdot)|_I \leqq 2N\eta;$$

whence by (6.5)

$$(6.9) \qquad |\tilde{\gamma}(\cdot)-\delta(\cdot)|_I \leqq 2N\eta+2C\int_0^1 k(t)\tilde{h}_N(t)\,dt.$$

For each k we may select $\dot{\gamma}_k(\cdot)\in L_1(I_k,\mathbb{R}^n)$ such that $\dot{\gamma}_k(t)\in G(t,g(t))$, a.e. on I_k, and

$$(6.10) \qquad |\dot{g}_k(t)-\dot{\gamma}_k(t)|=\operatorname{dist}(\dot{g}_k(t),G(t,g(t))), \qquad t\in I_k.$$

This is possible by Lemma 2.5 of [3], which extends a selection lemma of Hermes to permit unbounded set-valued functions. Since

$$\begin{aligned} \operatorname{dist}(\dot{g}_k(t),G(t,g(t))) &\leqq k(t)|g_k(t)-g(t)| \\ &\leqq k(t)(|g_k(t)-d_k(t)|+|d_k(t)-d(t)|+|d(t)-g(t)|) \\ &\leqq k(t)\left(2\eta+2\int_{t_k}^t h(\tau)\,d\tau\right), \qquad t\in I_k, \end{aligned}$$

it follows from (6.10) that

$$(6.11) \qquad |\dot{g}_k(t)-\dot{\gamma}_k(t)|\leqq 2k(t)\left(\eta+\int_{t_k}^t h(\tau)\,d\tau\right) \quad \text{a.e. on } I_k.$$

We define the function $\gamma_k(\cdot)$ on I_k as

$$\gamma_k(t)=g_k(t_k)+\int_{t_k}^t \dot{\gamma}_k(\tau)\,d\tau, \qquad t\in I_k.$$

By (6.11)

$$(6.12) \qquad |\gamma_k(\cdot)-g_k(\cdot)|_{I_k}\leqq 2\eta\int_{t_k}^{t_{k+1}} k(t)\,dt+2\int_{t_k}^{t_{k+1}} k(t)\int_{t_k}^t h(\tau)\,d\tau\,dt.$$

Finally we define the function $\gamma(\cdot)$ on I by piecing together the $\gamma_k(\cdot)$'s to give a continuous function:

$$\begin{aligned} &\gamma(0)=a, \\ (6.13) \qquad &\gamma(t)=\gamma_k(t)-\gamma_k(t_k)+\gamma(t_k), \qquad t\in I_k \quad \text{for } k=0,1,\cdots,N-1. \end{aligned}$$

Observe that now $\gamma(\cdot)\in\{a\}+\mathscr{I}G(\cdot,g(\cdot))$. Since $\tilde{\gamma}(\cdot)$ was constructed by piecing together the $g_k(\cdot)$'s and $\gamma(\cdot)$ by piecing together the $\gamma_k(\cdot)$'s, we have by (6.12)

$$|\gamma(\cdot)-\tilde{\gamma}(\cdot)|_I\leqq 2\eta\int_0^1 k(t)\,dt+2\int_0^1 k(t)\tilde{h}_N(t)\,dt$$

with $\tilde{h}_N(\cdot)$ as in (6.6). By (6.9)

$$(6.14) \qquad |\gamma(\cdot)-\delta(\cdot)|_I\leqq 2\eta\left(N+\int_0^1 k(t)\,dt\right)+2(C+1)\int_0^1 k(t)\tilde{h}_N(t)\,dt.$$

Now taking note of the manner in which $\tilde{h}_N(\cdot)$ is defined in terms of the integrable function $h(\cdot)$, we see that as $N \to \infty$, $\tilde{h}_N(\cdot) \to 0$ uniformly. Given $\varepsilon > 0$ then, it is clear from (6.14) that N and η may be chosen so that

$$|\gamma(\cdot) - \delta(\cdot)|_I \leqq \varepsilon.$$

It follows that $\overline{\text{dist}}\,(\mathscr{I}D(\cdot\,, d(\cdot)), \mathscr{I}G(\cdot\,, g(\cdot))) \leqq \varepsilon$. But the roles of d and g can be interchanged in the above arguments. It is clear then, for the same choice of N and η, $\text{dist}\,(\mathscr{I}D(\cdot, d(\cdot)), \mathscr{I}G(\cdot, g(\cdot))) \leqq \varepsilon$. The proof is complete. □

Proof of Lemma 5.3. We first show that $\overline{\text{dist}}\,(\mathscr{D}, \mathscr{G}) \leqq \eta$, whenever $\text{dist}\,(\mathscr{I}G(\cdot, g(\cdot)), \mathscr{I}D(\cdot, g(\cdot))) \leqq \varepsilon$ for all $g(\cdot) \in \mathscr{G}$, and $\varepsilon > 0$ is sufficiently small. Take any $g(\cdot) \in \mathscr{G}$. Then $g(\cdot) \in \{a\} + \mathscr{I}G(\cdot, g(\cdot))$ and there exists $\delta(\cdot) \in \{a\} + \mathscr{I}D(\cdot, g(\cdot))$ such that $|\delta(\cdot) - g(\cdot)|_I \leqq \varepsilon$ by hypothesis. We have

$$\int_0^1 \text{dist}\,(\dot{\delta}(t), D(t, \delta(t)))\, dt \leqq \int_0^1 k(t)|g(t) - \delta(t)|\, dt \leqq \varepsilon \int_0^1 k(t)\, dt.$$

By Theorem 5.1, there exists $d(\cdot) \in \mathscr{D}$ such that

$$|d(\cdot) - \delta(\cdot)|_I \leqq C\varepsilon \int_0^1 k(t)\, dt$$

and

$$|d(\cdot) - g(\cdot)|_I \leqq \varepsilon\left(C \int_0^1 k(t)\, dt + 1\right).$$

Thus it suffices to take $\varepsilon = \eta(C \int_0^1 k(t)\, dt + 1)^{-1}$.

We proceed to show that $\varepsilon > 0$ may be chosen so that additionally $\overline{\text{dist}}\,(\mathscr{D}, \mathscr{G}) \leqq \eta$. Let us suppose without loss of generality that $k(\cdot)$ has been chosen so that $k(t) \geqq \theta > 0$ a.e. on I. We write $K = \int_0^1 k(t)\, dt$ and let $\{t_0 = 0, 1, \cdots, t_N = 1\}$ partition I onto N intervals in such a way that

$$\int_{t_k}^{t_{k+1}} k(t)\, dt = \frac{K}{N}, \qquad k = 0, 1, \cdots, N-1.$$

It is clear then, with the assumption on $k(\cdot)$, that the lengths of the intervals within the partition tend to zero uniformly as $N \to \infty$. Take any $d(\cdot) \in \mathscr{D}$. We shall construct a sequence $g_0(\cdot), \cdots, g_N(\cdot)$ of elements in $\mathscr{G}$ having the property that $|g_{k+1}(\cdot) - d(\cdot)|_{[0,t_{k+1}]}$ is suitably estimated in terms of $|g_k(\cdot) - d(\cdot)|_{[0,t_k]}$, $k = 0, 1, \cdots, N-1$. In consequence an estimate for $|g_N(\cdot) - d(\cdot)|_I$ in terms of N, ε is obtained, on the basis of which we conclude that $|g_N(\cdot) - d(\cdot)|_I$ may be made arbitrarily small by choosing appropriately these parameters.

The sequence $g_0(\cdot), \cdots, g_N(\cdot)$ is defined inductively. $g_0(\cdot)$ is taken as an arbitrary element in $\mathscr{G}$, since only its value a at $t = 0$ matters. Suppose $g_k(\cdot) \in \mathscr{G}$ is given and set

$$\xi_k = |d(\cdot) - g_k(\cdot)|_{[0,t_k]}.$$

By hypothesis $\text{dist}\,(\mathscr{I}G(\cdot, g(\cdot)), \mathscr{I}D(\cdot, g(\cdot))) \leqq \varepsilon$ for all $g(\cdot) \in \mathscr{G}$. Since $g_k(\cdot) \in \{a\} + \mathscr{I}G(\cdot, g_k(\cdot))$, there exists a function $\delta_k(\cdot)$ on $[0, t_k]$, being the restriction to $[0, t_k]$ of some element in $\{a\} + \mathscr{I}D(\cdot, g_k(\cdot))$ satisfying

$$|\delta_k(\cdot) - g_k(\cdot)|_{[0,t_k]} \leqq \varepsilon.$$

We extend the domain of definition of $\delta_k(\cdot)$ to $[0, t_{k+1}]$, by setting

$$\delta_k(t) = \delta_k(t_k) + \int_{t_k}^{t} \dot{\delta}_k(\tau)\, d\tau$$

and choosing $\delta_k(\cdot)\in L_1(I_k, \mathbb{R}^n)$ such that

$$\dot\delta_k(t)\in D(t, g_k(t)) \quad \text{a.e. on } I_k$$

and

$$|\dot d(t)-\dot\delta_k(t)| = \operatorname{dist}(\dot d(t), D(t, g_k(t))) \quad \text{a.e. on } I_k.$$

Such a choice is possibly by Lemma 2.5 in [3]. $\delta_k(\cdot)$ is now the restriction to $[0, t_{k+1}]$ of some element of $\{a\}+\mathscr{I}D(\cdot, g_k(\cdot))$. For $t\in I_k$ we have

$$\begin{aligned}|\delta_k(t)-d(t)| &\leqq \xi_k+\varepsilon+\int_{t_k}^{t} |\dot\delta_k(\tau)-\dot d(\tau)|\,d\tau\\ &\leqq \xi_k+\varepsilon+\int_{t_k}^{t} \operatorname{dist}(D(\tau, g_k(\tau)), D(\tau, d(\tau)))\,d\tau\\ &\leqq \xi_k+\varepsilon+\int_{t_k}^{t} k(\tau)|g_k(\tau)-d(\tau)|\,d\tau\end{aligned}$$

and

$$\begin{aligned}|\delta_k(\cdot)-d(\cdot)|_{I_k} &\leqq \xi_k+\varepsilon+\int_{t}^{t_{k+1}} k(t)\Big(\xi_k+\int_{t_k}^{t} |\dot g_k(\tau)-\dot d(\tau)|\,d\tau\Big)\,dt\\ &\leqq \Big(1+\int_{t_k}^{t_{k+1}} k(t)\,dt\Big)\xi_k+2\int_{t_k}^{t_{k+1}} k(t)\int_{t_k}^{t} h(\tau)\,d\tau\,dt+\varepsilon.\end{aligned} \tag{6.15}$$

By hypothesis, there exists a function $\tilde\gamma_k(\cdot)$ on I_k, which is the restriction to I_k of some element in $\{a\}+\mathscr{I}G(\cdot, g_k(\cdot))$ and which satisfies

$$|\delta_k(\cdot)-\tilde\gamma_k(\cdot)|_{I_k}\leqq\varepsilon. \tag{6.16}$$

Set

$$\gamma_k(t)=\tilde\gamma_k(t)-\tilde\gamma_k(t_k)+g_k(t_k), \qquad t\in I_k.$$

Then

$$|\tilde\gamma_k(t_k)-g_k(t_k)|\leqq|\tilde\gamma_k(t_k)-\delta_k(t_k)|+|\delta_k(t_k)-g_k(t_k)|\leqq 2\varepsilon \tag{6.17}$$

and by (6.16) and (6.17)

$$|\gamma_k(\cdot)-\delta_k(\cdot)|_{I_k}\leqq|\tilde\gamma(\cdot)-\delta_k(\cdot)|_{I_k}+|\tilde\gamma(t_k)-g_k(t_k)|\leqq 3\varepsilon.$$

It follows from (6.15)

$$|\gamma_k(\cdot)-d(\cdot)|_{I_k}\leqq\Big(1+\int_{t_k}^{t_{k+1}} k(t)\,dt\Big)\xi_k+2\int_{t_k}^{t_{k+1}} k(t)\int_{t_k}^{t} h(\tau)\,d\tau\,dt+4\varepsilon. \tag{6.18}$$

Since $\gamma_k(\cdot)$ is the restriction to I_k of some element in $\{a\}+\mathscr{I}G(\cdot, g_k(\cdot))$,

$$\begin{aligned}\int_{t_k}^{t_{k+1}} \operatorname{dist}(\dot\gamma_k(t), G(t, \gamma_k(t)))\,dt &\leqq \int_{t_k}^{t_{k+1}} k(t)|g_k(t)-\gamma_k(t)|\,dt\\ &\leqq 2\int_{t_k}^{t_{k+1}} k(t)\int_{t_k}^{t} h(\tau)\,d\tau\,dt.\end{aligned}$$

By Theorem 5.1, then there exists some absolutely continuous function $\tilde g_k(\cdot)$ on I_k such that

$$\dot{\tilde g}_k(t)\in G(t, \tilde g_k(t)) \quad \text{a.e. on } I_k; \qquad \tilde g_k(t_k)=\gamma_k(t_k)(=g_k(t_k))$$

and

$$(6.19)\qquad |\tilde{g}_k(\cdot)-\gamma_k(\cdot)|_{I_k} \leqq 2C\int_{t_k}^{t_{k+1}} k(t)\int_{t_k}^{t} h(\tau)\,d\tau\,dt.$$

Now define the function $g_{k+1}(\cdot)$ on I by setting

$$g_{k+1}(t)=g_k(t),\qquad t\in[0,t_k],$$

$$g_{k+1}(t)=\tilde{g}_k(t),\qquad t\in[t_k,t_{k+1}],$$

and assigning the values of $g_{k+1}(\cdot)$ arbitrarily on $(t_{k+1},1]$, but so as to ensure that $g_{k+1}(\cdot)\in\mathcal{G}$. This is possible since $g_{k+1}(\cdot)$ on $[0,t_{k+1}]$ is a piecing together of restrictions of elements in $\mathcal{G}$. By (6.18) and (6.19),

$$|g_{k+1}(\cdot)-d(\cdot)|_{[0,t_{k+1}]}\leqq\left(1+\int_{t_k}^{t_{k+1}}k(t)\,dt\right)\xi_k+2(C+1)\int_{t_k}^{t_{k+1}}k(t)\int_{t_k}^{t}h(\tau)\,d\tau\,dt+4\varepsilon.$$

This completes the definition of the $g_k(\cdot)$'s. We have along the way obtained

$$\xi_{k+1}=|g_k(\cdot)-d(\cdot)|_{[0,t_{k+1}]}$$
$$=\left(1+\int_{t_k}^{t_{k+1}}k(t)\,dt\right)\xi_k+2(C+1)\int_{t_k}^{t_{k+1}}k(t)\int_{t_k}^{t}h(\tau)\,d\tau\,dt+4\varepsilon$$

for $k=0,1,\cdots,N-1$. Notice that $\xi_0=0$ and that in particular

$$|g_N(\cdot)-d(\cdot)|_I\leqq\xi_N.$$

It remains to demonstrate that ξ_N may be made arbitrarily small by suitable choice of ε and N. To that end let us define the nonnegative numbers $\{\zeta_k\}_{k=0}^N$ recursively as

$$\zeta_{k+1}=\left(1+\frac{K}{N}\right)\zeta_k+\frac{2(C+1)H_NK}{N}+4\varepsilon,\qquad k=0,1,\cdots,N-1,$$

$$\zeta_0=0,$$

where $H_N=\max_{0\leqq k\leqq N-1}\{\int_{t_k}^{t_{k+1}}h(\tau)\,d\tau\}$. We then have that $\xi_k\leqq\zeta_k\leqq\zeta_N$ for $k=0,1,\cdots,N-1$ and moreover

$$\zeta_N=\left[2(C+1)H_N+\frac{4N}{K}\varepsilon\right]\cdot\left[\left(1+\frac{K}{N}\right)^N-1\right]\leqq\left[2(C+1)H_N+\frac{4N}{K}\varepsilon\right](e^K-1).$$

But $h(\cdot)$ is integrable. Since, as previously pointed out, the lengths of the intervals in the partition tend to zero uniformly as $N\to\infty$, we may conclude that $H_N\to0$. Let $\tilde{N}$ be the smallest integer such that $2(e^K-1)(C+1)H_{\tilde{N}}\leqq\eta/2$ and take $\varepsilon\leqq K\eta/(8\tilde{N}(e^K-1))$. With this choice $\xi_{\tilde{N}}\leqq\zeta_{\tilde{N}}\leqq\eta$ and the proof is complete. □

7. Necessary and sufficient conditions for convergence of solution sets of differential inclusions. Proposition 5.4 which replaces the lemma of equivalent approximations for singleton-valued functions and Theorem 4.2 now come together to give our main result. This asserts that convergence of solution sets is fully characterized by weak convergence of the corresponding right hand sides evaluated on trajectories in the limiting solution set. Again, $\mathcal{D}_i$ denotes the solution set corresponding to $D_i(\cdot,\cdot)$.

THEOREM 7.1. *Let $D_0(\cdot,\cdot)$, $\{D_i(\cdot,\cdot)\}_{i=1}^\infty$ belong to $\mathcal{M}$. Then for $\{\mathcal{D}_i\}_{i=1}^\infty$ to converge to $\mathcal{D}_0$ it is necessary and sufficient that $\{D_i(\cdot,d_0(\cdot))\}_{i=1}^\infty$ converges to $D_0(\cdot,d_0(\cdot))$ weakly in $L_1(I,\mathcal{R})$ for every $d_0(\cdot)\in\mathcal{D}_0$.*

Proof. Sufficiency of the condition follows from Proposition 5.4 and Theorem 4.2. We prove necessity. Suppose that $\mathscr{D}_i \to \mathscr{D}_0$ in $\mathscr{C}$ and take any $d_0(\cdot) \in \mathscr{D}_0$. For each i choose $d_i(\cdot) \in \mathscr{D}_i$ such that $|d_i(\cdot) - d_0(\cdot)|_I \leqq \operatorname{dist}(\mathscr{D}_i, \mathscr{D}_0)$. Notice that $d_i(\cdot) \to d_0(\cdot)$ uniformly. By Proposition 5.4 then, $\mathscr{I}D_i(\cdot, d_i(\cdot)) \to \mathscr{I}D_0(\cdot, d_0(\cdot))$ in $\mathscr{C}$. But this implies $\mathscr{I}D_i(\cdot, d_0(\cdot)) \to \mathscr{I}D_0(\cdot, d_0(\cdot))$ in $\mathscr{C}$ in view of assumption A2. By assumption A3 the family $\{D_i(\cdot, d_0(\cdot))\}_{i=1}^{\infty}$ is uniformly integrable; we conclude from Theorem 4.3 that $D_i(\cdot, d_0(\cdot)) \to D_0(\cdot\, d_0(\cdot))$ weakly in $L_1(I, \mathscr{R})$ as required. □

The above condition for convergence of the solution sets is of course the most general possible in the present context, but it is difficult to use because it requires knowledge of the limiting solution set. We deduce from the theorem however the following more readily applicable, though more restrictive, condition for convergence of the solution sets:

COROLLARY 7.2. *Let $D_0(\cdot, \cdot)$ and $\{D_i(\cdot, \cdot)\}_{i=1}^{\infty}$ belong to $\mathscr{M}$. Suppose that*

$$\int_t^{t+\delta} D_i(\tau, x)\, d\tau \to \int_t^{t+\delta} D_0(\tau, x)\, d\tau$$

for each (t, x, δ) such that $(t, x) \in P$ and $t \leqq t + \delta \leqq 1$. Then

$$\mathscr{D}_i \to \mathscr{D}_0$$

in $\mathscr{C}$.

Proof. We merely sketch the steps in the proof, since they are standard ones. Let $d(\cdot) \in \mathscr{D}_0$. Let $\{t_0 = 0, t_1, \cdots, t_n, t_{n+1} = 1\}$ be a uniform partition of $[0, 1]$ and let $d_n(\cdot)$ be the piecewise constant function defined by $d_n(t) = d(t_k)$ for $t \in [t_k, t_{k+1})$, $k = 0, \cdots, n$. Suppose that the hypothesis of the Corollary holds. Then for each t, each n,

$$\operatorname{dist}\left(\int_{[0,t]} D_i(\tau, d_n(\tau))\, d\tau, \int_{[0,t]} D_0(\tau, d_n(\tau))\, d\tau\right) \to 0$$

as $i \to \infty$. But $d(\cdot)$ is absolutely continuous whence $|d_n(\cdot) - d(\cdot)|_I \to 0$ as $n \to \infty$. Bearing in mind the restrictions imposed on elements in the set $\mathscr{M}$ we easily show that, for each t,

$$\operatorname{dist}\left(\int_{[0,t]} D_i(\tau, d(\tau))\, d\tau, \int_{[0,t]} D_i(\tau, d_n(\tau))\, d\tau\right) \to 0$$

as $n \to \infty$, uniformly in i, and by another simple step that

$$\operatorname{dist}\left(\int_{[0,t]} D_i(\tau, d(\tau))\, d\tau, \int_{[0,t]} D_0(\tau, d(\tau))\, d\tau\right) \to 0 \tag{7.1}$$

as $i \to \infty$, for each t. Since the $D_i(\cdot, \cdot)$'s are uniformly integrably bounded, (7.1) implies that

$$D_i(\cdot, d(\cdot)) \to D_0(\cdot, d(\cdot))$$

weakly. But $d(\cdot) \in \mathscr{D}_0$ was arbitrary and the corollary now follows from Theorem 7.1. □

8. Dynamics affine in their control dependence. Now consider dynamical systems affine in their u-dependence:

$$\begin{aligned} &\dot{x}(t) = f(t, x(t), u(t)) \quad \text{a.e. } t \in I, \\ &f(t, x, u) = f_1(t, x) + f_2(t, x)u \quad \text{with } u(\cdot) \text{ a selector of } U(\cdot),\ x(0) = a. \end{aligned} \tag{8.1}$$

Theorem 7.1 specializes to characterize the perturbations of the control constraint set under which the solution set is stable.

Let $h_\Omega(\cdot)$ be a fixed nonnegative valued, integrable function on I. Denote by $\mathcal{R}_m$ the nonempty, convex, compact subsets of $\mathbb{R}^n$, and define

$$\Omega=\{U(\cdot)\in L_1(I,\mathcal{R}_m)\,|\,\|U(t)\|\leqq h_\Omega(t)\text{ a.e.}\}.$$

Now Ω may be equipped with a metric topology τ compatible with weak convergence such that (Ω,τ) is compact. To see this define

$$s(p;U)=\max\{p'u\,|\,u\in U\}$$

for $p\in\mathbb{R}^m$ and $U\in\mathcal{R}_m$ (p' denotes p transpose). Let $\{p_i\}_{i=1}^{\infty}$ be a dense subset in $\mathbb{R}^m$. Then if we set

$$\rho_\nu(U(\cdot),V(\cdot))=\sup_{t\in I}\left|\int_0^t[s(p_\nu;U(\xi))-s(p_\nu;V(\xi))]\,d\xi\right|,$$

the metric on $L_1(I,\mathcal{R}_m)$ defined by

$$\rho(U(\cdot),V(\cdot))=\sum_{\nu=1}^{\infty}\frac{1}{2^\nu}\cdot\frac{\rho_\nu(U(\cdot),V(\cdot))}{1+\rho_\nu(U(\cdot),V(\cdot))}$$

serves the purpose. Indeed, using the characterization of weak convergence given in [1, Thm. 4.1(iii)] and simple continuity-density arguments, we establish that Ω is weakly sequentially complete, that it is complete also with respect to the metric $\rho(\cdot,\cdot)$ and that weak convergence is equivalent to convergence in the metric $\rho(\cdot,\cdot)$. But Ω is weakly sequentially pre-compact [1, Prop. 4.9]. Ω, then, is sequentially compact, and τ, taken to be the topology on Ω induced by $\rho(\cdot,\cdot)$, has the desired properties.

Let $\mathcal{O}$ be some relatively open subset of $I\times\mathbb{R}^n$ containing P (P as in §3). The functions $f_1(\cdot,\cdot)$, $f_2(\cdot,\cdot)$ on $\mathcal{O}$ take values in $\mathbb{R}^n$ and the space of m by n matrices respectively.

We assume that

$$D(t,x)=\{f_1(t,x)\}+f_2(t,x)U(t) \tag{8.2}$$

is an element in $\mathcal{M}$ for each $U(\cdot)\in\Omega$. We also assume that for all continuous functions $x(\cdot)$ with graph in P, $f_2(\cdot,x(\cdot))$ is essentially bounded.

We write $\mathcal{C}_\Omega$ for the family of solution sets of the differential inclusion

$$\begin{aligned}&\dot{x}(t)\in\{f_1(t,x(t))\}+f_2(t,x(t))U(t),\\&x(0)=a\end{aligned} \tag{8.3}$$

obtained by allowing $U(\cdot)$ to range over Ω, and we write s for the map carrying $U(\cdot)$ into the solution set of (8.3).

PROPOSITION 8.1. *s: $(\Omega,\tau)\to\mathcal{C}$ is continuous, and if each element in $\mathcal{C}_\Omega$ contains a member $d(\cdot)$ such that $f_2(\cdot,d(\cdot))$ is a.e. one-to-one, then s is a homeomorphism as a mapping from (Ω,τ) to $\mathcal{C}_\Omega$, with the topology induced by $\mathcal{C}$.*

Proof. Since (Ω,τ), $\mathcal{C}$ are metric spaces, we need establish continuity only in the sense of sequences.

Consider $\{U_i(\cdot)\}_{i=1}^{\infty}$, $U_0(\cdot)$ in Ω and write $\{D_i(\cdot,\cdot)\}_{i=1}^{\infty}$, $D_0(\cdot,\cdot)$ for the corresponding right hand sides, as defined by (8.2). Our assumption on $f_2(\cdot,\cdot)$ permits direct application of [1, Prop. 6.3] which gives that $U_i(\cdot)\to U_0(\cdot)$ weakly in $L_1(I,\mathcal{R}_m)$ implies $D_i(\cdot,d_0(\cdot))\to D_0(\cdot,d_0(\cdot))$ weakly in $L_1(I,\mathcal{R})$ for every $d_0(\cdot)\in\mathcal{D}_0$. Conversely if $D_i(\cdot,d_0(\cdot))\to D_0(\cdot,d_0(\cdot))$ weakly in $L_1(I,\mathcal{R})$ for some $d_0(\cdot)\in\mathcal{D}_0$ such that

$f_2(\cdot, d_0(\cdot))$ is a.e. one-to-one, the same proposition asserts that $U_i(\cdot) \to U_0(\cdot)$ weakly in $L_1(I, \mathcal{R}_m)$. We may now apply Theorem 7.1. □

The assumption of affine dependence cannot be dropped. Indeed suppose that $V \subset \mathbb{R}^m$ is a convex subset and $f(\cdot, \cdot, \cdot)$: $\mathbb{R}^{1+n} \times V \to \mathbb{R}^n$ is a continuous function such that $F(\cdot, \cdot)$ defined by

$$F(t, x) = \{f(t, x, u(t))\} \tag{8.4}$$

satisfies assumptions A1, A2, A3 for every control $u(\cdot)$, that is to say a measurable function taking values in V. Here each right hand side of the form (8.4) gives rise to a solution set $\{x(\cdot)\}$, where $x(\cdot)$ is the unique solution of

$$\dot{x}(t) = f(t, x(t), u(t)),$$

$$u(t) \in V \quad \text{a.e. } t \in I, x(0) = a.$$

Let $x_0(\cdot)$ be the solution corresponding to the control $u_0(\cdot)$, and suppose that $f(t, x_0(t), \cdot)$ is not the restriction to V of an affine function for all $t \in I$. Then, as is easily shown, we can always find a sequence $\{u_i(\cdot)\}_{i=1}^{\infty}$ converging to $u_0(\cdot)$ weakly, such that

$$\{f(\cdot, x_0(\cdot), u_i(\cdot))\} \not\to \{f(\cdot, x_0(\cdot), u_0(\cdot))\}$$

weakly in $L_1(I, \mathcal{R})$ and hence $\{x_i(\cdot)\} \not\to \{x_0(\cdot)\}$ in $\mathcal{C}$, by Theorem 7.1. These observations illustrate that, in the absence of assumptions on affine dependence, the "weak" topology τ on Ω is too weak to describe perturbations with respect to which the solution set is stable. It is shown in [10] that strong convergence of the $U(\cdot)$'s in $L_1(I, \mathcal{R}_m)$ is necessary for stability of solution sets for functions $f(\cdot, \cdot, \cdot)$ in some prespecified class which permits the u-dependence to be nonaffine.

Given a right hand side in $\mathcal{M}$ with solution set $\mathcal{D}$, the corresponding attainable set $\mathcal{D}(\cdot)$ is taken to be the set-valued function defined as

$$\mathcal{D}(t) = \{d(t) | d(\cdot) \in \mathcal{D}\}, \qquad t \in I.$$

In general, convergence of the solution sets is a stronger property than pointwise convergence of the attainable sets, thus Theorem 7.1 and Proposition 8.1 provide only sufficient conditions for convergence of the attainable sets. Only when $f(\cdot, \cdot, \cdot)$ is affine in (x, u) can one immediately assert that pointwise convergence of the attainable sets is equivalent to convergence of the solution sets. This equivalence easily follows from Corollary 4.3 and the representation of the attainable set through the variation of constants formula.

9. Pointwise conditions on the right hand sides. It has been established (Theorem 7.1) that convergence of solution sets and weak convergence of right hand sides $D_i(\cdot, d_0(\cdot))$, evaluated on all $d_0(\cdot)$'s in the limiting solution set, are equivalent. We have seen (Corollary 7.2) that the Theorem leads to a simple sufficient condition for convergence of the solution sets. In this section we show that, under additional hypotheses, such a condition, in this case weak convergence of the $D_i(\cdot, x)$'s for each $x \in \mathbb{R}^n$, is also necessary for convergence of the solution sets when we require convergence of the solution sets for *every* initial condition.

We introduce the new hypotheses. Let $D(\cdot, \cdot)$ be a right hand side.

$\overline{\text{A1}}$, $\overline{\text{A2}}$, $\overline{\text{A3}}$: $D(\cdot, \cdot)$ satisfies A1, A2, A3, respectively, with $\mathcal{O}$ taken as

$$\mathcal{O} = [0, 1] \times \mathbb{R}^n.$$

Notice that $\overline{\text{A3}}$ is a significant strengthening of A3, since it excludes right hand sides "unbounded in their x-dependence".

$\overline{\mathrm{A4}}$: There exists an $\mathbb{R}^n$-valued function $d(\cdot,\cdot)$ such that $\{d(\cdot,\cdot)\}$ satisfies $\overline{\mathrm{A1}}$, $\overline{\mathrm{A2}}$ and $\overline{\mathrm{A3}}$ and, for a.e. $t\in I$,

$$d(t,x)\in D(t,x) \quad \text{for all } x\in\mathbb{R}^n.$$

Given a right hand side which satisfies $\overline{\mathrm{A1}}$, $\overline{\mathrm{A2}}$, $\overline{\mathrm{A3}}$, it is always possible to find some "selection" $d(\cdot,\cdot)$ such that $\{d(\cdot,\cdot)\}$ satisfies $\overline{\mathrm{A1}}$, $\overline{\mathrm{A3}}$ but such that $d(t,x)$ is merely continuous in x. This follows simply from [6, Lemma 6] and [3, Lemma 2.5]. The substance of assumption $\overline{\mathrm{A4}}$ is that $d(t,x)$ may be chosen Lipschitz continuous in x. $\overline{\mathrm{A4}}$ is of course satisfied in the situation of primary interest, namely when

$$D(t,x)=f(t,x,U(t))$$

under mild assumptions on $f(\cdot,\cdot,\cdot)$ and $U(\cdot)$. A rather general condition for $D(\cdot,\cdot)$ to have a "Lipschitz continuous" selection $d(\cdot,\cdot)$ may be derived from a recent result of Ioffe [11].[2]

Under assumptions $\overline{\mathrm{A1}}$, $\overline{\mathrm{A2}}$, $\overline{\mathrm{A3}}$ on $D(\cdot,\cdot)$, assumptions A1, A2, A3 are satisfied for *arbitrary* initial condition $a\in\mathbb{R}^n$. Recall that a enters the assumptions A1–A3 through definition of the set P. We need to emphasize a in the notation, since it is no longer fixed. Let $\mathcal{D}^a$ be the solution set for

$$\dot{x}(t)\in D(t,x(t)),$$
$$x(0)=a,$$

in which $D(\cdot,\cdot)$ satisfies $\overline{\mathrm{A1}}$–$\overline{\mathrm{A4}}$. We write $\mathcal{D}^a(t)=\{x(t)|x(\cdot)\in\mathcal{D}^a\}$ for all $t\in I$. The important implication of $\overline{\mathrm{A4}}$ is that

$$\bigcup_{a\in\mathbb{R}^n}\mathcal{D}^a(t)=\mathbb{R}^n,\quad \text{all } t\in I. \tag{9.1}$$

Equation (9.1) is true when $D(\cdot,\cdot)$ is singleton-valued (see application of Lemma 6.4 in concluding lines of the proof of Theorem 6.1 in [2]) and is true a foriori when $D(\cdot,\cdot)$ is set-valued. We write $\bar{\mathcal{D}}$ for the subset of $C(I,\mathbb{R}^n)$:

$$\bar{\mathcal{D}}=\bigcup_{a\in\mathbb{R}^n}\mathcal{D}^a.$$

The following lemma is essentially a restatement in our setting of Lemma 6.2 of [2].

LEMMA 9.1. *Suppose that the right hand sides* $\{D_i(\cdot,\cdot)\}_{i=1}^{\infty}$, $D_0(\cdot,\cdot)$ *satisfy* $\overline{\mathrm{A1}}$–$\overline{\mathrm{A3}}$. *Suppose also that* $D_0(\cdot,\cdot)$ *satisfies* $\overline{\mathrm{A4}}$. *Then the following two statements*:

(i) $D_i(\cdot,\bar{d}(\cdot))\to D_0(\cdot,\bar{d}(\cdot))$ *weakly in* $L_1(I,\mathcal{R})$ *for every* $\bar{d}(\cdot)\in\bar{\mathcal{D}}_0$, *and*

(ii) $\int_0^t D_i(\tau,x)\,d\tau\to\int_0^t D_0(\tau,x)\,d\tau$, *for every* $(t,x)\in\mathcal{O}$

are equivalent.

Proof. Since the family $\{D_i(\cdot,\bar{d}(\cdot))\}_{i=1}^{\infty}$ is uniformly integrably bounded (i) is equivalent to

$$\int_0^t D_i(\tau,\bar{d}(\tau))\,d\tau\to\int_0^t D_0(\tau,\bar{d}(\tau))\,d\tau \tag{9.2}$$

for every $t\in I$, $\bar{d}(\cdot)\in\bar{\mathcal{D}}_0$.

Given any constant function $x(t)=x$, all $t\in I$, and any positive number $\varepsilon>0$, then there exists a continuous function $\bar{d}^*(\cdot)$ and a finite partition of I into intervals I_i such that $|x(\cdot)-\bar{d}^*(\cdot)|_I\leqq\varepsilon$ and such that, for each i, $\bar{d}^*(\cdot)$ coincides on I_i with some element

[2] We thank a reviewer for calling our attention to this reference.

in $\bar{\mathscr{D}}_0$. This is easily shown to follow from the property that

$$\bigcup_{a\in\mathbb{R}^n} \mathscr{D}^a(t)=\mathbb{R}^n \quad \text{for all } t\in I$$

and the bound on the derivative $\dot{d}(\cdot)$, $\dot{d}(t)\leqq h(t)$, a.e. $t\in I$, which is implied by assumption $\overline{\text{A3}}$.

Suppose (i) holds. By the uniform Lipschitz condition $\overline{\text{A2}}$, we have for $i=0,1,\cdots,$

$$\left|\int_0^t D_i(\tau,x)\,d\tau-\int_0^t D_i(\tau,\bar{d}^*(\tau))\,d\tau\right|\leqq\varepsilon\int_0^t k(\tau)\,d\tau, \qquad t\in I,$$

and thus (ii) follows since ε can be taken arbitrarily small.

We show that (ii) implies (9.2) following the standard argument outlined in the proof of Corollary 7.2. □

In view of Lemma 9.1, the following theorem now follows immediately from Theorem 7.1.

THEOREM 9.2. *Suppose that* $\{D_i(\cdot,\cdot)\}_{i=1}^{\infty}, D_0(\cdot,\cdot)$ *satisfy assumptions* $\overline{\text{A1}}, \overline{\text{A2}}, \overline{\text{A3}}$ *and that additionally* $D_0(\cdot,\cdot)$ *satisfies* $\overline{\text{A4}}$. *Then* $\mathscr{D}_i^a\to\mathscr{D}_0^a$, *for* any *initial condition* $a\in\mathbb{R}^n$, *if and only if*

$$\int_0^t D_i(\tau,x)\,d\tau\to\int_0^t D_0(\tau,x)\,d\tau$$

for all $(t,x)\in\mathscr{O}$.

10. Stability of a class of optimal control problems. In this final section we turn to a class of optimal control problems and apply Theorem 7.1 to characterize perturbations of the dynamics under which the minimum cost is stable.

We consider free endpoint problems:

$$(\mathscr{P}) \qquad \begin{array}{l}\text{minimize } c(x(\cdot)) \text{ subject to } \dot{x}(t)\in D(t,x(t)) \text{ a.e. on } I,\\ x(0)=a.\end{array}$$

Here $c(\cdot)$ is a continuous real-valued function on $C(I,\mathbb{R}^n)$ and $D(\cdot,\cdot)\in\mathscr{M}$.

$D_0(\cdot,\cdot)$ and $\{D_i(\cdot,\cdot)\}_{i=1}^{\infty}$ in $\mathscr{M}$ are given. With each $D_i(\cdot,\cdot)$ is associated an optimal control problem $\mathscr{P}_i$ obtained by substituting $D_i(\cdot,\cdot)$ for $D(\cdot,\cdot)$ in $\mathscr{P}$. $\mathscr{P}_0$ is interpreted as the unperturbed problem. We denote by J_i the corresponding minimum costs, and by $\mathscr{D}_i^*$ the corresponding minimizing sets, that is the subsets of $C(I,\mathbb{R}^n)$ on which the minima are achieved. The continuity of $c(\cdot)$ and the compactness of the solution sets $\mathscr{D}_i$ ensure that each $\mathscr{D}_i^*$ is nonempty, compact.

THEOREM 10.1. $J_i\to J_0$ *for every continuous* $c(\cdot)$, *if and only if* $D_i(\cdot,d_0(\cdot))\to D_0(\cdot,d_0(\cdot))$ *weakly in* $L_1(I,\mathscr{R})$, *for every* $d_0(\cdot)\in\mathscr{D}_0$. *Moreover if either of these equivalent conditions hold,* $\overline{\text{dist}}\,(\mathscr{D}_i^*,\mathscr{D}_0^*)\to 0$.

Proof. Suppose that $D_i(\cdot,d_0(\cdot))\to D_0(\cdot,d_0(\cdot))$ weakly in $L_1(I,\mathscr{R})$ for every $d_0(\cdot)\in\mathscr{D}_0$. By Theorem 7.1, $\mathscr{D}_i\to\mathscr{D}_0$ in $\mathscr{C}$.

Define the set valued mapping $\gamma(\cdot)$ on $\mathscr{C}$:

$$\gamma(\mathscr{V})=\{d(\cdot)\in C(I,\mathbb{R}^n)|d(\cdot)\in\mathscr{V}\}, \qquad \mathscr{V}\in\mathscr{C}.$$

Since $c(\cdot)$ is continuous on $C(I,\mathbb{R}^n)$ and γ is a continuous set-valued mapping in the sense of Berge [5, p. 114], we conclude from Berge's theorem of the maximum [5, p. 122] that the function on $\mathscr{C}$,

$$J(\mathscr{V})=\min\{c(x(\cdot))|x(\cdot)\in\gamma(\mathscr{V})\}$$

is continuous, and the set-valued mapping γ^* on $\mathscr{C}$:

$$\gamma^*(\mathscr{V})=\{x(\cdot)\in\mathscr{V}|c(x(\cdot))=J(\mathscr{V})\}$$

is upper semicontinuous (in the sense of Berge). Let $J_i=\gamma^*(\mathscr{D}_i)$. Since $\mathscr{D}_i\to\mathscr{D}_0$ we have that $J_i\to J_0$. But $\mathscr{D}_i^*=\gamma^*(\mathscr{D}_i)$ for $i=0,1,\cdots$, so that in addition, $\overline{\text{dist}}\,(\mathscr{D}_i^*,\mathscr{D}_0^*)\to 0$ by upper semicontinuity.

To establish the converse, suppose that $J_i\to J_0$, for every continuous $c(\cdot)$. Choosing in particular $c(x(\cdot))=|x(\cdot)-d_0(\cdot)|_I$, with $d_0(\cdot)\in\mathscr{D}_0$, $J_i\to J_0$ implies that $d_0(\cdot)$ is the limit (in $C(I,\mathbb{R}^n)$) of some sequence $\{d_i(\cdot)\}_{i=1}^{\infty}$ with $d_i(\cdot)\in\mathscr{D}_i^*\subset\mathscr{D}_i$. By the compactness of $\mathscr{D}_0$ then, noting that $d_0(\cdot)\in\mathscr{D}_0$ was arbitrary, we have $\overline{\text{dist}}\,(\mathscr{D}_0,\mathscr{D}_i)\to 0$. On the other hand, by pre-compactness of $\bigcup_{i=1}^{\infty}\mathscr{D}_i$, every sequence $\{d_i(\cdot)\}_{i=1}^{\infty}$ such that $d_i(\cdot)\in\mathscr{D}_i$, $i=1,2,\cdots$, contains a subsequence converging to some element $\bar{d}(\cdot)$ in $C(I,\mathbb{R}^n)$. Set $c(x(\cdot))=|x(\cdot)-\bar{d}(\cdot)|_I$. Then $J_i\to 0$, and, under the hypothesis, $J_0=0$. But this must mean $\bar{d}(\cdot)\in\mathscr{D}_0$, whence $\overline{\text{dist}}\,(\mathscr{D}_i,\mathscr{D}_0)\to 0$. We conclude that $\mathscr{D}_i\to\mathscr{D}_0$. By Theorem 7.1 then $D_i(\cdot,d_0(\cdot))\to D_0(\cdot,d_0(\cdot))$ weakly in $L_1(I,\mathscr{R})$ for any $d_0(\cdot)\in\mathscr{D}_0$. □

The final assertion of Theorem 10.1 has the interpretation: if the dynamics are slightly perturbed, then no solution for the perturbed problem will emerge which is not close to some solution of the unperturbed problem. In particular, if the control problems under consideration have unique solutions, then small perturbations of the dynamics (of the form described in Theorem 10.1) give rise to small perturbations of the solution.

We could state an analogous result to Theorem 10.1 under the stronger hypotheses of § 9, relating weak convergence of $\{D_i(\cdot,x)\}_{i=1}^{\infty}$, each $x\in\mathbb{R}^n$, to convergence of the minimum cost for all continuous functions $c(\cdot)$ and initial conditions $a\in\mathbb{R}^n$.

We have commented in § 8 that Theorem 7.1 concerns stability of the solution set $\mathscr{D}_0$, that is the subset of $C(I,\mathbb{R}^n)$ comprising "admissible trajectories", rather than stability of the attainable set $\mathscr{D}_0(\cdot)$, as in [1]. This is advantageous in applications to stability of control problems, since stability results are obtained for costs which are functionals on the whole trajectory.

We consider only free endpoint problems here. Conditions for stability in the presence of endpoint constraints are given in [10].

REFERENCES

[1] Z. ARTSTEIN, *Weak convergence of set-valued functions and control*, this Journal, 13 (1975), pp. 865–878.

[2] ———, *Continuous dependence on parameters: on the best possible results*, J. Differential Equations, 19 (1975), pp. 214–225.

[3] T. F. BRIDGLAND, JR., *Trajectory integrals of set-valued functions*, Pacific J. Math., 33 (1970), pp. 43–68.

[4] ———, *Contributions to the theory of generalized differential equations*, Math. Systems Theory, 3 (1969), pp. 17–50.

[5] C. BERGE, *Espaces Topologies*, Dunod, Paris, 1959.

[6] A. F. FILIPPOV, *Classical solutions of differential equations with multivalued right-hand side*, this Journal, 5 (1967), pp. 609–621.

[7] R. T. ROCKAFELLAR, *Measurable dependence of convex sets and functions on parameters*, J. Math. Anal. Appl., 28 (1969), pp. 4–25.

[8] L. C. YOUNG, *Lectures on the Calculus of Variations and Optimal Control Theory*, Saunders, Philadelphia, 1969.

[9] H. HERMES, *The generalized differential equation $\dot{x}\in R(t,x)$*, Advances in Math., 4 (1970), pp. 149–169.

[10] G. I. STASSINOPOULOS, *Weak topologies and stability of control problems under data perturbations*, Ph.D. Thesis, Imperial College University of London, England, 1977.

[11] A. D. IOFFE, *Representation theorems for multifunctions and analytic sets*, Bull. Amer. Math. Soc., 84 (1978), pp. 142–144.

[12] J. L. DAVY, *Properties of the solution set of a generalized differential equations*, Bull. Austr. Math. Soc., 6 (1972), pp. 379–398.

[13] J. T. MARKIN, *Stability of solution sets for generalized differential equations*, J. Math. Anal. Appl., 46 (1974), pp. 289–291.

[14] M. GRANGER, *Equations au contigent et problèmes d'optimisation*, Rev. CETHEDEC, 23 (1970), pp. 63–110.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 4, July 1979

0363-0129/79/1704-0001 $01.00

CANONICAL REALIZATION OF BILINEAR INPUT/OUTPUT MAPS*

J. G. PEARLMAN† AND M. J. DENHAM‡

Abstract. The idea of extending the comprehensive theory of state space realization for linear systems to certain classes of nonlinear systems has been the subject of much recent research in mathematical system theory. One such class are those systems with a bilinear input/output map, first considered by Kalman [*Pattern recognition properties of multi-linear machines*, IFAC Symposium on Technical and Biological Problems of Control, Yerevan, Armenian SSR, September, 1968] in 1968. Kalman's paper raised many questions concerning the canonical realization of such an input/output map; in particular it demonstrated that the proposed realization procedure could lead to a nonreachable realization. More recently, Sontag and Rouchaleau [*On discrete-time polynomial systems*, J. Non-linear Analysis, Theory, Methods, and Applications, 1 (1976), no. 1, pp. 55–64] have shown that the required concept to be applied in this case is "quasi-reachability," i.e. that the state set of the realization is the closure (in the Zariski topology) of the reachable set.

The present paper is a further contribution to this theory, providing a necessary and sufficient condition for quasi-reachability of a realization of a bilinear input/output map. Unexpectedly, this is given by the easily verifiable algebraic criterion of reachability of a corresponding linear system. This leads to a constructive procedure, analogous to the linear case, for reducing a realization of the bilinear map to a quasi-reachable one, thereby reducing the state dimension of the realization.

1. Introduction. The idea of extending the comprehensive theory of realization for linear systems to certain classes of nonlinear systems has been central to a great deal of recent research in mathematical system theory. One such class is the set of systems with a bilinear input/output map

$$f\colon U\times V\to Y \tag{1}$$

first considered from the realization viewpoint by Kalman [1] in 1968. Kalman's paper raised many questions concerning the canonical realization of such a system, demonstrating that in general the procedure proposed could lead to a realization with a state space which was not reachable. In § 2 of this paper, we outline this procedure and its module theoretic basis developed in [1].

More recently, Sontag and Rouchaleau [2] have defined the concept of "quasi-reachability" i.e. a system is quasi-reachable if its state set is the closure (in the Zariski topology) of the reachable set. They demonstrate that this is the correct property to apply to certain classes of nonlinear systems in a definition of a "canonical" realization, together with the notion of "algebraic observability" also defined in [2]. In § 3 of this paper, we introduce this property into our study of bilinear systems and prove the main result of the paper: that quasi-reachability of the bilinear system is equivalent to reachability of a corresponding linear system.

The condition for quasi-reachability leads naturally to a procedure for reducing any given realization of the bilinear system to one which is quasi-reachable. This is described in § 4.

2. Definitions and realization procedure. In this section we summarize the main results of [1] and the subsequent contributions of [3] based on these results.

* Received by the editors November 30, 1976, and in revised form August 25, 1978.

† Rothamsted Experimental Station, Harpenden, England. Formerly of the Department of Computing and Control, Imperial College of Science and Technology, London.

‡ School of Electronic Engineering and Computer Science, Kingston Polytechnic, Kingston-upon-Thames, England. Formerly of the Department of Computing and Control, Imperial College of Science and Technology, London.

We use U, V and Y to denote the following spaces:

$$U = \{u \in R^{z^-} \text{ with compact support}\},$$
$$V = \{v \in R^{z^-} \text{ with compact support}\},$$
$$Y = \{y \in R^{N-\{0\}}\}.$$

DEFINITION 2.1. A map $f\colon U \times V \to Y$ is a bilinear discrete-time stationary input/output map if it satisfies the following conditions:

(i) bilinearity: $f(k_1u_1 + k_2u_2, v) = k_1f(u_1, v) + k_2f(u_2, v)$

$$f(u, k_1v_1 + k_2v_2) = k_1f(u, v_1) + k_2f(u, v_2)$$

for all $k_1, k_2 \in R$; $u, u_1, u_2 \in U$; $v, v_1, v_2 \in V$;

(ii) stationarity: $f(\sigma u, \sigma v) = \sigma^*f(u, v)$
where σ and σ^* are shift operators:

$$\sigma u = \sigma(\cdots, u_1, u_0) = (\cdots, u_1, u_0, 0)$$
$$\sigma^* y = \sigma(y_1, y_2, \cdots) = (y_2, y_3, \cdots).$$

Introducing the backward shift operators z_1 and z_2, it is then possible to identify $U \times V$ with $R[z_1] \times R[z_2]$ and Y with the ring of formal power series $(z_1z_2)^{-1}R[[(z_1z_2)^{-1}]]$. The input/output map f can then be described by a power series:

$$s = (z_1z_2)^{-1} \sum_{i,j} s_{ij} z_1^{-i} z_2^{-j} \tag{2}$$

with inputs $u(z_1)$ and $v(z_2)$ producing an output

$$\begin{aligned} y(z_1z_2) &= \left[(z_1z_2)^{-1} \sum_{i,j} s_{ij} z_1^{-i} z_2^{-j} u(z_1) v(z_2)\right] \odot \sum_k (z_1z_2)^{-k} \\ &= \sum_k y_k (z_1z_2)^{-k} \end{aligned} \tag{3}$$

where the Hadamard product $\odot$ just picks out from the term in squared brackets all the terms in $(z_1z_2)^{-k}$. Clearly s_{ij} represents the output at time 1 due to unit inputs at time $-i$ in U and time $-j$ in V.

The intuitive notion of state space is introduced by means of the Nerode equivalence relation $\underset{N}{\sim}$, i.e.

$$(u_1, v_1) \underset{N}{\sim} (u_2, v_2) \quad \text{iff} \quad f(z_1^k u_1 + u, z_2^k v_1 + v) = f(z_1^k u_2 + u, z_2^k v_2 + v)$$

for all k and for all $u \in R[z_1]$, $v \in R[z_2]$ of degree less than k. This leads to the definition of Nerode equivalence classes

$$[u, v] = \{(u_1, v_1) \in U \times V \colon (u_1, v_1) \underset{N}{\sim} (u, v)\}$$

and the Nerode state space

$$X_N = \{[u, v] \colon (u, v) \in U \times V\}.$$

In order to analyze the abstract Nerode equivalence relation defining the canonical state set, the following three equivalence relations were introduced in [1]:

$$u_1 \underset{1}{\sim} u_2 \quad \text{iff } f(z_1^k u_1, v) = f(z_1^k u_2, v)$$

for all k and for all $v \in R[z_2]$ with degree less than k,

$$v_1 \underset{2}{\sim} v_2 \quad \text{iff} \quad f(u, z_2^k v_1) = f(u, z_1^k v_2)$$

for all k and for all $u \in R[z_1]$ with degree less than k,

$$(u_1, v_1) \underset{3}{\sim} (u_2, v_2) \quad \text{iff} \quad f(z_1^k u_1, z_2^k v_1) = f(z_1^k u_2, z_2^k v_2)$$

for all k.

These equivalence relations were introduced in [1] since together they turn out to be the same as Nerode equivalence, i.e.

$$(u_1, v_1) \underset{N}{\sim} (u_2, v_2) \quad \text{iff} \quad u_1 \underset{1}{\sim} u_2, v_1 \underset{2}{\sim} v_2 \quad \text{and} \quad (u_1, v_1) \underset{3}{\sim} (u_2 v_2).$$

A formal proof of this fact may be found in [5].

A realization procedure can then be formulated in terms of the following module morphisms:

$$\begin{aligned} R[z_1]\text{-morphism } f_1\colon & R[z_1] \to (z_1 z_2)^{-1} R[[(z_1 z_2)^{-1}]]^{1\times\infty} \\ & u(z_1) \mapsto [f(z_1 u, 1), f(z_1^2 u, 1), \cdots] \\ R[z_2]\text{-morphism } f_2\colon & R[z_2] \to (z_1 z_2)^{-1} R[[(z_1, z_2)^{-1}]]^{1\times\infty} \\ & v(z_2) \mapsto [f(1, z_2 v), f(1, z_2^2 v), \cdots] \\ R[z_1 z_2]\text{-morphism } f_\otimes\colon & R[z_1, z_2] \to (z_1 z_2)^{-1} R[[(z_1 z_2)^{-1}]]^{1\times\infty} \\ & w(z_1, z_2) \mapsto \sum_{i,j} s_{ij} z_1^{-i} z_2^{-j} w(z_1, z_2) \odot \sum_k (z_1 z_2)^{-k}. \end{aligned}$$

It follows that:

$$f_1(u_1) = f_1(u_2) \quad \text{iff} \quad u_1 \underset{1}{\sim} u_2,$$

$$f_2(v_1) = f_2(v_2) \quad \text{iff} \quad v_1 \underset{2}{\sim} v_2,$$

$$f_\otimes(u_1, v_1) = f_\otimes(u_2, v_2) \quad \text{iff} \quad (u_1, v_1) \underset{3}{\sim} (u_2, v_2).$$

By the standard but lengthy procedures from linear system realization theory, it can be shown [1] that the dynamical state transition equations for the bilinear system can be written in terms of state vectors $x^1 \in U/\ker f_1$, $x^2 \in V/\ker f_2$ and $x \in U \otimes V/\ker f_\otimes$. In closed form, the equations can be expressed as:

$$x_{k+1}^1 = A_1 x_k^1 + b_1 u_k, \tag{4}$$

$$x_{k+1}^2 = A_2 x_k^2 + b_2 v_k, \tag{5}$$

$$x_{k+1} = A x_k + Q_1 x_k^1 v_k + Q_2 x_k^2 u_k + b u_k v_k \tag{6}$$

$$y_k = h^T x_k. \tag{7}$$

These equations represent a solution of the realization problem when the property of finiteness of the state module is satisfied. The condition for this property was given in abstract module-theoretic terms in [1]. An equivalent transfer function criterion was given in [3], i.e. the above state space realization is finite dimensional if and only if the formal power series $s=(z_1z_2)^{-1}\sum_{i,j}s_{ij}z_1^{-i}z_2^{-j}$ can be expressed as:

$$s=\frac{N(z_1,z_2)}{p_1(z_1)p_2(z_2)p(z_1z_2)} \tag{8}$$

where $p_1(z_1)\in R[z_1]$, $p_2(z_2)\in R[z_2]$, $p(z_1z_2)\in R[z_1z_2]$ and $N(z_1,z_2)\in R[z_1,z_2]$.

The above realization however is not natural, or canonical, in the sense that the equivalence relations $\underset{1}{\sim}$, $\underset{2}{\sim}$ and $\underset{3}{\sim}$ were introduced in a somewhat arbitrary way in order to analyze X_N. As might be expected therefore a problem arises as to reachability of the resulting realization, as pointed out in [1]. In the following section we will make use of the recently introduced concept of "quasi-reachability" [2] to study this problem.

3. Reachability. In general, the state space realization described above is not reachable. In order to study this problem we will use two concepts. Firstly, we will concern ourselves with the property of quasi-reachability [2] rather than reachability.

DEFINITION 3.1. A state space realization of an input/output map is *quasi-reachable* if the state set is the closure (in the Zariski topology) of the reachable set.

Secondly, we define a realization of the bilinear map f which is slightly more general than that of (4)–(7), i.e.

$$x_{k+1}^1=A_1x_k^1+b_1u_k, \tag{9}$$

$$x_{k+1}^2=A_2x_k^2+b_2v_k, \tag{10}$$

$$x_{k+1}=Ax_k+Cx_k^1\otimes x_k^2+Q_1x_k^1v_k+Q_2x_k^2u_k+bu_kv_k, \tag{11}$$

$$y_k=h^Tx_k+d^Tx_k^1\otimes x_k^2. \tag{12}$$

Note the presence of the $x_k^1\otimes x_k^2$ terms in (11) and (12) and that the dimension of the realization is unchanged. It is easy to show by induction that, given zero initial state, y_k is a bilinear function of the inputs u_k and v_k and their past values. The advantage of introducing these terms will become apparent but in simple terms it is evident that the state $x\in U\otimes V/\ker f_\otimes$ in the realization (4)–(7) contains some knowledge of the tensor product of the states $x^1\otimes x^2$. By including this as an input in the transition equation for x we can eliminate that part of x relating to $x^1\otimes x^2$ and hence reduce the overall state dimension.

We can illustrate this with the simple example used in [1], i.e.

$$f\colon u(z_1)v(z_2)\mapsto h(z_1,z_2)u(z_1)v(z_2)$$

where

$$h(z_1,z_2)=\frac{1}{(z_1-a)(z_2-b)(z_1z_2-c)}.$$

Using the realization procedure described in the previous section, the equivalence classes corresponding to $\underset{1}{\sim}$, $\underset{2}{\sim}$, $\underset{3}{\sim}$ yield a four dimensional state space, with the following state transition equations:

$$x_{k+1}^1=ax_k^1+u_k,$$

$$x_{k+1}^2=bx_k^2+v_k,$$

$$\begin{bmatrix}x_{k+1,1}\\x_{k+1,2}\end{bmatrix}=\begin{bmatrix}ab&0\\1&c\end{bmatrix}\begin{bmatrix}x_{k,1}\\x_{k,2}\end{bmatrix}+\begin{bmatrix}a\\0\end{bmatrix}x_k^1v_k+\begin{bmatrix}b\\0\end{bmatrix}x_k^2u_k+\begin{bmatrix}1\\0\end{bmatrix}u_kv_k.$$

It is clear that if $x_{k,1}=x_k^1x_k^2=(ax_{k-1}^1+u_k)(bx_{k-1}^2+v_k)=(abx_{k-1}^1x_{k-1}^2+ax_{k-1}^1v_{k-1}$

$+bx^2_{k-1}u_{k-1}+u_{k-1}v_{k-1})$, as pointed out in [1], we can replace the last transition equation by:

$$\begin{aligned}x_{k+1,2}&=cx_{k,2}+x_{k,1}\\&=cx_{k,2}+abx_{k-1,1}+ax^1_{k-1}v_{k-1}+bx^2_{k-1}u_{k-1}+u_{k-1}v_{k-1}\\&=cx_{k,2}+x^1_kx^2_k.\end{aligned}$$

Hence the addition of $x^1_k\otimes x^2_k\,(=x^1_kx^2_k$ in this scalar case) has enabled the state dimension to be reduced by one, since $x_{k+1,1}$ can now be discarded as it plays no role in the output, $y_k=x_{k,2}$.

We now consider some intuitive ideas on reachability, first for linear systems. By a well known theorem we know that the system

$$x_{k+1}=Fx_k+gu_k$$

is not reachable if and only if there exists a row vector a^T such that $a^Tg=0$ and $a^TF=\lambda a$ for some complex scalar λ. In other words $a^Tx_{k+1}=\lambda a^Tx_k$ and, given zero initial state, the state space evolves on a hyperplane $a^Tx_k=0$.

With bilinear systems, using a certain amount of intuitive reasoning, we might expect the state to evolve on some hypersurface $p^Tx_k+q^Tx^1_k\otimes x^2_k=0$ if the realization (9)–(12) is not reachable. We make the initial assumptions, easily checked by a linear system criterion, that

(A1) (A_1, b_1) and (A_2, b_2) are reachable pairs,

(A2) (h^T, A) is an observable pair.

If these assumptions are not satisfied, we know from linear system theory how to reduce the dimension of x^1, x^2 and x to suit our requirements. In particular, (A2) tells us that the A matrix is cyclic, i.e. if A is diagonalized into Jordan form, there is only one Jordan block for each distinct eigenvalue of A. Hence, there is just one block corresponding to the zero eigenvalues of A, a fact we shall use later.

Before stating the main result, we prove the following lemma which ensures the linear independence of the transfer functions associated with the states x^1_k, x^2_k and x_k, under the condition of the main theorem.

LEMMA 3.1. *Let*

$$\begin{bmatrix}A_1\otimes A_2 & 0\\ C & A\end{bmatrix},\begin{bmatrix}A_1\otimes b_2 & b_1\otimes A_2 & b_1\otimes b_2\\ Q_1 & Q_2 & b\end{bmatrix}$$

be a reachable pair. Then the components of

$$h^1(z_1)\otimes h^2(z_2)=(z_1I-A_1)^{-1}b_1\otimes(z_2I-A_2)^{-1}b_2$$

and

$$\begin{aligned}h(z_1,z_2)=(z_1z_2I-A)^{-1}[C(z_1I-A_1)^{-1}b_1\otimes(z_2I-A_2)^{-1}b_2\\+Q_1(z_1I-A_1)^{-1}b_1+Q_2(z_2I-A_2)^{-1}b_2+b]\end{aligned}$$

are linearly independent.

Proof. Contrary to the statement of the lemma, suppose there exist p^T and q^T such that:

$$p^Th^1(z_1)\otimes h^2(z_2)+q^Th(z_1,z_2)=0. \tag{13}$$

The expansion of $h(z_1, z_2)$ in powers of $z_1^{-i}z_2^{-j}$ is given by:

$$h(z_1, z_2)=\sum_{k\geqq 0}(z_1z_2)^{-(k+1)}A^k\Bigg[C\sum_{i,j\geqq 0}(A_1^i\otimes A_2^j)(b_1\otimes b_2)z_1^{-(i+1)}z_2^{-(j+1)} \tag{14}$$
$$+Q_1\sum_{i\geqq 0}A_1^ib_1z_1^{-(i+1)}+Q_2\sum_{j\leqq 0}A_2^jb_2z_2^{-(j+1)}+b\Bigg].$$

Thus the coefficient of $(z_1z_2)^{-(r+1)}$, $r=0, 1, \cdots$, is given by

$$[0\quad I]\begin{bmatrix}A_1\otimes A_2 & 0\\ C & A\end{bmatrix}^r\begin{bmatrix}b_1\otimes b_2\\ b\end{bmatrix} \tag{15}$$

and of $(z_1z_2)^{-(r+1)}z_1^{-(s+1)}$, $r, s=0, 1, \cdots$, is given by

$$[0\quad I]\begin{bmatrix}A_1\otimes A_2 & 0\\ C & A\end{bmatrix}^r\begin{bmatrix}A_1^{s+1}b_1\otimes b_2\\ Q_1A_1^sb_1\end{bmatrix} \tag{16}$$

and of $(z_1z_2)^{-(r+1)}x_2^{-(s+1)}$, $r, s=0, 1, \cdots$, is given by

$$[0\quad I]\begin{bmatrix}A_1\otimes A_2 & 0\\ C & A\end{bmatrix}^r\begin{bmatrix}b_1\otimes A_2^{s+1}b_2\\ Q_2A_2^sb_2\end{bmatrix}. \tag{17}$$

The corresponding coefficients in the expansion of $h^1(z_1)\otimes h^2(z_2)$ are

$$(A_1\otimes A_2)^rb_1\otimes b_2,\qquad r=0, 1, \cdots, \tag{18}$$

$$(A_1\otimes A_2)^rA_1^{s+1}b_1\otimes b_2,\qquad r, s=0, 1, \cdots, \tag{19}$$

and

$$(A_1\otimes A_2)^rb_1\otimes A_2^{s+1}b_2,\qquad r, s=0, 1, \cdots. \tag{20}$$

Therefore, linear dependence implies that:

$$[p^T\quad q^T]\begin{bmatrix}A_1\otimes A_2 & 0\\ C & A\end{bmatrix}^r\begin{bmatrix}b_1\otimes b_2\\ b\end{bmatrix}=0,\qquad r=0, 1, \cdots, \tag{21}$$

$$[p^T\quad q^T]\begin{bmatrix}A_1\otimes A_2 & 0\\ C & A\end{bmatrix}^r\begin{bmatrix}A_1\otimes b_2\\ Q_1\end{bmatrix}A_1^sb_1=0,\qquad r, s=0, 1, \cdots, \tag{22}$$

and

$$[p^T\quad q^T]\begin{bmatrix}A_1\otimes A_2 & 0\\ C & A\end{bmatrix}^r\begin{bmatrix}b_1\otimes A_2\\ Q_2\end{bmatrix}A_2^sb_2=0,\qquad r, s=0, 1, \cdots. \tag{23}$$

But since (A_1, b_1) and (A_2, b_2) are reachable pairs, (22) and (23) reduce to:

$$[p^T\quad q^T]\begin{bmatrix}A_1\otimes A_2 & 0\\ C & A\end{bmatrix}^r\begin{bmatrix}A_1\otimes b_2\\ Q_1\end{bmatrix}=0,\qquad r=0, 1, \cdots, \tag{24}$$

and

$$[p^T\quad q^T]\begin{bmatrix}A_1\otimes A_2 & 0\\ C & A\end{bmatrix}^r\begin{bmatrix}b_1\otimes A_2\\ Q_2\end{bmatrix}=0,\qquad r=0, 1, \cdots, \tag{25}$$

which, together with (21), provide a contradiction to the condition of the lemma.

We now state the main result of this paper.

THEOREM 3.1. *Under the assumptions* (A1) *and* (A2), *the realization* (9)–(12) *is quasi-reachable if and only if*

$$\begin{bmatrix} A_1\otimes A_2 & 0 \\ C & A \end{bmatrix}, \begin{bmatrix} A_1\otimes b_2 & b_1\otimes A_2 & b_1\otimes b_2 \\ Q_1 & Q_2 & b \end{bmatrix} \triangleq (F, G)$$

is a reachable pair.

Thus, quasi-reachability of the bilinear system realization can be checked simply by a linear system criterion.

Proof. Necessity. Suppose that (F, G) is not a reachable pair. Then there exist row vectors p^T and q^T such that

$$(26)\qquad [p^T \quad q^T]\begin{bmatrix} A_1\otimes A_2 & 0 \\ C & A \end{bmatrix} = \lambda[p^T \quad q^T]$$

and

$$(27)\qquad [p^T \quad q^T]\begin{bmatrix} A_1\otimes b_2 & b_1\otimes A_2 & b_1\otimes b_2 \\ Q_1 & Q_2 & b \end{bmatrix} = 0.$$

If we expand x_{k+1} and $x^1_{k+1}\otimes x^2_{k+1}$ in terms of x_k, x^1_k, x^2_k, u_k and v_k using (9)–(12), then (26) and (27) imply that:

$$(28)\qquad p^Tx_{k+1}+q^Tx^1_{k+1}\otimes x^2_{k+1} = \lambda(p^Tx_{k+1}+q^Tx^1_{k+1}\otimes x^2_{k+1}).$$

Therefore, given zero initial state, i.e. $x_0 = x^1_0 = x^2_0 = 0$, the state of the realization evolves on a hypersurface

$$(29)\qquad p^Tx_k+q^Tx^1_k\otimes x^2_k = 0$$

for all k. Thus the realization is not quasi-reachable and necessity is proved.

Sufficiency. The proof of sufficiency will be carried out constructively, i.e. by specifying a desired state and then constructing input sequences $u\in U$ and $v\in V$ to reach this state at time $k=1$. Note that the state x_k at $k=1$ is given by the vector coefficient of $(z_1z_2)^{-1}$ in the expansion of $x(z_1, z_2) = h(z_1, z_2)u(z_1)v(z_2)$, where, using (9)–(11),

$$(30)\qquad \begin{aligned} h(z_1, z_2) = (z_1z_2I-A)^{-1}[&C(z_1I-A_1)^{-1}b_1(z_2I-A_2)^{-1}b_2 \\ &+Q_1(z_1I-A_1)^{-1}b_1+Q_2(z_2I-A_2)^{-1}b_2+b]. \end{aligned}$$

We assume, of course, zero initial state at the beginning of the input sequences, i.e. if J is the length of the input sequences, we assume $x_{-J} = x^1_{-J} = x^2_{-J} = 0$.

A preliminary consideration using linear system theory will determine what flexibility exists in choosing the required input sequences. Let $\psi_1(z)$ and $\psi_2(z)$ be the characteristic polynomials of A_1 and A_2 respectively. Then, given desired states x^1_1 and x^2_1, we know that there exist unique polynomials $q_1(z)$ and $q_2(z)$ with degree $(q_i) <$ degree (ψ_i), $i = 1, 2$ such that the input sequences

$$(31)\qquad u(z_1) = p_1(z_1)\psi_1(z_1)+q_1(z_1),$$

$$(32)\qquad v(z_2) = p_2(z_2)\psi_2(z_2)+q_2(z_2)$$

applied to (9) and (10) respectively reach x^1_1 and x^2_1 for all $p_1(z_1)$ and $p_2(z_2)$. Hence, for a given x^1_1, x^2_1 and x_1, the reachability problem is to choose $p_1(z_1)$ and $p_2(z_2)$ such that the desired state x_1 is reached.

The construction of these polynomials is fairly long and detailed and therefore we outline the two major stages in the proof of sufficiency of the theorem:

(i) For a suitable choice of matrix T, we apply a similarity transformation to A in (11) such that

$$TAT^{-1}=\begin{bmatrix} J_1 & O \\ 0 & J_0 \end{bmatrix} \tag{33}$$

where J_0 is the Jordan block corresponding to the zero eigenvalues of A and J_1 consists of the remaining Jordan blocks.

We then show, via Lemma 3.2, that the subset of states corresponding to J_0 is quasi-reachable and construct the input sequences necessary to reach the desired state, i.e. $p_1(z_1)$ and $p_2(z_2)$ in (31) and (32).

(ii) We then construct a further input sequence in Lemma 3.3 which reaches the remaining components of the desired x_1 corresponding to J_1. In fact we show that these states are reachable rather than quasi-reachable. Hence, for a given bilinear system, if A has no zero eigenvalues then the realization (9)–(12) is reachable under the condition of Theorem 3.1, not just quasi-reachable.

LEMMA 3.2. *Let $A=J_0\in R^{m\times m}$ in* (11). *Then the realization* (9)–(12) *is quasi-reachable if and only if* (F, G) *as defined in Theorem* 3.1, *is a reachable pair.*

Proof. Using (9)–(11), the state sequence due to inputs $u(z_1)$ and $v(z_2)$ is given by the m-vector

$$x(z_1z_2)=(z_1z_2I-J_0)^{-1}[C(z_1I-A_1)^{-1}b_1\otimes(z_2I-A_2)^{-1}b_2$$
$$+Q_1(z_1I-A_1)^{-1}b_1+Q_2(z_2I-A_2)^{-1}b_2+b]u(z_1)v(z_2)\odot\sum_k(z_1z_2)^{-k}$$

$$=\frac{1}{(z_1z_2)^m}\begin{bmatrix}(z_1z_2)^{m-1} & (z_1z_2)^{m-2}\cdots & 1\\ 0 & (z_1z_2)^{m-1}\cdots & (z_1z_2)\\ \vdots & \vdots & \vdots \\ 0 & 0\quad\cdots & (z_1z_2)^{m-1}\end{bmatrix}\begin{bmatrix}r_1(z_1,z_2)\\ \vdots \\ r_m(z_1,z_2)\end{bmatrix}\frac{u(z_1)v(z_2)}{\psi_1(z_1)\psi_2(z_2)}\odot\sum_k(z_1z_2)^{-k}. \tag{34}$$

Therefore, the ith component of $x(z_1z_2)$ is given by

$$x_i(z_1z_2)=\left(\sum_{j=1}^{m-i+1}(z_1z_2)^{-j}r_{i+j-1}(z_1,z_2)\right)\frac{(\alpha(z_1)\psi_1(z_1)+q_1(z_1))}{\psi_1(z_1)}$$
$$\cdot\frac{(\beta(z_2)\psi_2(z_2)+q_2(z_2))}{\psi_2(z_2)}\odot\sum_k(z_1z_2)^{-k}. \tag{35}$$

From this expression, we make the following observation: in forming the products

$$(z_1z_2)^{-j}r_{i+j-1}(z_1,z_2)[\alpha_0+\alpha_1z_1+\alpha_2z_1^2+\cdots+\alpha_sz_1^s],\qquad j=1,\cdots,m-i+1,$$

we can ignore all terms of $\alpha(z_1)$ in z_1^k for $k>m-i$, since these terms will only result in products involving zero or positive powers of z_1. The operation of the Hadamard product will eliminate these in $x_i(z_1z_2)$. A similar observation is true for $\beta(z_2)$; hence we write these polynomials as

$$\alpha(z_1)=\alpha_0+\alpha_1z_1+\alpha_2z_1^2+\cdots+\alpha_{m-i}z_1^{m-i},$$
$$\beta(z_2)=\beta_0+\beta_1z_2+\beta_2z_2^2+\cdots+\beta_{m-i}z_2^{m-i}$$

with no loss of generality.

We now rewrite the summation of (35) in two parts:

$$\sum_{j=1}^{m-i} (z_1z_2)^{-j} r_{i+j-1}(z_1, z_2) + (z_1z_2)^{-(m-i+1)} r_m(z_1z_2)$$

and, by the same reasoning as above, we observe that the products of $\alpha_{m-i}z_1^{m-i}$ and $\beta_{m-i}z_2^{m-i}$ with the first part make no contribution to $x_i(z_1z_2)$. We can rewrite (35) therefore in the form

$$\begin{aligned} x_i(z_1z_2) = \hat{x}_i(z_1z_2) + (z_1z_2)^{-(m-i+1)} r_m(z_1, z_2)\left[\alpha(z_1) + \frac{q_1(z_1)}{\psi_1(z_1)}\right] \\ \cdot\left[\beta(z_2) + \frac{q_2(z_2)}{\psi_2(z_2)}\right] \odot \sum_k (z_1z_2)^{-k}, \end{aligned} \tag{36}$$

where $\hat{x}_i(z_1z_2)$ does not depend on α_{m-i} or β_{m-i}. If we now lump all the remaining terms not involving α_{m-i} or β_{m-i} in $\hat{x}_i(z_1z_2)$ also, we get

$$\begin{aligned} x_1(z_1z_2) &= \hat{x}_i(z_1z_2) + (z_1z_2)^{-(m-i+1)} r_m(z_1, z_2)\Big[\alpha_{m-i}\beta_{m-i}(z_1z_2)^{m-i} \\ &\quad + \alpha_{m-i}z_1^{m-i}(\beta_0 + \beta_1 z_2 + \cdots + \beta_{m-i-1}z_2^{m-i-1}) + \frac{\alpha_{m-i}z_1^{m-i}q_2(z_2)}{\psi_2(z_2)} \\ &\quad + \beta_{m-i}z_2^{m-i}(\alpha_0 + \alpha_1 z_1 + \cdots + \alpha_{m-i-1}z_1^{m-i-1}) + \frac{\beta_{m-i}z_2^{m-i}q_1(z_1)}{\psi_1(z_1)}\Big] \odot \sum_k (z_1z_2)^{-k} \\ &\triangleq \hat{x}_i(z_1z_2) + \tilde{x}_i(z_1z_2). \end{aligned} \tag{37}$$

Since $\hat{x}_i(z_iz_2)$ does not depend on α_{m-i} or β_{m-i}, we can prove the required result if $\tilde{x}_i(z_1z_2)$ can be made to have an arbitrary term in $(z_1z_2)^{-1}$ by suitable choice of α_{m-i} and β_{m-1}.

We now attempt to eliminate more terms from $\tilde{x}_i(z_1z_2)$, using the action of the Hadamard product. Rewriting $\tilde{x}_i(z_1z_2)$ in the form

$$\begin{aligned} \tilde{x}_i(z_1z_2) = (z_1z_2)^{-1} r_m(z_1, z_2)\Big[\alpha_{m-i}\beta_{m-i} + \frac{\alpha_{m-i}}{z_2^{m-i}}(\beta_0 + \beta_1 z_2 + \cdots + \beta_{m-i-1}z_2^{m-i-1}) \\ + \frac{\alpha_{m-i}q_2(z_2)}{z_2^{m-i}\psi_2(z_2)} + \frac{\beta_{m-i}}{z_1^{m-i}}(\alpha_0 + \alpha_1 z_1 + \cdots + \alpha_{m-i-1}z_1^{m-i-1}) \\ + \frac{\beta_{m-i}q_1(z_1)}{z_1^{m-i}\psi_1(z_1)}\Big] \odot \sum_k (z_1z_2)^{-k}, \end{aligned} \tag{38}$$

we observe that any term in $r_m(z_1, z_2)$ which cancels out the $(z_1z_2)^{-1}$ term will not contribute to $\tilde{x}_i(z_1z_2)$, since no negative powers of z_1z_2 will result. We can therefore, with no loss of generality, write $r_m(z_1, z_2)$ in the form

$$r_m(z_1, z_2) = a(z_1) + b(z_2) + c \tag{39}$$

where $a(z_1) = \sum_{j=1}^{n_1} a_j z_1^j$, $b(z_2) = \sum_{j=1}^{n_2} b_j z_2^j$ i.e. involving no terms in $z_1^i z_2^j$, $i, j \neq 0$.

We can now directly identify the required term in $\tilde{x}_i(z_1z_2)$, the coefficient of $(z_1z_2)^{-1}$, as

$$\begin{aligned} \tilde{x}_i^1 &= \alpha_{m-i}\beta_{m-i}c + \alpha_{m-i}(\beta_0 b_{m-i} + \beta_1 b_{m-i-1} + \cdots + \beta_{m-i-1}b_1) \\ &\quad + \alpha_{m-i}g_{m-i} + \beta_{m-i}(\alpha_0 a_{m-i} + \alpha_1 a_{m-i-1} + \cdots + \alpha_{m-i-1}a_1) + \beta_{m-i}f_{m-i} \end{aligned} \tag{40}$$

where g_{m-i} and f_{m-i} are the coefficients of z_2^{m-i} and z_1^{m-i} in the expansions of

$$\frac{b(z_2)q_2(z_2)}{\psi_2(z_2)} \quad \text{and} \quad \frac{a(z_1)q_1(z_1)}{\psi_1(z_1)}$$

respectively.

The following special cases can now be identified:

(i) $c \neq 0$: in this case α_{m-i} and β_{m-i} can be chosen arbitrarily to give any x_i^1 and the lemma is proved. Moreover, in this case we have reachability not just quasi-reachability.

(ii) $c = 0$: in this case we require that $\alpha(z_1)$ and $\beta(z_2)$ can be chosen so that

$$\beta_0 b_{m-i} + \cdots + \beta_{m-i-1} b_1 + g_{m-i} \neq 0 \tag{41}$$

or

$$\alpha_0 a_{m-i} + \cdots + \alpha_{m-i-1} a_1 + f_{m-i} \neq 0 \tag{42}$$

for all $i = 1, \cdots, m$.

We first note that $\alpha(z_1)$ and $\beta(z_2)$ cannot both be zero for $c = 0$. Otherwise, we could express $r_m(z_1, z_2)$ in (34) as $(z_1 z_2) r_m'(z_1, z_2)$, recalling our reasons for choosing $r_m(z_1, z_2)$ in the form (39). Thus, the mth component of $h(z_1, z_2)$ has the form

$$h_m(z_1, z_2) = \frac{(z_1 z_2)^{-1} r_m(z_1, z_2)}{\psi_1(z_1)\psi_2(z_2)} = \frac{r_m'(z_1, z_2)}{\psi_1(z_1)\psi_2(z_2)}$$

which is linearly dependent on the components of $h^1(z_1) \otimes h^2(z_2)$, in contradiction of Lemma 3.1.

Next, assume that $a(z_1) \neq 0$ and $b(z_2) = 0$. Further, let $a_j = 0$, $j = 1, \cdots, t < m-1$. Clearly, since $g_{m-i} = 0$ for all i and $b(z_2) = 0$, (41) cannot be satisfied. However, given the condition on $a(z_1)$, (29) requires that

$$f_j \neq 0, \qquad j = 0, \cdots, t, \tag{43}$$

$$\begin{gathered} \alpha_0 a_{t+1} + f_{t+1} \neq 0 \\ \vdots \\ \alpha_0 a_{m-1} + \cdots + \alpha_{m-t-2} a_{t+1} + f_{m-i} \neq 0. \end{gathered} \tag{44}$$

We can interpret (43) as a constraint on the coefficients of $q_1(z_1)$ and hence on the x^1 state, whereby x^1 is restricted not to lie in a union of hyperplanes in R^{n_1}. Apart from this constraint it is possible to choose α_i, $i = 0, \cdots, m-t-2$, such that (43), (44) are satisfied and therefore quasi-reachability is proved.

The same argument holds for the cases $a(z_1) = 0$, $b(z_2) \neq 0$ and $a(z_1) \neq 0$, $b(z_2) \neq 0$ and hence the lemma is proved in all cases.

We can now proceed to the second stage of the proof of Theorem 3.1, i.e. the construction of a further input sequence which reaches the remaining components of the desired x^1 corresponding to J_1. We recall from the proof of Lemma 3.2 that coefficients α_j in $\alpha(z_1)$ and β_j in $\beta(z_2)$, for $j \geqq m-i$ have no effect on $x_i(z_1 z_2)$. Thus all inputs of the form $z_1^j \psi_1(z_1)$ and $z_2^j \psi_2(z_2)$, $j > m-1$, have no effect on the J_0 subsystem or on the x^1 or x^2 states.

LEMMA 3.3. *Let* $A = J_1 \in R^{(n-m)\times(n-m)}$ *in* (11), *where* J_1 *has no zero eigenvalues. Then the realization* (9)–(12) *is reachable if and only if* (F, G) *as defined in Theorem* 3.1 *is a reachable pair.*

Proof. The transfer function of (9)–(11) will have the form

$$h(z_1, z_2) = \frac{s(z_1, z_2)}{\phi(z_1z_2)\psi_1(z_1)\psi(z_2)} \in R^{n-m}[(z_1, z_2)^{-1}] \tag{45}$$

where $\phi(z_1z_2)$ is the characteristic polynomial of J_1, $\phi(z) = \phi_0 + \phi_1 z + \cdots + z^{n-m}$. We can rewrite this in the form

$$h(z_1, z_2) = \frac{s(z_1, z_2)(z_1z_2)^m}{\phi(z_1z_2)\bar{\psi}_1(z_1)\bar{\psi}_2(z_2)} \tag{46}$$

where $\bar{\psi}_1(z_1) = z_1^m\psi_1(z_1)$, $\bar{\psi}_2(z_2) = z_2^m\psi_2(z_2)$.

We can now consider constructing input sequences of the form

$$u(z_1) = p_1(z_1)\bar{\psi}_1(z_1) + \bar{q}_1(z_1), \tag{47}$$

$$v(z_2) = p_2(z_2)\bar{\psi}_2(z_2) + \bar{q}_2(z_2) \tag{48}$$

where

$$\bar{q}_1(z_1) = \alpha(z_1)\psi_1(z_1) + q_1(z_1), \qquad \bar{q}_2(z_2) = \beta(z_2)\psi_2(z_2) + q_2(z)$$

are as constructed in Lemma 3.2 and due to the choice of $\bar{\psi}_1(z_1)$ and $\bar{\psi}_2(z_2)$, the remainder of the sequences has no effect on the J_0 subsystem. If the bilinear map (46) is represented by

$$g: R[z_1] \times R[z_2] \to R^{n-m}[(z_1z_2)^{-1}]$$
$$:(u(z_1), v(z_2)) \mapsto h(z_1, z_2)u(z_1)v(z_2) \odot \sum_k (z_1z_2)^{-k}$$

it follows that using (47) and (48) we find

$$\begin{aligned} g(u(z_1), v(z_2)) &= g(p_1(z_1)\bar{\psi}_1(z_1), p_2(z_2)\bar{\psi}_2(z_2)) \\ &\quad + g(p_1(z_1)\bar{\psi}_1(z_1), \bar{q}_2(z_2)) + g(\bar{q}_1(z_1), p_2(z_2)\bar{\psi}_2(z_2)) \\ &\quad + g(\bar{q}_1(z_1), \bar{q}_2(z_2)). \end{aligned} \tag{49}$$

We now choose $p_1(z_1)$ and $p_2(z_2)$ in such a way that the middle two terms of (49) become zero, and use only the first term to achieve reachability, assuming the final term is fixed.

Consider then

$$\begin{aligned} g(p_1(z_1)\bar{\psi}_1(z_1), \bar{q}_2(z_2)) &= \frac{(z_1z_2)^m s(z_1, z_2)}{\phi(z_1z_2)\bar{\psi}_1(z_1)\bar{\psi}_2(z_2)} p_1(z_1)\bar{\psi}_1(z_1)\bar{q}_2(z_2) \odot \sum_k (z_1z_2)^{-k} \\ &= \frac{(z_1z_2)^m s(z_1, z_2)}{\phi(z_1z_2)\bar{\psi}_2(z_2)} p_1(z_1)\bar{q}_2(z_2) \odot \sum_k (z_1z_2)^{-k}. \end{aligned} \tag{50}$$

It is possible to set (50) to zero by choosing $p_1(z_1) = z_1^{m_1}\bar{p}_1(z_1)$, where m_1 is sufficiently large that all terms in $z_1^{m_1}s(z_1, z_2)$ have order in z_1 greater or equal to that in z_2. This choice ensures that the expansion of (50) before the Hadamard product has no terms in $(z_1z_2)^{-i}$, so that $g(p_1(z_1)\bar{\psi}_1(z_1), \bar{q}_2(z_2)) = 0$.

Similarly, we set $p_2(z_2) = z_2^{m_2}\bar{p}_2(z_2)$ such that $g(\bar{q}_1(z_1), p_2(z_2)\bar{\psi}_2(z_2)) = 0$ in (49).

We must now show that with this choice of $p_1(z_1)$ and $p_2(z_2)$ it is possible to obtain reachability using

$$\begin{aligned} &g(p_1(z_1)\bar{\psi}_1(z_1), p_2(z_2)\bar{\psi}_2(z_2)) \\ &\quad = \frac{z_1^{m_1}z_2^{m_2}(z_1z_2)^m s(z_1, z_2)}{\phi(z_1z_2)} \bar{p}_1(z_1)\bar{p}_2(z_2) \odot \sum_k (z_1z_2)^{-k}. \end{aligned} \tag{51}$$

We write

$$z_1^{m_1}z_2^{m_2}(z_1z_2)^m s(z_1, z_2)=N(z_1, z_2)+\phi(z_1z_2)M(z_1, z_2) \tag{52}$$

such that the highest order term in (z_1z_2) of $N(z_1, z_2)$ is of order $n-m=$ degree of $\phi(z_1z_2)$. We maintain that the components of $N(z_1, z_2)$ are linearly independent. Supposing the contrary, then there exists C^T such that $C^TN(z_1, z_2)=0$. Thus

$$z_1^{m_1}z_2^{m_2}(z_1z_2)^m C^T s(z_1, z_2)=\phi(z_1z_2)C^TM(z_1z_2) \tag{53}$$

and hence $\phi(z_1z_2)$ divides $C^Ts(z_1, z_2)$ since $\phi(z_1z_2)$ has no zero roots. Thus, by (32):

$$\begin{aligned} C^Th(z_1, z_2) &= \frac{C^Ts(z_1, z_2)}{\phi(z_1z_2)\psi_1(z_1)\psi_2(z_2)} \\ &= \frac{k(z_1, z_2)}{\psi_1(z_1)\psi_2(z_2)} \\ &= d^T(h^1(z_1)\otimes h^2(z_2)). \end{aligned} \tag{54}$$

for some d^T, and the linear independence condition of Lemma 3.1 is violated. If (52) is now substituted into (51), we get:

$$g(p_1(z_1)\bar{\psi}_1(z_1), p_2(z_2)\bar{\psi}_2(z_2))=\frac{N(z_1, z_2)}{\phi(z_1z_2)}\bar{p}_1(z_1)\bar{p}_2(z_2)\odot\sum_k (z_1z_2)^{-k} \tag{55}$$

since $M(z_1, z_2)\phi(z_1z_2)$ does not contribute. It now remains to show that $\bar{p}_1(z_1)$ and $\bar{p}_2(z_2)$ can be chosen such that the coefficient of $(z_1z_2)^{-1}$ in (55) has any specified value.

The terms of $N(z_1, z_2)$ will be of the form $(z_1z_2)^kz_1^j$, $k=0, 1, \cdots, n-m$; $j=0, 1, \cdots, l_1$, and $(z_1z_2)^kz_2^j$, $k=0, 1, \cdots, n-m$; $j=0, 1, \cdots, l_2$, for some l_1, l_2.

We therefore define:

$$e_j=[z_2^{-j}\quad (z_1z_2)z_2^{-j}\cdots(z_1z_2)^{n-m}z_2^{-j}]^T, \qquad j=-l_2, \cdots, 0,$$

and

$$e_j=[z_1^j\quad (z_1z_2)z_1^j\cdots(z_1z_2)^{n-m}z_1^j]^T, \qquad j=1, \cdots, l_1.$$

Then, since $N(z_1, z_2)$ is composed of independent terms from e_j, $j=0, \cdots, -l_2$ and e_j, $j=1, \cdots, l_1$, it follows from (55) that the realization of $h(z_1, z_2)$ in (45) is reachable if and only if the coefficient of $(z_1z_2)^{-1}$ in the response

$$w(z_1z_2)=\frac{1}{\phi(z_1z_2)}\begin{bmatrix} e_{-l_2} \\ \vdots \\ e_0 \\ e_1 \\ \vdots \\ e_{l_1} \end{bmatrix}\bar{p}_1(z_1)\bar{p}_2(z_2)\odot\sum_k (z_1z_2)^{-k} \tag{56}$$

can be chosen arbitrarily by a suitable selection of $\bar{p}_1(z_1)\bar{p}_2(z_2)$.

Let (C^T, A, b) be a minimal realization of $1/\phi(z_1z_2)$ and let $N>l_1+l_2$ be chosen such that (A^N, A^kb) is a reachable pair for all k. Such an integer N always exists if and only if A is nonsingular; this result is proven in [4].

We now choose $\bar{p}_1(z_1)$ and $\bar{p}_2(z_2)$ in the following way:

$$\bar{p}_1(z_1)=z_1^{l_2}(1+z_1^N+z_1^{2N}+\cdots+z_1^{(n-m-1)N}), \tag{57}$$

$$\bar{p}_2(z_2)=\alpha_1(z_2)+z_2^N\alpha_2(z_2)+\cdots+z_2^{(n-m-1)N}\alpha_{n-m}(z_2) \tag{58}$$

where

$$\alpha_j(z_2)=\alpha_{-l_2,j}+\alpha_{-l_2+1,j}z_2+\cdots+\alpha_{0,j}z_2^{l_2}+\alpha_{1,j}z_2^{l_2+1}+\cdots+\alpha_{l_1,j}z_2^{l_2+l_1}$$

for $j=1,\cdots,n-m$.

The following diagram illustrates the form of the input sequences $\bar{p}_1$ and $\bar{p}_2$ corresponding to time steps k taken from the beginning of the input sequences.

k	$\bar{p}_1$	$\bar{p}_2$
0	0	$\alpha_{-l_2,1}$
⋮	⋮	⋮
l_2-1	0	$\alpha_{-1,1}$
l_2	1	$\alpha_{0,1}$
l_2+1	0	$\alpha_{1,1}$
⋮	⋮	⋮
l_1+l_2	0	$\alpha_{l_1,1}$
l_1+l_2+1	0	0
⋮	⋮	⋮
N	0	$\alpha_{-l_2,2}$
⋮	⋮	⋮
l_2+N	1	$\alpha_{0,2}$
⋮	⋮	⋮
l_1+l_2+N	0	$\alpha_{l_1,2}$
etc.		

Note that by our choice of $N>l_1+l_2$, there is no overlapping of the α_j sequences in $\bar{p}_2$.

Considering the response (56) to these input sequences, we see that its components have the form

$$\begin{aligned}\frac{(z_1z_2)^i z_1^j}{\phi(z_1z_2)}\bar{p}_1(z_1)\bar{p}_2(z_2)\odot\sum_k (z_1z_2)^{-k}\\ =\frac{(z_1z_2)^{i+l_2+j}}{\phi(z_1z_2)}(\alpha_{j,1}+(z_1z_2)^N\alpha_{j,2}+\cdots+(z_1z_2)^{(n-m-1)N}\alpha_{j,n-m})\end{aligned} \tag{59}$$

for $i=0,\cdots,n-m-1;\ j=1,\cdots,l_1$, and

$$\begin{aligned}\frac{(z_1z_2)^i z_2^j}{\phi(z_1z_2)}\bar{p}_1(z_1)\bar{p}_2(z_2)\odot\sum_k (z_1z_2)^{-k}\\ =\frac{(z_1z_2)^{i+l_2+j}}{\phi(z_1z_2)}(\alpha_{j,1}+(z_1z_2)^N\alpha_{j,2}+\cdots+(z_1z_2)^{(n-m-1)N}\alpha_{j,n-m})\end{aligned} \tag{60}$$

for $i=0,\cdots,n-m-1; j=-l_2,\cdots,0$. Hence, for all j, the terms $\alpha_{j,i}$, $i=0,\cdots,n-m-1$, only appear in the component in the response $w(z_1z_2)$ in (56) corresponding to e_j.

If we now label the coefficient of $(z_1z_2)^{-1}$ in $w(z_1z_2)$ corresponding to e_j by $[w_{j,0}\cdots w_{j,n-m-1}]^T$, it follows from (59) and (60) that

$$\begin{aligned} w_{j,i} &= \sum_{k=0}^{n-m-1} C^T A^{i+l_2+j} A^{kN} b\alpha_{j,k+1} \\ &= C^T A^i [A^{l_2+j} b A^N A^{l_2+j} b \cdots A^{(n-m)N} A^{l_2+j} b] \begin{bmatrix} \alpha_{j,1} \\ \alpha_{j,2} \\ \vdots \\ \alpha_{j,n-m} \end{bmatrix} \qquad (61) \\ &\triangleq C^T A^i P_j \alpha_j \end{aligned}$$

for $j=-l_2,\cdots,0,\cdots,l_1; i=0,\cdots,n-m-1$. Hence

$$\begin{bmatrix} w_{j,0} \\ \vdots \\ w_{j,n-m-1} \end{bmatrix} = \begin{bmatrix} C^T \\ \vdots \\ C^T A^{n-m-1} \end{bmatrix} P_j \alpha_j. \qquad (62)$$

Since (C^T, A) is observable and $(A^N, A^{l_2+j}b)$ is controllable it follows that, for all $j=-l_2,\cdots,0,\cdots,l_1$, α_j can be chosen to achieve any desired output vector $[w_{j,0},\cdots,w_{j,n-m-1}]^T$. Thus the coefficient of $(z_1z_2)^{-1}$ in the output response $w(z_1z_2)$ in (56) can be chosen arbitrarily by selection of $\bar{p}_1(z_1)$ and $\bar{p}_2(z_1)$ given by (57) and (58).

Recalling that this ensures reachability of the realization (9)–(12) of $h(z_1,z_2)$ in (45), we see the lemma is proved.

Since the choice of input sequences $u(z_1)$ and $v(z_2)$ in Lemma 3.3, for the case $A=J_1$, was consistent with the choice of the input sequences in Lemma 3.2, for the case $A=J_0$, the proof of sufficiency of Theorem 3.1 is now also complete.

It is interesting to recall the form of the resultant input sequences constructed by Lemmas 3.2 and 3.3, namely

$$u(z_1)=\alpha(z_1)\psi_1(z_1)+q_1(z_1)+p_1(z_1)z_1^m\psi_1(z_1), \qquad (63)$$

$$v(z_2)=\beta(z_2)\psi_2(z_2)+q_2(z_2)+p_2(z_2)z_2^m\psi_2(z_2) \qquad (64)$$

where

$q_1(z_1)=$ input required to reach x_1^1 in (9),

$q_2(z_2)=$ input required to reach x_1^2 in (10),

$\alpha(z_1)$, $\beta(z_2)=$ inputs required to reach subset of x_1 in (11) corresponding to J_0 in (33),

$p_1(z_1)$, $p_2(z_2)=$ inputs required to reach subset of x_1 in (11) corresponding to J_1 in (33).

We note, for example in (63), that

$$u(z_1)=[\alpha(z_1)+p_1(z_1)z_1^m]\psi_1(z_1)+q_1(z_1) \qquad (65)$$

where degree $(\alpha(z_1))<m$ since higher order terms have no effect on the subset of x_1 corresponding to J_0. Hence $\alpha(z_1)+p_1(z_1)z_1^m$ represents the concatenation of two sequences, used together to reach (in the quasi-reachable sense) the whole of the x_1 state. The factor $\psi_1(z_1)$ ensures that this sequence has no effect on the x^1 state.

4. Reduction to quasi-reachable form. As in the linear system case, the procedure of reducing a realization which is not quasi-reachable to one that is, is based on the use of state transformations. We first note the existence of the three obvious similarity transformations, namely

$$x_k^1 \to T_1 x_k^1,$$
$$x_k^2 \to T_2 x_k^2,$$
$$x_k \to T x_k$$

with the associated transition matrices $A_1 \to T_1 A_1 T_1^{-1}$ etc. However there is a further transformation described in the following lemma:

LEMMA 4.1. *Let* (9)–(12) *be a realization of a bilinear input/output map* f: $U \times V \to Y$. *Then, if we transform*

$$C \to W(A_1 \otimes A_2) + C - AW,$$
$$Q_1 \to Q_1 + W(A_1 \otimes b_2),$$
$$Q_2 \to Q_2 + W(b_1 \otimes A_2),$$
$$b \to b + W(b_1 \otimes b_2),$$
$$d^T \to d^T - h^T W$$

for any $W \in R^{n \times n_1 n_2}$, *the resultant set of equations* (9)–(12) *is also a realization of* f.

Proof. The proof is by direct computation of the transfer function of f with and without the above transformation.

Note that this transformation is equivalent to the transformation matrix $\begin{bmatrix} I \otimes I & 0 \\ W & I \end{bmatrix}$ applied to the linear system described by the triple

$$(H, F, G) = \left([d^T \quad h^T], \begin{bmatrix} A_1 \otimes A_2 & 0 \\ C & A \end{bmatrix}, \begin{bmatrix} A_1 \otimes b_2 & A_2 \otimes b_1 & b_1 \otimes b_2 \\ Q_1 & Q_2 & b \end{bmatrix} \right).$$

Before describing the reduction procedure, we prove the following lemma.

LEMMA 4.2. *If* (A_1, b_1) *and* (A_2, b_2) *are reachable pairs, then* $(A_1 \otimes A_2, [A_1 \otimes b_2 \quad b_1 \otimes A_2 \quad b_1 \otimes b_2])$ *is a reachable pair.*

Proof. Suppose otherwise. Then there exists $v \in R^{n_1 n_2}$ such that

$$v^T A_1 \otimes A_2 = \lambda v^T \tag{66}$$

for some complex number λ, and

$$v^T [A_1 \otimes b_2 \quad b_1 \otimes A_1 \quad b_1 \otimes b_2] = 0. \tag{67}$$

In particular, $v^T b_1 \otimes A_2 = 0$ implies that

$$v^T b_1 \otimes A_2^k b_2 = 0 \tag{68}$$

and $v^T A_1 \otimes b_2$ implies that

$$v^T A_1^k b_1 \otimes b_2 = 0 \tag{69}$$

for all k. Then, using (66), we have

$$v^T A_1^{j+k} b_1 \otimes A_2^j b_2 = \lambda^j v^T A_1^k b_1 \otimes b_2 = 0 \tag{70}$$

and

$$v^T A_1^j b_1 \otimes A_2^{j+k} b_2 = \lambda^j v^T b_1 \otimes A_2^k b_2 = 0 \tag{71}$$

for all j, k.

Now (A_1, b_1), (A_2, b_2) reachable implies that

$$\{A_1^i b_1 \otimes A_2^j b_2 \colon i = 1, \cdots, n_1 - 1; j = 1, \cdots, n_2 - 1\}$$

is a basis for $R^{n_1 n_2}$. Hence (70) and (71) together imply that $v^T = 0$, and the lemma is proved.

To describe the reduction procedure, assume the system (9)–(12) is not quasi-reachable. Let L denote the controllability matrix of the pair

$$\left(\begin{bmatrix} A_1 \otimes A_2 & 0 \\ C & A \end{bmatrix}, \begin{bmatrix} A_1 \otimes b_2 & b_1 \otimes A_2 & b_1 \otimes b_2 \\ Q_1 & Q_2 & b \end{bmatrix} \right).$$

By Theorem 3.1, L does not have full rank.

Using Lemma 4.2 we normalize L to the form $\begin{bmatrix} I & 0 \\ L_1 & L_2 \end{bmatrix}$, where I is the identity matrix of $R^{n_1 n_2 \times n_1 n_2}$.

Using the standard property of controllability matrices

$$\begin{bmatrix} A_1 \otimes A_2 & 0 \\ C & A \end{bmatrix} \begin{bmatrix} I & 0 \\ L_1 & L_2 \end{bmatrix} = \begin{bmatrix} I & 0 \\ L_1 & L_2 \end{bmatrix} \begin{bmatrix} A_1 \otimes A_2 & 0 \\ E_1 & E_2 \end{bmatrix} \tag{72}$$

for some E_1, E_2. Also

$$\begin{bmatrix} A_1 \otimes b_2 & b_1 \otimes A_1 & b_1 \otimes b_2 \\ Q_1 & Q_2 & b \end{bmatrix} = \begin{bmatrix} I & 0 \\ L_1 & L_2 \end{bmatrix} \begin{bmatrix} A_1 \otimes b_2 & b_1 \otimes A_2 & b_1 \otimes b_2 \\ & E & \end{bmatrix} \tag{73}$$

for some E.

We now append the matrix $\begin{bmatrix} o \\ L_3 \end{bmatrix}$ to L, where L_3 is independent of L_2, and calculate

$$\begin{bmatrix} I & 0 & 0 \\ L_1 & L_2 & L_3 \end{bmatrix}^{-1} = \begin{bmatrix} I & 0 \\ -N_1 L_1 & N_1 \\ -N_2 L_1 & N_2 \end{bmatrix}. \tag{74}$$

Then, using (72), we calculate

$$\begin{aligned} &\begin{bmatrix} I & 0 \\ -N_1 L_1 & N_1 \\ -N_2 L_1 & N_2 \end{bmatrix} \begin{bmatrix} A_1 \otimes A_2 & 0 \\ C & A \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ L_1 & L_2 & L_3 \end{bmatrix} \\ &\quad = \begin{bmatrix} A_1 \otimes A_2 & 0 & 0 \\ N_1[C + AL_1 - L_1(A_1 \otimes A_2)] & N_1 A L_2 & N_1 A L_3 \\ 0 & 0 & N_2 A L_3 \end{bmatrix} \end{aligned} \tag{75}$$

and

$$\begin{aligned} &\begin{bmatrix} I & 0 \\ -N_1 L_1 & N_1 \\ -N_2 L_1 & N_2 \end{bmatrix} \begin{bmatrix} A_1 \otimes b_2 & b_1 \otimes A_2 & b_1 \otimes b_2 \\ Q_1 & Q_2 & b \end{bmatrix} \\ &\quad = \begin{bmatrix} A_1 \otimes b_2 & b_1 \otimes A_2 & b_1 \otimes b_2 \\ N_1[Q_1 - L_1(A_1 \otimes b_2)] & N_1[Q_2 - L_1(b_1 \otimes A_2)] & N_1[b - L_1(b_1 \otimes b_2)] \\ 0 & 0 & 0 \end{bmatrix}. \end{aligned} \tag{76}$$

Thus the unreachable part N_2AL_3 has been separated, and it follows that:

$$\left(\begin{bmatrix} A_1 \otimes A_2 & 0 \\ N[C+AL_1-L_1(A_1\otimes A_2)] & N_1AL_2 \end{bmatrix},\right.$$
$$\left.\begin{bmatrix} A_1\otimes b_2 & b_1 \otimes A_2 & b_1\otimes b_2 \\ N_1[Q_1-L_1(A_1\otimes b_2)] & N_1[Q_2-L_1(b_1\otimes A_2)] & N_1[b-L_1(b_1\otimes b_2)]\end{bmatrix}\right)$$

is a reachable pair, representing a realization of the bilinear map f which is quasi-reachable by Theorem 3.1. This realization has the form (9)–(12) with (11) and (12) replaced respectively by

$$\begin{aligned}\hat{x}_{k+1} = N_1AL_2\hat{x}_k + N_1[C+AL_1-L_1(A_1\otimes A_2)]x_k^1\otimes x_k^2 \\ +N_1[Q_1-L_1(A_1\otimes b_2)]x_k^1v_k + N_1[Q_2-L_1(b_1\otimes A_2)]x_k^2u_k \\ +N_1[b-L_1(b_1\otimes b_2)]u_kv_k,\end{aligned} \tag{77}$$

$$y_k' = (d^T + h^TL_1)x_k^1\otimes x_k^2 + h^TL_1\hat{x}_k. \tag{78}$$

This corresponds to transforming the realization (9)–(12), according to the transformation described in Lemma 4.1 with $W=-L_1$, followed by the transformation of the basis for x_k by

$$\hat{x}_k = \begin{bmatrix} N_1 \\ N_2 \end{bmatrix} x_k$$

eliminating the component N_2x_k of $\hat{x}_k$.

5. Concluding remarks. The realization of nonlinear input/output maps will undoubtedly continue to attract research effort for some time to come. In this paper we have considered a simple bilinear form of input/output map and reached the conclusion that, somewhat unexpectedly, reachability of a realization of this map depends on the reachability of a corresponding *linear* system. This result can be extended also to the more complex multi-linear case [4]. A more detailed account of the work presented here, together with extension to multi-output bilinear systems and results on stability and observability of bilinear realizations can also be found in [4]. One of the major questions to be answered now is whether any other simple form of nonlinear input/output map can be approached in a similar way, yielding similar results and easily verifiable algebraic conditions for reachability, observability, etc. Such a class of systems may well be the discrete-time polynomial systems described in [2].

The constructional proof used here is long and tortuous and takes the transfer function approach, heavily influenced by the approach used in [3]. A more elegant and concise result might be obtainable if the module theoretic approach of [1] were to be developed with the additional insight now available into the mechanism by which nonreachable states occur.

Acknowledgments. The authors would like to thank Prof. R. E. Kalman for his suggestions on improving the original draft of this paper.

REFERENCES

[1] R. E. Kalman, *Pattern recognition properties of multi-linear machines*, IFAC Symposium on Technical and Biological Problems of Control (Yerevan, Armenian SSR), September, 1968.

[2] E. D. Sontag and Y. Rouchaleau, *On discrete-time polynomial systems*, J. Non-linear Analysis, Theory, Methods and Applications, 1 (1976), no. 1, pp. 55–64.

[3] E. FORNASINI AND G. MARCHESINI, *Algebraic realization theory of bilinear discrete-time input-output maps*, J. Franklin Institute, 301 (1976), pp. 143–160.

[4] J. G. PEARLMAN, *Internal description of multi-linear systems*, Ph.D. thesis, Imperial College, University of London, April 1977.

[5] M. A. ARBIB, *A characterization of multilinear systems*, IEEE Trans. Automatic Control, AC-14 (1969), pp. 699–702.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 4, July 1979

0363-0129/79/1704-0002 $01.00/0

DETERMINATION OF UNKNOWN FUNCTIONS FOR A CLASS OF DISTRIBUTED PARAMETER SYSTEMS*

TOSHIHIRO KOBAYASHI†

Abstract. The aim of this paper is to investigate the determination of unknown functions which give the input distributions for distributed parameter systems. After the problem formulation, the identifiability of input-distribution functions is discussed. The identifiability of the system does not necessarily assure that the problem of input-function determination is well posed. A well-posed approximation method is presented and discussed from the standpoint of regularization. The relation between the observability and the identifiability is also clarified for the system described by a linear parabolic partial differential equation.

1. Introduction. The construction of a mathematical model for a system using measurement data is a very important problem from both a theoretical and a practical point of view. Even for those systems for which the system equations have been postulated there still remains the equally important problem of identifying the initial state and some unknown functions from the measurement data. Kobayashi [1], [2] has studied the problem of initial-state determination for a class of distributed parameter systems. Since this problem is not well posed in general, he has presented a well posed approximation method by regularization.

We consider in this paper the problem of the determination of unknown functions which give the input distribution for a class of distributed parameter systems. We seek identifiability conditions which ensure that an unknown function can be uniquely determined from the measurement data. Since the space of unknown functions is an infinite dimensional one, identifiability does not generally assure that the function determined depends continuously on the measurement data. That is, the problem of unknown-function determination for a distributed parameter system is not necessarily well posed.

From the above facts, it is not sufficient to investigate only the identifiability of the distributed parameter systems when we consider the problem of the unknown-function determination. An approximation method is necessary which reduces the ill-posed problem to a well-posed one.

2. System description and problem statement. In this section, system description is following Lions [3]. So let H and V be two Hilbert spaces with

$$V \subset H, \quad V \text{ dense in } H; \tag{2.1}$$

the sign $\subset$ denotes both algebraic and topological inclusion. This means that the identity mapping of V in H is continuous. 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). Let V' be the dual of V; we identify H with its dual so that

$$V \subset H \subset V'. \tag{2.2}$$

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.

* Received by the editors October 18, 1977 and in revised form November 22, 1978.

† Department of Control Engineering, Kyushu Institute of Technology, Tobata, Kitakyushu, Japan.

We are given a continuous bilinear form $a(u, v)$ on V. For fixed u in V, the linear form

$$v \to a(u, v)$$

is continuous on V; therefore it can be written

$$a(u, v) = (Au, v), \qquad Au \in V'.$$

We deduce also that

$$\|Au\|_{V'} \leqq L\|u\|_V, \qquad u \in V,$$

where $\|\cdot\|_{V'}$ is the dual norm of $\|\cdot\|_V$. Suppose the family of operators $A \in \mathscr{L}(V; V')$ is coercive; that is

there exists β and $\alpha > 0$ such that

$$a(u, u) + \beta\|u\|_H^2 \geqq \alpha\|u\|_V^2, \qquad u \in V. \tag{2.3}$$

Consider now the distributed parameter system described by the following equation of evolution

$$\frac{du(t)}{dt} + Au(t) = gf(t), \qquad t \in (0, T), \tag{2.4}$$

and

$$u(0) = u_0, \qquad u_0 \text{ given in } H. \tag{2.5}$$

Here $u' = du/dt$ is taken in the sense of distribution on $(0, T)$. The function f belongs to the space $L^2(0, \infty)$. Moreover $g \in V'$ is the function to be determined.

We have the following existence and uniqueness lemma.

LEMMA 1 (Lions [3]). *Under the assumption* (2.3), *the system* (2.4) *and* (2.5) *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 the data* u_0, *f and g.*

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 $\int_0^T \|f(t)\|_F^2\, dt < \infty$.

From Lemma 1 the solution u of the system (2.4) and (2.5) is written

$$u(t) = U(t)u_0 + S(t)g, \qquad t \in (0, T), \tag{2.6}$$

where $U(t) \in \mathscr{L}(H; H)$, $S(t) \in \mathscr{L}(V'; H)$.

In physical situations, the space of observations K is finite dimensional. The outputs of the system are given by

$$z(t) = Mu(t) + m(t), \qquad t > 0, \tag{2.7}$$

where $M \in \mathscr{L}(H; K)$ and m is a measurement error such that $m \in L^2(0, T; K)$. By virtue of Lemma 1, we see that $z \in L^2(0, T; K)$.

We denote by $J(g)$ a functional which measures the distance between the observation z and the output Mu computed for each input distribution function g from the system (2.4) and (2.5). Then the problem of unknown function determination can be formulated as that of minimizing $J(g)$ with respect to g under the constraints (2.4) and (2.5). In this paper we take the functional $J(g)$ to be

$$J(g) = \int_0^T \|z(t) - Mu(t)\|_K^2\, dt. \tag{2.8}$$

3. Identifiability. In this section we investigate identifiability of the dynamical system described by (2.4) and (2.5) with the output equation (2.7).

We start with the following definition.

DEFINITION 1. The *system* described by (2.4) and (2.5) with the output equation (2.7) is said to be identifiable if the observation $Mu(t)$, $t>0$ implies the function g is unique. Moreover the system is said to be identifiable at time T if the observation $Mu(t)$ on $[0, T]$ implies the function g is unique.

From (2.6) we obtain

$$Mu(t)=MU(t)u_0+MS(t)g. \tag{3.1}$$

Since the initial state u_0 is known, the system (2.4), (2.5) and (2.7) is identifiable if and only if $MS(t)g=0$, $t>0$ implies $g=0$. The system is identifiable at time T if and only if $MS(t)g=0$ on $[0, T]$ implies $g=0$.

Let us consider the function $y(t)$ defined by

$$y(t)=Mu(t)-MU(t)u_0. \tag{3.2}$$

The problem of determining g from the observation $Mu(t)$ on $[0, T]$ is equivalent to the problem of determining g from $y(t)$ on $[0, T]$. The function y belongs to $L^2(0, T; K)$. Thus

$$\begin{aligned}\|y\|^2_{L^2(0,T;K)}&=\int_0^T \|y(t)\|^2_K\,dt\\ &=\int_0^T (MS(t)g, MS(t)g)_K\,dt \\ &=\int_0^T (S^*(t)M^*MS(t)g, g)\,dt\end{aligned} \tag{3.3}$$

Here $(\cdot)^*$ denotes the adjoint operator of an operator $(\cdot)$. Let us define an operator W: $V'\to L^2(0, T;K)$ by

$$Wg=MSg, \qquad g\in V'. \tag{3.4}$$

We get $W\in\mathscr{L}(V'; L^2(0, T; K))$. Then from (3.3) we obtain

$$\|y\|^2_{L^2(0,T;K)}=(W^*Wg, g). \tag{3.5}$$

This shows that the operator $W^*W\in\mathscr{L}(V'; V)$ is nonnegative. Thus the function g is uniquely determined if and only if the operator W^*W is positive, that is, for any $g\in V'$

$$(W^*Wg, g)\geqq 0 \quad\text{and}\quad (W^*Wg, g)=0 \quad\text{implies}\quad g=0. \tag{3.6}$$

Moreover W^*W is positive if and only if the nullspace of W is $\{0\}$. We have got the following theorem.

THEOREM 1. *The following three conditions are equivalent*:

(i) *the system* (2.4) *and* (2.5) *with* (2.7) *is identifiable at time* T;

(ii) W^*W *is positive*;

(iii) *the nullspace of* W *is* $\{0\}$.

4. Minimization of $J(g)$. We will show that a minimizing solution of $J(g)$ uniquely exists if the system (2.4) and (2.5) with (2.7) is identifiable at time T.

Since $J(g)$ is

$$J(g)=\int_0^T \|y_m(t)-MS(t)g\|^2_K\,dt, \qquad y_m(t)=z(t)-MU(t)u_0, \tag{4.1}$$

and the operators $S(t)$ and M are continuous, the functional $J(g)$ is differentiable and convex. Hence the necessary condition for optimality is

$$J'(g)\cdot h=0 \quad \text{for any } h\in V'. \tag{4.2}$$

In this case $J'(g)$ is explicitly calculated, and then (4.2) becomes

$$\int_0^T (MS(t)g-y_m(t), MS(t)h)_K\,dt=0, \qquad h\in V', \tag{4.3}$$

that is,

$$(W^*Wg-W^*y_m, h)=0, \qquad h\in V'. \tag{4.4}$$

Since (4.5) must hold for all $h\in V'$, the minimizing solution must satisfy

$$W^*Wg=W^*y_m. \tag{4.5}$$

If the system (2.4) and (2.5) with (2.7) is identifiable at time T, the optimal solution g_0 for $J(g)$ is uniquely determined by

$$g_0=(W^*W)^{-1}W^*y_m=G^{-1}W^*y_m. \tag{4.6}$$

However the inverse G^{-1} is not continuous in general. Then the solution g_0 does not necessarily depend continuously on the measurement data z.

Now we should consider a new approximation method which presents an approximate function for g_0 depending continuously on the measurement data.

5. A well-posed approximation method. In this section, by the method of regularization [5], we shall present the approximation method which gives constructively approximate functions for g_0 depending continuously on the measurement data. This approximation method corresponds to approximating the positive operator G by a family of positive definite ones.

We first introduce a regularized functional $J_\varepsilon(g)$ corresponding to $J(g)$:

$$J_\varepsilon(g)=J(g)+\varepsilon(\Phi g, g), \qquad \varepsilon>0, \tag{5.1}$$

where Φ is a continuous linear operator from V' to V such that

$$(\Phi h, h)\geqq \mu\|h\|^2_{V'}, \qquad h\in V', \qquad \mu>0. \tag{5.2}$$

We can see that for $J_\varepsilon(g)$ there exists a unique minimizing solution g_ε determined by

$$g_\varepsilon=(G+\varepsilon\Phi)^{-1}W^*y_m. \tag{5.3}$$

The operator $G_\varepsilon=G+\varepsilon\Phi$ is a continuous linear operator from V' to V and

$$(G_\varepsilon h, h)\geqq \varepsilon\|h\|^2_{V'}. \tag{5.4}$$

From this G_ε^{-1} is continuous. Therefore g_ε depends continuously on the measurement data z.

We now proceed to prove the following theorem.

THEOREM 2. *If the system* (2.4), (2.5) *and* (2.7) *is identifiable at time T, g_ε satisfies the convergence property*

$$\lim_{\varepsilon\to 0}\|g_\varepsilon-g_0\|_{V'}=0. \tag{5.5}$$

Proof. From (4.6) and (5.3) we obtain

$$(Gg_0-W^*y_m, h)=0, \qquad h\in V', \tag{5.6}$$

and

$$(Gg_\varepsilon - W^*y_m, h) + \varepsilon(\Phi g_\varepsilon, h) = 0, \qquad h \in V'. \tag{5.7}$$

Putting $h = g_0 - g_\varepsilon$ in (5.6), $h = g_\varepsilon - g_0$ in (5.7), and adding, respectively, both sides of two equations, we obtain

$$(G(g_\varepsilon - g_0), g_\varepsilon - g_0) + \varepsilon(\Phi g_\varepsilon, g_\varepsilon - g_0) = 0. \tag{5.8}$$

From this equation we have

$$(\Phi g_\varepsilon, g_\varepsilon) \leqq (\Phi g_\varepsilon, g_0), \tag{5.9}$$

since $\varepsilon > 0$. Using (5.2), we obtain

$$\|g_\varepsilon\|_{V'} \leqq \frac{1}{\mu}\|\Phi\| \cdot \|g_0\|_{V'} \tag{5.10}$$

Thus from every sequence of $\varepsilon \to 0$, we can extract a subsequence η such that $g_\eta \to w$ weakly in V'. As $\eta \to 0$, (5.7) becomes

$$(Gw, h) = (W^*y_m, h), \qquad h \in V'. \tag{5.11}$$

From (5.6) and (5.11), we get

$$(G(w - g_0), h) = 0, \qquad h \in V'. \tag{5.12}$$

Here putting $h = w - g_0$, we obtain

$$(G(w - g_0), w - g_0) = 0. \tag{5.13}$$

From the positiveness of G (the hypothesis of the system being identifiable at time T), we have $w = g_0$. Here $\{g_\eta\}$ is an arbitrary, weakly convergent subsequence and its weak limit g_0 does not depend on the subsequence. Thus the extraction of a subsequence is unnecessary and $g_\varepsilon \to g_0$ weakly in V'. Moreover from (5.9)

$$(\Phi(g_\varepsilon - g_0), g_\varepsilon - g_0) \leqq -(\Phi g_0, g_\varepsilon - g_0).$$

This implies that $g_\varepsilon \to g_0$ strongly in V'.

Now we notice that g_0 and g_ε are the minimizing solutions of $J(g)$ and $J_\varepsilon(g)$ respectively with the measurement error m. Thus g_0 is not an actual input-distribution function (denote it by g^*). We should evaluate $\|g_\varepsilon - g^*\|_{V'}$. We can obtain the next theorem.

THEOREM 3. *Suppose that the system* (2.4), (2.5) *and* (2.7) *is identifiable at time T. If the measurement error can be evaluated by*

$$\|m\|^2_{L^2(0,T;K)} \leqq \delta^2, \tag{5.14}$$

$$\lim_{\varepsilon,\delta \to 0} \|g_\varepsilon - g^*\|_{V'} = 0 \tag{5.15}$$

when $\delta/\sqrt{\varepsilon}$ goes to 0 *as* $\varepsilon \to 0$.

Proof. Define g^*_ε by

$$G_\varepsilon g^*_\varepsilon = W^*y. \tag{5.16}$$

Then

$$\|g_\varepsilon - g^*\| \leqq \|g_\varepsilon - g^*_\varepsilon\| + \|g^*_\varepsilon - g^*\|. \tag{5.17}$$

For the second term on the right-hand side, we can apply Theorem 2 in the case of $y_m = y$. As a result we obtain

$$\lim_{\varepsilon\to 0} \|g_\varepsilon^* - g^*\|_{V'} = 0. \tag{5.18}$$

Next, as for the first term, we have

$$(G_\varepsilon(g_\varepsilon - g_\varepsilon^*) - W^*(y_m - y), h) = 0, \qquad h \in V', \tag{5.19}$$

from (5.3) and (5.16). This implies that the element $(g_\varepsilon - g_\varepsilon^*)$ realizes the lower bound of the functional

$$I(g) = \int_0^T \|m(t) - MS(t)g\|_K^2\, dt + \varepsilon(\Phi g, g), \qquad \varepsilon > 0. \tag{5.20}$$

Thus

$$I(g_\varepsilon - g_\varepsilon^*) \leqq I(0) = \int_0^T \|m(t)\|_K^2\, dt \leqq \delta^2. \tag{5.21}$$

From this and (5.2), it follows that

$$\varepsilon\mu \|g_\varepsilon - g_\varepsilon^*\|_{V'}^2 \leqq \delta^2. \tag{5.22}$$

That is,

$$\|g_\varepsilon - g_\varepsilon^*\|_{V'}^2 \leqq \frac{\delta}{\sqrt{\mu\varepsilon}}. \tag{5.23}$$

We have proved the theorem.

6. The conditions for identifiability. In this section we seek concrete conditions of identifiability for the following system:

$$\frac{du(t)}{dt} + Au(t) = gf(t), \qquad u(0) = 0, \tag{6.1}$$

$$z_p(t) = (w_p, u(t))_H, \qquad p = 1, 2, \cdots, r, \qquad t > 0, \tag{6.2}$$

where w_p, $p = 1, 2, \cdots, r$ are given elements in H.

We assume

HYPOTHESIS 1([3], [4]). *The operator A is symmetric and the injection map of V into H is compact.*

Then there exists a sequence $\{\lambda_n, \phi_{nm}; m = 1, 2, \cdots, m_n, n = 1, 2, \cdots\}$ of eigenvalues and eigenvectors satisfying the following conditions:

(i) *For a constant C*

$$C \geqq \lambda_1 > \lambda_2 > \cdots > \lambda_n > \cdots, \qquad \mathit{lim}_{n\to\infty}\, \lambda_n = -\infty. \tag{6.3}$$

(ii) $\{\phi_{nm}; m = 1, 2, \cdots, m_n, n = 1, 2, \cdots\}$ *is a complete orthonormal basis in H, where the positive integers m_n are assumed to be finite for any n.*

(iii) *Each ϕ_{nm} satisfies*

$$A\phi_{nm} = -\lambda_n \phi_{nm}, \qquad \phi_{nm} \in V. \tag{6.4}$$

(iv) *For any $g \in V'$, the solution of* (6.1) *is given by*

$$u(t)=\int_0^t \sum_{n=1}^{\infty} e^{\lambda_n(t-\tau)} \sum_{m=1}^{m_n} g_{nm}\phi_{nm} f(\tau)\, d\tau, \tag{6.5}$$

where $g_{nm}=(g,\phi_{nm})$.

From (6.2) and (6.5) we obtain

$$z_p(t)=\int_0^t \sum_{n=1}^{\infty} e^{\lambda_n(t-\tau)} \sum_{m=1}^{m_n} w_{nm}^p g_{nm} f(\tau)\, d\tau, \tag{6.6}$$

where $w_{nm}^p=(w^p,\phi_{nm})_H$.

Defining $r\times m_n$ matrices W_n by

$$W_n=\begin{pmatrix} w_{n1}^1 & w_{n2}^1 & \cdots & w_{nm_n}^1 \\ w_{n1}^2 & w_{n2}^2 & \cdots & w_{nm_n}^2 \\ \cdot & \cdot & & \cdot \\ w_{n1}^r & w_{n2}^r & \cdots & w_{nm_n}^r \end{pmatrix}, \qquad n=1,2,\cdots, \tag{6.7}$$

we get the following theorem.

THEOREM 4. *The system* (6.1), (6.2) *is identifiable if and only if the following two conditions* (i) *and* (ii) *hold.*

(i) *rank $W_n=m_n\leqq r$ for all $n=1,2,\cdots$.*

(ii) *The input function $f(t)$ is not identically zero.*

Proof. To prove sufficiency, assume that

$$z_p(t)=\int_0^t \sum_{n=1}^{\infty} e^{\lambda_n(t-\tau)} \sum_{m=1}^{m_n} w_{nm}^p g_{nm} f(\tau)\, d\tau=0, \qquad t>0, \qquad p=1,2,\cdots,r. \tag{6.8}$$

Applying the Laplace transformation, we obtain

$$\sum_{n=1}^{\infty} \frac{F(s)}{s-\lambda_n}\left(\sum_{m=1}^{m_n} w_{nm}^p g_{nm}\right)=0, \qquad p=1,2,\cdots,r, \tag{6.9}$$

for any s such that Re $(s)>\max(\lambda_1,\xi)$, where ξ is the convergence coordinate of $F(s)$. Since $f(t)$ is not identically zero, $F(s)\not\equiv 0$. Moreover $F(s)$ is an analytic function on (ξ,∞). $F(s)\neq 0$ for almost every $s\in(\xi,\infty)$. Then we have

$$\sum_{n=1}^{\infty} \frac{1}{s-\lambda_n}\left(\sum_{m=1}^{m_n} w_{nm}^p g_{nm}\right)=0, \qquad p=1,2,\cdots,r, \tag{6.10}$$

for almost every s such that Re $(s)>\max(\lambda_1,\xi)$. By analytic continuation we see that (6.10) holds for all s such that $s\neq\lambda_n$, $n=1,2,\cdots$. Let C_n be a circle in the complex plane which includes only one pole $s=\lambda_n$. We obtain

$$0=\int_{C_n} \sum_{n=1}^{\infty} \frac{1}{s-\lambda_n}\left(\sum_{m=1}^{m_n} w_{nm}^p g_{nm}\right) ds=2\pi\sqrt{-1}\sum_{m=1}^{m_n} w_{nm}^p g_{nm}, \qquad p=1,2,\cdots,r, \tag{6.11}$$

for all $n=1,2,\cdots$. If rank $W_n=m_n\leqq r$, $n=1,2,\cdots$, then (6.11) implies that $g_{n1}=g_{n2}=\cdots=g_{nm_n}=0$ for all $n=1,2,\cdots$. Thus we have $g=0$ which implies that the system is identifiable.

To prove necessity, suppose that rank $W_n<m_n$ for some n. Then there exists a nonzero m_n-vector $\bar{g}_n=(g_{n1},g_{n2},\cdots,g_{nm_n})^T$ satisfying (6.11). This means the existence of a nonzero element g in V' which satisfies (6.8). Therefore the system is not identifiable. The necessity of the condition (ii) is evident.

Remark 2. In Theorem 4, (i) is the condition for the observability of the system. Therefore, if an input function $f(t)$ is not identically zero, the identifiability of the system (6.1), (6.2) is equivalent to the observability of the system.

Remark 3. If an input function $f(t)$ is an analytic function in $t \in (0, \infty)$,

$$z_p(t) = \int_0^t \sum_{n=1}^{\infty} e^{\lambda_n(t-\tau)} \sum_{m=1}^{m_n} w_{nm}^p g_{nm} f(\tau)\, d\tau \equiv 0$$

over an arbitrary interval $0 < t \leqq T$ implies that $z_p(t) \equiv 0$ for all $t > 0$. In other words, the system (6.1), (6.2) is identifiable at any time T if and only if it is identifiable.

REFERENCES

[1] T. KOBAYASHI, *Initial state determination for distributed parameter systems*, this Journal 14 (1976), pp. 934–944.

[2] ———, *A well-posed approximate method for initial state determination of discrete-time distributted parameter systems*, this Journal, 15 (1977), pp. 947–958.

[3] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.

[4] S. MIZOHATA, *The Theory of Partial Differential Equations*, Cambridge University Press, London, 1973.

[5] S. ROLEWICZ, *On optimal observability of linear systems with infinite-dimensional states*, Studia Math., 48 (1972), pp. 411–416.

[6] A. N. TIHONOV, *Solution of incorrectly formulated problems and the regularization method*, Dokl. Akad. Nauk SSSR, 151 (1963), pp. 1035–1038.

[7] S. KITAMURA AND S. NAKAGIRI, *Identifiability of spatially-varying and constant parameters in distributed systems of parabolic type*, this Journal. 15 (1977), pp. 785–802.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 4, July 1979

0363-0129/79/1704-0003 $01.00/0

ON CONSTRAINT DROPPING SCHEMES AND OPTIMALITY FUNCTIONS FOR A CLASS OF OUTER APPROXIMATIONS ALGORITHMS*

C. GONZAGA† AND E. POLAK‡

Abstract. This paper presents a new class of outer approximations algorithms which incorporate constraint dropping schemes. The algorithms are based on the use of certain types of optimality functions, which are commonly used in minimization algorithms, for defining stationary points. The algorithms are implementable in that all the inner minimizations and maximizations need to be carried out only approximately. It is shown that any accumulation point constructed by these algorithms is both feasible and stationary.

1. Introduction. After their introduction in 1960, by Cheney and Goldstein [1] and Kelley [2], in the form of cutting plane methods, and, in 1966, by Levitin and Polyak [3], who treated them in a more abstract setting, outer approximations algorithms went through a decade of stagnation. The reason for this was simple. These methods were intended to solve problems of the form

$$P\colon \min\{f(x)|x\in X\} \tag{1}$$

where $X\subset\mathbb{R}^n$ had a very complicated description, e.g., $X=\{x\,|\,\phi(x,\omega)\leqq 0,\ \omega\in\Omega\}$, with $\Omega\subset\mathbb{R}^m$ a set of infinite cardinality (i.e. X is defined by a continuum of inequalities). The approach was to substitute for P a sequence of approximating problems

$$P_k\colon \min\{f(x)\,|\,x\in X_k\},\qquad k=0,1,2,\cdots, \tag{2}$$

where $X\subset X_0\subset X_1\subset X_2\subset\cdots$ and the X_k had relatively simple descriptions, e.g. by a finite set of inequalities, $X_k=\{x\,|\,\phi(x,\omega)\leqq 0,\ \omega\in\Omega_k\subset\Omega\}$ with Ω_k a discrete set. Under certain rules defining the properties of the X_k, one could then show that the accumulation points of the sequence of solutions $\{x_k\}$, of the problems P_k, were solutions of P. Unfortunately, in all the specific schemes, the complexity of the description of the X_k (i.e. the number of inequalities involved) grew rapidly with k and quite quickly the problems P_k became almost as difficult as the original problem P.

The first breakthrough came when Topkis [4], [5] and Eaves and Zangwill [6] proposed constraint dropping schemes which broke the monotonic growth of the descriptions of the X_k. The Eaves and Zangwill theory in terms of cut set maps is particularly elegant. An interesting further generalization was given by Hogan [7]. Although from a theoretical point of view the work in [4], [6], [7] was of great importance, it still had several drawbacks from a practical point of view. These are easiest to explain in the Eaves–Zangwill framework, using a simple problem, e.g.

$$P\colon \min\{f(x)\,|\,\phi(x,\omega)\leqq 0,\ \omega\in\Omega\}, \tag{3}$$

where f and ϕ are both differentiable and $x\in\mathbb{R}^n$, $\Omega\subset\mathbb{R}^m$. The Eaves–Zangwill theory requires that we solve, exactly, two problems at each iteration.

$$P_k\colon \min\{f(x)\,|\,\phi(x,\omega)\leqq 0,\ \omega\in\Omega_k\}, \tag{4}$$

* Received by the editors December 1, 1977 and in revised form November 14, 1978. This work was supported in part by the Joint Services Electronics Program F44620-76-C-0100 and the National Science Foundation (RANN) under Grant ENU 76-04264.

† Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, California 94720. On leave from COPPE-UFRJ, Rio de Janeiro, Brazil.

‡ Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, California 94720.

where Ω_k is a discrete subset of Ω, to obtain a solution x_k and then

$$(5) \qquad \max\{\phi(x_k, \omega) \mid \omega \in \Omega\}$$

to produce a point ω_k. Assuming that $\phi(x_k, \omega_k) > 0$, we see that x_k is not optimal for P. The point ω_k is then added to Ω_k to form Ω_{k+1}. In a constraint dropping scheme, some of the points ω_j, acquired earlier, may be dropped from Ω_{k+1}. Now, in the absence of convexity, and since only a finite number of iterations of a program for solving (4) and (5) can be used, the best one can hope to achieve is to find an approximation to a *stationary* point for P_k (rather than to a solution x_k) and, perhaps, an approximation to ω_k. The Eaves–Zangwill theory does not apply to this situation. Moreover, their constraint dropping scheme (i.e., the dropping of points from Ω_k) depends on the rate of growth of the cost sequence $\{f(x_k)\}$ relative to the constraint violation sequence $\{\phi(x_k, \omega_k)\}$. As a result, unless a problem is extremely well-scaled, their constraint dropping scheme may fail to operate. The third objection to the early constraint dropping schemes is that when constraint dropping is in operation, only the subsequence of $\{x_k\}$ at which constraints were dropped can be shown to have accumulation points which are solutions of P. The first two of the above described drawbacks were overcome to some degree by Mayne, Polak and Trahan [8], in the framework of an algorithm for computer-aided-design.

In this paper we shall present a new class of implementable[1] outer approximations algorithms aimed at problems with constraints of the special form appearing in (3). This type of constraint commonly occurs in engineering design problems. For example, suppose that a design specification is stated as an inequality on the *nominal* design: $h(x) \leqq 0$. However, when the device is built, it can only be realized with a certain tolerance $\tau \in T$ (e.g. ±10%) and hence the design constraint becomes $h(x+\tau) \leqq 0 \ \forall \tau \in T$. Or suppose that a device is required to function over a range of temperatures or altitudes; then, again we get design specifications in the form (5). The algorithms which we shall present (i) utilize approximate solutions to the subproblems (4) and (5), (ii) eliminate dependence on scaling, and (iii) have much better convergence properties in the presence of constraint dropping than any earlier scheme. (Any accumulation point of $\{x_k\}$ is a solution to P rather than any accumulation point of the *subsequence* $\{x_{k_i}\}$ at which constraints were dropped, as in [6], [8].) Each algorithm in our class is completely characterized by three features: (i) the scheme for constructing the constraint sets Ω_k in (4), (ii) the algorithm for solving (4), and (iii) the algorithm for solving (5).

We present two constraint construction schemes, two master algorithms, two compatible characterizations of the algorithms for solving (4), and one requirement on the algorithms for solving (5). The characterization of the algorithms for solving (4) is in terms of the optimality functions on which they are based, while the characterization of the algorithms for solving (5) is in terms of the convergence property that they must have.

In our first constraint construction scheme, the user specifies a sequence $\{\varepsilon_k\}_{k=0}^{\infty}$, $\varepsilon_k > 0$, $\varepsilon_k \to 0$ (e.g. $\varepsilon_k = \varepsilon_0/(k+1)^{1/10}$, with ε_0 large), selects initial Ω_0, and then sets $\Omega_{k+1} = \{\omega_k\} \cup \Omega_k$ if $\phi(x_k, \omega_k) > \varepsilon_k$ and $\Omega_{k+1} = \Omega_k$ otherwise.

In the second scheme, the user specifies a double subscripted sequence ε_{kj}, (e.g. $\varepsilon_{kj} = \varepsilon_j - \varepsilon_k$) such that $\varepsilon_{kj} \nearrow \varepsilon_j$ as $k \to \infty$ and retains $\omega_j (j \leqq k)$ in Ω_{k+1} only as long as $\phi(x_j, \omega_j) > \varepsilon_{kj}$ holds.

[1] By *implementable* we mean that all the required computations in each iteration or stage can be carried out by means of a finite number of operations on a digital computer.

The first scheme is intended for use in an interactive computing factility for very difficult problems such as those found in computer aided design, where function evaluations take *minutes* because of the need to integrate large systems of nonlinear differential equations. In such a situation, the designer can keep up and successfully interact with the computing process. He chooses a sequence $\{\varepsilon_k\}_{k=0}^{\infty}$ which forces very few ω_k to be included in Ω_{k+1}. He monitors the outcome of the computation on a graphics display terminal and adds a certain number of points to Ω_{k+1} on the basis of this observation. He usually adds ω_k to Ω_{k+1}. He may subsequently remove from Ω_j, $j > k+1$, the points which he added to Ω_{k+1} on the basis of judgement without affecting the theoretical convergence of the algorithm. For example, suppose he monitors a step response for its peak overshoot. Suppose that the peak occurs at time $t_1 \in \Omega \subset \mathbb{R}$ ($\omega = t$ in this case), that there are no other local peaks, that the cardinality of Ω_k is small and that $\phi(x_k, t_1) < \varepsilon_k$. He will most probably choose to add t_1 to Ω_{k+1} even though he is not required to do so by the algorithm. Next, suppose this peak occurs at t_1, but there are also large local peaks at t_2 and t_3. Since when the peak at t_1 is pushed down the ones at t_2, t_3 may come up, he may decide to include t_2 and t_3 as well as t_1 in Ω_{k+1}.

The second scheme is intended for more automatic use. It retains points ω_j in $\Omega_k (k > j)$ for a certain number of subproblems and then drops them automatically when, presumably, they are no longer relevant. This scheme is a generalization of the one in [8]. Our computational experience with the scheme in [8] was that even in this case, a certain amount of discretionary intervention (to augment Ω_k) can be highly beneficial.

We characterize the algorithms for solving (4) by the optimality function[2] on which they are based (e.g., as in methods of feasible directions or of penalty functions). We state two assumptions which are sufficient conditions for such an optimality function to be useable in our algorithms and we show that a number of optimality functions, associated with Phase I–Phase II feasible directions algorithms or penalty function methods, satisfy our assumptions.

The notation in this treatment tends to become a little complex and we have collected all of our symbols in the Appendix, for the reader's convenience.

As was the case in [8], this work was strongly motivated by our experience in optimization-based computer-aided-design of shock resistant structures, control systems and electronic circuits. Given that more and more interactive computing facilities are being set up, and that there are no competing alternatives in the literature, we expect our algorithms to have an important technological impact.

2. New classes of outer approximations algorithms. The algorithms which we shall present are intended for the solution of problems of the form

$$(6) \qquad P_\Omega\colon \min \{f(x) \mid g^j(x) \leqq 0, j \in \mathbf{l};\ \phi^k(x, \omega^k) \leqq 0, \omega^k \in \Omega^k, k \in \mathbf{m}\}$$

where $\mathbf{l} \triangleq \{1, 2, \cdots, l\}$, $\mathbf{m} \triangleq \{1, 2, \cdots, m\}$, the functions $f(\cdot)$, $g^j(\cdot)$ and $\phi^k(\cdot,\cdot)$ are continuously differentiable[3] on $\mathbb{R}^n$ and on $\mathbb{R}^n \times \mathbb{R}^{p_k}$, respectively, and Ω^k is a *compact* subset of $\mathbb{R}^{p_k}$, $k \in \mathbf{m}$. The symbol Ω is used to denote $\Omega^1 \times \Omega^2 \times \cdots \times \Omega^m$. The problem form (7) is particularly important because many engineering design problems can be transcribed into it (see [8] for control examples).

[2] We say that $\theta\colon \mathbb{R}^n \to \mathbb{R}^1$ is an optimality function if $\theta(x) = 0$ for all x solving P and $\theta(x) \leqq 0$ for all $x \in \mathbb{R}^n$.

[3] Differentiability in ω is not required by our proofs, but is stipulated as an assumption which is usually required by algorithms which compute approximate solutions to $\max_{\omega^k} \{\phi^k(x, \omega^k) \mid \omega^k \in \Omega^k\}$.

We shall approximate the problem P_Ω by a sequence of simpler problems of the form (with $i = 1, 2, 3, \cdots$)

$$(7) \qquad P_{\Omega_i}\colon \min\{f(x) \mid g^j(x) \leqq 0, j \in \mathbf{l};\ \phi^k(x, \omega^k) \leqq 0, \omega^k \in \Omega_i^k, k \in \mathbf{m}\},$$

where $\Omega_i^k \subset \Omega^k$ are discrete sets. Our aim is to approximate feasible stationary points of P_Ω, i.e., points $\hat{x} \in \mathbb{R}^n$ such that

$$(8) \qquad g^j(\hat{x}) \leqq 0, \quad \forall j \in \mathbf{l}; \qquad \max_{\omega \in \Omega} \phi^k(\hat{x}, \omega) \leqq 0, \quad \forall k \in \mathbf{m},$$

and

$$(9) \qquad \theta_\Omega(\hat{x}) \triangleq \min_{\|h\|_\infty \leqq 1} \max\{\langle \nabla f(\hat{x}), h\rangle;\ g^j(\hat{x}) + \langle \nabla g^j(\hat{x}), h\rangle,$$
$$j \in \mathbf{l};\ \phi^k(\hat{x}, \omega^k) + \langle \nabla_x \phi^k(\hat{x}, \omega^k), h\rangle, \omega^k \in \Omega^k, k \in \mathbf{m}\} = 0.$$

We recognize (9) as the Topkis–Veinott [9] multiplier free form of the F. John condition for P_Ω (see p. 8 and p. 182 in [10]).

DEFINITION 1. We shall say that *point* $\hat{x} \in \mathbb{R}^n$ is desirable if (9) and (10) are satisfied at $\hat{x}$. We shall denote the set of all desirable points in $\mathbb{R}^n$ by Δ. □

We assume that we can "solve" the problems P_{Ω_i} approximately, to the extent of finding a point x_i for which the value of an appropriate optimality function $\theta^i_{\Omega_i}(x_i)$ is near zero. The superscript is introduced to allow for the possible use of penalty functions. The theory we are about to present is based on our knowledge of Phase I–Phase II type methods of feasible directions [11], [12] and penalty function methods [10], all of which utilize real valued optimality functions $\theta^i_{\Omega'}(\cdot)$, defined on $\mathbb{R}^n$ for discrete subsets $\Omega'^k \subset \Omega^k$, $k \in \mathbf{m}$, and all positive integers i. All of these functions have the property that $\theta^i_{\Omega'}(x) \leqq 0$ for all $x \in \mathbb{R}^n$ and that if x' is optimal for $P_{\Omega'}$, then $\lim_{i\to\infty} \theta^i_{\Omega'}(x') = 0$. Some of these optimality functions are continuous while others are not. Early examples of such optimality functions can be found in [10, p. 182]. Not all the existing optimality functions can be used in our outer approximations algorithms. Only the ones satisfying Assumptions 1 and 2 below are acceptable within the framework of our convergence theorems. We need the following definition. For any compact subset $\Omega' \in \Omega$, $\psi_{\Omega'}\colon \mathbb{R}^n \to \mathbb{R}^1$ is defined by

$$(10) \qquad \psi_{\Omega'}(x) \triangleq \max\{0;\ g^j(x), j \in \mathbf{l};\ \phi^k(x, \omega^k), \omega^k \in \Omega'^k, k \in \mathbf{m}\}.$$

ASSUMPTION 1. *Consider the family of optimality functions* $\{\theta^i_{\Omega'}(\cdot)\}$, *with* Ω' *a discrete subset of* Ω *and* i *a positive integer. For all* $x \in \mathbb{R}^n$, $x \notin \Delta$, *there exist* $\mu > 0$, $\rho > 0$, $N \geqq 0$ *and* $\delta \in (0, 1)$ (*possibly depending on* x), *such that for all* $x' \in B(x, \rho) \triangleq \{x' \mid \|x - x'\| \leqq \rho\}$ *and all discrete subsets* $\Omega'^k \subset \Omega^k$, $k = 1, 2, \cdots, m$, *satisfying* $\psi_{\Omega'}(x) \geqq \delta\psi_\Omega(x)$, *we have*

$$(11) \qquad \theta^i_{\Omega'}(x') \leqq -\mu \quad \textit{for all } i \geqq N.$$

To obtain an intuitive understanding for the reason for Assumption 1, consider a point $x \notin \Delta$. Let $\Omega_i \subset \Omega$, $i = 0, 1, 2, \cdots$, be a sequence of discrete sets such that the sets $\{x' \mid \psi_{\Omega_i}(x') \leqq 0\}$ are good approximations to the set $\{x' \mid \psi_\Omega(x') \leqq 0\}$, at x, in the sense that $\psi_{\Omega_i}(x) \geqq \delta\psi_\Omega(x)$, with $\delta \in (0, 1)$, and let x_i, $i = 0, 1, 2, \cdots$, be the corresponding stationary point of P_{Ω_i}. Then $\theta^i_{\Omega_i}(x_i) = 0$ for all i and hence the x_i cannot converge to x, since this would violate (11). Thus, provided we construct the P_{Ω_i} so that the sets $\{x' \mid \psi_{\Omega_i}(x') \leqq 0\}$ become good local approximations to the sets $\{x' \mid \psi_\Omega(x') \leqq 0\}$ at any limit point $\hat{x}$ of a solution sequence $\{x_i\}$, then $\hat{x}$ will have to be in Δ.

We shall later devote a separate section to showing that a number of common optimality functions satisfy Assumption 1. In the present section we shall present two schemes for constructing the Ω_i so that the desired convergence properties are ensured.

Our outer approximations algorithms based on Assumption 1 are of the form of the model below. They differ from one another only by the manner in which the discrete sets Ω_i^k, $k = 1, 2, \cdots, n$, are constructed. They all require that we have an algorithm for solving the problem P_{Ω_i}, with Ω_i a discrete set, and another one for approximating the values of the functions $\bar{\psi}^k_{\Omega'^k}: \mathbb{R}^n \to \mathbb{R}^1$ defined by

$$\bar{\psi}^k_{\Omega'^k}(x) \triangleq \max_{\omega^k \in \Omega'^k} \{\phi^k(x, \omega^k), k \in \mathbf{m}\}, \tag{12}$$

with $\Omega'^k \subseteq \Omega^k$. To complete our notation, we define $\bar{\psi}^0$: $\mathbb{R} \to \mathbb{R}^1$ by

$$\bar{\psi}^0(x) \triangleq \max \{g^1(x), g^2(x), \cdots, g^l(x)\}. \tag{13a}$$

MASTER ALGORITHM MODEL 1.

Parameters. An infinite sequence $\{\beta_i\}_{i=1}^{\infty}, \beta_i > 0, \beta_i \to 0.$[4] ($\theta^i_{\Omega'}(\cdot)$ is a family of optimality functions.)

Data. Discrete sets $\Omega_0^k \subset \Omega^k$, $k \in \mathbf{m}$.

Step 0. Set $i = 0$.

Step 1. Construct the discrete sets Ω_i^k, $k \in \mathbf{m}$.

Step 2. Compute an x_i such that

$$\theta^i_{\Omega_i}(x_i) \geqq -\beta_i. \tag{13b}$$

Step 3. Set $i = i + 1$ and go to Step 1.

Since our algorithms cannot be stated without ambiguity unless all the details of the problems (6) and (7) are preserved, we shall maintain these details in the algorithm statements. However, as far as the proofs are concerned, there is no great loss of generality, and a great gain in notational simplification, when one restricts oneself to the simplest case of (6) and (7), viz: $\mathbf{l}$ = empty set, $\mathbf{m} = \{1\}$. In that case the superscripts become redundant and we get

$$P_\Omega: \min \{f(x) \mid \phi(x, \omega) \leqq 0, \omega \in \Omega\}, \tag{14a}$$

$$P_{\Omega_i}: \min \{f(x) \mid \phi(x, \omega) \leqq 0, \omega \in \Omega_i\}, \tag{14b}$$

$$\psi_{\Omega'}(x) \triangleq \max \{0; \phi(x, \omega), \omega \in \Omega'\}, \qquad \Omega' \subseteq \Omega, \tag{14c}$$

$$\bar{\psi}_{\Omega'}(x) \triangleq \max_{\omega \in \Omega'} \phi(x, \omega), \qquad \Omega' \subseteq \Omega. \tag{14d}$$

We shall need the following result.

PROPOSITION 1. *Consider the simplified problem* (14a). *Let* $\{x_i\}_{i=0}^{\infty}$ *be any converging sequence in* $\mathbb{R}^n$ *with* $x_i \to \hat{x}$ *as* $i \to \infty$ *and let* $\Omega_i \subset \Omega$, $i = 1, 2, \cdots$, *be any sequence of compact sets contained in* Ω. *Then*

$$|\bar{\psi}_{\Omega_i}(x_i) - \bar{\psi}_{\Omega_i}(\hat{x})| \to 0 \quad \text{as } i \to \infty. \tag{15}$$

Proof. Suppose that (15) does not hold. Then there exists a $\hat{\delta} > 0$ and an infinite subsequence indexed by $K \subset \{1, 2, \cdots\}$ such that

$$|\bar{\psi}_{\Omega_i}(x_i) - \bar{\psi}_{\Omega_i}(\hat{x})| \geqq \hat{\delta} \quad \text{for all } i \in K. \tag{16}$$

Now, $\bar{\psi}_{\Omega_i}(x_i) = \phi(x_i, \omega_i)$ and $\bar{\psi}_{\Omega_i}(\hat{x}) = \phi(\hat{x}, \hat{\omega}_i)$ for some ω_i, $\hat{\omega}_i$ in Ω_i. Without loss of

[4] For example, $\beta_i = \varepsilon\delta^i$, for $\delta \in (0, 1)$, or $\beta_i = \varepsilon / i$.

generality, we may assume, therefore, that

$$\phi(x_i, \omega_i) \geqq \phi(\hat{x}, \hat{\omega}_i) + \hat{\delta} \quad \text{for all } i \in K. \tag{17}$$

Now, since $x_i \to \hat{x}$ and $\Omega_i \subset \Omega$, with Ω a compact set, there exists an i_0 such that

$$\phi(\hat{x}, \omega_i) \geqq \phi(x_i, \omega_i) - \hat{\delta}/2 \quad \text{for all } i \in K, \quad i \geqq i_0. \tag{18}$$

But (17) and (18) show that $\hat{\omega}_i$ is not a maximizer of $\phi(\hat{x}, \omega)$ over Ω_i, which is a contradiction. Hence the proposition is true. □

COROLLARY. *Consider the simplified problem* (14a). *The functions* $\psi_\Omega(\cdot)$ *are uniformly continuous in* $\Omega' \subset \Omega$, i.e., *for any* $\delta > 0$ *and* $x \in \mathbb{R}^n$ *there exists* $\varepsilon > 0$ *such that for any compact* $\Omega' \subset \Omega$, $|\psi_{\Omega'}(x) - \psi_{\Omega'}(x')| < \delta$ *whenever* $\|x - x'\| < \varepsilon$.

Proof. If the corollary were false, then there would exist $\hat{x} \in \mathbb{R}^n$, $\hat{\delta} > 0$, and sequences $\Omega_i \subset \Omega$ and $x_i \to \hat{x}$ such that $|\psi_{\Omega_i}(\hat{x}) - \psi_{\Omega_i}(x_i)| \geqq \hat{\delta}$. But this would contradict Proposition 1, so the corollary must be true. □

Our two constraint construction schemes define the operations to be performed in step 1 of the master algorithm model 1.

Constraint construction scheme 1: General case.

a) Specify a positive, decreasing sequence $\{\varepsilon_i\}_{i=0}^{\infty}$, with $\varepsilon_i \to 0$ as $i \to \infty$.

b) Given i, x_i,

(i) compute $\omega_i^k \in \Omega^k$, for all $k \in \mathbf{m}$, by approximately evaluating $\max_{\omega^k \in \Omega^k} \phi^k(x_i, \omega^k)$;

(ii) set

$$\psi_\Omega^i(x_i) \triangleq \max\{0, \bar{\psi}^0(x_i); \phi^k(x_i, \omega_i^k), k \in \mathbf{m}\}; \tag{19}$$

(iii) if $\psi_\Omega^i(x_i) > \varepsilon_i$, include ω_i^k in Ω_j^k for all $j > i$, and all $k \in \mathbf{m}$ such that

$$\psi_\Omega^i(x_i) = \phi^k(x_i, \omega_i^k). \quad □ \tag{20}$$

In the case of Problem (14a) we get simplifications which we now state.

Constraint construction scheme 1 for problem (14a).

a) Specify a positive, decreasing sequence $\{\varepsilon_i\}_{i=0}^{\infty}$, with $\varepsilon_i \to 0$ as $i \to \infty$.

b) Given i, x_i,

(i) compute $\omega_i \in \Omega$ by approximately evaluating $\max_{\omega \in \Omega} \phi(x_i, \omega)$;

(ii) set

$$\psi_\Omega^i(x_i) = \phi(x_i, \omega_i); \tag{20a}$$

(iii) if $\psi_\Omega^i(x_i) > \varepsilon_i$, include ω_i in Ω_j for all $j > i$. □

In its most economical form, the constraint construction scheme 1 only requires that given the sets Ω_i^k, $k = 1, 2, \cdots, m$, $\Omega_{i+1}^k = \{\omega_i^k\} \cup \Omega_i^k$ if $\phi^k(x_i, \omega_i^k) = \psi_\Omega^i(x_i)$ and $\psi_\Omega^i(x_i) > \varepsilon_i$, and $\Omega_{i+1}^k = \Omega_i^k$ otherwise; i.e., the approximating constraint sets Ω_i^k are augmented only when the corresponding functional inequalities ($\max_{\omega \in \Omega^k} \phi^k(x, \omega) \leqq 0$) have been sufficiently violated.

Apart from this specified restriction, the construction of the Ω_i^k is arbitrary in the sense that any other points $\omega^k \subset \Omega^k$, not specifically covered by the scheme, can be added to (or subtracted from) the sets Ω_i^k.

Assuming that our algorithm is being used in an interactive computing facility with a graphics display terminal, we expect the user to proceed as follows. He will choose a very slowly decaying sequence $\{\varepsilon_i\}_{i=0}^{\infty}$ to make the test $\psi_\Omega^i(x_i) \leqq \varepsilon_i$ easy to satisfy. For example, he may set

$$\varepsilon_i = K\psi_\Omega^0(x_0)/(i+1)^{1/L} \tag{21}$$

with $K \geqq 100, L \geqq 2$. This forces very few ω_i^k to be added to Ω_i^k in forming Ω_{i+1}^k. However, so as to ensure better computational behavior, we expect him to add ω_i^k to Ω_{i+1}^k if a constraint violation occurred at ω_i^k (i.e. $\phi^k(x_i, \omega_k^k) > 0$) and to keep it in $\Omega_j^k, j \geqq i+1$, for a few more problems P_j on a discretionary basis, after which he will probably find it reasonable to drop it. He may do this, without violating the hypotheses for convergence in Theorem 1, below.

Returning again to the special case of problem (14a), we find that here we compute only one ω_i and hence

$$\psi_\Omega^i(x_i) = \max\{0, \phi(x_i, \omega_i)\}. \tag{22}$$

THEOREM 1. *Consider a sequence* $\{x_i\}_{i=0}^\infty$ *constructed by the master algorithm model* 1, *using the constraint construction scheme* 1. *Suppose that* (i) $|\psi_{\Omega_i}^i(x_i) - \psi_{\Omega_i}(x_i)| \to 0$ *as* $i \to \infty$ *and* (ii) *the optimality functions* $\theta_{\Omega'}^i(\cdot)$ *used in the master algorithm model* 1 *satisfy Assumption* 1. *Then any accumulation point of* $\{x_i\}_{i=0}^\infty$ *is in* Δ.

Proof. Without essential loss of generality, it suffices to prove the theorem for the special case of problem (14a). To obtain a contradiction, suppose that $x_i \xrightarrow{K} \hat{x}$, where $K \subset \{1, 2, 3, \cdots\}$, and $\hat{x} \notin \Delta$. We consider the various possibilities.

(i) Suppose that $\psi_\Omega(\hat{x}) = 0$, i.e. $\hat{x}$ is feasible for P_Ω. Let $\hat{\delta} \in (0, 1)$, $\hat{\rho} > 0$, $\hat{N} > 0$ and $\hat{\mu} > 0$ be as specified in Assumption 1 for $\hat{x}$. Then, since $\psi_{\Omega_i}(x) \geqq 0$ for all x and any $\Omega_i \subset \Omega$,

$$\psi_{\Omega_i}(\hat{x}) \geqq \hat{\delta}\psi_\Omega(\hat{x}) \quad \text{for all } i. \tag{23}$$

Consequently, there exists an integer $i_0 \geqq \hat{N}$ such that $x_i \in B(\hat{x}, \hat{\rho})$ for $i \geqq i_0$, $i \in K$ and

$$\theta_{\Omega_i}^i(x_i) < -\hat{\mu} \leqq -\beta_i \quad \text{for all } i \geqq i_0, \quad i \in K, \tag{24}$$

which contradicts the construction in step 2 of master algorithm model 1.

(ii) Suppose that $\psi_\Omega(\hat{x}) > 0$. Then, since $\varepsilon_i \to 0$ as $i \to \infty$, since $\psi_\Omega(x_i) \xrightarrow{K} \psi_\Omega(\hat{x}) > 0$ by continuity, and since $|\psi_\Omega^i(x_i) - \psi_\Omega(x_i)| \to 0$ as $i \to \infty$, there exists an i_0 such that $\varepsilon_i < \psi_\Omega^i(x_i) = \phi(x_i, \omega_i)$ for all $i \in K$, $i \leqq i_0$. Furthermore

$$\phi(x_i, \omega_i) \xrightarrow{K} \psi_\Omega(\hat{x}) \quad \text{as } i \to \infty. \tag{25}$$

Since Ω is compact, we must have

$$|\phi(x_j, \omega_i) - \phi(x_i, \omega_i)| \to 0 \quad \text{as } i \to \infty, \quad i, j \in K, \quad j > i. \tag{26}$$

Now, $\omega_i \in \Omega_j$ for all $i \in K, j > i \geqq i_0$, by construction. Hence we obtain that

$$\psi_\Omega(x_j) \geqq \psi_{\Omega_j}(x_j) \geqq \phi(x_j, \omega_i), \tag{27}$$

for all $i, j \in K, j > i \geqq i_0$. Taking (25) and (26) into account, we conclude that

$$\psi_{\Omega_i}(x_i) \xrightarrow{K} \psi_\Omega(\hat{x}) \quad \text{as } i \to \infty. \tag{28}$$

Making use of Proposition 1, we now conclude that $\psi_{\Omega_i}(\hat{x}) \xrightarrow{K} \psi_\Omega(\hat{x})$ as $i \to \infty$. Let $\hat{\delta} \in (0, 1)$, $\hat{N} > 0$, $\hat{\mu} > 0$ be as specified in Assumption 1. Because $\psi_{\Omega_i}(\hat{x}) \xrightarrow{K} \psi_\Omega(\hat{x})$, there exists an $i_1 \geqq \hat{N}$ such that for $i \in K$, $i \geqq i_1$,

$$\psi_{\Omega_i}(\hat{x}) \geqq \hat{\delta}\psi_\Omega(\hat{x}) \tag{29a}$$

and

(29b) $$\beta_i < \hat{\mu},$$

and therefore

(29c) $$\theta^i_{\Omega_i}(x_i) \leqq -\hat{\mu} < -\beta_i \quad \text{for all } i \in K, \quad i \geqq i_1,$$

which contradicts the construction in step 2 of the algorithm model.

Since (i) and (ii) are the only two possibilities, we conclude that the theorem holds. □

Our second constraint construction scheme is a generalization of the ones proposed by Eaves and Zangwill [6] and by Mayne, Trahan and Polak [8]. Like those schemes, it will retain a particular constraint for a certain number of approximating problems, and then drop it. However, it utilizes somewhat more information than the schemes in [6] and [8] and therefore leads to a more interesting convergence theorem. On the basis of our experience with the algorithm in [8], we predict that the constraint dropping scheme below, when used in an interactive computing facility, is bound to result in much better computational behavior than the earlier outer approximations algorithms.

Constraint construction scheme 2: General case.

a) Specify a double indexed sequence $\{\varepsilon_{ij}\}_{i=0,j\leqq i}^{\infty}$ such that

(i) $\varepsilon_{ii} = 0$, $\varepsilon_{ij} > 0$ for all $i, j < i$;

(ii) $\varepsilon_{ij} \to \bar{\varepsilon}_j$ as $i \to \infty$, uniformly in j;

(iii) $\bar{\varepsilon}_j > \varepsilon_{ij}$ for $i \geqq j$, and $\bar{\varepsilon}_j \to 0$ as $j \to \infty$. (For example, $\varepsilon_{ij} = \delta^j - \delta^i$, with $\delta \in (0, 1)$, or $\varepsilon_{ij} = \bar{\varepsilon}_j - \bar{\varepsilon}_i$, where $\bar{\varepsilon}_i \downarrow 0$.)

b) Given x_i,

(i) compute $\omega_i^k \in \Omega^k$, $k = 1, 2, \cdots, m$, by approximately evaluating $\max_{\omega^k \in \Omega^k} \phi^k(x_i, \omega^k)$;

(ii) set $\psi^i_\Omega(x_i)$ as in (19);

(iii) for all $j \in \{1, 2, \cdots, i\}$ such that $\psi^i_\Omega(x_j) > \varepsilon_{ij}$, include ω_j^k in Ω^k_{i+1}, for all $k \in \mathbf{m}$ such that $\psi^i_\Omega(x_j) = \phi^k(x_j, \omega_j^k)$. □

Again, for problem (14a) we get a simplification:

Constraint construction scheme 2 for problem (14a).

a) Specify a double indexed sequence $\{\varepsilon_{ij}\}_{i=0,j\leqq i}^{\infty}$ such that

(i) $\varepsilon_{ii} = 0$, $\varepsilon_{ij} > 0$ for all $i, j < i$;

(ii) $\varepsilon_{ij} \to \bar{\varepsilon}_j$ as $i \to \infty$, uniformly in j;

(iii) $\varepsilon_{ij} < \bar{\varepsilon}_j$ $\forall i \geqq j$ and $\bar{\varepsilon}_j \to 0$ as $j \to \infty$.

b) Given x_i,

(i) compute $\omega_i \in \Omega$ by approximately evaluating $\max_{\omega\in\Omega} \phi(x_i, \omega)$ and set $\psi^i_\Omega(x_i) = \phi(x_i, \omega_i)$;

(ii) for all $j \in \{0, 1, 2, \cdots, i\}$ such that $\psi^i_\Omega(x_j) > \varepsilon_{ij}$, include ω_j in Ω_{i+1}. □

Note that once the test $\psi^i_\Omega(x_j) \leqq \varepsilon_{ij}$ is satisfied, the pair $(\psi^i_\Omega(x_j), \omega_j)$ need no longer be stored.

For a comparison with the Eaves–Zangwill scheme [6], we set $\varepsilon_{ij} \triangleq f(x_i) - f(x_j)$, $i \geqq j$. Their rule is to store only the last $\psi_\Omega(x_j)$, ω_j^k and $f(x_j)$ at which constraints were dropped and to include all ω_l^k in Ω_l^k, $j \leqq l < i$ for all $k \in \mathbf{m}$ satisfying $\psi_\Omega(x_l) = \phi^k(x_l, \omega_l^k)$, whenever $\psi_\Omega(x_j) > \varepsilon_{ij}$. The Mayne–Polak–Trahan scheme [8] is similar to the Eaves–Zangwill one, except that it sets

$$\varepsilon_{ij} = \frac{f(x_i) - f(x_j) + \mu\beta^i}{\tau(1 - \beta^i)}$$

where $\beta \in (0, 1)$ and $\tau > 0$, $\mu > 0$. Thus, the schemes in [6] and in [8] slowly accumulate constraints (i.e. ω_j^k), then drop them *en masse*, then accumulate them again. This type of oscillatory behavior results in poor computational properties. Also, since only one $\psi_\Omega(x_j)$ is utilized at any time, convergence properties in [6], [8] can only be established for the subsequence at which constraints were dropped, rather than for the whole sequence. Our constraint construction scheme 2 was evolved to avoid the type of oscillatory behavior mentioned above and, in addition, to enable the establishment of convergence properties for the entire sequence $\{x_i\}$. It shares with the schemes in [6], [8] the property that it retains a certain ω_j^k in Ω_i^k until $i-j$ has become sufficiently large for $\psi_\Omega^i(x_j) \leqq \varepsilon_{ij}$ to take place, and then drops it.

THEOREM 2. *Consider a sequence $\{x_i\}_{i=1}^\infty$ constructed by master algorithm model* 1, *using the constraint construction scheme* 2. *Suppose that*

(i) $|\psi^i_{\Omega_i}(x_i) - \psi_\Omega(x_i)| \to 0$ *as* $i \to \infty$,

(ii) *the optimality functions $\theta^i_{\Omega'}(\cdot)$ used in the algorithm model* 1 *satisfy Assumption* 1.

Then any accumulation point of the sequence $\{x_i\}_{i=1}^\infty$ is in Δ.

Proof. We note that since $\varepsilon_{ij} < \bar{\varepsilon}_j$ for all $i \geqq j$ and all j, when scheme 2 is used, a point ω_j^k satisfying $\psi^j_\Omega(x_j) = \phi^k(x_i, \omega_j^k)$ is always included in all Ω_i^k, $i > j$, whenever $\psi^j_\Omega(x_j) > \bar{\varepsilon}_j$. Since the $\bar{\varepsilon}_j$ satisfy the properties of the $\{\varepsilon_i\}$ specified in scheme 1, Theorem 2 follows directly from Theorem 1. □

It may sometimes be difficult to show that an optimality function $\theta^i_{\Omega'}(\cdot)$ satisfies Assumption 1. In that case one can make use of Assumption 2, below. It is satisfied by the optimality functions used in [8].

ASSUMPTION 2. *Consider the family of optimality functions $\{\theta^i_{\Omega_i}(\cdot)\}$, where the Ω_i are discrete subsets of Ω. If $\{x_i\}_{i=1}^\infty$ is a sequence in $\mathbb{R}^n$ such that $x_i \to \hat{x}$, with $\psi_\Omega(\hat{x}) = 0$, and $\theta^i_{\Omega_i}(x_i) \to 0$ as $i \to \infty$, then $\hat{x} \in \Delta$.*

When Assumption 2 is in force, we use a different algorithm model.

MASTER ALGORITHM MODEL 2.

Parameters. An infinite sequence $\{\beta_i\}_{i=0}^\infty$, $\beta_i > 0$, $\beta_i \to 0$.

Data. Discrete sets Ω_0^k, $k \in \mathbf{m}$ contained in Ω.

Step 0. Set $i = 0$.

Step 1. Construct the discrete sets Ω_i^k, $k \in \mathbf{m}$.

Step 2. Compute an x_i such that

$$-\beta_i \leqq \theta_{\Omega_i}(x_i) \leqq 0 \text{ and } \psi_{\Omega_i}(x_i) \leqq \beta_i. \tag{30}$$

Step 3. Set $i = i+1$ and go to Step 1. □

THEOREM 3. *Let $\{x_i\}_{i=0}^\infty$ be a sequence constructed by master algorithm model* 2, *using the constraint construction scheme* 1 *or* 2. *Suppose that*

(i) $|\psi^i_{\Omega_i}(x_i) - \psi_\Omega(x_i)| \to 0$ *as* $i \to \infty$;

(ii) *the optimality functions $\theta^i_{\Omega^i}(\cdot)$ used in the master algorithm model* 2 *satisfy Assumption* 2.

Then any accumulation point of the sequence $\{x_i\}_{i=1}^\infty$ is in Δ.

Proof. We only need to prove this theorem for the case where the constraint construction scheme 1 is used, since scheme 2 is a special case of it. Without essential loss of generality, we restrict ourselves to the special case of problem (14a).

Thus, suppose that $x_i \xrightarrow{K} \hat{x}$, with $K \subset \{1, 2, 3, \cdots\}$. First, suppose that $\psi_\Omega(\hat{x}) = 0$. Since $\theta^i_{\Omega_i}(x_i) \to 0$ as $i \to \infty$ by construction, it follows from Assumption 2 that $\hat{x} \in \Delta$. Hence we only need to show that assuming $\psi_\Omega(\hat{x}) > 0$ leads to a contradiction.

Therefore, suppose that $\psi_\Omega(\hat{x}) > 0$. Then, since $\psi_\Omega(x_i) \xrightarrow{K} \psi_\Omega(\hat{x})$ by continuity, and $|\psi^i_\Omega(x_i) - \psi_\Omega(x_i)| \to 0$ as $i \to \infty$ by assumption, we must have $\psi^i_\Omega(x_i) \xrightarrow{K} \psi_\Omega(\hat{x})$. Therefore, since $\varepsilon_i \to 0$ as $i \to \infty$, there exists an integer $i_0 > 0$ such that $\psi^i_\Omega(x_i) = \phi(x_i, \omega_i) > \varepsilon_i$ for all $i \geqq i_0$, $i \in K$, and thus, $\omega_i \in \Omega_j$ for all $j > i \geqq i_0$, $i \in K$. Consequently,

$$\psi_{\Omega_j}(x_j) \geqq \phi(x_j, \omega_i) \quad \text{for all } j > i \geqq i_0, \quad i \in K. \tag{31}$$

Now, because ϕ is continuous and Ω is compact,

$$|\phi(x_j, \omega_i) - \phi(x_i, \omega_i)| \to 0 \tag{32}$$

as $i, j \to \infty$, $i, j \in K$, $j > i$. Since $\psi^i_\Omega(x_i) = \phi(x_i, \omega_i) \xrightarrow{K} \psi_\Omega(\hat{x})$ as $i \to \infty$, we obtain from (32) that

$$\phi(x_j, \omega_i) \to \psi_\Omega(\hat{x}) \tag{33}$$

as $i, j \to \infty$, $i, j \in K$, $j > i$. Hence, because of (31) and (33), there exists an $i_1 \geqq i_0$ such that

$$\psi_{\Omega_i}(x_i) \geqq \psi_\Omega(\hat{x})/2 > \beta_i \quad \text{for all } i \geqq i_1, \quad i \in K. \tag{34}$$

But this contradicts (30) and hence we are done. □

3. Optimality functions for outer approximations algorithms. We shall now present a few optimality functions which satisfy Assumptions 1 and 2. To avoid even more subscripts or superscripts, we shall denote them all by the same symbol $\theta^i_\Omega(\cdot)$ or $\theta_\Omega(\cdot)$. They will be treated one at a time and so no confusion should arise among them. First we show that any family of optimality functions satisfying Assumption 1 must also satisfy Assumption 2.

PROPOSITION 2. *Suppose* $\{\theta^i_{\Omega_i}(\cdot)\}$, $\Omega_i \subset \Omega$, *is a sequence of optimality functions satisfying Assumption* 1. *Then it also satisfies Assumption* 2.

Proof. Suppose that $x_i \to \hat{x}$ as $i \to \infty$, with $\psi_\Omega(\hat{x}) = 0$, and that $\theta^i_{\Omega_i}(x_i) \to 0$ as $i \to \infty$. Then $\hat{x} \in \Delta$, for otherwise, by Assumption 1 (since $\psi_{\Omega_i}(\hat{x}) \geqq \hat{\delta}\psi_\Omega(\hat{x})$ for any $\hat{\delta} \in (0, 1)$ and all i) there exists a $\hat{\mu} > 0$ and an i_0 such that $\theta^i_{\Omega_i}(x_i) \leqq -\hat{\mu}$ for all $i \geqq i_0$, which contradicts $\theta^i_\Omega(x_i) \to 0$ as $i \to \infty$. □

The first two optimality functions that we consider are independent of the superscript i and hence we shall drop it for these cases. These optimality functions are normally used in methods of feasible directions (see [10], [11], [12]) for computing descent directions. Since these optimality functions satisfy Assumption 1, we conclude that methods of feasible directions based on these optimality functions are suitable for solving problems $P_{\Omega'}$ in a scheme based on either master algorithm model 1 or on master algorithm model 2.

Consider the functions, with $\Omega'^k \subseteq \Omega^k$, compact, introduced in [13] by Pironneau and Polak,

$$\begin{aligned} \theta_{\Omega'}(x) \triangleq \min_h \{\tfrac{1}{2}\|h\|^2 + \max\{&\langle \nabla f(x), h\rangle;\ g^j(x) + \langle \nabla g^j(x), h\rangle,\ j \in \mathbf{l}; \\ &\phi^k(x, \omega^k) + \langle \nabla_x \phi^k(x, \omega^k), h\rangle,\ \omega^k \in \Omega'^k,\ k \in \mathbf{m}\}\} - \psi_{\Omega'}(x). \end{aligned} \tag{35}$$

Since (35) is an extremely messy expression, we shall show (without much loss of generality) that it satisfies Assumption 1 by considering only the special case where $m = 1$ and $l = 0$, i.e., problem (14a). In this case superscripts can be dropped, and (35) simplifies to

$$\theta_{\Omega'}(x) = \min_h \{\tfrac{1}{2}\|h\|^2 + \max\{\langle \nabla f(x), h\rangle;\ \phi(x, \omega) + \langle \nabla_x \phi(x, \omega), h\rangle,\ \omega \in \Omega'\}\} - \psi_{\Omega'}(x). \tag{36}$$

ASSUMPTION 3. *For every* $x \in \mathbb{R}^n$, $0 \notin \text{co}_{\omega \in \Omega_0(x)} \nabla_x \phi(x, \omega)$, *where*

$$\Omega_0(x) \triangleq \{\omega \in \Omega \mid \phi(x, \omega) = \psi_\Omega(x)\} \tag{37}$$

and co *denotes the convex hull of the set in question.*

This assumption states that the gradients of the constraint function at the most violated points are positive linearly independent. We recognize it as a sufficient condition for the Kuhn–Tucker constraint qualification to be satisfied at every $x \in \mathbb{R}^n$.

THEOREM 4. *Suppose that Assumption* 3 *is satisfied. Then the family of optimality functions defined by* (35) *satisfies Assumption* 1.

Proof. We shall only give a proof for the special case (36). It is quite easy to see that $\Delta = \{x \mid \theta_\Omega(x) = 0, \psi_\Omega(x) = 0\}$. Since by assumption $0 \notin \text{co}_{\omega \in \Omega_0(x)} \nabla_x \phi(x, \omega)$ for all $x \in \mathbb{R}^n$, and $\text{co}_{\omega \in \Omega_0(x)} \nabla_x \phi(x, \omega)$ is closed, there exists an $\bar{h} \in \mathbb{R}^n$ such that $\langle \nabla_x \phi(x, \omega), \bar{h} \rangle < 0$ for all $\omega \in \Omega_0(x)$. Hence it is easy to see that $\theta_\Omega(x) < 0$ for all $x \in \mathbb{R}^n$ such that $\psi_\Omega(x) > 0$ and therefore

$$\Delta = \{x \mid \theta_\Omega(x) = 0\}. \tag{38}$$

Now, suppose that $\hat{x} \notin \Delta$; therefore $\theta_\Omega(\hat{x}) < 0$. Then, for any $x \in \mathbb{R}^n$, $\Omega' \subset \Omega$, compact, we get from (36)

$$\theta_{\Omega'}(x) \leqq \theta_\Omega(x) + [\psi_\Omega(x) - \psi_{\Omega'}(x)]. \tag{39}$$

Since $\theta_\Omega(\cdot)$ is continuous and $\psi_{\Omega'}(\cdot)$ is continuous uniformly in Ω' (by the corollary to Proposition 1), there exist $\hat{\delta} \in (0, 1)$ and $\hat{\rho} > 0$ such that

$$\hat{\delta}\psi_\Omega(\hat{x}) \geqq \psi_\Omega(\hat{x}) + \tfrac{1}{4}\theta_\Omega(\hat{x}), \tag{40}$$

$$\theta_\Omega(x) \leqq \tfrac{1}{2}\theta_\Omega(\hat{x}) \quad \text{for all } x \in B(\hat{x}, \hat{\rho}), \tag{41}$$

and for any $\Omega' \subset \Omega$ such that $\psi_{\Omega'}(\hat{x}) \geqq \hat{\delta}\psi_\Omega(\hat{x}) \geqq \psi_\Omega(\hat{x}) + \frac{1}{4}\theta_\Omega(\hat{x})$,

$$\psi_{\Omega'}(x) \geqq \psi_\Omega(x) + \tfrac{1}{2}\theta_\Omega(x) \quad \text{for all } x \in B(\hat{x}, \hat{\rho}). \tag{42}$$

Hence, from (39) and (42), for all $x \in B(\hat{x}, \hat{\rho})$,

$$\theta_{\Omega'}(x) \leqq \theta_\Omega(x) - \tfrac{1}{2}\theta_\Omega(x) \leqq \tfrac{1}{4}\theta_\Omega(\hat{x}) \triangleq -\hat{\mu}, \tag{43}$$

which completes our proof. □

Next, continuing in the simplified framework of the problem P_Ω: $\min \{f(x) \mid \phi(x, \omega) \leqq 0, \omega \in \Omega\}$, which results in no essential loss of generality, we define a new optimality function, which we obtain from the test in Polak's method of feasible directions [10, (p. 164)] as follows. For any $\varepsilon \geqq 0$, $\Omega' \subseteq \Omega$ compact, and $x \in \mathbb{R}^n$, let the set of "ε-active" $\omega \in \Omega$ be defined by

$$\Omega'_\varepsilon(x) \triangleq \{\omega \in \Omega' \mid \phi(x, \omega) \geqq \psi_{\Omega'}(x) - \varepsilon\} \tag{44}$$

and let

$$\gamma^\varepsilon_{\Omega'}(x) \triangleq \min_{\|h\|_\infty \leqq 1} \max \{\langle \nabla f(x), h \rangle - \psi_{\Omega'}(x); \langle \nabla_x \phi(x, \omega), h \rangle, \omega \in \Omega'_\varepsilon(x)\}. \tag{45}$$

This function is used to find a descent direction in one of the Phase I–Phase II algorithms in [11]. Unfortunately, it is not continuous, which prompts the development, below. Let $\beta \in (0, 1)$, $\rho > 0$ be given.

$$\theta_{\Omega'}(x) \triangleq \min \{-\varepsilon \mid \gamma^\varepsilon_\Omega(x) \leqq -\varepsilon, \varepsilon \in \{0\} \cup \{\beta^k \rho \mid k = 0, 1, 2, 3, \cdots\}\}. \tag{46}$$

It is easy to show (by extension of the results in § 4.4 of [10]) that an equivalent characterization of Δ (see Definition 1) is

$$\Delta = \{x \in \mathbb{R}^n \mid \gamma^0_\Omega(x) = 0,\ \psi_\Omega(x) = 0\}. \tag{47}$$

Since $\gamma^\varepsilon_\Omega(x) \geqq \gamma^0_\Omega(x)$ for all $\varepsilon \geqq 0$ and $0 \geqq \gamma^\varepsilon_\Omega(x)$ always holds, we must have also that

$$\Delta = \{x \in \mathbb{R}^n \mid \theta_\Omega(x) = 0,\ \psi_\Omega(x) = 0\}. \tag{48}$$

LEMMA 1. *Suppose that Assumption 3 is satisfied and let $\Omega' \subseteq \Omega$ be any compact set. Under these assumptions, we have the following*:

a) *If $\hat{x}$ is optimal for $P_{\Omega'}$ then $\theta_{\Omega'}(\hat{x}) = 0$. Furthermore,*

$$.\Delta = \{x \in \mathbb{R}^n \mid \theta_\Omega(x) = 0\}. \tag{49}$$

b) *For any $\hat{x} \in \mathbb{R}^n$ such that $\theta_{\Omega'}(\hat{x}) < 0$, there exist $\hat{\rho} < 0$ and $\hat{\varepsilon} > 0$ such that $\theta_{\Omega'}(x) \leqq -\hat{\varepsilon}$ for all $x \in B(\hat{x}, \hat{\rho})$.*

c) *If $\hat{x}$ is such that $\theta_{\Omega'}(\hat{x}) = 0$ and Ω' is discrete, then $\theta_{\Omega'}(\cdot)$ is continuous at $\hat{x}$.*

Proof. a) Referring to [10, p. 181], we see that if $\hat{x}$ is optimal for $P_{\Omega'}$ then $\gamma^0_{\Omega'}(\hat{x}) = 0$. Hence, since $\gamma^0_{\Omega'}(\hat{x}) \leqq \gamma^\varepsilon_{\Omega'}(\hat{x})$ for all $\varepsilon \geqq 0$, it follows from (46) that $\theta_{\Omega'}(\hat{x}) = 0$. The fact that (49) holds follows from (48) and Assumption 3, which guarantees that $\gamma^0_\Omega(x) < 0$ for all x such that $\psi_\Omega(x) > 0$.

b) Suppose that $\theta_{\Omega'}(\hat{x}) < 0$ (i.e. $\gamma^0_{\Omega'}(\hat{x}) < 0$). Our first observation is that the map $(x, \varepsilon) \rightarrow \Omega'_\varepsilon(x)$ is upper semi-continuous, i.e., given $(\hat{x}, 0)$ and $\hat{\delta} > 0$, there exist $\hat{\varepsilon}_0 > 0$ and $\hat{\rho}_0 > 0$ such that

$$\Omega'_\varepsilon(x) \subset N_{\hat{\delta}}(\hat{x}) \quad \text{for all } \varepsilon \in [0, \hat{\varepsilon}_0], \quad x \in B(\hat{x}, \hat{\rho}_0), \tag{50}$$

where $N_{\hat{\delta}}$ is a neighborhood of $\Omega'_0(\hat{x})$ defined by

$$N_{\hat{\delta}} \triangleq \{\omega \in \mathbb{R}^{p} \mid \|\omega - \omega'\| \leqq \hat{\delta}, \text{ for some } \omega' \in \Omega'_0(\hat{x})\}. \tag{51}$$

Let $\hat{\delta} > 0$ be such that for

$$\bar{\gamma}^{\hat{\delta}}_{\Omega'}(x) \triangleq \min_{\|h\|_\infty \leqq 1} \max \{\langle \nabla f(x), h\rangle - \psi_{\Omega'}(x);\ \langle \nabla_x \phi(x, \omega), h\rangle,\ \omega \in N_{\hat{\delta}}\} \tag{52}$$

we have

$$\bar{\gamma}^{\hat{\delta}}_\Omega(\hat{x}) \leqq \gamma^0_{\Omega'}(\hat{x})/2. \tag{53}$$

Note that $\bar{\gamma}^{\hat{\delta}}_{\Omega'}(\cdot)$ is a continuous function.

Now, let $\hat{\varepsilon}_0 > 0, \hat{\rho}_0 > 0$ be such that (50) holds and let $\hat{\rho} \in (0, \hat{\rho}_0]$ be such that $\bar{\gamma}^{\hat{\delta}}_{\Omega'}(x) \leqq \gamma^0_{\Omega'}(\hat{x})/4$ for all $x \in B(\hat{x}, \hat{\rho})$. Then, for all $\varepsilon \in [0, \hat{\varepsilon}_0]$ and for all $x \in B(\hat{x}, \hat{\rho})$,

$$\gamma^\varepsilon_{\Omega'}(x) \leqq \bar{\gamma}^{\hat{\delta}}_{\Omega'}(x) \leqq \gamma^0_{\Omega'}(\hat{x})/4$$

where the first inequality holds because $\Omega_\varepsilon(x) \subset N_{\hat{\delta}}(\hat{x})$. Let $\hat{k} \geqq 0$ be any integer such that $\gamma^0_{\Omega'}(\hat{x})/4 \leqq -\beta^{\hat{k}}\rho \triangleq -\hat{\varepsilon}$ and $\hat{\varepsilon} \in (0, \hat{\varepsilon}_0]$. Then for all $x \in B(\hat{x}, \hat{\rho})$,

$$\gamma^{\hat{\varepsilon}}_{\Omega'}(x) \leqq -\hat{\varepsilon} \tag{54}$$

and therefore, by definition, $\theta_{\Omega'}(x) \leqq -\hat{\varepsilon}$ for all $x \in B(\hat{x}, \hat{\rho})$.

c) Now suppose that $\theta_{\Omega'}(\hat{x}) = 0$, and, for the sake of contradiction, suppose that $\theta_{\Omega'}(\hat{x})$ is not continuous at $\hat{x}$. Then there exists a sequence $x_i \rightarrow \hat{x}$ as $i \rightarrow \infty$ and a $\delta = \beta^k\rho > 0$ such that

$$\theta_{\Omega'}(x_i) \leqq -\delta < 0 \quad \text{for all } i. \tag{55}$$

Since Ω' is discrete, there exists a $\hat{\rho} > 0$ such that $\Omega'_{\delta/2}(x) \supset \Omega'_0(\hat{x})$ for all $x \in B(\hat{x}, \hat{\rho})$. Hence, by continuity of $\bar{\gamma}^0_{\Omega'}(\cdot)(\hat{\delta} = 0$ in (52)), and because $\bar{\gamma}^0_{\Omega'}(\hat{x}) = \gamma^0_{\Omega'}(\hat{x}) = 0$, there exists an $i_0 \geqq 0$ such that

$$-\beta\delta < \bar{\gamma}^{0}_{\Omega'}(x_i) \leqq \gamma^{\beta\delta}_{\Omega'}(x_i) \quad \text{for all } i \geqq i_0. \tag{56}$$

But this implies that $\theta_{\Omega'}(x_i) \geqq -\beta\delta$ for all $i \geqq i_0$ which contradicts (55). This completes our proof. □

THEOREM 5. *Suppose that Assumption* 3 *is satisfied, then the optimality functions* $\theta_{\Omega'}(\cdot)$ *defined by* (46) *satisfy Assumption* 1.

Proof. Suppose $\hat{x} \notin \Delta$. Then, by Lemma 1, $\theta_{\Omega}(\hat{x}) < 0$ and there exist $\rho_1 > 0$ and $\hat{\varepsilon} > 0$ such that

$$\theta_{\Omega}(x) \leqq -\hat{\varepsilon} \quad \text{for all } x \in B(\hat{x}, \rho_1). \tag{57}$$

Let $\hat{\delta} \in (0, 1)$ be such that

$$\hat{\delta}\psi_{\Omega}(\hat{x}) \geqq \psi_{\Omega}(\hat{x}) - \hat{\varepsilon}/4. \tag{58}$$

Since Ω is compact, $\phi(\cdot, \omega)$ is continuous, uniformly in $\omega \in \Omega$, and hence there exists a $\hat{\rho} \in (0, \rho_1]$ such that for any discrete subset $\Omega' \subset \Omega$ and for all $x \in B(\hat{x}, \hat{\rho})$

$$\psi_{\Omega'}(x) - \psi_{\Omega'}(\hat{x}) \geqq \phi(x, \hat{\omega}) - \phi(\hat{x}, \hat{\omega}) \geqq -\hat{\varepsilon}/8, \tag{59}$$

where $\hat{\omega} \in \arg\max_{\omega \in \Omega'} \phi(\hat{x}, \omega)$, and also (by continuity of $\psi_{\Omega}(\cdot)$)

$$\psi_{\Omega}(x) \leqq \psi_{\Omega}(\hat{x}) + \hat{\varepsilon}/8. \tag{60}$$

Now suppose that $\Omega' \subset \Omega$ is a finite set satisfying

$$\psi_{\Omega'}(\hat{x}) \geqq \hat{\delta}\psi_{\Omega}(\hat{x}) \geqq \psi_{\Omega}(\hat{x}) - \hat{\varepsilon}/4, \tag{61}$$

and suppose that $x \in B(\hat{x}, \hat{\rho})$. Then, making use of (59), (60) and (61), we obtain

$$\psi_{\Omega'}(x) \geqq \psi_{\Omega'}(\hat{x}) - \hat{\varepsilon}/8 \geqq \psi_{\Omega}(\hat{x}) - \hat{\varepsilon}/8 - \hat{\varepsilon}/4 \geqq \psi_{\Omega}(x) - \hat{\varepsilon}/2. \tag{62}$$

Therefore, since $\Omega'_{\hat{\varepsilon}/2}(x) \subset \Omega'_{\hat{\varepsilon}}(x) \subset \Omega_{\hat{\varepsilon}}(x)$ always holds, for all $x \in B(\hat{x}, \hat{\rho})$ we obtain

$$\begin{aligned} \gamma^{\hat{\varepsilon}/2}_{\Omega'}(x) &= \min_{\|h\|_\infty \leqq 1} \max \{\langle \nabla f(x), h\rangle - \psi_{\Omega'}(x); \langle \nabla_x \phi(x, \omega), h\rangle, \omega \in \Omega'_{\hat{\varepsilon}/2}(x)\} \\ &\leqq \min_{\|h\|_\infty \leqq 1} \max \{\langle \nabla f(x), h\rangle - \psi_{\Omega}(x) + \hat{\varepsilon}/2; \hat{\varepsilon}/2 + \langle \nabla_x \phi(x, \omega), h\rangle, \omega \in \Omega_{\hat{\varepsilon}}(x)\} \\ &= \gamma^{\hat{\varepsilon}}_{\Omega}(x) + \hat{\varepsilon}/2 \leqq -\hat{\varepsilon}/2. \end{aligned} \tag{63}$$

The last inequality follows from (57) and (46) because, with $\varepsilon(x) \triangleq -\theta_{\Omega}(x)$, for all $x \in B(\hat{x}, \hat{\rho})$,

$$\gamma^{\varepsilon(x)}_{\Omega}(x) \leqq -\varepsilon(x) \leqq -\hat{\varepsilon}, \tag{64}$$

which implies that $\hat{\varepsilon} \leqq \varepsilon(x)$ for all $x \in B(\hat{x}, \hat{\rho})$ and therefore,

$$\gamma^{\hat{\varepsilon}}_{\Omega}(x) \leqq \gamma^{\varepsilon(x)}_{\Omega}(x) \leqq -\hat{\varepsilon} \quad \text{for all } x \in B(\hat{x}, \hat{\rho}). \tag{65}$$

It now follows from (46) and (63) that

$$\theta_{\Omega'}(x) \leqq -\hat{\varepsilon}/2 \triangleq -\hat{\mu} < 0, \quad \text{for all } x \in B(\hat{x}, \hat{\rho}), \tag{66}$$

which completes our proof. □

To conclude this section, we show that penalty function algorithms can also be used for solving the problems $P_{\Omega'}$ in our outer approximations methods. We continue to restrict ourselves to the special case where P_{Ω} is $\min\{f(x) \mid \phi(x, \omega) \leqq 0, \omega \in \Omega\}$, since there is no essential loss of generality in doing so, but the notational simplification is great.

Let $\{s_i\}_{i=1}^{\infty}$ be an infinite sequence such that $s_i > 0$ and $s_i \to 0$ as $i \to \infty$ (e.g. $s_i = s_0/i$, or $s_i = \beta^i$, with $\beta \in (0, 1)$), and let $\Omega' \subset \Omega$ be any discrete set with cardinality $\nu_{\Omega'}$. Then we define $p_{\Omega'}: \mathbb{R}^n \to \mathbb{R}^1$ and, for $i = 1, 2, 3, \cdots, f^i_{\Omega'}: \mathbb{R}^n \to \mathbb{R}^1$ by

$$p_{\Omega'}(x) \triangleq \frac{1}{\nu_{\Omega'}} \sum_{\omega \in \Omega'} [\max\{0, \phi(x, \omega)\}]^2 \tag{67}$$

and

$$f^i_{\Omega'}(x) \triangleq f(x) + \frac{1}{s_i} p_{\Omega'}(x). \tag{68}$$

Next, we define the optimality functions $\theta^i_{\Omega'}(\cdot)$ by

$$\theta^i_{\Omega'}(x) \triangleq -\|\nabla f^i_{\Omega'}(x)\|, \qquad \Omega' \subset \Omega, \quad i = 1, 2, 3, \cdots, \tag{69}$$

with Ω' always a discrete subset of Ω. A standard assumption in penalty function methods is that for any x such that $\psi_{\Omega'}(x) > 0, \nabla p_{\Omega'}(x) \neq 0$ or the somewhat stronger assumption that $0 \notin \operatorname{co}_{\omega \in \Omega'(x)_+} \nabla\phi(x, \omega)$, for all x such that $\bar{\psi}_{\Omega'}(x) \geqq 0$, where

$$\Omega'(x)_+ \triangleq \{\omega \in \Omega' \mid \phi(x, \omega) \geqq 0\} \tag{69a}$$

and co denotes the convex hull of the set specified. When extended to the problem P_Ω, the latter assumption becomes $0 \notin \operatorname{co}_{\omega \in \Omega(x)_+} \nabla\phi(x, \omega)$ for all x such that $\bar{\psi}_\Omega(x) \leqq 0$. This, in turn, leads to the following strengthened assumptions which we shall need to show that the optimality functions (69) satisfy Assumption 1.

ASSUMPTION 4. (i) *For all* $x \in \mathbb{R}^n$ *such that* $\bar{\psi}_\Omega(x) \geqq 0$, $0 \notin \operatorname{co}_{\omega \in \Omega(x)_+} \nabla\phi(x, \omega)$. (ii) *For every* $\varepsilon > 0$, *there exists an* $\eta > 0$ *such that for any* $x \in \mathbb{R}^n$ *and any* $\Omega' \subset \Omega$ *finite, if* $\psi_{\Omega'}(x) \geqq \varepsilon$, *then* $\|\nabla p_{\Omega'}(x)\| \geqq \eta$.

THEOREM 6. *Suppose Assumption* 4 *is satisfied. Then the family of optimality functions defined by* (69) *satisfies Assumption* 1.

Proof. Let $\hat{x} \in \mathbb{R}^n$ be such that $\hat{x} \notin \Delta$.

a) Suppose that $\psi_\Omega(\hat{x}) = 0$ and that the $\theta^i_{\Omega'}(\cdot)$ do not satisfy Assumption 1 at $\hat{x}$. Then, since $\psi_{\Omega'}(\hat{x}) \geqq \hat{\delta}\psi_\Omega(\hat{x})$ for any $\hat{\delta} > 0$ and compact $\Omega' \subset \Omega$, we can construct sequences $\{x_i\}_{i=1}^{\infty}, \{\Omega_i\}_{i=1}^{\infty}, \{\mu_i\}_{i=1}^{\infty}$, such that $x_i \to \hat{x}$ as $i \to \infty$, Ω_i are discrete subsets of Ω, $\mu_i > 0$, $\mu_i \to 0$ as $i \to \infty$, and

$$\theta^i_{\Omega_i}(x_i) = -\left\|\nabla f(x_i) + \frac{1}{\nu_i} \sum_{\omega \in \Omega_i} \frac{2}{s_i} \max\{0, \phi(x, \omega)\} \nabla_x \phi(x_i, \omega)\right\| \geqq -\mu_i \tag{70}$$

where $\nu_i = \nu_{\Omega_i}$.

If there exists an infinite subsequence $\{\Omega_i\}_{i \in K}, K \subset \{1, 2, \cdots\}$ such that either $\Omega_i = \phi$ or $\psi_{\Omega_i}(x_i) = 0$, for all $i \in K$, then (70) implies that $\nabla f(x_i) \xrightarrow{K} 0$ as $i \to \infty$ and hence that $\nabla f(\hat{x}) = 0$. But this is impossible since $\hat{x} \notin \Delta$. Hence no subsequence satisfying (70) can exist. Therefore, we assume that $\Omega_i \neq \phi$, $\psi_\Omega(x_i) \geqq \psi_{\Omega_i}(x_i) > 0$ for all i. Now let

$$\pi_i \triangleq \frac{1}{\nu_i} \sum_{\omega \in \Omega_i(x_i)_+} \frac{2}{s_i} \phi(x, \omega) \nabla_x \phi(x, \omega). \tag{71}$$

Then, since $\nabla f(x_i) \to \nabla f(\hat{x})$, as $i \to \infty$, by continuity, (70) implies that $\pi_i \to \nabla f(\hat{x})$ as $i \to \infty$, which shows that the $\|\pi_i\|$ are bounded. Next, since $0 \notin \operatorname{co}_{\omega \in \Omega(x)_+} \nabla_x \phi(x, \omega)$ for $x \in \{\hat{x}, x_1, x_2, \cdots\}$, it is easy to see from a compactness argument that

$$\inf \min \{\|y\| \mid y \in \operatorname*{co}_{\omega \in \Omega(x_i)_+} \nabla_x \phi(x_i, \omega)\} = d > 0. \tag{72}$$

Writing π_i as

$$(73)\qquad \pi_i=\frac{1}{s_i\nu_i}\Big(\sum_{\omega\in\Omega_i(x_i)_+}\phi(x_i,\omega)\Big)\Big(\sum_{\omega\in\Omega_i(x_i)_+}\lambda_i(\omega)\nabla_x\phi(x_i,\omega)\Big)$$

where $\lambda_i(\omega)=\phi(x_i,\omega)/\sum_{\omega\in\Omega_i(x_i)_+}\phi(x_i,\omega)$, we conclude, since $\|\sum_{\omega\in\Omega_+(x_i)_+}\lambda_i(\omega)\cdot\nabla_x\phi(x_i,\omega)\|\geqq d$ by (72), and since $\pi_i\to\nabla f(\hat{x})$, that the coefficients $(1/s_i\nu_i)\sum_{\omega\in\Omega_i(x_i)_+}\phi(x_i,\omega)$ are bounded from above for all i. Consequently, there exists a bound M, such that, because of Carathéodory's theorem [14],

$$(74)\qquad \pi_i=\sum_{k=1}^{p+1}\lambda_i^k\nabla_x\phi(x_i,\omega_i^k),\qquad i=1,2,3,\cdots,$$

with $\omega_k^i\in\Omega_i(x_i)_+$ and $0\leqq\lambda_i^k\leqq M$. Since Ω is compact, there must exist an infinite subset $K'\subset\{1,2,3,\cdots\}$ such that $\omega_i^k\xrightarrow{K'}\hat{\omega}^k$, $k=1,2,\cdots,p+1$, as $i\to\infty$, and $\lambda_i^k\to\hat{\lambda}^k\geqq0$, $k=1,2,\cdots,p+1$, as $i\to\infty$. Substituting into (70) and taking limits, we get that

$$(75)\qquad \nabla f(\hat{x})+\sum_{k=1}^{p+1}\hat{\lambda}^k\nabla_x\phi(\hat{x},\hat{\omega}^k)=0.$$

Also, since $\phi(x_i,\omega_i^k)\geqq0$ for all $i=0,1,2,\cdots$, and $k=1,2,\cdots,p+1$, and since $\psi_\Omega(\hat{x})=0$, we must have $\phi(\hat{x},\hat{\omega}^k)=0$ for $k=1,2,\cdots,p+1$. But this shows that $\hat{x}$ satisfies the Kuhn–Tucker conditions and therefore $\hat{x}\in\Delta$(which only requires the F. John condition), and hence we get a contradiction. Thus the $\theta_{\Omega'}^i(\cdot)$ satisfy Assumption 1 at any $\hat{x}\notin\Delta$ such that $\psi_\Omega(\hat{x})=0$.

b) Now suppose that $\psi_\Omega(\hat{x})>0$. Let $\hat{\delta}>0$ be arbitrary. Then, by Assumption 4, there exists a $\hat{\mu}>0$ such that $\|\nabla p_{\Omega'}(x)\|\geqq\hat{\mu}$ for all $x\in\mathbb{R}^n$ and all $\Omega'\subset\Omega$ finite, for which $\psi_{\Omega'}(x)>(\hat{\delta}/2)\psi_\Omega(\hat{x})$.

Since the $\psi_{\Omega'}(\cdot)$ functions are continuous, uniformly in Ω' (see the corollary to Proposition 1), there exists a $\hat{\rho}>0$ such that if $\Omega'\subset\Omega$ is finite and if $\psi_{\Omega'}(\hat{x})\geqq\hat{\delta}\psi_\Omega(\hat{x})$, then $\psi_{\Omega'}(x)\geqq\frac{1}{2}\hat{\delta}\psi_\Omega(\hat{x})$ for all $x\in B(\hat{x},\hat{\rho})$ and consequently $\|\nabla p_{\Omega'}(x)\|\geqq\hat{\mu}$ for all $x\in B(\hat{x},\hat{\rho})$. Hence, if $\psi_{\Omega'}(\hat{x})\geqq\hat{\delta}\psi_\Omega(\hat{x})$ and $x\in B(\hat{x},\hat{\rho})$,

$$(76)\qquad \theta_{\Omega'}^i(x)=-\Big\|\nabla f(x)+\frac{1}{s_i}\nabla p_{\Omega'}(x)\Big\|\leqq-\frac{1}{s_i}\|\nabla p_{\Omega'}(x)\|+\|\nabla f(x)\|\leqq-\frac{1}{s_i}\hat{\mu}+M'$$

where $M'=\max\{\|\nabla f(x)\|\,|\,x\in B(\hat{x},\hat{\rho})\}$. Since $s_i\to0$ as $i\to\infty$, there exists an $\hat{N}$ such that $-((1/s_i)\hat{\mu}+M)\leqq-\hat{\mu}$ for all $i\geqq\hat{N}$ and hence we see that Assumption 1 holds at $\hat{x}$. This completes our proof. □

Conclusion. The algorithms described in this paper differ from the earlier versions of outer approximations algorithms in three important respects. They have better convergence properties, they are implementable, and they are designed for use in an interactive computing facility. The last property makes them particularly suitable for use in solving computer-aided engineering design problems in which function evaluations are extremely expensive and for which it is generally not possible to specify in advance scaling and algorithm parameters which ensure good computational behavior.

Appendix: List of symbols.

A.1. General problem (7).

f: $\mathbb{R}^n\to\mathbb{R}^1$, cost function

g^j: $\mathbb{R}^n\to\mathbb{R}^1$, simple constraint function

$\phi^k: \mathbb{R}^n \times \mathbb{R}^{p_k} \to \mathbb{R}^1$, functional constraint function

Ω^k compact subset of $\mathbb{R}^{p_k}$

$\Omega \triangleq \Omega^1 \times \Omega^2 \times \cdots \Omega^m$

$\Omega_i^k \subset \Omega^k$, discrete subset

$$\psi_{\Omega'}(\mathrm{x}) = \max\{0; g^j(x), j \in \mathbf{l}; \phi^k(x, \omega^k), \omega^k \in \Omega'^k, k \in \mathbf{m}\}, \Omega' \subseteq \Omega \tag{10}$$

$$\bar{\psi}_{\Omega'^k}(x) \triangleq \max_{\omega \in \Omega'^k} \phi^k(x, \omega^k), k \in \mathbf{m}, \Omega'_k \subseteq \Omega_k \tag{12}$$

$$\bar{\psi}^0(x) \triangleq \max\{g^1(x), g^2(x), \cdots, g^l(x)\} \tag{13a}$$

$$P_\Omega: \min\{f(x) \mid \psi_\Omega(x) \leqq 0\} \tag{6}$$

$$P_{\Omega_i}: \min\{f(x) \mid \psi_{\Omega_i}(x) \leqq 0\}, \Omega_i \subset \Omega \tag{7}$$

$$\theta_\Omega(\hat{x}) \triangleq \min_{\|h\|_\infty \leqq 1} \max\{\langle \nabla f(\hat{x}), h\rangle; g^j(\hat{x}) + \langle \nabla g^j(\hat{x}), h\rangle, j \in \mathbf{l};$$
$$\phi^k(\hat{x}, \omega^k) + \langle \nabla_x \phi(\hat{x}, \omega^k), h\rangle, \omega^k \in \Omega^k, k \in \mathbf{m}\} \tag{9}$$

$\theta_{\Omega_i}^i(x)$: optimality function

$\Delta \triangleq \{x \in \mathbb{R}^n \mid \psi_\Omega(x) \leqq 0, \theta_\Omega(x) = 0\}$

$B(x, \rho) = \{x' \in \mathbb{R}^n \mid \|x' - x\| \leqq \rho\}$

$\mathbf{l} \triangleq \{1, 2, \cdots, l\}$

$\mathbf{m} \triangleq \{1, 2, \cdots, m\}$

A.2. Simplified problem (14a).

$f: \mathbb{R}^n \to \mathbb{R}^1$ cost function

$\phi: \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^1$, functional constraint function

$\Omega \subset \mathbb{R}^p$

$\Omega_i \subseteq \Omega$ discrete subset

$$\psi_{\Omega'}(x) \triangleq \max\{0; \phi(x, \omega), \omega \in \Omega'\}, \Omega' \subseteq \Omega \tag{14c}$$

$$\bar{\psi}_{\Omega'}(x) \triangleq \max_{\omega \in \Omega'} \phi(x, \omega), \Omega' \subseteq \Omega \tag{14d}$$

$$P_\Omega: \min\{f(x) \mid \psi_\Omega(x) \leqq 0\} \tag{14a}$$

$$P_{\Omega i}: \min\{f(x) \mid \psi_{\Omega_i}(x) \leqq 0\} \tag{14b}$$

$$\Omega'_\varepsilon(x) \triangleq \{\omega \in \Omega' \mid \phi(x, \omega) \geqq \psi_{\Omega'}(x) - \varepsilon\} \tag{44}$$

$$\Omega'(x)_+ \triangleq \{\omega \in \Omega' \mid \phi(x, \omega) \geqq 0\} \tag{69a}$$

$$p_{\Omega'}(x) \triangleq \frac{1}{\nu_\Omega} \sum_{\omega \in \Omega'} [\max\{0, \phi(x, \omega)\}]^2, \Omega' \subseteq \Omega \tag{67}$$

$$f_{\Omega'}^i(x) \triangleq f(x) + \frac{1}{s_i} p_{\Omega'}(x) \tag{68}$$

A.3. Optimality functions for simplified problem (14a).

Topkis–Veinott:

$$\theta_{\Omega'}(x) = \min_{h \in \|h\|_\infty \leqq 1} \max \{\langle \nabla f(x), h\rangle, \phi(x, \omega) + \langle \nabla \phi(x, \omega), h\rangle, \omega \in \Omega'\}, \Omega' \subseteq \Omega \tag{9}$$

Pironneau–Polak:

$$\theta_{\Omega'}(x) \triangleq \min_h \{\tfrac{1}{2}\|h\|^2 + \max \{\langle \nabla f(x), h\rangle; \phi(x, \omega) + \langle \nabla_x \phi(x, \omega), h\rangle, \omega \in \Omega'\}\} - \psi_{\Omega'}(x) \tag{36}$$

Zoutendijk:

$$\gamma^\varepsilon_{\Omega'}(x) \triangleq \min_{h \in \|h\|_\infty \leqq 1} \max \{\langle \nabla f(x), h\rangle - \psi_{\Omega'}(x); \langle \nabla_x \phi(x, \omega), h\rangle, \omega \in \Omega'_\varepsilon(x)\}, \Omega' \subseteq \Omega \tag{45}$$

Gonzaga–Polak:

$$\theta_{\Omega'}(x) \triangleq \min \{-\varepsilon \mid \gamma^\varepsilon_\Omega(x) \leqq -\varepsilon, \varepsilon \in \{0\} \cup \{\beta^k \rho \mid k = 0, 1, 2, 3, \cdots\}\} \tag{46}$$

Penalty function methods:

$$\theta^i_{\Omega'}(x) \triangleq -\|\nabla f^i_{\Omega'}(x)\|, \Omega' \subset \Omega, i = 1, 2, 3, \cdots \tag{69}$$

Acknowledgment. The authors wish to thank Dr. David M. Gay, of the Center for Computational Research in Economics and Management Science at the Massachusetts Institute of Technology, for his many valuable suggestions which have led to a substantial improvement in the readability of the manuscript and for pointing out an error in one of the proofs.

REFERENCES

[1] E. W. Cheney and A. A. Goldstein, *Newton's method for convex programming and Tchebycheff approximation*, Numer. Math., I (1959), pp. 253–68.

[2] J. E. Kelley, *The cutting-plane method for solving convex programs*, J. Soc. Indust. Appl. Math., 8 (1960), pp. 703–12.

[3] E. S. Levitin and B. T. Polyak, *Constrained minimization methods*, Ž. Vyčisl. Mat. i Mat. Fiz., 6 (1966), pp. 781–823.

[4] D. M. Topkis, *Cutting plane methods without nested constraints*, Operations Res., 18 (1970), pp. 404–413.

[5] ———, *A note on cutting plane methods without nested constraints*, Operations Res. Center, University of California, Report No. 69-36, 1969.

[6] B. C. Eaves and W. I. Zangwill, *Generalized cutting plane algorithms*, this Journal, 9 (1971), pp. 529–542.

[7] W. W. Hogan, *Applications of a general convergence theory for outer approximations algorithms*, Math. Programming, 5 (1973), pp. 151–168.

[8] D. Q. Mayne, E. Polak and R. Trahan, *An outer approximations algorithm for computer-aided design problems*, Memo #M77-10, Electronics Res. Lab., University of California, Berkeley, Feb. 1977, J.O.T.A., in press.

[9] D. M. Topkis and A. Veinott, *On the convergence of some feasible directions algorithms for nonlinear programming*, this Journal, 5 (1967), pp. 268–79.

[10] E. Polak, *Computational Methods in Optimization*, Academic Press, New York, 1971.

[11] E. Polak, R. Trahan and D. Q. Mayne, *Combined phase I-phase II methods of feasible directions*, Memo #M77-39, Electronics Res. Lab., University of California, Berkeley, 1977, Math. Programming, to appear.

[12] E. Polak and D. Q. Mayne, *An algorithm for optimization problems with functional inequality constraints*, IEEE Trans. Automatic Control, AC-21, pp. 184–193, 1976.

[13] O. Pironneau and E. Polak, *Rate of convergence of a class of methods of feasible directions*, SIAM J. Numer. Anal., 10 (1973), pp. 161–173.

[14] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 4, July 1979

0363-0129/79/1704-004$01.00/0

UNIQUE IDENTIFICATION OF EIGENVALUES AND COEFFICIENTS IN A PARABOLIC PROBLEM*

ALAN PIERCE†

Abstract. This paper discusses uniqueness questions for identification of coefficients in a second-order, linear, one-dimensional, parabolic partial differential equation. Here, the unknowns are spatially-varying coefficients appearing in the equation. The solution of an initial-boundary value problem is observed at one point over a finite time interval. Conditions are given under which the eigenvalues associated with the problem are uniquely determined by such an observation. The coefficients are not uniquely determined. If, however, the equation is in normal form, the single coefficient which appears is, in certain cases, uniquely determined. This can be established by obtaining the spectral function or by obtaining the eigenvalues for two different boundary value problems and applying existing results (I. M. Gelfand and B. M. Levitan (1959), N. Levison (1949)).

1. Introduction. As part of their work on parameter identification in parabolic problems, Kitamura and Nakagiri [6] gave conditions under which coefficients in a parabolic equation could be determined by point-wise measurement. If these coefficients are constant, observation of the solution of the equation at one point for all time, under conditions they specified, determines eigenvalues of the equation. The coefficients in the equation they considered are then easily obtained.

The situation is naturally more difficult in the case of nonconstant coefficients. A number of results exist for problems whose coefficients depend only on the solution [1] or on the time variable [5]. For problems whose coefficients are functions of the space variable, it is still possible to carry out the program of Kitamura and Nakagiri to a certain extent.

In this paper it is shown that, under certain conditions, eigenvalues of a parabolic partial differential equation are uniquely determined by an observation of the solution at one point in space. The coefficients of the equation may be functions of the space variable. In addition, the observed solution need not be known for all time but only on a finite interval.

In contrast to the case in which the coefficients are constant, knowledge of the eigenvalues does not directly yield the coefficients. If, however, the equation is in normal form, in which one coefficient appears, then this coefficient can be uniquely determined by observation at one point in space.

2. Eigenvalue identification. Consider the problem

$$(au_x)_x - bu_t - cu = Q \qquad \text{on } (0,1)\times(0,T], \tag{1}$$

$$u = u_0, \qquad t = 0, \tag{2}$$

$$\alpha_0 u - (1-\alpha_0)u_x = q_0, \qquad x = 0, \tag{3}$$

$$\alpha_1 u + (1-\alpha_1)u_x = q_1, \qquad x = 1. \tag{4}$$

In all that follows, it will be supposed that a and b are positive and that they, with c, Q, u_0, q_0, and q_1, are sufficiently smooth to guarantee existence of a unique solution u with continuous partial derivatives on $[0,1]\times[0,T]$ (see, e.g., [7, pp. 320–321] for appropriate Hölder continuity and compatibility conditions).

THEOREM 1. *Let a, b, c be unknown functions of x in $[0,1]$ and let α_0, α_1 be given. Let $\{\phi_n\}$ be the nontrivial solutions of*

$$(a\phi_n')' + (\lambda_n b - c)\phi_n = 0, \tag{5}$$

* Received by the editors April 13, 1978 and in revised form October 9, 1978.

† Amoco Production Company, Tulsa, Oklahoma 74102.

$$\alpha_0\phi_n-(1-\alpha_0)\phi_n'=0, \qquad x=0, \tag{6}$$

$$\alpha_1\phi_n+(1-\alpha_1)\phi_n'=0, \qquad x=1, \tag{7}$$

ordered as $\lambda_0<\lambda_1<\lambda_2<\cdots$. *Let u be a solution of* (1), (2), (3), (4). *Suppose* x_0 *is not in the countable set* $N=\{x\in[0,1]\,|\,\phi_n(x)=0$ *for some* $n\}$. *Assume that exactly one of the inputs* Q, u_0, q_0, q_1 *is nonzero. Then the sequence of eigenvalues* $\{\lambda_n\}$ *is uniquely determined by knowledge of that single input and of* $u(x_0,t)$ *for* $0\leqq t\leqq T$ *in the following cases*:

(i) *if* Q *has the form* $h(x)q(t)$ *with* q *vanishing in no interval about* $t=0$ *and if, for all* n, $\int_0^1 h(y)\phi_n(y)\,dy\neq 0$,

(ii) *if, for all* n,

$$\int_0^1 b(y)u_0(y)\phi_n(y)\,dy\neq 0,$$

(iii) *if* q_0 *vanishes in no interval about* $t=0$,

(iv) *if* q_1 *vanishes in no interval about* $t=0$.

Proof. If the coefficients a, b, and c in (1) are not known, then (5), (6), (7) cannot be solved directly. Nevertheless, the eigenfunctions may be chosen to be a complete orthonormal set on [0, 1] (with weight function b). Further, the generalized Fourier expansion

$$u(x,t)=\sum_{n=0}^{\infty}\left[\int_0^1 b(y)u(y,t)\phi_n(y)\,dy\right]\phi_n(x)$$

may be rewritten as

$$\begin{aligned} u(x,t)=&\sum_{n=0}^{\infty}(u_0,\phi_n)\,e^{-\lambda_n t}\phi_n(x)\\ &-\int_0^t\left\{\sum_{n=0}^{\infty}e^{-\lambda_n(t-\tau)}Q_n(\tau)\phi_n(x)\right\}d\tau\\ &+\int_0^t\left\{\sum_{n=0}^{\infty}a(0)[\phi_n(0)+\phi_n'(0)]\,e^{-\lambda_n(t-\tau)}\phi_n(x)\right\}q_0(\tau)\,d\tau\\ &+\int_0^t\left\{\sum_{n=0}^{\infty}a(1)[\phi_n(1)-\phi_n'(1)]\,e^{-\lambda_n(t-\tau)}\phi_n(x)\right\}q_1(\tau)\,d\tau \end{aligned} \tag{8}$$

[11, pp. 215–216]. Here

$$(u_0,\phi_n)=\int_0^1 b(y)u_0(y)\phi_n(y)\,dy$$

and

$$Q_n(t)=\int_0^1 Q(y,t)\phi_n(y)\,dy.$$

Case (i). Now let u_0, q_0, and q_1 be zero. Let $Q(x,t)=h(x)q(t)$ be known. Then, from (8),

$$u(x_0,t)=-\int_0^t\left\{\sum_{n=0}^{\infty}e^{-\lambda_n(t-\tau)}\phi_n(x_0)\int_0^1 h(y)\phi_n(y)\,dy\right\}q(\tau)\,d\tau$$

for $0\leqq t\leqq T$. Since q does not vanish identically in any neighborhood of $t=0$, the integral equation

$$u(x_0, t) = \int_0^t f(t-\tau)q(\tau)\,d\tau$$

has a unique solution f [12, pp. 324–325]. By the above, this solution must be

$$f(t) = -\sum_{n=0}^{\infty} e^{-\lambda_n t}\phi_n(x_0)\int_0^1 h(y)\phi_n(y)\,dy.$$

Whatever the coefficients a, b, and c, the eigenfunctions are uniformly bounded and the eigenvalues are asymptotic to a multiple of n^2 [4, pp. 270–273]. Thus this Dirichlet series converges for all $t > 0$ and, in fact, on the right half of the complex plane, so f can be extended to an analytic function on the right half plane. Since a function can be represented in at most one way by a Dirichlet series [10, p. 435], the exponents $\{\lambda_n\}$ (as well as the coefficients) are uniquely determined.

Of course, only those exponents actually appearing in the series are determined in this way. But x_0 is not in N, so no $\phi_n(x_0)$ is 0. Likewise, each of the integrals is nonzero. Thus, the observation $u(x_0, \cdot)$ uniquely determines all the eigenvalues.

Case (ii). Suppose now that u_0 is nonzero but that Q, q_0, and q_1 are everywhere zero. In this case, the series in (8) is already a Dirichlet series:

$$u(x_0, t) = \sum_{n=0}^{\infty} (u_0, \phi_n)\,e^{-\lambda_n t}\phi_n(x_0).$$

As before, the exponents which appear are uniquely determined by the observation $u(x_0, \cdot)$. By hypothesis, no coefficient is zero so all the eigenvalues are determined.

Case (iii). Suppose that Q, u_0 and q_1 vanish everywhere. In this case (8) becomes

$$u(x_0, t) = \int_0^t \left\{\sum_{n=0}^{\infty} a(0)[\phi_n(0) + \phi_n'(0)]\,e^{-\lambda_n(t-\tau)}\phi_n(x_0)\right\}q_0(\tau)\,d\tau.$$

In the same manner as before, if q_0 does not vanish identically in any interval about $t = 0$, then this relation uniquely determines the exponents appearing in the Dirichlet series

$$\sum_{n=0}^{\infty} a(0)[\phi_n(0) + \phi_n'(0)]\phi_n(x_0)\,e^{-\lambda_n t}.$$

A full determination of the eigenvalues from this series requires, once more, that each of the coefficients be nonzero. Here, however, the form in which the driving force is supplied to the system is sufficiently "impulsive" to avoid an assumption involving the unknown sequence $\{\phi_n\}$.

The function a was assumed to be positive so there is no problem with $a(0)$. The factor $\phi_n(0) + \phi_n'(0)$ is always nonzero. For, otherwise, it follows from (6) that $\phi_n(0) = \phi_n'(0) = 0$ from which, in turn, the eigenfunction ϕ_n would have to vanish identically. Again, each $\phi_n(x_0)$ is nonzero. Thus the observation $u(x_0, \cdot)$ uniquely determines all the eigenvalues.

In particular, if $\alpha_0 \neq 1$ it follows from (6) that $\phi_n(0) \neq 0$. Likewise, if $\alpha_1 \neq 1$, then $\phi_n(1) \neq 0$, so that an observation at $x_0 = 0$ or $x_0 = 1$, respectively, is sufficient to determine the eigenvalues.

Case (iv). The situation in which $q_0 \equiv 0$ but q_1 is nonzero may be treated in the same way as Case (iii).

3. Identification of a coefficient. In additional contrast to the case of constant coefficients (cf. [6]), the determination even of all the eigenvalues associated with (1), (2), (3), (4) does not identify the coefficients a, b and c. For, by the transformation

$$z=\int_0^x [b(y)/a(y)]^{1/2}\,dy \qquad \left(l=\int_0^1 [b(y)/a(y)]^{1/2}\,dy\right),$$

$$v=[ab]^{1/4}u,$$

the operator of (1) may be brought to Liouville normal form

$$v_{zz}-v_t-dv$$

on $0<z<l$, $0<t\leqq T$ (see, e.g., [2, p. 292]). Clearly, the system is governed by the coefficient d, rather than a, b, and c themselves.

Suppose that the problem is given in this form, that is, $a\equiv 1$ and $b\equiv 1$ in (1). If neither α_0 nor α_1 is 1, then the problem (1), (2), (3), (4) may be taken to be

$$L(c)u=u_{xx}-u_t-cu=Q, \tag{9}$$

$$u=u_0, \qquad t=0, \tag{10}$$

$$B_0(h)u=u_x(0,t)-hu(0,t)=q_0, \tag{11}$$

$$B_1(H)u=u_x(1,t)+Hu(1,t)=q_1. \tag{12}$$

Consider the following result of Levinson [8] for the associated Sturm–Liouville problem: if the eigenvalues of

$$\psi''+(\lambda-c)\psi=0$$

are known for the boundary conditions

$$B_0(h)\psi=0, \qquad B_1(H)\psi=0$$

and for the boundary conditions

$$B_0(h)\psi=0, \qquad B_1(H_1)\psi=0$$

where $H\neq H_1$, then c is uniquely determined.

Eigenvalues are uniquely determined by $u(x_0, t)$ under the circumstances given in Theorem 1. If the system can be observed under conditions described by two different boundary conditions in the form (11) and (12)—that is, for two different pairs (h, H) and (h, H_1)—then this shows that the coefficient c is uniquely determined.

The following result describes circumstances under which observation at a point for a single set of boundary conditions is sufficient to uniquely determine the coefficient c. (For a similar approach, in which the coefficient a in (1) is determined under the assumption that $b\equiv 1$ and $c\equiv 0$, see [9].)

THEOREM 2. *Suppose u_1 and u_2 satisfy*

$$L(c_1)u_1=L(c_2)u_2=0$$

for some continuously differentiable c_1 and c_2, and

$$u_1=u_2=0, \qquad t=0,$$

$$B_0(h)u_1=B_0(h)u_2=q_0,$$

$$B_1(H)u_1=B_1(H)u_2=0.$$

Let q_0 be continuous and not identically zero in any interval about $t=0$. Suppose further *that observations of u_1 and u_2 at* 0 *agree, i.e., that* $u_1(0, t) = u_2(0, t)$ *for* $0 \leqq t \leqq T$. *Then* $c_1 = c_2$.

Proof. Let $\{\lambda_{n,1}\}$, $\{\psi_{n,1}\}$, and $\{\lambda_{n,2}\}$, $\{\psi_{n,2}\}$ satisfy the associated homogeneous Sturm–Liouville problems, i.e., let

$$\psi''_{n,i} + (\lambda_{n,i} - c_i)\psi_{n,i} = 0,$$
$$B_0(h)\psi_{n,i} = 0,$$
$$B_1(H)\psi_{n,i} = 0,$$

for $i = 1, 2$. Let the orthogonal functions $\{\psi_{n,i}\}$ be normalized so that

$$\psi_{n,i}(0) = 1$$

and let

$$\rho_{n,i} = 1\bigg/\int_0^1 \psi_{n,i}(y)^2\,dy.$$

In the same way that (8) was obtained, it follows that

$$u_i(x, t) = -\int_0^t \sum_{n=0}^{\infty} \rho_{n,i}\psi_{n,i}(0)\psi_{n,i}(x)\,e^{-\lambda_{n,i}(t-\tau)}q_0(\tau)\,d\tau.$$

In particular,

$$u_i(0, t) = -\int_0^t \left\{\sum_{n=0}^{\infty} \rho_{n,i}\,e^{-\lambda_{n,i}(t-\tau)}\right\}q_0(\tau)\,d\tau.$$

But, since $u_1(0, t) = u_2(0, t)$, it must follow that these two integral equations have the same solution:

$$\sum_{n=0}^{\infty} \rho_{n,1}\,e^{-\lambda_{n,1}t} = \sum_{n=0}^{\infty} \rho_{n,2}\,e^{-\lambda_{n,2}t}$$

for $0 < t \leqq T$. As noted in the proof of Theorem 1, the two Dirichlet series must have the same exponents and coefficients. Thus

$$\lambda_{n,1} = \lambda_{n,2}$$

and

$$\rho_{n,1} = \rho_{n,2}$$

for all n. Let $\{\lambda_n\}$ and $\{\rho_n\}$ denote these two sequences.

The problem may be converted, if necessary, to one with a nonnegative spectrum by considering

$$\psi''_{n,i} + (\mu_n - (c_i - \lambda_0))\psi_{n,i} = 0$$

where $\mu_n = \lambda_n - \lambda_0$. The spectral function for both of these problems is

$$\rho(\mu) = \sum_{\mu_n \leqq \mu} \rho_n.$$

The proof is completed by following Gelfand and Levitan [3] and considering the function

$$F(x, y) = \int_0^{\infty} \frac{\sin\sqrt{\mu}x \sin\sqrt{\mu}y}{\mu}\,d\sigma(\mu)$$

where $\sigma(\mu)=\rho(\mu)-(2/\pi)\sqrt{\mu}$. As shown in [3],

$$f=\partial^2F/\partial x\partial y$$

is continuous and the integral equation

$$f(x,y)+\int_0^x K(x,s)f(s,y)\,ds+K(x,y)=0 \qquad (0\leqq y\leqq x\leqq 1)$$

has, for each fixed x, a unique solution $K(x,y)$.

This solution K satisfies

$$\psi_{n,i}(x)=\cos\sqrt{\mu_n}\,x+\int_0^x K(x,t)\cos\sqrt{\mu_n}\,t\,dt$$

and

$$c_i(x)-\lambda_0=2\frac{d}{dx}K(x,x)$$

for $i=1,2$. It follows from either of these relations that $c_1=c_2$.

Acknowledgment. The author is indebted to Dr. Kenneth R. Driessel for a number of helpful discussions of this problem. Thanks are also due to the reviewer for suggesting improvements in the exposition.

REFERENCES

[1] J. R. CANNON AND P. DUCHATEAU, *Determining unknown coefficients in a nonlinear heat conduction problem*, SIAM J. Appl. Math., 24 (1973), pp. 298–314.

[2] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, vol. I, Interscience, New York, 1953.

[3] I. M. GELFAND AND B. M. LEVITAN, *On the determination of a differential equation from its spectral function*, Izvestiya Akad. Nauk SSSR, 15 (1951), pp. 309–360; English transl., Amer. Math. Soc. Translations, 1 (1955), pp. 253–304.

[4] E. L. INCE, *Ordinary Differential Equations*, Dover, New York, 1944.

[5] B. F. JONES, JR., *The determination of a coefficient in a parabolic differential equation*, Part I: Existence and Uniqueness, J. Math. Mech., 11 (1962), pp. 907–918.

[6] S. KITAMURA AND S. NAKAGIRI, *Identifiability of spatially-varying and constant parameters in distributed systems of parabolic type*, this Journal, 15 (1977), pp. 785–802.

[7] O. A. LADYZENSKAJA, V. A. SOLONNIKOV AND N. N. URAL'CEVA, *Linear and Quasi-linear Equations of Parabolic Type*, Amer. Math. Soc., Providence, R.I., 1968.

[8] N. LEVINSON, *The inverse Sturm–Liouville problem*, Mat. Tidsskr. B, (1949), pp. 25–30.

[9] H.-P. LINKE, *Über ein inverses Problem für die Wärmeleitungsgleichung*, Wiss. Z. Techn. Hochsch. Karl-Marx-Stadt, 18 (1976), pp. 445–447.

[10] S. SAKS AND A. ZYGMUND, *Analytic Functions*, Monografie Matematyczne, Warsaw, 1965.

[11] I. STAKGOLD, *Boundary Value Problems of Mathematical Physics*, vol. II, Macmillan, New York, 1968.

[12] E. C. TITCHMARSH, *Introduction to the Theory of Fourier Integrals*, 2nd Ed., Clarendon Press, Oxford, 1948.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 4, July 1979

0363-0129/79/1704-0006$01.00/0

ALGEBRAIC THEORY OF LINEAR TIME-VARYING SYSTEMS*

EDWARD W. KAMEN† AND KHALED M. HAFEZ‡

Abstract. An algebraic theory of linear time-varying discrete-time systems is developed in terms of a module structure defined over a noncommutative polynomial ring. The module setup is induced from a semilinear transformation that is derived from the given system. Various structural properties of the module framework are explored including the concepts of cyclicity and n-cyclicity. The module theory is then applied to the study of reachability and state feedback. Results on the construction of feedback controllers are obtained that resemble pole or coefficient assignability in the theory of time-invariant systems.

1. Introduction. In 1965, R. E. Kalman [1] presented an algebraic theory for the class of linear time-invariant discrete-time systems defined over a field K. The central component of Kalman's theory is a module structure on the state space defined over the commutative ring $K[z]$ of polynomials in z with coefficients in K. In this paper we present an algebraic theory for linear time-varying discrete-time systems, which is also based on a module structure. However, in contrast to Kalman's theory, the approach developed here is based on the concept of a semilinear transformation S defined on a function space V containing all possible state trajectories. The main component of the algebraic framework is an S-induced module structure on V defined over a skew (noncommutative) polynomial ring.

Skew polynomial rings have been applied to various problems in the theory of time-varying networks and systems, such as network synthesis (Newcomb [2]), system realization (Kaman [3]), and structural analysis of input/output operators (Salovaara and Blomberg [4], Ylinen [5]). The utilization of skew polynomial rings in the state-space theory of linear time-varying systems was first considered by Kamen in an unpublished report [6]. Several of the results in the present paper were first given in the Ph.D. thesis of Hafez [14].

In the present work, time-varying systems are specified by a triple of matrices defined over a ring of time functions. This leads to a characterization of systems in terms of a semilinear transformation S defined in § 2. In § 3, we use S to induce a module structure over a skew polynomial ring. Here we introduce the concepts of cyclicity and n-cyclicity. As shown in § 3, n-cyclicity is equivalent to the existence of a canonical form which is the discrete-time counterpart to the phase-variable form constructed by Silverman [7] in the theory of linear time-varying continuous-time systems.

In the last section of the paper we study reachability and state feedback in terms of S and the induced module structure. Using the concept of n-cyclicity, we present a new approach to the construction of state-feedback controllers. The theory yields results that resemble coefficient assignability of the characteristic polynomial in the theory of time-invariant systems. This framework is then compared to pole assignment-type results based on a discrete-time version of index invariance defined by Morse and Silverman [8] for continuous-time systems.

2. System description and basic properties. Let Z denote the set of integers and let K denote a fixed field (finite or infinite). Let R denote the set of all functions defined on Z with values in K. With pointwise addition and multiplication given by $(a+b)(t)=$

* Receiv̇ed by the editors July 7, 1978, and in revised form December 19, 1978. This work was supported by the United States Army Research Office, under Grants DA-ARO-D-31-124-73-G171 and DAAG29-77-G-0085.

† School of Electrical Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332.

‡ Department of Mathematical Sciences, Memphis State University, Memphis, Tennessee 38153.

$a(t)+b(t)$, $(ab)(t)=a(t)b(t)$, $a, b \in R$, R is a commutative ring. The function $1: Z \to Z: t \to 1$ is the multiplicative identity of the ring R.

Let $\sigma: R \to R$ denote the right-shift operator on R given by $\sigma: a(t) \to a(t-1)$. The operator σ is a (ring) automorphism on R; that is, σ is bijective and $\sigma(a+b)=\sigma a+\sigma b$, $\sigma(ab)=(\sigma a)(\sigma b)$ for all $a, b \in R$. The inverse σ^{-1} of σ is the left-shift operator on R. The ring R with the automorphism σ is said to be a difference ring.

Finally, for any positive integer n, let K^n denote the vector space consisting of all n-element column vectors over the field K. In terms of these constructions, we have the following notion of a system.

DEFINITION 2.1. Let m, n, p be fixed positive integers. An *m-input p-output n-dimensional linear time-varying discrete-time system over the difference ring* R is a triple (F, G, H) of $n \times n$, $n \times m$, $p \times n$ matrices over R, together with the dynamical equations

$$x(t+1)=F(t)x(t)+G(t)u(t), \tag{2.1}$$

$$y(t)=H(t)x(t) \tag{2.2}$$

where $x(t) \in K^n$ is the state at time $t \in Z$, $u(t) \in K^m$ is the input or control at time t, and $y(t) \in K^p$ is the output at time t.

For notational convenience, we shall work with the following modified version of the standard equation (2.1). Apply the right-shift operator to both sides of (2.1) and let $D(t)=F(t-1)$, $E(t)=G(t-1)$, so that we have

$$x(t)=D(t)x(t-1)+E(t)u(t-1). \tag{2.3}$$

From here on, we shall work with the system representation given by the dynamical equations (2.2)–(2.3). The system given by (2.2)–(2.3) will be denoted by the triple (D, E, H).

Our first objective is to characterize systems over R in terms of a semilinear transformation defined as follows. Let R^n denote the free R-module consisting of all n-element column vectors over R. Given an $n \times n$ matrix M over R, let S denote the operator on R^n defined by $S: R^n \to R^n: v \to M(\sigma v)$, where $(\sigma v)(t)=v(t-1)$. The operator S is clearly additive; i.e., $S(v_1+v_2)=S(v_1)+S(v_2)$ for all $v_1, v_2 \in R^n$. However, given $a \in R$ and $v \in R^n$, we have that $S(av)=M(\sigma(av))=M(\sigma a)(\sigma v)=(\sigma a)S(v)$, so S is *not linear* with respect to the R-module structure on R^n. The operator S is said to be a semilinear transformation relative to σ (see [9]).

Now let (D, E, H) be an n-dimensional system over R. We shall refer to the operator $S: R^n \to R^n: v \to D(\sigma v)$ as the *semilinear transformation* (s.l.t) *of the system* (D, E, H).

As we now show, the state and output responses of a system over R can be expressed in terms of the system's s.l.t. First, given $t_0 \in Z$ and $x_0 \in K^n$, let $\hat{x}_0$ denote the element in R^n defined by $\hat{x}_0(t)=x_0$ when $t=t_0$ and $\hat{x}_0(t)=0$ when $t \neq t_0$. Let $S^0=I=$ identity operator on R^n, and for $i=1, 2, \cdots$, define $(S^iE)(t)=D(t)(S^{i-1}E)(t-1)$.

PROPOSITION 2.2. *Let* (D, E, H) *be a system over* R *given by the dynamical equations* (2.2)–(2.3). *The solution* $x(t)$ *of* (2.3) *resulting from initial state* $x_0 \in K^n$ *at initial time* $t_0 \in Z$ *and input* $u(t)$, $t \geqq t_0$, *can be expressed in the form*

$$x(t)=(S^{t-t_0}\hat{x}_0)(t)+\sum_{i=t_0}^{t-1}(S^{t-i-1}E)(t)u(i), \qquad t>t_0. \tag{2.4}$$

Proof. Solve (2.3) using iteration and then apply the definition of S. Q.E.D.

COROLLARY 2.3. *Let $u(t) \in R^m$ be an input function with support bounded on the left. Then the output response $y(t)$ resulting from input $u(t)$ with zero initial state is given by*

$$y(t) = \sum_{i=-\infty}^{t-1} (HS^{t-i-1}E)(t)u(i). \tag{2.5}$$

Proof. Combine (2.2) and (2.4). Q.E.D.

The expressions (2.4)–(2.5) are very interesting since they are identical in form to the expressions for the state and output responses in the time-invariant case. In fact, if the matrix functions $D(t)$, $E(t)$, $H(t)$ are constant, replacing S by D and $\hat{x}_0$ by x_0 in (2.4)–(2.5) we get the response functions for the time-invariant case.

Since there is an operator theory for time-invariant systems based on a D-induced module structure (i.e. Kalman's $K[z]$-module theory), the expressions (2.4)–(2.5) suggest the possibility of an operator theory for time-varying systems constructed in terms of an S-induced module structure. The primary objective of this paper is to develop such a theory. But before we begin to do this, we need to consider the notion of system equivalence.

DEFINITION 2.4. Let (D, E, H) and $(\bar{D}, \bar{E}, \bar{H})$ be m-input p-output n-dimensional systems over R. Then (D, E, H) and $(\bar{D}, \bar{E}, \bar{H})$ are *equivalent* if there exists an $n \times n$ matrix P over R, with inverse P^{-1} also over R, such that $\bar{D} = P^{-1}D(\sigma P)$, $\bar{E} = P^{-1}E$, and $\bar{H} = HP$, where $(\sigma P)(t) = P(t-1)$.

As is well known, equivalent systems are related by a coordinate change in the state space K^n: Let $x(t)$ (resp. $\bar{x}(t)$) denote the state of (D, E, H) (resp. $(\bar{D}, \bar{E}, \bar{H})$) resulting from initial state $x(t_0)$ (resp. $\bar{x}(t_0)$ with $\bar{x}(t_0) = P^{-1}(t_0)x(t_0)$) and input $u(t)$, $t \geqq t_0$. Then $x(t) = P(t)\bar{x}(t)$ for all $t \geqq t_0$. Since $H(t)x(t) = H(t)P(t)\bar{x}(t) = \bar{H}(t)\bar{x}(t)$ for $t \geqq t_0$, equivalent systems produce the same output response when excited by the same input with zero initial state.

Let S (resp. $\bar{S}$) denote the s.l.t. of the system (D, E, H) (resp. $(\bar{D}, \bar{E}, \bar{H})$). If there exists an invertible matrix P over R such that $\bar{D} = P^{-1}D(\sigma P)$, the s.l.t.'s S and $\bar{S}$ are said to be *similar*. Thus, equivalent systems have similar s.l.t.'s.

3. S-Induced module structure. Given an s.l.t. S on R^n, in this section we use S to induce a module structure on R^n defined over a skew polynomial ring, and then we study various aspects of the resulting operator framework. The results obtained here will be utilized in the next section in the study of reachability and state feedback.

Let $R[z]$ denote the set of all finite sums of the form $\sum_i a_i z^i$ in the symbol z with coefficients $a_i \in R$ written on the left. With the usual addition and with multiplication defined by $z^i z^j = z^{i+j}$, $za = (\sigma a)z$, $a \in R$, $R[z]$ is a noncommutative ring, called a (left) skew polynomial ring [10].

Let $\pi(z) = \sum_i a_i z^i$ be an element of $R[z]$. We define the degree of $\pi(z)$ by $\deg \pi(z) = \max\{i : a_i \neq 0\}$ when $\pi(z) \neq 0$, and $\deg \pi(z) = -\infty$ when $\pi(z) = 0$. If $\deg \pi(z) = n$ and $a_n = 1$, $\pi(z)$ is said to be *monic*. Monic polynomials are nonzero divisors, i.e., if $\pi(z)$ is monic and $\pi(z)\theta(z) = 0$ for some $\theta(z) \in R[z]$, then $\theta(z) = 0$.

Given $\pi(z)$, $\theta(z) \in R[z]$, it is easily verified that $\deg(\pi(z) + \theta(z)) \leqq \max\{\deg \pi(z), \deg \theta(z)\}$. However, in general $\deg(\pi(z)\theta(z)) \neq \deg \pi(z) + \deg \theta(z)$, since R is not an integral domain.

Now let S be the s.l.t. of the system (D, E, H). Then S induces a left $R[z]$-module structure on R^n with addition and scalar multiplication given by

$$(v + w)(t) = v(t) + w(t),$$

$$\left(\sum_i a_i z^i\right) v = \sum_i a_i S^i(v)$$

where $v, w \in R^n$, $a_i \in R$.

Note that this construction resembles the manner in which a $K[z]$-module structure is induced on the state space in Kalman's algebraic theory of time-invariant systems [1], [11].

Now R^n has both an R-module structure and an $R[z]$-module structure. As will be seen, there is a good deal of interplay between these two structures. In studying the $R[z]$-module framework, we shall need the following results involving the R-module structure on R^n.

Given a positive integer q and elements $v_1, v_2, \cdots, v_q$ belonging to R^n, let $\langle v_1, \cdots, v_q\rangle_R$ denote the R-submodule of R^n consisting of all R-linear combinations of $v_1, \cdots, v_q$. Let $[v_1, \cdots, v_q]$ denote the $n\times q$ matrix whose ith column is equal to v_i. Finally, we say that $v_1, \cdots, v_q$ are R-independent if $\sum_i a_iv_i = 0$ implies that $a_i = 0$ for all i. We then have the following (known) results.

PROPOSITION 3.1. *Suppose that $q \geqq n$. Then the following are equivalent*:

(1) $\langle v_1, \cdots, v_q\rangle_R = R^n$;

(2) rank $[v_1, \cdots, v_q](t) = n$ *for all* $t\in Z$;

(3) *for every $t\in Z$, there is an $n\times n$ submatrix of $[v_1, \cdots, v_q](t)$ with nonzero determinant.*

PROPOSITION 3.2. *The following are equivalent*:

(1) $\langle v_1, \cdots, v_n\rangle_R = R^n$;

(2) $[v_1, \cdots, v_n]$ *is invertible over* R;

(3) *the determinant of* $[v_1, \cdots, v_n](t)$ *is nonzero for all* $t\in Z$;

(4) $v_1, \cdots, v_n$ *are R-independent.*

We begin the study of the $R[z]$-module structure by considering the notion of cyclicity: An s.l.t. S on R^n is said to be cyclic with generator $g\in R^n$ if R^n is cyclic as an $R[z]$-module with generator g. That is, any $v\in R^n$ can be expressed in the form $v = \pi_v(z)g$, where $\pi_v(z)$ is some element of $R[z]$ depending on v. If R^n is cyclic with generator g, then $R^n = R[z]g = \{\pi(z)g: \pi(z)\in R[z]\}$.

A necessary and sufficient condition for S to be cyclic with generator g can be given in terms of the R-module structure on R^n as follows.

PROPOSITION 3.3. *An s.l.t. S on R^n is cyclic with generator g if and only if there is a positive integer q such that* $\langle g, Sg, \cdots, S^{q-1}g\rangle_R = R^n$.

Proof. If $\langle g, Sg, \cdots, S^{q-1}g\rangle_R = R^n$, for any $v\in R^n$ there exist $a_0, a_1, \cdots, a_{q-1}\in R$ such that $v = \sum_{i=0}^{q-1} a_iS^ig = (\sum_i a_iz^i)g$. Conversely, suppose that S is cyclic with generator g. Let $w_1, \cdots, w_n$ be a basis for R^n. Then for each i, there is a $\pi_i(z)\in R[z]$ such that $w_i = \pi_i(z)g$. Since the w_i generate R^n, for any $v\in R^n$, there exist $a_1, a_2, \cdots, a_n\in R$ such that $v = \sum_{i=1}^n a_iw_i = \sum_{i=1}^n a_i\pi_i(z)g = (\sum_i a_i\pi_i(z))g$. Hence R is generated by $g, Sg, \cdots, S^{q-1}g$, where $q = \max\{\deg \pi_i\}$. Q.E.D.

Suppose that S is cyclic with generator g, and let q be the smallest integer for which $\langle g, Sg, \cdots, S^{q-1}g\rangle_R = R^n$. It is very possible that q is strictly greater than n. This situation can occur because, in contrast to the theory of linear transformations, there is no Cayley–Hamilton theorem for semilinear transformations. In other words, it may not be possible to express S^n as an R-linear combination of $I, S, \cdots, S^{n-1}$ (see [10, p. 300]). A class of s.l.t.'s for which $q = n+1$ is constructed in the following example.

Example 3.4. Given the s.l.t. $S: R^n \to R^n: v \to D(\sigma v)$ and $g\in R^n$, let $C(t)$ denote the $n\times n$ matrix $[g, Sg, \cdots, S^{n-1}g](t)$. Suppose that the determinant of $C(t)$, denoted by $\det C(t)$, is zero when $t = t_0$ and nonzero when $t\neq t_0$. Then by Proposition 3.2, $\langle g, Sg, \cdots, S^{n-1}g\rangle_R \neq R^n$. Now $SC = D(\sigma C)$, so $\det(SC) = (\det D)(\det(\sigma C)) = (\det D)(\sigma(\det C))$. Therefore, if $\det D(t_0)\neq 0$, $\det(SC)(t_0)\neq 0$. Hence for every $t\in Z$, there is an $n\times n$ submatrix of $[g, Sg, \cdots, S^{n-1}g, S^ng](t)$ with nonzero determinant. Thus by Proposition 3.1, $\langle g, Sg, \cdots, S^ng\rangle_R = R^n$, showing that $q = n+1$.

If the elements $g, Sg, \cdots, S^{n-1}g$ generate R^n, the s.l.t. S is said to be *n-cyclic* with generator g. In this paper we restrict attention to cyclic s.l.t.'s that are n-cyclic. The theory of cyclic s.l.t.'s with q strictly greater than n will be developed in a separate paper using results derived below for the n-cyclic case.

We shall now characterize n-cyclicity in terms of the $R[z]$-module structure on R^n. Again let S be a fixed s.l.t. on R^n. Given $w \in R^n$, let Ann w denote the annihilator of w defined by Ann $w = \{\pi(z) \in R[z]: \pi(z)w = 0\}$. It is easily checked that Ann w is a left ideal of the ring $R[z]$.

THEOREM 3.5. *Given $g \in R^n$, the s.l.t. S is n-cyclic with generator g if and only if* Ann $g = R[z]\psi(z)$, *where $\psi(z)$ is a monic polynomial of degree n.*

Proof. Suppose that S is n-cyclic with generator g. Then since $\langle g, Sg, \cdots, S^{n-1}g\rangle_R = R^n$, there exist $a_i \in R$, $i = 0, 1, \cdots, n-1$, such that $S^n g = \sum_{i=0}^{n-1} a_i S^i g$. Thus $(z^n - \sum_i a_i z^i)g = 0$, so $\psi(z) \triangleq z^n - \sum_i a_i z^i \in$ Ann g. Clearly, $R[z]\psi(z) \subset$ Ann g, so it must be shown that Ann $g \subset R[z]\psi(z)$: Let $\pi(z) \in$ Ann g. Since $\psi(z)$ is monic, there exist polynomials $q(z)$ and $r(z)$ such that $\pi(z) = q(z)\psi(z) + r(z)$, with $\deg r(z) < \deg \psi(z)$. The existence of $q(z)$, $r(z)$ can be shown by a straightforward modification of the proof of the polynomial division theorem [12, p. 120] for commutative polynomial rings. Now since $\pi(z)$ and $q(z)\psi(z)$ belong to Ann g, $r(z)$ must also belong to Ann g. But since $\deg r(z) \leqq n-1$ and the elements $g, Sg, \cdots, S^{n-1}g$ are R-independent (by Proposition 3.2), $r(z)$ must be zero. Thus $\pi(z) \in R[z]\psi(z)$. Conversely, suppose that Ann $g = R[z]\psi(z)$ with $\psi(z) = z^n + \sum_{i=0}^{n-1} b_i z^i$, $b_i \in R$. Then $S^n g = -\sum_{i=0}^{n-1} b_i S^i g$, which implies that $\langle g, Sg, S^2 g, \cdots\rangle_R = \langle g, S, \cdots, S^{n-1}g\rangle_R$. Now suppose that $\langle g, Sg, \cdots, S^{n-1}g\rangle_R \neq R^n$. Then by Proposition 3.2, $g, Sg, \cdots, S^{n-1}g$ are R-dependent. Hence Ann g contains a nonzero polynomial with degree strictly less than n. But $\deg \pi(z)\psi(z) \geqq \deg \psi(z) = n$ for $\pi(z) \in R[z]$, since $\psi(z)$ is monic with degree n, so that the degree of every polynomial in Ann g is strictly greater than $n-1$, resulting in a contradiction. Q.E.D.

Suppose that S is n-cyclic with generator g so that Ann $g = R[z]\psi(z)$ by Theorem 3.5. We shall call $\psi(z)$ the *order* of g and write ord $g = \psi(z)$. We have the following result on the computation of ord g.

PROPOSITION 3.6. *Let $\psi(z) = z^n + \sum_{i=0}^{n-1} a_i z^i$ denote the order of g and let $\alpha = (a_0, a_1, \cdots, a_{n-1})^{\mathrm{TR}}$, where* TR *denotes the transpose. Then $\alpha = -C^{-1}(S^n g)$ where $C = [g, Sg, \cdots, S^{n-1}g]$.*

Proof. By definition of $\psi(z)$ and α, $S^n g = -\sum_{i=0}^{n-1} a_i S^i g = -C\alpha$. Since S is n-cyclic, C is invertible, so $\alpha = -C^{-1}(S^n g)$. Q.E.D.

The next result interconnects n-cyclicity and similarity.

THEOREM 3.7. *Let S: $v \to D(\sigma v)$ and $\bar{S}$: $v \to \bar{D}(\sigma v)$ be s.l.t.'s on R^n, and suppose that S is n-cyclic with generator g. If S and $\bar{S}$ are similar, so that there is an $n \times n$ invertible matrix P with $\bar{D} = P^{-1}D(\sigma P)$, then $\bar{S}$ is n-cyclic with generator $\bar{g} = P^{-1}g$ and* ord $\bar{g}$ = ord g. *Conversely, if $\bar{S}$ is n-cyclic with generator $\bar{g}$ and* ord $\bar{g}$ = ord g, *then there is an $n \times n$ invertible matrix P such that $\bar{D} = P^{-1}D(\sigma P)$ and $\bar{g} = P^{-1}g$.*

Proof. Let S and $\bar{S}$ be s.l.t.'s on R^n and suppose that S is n-cyclic with generator g and ord $g = \psi(z)$. Let R^n_S (resp. $R^n_{\bar{S}}$) denote R^n with the $R[z]$-module structure induced by S (resp. $\bar{S}$). Now S and $\bar{S}$ are similar if and only if R^n_S and $R^n_{\bar{S}}$ are isomorphic as $R[z]$-modules [9]. Since S is n-cyclic, it follows from Theorem 3.5 that R^n_S is isomorphic to the left $R[z]$-quotient module $R[z]/R[z]\psi(z)$. Thus S and $\bar{S}$ are similar if and only if $R^n_{\bar{S}}$ is isomorphic to $R[z]/R[z]\psi(z)$, and this is the case if and only if $\bar{S}$ is n-cyclic with generator $\bar{g}$ for some $\bar{g} \in R^n$ with ord $\bar{g}$ = ord g. Now suppose that R^n_S is isomorphic to $R^n_{\bar{S}}$ and let ϕ: $R^n_{\bar{S}} \to R^n_S$ denote the isomorphism. Then there is an $n \times n$ invertible matrix P over R such that $\phi(v) = Pv$ for all $v \in R^n_{\bar{S}}$, and $P\bar{D} = D(\sigma P)$. Lastly, since R_S^n is n-cyclic with generator g, $R_{\bar{S}}^n$ must be n-cyclic with generator $\bar{g} = \phi^{-1}(g)$, so $\bar{g} = P^{-1}g$. Q.E.D.

As we now show, n-cyclicity of S is equivalent to the existence of a canonical representation (which is very useful in the study of state feedback).

THEOREM 3.8. *Given an n-cyclic s.l.t. S on R^n with generator g and* ord $g = \psi(z) = z^n + \sum_{i=0}^{n-1} a_i z^i$, *let $\bar{S}$ denote the s.l.t. on R^n defined by $\bar{S}(v) = \bar{D}(\sigma v)$ where*

$$\bar{D} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & & 1 \\ -a_0 & -\sigma^{-1}a_1 & -\sigma^{-2}a_2 & \cdots & -\sigma^{-(n-1)}a_{n-1} \end{bmatrix}. \tag{3.1}$$

Let $C = [g, Sg, \cdots, S^{n-1}g]$ and $\bar{C} = [\bar{g}, \bar{S}\bar{g}, \cdots, \bar{S}^{n-1}\bar{g}]$ where $\bar{g} = (0 \quad 0 \cdots 0 \quad 1)^{\mathrm{TR}}$. Then $\bar{S}$ is similar to S with $\bar{D} = P^{-1}D(\sigma P)$ and $\bar{g} = P^{-1}g$, where $P = C(\bar{C})^{-1}$.

Proof. By direct computation, it can be shown that the jth element of $\bar{S}^i\bar{g}$ is 0 for $j = 1, 2, \cdots, n-i-1$ and 1 for $j = n-i$ and that $\bar{S}^n\bar{g} = -\sum_{i=0}^{n-1} a_i\bar{S}^i\bar{g}$. Hence det $\bar{C} = -1$ and $\psi(z)\bar{g} = 0$, so $\bar{S}$ is n-cyclic with generator $\bar{g}$ and ord $\bar{g} = \psi(z)$. The desired result then follows from Theorem 3.7. Q.E.D.

The last result of this section is the converse of Theorem 3.8.

PROPOSITION 3.9. *Let S be an s.l.t. on R^n and suppose that S is similar to $\bar{S}$: $v \to \bar{D}(\sigma v)$ with $\bar{D}$ given by* (3.1) *for some $a_i \in R$. Then S is n-cyclic.*

Proof. Apply Theorem 3.7. Q.E.D.

4. Reachability and state feedback. Let (D, E, H) be a system over R with s.l.t. S: $v \to D(\sigma v)$. In this section we first study reachability in terms of S and the $R[z]$-module structure on R^n induced by S. We then present a new approach to state feedback based on the concept on n-cyclicity. The results obtained below are very similar to results in the algebraic theory of time-invariant systems.

We start with the usual definition of reachability.

DEFINITION 4.1. The system (D, E, H) is *completely reachable* at time t if, for any $x(t) \in K^n$, there is an integer $N > 0$ and inputs $u(t-N)$, $u(t-N+1), \cdots, u(t-1)$ that drive the system from the zero state at time $t-N$ to the state $x(t)$ at time t. If there is a fixed N for all $x(t) \in K^n$, (D, E, H) is completely reachable in N steps at time t. The system (D, E, H) is completely reachable in N steps at *all* times if it is completely reachable in N steps at each $t \in Z$.

Suppose that the system (D, E, H) is completely reachable at time t. Let $x_1, x_2, \cdots, x_n$ be a basis of K^n. Then x_i can be reached in N_i steps for some $N_i > 0$, so any $x \in K^n$ can be reached in $N = \max\{N_i\}$ steps. Thus (D, E, H) is completely reachable at time t if and only if it is completely reachable in N steps at time t for some $N > 0$.

Criteria for reachability, expressed in terms of the system's s.l.t. S and the induced $R[z]$-module structure, are given in the following two propositions.

PROPOSITION 4.2. *Let t be a fixed element of Z. An n-dimensional system (D, E, H) is completely reachable in N steps at time t if and only if* rank $[E, SE, \cdots, S^{N-1}E](t) = n$.

Proof. Given a fixed t and $N > 0$, from (2.4) the state $x(t)$ at time t starting from the zero state at time $t-N$ is equal to $\sum_{i=t-N}^{t-1} (S^{t-i-1}E)(t)u(i)$. This expression defines a map from $(K^m)^N$ into K^n with matrix representation $[E, SE, \cdots, S^{N-1}E](t)$. Obviously, the system is completely reachable in N steps at time t if and only if this map is onto, which is the case if and only if rank $[E, SE, \cdots, S^{N-1}E](t) = n$.

PROPOSITION 4.3. *Write $E = [e_1, \cdots, e_m]$ where e_i is the i-th column of E. There is an integer $N > 0$ such that (D, E, H) is completely reachable in N steps at all times if and only if $R^n = \sum_{i=1}^m R[z]e_i$; that is, the columns of E generate R^n as an $R[z]$-module.*

Proof. It follows from the definition of the $R[z]$-module structure that R^n can be generated from the columns of E if and only if there is an integer $N > 0$ such that $\langle E, SE, \cdots, S^{N-1}E\rangle_R = R^n$. By Proposition 3.1, $\langle E, SE, \cdots, S^{N-1}E\rangle_R = R^n$ if and only if rank $[E, SE, \cdots, S^{N-1}E](t) = n$ for all $t \in Z$, and by Proposition 4.2 the rank condition is equivalent to reachability of the system in N steps at all times. Q.E.D.

COROLLARY 4.4. *Let (D, e, H) be a single-input $(m = 1)$ n-dimensional system. Then (D, e, H) is completely reachable in N steps at all times for some $N > 0$ (resp. completely reachable in n steps at all times) if and only if S is cyclic (resp. n-cyclic) with generator e.*

Given the system (D, E, H) with the dynamical equation $x(t) = D(t)x(t-1) + E(t)u(t-1)$, we now consider state feedback by setting $u(t-1) = -W(t)x(t-1) + r(t-1)$, where W is an $m \times n$ matrix over R, called the feedback matrix, and $r(t)$ is an external input or disturbance. The resulting closed-loop system is given by the triple $(D - EW, E, H)$ which defines the following dynamical equations

$$x(t) = [D(t) - E(t)W(t)]x(t-1) + E(t)r(t-1),$$
$$y(t) = H(t)x(t).$$

We shall let S_W denote the s.l.t. of the closed-loop system, i.e. $S_W(v) = D_W(\sigma v)$ for all $v \in R^n$ where $D_W = D - EW$.

Now suppose that there is a feedback matrix Q over R such that S_Q is n-cyclic with generator $g = Eu$ for some $u \in R^m$. Let $\psi(z) = z^n + \sum_{i=0}^{n-1} a_i z^i$ denote the order of g. We shall show that, given any monic polynomial $\chi(z)$ of degree n, there is a feedback matrix W such that S_W is n-cyclic with generator $g = Eu$ and ord $g = \chi(z)$. In other words, we claim that it is possible to assign the coefficients of the order of g using feedback.

Let $\bar{S}_Q$ denote the s.l.t. on R^n defined by $\bar{S}_Q(v) = \bar{D}_Q(\sigma v)$ where $\bar{D}_Q$ is given by (3.1). Let $C = [g, S_Q g, \cdots, S_Q^{n-1} g]$ and $\bar{C} = [\bar{g}, \bar{S}_Q \bar{g}, \cdots, \bar{S}_Q^{n-1}\bar{g}]$ where $\bar{g} = (0 \quad 0 \cdots 0 \quad 1)^{\mathrm{TR}}$. By Theorem 3.8, $\bar{S}_Q$ is similar to S_Q with $\bar{D}_Q = P^{-1}D_Q(\sigma P)$ and $\bar{g} = P^{-1}g$ where

$$P = C(\bar{C})^{-1}. \tag{4.1}$$

Now pick $\chi(z) = z^n + \sum_{i=0}^{n-1} b_i z^i$, $b_i \in R$, and let

$$\beta = (a_0 - b_0 \quad \sigma^{-1}(a_1 - b_1) \cdots \sigma^{-n+1}(a_{n-1} - b_{n-1})). \tag{4.2}$$

In terms of these constructions, we have the result claimed above.

THEOREM 4.5. *Suppose that S_Q is n-cyclic with generator $g = Eu$. Given $b_0, b_1, \cdots, b_{n-1} \in R$, let $W = Q - u\beta(\sigma P)^{-1}$ where P is given by* (4.1) *and β is given by* (4.2). *Then S_W is n-cyclic with generator $g = Eu$ and* ord $g = \chi(z) = z^n + \sum_{i=0}^{n-1} b_i z^i$.

Proof. Let $W = Q - u\beta(\sigma P)^{-1}$ and define $\bar{D}_W = P^{-1}D_W(\sigma P)$ where $D_W = D - EW$. Then

$$\bar{D}_W = P^{-1}D(\sigma P) - P^{-1}E(Q - u\beta(\sigma P)^{-1})(\sigma P),$$

$$\bar{D}_W = \bar{D}_Q + \bar{g}\beta,$$

$$\bar{D}_W = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -b_0 & -\sigma^{-1}b_1 & -\sigma^{-2}b_2 & \cdots & -\sigma^{-(n-1)}b_{n-1} \end{bmatrix}. \tag{4.3}$$

It follows that the s.l.t. $\bar{S}_W : v \to \bar{D}_W(\sigma v)$ is n-cyclic with generator $\bar{g}$ and ord $\bar{g} = \chi(z)$. Since S_W and $\bar{S}_W$ are similar by construction, it follows from Theorem 3.7 that S_W is n-cyclic with generator $P\bar{g} = g$ and ord $g =$ ord $\bar{g} = \chi(z)$. Q.E.D.

The above result resembles coefficient assignability of the characteristic polynomial in the theory of time-invariant systems. In fact, as in the time-invariant case, assignability implies that we can specify (up to an invertible matrix) the free response of the closed-loop system:

THEOREM 4.6. *Suppose that there is a feedback matrix Q over R such that S_Q is n-cyclic with generator $g = Eu$ and* ord $g = z^n + \sum_{i=0}^{n-1} a_i z^i$. *Given $b_0, b_1, \cdots, b_{n-1} \in R$, let $W = Q - u\beta(\sigma P)^{-1}$, where P is given by* (4.1) *and β is given by* (4.2). *Then the state $x(t)$ of the closed-loop system $(D - EW, E, H)$ resulting from initial state $x(t_0) \in K^n$, with zero input for $t \geqq t_0$, is given by*

$$x(t) = P(t)\begin{bmatrix} \gamma(t-n+1) \\ \gamma(t-n+2) \\ \vdots \\ \gamma(t) \end{bmatrix}$$

where $\gamma(t)$ is the solution of the n-th-order difference equation

$$\gamma(t) + \sum_{i=0}^{n-1} b_i(t+i)\gamma(t-n+i) = 0, \qquad t > t_0,$$

with initial data $\gamma(t_0 - n + j) = j$-th component of $P^{-1}(t_0)x(t_0)$ for $j = 1, 2, \cdots, n$.

Proof. Let $\bar{x}(t)$ denote the state of the system $(\bar{D}_W, \bar{E}, H)$ where $\bar{D}_W$ is given by (4.3). The theorem follows from the form of $\bar{D}_W$ and the relation $x(t) = P(t)\bar{x}(t)$. Q.E.D.

As a special case of the above result, note that if we set $b_i = 0$ for all i, then $\gamma(t) = 0$ for $t > t_0$, so the free response $x(t)$ is zero for all $t > t_0 + n$. This is sometimes referred to as "dead-beat control".

In the remainder of this section, we shall consider conditions under which there exists a feedback matrix Q such that S_Q is n-cyclic with generator $g = Eu$. We begin with the single-input case $(E = e)$.

PROPOSITION 4.7. *Given the single-input system (D, e, H), there is a Q over R such that S_Q is n-cyclic with generator $g = eu$ for some $u \in R$ if and only if (D, e, H) is completely reachable in n steps at all times, in which case Q can be taken to be zero and u can be taken to be the unit function.*

Proof. By multilinearity of the determinant function, it follows that S_Q is n-cyclic with generator $g = eu$ if and only if S is n-cyclic with generator e. By Corollary 4.4, the latter condition is equivalent to complete reachability of (D, e, H) in n steps at all times. Q.E.D.

One might expect that complete reachability in n steps at all times is also a necessary and sufficient condition in the multi-input case. Reachability is necessary (Corollary 4.10), but it is *not* sufficient as seen from the following example.

Example 4.8. Let $D(t)=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$ and $E(t)=\begin{bmatrix}t & 0\\ 0 & t\end{bmatrix}$. Then

$$[E, SE](t)=\begin{bmatrix}t & 0 & t-1 & 0\\ 0 & t & 0 & t-1\end{bmatrix}$$

which has rank 2 for all $t\in Z$. Thus the system is completely reachable in $n=2$ steps at all times. Now let Q be a 2×2 matrix over R and let $g=Eu$, $u\in R^2$. Then $g=tu$ and $\det[g, S_Qg]=t(\det[u, S_Qg])$. Hence $\langle g, S_Qg\rangle_R\neq R^2$, so S_Q is not n-cyclic.

The next result gives a necessary and sufficient condition for the existence of an n-cyclic S_Q.

THEOREM 4.9. *There is a Q over R such that S_Q is n-cyclic with generator $g=Eu$ for some $u\in R^m$ if and only if there exist elements $u_0, u_1, \cdots, u_{n-1}\in R^m$ such that $\langle g_0, g_1, \cdots, g_{n-1}\rangle_R=R^n$ where $g_0=Eu_0$ and $g_i=D(\sigma g_{i-1})+Eu_i$ for $i=1, 2, \cdots, n-1$.*

Proof. Suppose that there exist Q and $g=Eu$ such that $\langle g, S_Qg, \cdots, S_Q^{n-1}g\rangle_R=R^n$. Let $u_0=u$ and $u_i=-Q(\sigma g_{i-1})$ for $i=1, 2, \cdots, n-1$, where $g_0=Eu_0$ and $g_i=D(\sigma g_{i-1})+Eu_i$. Then $g_i=D(\sigma g_{i-1})-EQ(\sigma g_{i-1})=(D-EQ)(\sigma g_{i-1})$ for $i=1, 2, \cdots, n-1$. Thus $g_i=S_Q^ig$ for $i=0, 1, 2, \cdots, n-1$, so $\langle g_0, g_1, \cdots, g_{n-1}\rangle_R=R^n$. Conversely, suppose that there exist $u_0, u_1, \cdots, u_{n-1}\in R^m$ such that $\langle g_0, g_1, \cdots, g_{n-1}\rangle_R=R^n$ where $g_0=Eu_0$ and $g_i=D(\sigma g_{i-1})+Eu_i$. Let $Q=-[u_1, u_2, \cdots, u_{n-1}, 0][\sigma g_0, \sigma g_1, \cdots, \sigma g_{n-1}]^{-1}$. It will be shown that $\langle g_0, S_Q, g_0, \cdots, S_Q^{n-1}g_0\rangle_R=R^n$: By definition of Q, $Q(\sigma g_{i-1})=-u_i$ for $i=1, 2, \cdots, n-1$. Then for $i=0, 1, 2, \cdots, n-2$, $S_Qg_i=(D-EQ)(\sigma g_i)=D(\sigma g_i)-EQ(\sigma g_i)=D(\sigma g_i)+Eu_{i+1}=g_{i+1}$. Thus $S_Q^ig_0=g_i$ for $i=0, 1, 2, \cdots, n-1$, which proves the claim. Q.E.D.

COROLLARY 4.10. *Given the system (D, E, H), there is a Q over R such that S_Q is n-cyclic with generator $g=Eu$ for some $u\in R^m$ only if the system is completely reachable in n steps at all times.*

Proof. Suppose that there exist $u_0, u_1, \cdots, u_{n-1}$ such that $\langle g_0, g_1, \cdots, g_{n-1}\rangle_R=R^n$ where $g_0=Eu_0$ and $g_i=D(\sigma g_{i-1})+Eu_i$. By definition, g_i is an R-linear combination of the columns of $E, SE, \cdots, S^iE$. Then since the g_i generate R^n, the columns of $E, SE, \cdots, S^{n-1}E$ must generate R^n. By Proposition 4.2, the latter condition implies that the system is completely reachable in n steps at all times. Q.E.D.

It is interesting to note that Heymann's result [13] on the existence of cyclic transformations in the control theory of time-invariant systems resembles the result given in Theorem 4.9 with $u_i\in\{0, s_1, s_2, \cdots, s_m\}$ where $s_j=(0\cdots 0\quad 1\quad 0\cdots 0)^{\mathrm{TR}}\in R^m$ with 1 in the jth position. However, here the constraint that the u_i belong to $\{0, s_1, \cdots, s_m\}$ is too severe.

A procedure is given below for determining a set of elements $u_0, u_1, \cdots, u_{n-1}, g_0, g_1, \cdots, g_{n-1}$ satisfying the condition in Theorem 4.9 (if such a set exists).

Given the system (D, E, H), first compute a $\bar{u}_0\in R^m$ such that $\bar{g}_0(t)=E(t)\bar{u}_0(t)\neq 0$ for all $t\in Z$. Such a $\bar{u}_0$ must exist if $\bar{g}_0=E\bar{u}_0$ is to be a basis element of R^n. Now for some fixed $i\in\{2, 3, \cdots, n-1\}$, suppose that $\bar{u}_0, \bar{u}_1, \cdots, \bar{u}_{i-1}\in R^m$ have been found such that $\operatorname{rank}[\bar{g}_0(t), \bar{g}_1(t), \cdots, \bar{g}_{i-1}(t)]=i$ for all $t\in Z$, where $\bar{g}_0=E\bar{u}_0$ and $\bar{g}_i=$

$D(\sigma\bar{g}_{j-1})+E\bar{u}_j$ for $j=1,2,\cdots,i-1$. We shall attempt to find elements $u_0, u_1, \cdots, u_{i-1}, u_i \in R^m$ such that rank $[g_0(t),\cdots,g_{i-1}(t),g_i(t)]=i+1$ for all $t\in Z$, where $g_0=Eu_0$ and $g_j=D(\sigma g_{j-1})+Eu_j$ for $j=1,2,\cdots,i$. Let T denote the subset of Z consisting of all t such that rank $[\bar{g}_0(t),\cdots,\bar{g}_{i-1}(t),S(\bar{g}_{i-1})(t)]=i$. There are three cases to consider:

Case 3. If J is not empty, find elements $\varphi,\omega\in R^m$ such that rank so that $g_i=S(\bar{g}_{i-1})$ and $g_j=\bar{g}_j$ for $j=0,1,\cdots,i-1$.

Case 2. Suppose that T is not empty. Let J denote the subset of T consisting of all t for which there is no $\lambda_t\in K^m$ such that rank $[\bar{g}_0(t),\cdots,\bar{g}_{i-1}(t),S(\bar{g}_{i-1})(t)+E(t)\lambda_t]=i+1$. If J is the empty set, we can take $u_i(t)=\lambda_t$, $t\in T$, $u_i(t)=0$, $t\in Z-T$ and $u_j=\bar{u}_j$ for $j=0,1,\cdots,i-1$. Thus $g_i=S(\bar{g}_{i-1})+Eu_i$ and $g_j=\bar{g}_j$ for $j=0,1,\cdots,i-1$.

Case 3. If J is not empty, find elements $\varphi,\omega\in R^m$ such that rank $[\bar{g}_0(t),\cdots,\bar{g}_{i-2}(t),\bar{g}_{i-1}(t)+E(t)\varphi(t),S(\bar{g}_{i-1}+E\varphi)(t)+E(t)\omega(t)]=i+1$ for all $t\in Z$. Then we can take $u_{i-1}=\bar{u}_{i-1}+\varphi$, $u_i=\omega$, and $u_j=\bar{u}_j$ for $j=0,1,\cdots,i-2$. If no such φ and ω exist, set $u_{i-2}=\bar{u}_{i-2}+\eta$, $u_{i-1}=\bar{u}_{i-1}+\varphi$, and $u_i=\omega$ for some η, φ, ω belonging to R^m, and so on.

This procedure works well if Cases 1 and 2 are the only ones that arise in the computation of a solution. If it is necessary to consider Case 3, the "rate" at which the procedure converges to a solution (assuming one exists) depends on the manner in which φ, ω, etc. are selected. This issue will be considered in a separate paper, as the details appear to require an extensive amount of development.

Example 4.11. Consider the system (D,E,H) where

$$D(t)=\begin{bmatrix} t-1 & 0 & 0\\ -1 & 1 & 0\\ 0 & t & -1\end{bmatrix} \quad\text{and}\quad E(t)\begin{bmatrix}1 & 0\\ 0 & t\\ 0 & -t\end{bmatrix}.$$

Take $\bar{u}_0=\begin{bmatrix}1\\0\end{bmatrix}$, so that

$$\bar{g}_0=\begin{bmatrix}1\\0\\0\end{bmatrix} \quad\text{and}\quad [\bar{g}_0,S(\bar{g}_0)]=\begin{bmatrix}1 & t-1\\ 0 & -1\\ 0 & 0\end{bmatrix}.$$

The rank of $[\bar{g}_0(t),S(\bar{g}_0)(t)]$ is equal to 2 for all $t\in Z$, so we can take $\bar{u}_1=0$ and $\bar{g}_1=S(\bar{g}_0)$. Then

$$[\bar{g}_0(t),\bar{g}_1(t),S(\bar{g}_1)(t)]=\begin{bmatrix}1 & t-1 & t^2-3t+2\\ 0 & -1 & -t+1\\ 0 & 0 & -t\end{bmatrix}.$$

The rank of this matrix is 3 for all $t\in Z$ except $t=0$. Thus we need to find φ, $\omega\in R^2$ such that rank $[\bar{g}_0(t),\bar{g}_1(t)+E(t)\varphi(t),S(\bar{g}_1+E\varphi)(t)+E(t)\omega(t)]=3$ for all $t\in Z$. This is satisfied with $\varphi(t)=\omega(t)=\begin{bmatrix}0\\1\end{bmatrix}$. We then have the following elements which satisfy the condition in Theorem 4.9:

$$u_0=\bar{u}_0=\begin{bmatrix}1\\0\end{bmatrix},\quad u_1=\bar{u}_1+\varphi=\begin{bmatrix}0\\1\end{bmatrix},\quad u_2=\omega=\begin{bmatrix}0\\1\end{bmatrix},\quad g_0=\bar{g}_0=\begin{bmatrix}1\\0\\0\end{bmatrix},$$

$$g_1=\bar{g}_1+E\varphi=\begin{bmatrix}t-1\\t-1\\-t\end{bmatrix},\qquad g_2=S(\bar{g}_1+E\varphi)+E\omega=\begin{bmatrix}t^2-3t+2\\t\\t^2-2t-1\end{bmatrix}.$$

In contrast to the approach given above, assignability-type results for time-varying systems can be obtained by requiring that the given system with feedback be equivalent to a pole-assignable time-invariant system (the output matrix may be time-varying). For time-varying continuous-time systems, this approach is presented in [8] where it is required that the given system be index invariant. In the discrete-time case, a system (D, E, H) is index invariant if, for each $i=1,2,\cdots,n+1$, rank $[E, SE,\cdots, S^{i-1}E(t)=q_i=$ constant for all $t\in Z$ and $q_n=q_{n+1}$. Clearly, index invariance is a very restrictive requirement, which is not necessary for the existence of an n-cyclic S_Q with generator $g=Eu$ (for instance, the system in Example 4.11 is not index invariant).

REFERENCES

[1] R. E. KALMAN, *Algebraic structure of linear dynamical systems. I. The module of* Σ, Proc. Nat. Acad. Sci. U.S.A., 54 (1965), pp. 1503–1508.

[2] R. W. NEWCOMB, *A local time-variable synthesis*, Fourth Colloquium on Microwave Communications, Budapest, 1970.

[3] E. W. KAMEN, *Representation and realization of operational differential equations with time-varying coefficients*, J. Franklin Institute, 301 (1976), pp. 559–571.

[4] S. SALOVAARA AND H. BLOMBERG, *On an algebraic theory of ordinary linear time-varying differential systems with generalized stochastic processes as inputs and outputs*, Advances in Cybernetics and Systems Research, vol. I, F. Pichler and R. Trappl eds., Transcripta Books, London, 1973.

[5] R. YLINEN, *On the algebraic theory of linear differential and difference systems with time-varying or operator coefficients*, Report B 23, Helsinki University of Technology, Helsinki, Finland, 1975.

[6] E. W. KAMEN, *A new algebraic approach to linear time-varying systems*, Technical Report, Georgia Institute of Technology, Atlanta, 1974.

[7] L. M. SILVERMAN, *Transformation of time-variable systems to canonical (phase variable) form*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 300–303.

[8] A. S. MORSE AND L. M. SILVERMAN, *Structure of index-invariant systems*, this Journal, 11 (1973), pp. 215–225.

[9] N. JACOBSON, *Pseudo-linear transformations*, Ann. of Math., 38 (1937), pp. 484–507.

[10] P. M. COHN, *Free Rings and Their Relations*, Academic Press, London, 1971.

[11] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.

[12] S. LANG, *Algebra*, Addison-Wesley, Reading, MA, 1965.

[13] M. HEYMANN, *Comments on pole assignment in multi-input controllable linear systems*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 748–749.

[14] K. M. HAFEZ, *New results on discrete-time time-varying linear systems*, Ph.D. dissertation, Georgia Institute of Technology, Atlanta, 1975.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 4, July 1979

0363-0129/79/1704-0005$01.00/0

EXISTENCE OF OPTIMAL CONTROLS FOR STOCHASTIC JUMP PROCESSES*

C. B. WAN† AND M. H. A. DAVIS†

Abstract. Sufficient conditions are given for existence of an optimal control policy for a class of controlled jump processes. The processes are specified by a family of "local descriptions" depending on a control which is a function of the complete past of the process. Conditions for optimality were given in a previous paper [M. H. A. Davis and R. J. Elliott, *Optimal control of a jump process*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 40 (1977), pp. 183–202]. Here it is shown that, under fairly stringent conditions on the form of the local descriptions, an optimal policy can be constructed as long as a certain "Hamiltonian" function can be minimized.

1. Introduction. In this paper sufficient conditions are given for the existence of an optimal control policy for a controlled family of stochastic jump processes. This work is a continuation of that started in [8]. The formulations of the controlled jump process and the optimal control problem will be essentially the same as there.

The jump process x_t is first specified under a basic probability measure P to which corresponds a pair of entities (Λ, λ) called the local descriptions of the process; Λ determines the rate of occurrence of jumps while λ determines their positions. By using an indexed pair of Radon–Nikodym derivatives (α^u, β^u), we achieve control of the jump process x_t through mutually absolutely continuous transformation of the local descriptions from (Λ, λ) to (Λ^u, λ^u).

The control policy u will be assumed to be nonanticipative and to depend on the complete information of the past of the process. The optimal control problem is then to select a control policy so as to minimize a cost function of the form:

$$J(u) = E_u\left\{\int_0^{T_f} c(s, u_s, \omega)\, d\Lambda_s + G_f\right\}. \tag{1.1}$$

(See Definition 3.2.)

Necessary and sufficient Hamilton–Jacobi type conditions for optimality for the above problem can be found in [15] and also in [8] for a slightly different cost structure. Here we use these results to construct an optimal control policy, using an approach similar to that developed by Davis [6] for controlled processes described by stochastic differential equations. For that problem an alternative method, used by Beneš [2] and Duncan and Varaiya [10], is to prove compactness of the set of densities corresponding to admissible controls. Such an approach cannot generally be used here; see § 5 below for further comments on this point.

In § 2 we give the mathematical formulation of the jump process and some related results; this is a summary of material in [5] and [8]. In § 3 we formulate the control problem and give an example (Example 3.1, to which the reader may refer immediately) which illustrates the type of problem to which our formulation applies. The main result, stating conditions for existence of an optimal policy, is Theorem 4.2 in § 4. This depends on some technical results on compactness of certain sets of densities; these are collected in an Appendix.

2. Outline of mathematical description of jump process and some related results. The jump process formulation is very similar to that in [5], [8] but we include a brief outline here in order to establish notation and to make this paper more self-contained.

* Received by the editors January 31, 1978, and in revised form July 7, 1978.

† Department of Computing and Control, Imperial College, London, SW7 2BZ.

We consider piecewise-constant processes $\{x_t\}$ taking values in a Blackwell space $(X, \mathcal{S})$ and having isolated discontinuities. Such a process is defined by a countable family $\{S_i, Z_i, i = 1, 2 \cdots\}$ of random variables, where $\{S_i\}$ are the "interarrival times" and $\{Z_i\}$ the "states". Formally, let z_0 be a fixed element of X, define

$$(Y, \mathcal{Y}) = ((R^+ \times X) \cup \{(\infty, z_\infty)\}, \sigma\{\underline{B}(R^+) * \mathcal{S}, \{\infty, z_\infty\}\}),$$

and let $(Y^i, \mathcal{Y}^i)$ be a copy of $(Y, \bar{\mathcal{Y}})$ for $i = 1, 2, \cdots$.

Then the basic measurable space is $(\Omega, \mathcal{F}^0)$ where

$$\Omega = \prod_{i=1}^{\infty} Y^i, \qquad \mathcal{F}^0 = \sigma\left\{\prod_{i=1}^{\infty} \mathcal{Y}^i\right\}.$$

Let $(S_i, Z_i) : \Omega \to Y^i$ be the coordinate mapping and $\omega_k : \Omega_k = \prod_{i=1}^k Y^i$ be the projection onto Ω_k, i.e.

$$\omega_k(\omega) = (S_1(\omega), Z_1(\omega) \cdots S_k(\omega), Z_k(\omega)).$$

We also define $Z_0(\omega) = z_0$ (another fixed element of X). Now let

$$T_o(\omega) = 0,$$

$$T_k(\omega) = \sum_{i=1}^{k} S_i(\omega),$$

$$T_\infty(\omega) = \lim_{k \to \infty} T_k(\omega).$$

The sample path of the process $\{x_t\}$ is given by

$$x_t(\omega) = \begin{cases} Z_i(\omega), & t \in [T_i(\omega), T_{i+1}(\omega)[, \\ z_\infty, & t \geqq T_\infty(\omega), \end{cases}$$

A measure P on $(\Omega, \mathcal{F}^0)$ is given in the following way: The interarrival times $\{S_k\}$ are independent with survivor functions $F_t^k = P[S_k > t]$ determined by the corresponding "integrated rates" which are functions Λ^k: $[0, d^k[\to R^+ (0 < d^k \leqq \infty)$ satisfying:

(i) $\Lambda^k(0) = 0$; $\Lambda^k(\cdot)$ is increasing and right-continuous;

(ii) $\Lambda^k(t) \uparrow \infty$ as $t \uparrow \infty$ if $d^k = \infty$;

(iii) $\Delta\Lambda^k(s) < 1$ where $\Delta\Lambda^k(s) = \Lambda^k(s) - \Lambda^k(s-)$;

(iv) There exist constants $\theta_1, \theta_2 > 0$ such that

$\Lambda^k(t) \leqq \theta_2$ for $t \in [0, \theta_1]$ and $k \in \mathcal{K}$, where

$\mathcal{K}$ is an infinite subset of the integers $1, 2, \cdots$.

In terms of Λ^k, F^k is given by

$$F_t^k = \exp(-\Lambda^{kc}(t)) \prod_{s \leqq t} (1 - \Delta\Lambda^k(s)). \tag{2.1}$$

(The product is over the countable set $\{s \leqq t : \Delta\Lambda^k(s) \neq 0\}$, and $\Lambda^{kc}(t) = \Lambda^k(t) - \sum_{s \leqq t} \Delta\Lambda^k(s)$.)

This formulation enables continuous- and discrete-time models to be handled simultaneously; for example if $\Lambda^k(t) = t$ then the T_k sequence forms a Poisson process, whereas if $\Lambda^k(t) = \frac{1}{2}$[integer part t] then the S_k's can only take integer values. Note that

$$d^k = \inf\{t : F_t^k = 0\}.$$

Condition (iv) is introduced to ensure that the T_k's do not accumulate at any finite time, as the following result shows.

LEMMA 2.1.

$$T_\infty = \infty \quad a.s.$$

Proof. For $t \leqq \theta_1$ and $k \in \mathcal{K}$, $\sum_{s \leqq t} \Delta\Lambda^k(s) \leqq \Delta^k(t) \leqq \theta_2$ and hence

$$\prod_{s \leqq t} (1 - \Delta\Lambda^k(s)) > 0.$$

Thus for $k \in \mathcal{K}$ there exists a constant $\theta_3 > 0$ such that $F_t^k \geqq \theta_3$ for $t \leqq \theta_1$. Let $A_k = \{\omega : S_k > \theta_1\}$. Then $PA_k = F_{\theta_1}^k$ and hence $\sum_k PA_k = \infty$. Since the A_k are independent, the Borel zero-one law gives $P[\limsup A_k] = 1$ and this implies that $T_\infty \to \infty$ a.s.

The specification of P is completed by giving a family of functions $\lambda^k : \Omega_{k-1} \times R^+ \times \mathcal{S} \to [0, 1]$ satisfying the following conditions:

(i) $\lambda^k(\cdot, \cdot, A)$ is measurable for each $A \in \mathcal{S}$;

(ii) $\lambda^k(\omega_{k-1}, t, \cdot)$ is a probability measure on $\mathcal{S}$ for each $(\omega_{k-1}, t) \in \Omega_{k-1} \times]0, d^k]$ such that $\lambda(\omega_{k-1}, t, \{Z_{k-1}(\omega)\}) = 0$.

λ^k specifies the conditional distribution of Z_k given the past. Formally, the measure P is defined recursively by setting

$$P[S_k > t, Z_k \in A | F_{T_{k-1}}^0] = -\int_{]t,\infty]} \lambda_k(\omega_{k-1}, s, A)\, dF_s^k.$$

According to [5], this procedure defines uniquely a measure P on $(\Omega, \mathcal{F}^0)$. Now let $\mathcal{F}_t$ be $\mathcal{F}_t^0$ completed with all P-null sets of $\mathcal{F}^0$.

The fundamental family of martingales associated with the jump process $\{x_t\}$ is given as follows. For $A \in \mathcal{S}$, $t \geqq 0$ let

$$p(t, A) = \sum_i I_{(t \geqq T_i)} I_{(Z_i \in A)}.$$

Now define

$$\Lambda_t(\omega) = \Lambda^1(S_1) + \Lambda^2(S_2) + \cdots + \Lambda^{k-1}(S_{k-1}) + \Lambda^k(t - T_{k-1}) \quad \text{for } t \in]T_{k-1}, T_k]$$

and

$$\lambda(t, A)(\omega) = \sum_{k=1}^{\infty} I_{(t \in]T_{k-1}, T_k])} \lambda^k(\omega_{k-1}; t - T_{k-1}, A).$$

The pair (Λ, λ) is called the "local description" of the process $\{x_t\}$. Now let

$$\tilde{p}(t, A) = \int_{]0,t]} \lambda(t, A)\, d\Lambda_t.$$

Then

$$q(t, A) = p(t, A) - \tilde{p}(t, A)$$

is a local martingale of $\mathcal{F}_t$ and the martingale representation theorem [5, Thm. 2] states that any local martingale $\{M_t\}$ of $\mathcal{F}_t$ is of the form

$$M_t = \int_{]0,t] \times X} g(s, z, \omega) q(ds, dz)$$

$$= \int_{]0,t] \times X} g(s, z, \omega) p(ds\, dz) - \int_{]0,t] \times X} g(s, z, \omega)\, \tilde{p}(ds, dz)$$

where the integrals are Stieltjes integrals in the sample paths and the integrand g is a predictable process belonging to $L_1^{\text{loc}}(p)$ (see [5]).

3. Formulation of the control problem and some previous results. The method of control of the jump process is through absolutely continuous transformations of the local description from the basic pair to a pair (Λ^u, λ^u), where $u \in \mathcal{U}$ is the control variable and $\mathcal{U}$ is the class of admissible controls. This transformation is achieved through a pair of controlled Radon–Nikodym derivatives (α^u, β^u), where $\alpha^u = d\Lambda^u/d\Lambda$, $\beta^u = d\lambda^u/d\lambda$. To render this precise, we first give some definitions. Let $(U, \mathcal{B}_u)$ be a measurable space.

DEFINITION 3.1. The class of admissible controls $\mathcal{U}$ is the set of $\mathcal{F}_t$-predictable processes with values in U.

Now let $\alpha : R^+ \times U \times \Omega \to R^+$ and $\beta : R^+ \times X \times U \times \Omega \to R^+$ be functions measurable with respect to the appropriate product σ-fields and satisfying the following conditions:

(i) For each $(x, u) \in X \times U$, $\alpha(t, u, \omega)$ and $\beta(t, x, u, \omega)$ are $\mathcal{F}_t$-predictable processes;

(ii) There exist positive constants c_1, c_2, c_3, δ' such that

$$0 < c_1 \leqq \alpha(t, u, \omega) \leqq \min\left(c_2, \frac{1-\delta'}{\Delta\Lambda_t}\right),\tag{3.1}$$
$$0 < c_1 \leqq \beta(t, x, u, \omega) \leqq c_3$$

for all $(t, x, u, \omega) \in R^+ \times X \times U \times \Omega$;

(iii) $\int_X \beta(t, x, u, \omega)\lambda(dx, t, \omega) = 1$ for all $(t, u, \omega) \in R^+ \times U \times \Omega$.

As is customary, the ω-dependence of α and β will often be suppressed.

For each $u \in \mathcal{U}$, let α^u, β^u be the functions defined by

$$\alpha^u(t, \omega) = \alpha(t, u(t, \omega), \omega),$$

$$\beta^u(t, x, \omega) = \beta(t, x, u(t, \omega), \omega).$$

In view of (3.1) (i) above and the structure of $\mathcal{F}_t$-predictable processes, there exist, for each $k \in Z^+$ and $u \in \mathcal{U}$, functions $\alpha_k^u : \Omega_{k-1} \times R^+ \to R$ and $\beta_k^u : \Omega_{k-1} \times X \times R^+ \to R$ such that

$$\alpha^u(t, \omega) = \sum_k \alpha_k^u(\omega_{k-1}, t - T_{k-1}(\omega))I_{(t \in]T_{k-1}, T_k])},\tag{3.2}$$

$$\beta^u(t, x, \omega) = \sum_k \beta_k^u(\omega_{k-1}, x, t - T_{k-1}(\omega))I_{(t \in]T_{k-1}, T_k])}.\tag{3.3}$$

For a given $u \in \mathcal{U}$, a measure P_u is defined on $(\Omega, \mathcal{F}_\infty)$ by giving the Radon–Nikodym derivative of its restriction to $\mathcal{F}_{T_N}$, $N = 1, 2, \cdots$, as

$$\left.\frac{dP_u}{dP}\right|_{\mathcal{F}_{T_N}} = \prod_{k=1}^{N} L_k(\omega)\tag{3.4}$$

where

$$L_k(\omega_k) = \alpha_k^u(\omega_{k-1}, S_k)\beta_k^u(\omega_{k-1}, Z_k, S_k)$$

and

$$\cdot \exp\left\{-\int_{]0, S_k]} (\alpha_k^u(\omega_{k-1}, s) - 1)\, d\Lambda^{kc}(s)\right\}\pi_{S_k}^{k-} I_{(S_k \leqq d^k)}$$

$$\pi_t^k = \prod_{s \leqq t} \frac{(1 - \alpha_k^u(\omega_{k-1}, s)\Delta\Lambda^k(s))}{(1 - \Delta\Lambda^k(s))}.$$

It is shown in [8] that P_u is the probability measure corresponding to a jump process whose local description (Λ^u, λ^u) is related to that of P by

$$\Lambda_t^u = \int_{]0,t]} \alpha^u(s)\, d\Lambda_s, \qquad \lambda^u(t, A) = \int_{]0,t]} \beta^u(t, x)\lambda(t, dx). \tag{3.5}$$

The fundamental family of martingales of the jump process under measure P_u is $\{q^u(t, A) \vdots A \in \mathcal{S}\}$ where

$$q^u(t, A) = p(t, A) - \tilde{p}^u(p, A) \quad \text{and} \quad \tilde{p}^u(t, A) = \int_X \lambda^u(t, A)\, d\Lambda_t^u.$$

Due to the bounds (3.1) (ii), P and P_u are mutually absolutely continuous on $\mathcal{F}_{T_N}$ for each N. It is shown in Lemma A.1 below that $P_u[T_\infty = \infty] = 1$ and it follows that P, P_u are mutually absolutely continuous on $\mathcal{F}_t$ for fixed $t > 0$, since any set $A \in \mathcal{F}_t$ differs from $\bigcup_k A \cap B_k$ by a P_u-null set, where $B_k = (T_{k-1} \leqq t < T_k) \in \mathcal{F}_{T_k}$.

In the proof of the main theorem (Theorem 4.2) we need the following technical result. It follows directly from some compactness properties of densities (see Lemma A.2; details are relegated to the Appendix since they are messy and the results are of only subsidiary interest.)

LEMMA 3.1. *For $u \in \mathcal{U}$ and $N \in Z^+$ let $L^{(N)}(u)$ denote the restriction of dP_u/dP to $\mathcal{F}_{T_N}$ as given by* (3.4) *above. (Then $L^{(N)}(u) \in L_1(\Omega, \mathcal{F}_{T_N}, P)$). If $\{u_n\}$ is any sequence of admissible controls then there is a subsequence $\{u_{n_k}\}$ and an element ρ of $L_1(\Omega, \mathcal{F}_{T_N}, P\}$ such that $\rho > 0$ a.s. and $L^{(N)}(u_{n_k}) \to \rho$ weakly in $L_1(\Omega, \mathcal{F}_{T_N}, P)$ as $k \to \infty$.*

We now define the cost structure of the control problem, which takes place on a finite interval $[0, T_f]$.

DEFINITION 3.2. The *cost rate* is a function $c : [0, T_f] \times U \times \Omega \to R^+$ satisfying:

(i) $c(t, u, \omega)$ is an $\mathcal{F}_t$-predictable function of (t, ω) for each $u \in U$;

(ii) There is a constant c_4 such that

$$0 \leqq c(t, u, \omega) \leqq c_4 \quad \text{for all } (t, u, \omega) \in [0, T_f] \times U \times \Omega.$$

The *terminal cost* G_f is a nonnegative $\mathcal{F}_{T_f}$-measurable random variable, also bounded by c_4. The *cost* corresponding to $u \in \mathcal{U}$ is then

$$J(u) = E_u\left\{\int_{]0,T_f]} c(s, u_s, \omega)\, d\Lambda(s, \omega) + G_f(\omega)\right\}. \tag{3.6}$$

The optimal control problem is to find a control $u \in \mathcal{U}$ such that

$$J(u^*) = J^* = \inf_{u \in \mathcal{U}} J(u). \tag{3.7}$$

In [3] the integral part of the cost function was of the form

$$E_u \int_{]0,T_f] \times X} \kappa(x, s, u_s)\tilde{p}^u(ds, dx) \tag{3.8}$$

where $\tilde{p}^u(t, A)$ is the compensator for $p(t, A)$ under measure P_u, as defined above. Expression (3.8) is equal to

$$E_u \int_{]0,T_f]} \left(\int_X \kappa(x, s, u_s)\beta^u(s, x)\lambda(s, dx)\right)\alpha^u(s)\, d\Lambda_s$$

and is thus included in our framework (3.6), taking

$$c(s, u, \omega) = \alpha^u(s, \omega) \int_X \kappa(x, s, u, \omega)\beta^u(s, x, \omega)\lambda(s, dx, \omega).$$

Before proceeding with the formal development, let us consider the following example which illustrates the type of control problem which can be formulated in the above framework.

Example 3.1 (The market trader's problem). A market trader starts the day with a stock of N slightly over-ripe pineapples which, if not sold by the end of the day, must be thrown away. He can vary the price continually throughout the day and this will affect the number of customers and their buying patterns. How should he set the price so as to maximize his revenue?

Here the jump process $\{x_t\}$ is the number of items in stock and the control is the price per item $u_t \in [0, \bar{u}]$ where $\bar{u}$ is a maximum reasonable price. We assume that the arrival of customers at the stall is a point process with rate $l(t, u_t)$ which in an simple model might be of the form $l(t)(1-\phi(u_t))$ where $l(t)$ is the rate of arrival of potential customers and $\phi(u)$ the fraction who are turned off by a price of u. The customers are assumed to buy at most M items ($M \ll N$), their propensity to buy being measured by a price-dependent probability distribution $q_1(u) \cdots q_M(u)$ on $\{1, \cdots, M\}$, $q_i(u)$ being the probability of buying i items. This will have various obvious properties: for example the average purchase $\sum iq_i(u)$ should be decreasing with increasing u. There is a disposal cost $d(x)$ for x left over items (for example, $d(x) = dI_{(x>0)}$ where d is the cost of a trip to the dump).

In terms of our abstract model, we take $X = \{0, \cdots, N\}$ and $z^0 = N$. For $Z_{k-l}(\omega) \geqq 1$ we define $\Lambda^k(t) = t$ and

$$\lambda^k(\omega_{k-1}, t, \{i\}) = 1/M, \qquad i = Z_{i-1}-1, Z_{i-1}-2, \cdots, (1 \vee (Z_{i-1}-M)),$$

$$\lambda^k(\omega_{k-1}, t, \{0\}) = \begin{cases} (M - Z_{k-1}+1)/M, & M \geqq Z_{k-1}, \\ 0, & M < Z_{k-1}. \end{cases}$$

For $Z_{k-1} = 0$ we take $\Lambda^k(t) = 0$ and λ^k arbitrary (This means that Λ^k actually depends on ω_{k-1}, which was not allowed for earlier; but the lack of ω_{k-1}-dependence in the general model is only used in Lemma 2.1 whose conclusion holds here anyway.)

Now let

$$\alpha(t, u, \omega) = l(t, u),$$

$$\beta(t, i, u, \omega) = Mq_{(x_{t-}-i)}(u), \qquad i = x_{t-}-1, \cdots, ((x_{t-}-M) \vee 1),$$

$$\beta(t, 0, u, \omega) = \frac{M}{M - x_{t-}+1}\left(\sum_{j=x_{t-}}^{M} q_j(u)\right) I_{(x_{t-} \leqq M)}.$$

These satisfy conditions (3.1) as long as $l(t, u)$ is bounded and $l(t, u), q_i(u) \geqq c_1$ for some $c_1 > 0$. These are not unreasonable assumptions. Then under measure P_u constructed as above the jump rate is $l(t, u_t)$ and the jump distribution corresponds to a demand distribution $\{q_i(u)\}$ as it should. The trader's gross revenue for the day's trading of T_f hours is then

$$R(u) = \sum_{\substack{s \leqq T_f \\ x_s \neq x_{s-}}} -u_s \Delta x_s - d(x_{T_f})$$

and the control problem is thus to minimize $J(u) = -E_u R(u)$. To get this in the form (3.2) we have to introduce the fundamental martingales associated with the process, but in this case it is more economical (since $M < N$) to classify them by jump *size* rather than, as above, by jump *location*. Thus we define for $i = -1, -2, \cdots, -M$

$$p_i(t) = \sum_{s \leqq t} I_{(\Delta x_s = i)}.$$

This is a point process whose rate under measure P_u is

$$r_i(t, x_{t-}, u_t) = \lambda(t, u_t)\left[q_i(u_t) I_{(x_{t-}>i)} + \left(\sum_{j=i}^{M} q_j(u_t)\right) I_{(x_{t-}=i)}\right].$$

The expected cost is thus

$$J(u) = E_u\left[-\sum_{i=1}^{M} \int_0^{T_f} u_s i\, dp_i(s) + d(x_{T_f})\right]$$

$$= E_u\left[-\int_0^{T_f} u_s r(s, x_{s-}, u_s)\, ds + d(x_f)\right]$$

where $r(t, x, u) = \sum_i i r_i(t, x, u)$.

The formulation of the market trader's problem is now entirely in accordance with our abstract model.[1]

4. Existence of optimal controls. The *value function* for the control problem is the process $\{W_t\}$ defined by

$$W_t = \bigwedge_{u \in \mathcal{U}} E_u\left[\left[\int_{]t, T_f]} c(s, u_s)\, d\Lambda_s + G_f \middle| \mathcal{F}_t\right].\right.$$

(This is the lattice infimum in $L_1(\Omega, \mathcal{F}_t, P_{u_0})$ for arbitrary $u_0 \in \mathcal{U}$; see [3], [8].). Note that

$$W_0 = J^* = \lim_{u \in \mathcal{U}} J(u).$$

The value function satisfies the following "principle of optimality" [8, Thm. 4.6].

THEOREM 4.1. *For any $u \in \mathcal{U}$ the process $\{M_t^u\}$ defined by*

$$M_t^u = \int_{]0,t]} c(s, u_s)\, d\Lambda_s + W_t$$

is an $(\mathcal{F}_t, P_u)$ submartingale. It is a martingale if and only if u is optimal.

Take any $u \in \mathcal{U}$. Since $\{M_t^u\}$ is a bounded submartingale it admits a Doob–Meyer decomposition

$$M_t^u = J^* + N_t^u + a_t^u \tag{4.1}$$

where $\{N_t^u\}$ is an $\{\mathcal{F}_t, P_u\}$ martingale and $\{a_t^u\}$ a predictable increasing process with $a_0^u = 0$. The martingale representation theorem [5, Thm. 2] can thus be applied to $\{N_t^u\}$ to give

$$N_t^u = \int_{]0,t] \times X} g(s, x) q^u(ds, dx) \tag{4.2}$$

for some $g \in L_1^{\text{loc}}(p)$. The crucial point is that due to the form of the decomposition (4.1) and the definition of $q^u(t, \Lambda)$, g does not depend on u. (This is seen in the proof of Theorem 4.2 below, where we calculate the decomposition corresponding to another control u^*.) Now define the following "Hamiltonian" process

$$H(t, u, \omega) = c(t, u, \omega) + \alpha(t, u, \omega) \int_X g(t, x, \omega) \lambda(dx, t, \omega).$$

The minimum principles given in [3], [8] state that an optimal control must minimize this Hamiltonian pointwise. Since the function g is defined independently of the existence of an optimal control this suggests that such a control can be constructed

[1] This example is included mainly to demonstrate how such problems can be formulated in terms of measure transformations. As it stands, the example is a Markovian decision problem which could be handled by the methods of, for example, [14], but it is clear that various forms of non-Markovian dependence could be introduced into it without leaving our general framework.

simply by picking out, for each (t, ω), the value of u that minimizes $H(t, u, \omega)$. This is the idea used below, and it is similar to the argument used for the stochastic differential equation case in [6]. We need the following additional conditions:

(4.3)
(i) For each $(s, x, \omega) \in R^+ \times X \times \Omega$, $\alpha(s, \cdot, \omega)$, $\beta(s, x, \cdot, \omega)$, $c(s, x, \cdot, \omega)$ are continuous on U.
(ii) For each $(t, \omega) \in R^+ \times \Omega$ there is a $u_0 \in U$ such that

$$\mathscr{H}(t, \omega) = H(t, u_0, \omega) = \inf_{u \in U} H(t, u, \omega).$$

This assumption is satisfied if for example U is compact.

THEOREM 4.2. *With the formulation of* § 3 *and the assumptions* (4.3), *an optimal control* u^* *exists in the class* $\mathscr{U}$ *of* $\mathscr{F}_t$ *predictable controls.*

Proof. Let $\mathscr{P}$ denote the predictable σ-field on $[0, T_f] \times \Omega$. From (4.3)(i), H is continuous in u for fixed (t, ω) and hence

$$\mathscr{H}(t, \omega) = \inf_{u \in S} H(t, u, \omega)$$

where S is a countable dense subset of U, so that for any constant a,

$$\{(t, \omega) : \mathscr{H}(t, \omega) < a\} = \bigcup_{u \in S} \{(t, \omega) : H(t, u, \omega) < a\}.$$

Since $H(\cdot, u, \cdot)$ is a predictable process for each $u \in U$, this shows that $\mathscr{H}$ is $\mathscr{P}$-measurable. Also (4.3)(i) means that

$$\mathscr{H}(t, \omega) \in H(t, U, \omega) \quad \text{for all } (t, \omega).$$

According to Beneš' implicit function lemma in [1] these facts guarantee the existence of a $\mathscr{P}$-measurable mapping $u^* : [0, T_f] \times \Omega \to U$ such that

$$\mathscr{H}(t, \omega) = H(t, u^*(t, \omega), \omega). \tag{4.4}$$

u^* is an admissible control in accordance with Definition 3.1.

We now show that u^* is optimal in $\mathscr{U}$. From (4.1) and (4.2)

$$M_t^u = J^* + \int_{]0,t] \times X} g(s, x)\, dq^u + a_t^u. \tag{4.5}$$

Using the control u^* constructed in (4.4) we obtain, using obvious shorthands (* stands for u^*),

$$\begin{aligned} M_t^* &= \int_0^t c(s, x_s, u^*)\, d\Lambda_s + W(t) \\ &= M_t^u + \int_0^t c_s^*\, d\Lambda_s - \int_0^t c_s^u\, d\Lambda_s \\ &= J^* + \int_{]0,t] \times X} dq^u + a_t^u + \int_0^t (c_s^* - c_s^u)\, d\Lambda_s \\ &= J^* + \int_{]0,t] \times X} g\, dq^* + \int_{]0,\, t] \times X} g(\alpha^* \beta^* - \alpha^u \beta^u)\, d\lambda_s\, d\Lambda_s + a_t^u \\ &\quad + \int_0^t (c_s^* - c_s^*)\, d\Lambda_s \\ &= J^* + \int_{]0,t] \times X} g\, dq^* + a_t^* \end{aligned} \tag{4.6}$$

where

$$a_t^* = a_t^u - \bar{a}_t^u$$

and

$$\bar{a}_t^u = \int_0^t \left\{ (c_s^u - c_s^*) + \int_X g(\alpha^u \beta^u - \alpha^* \beta^*)\, d\lambda_s \right\} d\Lambda_s. \tag{4.7}$$

The calculations above do not depend on the particular form of u^*, and hence (4.6) confirms our earlier assertion that the integrand g is not control-dependent. What we do get from the construction of u^* as in (4.4), however, is that $\{\bar{a}_t^u\}$ is a predictable *increasing process*. To prove that u^* is optimal, it suffices to show that $a_{T_f}^* = 0$ a.s. since then M_t^* is a martingale (see theorem 4.1). From (4.5) and (4.7) we have

$$M_{T_f}^u = J^* + \int_{]0,T_f]\times X} g\, dq^* + a_{T_f}^* + \bar{a}_{T_f}^u$$

and hence

$$J(u) = E_u M_{T_f}^u \geqq J^* + E_u a_{T_f}^* \geqq 0. \tag{4.8}$$

Now, since $J^* = \inf_u J(u)$, we can choose a sequence of controls $\{u_n\}$ such that $J(u_n)\downarrow J^*$, which from (4.8) implies that

$$E_{u_n}[a_{T_f}^*] \to 0, \qquad n \to \infty. \tag{4.9}$$

Fix $N, K \in Z^+$ and define

$$Y = (a_{T_f \wedge T_N}^*) \wedge K;$$

then $Y \in L_\infty(\Omega, F_{T_N}, P)$ and $E_{u_n}[Y] \to 0$ in view of (4.9) and the fact that a_t^* is an increasing process. But now we have

$$E_{u_n}[Y] = E[L^{(N)}(u_n) Y]$$

(see Lemma 3.1), and according to Lemma 3.1 there is a subsequence $\{u_{n_k}\}$ such that $L^{(N)}(u_{n_k})$ converges weakly to an a.s. positive random variable ρ as $k \to \infty$. Thus

$$0 = \lim_{k\to\infty} E[L^{(N)}(u_{n_k}) Y] = E[\rho Y]$$

and hence $Y = 0$ a.s. Since this holds for every N and K, letting N and K tend to infinity (in that order) gives

$$a_{T_f}^* = 0 \quad \text{a.s.}$$

This completes the proof.

5. Concluding remarks. Theorem 5.1 is a much less satisfactory result than the corresponding theorem in [6] concerning control of systems described by stochastic differential equations. There, only minimal conditions on the system functions were needed to ensure that the measures corresponding to all admissible controls are mutually absolutely continuous whereas here we are obliged to impose the rather unnatural conditions (3.1) to achieve this. These conditions, though essential to our line of argument, are probably more stringent than necessary in particular cases. (For example, they are not satisfied in Pliska's Markovian jump process formulation [14]).

In [2] and [10] existence for the stochastic differential equation case was proved by establishing a compactness property of the set of densities corresponding to admissible controls. The approach is not generally applicable here, since the crucial convexity property of the set of densities only holds in special circumstances. Indeed let $L_t(u)$ denote $E[dP_u/dP|\mathcal{F}_t]$ and consider the process

$$L_t = \theta L_t(u_1) + (1-\theta)L_t(u_2)$$

where $u_1, u_2 \in \mathcal{U}$ and $0<\theta<1$. Using a similar approach to that employed by Beneš in [2] one can show that L_t is the density obtained by replacing α^u, β^u in (3.4) by $\tilde{\alpha}, \tilde{\beta}$ defined as follows. Let

$$\mu_t = \theta L_{t-}(u_1)/L_{t-}.$$

Then

$$\tilde{\alpha}(t,\omega) = \mu_t\alpha^{u_1}(t,\omega) + (1-\mu_t)\alpha^{u_2}(t,\omega)$$

and

$$\tilde{\beta}(t,x,\omega) = \nu_t\beta^{u_1}(t,x,\omega) + (1-\nu_t)\beta^{u_2}(t,x,\omega)$$

where

$$\nu_t = \mu_t\alpha^{u_1}/\tilde{\alpha}.$$

Thus *different* convex combinations are required for $\tilde{\alpha}$ and $\tilde{\beta}$, which will therefore not be equal to α^u, β^u for any $u \in \mathcal{U}$, even if $\alpha(t,U,\omega)$ and $\beta(t,x,U,\omega)$ are convex, except in special circumstances. This argument does, however, show that the sets $D^N(\mathcal{G})$ introduced below in the Appendix are convex.

Appendix. The purpose of this Appendix is to give a proof of Lemma 3.1. Indeed, Lemma 3.1 follows immediately from Lemmas A.3 and A.4 below.

DEFINITION A.1. Let $\mathcal{G}_1$ and $\mathcal{G}_2$ denote respectively the sets of measurable functions $\tilde{\alpha}: R^+\times\Omega\to R^+$ and $\tilde{\beta}: R^+\times X\times\Omega\to R^+$ satisfying conditions (3.1) with the u-dependence deleted. Now define

$$\mathcal{G} = \{(\tilde{\alpha},\tilde{\beta}): \tilde{\alpha}\in\mathcal{G}_1, \tilde{\beta}\in\mathcal{G}_2\}.$$

Note in particular that $(\alpha^u,\beta^u)\in\mathcal{G}$ for each $u\in\mathcal{U}$. To each $(\tilde{\alpha},\tilde{\beta})\in\mathcal{G}$ there correspond families of functions $\{(\tilde{\alpha}_k,\tilde{\beta}_k): k\in Z^+\}$ related to $(\tilde{\alpha},\tilde{\beta})$ as in (3.2), (3.3) and we can define a measure $\tilde{P}$ on $(\Omega,\mathcal{F}_\infty)$ by specifying its Radon–Nikodym derivative in a manner analogous to (3.4). For $N\in Z^+$ let $\tilde{P}_N$ be the restriction of $\tilde{P}$ to $\mathcal{F}_{T_N}$ defined by this procedure, and define

$$L^N(\tilde{\alpha},\tilde{\beta}) = \frac{d\tilde{P}_N}{dP_N}$$

and

$$D^N(\mathcal{G}) = \{L^N(\tilde{\alpha},\tilde{\beta}): (\tilde{\alpha},\tilde{\beta})\in\mathcal{G}\}.$$

Then $D^N(\mathcal{G})\subset L_1(\Omega,\mathcal{F}_{T_N},P_N)$.

LEMMA A.1. *For any* $(\tilde{\alpha},\tilde{\beta})\in\mathcal{G}$, $T_\infty=\infty$ *a.s.*$(\tilde{P})$.

Proof. From (3.1)(ii) $\tilde{\alpha}_k(t)\leqq c_2$ so that

$$\tilde{\Lambda}^k_t = \int_{]0,t-T_{k-1}]} \tilde{\alpha}_k(s+T_{k-1})\,d\Lambda^k_s \leqq c_2\Lambda^k(t-T_{k-1}).$$

Thus for $k\in\mathcal{K}$,

$$\tilde{\Lambda}^k_{s+T_{k-1}} \leqq c_2\Lambda^k(s) \leqq c_2\theta_2 \quad \text{for } s\in[0,\theta_1].$$

It follows as in Lemma 2.1 that there is a constant $\theta'_3 > 0$ such that

$$\tilde{P}[S_k > \theta_1 | \mathcal{F}_{T_{k-1}}] \geqq \theta'_3 \quad \text{a.s. } (\tilde{P})$$

and hence $\sum_k \tilde{P}[B_k \mid \mathcal{F}_{T_{k-1}}] = \infty$ a.s., where $B_k = (S_k > \theta_1)$. Now by a result of Doob, [9, Cor. 1, p. 323] this series converges on the same set (modulo a null set) as the series $\sum_k I_{B_k}$. Thus B_k occurs infinitely often and consequently $T_k \to \infty$ a.s. $(\tilde{P})$.

LEMMA A.2. *For each $N \in Z^+$, $\varepsilon > 0$ there exist $\delta > 0$, $\rho < \infty$, $D \in \mathcal{F}_{T_N}$ such that, for all $(\tilde{\alpha}, \tilde{\beta}) \in \mathcal{G}$, $P_N(D) \geqq 1 - \varepsilon$ and $\delta \leqq L^N(\tilde{\alpha}, \tilde{\beta})(\omega) \leqq \rho$ for $\omega \in D$.*

Proof. First note from (3.4) that for $(\tilde{\alpha}, \tilde{\beta}) \in \mathcal{G}$ we can write

$$L^N(\tilde{\alpha}, \tilde{\beta}) = \prod_{k-1}^{N} L_k(\omega_{k-1}; S_k, Z_k)$$

where

$$\begin{aligned} L_k(\omega_{k-1}; t, x) = \tilde{\alpha}_k(t)\tilde{\beta}_k(t, x) \exp\left\{-\int_0^t (\tilde{\alpha}_k(s) - 1)\, d\Lambda_s^{kc}\right\} \\ \cdot \prod_{s<t} \frac{(1 - \tilde{\alpha}_k(s)\Delta\Lambda_s^k)}{(1 - \Delta\Lambda_s^k)} I_{(t \leqq \tilde{d}^k)}. \end{aligned} \tag{A.1}$$

Fix k and for notational convenience, let $\{s : \Delta\Lambda_s^k > 0\} = \{t_1, t_2 \cdots\}$ and $a_i = \Delta\Lambda_{t_i}^k$. Observe that $0 \leqq a_i < 1$, so that from standard analysis

$$\prod_{i=1}^{\infty} (1 - a_i) > 0 \quad \Leftrightarrow \quad \sum_{n=1}^{\infty} a_n < \infty.$$

Hence for any constant c, $0 \leqq c < 1$, we have

$$\prod_{i=1}^{\infty} (1 - a_i) > 0 \quad \Leftrightarrow \quad \prod_{i=1}^{\infty} (1 - ca_i) > 0 \quad \Leftrightarrow \quad \sum_{i=1}^{\infty} a_i < \infty.$$

Denote

$$\gamma^k(s) = \min(c_2\Delta\Lambda_s^k, 1 - \delta') < 1.$$

Now from (A.1) and (3.5)

$$\begin{aligned} \tilde{F}^k(\omega_{k-1}; t) &= \tilde{P}_N[S_k > t | \mathcal{F}_{T_{k-1}}] \\ &= \exp\left\{-\int_0^t \tilde{\alpha}_k(s)\, d\Lambda_s^{kc}\right\} \prod_{s \leqq t} (1 - \tilde{\alpha}_k(s)\Delta\Lambda_s^k) \\ &\leqq e^{-c_1\Lambda_t^{kc}} \prod_{t_i \leqq t} (1 - c_1a_i) \leqq \prod_{t_i \leqq t} (1 - c_1a_i), \end{aligned} \tag{A.2}$$

and from (A.1)

$$c_1^2 e^{-(c_2-1)\Lambda_t^{kc}} \prod_{t_i<t} \frac{(1 - \gamma^k(t_i))}{(1 - a_i)} \leqq L_k(\omega_{k-1}; t, x) \leqq c_2c_3\, e^{(1-c_1)\Lambda_t^{kc}} \prod_{t_i \leqq t} \frac{(1 - c_1a_i)}{(1 - a_i)}. \tag{A.3}$$

There are now two separate cases to consider:

Case 1. $d^k = \infty$, or $d^k < \infty$, $\Lambda^k(d^k -) = \infty$. Clearly in this case $\Lambda_t^k \uparrow \infty$ as $t \uparrow d^k$. Suppose $\sum_{t_i \leqq d^k} a_i = \infty$ so that

$$\prod_{t_i \leqq d^k} (1 - \alpha^k(t_i)) = 0 = \prod_{t_i \leqq d^k} (1 - a_i) = \prod_{t_i \leqq d^k} (1 - c_1a_i).$$

Then for any $\varepsilon' > 0$ there exists a $\tau^k < d^k$ such that

$$\prod_{t_i \leqq \tau^k} (1 - c_1 a_i) < \varepsilon' \tag{A.4}$$

and

$$\prod_{t_i \leqq \tau^k} (1 - \gamma^k(t_i)) > 0, \qquad \prod_{t_i \leqq \tau^k} (1 - a_i) > 0.$$

Now define a set $D_k \subset \Omega_k$ and numbers δ_k, ρ_k by

$$\begin{aligned} D_k &= \Omega_{k-1} \times [0, \tau^k] \times X, \\ \delta_k &= c_1^2 \, e^{-(c_2 - 1)\Lambda^{kc}_{\tau^k}} \prod_{t_i < \tau^k} \frac{(1 - \gamma^k(t_i))}{(1 - a_i)}, \\ \rho_k &= c_2 c_3 \, e^{(1 - c_1)\Lambda^{kc}_{\tau^k}} \prod_{t_i < \tau^k} \frac{(1 - c_1 a_i)}{(1 - a_i)}. \end{aligned} \tag{A.5}$$

Since Λ^k is deterministic δ_k, ρ_k, τ^k are not dependent on ω_{k-1}, and depend only on ε' and k.

Using (A.2)–(A.4), we now conclude that

$$0 < \delta_k \leqq L_k(\omega_k) \leqq \rho_k < \infty, \quad \text{all } \omega_k \in D_k \tag{A.6}$$

with

$$\tilde{P}_N(\omega_k(\omega) \in D_k) \geqq 1 - \varepsilon'.$$

On the other hand if $\sum_{t_i \leqq d^k} a_i < \infty$, so that

$$\prod_{t_i < d^k} (1 - \alpha^k(t_i)) > 0, \qquad \prod_{t_i < d^k} (1 - a_i) > 0,$$

$$\prod_{t_i < d^k} (1 - c_1 a_i) > 0,$$

we must then have

$$\Lambda^{kc}_t \uparrow \infty \quad \text{as} \quad t \uparrow d^k.$$

Hence for given $\varepsilon' > 0$, there must exist a $\tau^k < d^k$ such that $e^{-c_1 \Lambda^{kc}_{\tau^k}} < \varepsilon'$. Again defining D_k, δ_k, ρ_k as in (A.5) we see that equation (A.6) applies.

Case 2. $d^k < \infty$, $\Lambda^k(d^k -) < \infty$. In this case since $\sum_{t_i \leqq d^k} a_i < \infty$,we know that

$$\prod_{t_i \leqq d^k} (1 - a_i), \qquad \prod_{t_i \leqq d^k} (1 - c_1 a_i), \qquad \prod_{t_i \leqq d^k} (1 - \alpha^k(t_i - 1)$$

and all strictly positive, and $\Lambda^{kc}_{d^k} < \infty$.

Thus choosing $\tau^k = d^k$, equation (A.6) applies with $\tilde{P}_N(\omega_k(\omega) \in D_k) = 1$, since $\tilde{F}^k_{d_k} = 0$.

We have thus shown that (A.6) is true in general, for each $k = 1, \cdots, N$.

We can now apply the above argument for each $k = 1, \cdots, N$ and for each $\varepsilon' > 0$ pick times $\tau^1, \tau^2, \cdots, \tau^N$ such that

$$\tilde{P}_N(S_k > \tau^k) < \varepsilon', \quad \text{all } L^N \in D^N(\mathscr{G}).$$

Hence for any $\varepsilon > 0$, we can choose ε' small enough so that

$$\tilde{P}_N\Big(\bigcap_{k=1}^{N} \{S_k \leqq \tau^k\}\Big) \geqq 1 - \varepsilon,$$

and since on D_k, L_k has bounds δ_k, ρ_k, we conclude that

$$0<\delta \leqq L^k(\omega_N) \leqq \rho < \infty, \quad \text{for } \omega_N \in D \text{ all } L^N \in D^N(\mathscr{G}),$$
$$\tilde{P}_N(D) \geqq 1-\varepsilon,$$

where

$$\delta = \prod_{k=1}^{N} \delta_k,$$

$$\rho = \prod_{k=1}^{N} \rho_k,$$

$$D = \bigcap_{k=1}^{N} \{T_k \leqq \tau^k\}.$$

This proves the lemma.

LEMMA A.3. *For each $N \in Z^+$, the set of densities $D^N(\mathscr{G}) \subset L_1(\Omega_N, \mathscr{F}_{T_N}, P_N)$ is weakly sequentially compact.*

Proof. It is sufficient to show that $D^N(\mathscr{G})$ is uniformly integrable, i.e. for each $\varepsilon > 0$, there exists a $\delta > 0$ such that $P_N(A) < \delta$ implies $\tilde{P}_N(A) < \varepsilon$, where $A \in \mathscr{F}_{T_N}$.

Let $D \in \mathscr{F}_{T_N}$ be as in Lemma A.2; then

$$\tilde{P}_N(A) = \int_{A \cap D} L^N \, dP_N + \int_{A \cap D^c} L^N \, dP_N.$$

From Lemma A.2 we then have for $\varepsilon' > 0$

$$\tilde{P}_N(A) \leqq \rho' P_N(A) + \varepsilon' \quad \text{all } L^N \in D^N(\mathscr{G}) \quad \text{some } \rho' < \infty.$$

Hence, for $\varepsilon > 0$, choosing

$$\delta = \frac{\varepsilon}{2\rho'}, \qquad \varepsilon' = \frac{\varepsilon}{2},$$

we see that if $P_N(A) < \delta$ then

$$\tilde{P}_N(A) < \varepsilon, \quad \text{all } L^N \in D^N(\mathscr{G}), \quad A \in \mathscr{F}_{T_N}.$$

This proves the lemma.

LEMMA A.4. *Suppose $\{L_n\}$ is a weakly convergent sequence in $D^N(\mathscr{G})$ with limit $L \in L_1(\Omega^N, \mathscr{F}_{T_N}, P_N)$. Then $L > 0$ a.s. P_N.*

Proof. Let $A = (L = 0)$ and take $\varepsilon > 0$, $D \in \mathscr{F}_{T_N}$ as in Lemma A.2. Then $P_N(A \cap D^c) < \varepsilon$ and

$$0 = \int_A L \, dP_N = \lim_{n \to \infty} \int_A L_n \, dP_N \geqq \lim_{n \to \infty} \int_{A \cap D} L_n \, dP_N \geqq P_N(A \cap D).$$

Thus $P_N(A \cap D) = 0$, so that $P_N A < \varepsilon$. The result follows.

REFERENCES

[1] V. E. BENEŠ,, *Existence of optimal strategies based on specified information for a class of stochastic decision problems*, this Journal, 8 (1970), pp. 179–188.

[2] ———, *Existence of optimal stochastic control laws*, this Journal, 9 (1971), pp. 446–475.

[3] R. BOEL AND P. P. VARAIYA, *Optimal control of jump processes*, this Journal, 15 (1977), pp. 92–119.

[4] R. Boel, P. Varaiya and E. Wong, *Martingales on jump processes II: Applications*, this Journal, 13 (1975), pp. 1022–1061.

[5] M. H. A. Davis, *The representation of martingales of jump processes*, this Journal, 14 (1976), pp. 623–638.

[6]———, *On the existence of optimal policies in stochastic control*, this Journal, 11 (1973), pp. 587–594.

[7] M. H. A. Davis and P. P. Varaiya, *Dynamic programming conditions for partially observable stochastic systems*, this Journal, 11 (1973), pp. 226–261.

[8] M. H. A. Davis and R. J. Elliott, *Optimal control of a jump process*, Z. Wahrscheinlichkeitstheorie verw. Gebiete, 40 (1977), pp. 183–202.

[9] J. L. Doob, *Stochastic Processes*, John Wiley, New York, 1953.

[10] T. Duncan and P. P. Varaiya, *On the solutions of a stochastic control system*, this Journal, 9 (1971), pp. 354–371.

[11] N. Dunford and J. T. Schwartz, *Linear Operators* 1, Interscience, New York, 1958.

[12] R. J. Elliott, *Levy systems and absolutely continuous changes of measure for a jump process*, J. Math. Anal. Appl., 61 (1977), pp. 785–796.

[13] P. A. Meyer, *Probability and Potentials*, Blaisdell, Waltham, MA, 1966.

[14] S. R. Pliska, *Controlled jump processes*, Stochastic Processes Appl. 3 (1975), pp. 259–282.

[15] C. B. Wan, *The optimal control of stochastic jump processes: a martingale representational approach*, PhD. thesis, Dept. of Computing and Control, Imperial College, London, 1977.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 4, July 1979

0363-0129/79/1704-0007$01.00/0

ARMA SPLINES, SYSTEM INVERSES, AND LEAST-SQUARES ESTIMATES*

H. L. WEINERT,† U. B. DESAI† AND G. S. SIDHU‡

Abstract. Many of the optimal curve fitting problems arising in approximation theory and numerical analysis have the same structure as certain problems of least-squares estimation of stochastic processes. This structural correspondence implies that optimal curve fits (splines) are sample functions of linear least-squares estimates. As a result, recursive estimation techniques can be used to solve these spline problems. Previous work has dealt with splines determined by differential operators; these so-called Lg-splines are sample functions of estimates of autoregressive stochastic processes (generated by all-pole systems in response to white noise). The present work examines splines determined by certain integro-differential operators defined in terms of the system inverse; these splines are sample functions of estimates of autoregressive-moving average (ARMA) processes (generated by systems with zeros).

1. Introduction. In this paper we investigate splines determined by certain integro-differential operators. These are natural generalizations of Lg-splines [1], which are determined by differential operators. This generalization was motivated by the following considerations. The properties of, and recursive algorithms for, Lg-splines are intimately connected with dynamical systems whose transfer functions have no zeros. This is a rather special class of systems, and one might wonder about the properties of splines associated with systems having numerator dynamics (zeros). One problem for which such generalized splines are needed is the minimum-energy control of a linear system having functional constraints on its output [2]. (The zero-free case is discussed in [3], [4].)

The particular integro-differential operators we consider are defined in terms of reduced-order system inverses [5]. The generalized spline of interest is defined as the solution to an optimization problem involving this operator. We have named these splines autoregressive-moving average (ARMA) splines because, as is also shown, they are sample functions of least-squares estimates of ARMA random processes. This stochastic correspondence is used to develop recursive algorithms for ARMA splines. Finally, the structural and continuity properties of ARMA splines are investigated.

Splines determined by general continuous linear operators were apparently first considered by Atteia [6], [7] and then Anselone and Laurent [8] and Sard [9]. At this level of generality, not much can be said about actual spline algorithms or structural properties, although existence and uniqueness results can be established, and the spline can be characterized as a projection. Results similar to some of ours have been obtained recently by de Figueiredo [10], who relied heavily on our earlier work [11] concerning Lg-splines. He takes a different approach to the ARMA spline problem by using two separate Green's functions rather than the reduced-order system inverse. No results on structural and continuity properties of the splines are given. We note that the foundations of our present work were laid in [2].

2. ARMA splines and system inverses. In order to motivate our definition of the ARMA spline, let us review the definition of the Lg-spline. Let H^n be the Hilbert space of functions whose nth derivatives are square-integrable on the interval $[0, 1]$. Let $\{\lambda_i\}_1^N$

* Received by the editors August 23, 1978.

† Department of Electrical Engineering, The Johns Hopkins University, Baltimore, Maryland 21218.

‡ Instituto de Investigaciones en Matematicas Aplicadas y en Sistemas, Universidad Nacional Autonoma de Mexico, Mexico City, Mexico.

be linear functionals, let $\{r_j\}_1^N$ be real numbers, and let L be the differential operator

(1) $$L = D^n + \sum_{j=0}^{n-1} a_j D^j.$$

The Lg-spline interpolating $\{r_j\}_1^N$ with respect to $\{\lambda_j\}_1^N$ is a function $\sigma(\cdot) \in H^n$ that satisfies the constraints $\lambda_j \sigma = r_j$, $1 \leqq j \leqq N$, and that minimizes $\int_0^1 (Lf)^2$ among all functions $f(\cdot) \in H^n$ that satisfy the foregoing constraints. The operator L determines the functional form of the spline, and is chosen by the analyst. Lg-splines are intimately connected with the nth order linear dynamical system having transfer function $1/L$. Note that if $u(\cdot)$ is a square-integrable input to such a system, and $f(\cdot) \in H^n$ is the corresponding output, then $u(\cdot)$ can be recovered from $f(\cdot)$ via

(2) $$u = Lf.$$

Equation (2) represents the reduced-order system inverse (see the Appendix) for the system with transfer function $1/L$.

Now let us introduce another differential operator

(3) $$M = \sum_{j=0}^{m} c_j D^j$$

with $c_m \neq 0$ and $m < n$. We assume that L and M have no common factors. Consider the system with transfer function M/L. This nth order system can be written in state variable form as (see the Appendix)

(4) $$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{b}u(t) \\ f(t) &= \mathbf{c}\mathbf{x}(t) \end{aligned}$$

where

(5a) $$\mathbf{A} = \left[\begin{array}{c|ccc} \mathbf{0} & & \mathbf{I} & \\ \hline -a_0 & -a_1 & \cdots & -a_{n-1} \end{array}\right]$$

(5b) $$\mathbf{c} = [1, 0, \cdots, 0]$$

(5c) $$\mathbf{b} = [0, \cdots, 0, b_\alpha, \cdots, b_n]'.$$

Here the parameter $\alpha = n - m$, and is called the *relative order* of the system (4). The parameters $\{b_j\}$ are found via the recursions

(6) $$\begin{aligned} b_\alpha &= c_m \\ b_{\alpha+j} &= c_{m-j} - \sum_{k=0}^{j-1} a_{n-j+k} b_{\alpha+k}, \qquad 1 \leqq j \leqq m. \end{aligned}$$

The state vector can be written as

$$\mathbf{x}(\cdot) = [x_1(\cdot), \cdots, x_n(\cdot)]'$$

where in this particular coordinate system the first α state variables are determined solely by the output and its derivatives; i.e.,

(7) $$x_k(\cdot) = f^{(k-1)}(\cdot), \qquad 1 \leqq k \leqq \alpha.$$

Determination of the remaining state variables requires knowledge of both output and input. The initial value $\mathbf{x}(t_0)$ for (4) will be specified at some point $t_0 \in [0, 1]$ to be chosen later.

As shown in [5] and summarized in the Appendix, the reduced-order inverse of (4) is the mth order system:

$$\begin{aligned}\dot{\boldsymbol{\theta}}(t) &= \hat{\mathbf{A}}\boldsymbol{\theta}(t) + \hat{\mathbf{B}}\mathbf{F}(t)\\ u(t) &= \hat{\mathbf{c}}\boldsymbol{\theta}(t) + \hat{d}f^{(\alpha)}(t)\end{aligned} \tag{8}$$

where

$$\mathbf{F}(t) = [f(t), f^{(1)}(t), \cdots, f^{(\alpha)}(t)]'.$$

The parameters in (8) are given in the Appendix in terms of the $\{a_i\}$ and $\{b_i\}$. The initial value for (8) is $\boldsymbol{\theta}(t_0) = [x_{\alpha+1}(t_0), \cdots, x_n(t_0)]'$. If $f(\cdot) \in H^\alpha$ is the output of (4) in response to a square-integrable input $u(\cdot)$, we can recover $u(\cdot)$ from $f(\cdot)$ via (8) only if we know $\boldsymbol{\theta}(t_0)$. Since this initial condition cannot be determined from $f(\cdot)$ alone, and since its value has nothing to do with the choice of L and M, we set it to zero. That is,

$$\boldsymbol{\theta}(t_0) = \mathbf{0} \tag{9}$$

and thus

$$\mathbf{x}(t_0) = [f(t_0), \cdots, f^{(\alpha-1)}(t_0), 0, \cdots, 0]'. \tag{10}$$

The choice in (9) does not restrict the range of the system (4); that is, every function in H^α can be generated by (4), (10) in response to some square-integrable input. That input is simply the output of the inverse system (8), (9). As we shall see later, the choice in (9) is also essential to the development of the proper topology for H^α. A similar initial condition problem arises in the study of innovations representations for smooth processes [12], [13].

It is clear now that the reduced-order inverse (8), (9) can be written as

$$u = Tf \tag{11}$$

where T is a bounded linear integro-differential operator that maps H^α onto the space of functions square-integrable on [0, 1]. We can now state the definition of the autoregressive-moving average (ARMA) spline.

DEFINITION 2.1. A function $s(\cdot) \in H^\alpha$ is an *ARMA spline* interpolating $\{r_j\}_1^N$ with respect to $\{\lambda_j\}_1^N$ and T if

$$\lambda_j s = r_j, \qquad 1 \leqq j \leqq N, \tag{12}$$

and

$$\int_0^1 (Ts)^2 = \min_{f \in U} \int_0^1 (Tf)^2, \tag{13}$$

where

$$U = \{f \in H^\alpha : \lambda_j f = r_j,\ 1 \leqq j \leqq N\}.$$

We are assuming here that $\{\lambda_j\}_1^N$ are linearly independent and bounded on H^α. We know from [8] and [1] that a solution to the optimization problem (12), (13) always exists, and that the solution is unique if $\{\lambda_j\}_1^\alpha$ are linearly independent on the null space of T, which, as can be seen by examining the zero-input response of (4), (10), has dimension α. By a straightforward generalization of the results of [14], it can be shown that this uniqueness condition is equivalent to a strong type of observability for the system (4), (10). In particular, $\{\lambda_j\}_1^\alpha$ are linearly independent on the null space of T if and only if the initial state (10) of (4) can be recovered uniquely from measurements $\{\lambda_j f\}_1^\alpha$ when the input $u(\cdot)$ is zero. We will assume in what follows that this uniqueness condition is satisfied. It should be clear that the ARMA spline reduces to the Lg-spline when $m=0$ and $c_0=1$.

3. Reproducing kernel structure. We now show that H^α is a reproducing kernel Hilbert space. This will allow us to compute the ARMA spline as a projection in H^α. First we need a certain representation for functions in H^α, but before proceeding, we will restrict our attention to a specific broad class of constraint functionals $\{\lambda_j\}_1^N$ called extended Hermite–Birkhoff functionals. These functionals have the form

$$\lambda_j f = \sum_{k=1}^{\alpha} \gamma_{jk} f^{(k-1)}(t_j), \qquad 1 \leqq j \leqq N, \tag{14}$$

where $0 \leqq t_1 \leqq t_2 \leqq \cdots \leqq t_N \leqq 1$ and the $\{\gamma_{jk}\}$ are known real numbers. For future reference let $\mathbf{h}_j$ be the n-vector

$$\mathbf{h}_j = [\gamma_{j1}, \gamma_{j2}, \cdots, \gamma_{j\alpha}, 0, \cdots, 0], \qquad 1 \leqq j \leqq N. \tag{15}$$

Also, as discussed later on, the best choice of initialization point is

$$t_0 = 0.$$

Now from (4) we have, for any $f \in H^\alpha$

$$f(t) = \mathbf{c}\boldsymbol{\phi}(t)\mathbf{x}(0) + \int_0^t \mathbf{c}\boldsymbol{\phi}(t-\tau)\mathbf{b}u(\tau)\,d\tau, \tag{16}$$

where

$$\boldsymbol{\phi}(t) = \exp(\mathbf{A}t).$$

Since (4), (7), (14) and (15) imply

$$\lambda_j f = \mathbf{h}_j \mathbf{x}(t_j), \tag{17}$$

we have

$$\lambda_j f = \mathbf{h}_j \boldsymbol{\phi}(t_j)\mathbf{x}(0) + \int_0^{t_j} \mathbf{h}_j \boldsymbol{\phi}(t_j - \tau)\mathbf{b}u(\tau)\,d\tau. \tag{18}$$

We can write (18) in matrix form as follows: let $\mathbf{W}$ be the $\alpha \times \alpha$ matrix formed from the first α columns of the $\alpha \times n$ matrix

$$\begin{bmatrix} \mathbf{h}_1 \boldsymbol{\phi}(t_1) \\ \vdots \\ \mathbf{h}_\alpha \boldsymbol{\phi}(t_\alpha) \end{bmatrix}$$

and let $\boldsymbol{\Delta}(\cdot)$ be the $\alpha \times n$ matrix with ith row $\boldsymbol{\Delta}_i'(\cdot)$ given by

$$\boldsymbol{\Delta}_i'(t) = \begin{cases} \mathbf{h}_i \boldsymbol{\phi}(t_i - t), & t \in [0, t_i], \\ \mathbf{0}, & \text{otherwise.} \end{cases}$$

If also

$$\mathbf{x}_0 = [x_1(0), \cdots, x_\alpha(0)]'$$

and

$$\boldsymbol{\eta} = [\lambda_1 f, \cdots, \lambda_\alpha f]',$$

then for $j = 1, 2, \cdots, \alpha$, (18) becomes

$$\boldsymbol{\eta} = \mathbf{W}\mathbf{x}_0 + \int_0^{t_\alpha} \boldsymbol{\Delta}(\tau)\mathbf{b}u(\tau)\, d\tau. \tag{19}$$

Our uniqueness assumption guarantees that $\mathbf{W}$ is invertible [14] and thus

$$\mathbf{x}_0 = \mathbf{W}^{-1}\boldsymbol{\eta} - \mathbf{W}^{-1}\int_0^{t_\alpha} \boldsymbol{\Delta}(\tau)\mathbf{b}u(\tau)\, d\tau. \tag{20}$$

Now substituting (20) into (16), and recalling (10), we have

$$f(t) = \sum_{j=1}^{\alpha} (\lambda_j f) z_j(t) - \mathbf{z}'(t) \int_0^{t_\alpha} \boldsymbol{\Delta}(\tau)\mathbf{b}u(\tau)\, d\tau + \int_0^t \mathbf{c}\boldsymbol{\phi}(t-\tau)\mathbf{b}u(\tau)\, d\tau, \tag{21}$$

where

$$\mathbf{z}'(t) = [z_1(t), \cdots, z_\alpha(t)] = \mathbf{c}\boldsymbol{\phi}(t)\left[\frac{\mathbf{W}^{-1}}{\mathbf{0}}\right]. \tag{22}$$

Now let

$$g(t-\tau) = \begin{cases} \mathbf{c}\boldsymbol{\phi}(t-\tau)\mathbf{b}, & t \geqq \tau, \\ 0, & t < \tau. \end{cases}$$

Recalling the definition of $\boldsymbol{\Delta}(\cdot)$, we can write

$$f(t) = \sum_{j=1}^{\alpha} (\lambda_j f) z_j(t) + \int_0^1 G(t, \tau) u(\tau)\, d\tau, \tag{23}$$

where

$$G(t, \tau) = g(t-\tau) - \mathbf{z}'(t)\boldsymbol{\Delta}(\tau)\mathbf{b}. \tag{24}$$

From our discussion of the system inverse (11), we know that $u(\cdot)$ can be replaced by $Tf(\cdot)$ in (23); thus,

$$f(t) = \sum_{j=1}^{\alpha} (\lambda_j f) z_j(t) + \int_0^1 G(t, \tau)[Tf(\tau)]\, d\tau. \tag{25}$$

Equation (25) is the desired representation for functions in H^α. It can now be easily checked that the following is a valid inner product for H^α:

$$\langle e, f\rangle = \sum_{j=1}^{\alpha} (\lambda_j e)(\lambda_j f) + \int_0^1 (Te)(Tf). \tag{26}$$

The norm induced by this inner product is

$$\|f\|^2 = \sum_{j=1}^{\alpha} (\lambda_j f)^2 + \int_0^1 (Tf)^2. \tag{27}$$

Since the first term in (27) is fixed for all $f \in U$, we can restate the ARMA spline definition as

DEFINITION 3.1. The *ARMA spline* $s(\cdot) \in H^\alpha$ interpolating $\{r_j\}_1^N$ with respect to $\{\lambda_j\}_1^N$ and T satisfies

$$\lambda_j s = r_j, \qquad 1 \leqq j \leqq N, \tag{28}$$

and

$$\|s\|^2 = \min_{f \in U} \|f\|^2. \tag{29}$$

Now if $\{d_j\}_1^N$ are the representers of $\{\lambda_j\}_1^N$; i.e., if

$$\langle f, d_j \rangle = \lambda_j f, \qquad 1 \leqq j \leqq N,$$

then according to the projection theorem, the ARMA spline is the projection of any function in U onto the subspace of H^α spanned by $\{d_j\}_1^N$. If H^α has a reproducing kernel $K(\cdot, \cdot)$, the representers can be found via

$$d_j(t) = \lambda_j K(\cdot, t).$$

In fact, the reproducing kernel for H^α relative to the inner product (26) is given by

$$K(t, \tau) = \sum_{j=1}^{\alpha} z_j(t) z_j(\tau) + \int_0^1 G(t, \xi) G(\tau, \xi)\, d\xi. \tag{30}$$

To verify this, we must check that [15]

$$K(\cdot, t) \in H^\alpha \quad \text{for all } t \in [0, 1], \tag{31}$$

$$\langle f(\cdot), K(\cdot, t) \rangle = f(t) \quad \text{for all } f \in H^\alpha, \quad t \in [0, 1]. \tag{32}$$

A little algebra will show that

$$Tz_j = 0, \qquad \lambda_i z_j = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases} \quad 1 \leqq (i, j) \leqq \alpha, \tag{33}$$

$$TG(\cdot, \tau) = \delta(\tau - \cdot), \qquad \lambda_j G(\cdot, \tau) = 0, \qquad 1 \leqq j \leqq \alpha, \quad \tau \in [0, 1]. \tag{34}$$

Thus $\{z_j\}_1^\alpha$ span the null space of T and $G(\cdot, \cdot)$ is a Green's function for T. As a result of (33), (34),

$$\lambda_j K(\cdot, \tau) = z_j(\tau), \qquad 1 \leqq j \leqq \alpha \tag{35}$$

$$TK(\cdot, \tau) = G(\tau, \cdot). \tag{36}$$

These equations show that (30) is just a special case of (25), so that (31) is established. Equation (32) can be verified by simply substituting (30) into (26) and using (35), (36) and (25).

Inner products and reproducing kernels for H^α have also been characterized in terms of the individual Green's functions of L and M by Hajek [16], Parzen [17], Wahba [18], and de Figueiredo [10]. Our approach avoids the separate inversion of the operators L and M by using the reduced-order inverse of the system (4), (10). The inverse system (8), (9) can be constructed by simple algebraic manipulations of the original system. In addition, the basic equations (25)–(36) remain in precisely the same form as the corresponding equations in the Lg-spline special case as reported in [19], [20].

4. ARMA splines and least-squares estimates. Since the reproducing kernel $K(\cdot, \cdot)$ is symmetric and nonnegative definite, it is the covariance function of some zero-mean random process $\{y(t), t \in [0, 1]\}$. Proceeding in analogy to the Lg-spline case [19], we can establish the following theorem.

THEOREM 4.1. *Let $\hat{y}(t)$ be the linear least-squares estimate of $y(t)$ given random variables $\{\lambda_j y\}_1^N$ and let $\hat{\hat{y}}(t)$ be the sample function of $\hat{y}(t)$ obtained by setting $\lambda_j y = r_j$,*

$1 \leqq j \leqq N$. *Then*

$$s(t) = \hat{y}(t), \quad \textit{for all } t \in [0, 1]. \tag{37}$$

We see from the above that any algorithm for $\hat{y}(\cdot)$ is an algorithm for $s(\cdot)$ once we replace $\{\lambda_j y\}$ by $\{r_j\}$. We can develop a recursive algorithm for $s(\cdot)$ without computing $K(\cdot,\cdot)$ by using the system that generates $y(\cdot)$ in response to white noise. This system is

$$\begin{aligned} \dot{\mathbf{p}}(t) &= \mathbf{A}\mathbf{p}(t) + \mathbf{b}w(t), \\ y(t) &= \mathbf{c}\mathbf{p}(t), \end{aligned} \tag{38}$$

where $w(\cdot)$ is zero-mean, unit intensity white noise and $\mathbf{A}$, $\mathbf{b}$, $\mathbf{c}$ are as in (5). The initial conditions on (38) are

$$E[\mathbf{p}(0)] = \mathbf{0}, \qquad E[\mathbf{p}(0)w(t)] = \begin{cases} -\left[\begin{array}{c} \mathbf{W}^{-1} \\ \hline \mathbf{0} \end{array}\right] \boldsymbol{\Delta}(t)\mathbf{b}, & t \in [0, t_\alpha], \\ \mathbf{0}, & \text{otherwise}, \end{cases} \tag{39}$$

$$E[\mathbf{p}(0)\mathbf{p}'(0)] = \left[\begin{array}{c} \mathbf{W}^{-1} \\ \hline \mathbf{0} \end{array}\right] \left[\mathbf{I} + \int_0^{t_\alpha} \boldsymbol{\Delta}(\tau)\mathbf{b}\mathbf{b}'\boldsymbol{\Delta}'(\tau)\, d\tau\right] \left[\begin{array}{c} \mathbf{W}^{-1} \\ \hline \mathbf{0} \end{array}\right]'. \tag{40}$$

The derivation of (38)–(40) again follows that in [19]. The process $y(\cdot)$ is called an autoregressive-moving average process.

5. Recursive ARMA spline algorithm. We now give a recursive algorithm for the ARMA spline. The derivation, which is similar to that in [19], is based on the stochastic correspondence discussed in the previous section; that is, the estimation problem is solved recursively for $\hat{y}(\cdot)$ using (38)–(40), and then $\{\lambda_j y\}$ are replaced by their sample values $\{r_j\}$.

Now

$$s(t) = \mathbf{c}\mathbf{x}(t/N) \tag{41}$$

where the n-vector $\mathbf{x}(\cdot/N)$ is computed via the following steps.

Step 1. Initialization.

Let

$$\mathbf{x}(t_\alpha/\alpha) = \boldsymbol{\phi}(t_\alpha) \left[\begin{array}{c} \mathbf{W}^{-1} \\ \hline \mathbf{0} \end{array}\right] \mathbf{r}_0,$$

$$\mathbf{r}_0 = [r_1, r_2, \cdots, r_\alpha]'.$$

Let

$$\mathbf{P}(t_\alpha/\alpha) = \boldsymbol{\phi}(t_\alpha) \int_0^{t_\alpha} \left\{\left[\begin{array}{c} \mathbf{W}^{-1} \\ \hline \mathbf{0} \end{array}\right] \boldsymbol{\Delta}(\tau) - \boldsymbol{\phi}(-\tau)\right\} \mathbf{b}\mathbf{b}' \left\{\left[\begin{array}{c} \mathbf{W}^{-1} \\ \hline \mathbf{0} \end{array}\right] \boldsymbol{\Delta}(\tau) - \boldsymbol{\phi}(-\tau)\right\}' d\tau\, \boldsymbol{\phi}'(t_\alpha).$$

Step 2. Compute the following for $\alpha \leqq j \leqq N-1$:

$$\varepsilon_{j+1} = r_{j+1} - \mathbf{h}_{j+1}\mathbf{x}(t_{j+1}/j),$$

$$R_{j+1} = \mathbf{h}_{j+1}\mathbf{P}(t_{j+1}/j)\mathbf{h}'_{j+1},$$

$$\mathbf{K}_{j+1} = \mathbf{P}(t_{j+1}/j)\mathbf{h}'_{j+1},$$

where

$$\dot{\mathbf{x}}(t/j) = \mathbf{A}\mathbf{x}(t/j), \qquad t_j \leqq t \leqq t_{j+1},$$
$$\dot{\mathbf{P}}(t/j) = \mathbf{A}\mathbf{P}(t/j) + \mathbf{P}(t/j)\mathbf{A}' + \mathbf{b}\mathbf{b}', \qquad t_j \leqq t \leqq t_{j+1},$$
$$\mathbf{x}(t_{j+1}/j+1) = \mathbf{x}(t_{j+1}/j) + \mathbf{K}_{j+1}R_{j+1}^{-1}\varepsilon_{j+1},$$
$$\mathbf{P}(t_{j+1}/j+1) = \mathbf{P}(t_{j+1}/j) - \mathbf{K}_{j+1}R_{j+1}^{-1}\mathbf{K}'_{j+1}.$$

Step 3. Starting with the value of $\mathbf{x}(t_N/N)$ obtained from Step 2, compute $\mathbf{x}(t/N)$ by integrating

$$\dot{\mathbf{x}}(t/N) = \mathbf{A}\mathbf{x}(t/N) + \mathbf{b}\mathbf{b}'\boldsymbol{\mu}(t) \tag{42}$$

where for $1 \leqq j \leqq N$

$$\boldsymbol{\mu}(t) = \begin{cases} \boldsymbol{\mu}_j(t), & t_{j-1} < t < t_j, \\ \mathbf{0}, & t > t_N, \end{cases} \tag{43}$$

and

$$\dot{\boldsymbol{\mu}}_j(t) = -\mathbf{A}'\boldsymbol{\mu}_j(t), \qquad t_{j-1} \leqq t \leqq t_j, \tag{44}$$

$$\boldsymbol{\mu}_j(t_j) = \begin{cases} \mathbf{h}'_N R_N^{-1}\varepsilon_N, & j = N, \\ [\mathbf{I} - \mathbf{h}'_j R_j^{-1}\mathbf{K}'_j]\boldsymbol{\mu}_{j+1}(t_j) + \mathbf{h}'_j R_j^{-1}\varepsilon_j, & \alpha + 1 \leqq j \leqq N - 1, \\ \boldsymbol{\mu}_{j+1}(t_j) - \mathbf{h}'_j\beta_j, & 1 \leqq j \leqq \alpha, \end{cases} \tag{45}$$

$$[\beta_1, \beta_2, \cdots, \beta_\alpha]' = \left[\frac{\mathbf{W}^{-1}}{\mathbf{0}}\right]' \boldsymbol{\phi}'(t_\alpha)\boldsymbol{\mu}_{\alpha+1}(t_\alpha).$$

6. ARMA spline structural properties. In this section we shall restrict attention to point evaluation constraint functionals; that is

$$\lambda_j f = f(t_j), \qquad 1 \leqq j \leqq N,$$

or, equivalently,

$$\mathbf{h}_j = \mathbf{c}, \qquad 1 \leqq j \leqq N.$$

To begin, note that the structure of the matrices **A**, **b**, **c** given in (5) implies

$$\mathbf{c}\mathbf{A}^i\mathbf{b} = \begin{cases} 0, & 0 \leqq i \leqq \alpha - 2, \\ b_{i+1}, & \alpha - 1 \leqq i \leqq n - 1. \end{cases} \tag{46}$$

By successively differentiating (41) and using (42), (44) and (46), we get for $t \notin \{t_i\}_1^N$,

$$s^{(j)}(t) = \begin{cases} \mathbf{c}\mathbf{A}^j\mathbf{x}(t/N), & 0 \leqq j \leqq \alpha - 1, \\ \mathbf{c}\mathbf{A}^j\mathbf{x}(t/N) + \left[\sum_{k=\alpha-1}^{j-1} (-1)^{j-1-k}\mathbf{c}\mathbf{A}^k\mathbf{b}\mathbf{b}'(\mathbf{A}')^{j-1-k}\right]\boldsymbol{\mu}(t), & j \geqq \alpha. \end{cases} \tag{47}$$

Now the Cayley–Hamilton theorem and (5a) imply (here $a_n = 1$)

$$\sum_{j=0}^{n} a_j\mathbf{A}^j = \mathbf{0} = \sum_{j=0}^{n} a_j(\mathbf{A}')^j. \tag{48}$$

Then using (47) and (48),

$$
\begin{aligned}
Ls(t) &= \sum_{j=0}^{n} a_j s^{(j)}(t) \\
&= \mathbf{c}\left[\sum_{j=0}^{n} a_j \mathbf{A}^j\right]\mathbf{x}(t/N) + \left[\sum_{j=\alpha}^{n}\ \sum_{k=\alpha-1}^{j-1} (-1)^{j-1-k} a_j \mathbf{c}\mathbf{A}^k \mathbf{b}\mathbf{b}'(\mathbf{A}')^{j-1-k}\right]\boldsymbol{\mu}(t) \qquad (49)\\
&= \boldsymbol{\xi}'\boldsymbol{\mu}(t).
\end{aligned}
$$

Now if L^* is the formal adjoint of L,

$$L^*Ls(t) = \sum_{j=0}^{n} (-1)^j a_j (Ls)^{(j)}(t).$$

Differentiating (49) we get

$$(Ls)^{(j)}(t) = (-1)^j \boldsymbol{\xi}'(\mathbf{A}')^j \boldsymbol{\mu}(t)$$

and thus, using (48),

$$L^*Ls(t) = \boldsymbol{\xi}'\left[\sum_{j=0}^{n} a_j(\mathbf{A}')^j\right]\boldsymbol{\mu}(t) = 0, \qquad t \notin \{t_i\}_1^N. \tag{50}$$

Next we will examine the behavior of derivatives of $s(\cdot)$ at the points $\{t_i\}_1^N$. From (47), we have for $0 \leqq j \leqq \alpha - 1$,

$$s^{(j)}(t_i+) - s^{(j)}(t_i-) = 0, \qquad 1 \leqq i \leqq N, \tag{51}$$

and for $j \geqq \alpha$,

$$s^{(j)}(t_i+) - s^{(j)}(t_i-) = \begin{cases} \boldsymbol{\sigma}_j'[\boldsymbol{\mu}_{i+1}(t_i) - \boldsymbol{\mu}_i(t_i)], & 1 \leqq i \leqq N-1, \\ -\boldsymbol{\sigma}_j'\,\boldsymbol{\mu}_N(t_N), & i = N, \end{cases} \tag{52}$$

where the vector $\boldsymbol{\sigma}_j'$ is the bracketed term in (47). Now from (45),

$$\boldsymbol{\mu}_{i+1}(t_i) - \boldsymbol{\mu}_i(t_i) = \begin{cases} \mathbf{c}'\beta_i, & 1 \leqq i \leqq \alpha, \\ \mathbf{c}'[R_i^{-1}\mathbf{K}_i'\boldsymbol{\mu}_{i+1}(t_i) - R_i^{-1}\varepsilon_i], & \alpha+1 \leqq i \leqq N-1, \end{cases} \tag{53}$$

and $\boldsymbol{\mu}_N(t_N) = \mathbf{c}'R_N^{-1}\varepsilon_N$. Equation (46) implies

$$\boldsymbol{\sigma}_j'\mathbf{c}' = 0, \qquad \alpha \leqq j \leqq 2\alpha - 2,$$

and therefore, using (52), (53),

$$s^{(j)}(t_i+) - s^{(j)}(t_i-) = 0, \qquad \alpha \leqq j \leqq 2\alpha-2, \qquad 1 \leqq i \leqq N. \tag{54}$$

Now for $j \geqq 2\alpha - 1$,

$$\boldsymbol{\sigma}_j'\mathbf{c}' = \sum_{k=\alpha-1}^{j-\alpha} (-1)^{j-1-k}\mathbf{c}\mathbf{A}^k\mathbf{b}\mathbf{b}'(\mathbf{A}')^{j-1-k}\mathbf{c}'. \tag{55}$$

The alternating sign in (55) implies

$$\boldsymbol{\sigma}_j'\mathbf{c}' = 0, \qquad j = 2\alpha,\ 2\alpha+2, 2\alpha+4, \cdots,$$

and thus for $1 \leqq i \leqq N$,

$$s^{(j)}(t_i+) - s^{(j)}(t_i-) = 0, \qquad j = 2\alpha,\ 2\alpha+2,\ 2\alpha+4, \cdots.$$

The above results are summarized in the following theorem.

THEOREM 6.1. *The ARMA spline* $s(\cdot)$ *of Definition* 2.1 *satisfies*

$$\text{(i) } L^*Ls(t)=0, \qquad t\notin\{t_i\}_1^N,$$

$$\text{(ii) } s(\cdot)\in C^{2\alpha-2}[0,1].$$

Jumps in derivatives of odd order greater than $2\alpha-2$ can be computed via (52), (53), (55). Now recall [1] that Lg-splines interpolating point evaluation functionals satisfy condition (i), but instead of (ii), all of the first $2n-2$ derivatives are continuous on $[0, 1]$. Thus the introduction of numerator dynamics via the operator M of (3) does not change the functional form of the spline; it simply produces discontinuities in derivatives of order greater than $2\alpha-2$.

It is clear from § 2 that the operator T, and therefore the ARMA spline $s(\cdot)$, depends on the value of t_0. In § 3 we chose $t_0=0$ because any other choice produces a spline with unsymmetric continuity properties. For example, if $t_0=t_\alpha$, the spline satisfies conditions (i)–(ii) above except that at t_α only the first $\alpha-1$ derivatives are continuous.

7. Example. In this section we will solve for the ARMA spline in the case that $L=D^2$ and $M=D+1$, and $\lambda_j f=f(t_j)$, $j=1,2$. Thus $\alpha=1$ and $N=2$. This choice for L and M implies

$$\mathbf{A}=\begin{bmatrix}0&1\\0&0\end{bmatrix},\qquad \mathbf{b}=[1\quad 1]',\qquad \mathbf{c}=[1\quad 0],\qquad \boldsymbol{\phi}(t)=\begin{bmatrix}1&t\\0&1\end{bmatrix},$$

$$W=1,\qquad \boldsymbol{\Delta}(t)=[1\quad t_1-t]\quad \text{if } t\in[0,t_1].$$

From results in the Appendix, the operator T is given by

$$Tf(t)=\dot f(t)-\int_0^t e^{-(t-\tau)}\dot f(\tau)\,d\tau.$$

Using the algorithm of § 5, we have from Steps 1 and 2,

$$\mathbf{x}(t_1/1)=[r_1\quad 0]',\qquad \mathbf{P}(t_1/1)=\begin{bmatrix}0&0\\0&t_1\end{bmatrix},$$

$$\mathbf{x}(t_2/1)=[r_1\quad 0]',\qquad \varepsilon_2=r_2-r_1,$$

$$\mathbf{K}_2=\begin{bmatrix}(1/3)[(1+t_2-t_1)^3-1]+t_1(t_2-t_1)^2\\(1/2)[(1+t_2-t_1)^2-1]+t_1(t_2-t_1)\end{bmatrix},\qquad R_2=(1/3)[(1+t_2-t_1)^3-1]+t_1(t_2-t_1)^2,$$

$$\mathbf{x}(t_2/2)=[r_1\quad 0]'+\mathbf{K}_2R_2^{-1}(r_2-r_1)$$

and from Step 3,

$$\mathbf{b}'\boldsymbol{\mu}_2(t)=R_2^{-1}(1+t_2-t)(r_2-r_1),\qquad \mathbf{b}'\boldsymbol{\mu}_1(t)=R_2^{-1}(t_2-t_1)(r_2-r_1),$$

$$s(t)=\begin{cases} r_2+R_2^{-1}((1/2)(1+t_2-t_1)^2-1/2+t_1(t_2-t_1))(r_2-r_1)(t-t_2), & t\geqq t_2,\\ r_2+R_2^{-1}((1/6)(1+t_2-t)^3-(1/2)(1+t_2-t)^2+(t-t_2)((1/2)(1+t_2-t_1)^2 & \\ \qquad\qquad +t_1(t_2-t_1))+(1/3)(r_2-r_1), & t_1\leqq t\leqq t_2,\\ r_1+R_2^{-1}((1/2)(1+t-t_1)^2+t_1(t-t_1)(-1/2)(t_2-t_1)(r_2-r_1), & t\leqq t_1,\end{cases}$$

The structural properties of Theorem 6.1 can now easily be checked.

8. Conclusions. The familiar Lg-spline is associated with an all-pole linear system. We have generalized this by introducing an autoregressive-moving average spline that is associated with a linear system that has zeros (numerator dynamics). The concept of reduced-order system inverse was used to define the ARMA spline. Reproducing kernel Hilbert space methods were used to establish the fact that the ARMA spline is a sample function of a certain least-squares estimate. A recursive spline algorithm could then be derived using estimation techniques. Finally, we derived the structural properties of ARMA splines and thus showed the precise changes that occur in the Lg-spline when numerator dynamics are introduced.

Generalizations are possible in several directions. First, under mild conditions, our development will carry through for variable-coefficient operators L and M. Second, our results generalize to vector-valued ARMA splines along the lines of [21]. Structural properties for ARMA splines interpolating general extended Hermite–Birkhoff functionals can be derived using the techniques of [20]. When the measurements contain errors and exact interpolation is not desirable, the ARMA spline *smoothing* problem can be formulated and solved in a similar fashion to the Lg-spline case considered in [22].

Finally, interpolation and smoothing error bounds can be derived for ARMA splines with the methods of [23].

Appendix. In this appendix, we will outline the method detailed in [5] for inverting the linear system (4), (5). First, in order to obtain the representation (4)–(6), write the system having transfer function M/L in state-variable form as

$$\begin{aligned}\dot{\boldsymbol{\omega}} &= \mathbf{A}\boldsymbol{\omega} + \tilde{\mathbf{b}}u \\ f &= \tilde{\mathbf{c}}\boldsymbol{\omega}\end{aligned} \tag{A.1}$$

where $\mathbf{A}$ is as in (5a) and

$$\begin{aligned}\tilde{\mathbf{b}} &= [0, \cdots, 0, 1]' \\ \tilde{\mathbf{c}} &= [c_0, c_1, \cdots, c_m, 0, \cdots, 0].\end{aligned} \tag{A.2}$$

It can easily be verified that (A.1)–(A.2) has the designated transfer function. Next let

$$\mathbf{x} = \mathbf{Q}\boldsymbol{\omega}, \tag{A.3}$$

where

$$\mathbf{Q} = [\tilde{\mathbf{c}}'|\mathbf{A}'\tilde{\mathbf{c}}'| \cdots |(\mathbf{A}')^{n-1}\tilde{\mathbf{c}}']'.$$

Substituting for $\boldsymbol{\omega}$ in (A.1) using (A.3), one obtains (4)–(6). To invert (4), differentiate the output α times to obtain (recall (46))

$$f^{(\alpha)} = \mathbf{c}\mathbf{A}^{\alpha}\mathbf{x} + b_{\alpha}u = x_{\alpha+1} + b_{\alpha}u. \tag{A.4}$$

Now let

$$\begin{aligned}\boldsymbol{\theta} &= [x_{\alpha+1}, \cdots, x_n]' \\ \hat{\mathbf{b}} &= [b_{\alpha+1}, \cdots, b_n]'\end{aligned}$$

and partition $\mathbf{A}$ as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

where $\mathbf{A}_{11}$ is $\alpha \times \alpha$. Then, recalling (7), we have

$$\dot{\boldsymbol{\theta}} = \mathbf{A}_{21}[f, \cdots, f^{(\alpha-1)}]' + \mathbf{A}_{22}\boldsymbol{\theta} + \hat{\mathbf{b}}u. \tag{A.5}$$

If $\mathbf{e}_1' = [1, 0, \cdots, 0]$, then (A.4) implies

$$u = -b_\alpha^{-1}\mathbf{e}_1'\boldsymbol{\theta} + b_\alpha^{-1}f^{(\alpha)}.$$

Substituting the above into (A.5) yields the reduced-order inverse (8).

REFERENCES

[1] J. Jerome and L. L. Schumaker, *On Lg-splines*, J. Approximation Theory, 2 (1969), pp. 29–49.

[2] H. L. Weinert and G. S. Sidhu, *A spline-theoretic approach to minimum-energy control. Part II: Systems with numerator dynamics*. Tech. Report 75-11, Electrical Engineering Dept., The Johns Hopkins Univ., Baltimore, Md., 1975.

[3] H. L. Weinert and T. Kailath, *A spline-theoretic approach to minimum energy control*, IEEE Trans. Automatic Control, AC-21 (1976), pp. 391–393.

[4] R. J. P. de Figueiredo and A. N. Netravali, *On a class of minimum energy controls related to spline functions*, IEEE Trans. Automatic Control, AC-21 (1976), pp. 725–727.

[5] L. M. Silverman, *Properties and applications of inverse systems*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 436–437.

[6] M. Atteia, *Generalization of the definition and properties of spline functions*, C.R. Acad. Sci. Paris, 260 (1965), pp. 3550–3553.

[7] ———, *Generalized spline functions*, C.R. Acad. Sci. Paris, 261 (1965), pp. 2149–2152.

[8] P. M. Anselone and P. J. Laurent, *A general method for the construction of interpolating or smoothing spline functions*, Numer. Math., 12 (1968), pp. 66–82.

[9] A. Sard, *Optimal approximation*, J. Functional Analysis, 1 (1967), pp. 222–244.

[10] R. J. P. de Figueiredo, *LM-g splines*, J. Approximation Theory, 19 (1977), pp. 332–360.

[11] H. L. Weinert and G. S. Sidhu, *Recursive computation of L-splines based on stochastic least-squares estimation*. Proc. Johns Hopkins Conf. Information Sciences and Systems (1975), pp. 318–321.

[12] T. Kailath and R. Geesey, *An innovations approach to least-squares estimation—Part V: Innovations representations and recursive estimation in colored noise*, IEEE Trans. Automatic Control, AC-18 (1973), pp. 435–453.

[13] T. Kailath, R. Geesey and H. L. Weinert, *Some relations among RKHS norms, Fredholm equations, and innovations representations*, IEEE Trans. Information Theory, IT-18 (1972), pp. 341–348.

[14] H. L. Weinert and G. S. Sidhu, *On uniqueness conditions for optimal curve fitting*, J. Optimization Theory Appl., 23 (1977), pp. 211–216.

[15] N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68 (1950), pp. 337–404.

[16] J. Hajek, *On linear statistical problems in stochastic processes*, Czech. Math. J., 12 (1962), pp. 404–444.

[17] E. Parzen, *A new approach to the synthesis of optimal smoothing and prediction systems*, Mathematical Optimization Techniques, Univ. of Calif. Press, Berkeley, 1963, pp. 75–108.

[18] G. Wahba, *On the numerical solution of Fredholm integral equations of the first kind*, Tech. Report 217, Dept. of Statistics, Univ. of Wisconsin, Madison, WI, 1969.

[19] H. L. Weinert and G. S. Sidhu, *A stochastic framework for recursive computation of spline functions. Part I: Interpolating splines*, IEEE Trans. Information Theory, IT-24 (1978), pp. 45–50.

[20] G. S. Sidhu and H. L. Weinert, *Dynamical recursive algorithms for Lg-spline interpolation of EHB data*, Applied Math. Comput., (1979) to appear.

[21] ———, *Vector-valued Lg-splines, I: Interpolating splines*, Tech. Report 154, Institute for Research in Applied Mathematics and Systems, National Univ. of Mexico, Mexico City, 1977.

[22] H. L. Weinert, R. H. Byrd and G. S. Sidhu, *A stochastic framework for recursive computation of spline functions. Part II: Smoothing splines*, J. Optimization Theory Appl. (1979), to appear.

[23] H. L. Weinert, *Statistical methods in optimal curve fitting*, Communications in Statistics, B7 (1978), no. 4, pp. 417–435.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 4, July 1979

0363-0129/79/1704-0008$01.00/0

THE RICCATI INTEGRAL EQUATIONS FOR OPTIMAL CONTROL PROBLEMS ON HILBERT SPACES*

J. S. GIBSON†

Abstract. The two Riccati integral equations for linear-quadratic control problems involving evolution operators on Hilbert spaces are derived and shown to have a common solution, which yields the closed-loop structure of the optimal control. Riccati integral equations, instead of differential equations, arise because evolution operators are used to represent system dynamics. The operator representing the closed-loop control perturbs the evolution operator representing the uncontrolled system to produce a second evolution operator, representing the optimally controlled system, hence the two Riccati integral equations in terms of these two evolution operators, respectively. Having both Riccati integral equations facilitates the extension of the analysis of optimal control on finite time intervals to the analysis of optimal control on infinite time intervals, and then existence, uniqueness, and stability results for periodic solutions of the Riccati equations are obtained. Finally, sufficient conditions are given for convergence of approximate solutions of optimal control problems on both finite and infinite time intervals.

1. Introduction. In recent years, investigation of the optimal control problem for infinite dimensional, linear dynamical systems with quadratic cost criteria and unconstrained controls has led, as would be expected from experience with finite dimensional systems, to a feedback control, whose structure is given by the solution of a Riccati differential equation. The systems are usually described either by partial differential equations—see Lions [10], Lukes and Russell [11], and Vinter and Johnson [13] for example—or by functional differential equations—see Curtain [1], Datko [4], and Delfour and Mitter [7] for example. It has been found natural to model such systems by representing the state and control vectors as elements of infinite dimensional Hilbert spaces, so that the linear-quadratic nature of the control problem leads to the minimization of a quadratic functional on a Hilbert space.

Different types of differential equations on infinite dimensional spaces lead to different types of solutions—strong, weak, mild, etc.—and, for this reason, although experience suggests that almost any linear-quadratic control problem entails an optimal control with linear feedback structure, most results to date are limited to problems involving particular types of partial or functional differential equations. Thus, a mathematical model is needed that is sufficiently general to encompass a very wide variety of dynamical systems, but sufficiently structured to allow the important properties of the optimal control to be deduced. Such a model is available in the concept of an evolution operator, which is usually the fundamental solution, in an appropriate sense, of a differential equation. Lions, Lukes and Russell, Delfour and Mitter, and others showed that their systems could be represented by evolution operators, but, in varying degrees, these authors relied upon the properties of particular classes of differential equations. Of course, to obtain certain useful information for a given problem—for example, convergence of numerical computations or smoothness of solutions—it is necessary to refer to the specific type of differential equation involved; however, the feedback structure of the optimal control, the dependence of the cost on the initial conditions, and appropriate Riccati equations can be derived without reference to a differential equation, when the system dynamics are represented by a very general evolution operator. Actually, the closed-loop structure of the optimal control seems to follow more naturally from this formulation than from formulations explicitly involving differential equations, where the linear feedback control often appears to be a lucky

*Received by the editors May 18, 1977, and in revised form August 30, 1978.

†Mechanics and Structures Department, School of Engineering and Applied Science, University of California, Los Angeles, California 90024. This research was supported partially by NSF Grant ENG78-04753.

guess. From this more general formulation, the linear feedback structure of the optimal control follows as the direct result of two properties of the optimal control problem: its linear-quadratic nature, and the principle that any segment of an optimal trajectory is an optimal trajectory.

For control systems described by evolution operators, the input-output relations are integral equations, and the structure of the optimal control is given by the solution of Riccati integral equations. There are two Riccati integral equations, which are shown in this paper to have a common solution. One equation involves the evolution operator which represents the original dynamics of the control system, and the other involves the "perturbed" evolution operator which represents the dynamics of the system after it has been modified to implement the feedback control. The reader who is familiar with finite dimensional control theory should be reassured to know that, at least formally, both Riccati integral equations yield the same differential equation; however, in infinite dimensions, the Riccati differential equation is not always well posed. For instance, in [2], Curtain and Pritchard were unable to establish the uniqueness of a solution of their Riccati differential equation resulting from Lions' parabolic systems, although the solution of Curtain and Pritchard's Riccati integral equation for these systems is unique.

Lukes and Russell used the Riccati integral equation involving the semigroup representing the original dynamics of a class of autonomous distributed systems, and Delfour and Mitter derived the Riccati integral equation involving the perturbed evolution operator representing the optimal control scheme for their hereditary systems. But the most comprehensive results to date on Riccati integral equations are the results of Curtain and Pritchard in [2], where the Riccati integral equation involving the perturbed evolution operator was derived for the control problem on a finite time interval, using only the properties of a very general evolution operator. Also for control systems modeled by evolution operators, Datko, in [3] and [4], obtained some quite useful results on the structure of the optimal control for problems on both finite and infinite time intervals; however, Datko did not obtain a Riccati equation.

Although the technicalities involved in demonstrating that the solution of a particular type of differential equation is given by an evolution operator should not be underestimated, the references cited here and many others have established ample precedent for studying control systems represented by evolution operators. The purpose of this paper then is to derive as much as possible without reference to a differential equation, for only by avoiding the use of differential equations altogether can one hope to obtain results which hold for all differential systems. For problems on both finite and infinite intervals, the two Riccati integral equations and the relationships between their respective solutions are derived, using only the properties of an evolution operator of sufficient generality to represent almost any realistic linear dynamical control system. Having both Riccati integral equations is especially useful for analysis of optimal control on the infinite interval, and existence, uniqueness, and stability results for periodic solutions of the Riccati integral equations follow rather easily from this analysis. As a special case of the periodic problem, the Hilbert space version of the familiar Riccati algebraic equation is obtained. Finally, with a view to numerical computations, sufficient conditions are given for the solutions to a sequence of finite dimensional optimal control problems to converge to the solution to an infinite dimensional problem.

2. Evolution operators. In the theory of partial and functional differential equations, several definitions have been used for evolution operators. The variations in these definitions involve the types of continuity and differentiability assumed for the evolution operators. So that the results here will not depend upon any specific type of differentiation, we will make no assumption of differentiability for our evolution operators, and the properties we assume should leave us with a class of evolution operators of sufficient generality to include practically all well posed linear models of realistic dynamical systems.

DEFINITION 2.1 Let $-\infty < t_0 \leqq t_f < \infty$, and let H be a real Hilbert space. $T(\cdot, \cdot)$: $\{(t, s): t_0 \leqq s \leqq t \leqq t_f\} \to \mathscr{L}(H, H)$ is an *evolution operator* if

$$T(t, r)T(r, s) = T(t, s), \qquad t_0 \leqq s \leqq r \leqq t \leqq t_f, \tag{2.1}$$

$$T(t, t) = I, \tag{2.2}$$

(2.3) $T(t, s)$ is strongly contiriuous in s on $[t_0, t]$ and strongly continuous in t on $[s, t_f]$.

This definition includes the evolution operators of all the references cited in the Introduction, except the "mild" evolution operator of Curtain and Pritchard in [2], where weak continuity was used in (2.3). The careful reader should observe that the important results here concerning the optimal control and the Riccati integral equations hold if we replace strong continuity of $T(\cdot, \cdot)$ with weak continuity and require H to be separable. We need either strong continuity in (2.3) or weak continuity and separability of H to guarantee strong measurability of $T(\cdot, \cdot)$ in either argument (see Appendix A), which will be needed in our integral equations. We assume strong continuity in (2.3) because the response of a physical system normally should be continuous in an appropriate space. Also, since strong measurability of $T(\cdot, \cdot)$ implies only weak measurability of the adjoint operator $T^*(\cdot, \cdot)$, we must require additionally that $T^*(\cdot, \cdot)$ be strongly measurable in either argument. This is certainly the case if either $T^*(\cdot, \cdot)$ is strongly continuous, as in (2.3), or H is separable.

We will need one more condition on $T(\cdot, \cdot)$: we assume that there is a constant M_1 such that

$$\|T(t, s)\| \leqq M_1, \qquad t_0 \leqq s \leqq t \leqq t_f. \tag{2.4}$$

It should be noted that the uniform boundedness of $\|T(\cdot, \cdot)\|$ does not follow from Definition 2.1, as Curtain and Pritchard mistakenly supposed in [2]. (For a counter example, see Appendix B.) However, this inaccuracy does not diminish the usefulness of [2]; the norm of the evolution operator is uniformly bounded on finite intervals of time for practically any realistic linear system. If $T(\cdot, \cdot)$ is jointly weakly continuous, (2.4) follows from the principle of uniform boundedness; and, even without joint continuity, reasonable dependence on initial data will usually guarantee (2.4).

While we do not assume any differentiability for $T(\cdot, \cdot)$, the reader may find it helpful to compare the results of this paper with more classical theory by differentiating formally some of the integral equations here, using

$$\frac{d}{dt}T(t, s) = A(t)T(t, s), \qquad \frac{d}{ds}T(t, s) = -T(t, s)A(s). \tag{2.5}$$

Such formal differentiation of the Riccati integral equations should be especially enlightening. Of course, in finite dimensions, our equations may be differentiated immediately to obtain the standard results for control systems governed by ordinary differential equations.

The following perturbation theorem from [2] will be very useful.

THEOREM 2.1. *Let $T(\cdot, \cdot)$ be an evolution operator which is uniformly bounded as* *in* (2.4), *and let C be in* $\mathcal{B}_\infty(t_0, t_f; H, H)$.[1] Then the operator integral equation

$$S(t,s)x = T(t,s)x + \int_s^t T(t,\eta)C(\eta)S(\eta,s)x\,d\eta, \qquad x \in H, \tag{2.6}$$

has a unique solution $S(\cdot,\cdot)$ *in the class of strongly continuous (as in* (2.3)) *bounded linear operators on* H. $S(\cdot,\cdot)$ *is an evolution operator and is called the* perturbed evolution operator corresponding to the perturbation of $T(\cdot,\cdot)$ by C. $S(\cdot,\cdot)$ *is also the unique solution of*

$$S(t,s)x = T(t,s)x + \int_s^t S(t,\eta)C(\eta)T(\eta,s)x\,d\eta, \qquad x \in H; \tag{2.7}$$

i.e., $T(\cdot,\cdot)$ *is the perturbed evolution operator corresponding to the perturbation of* $S(\cdot,\cdot)$ *by* $-C$. *If* M_1 *is the uniform bound of* (2.4), *we have*

$$\|S(t,s)\| \leqq M_1 \exp\left(M_1\|C\|_{\mathcal{B}_\infty}(t-s)\right). \tag{2.8}$$

The theorem is a generalization of a similar perturbation result for semigroups, given by Phillips in [12] and later in [8]. The proof, which is given in detail in [2], is based upon the construction of $S(\cdot,\cdot)$ according to

$$S(t,s) = \sum_{n=0}^{\infty} S_n(t,s), \tag{2.9}$$

where

$$S_0(t,s) = T(t,s),\ S_n(t,s)x = \int_s^t T(t,\eta)C(\eta)S_{n-1}(\eta,s)x\,d\eta. \tag{2.10}$$

For analysis of optimal control on the infinite interval, we will need an asymptotic stability result for linear evolution operators. Suppose (2.1)–(2.3) hold for $t_0 \leqq s \leqq t < \infty$. We say that $T(\cdot,\cdot)$ is uniformly exponentially bounded if there are constants M_2 and α such that

$$\|T(t,s)x\| \leqq M_2\, e^{\alpha(t-s)}, \qquad t_0 \leqq s \leqq t < \infty. \tag{2.11}$$

We say that $T(\cdot,\cdot)$ is uniformly exponentially stable if (2.11) holds for some $\alpha < 0$. The following theorem is due to Datko [5].

THEOREM 2.2. *Suppose* (2.11) *holds for some* M_2 *and* α. *Then* $T(\cdot,\cdot)$ *is uniformly exponentially stable if and only if there exists a constant* M_3 *such that, for all* $x \in H$ *and* $t_0 \leqq s < \infty$.

$$\int_s^\infty \|T(t,s)x\|^2\,dt \leqq M_3\|x\|^2. \tag{2.12}$$

Datko proved this result for H a Banach space and noted that it remains valid if the power 2 is replaced by any p, $1 \leqq p < \infty$, in (2.12). Also, though Datko did not give a decay rate explicitly in terms of the original bounds M_2, M_3, and α, a rephrasing of his proof shows that, if $M_3\alpha \geqq 1$,

$$\|T(t,s)\| \leqq M_4\, e^{-\beta(t-s)}, \qquad t_0 \leqq s \leqq t < \infty, \tag{2.13}$$

[1] $\mathcal{B}_\infty(t_0, t_f; H, H)$ is the Banach space of strongly measurable, essentially bounded functions from (t_0, t_f) to $\mathcal{L}(H, H)$. Also, the vector valued integrals in Theorem 2.1 and throughout this paper are Bochner integrals. For the results concerning strong measurability and Bochner integration that are important to this paper, see Appendix A.

where

$$M_4 = 4M_2M_3\alpha \quad \text{and} \quad \beta = (16M_2^2M_3^2\alpha)^{-1}\ln 2. \tag{2.14}$$

While this is not the sharpest estimate possible, it will be quite useful for Theorem 5.3 to know that M_4 and β can be given in terms of M_2, M_3, and α.

3. The optimal control problem. We consider an evolution process defined by

$$x(t) = T(t, s)x(s) + \int_s^t T(t, \eta)B(\eta)u(\eta)\,d\eta, \qquad t_0 \leqq s \leqq t \leqq t_f < \infty, \tag{3.1}$$

where $x(t)$ is in a real Hilbert space H, $T(\cdot,\cdot)$ is an evolution operator on H, $u \in L_2(t_0, t_f; U)$ where U is a real Hilbert space, and $B \in \mathcal{B}_\infty(t_0, t_f; U, H)$ and $B^* \in \mathcal{B}_\infty(t_0, t_f; H, U)$.[2] The optimal control problem is to find a control u which minimizes the cost functional

$$J(t_0, x(t_0), u) = \langle Gx(t_f), x(t_f)\rangle_H + \int_{t_0}^{t_f} (\langle D(t)x(t), x(t)\rangle_H + \langle Q(t)u(t), u(t)\rangle_U)\,dt; \tag{3.2}$$

where $x(t)$ is given by (3.1) with $s = t_0$, $G \in \mathcal{L}(H, H)$ is self-adjoint and nonnegative, and $D \in \mathcal{B}_\infty(t_0, t_f; H, H)$ and $Q \in \mathcal{B}_\infty(t_0, t_f; U, U)$ are self-adjoint and nonnegative, with $Q(t) \geqq m$ for some $m > 0$, for almost all t.

Let us denote $L_2(t, t_f; H)$ by $\mathcal{H}_t$ and $L_2(t, t_f; U)$ by $\mathcal{U}_t$, for $t_0 \leqq t \leqq t_f$. Define $T_t \in \mathcal{L}(H, \mathcal{H}_t)$, $\mathcal{T}_t \in \mathcal{L}(\mathcal{H}_t, \mathcal{H}_t)$, and $\mathcal{F}_t \in \mathcal{L}(\mathcal{H}_t, H)$ by

$$(T_t x)(s) = T(s, t)x, \qquad x \in H, \tag{3.3}$$

$$(\mathcal{T}_t \phi)(s) = \int_t^s T(s, \eta)\phi(\eta)\,d\eta, \qquad \phi \in \mathcal{H}_t, \tag{3.4}$$

$$\mathcal{F}_t \phi = (\mathcal{T}_t \phi)(t_f), \qquad \phi \in \mathcal{H}_t. \tag{3.5}$$

Then we can write (3.2) as

$$\begin{aligned} J(t_0, x(t_0), u) = {}& \langle G(T(t_f, t_0)x(t_0) + \mathcal{F}_{t_0}Bu), (T(t_f, t_0)x(t_0) + \mathcal{F}_{t_0}Bu)\rangle_H \\ &+ \langle D(T_{t_0}x(t_0) + \mathcal{T}_{t_0}Bu), (T_{t_0}x(t_0) + \mathcal{T}_{t_0}Bu)\rangle_{\mathcal{H}_{t_0}} + \langle Qu, u\rangle_{\mathcal{U}_{t_0}}. \end{aligned} \tag{3.6}$$

Under the hypotheses on G, D, and Q, there exists a unique u which minimizes J and this u is the unique solution of

$$J'(t_0, x(t_0), u)v = 0, \quad \forall v \in \mathcal{U}_{t_0}, \tag{3.7}$$

where $J'(t_0, x(t_0), u)v$ means the Fréchet derivative of J at u, applied to v. From (3.6), we have

$$\begin{aligned} J'(t_0, x(t_0), u)v = {}& 2\langle G(T(t_f, t_0)x(t_0) + \mathcal{F}_{t_0}Bu), \mathcal{F}_{t_0}Bv\rangle_H \\ &+ 2\langle D(T_{t_0}x(t_0) + \mathcal{T}_{t_0}Bu), \mathcal{T}_{t_0}Bv\rangle_{\mathcal{H}_{t_0}} + 2\langle Qu, v\rangle_{\mathcal{U}_0}. \end{aligned} \tag{3.8}$$

By identifying H and U, and therefore $\mathcal{H}_t$ and $\mathcal{U}_t$, with their respective duals, we may write (3.8) as

$$J'(t_0, x(t_0), u)v = 2\langle \tilde{Q}_{t_0}u + \tilde{B}_{t_0}^* x(t_0), v\rangle_{\mathcal{U}_{t_0}}, \tag{3.9}$$

[2] $\mathcal{B}_\infty(t_0, t_f; U, H)$ is the Banach space of essentially bounded, strongly measurable functions from (t_0, t_f) to $\mathcal{L}(U, H)$. For the justification of the Bochner integral in (3.1), see Appendix A.

where

$$\tilde{Q}_{t_0} = (Q + B^*\mathscr{T}^*_{t_0} D \mathscr{T}_{t_0} B + B^*\mathscr{F}^*_{t_0} G \mathscr{F}_{t_0} B) \in \mathscr{L}(\mathscr{U}_{t_0}, \mathscr{U}_{t_0}) \tag{3.10}$$

and

$$\tilde{B}^*_{t_0} = (B^*\mathscr{T}^*_{t_0} D T_{t_0} + B^*\mathscr{F}^*_{t_0} G T(t_f, t_0)) \in \mathscr{L}(H, \mathscr{U}_{t_0}), \tag{3.11}$$

with $\mathscr{T}^*_{t_0}$ and $\mathscr{F}^*_{t_0}$ given by

$$(\mathscr{T}'_{t_0}\phi)(t) = \int_t^{t_f} T^*(\eta, t)\phi(\eta)\, d\eta \tag{3.12}$$

and

$$(\mathscr{F}^*_{t_0}x)(t) = T^*(t_f, t)x. \tag{3.13}$$

Note that the right sides of (3.12) and (3.13) do not involve t_0.

According to (3.9), a necessary and sufficient condition that u be the optimal control of (3.7) is that

$$u(t) = -(\tilde{Q}_{t_0}^{-1}\tilde{B}^*_{t_0})(t)x(t_0) \equiv -(\tilde{Q}_{t_0}^{-1}\tilde{B}^*_{t_0}x(t_0))(t), \quad \text{a.e. in}\,[t_0, t_f]. \tag{3.14}$$

Note that our hypotheses on the operators involved in (3.10) and (3.11) justify (3.14), where $\tilde{Q}_{t_0}^{-1}\tilde{B}^*_{t_0} \in \mathscr{B}_\infty(t_0, t_f; H, \mathscr{U})$. The composite operator $\tilde{Q}_{t_0}^{-1}\tilde{B}^*_{t_0}$ is clearly in $\mathscr{L}(H, \mathscr{U}_t)$ because $\tilde{Q}_{t_0}^{-1} \in \mathscr{L}(\mathscr{U}_{t_0}, \mathscr{U}_{t_0})$ and $\tilde{B}^*_{t_0} \in \mathscr{B}_\infty(t_0, t_f; H, \mathscr{U})$; since $Q^{-1} \in \mathscr{B}_\infty(t_0, t_f; H, \mathscr{U})$ (see Property A.4 of Appendix A), the stronger statement, $\tilde{Q}_{t_0}^{-1}\tilde{B}^*_{t_0} \in \mathscr{B}_\infty(t_0, t_f; H, \mathscr{U})$, can be proved using (3.10).

By observing that, if $x(\cdot)$ is the optimal trajectory, then, for any $s \in [t_0, t_f]$, the optimal control $u(\cdot)$ of (3.7) and (3.14) must coincide on $[s, t_f]$ with the unique optimal control corresponding to the initial time s and initial state $x(s)$, we see that all of our equations thus far must hold with t_0 replaced by any $s \in [t_0, t_f]$. We can then deduce the feedback structure of the optimal control. We write formally

$$u(s) = -(\tilde{Q}_s^{-1}\tilde{B}^*_s)(s)x(s). \tag{3.15}$$

However, unless we require $B(\cdot)$, $D(\cdot)$, and $Q(\cdot)$ to be piecewise continuous, the meaning of (3.15) is questionable because we cannot be certain about the set of values of s for which $(\tilde{Q}_s^{-1}\tilde{B}^*_s)(s)$ is defined. The only purpose (3.15) will serve here is to show that the linear feedback structure of the optimal control arises quite naturally in the linear-quadratic control problem. We will proceed on the basis of (3.14) only.

From the boundedness assumptions on $D(\cdot)$, $Q(\cdot)$, $B(\cdot)$, and $T(\cdot,\cdot)$, we know that there is a constant M_5 such that

$$\|(\tilde{Q}_s^{-1}\tilde{B}^*_s)(t)\| \leqq M_5, \qquad t_0 \leqq s \leqq t \leqq t_f \quad \text{(a.e.)}. \tag{3.16}$$

Now, if $u \in \mathscr{U}_s$ minimizes $J(s, x, u)$ and $x(t)$ is the corresponding optimal trajectory, we have

$$x(t) = S(t, s)x, \qquad s \leqq t \leqq t_f, \tag{3.17}$$

where $S(t, s) \in \mathscr{L}(H, H)$ is defined by

$$S(t, s)x = T(t, s)x - \int_s^t T(t, \eta)B(\eta)(\tilde{Q}_s^{-1}\tilde{B}^*_s)(\eta)x\, d\eta, \qquad t_0 \leqq s \leqq t \leqq t_f. \tag{3.18}$$

The standard observation that any segment of an optimal trajectory must be an optimal trajectory implies that, for $s \leqq r \leqq t$, the $x(t)$ defined by (3.17) must coincide with the unique optimal trajectory for the initial time r and initial state $x(r)=S(r,s)x$. Therefore,

$$S(t,r)S(r,s)=S(t,s), \qquad t_0 \leqq s \leqq r \leqq t \leqq t_f. \tag{3.19}$$

From (3.18), we see that $S(t,t)=I$; and, using (2.3), (2.4), (3.16), and (3.18), it is easy to show that, for each s, $S(t,s)$ is strongly continuous in t. To conclude that $S(t,s)$ is an evolution operator, we have only to show that, for each t, it is strongly continuous in s.

Since the optimal trajectory is given by (3.17) and the optimal control is given by (3.14) with t_0 replaced by any $s \in [t_0, t_f]$, we have

$$(\tilde{Q}_{s_2}^{-1}\tilde{B}_{s_2}^*)(t)S(s_2,s_1)=(\tilde{Q}_{s_1}^{-1}\tilde{B}_{s_1}^*)(t), \qquad t_0 \leqq s_1 \leqq s_2 \leqq t \leqq t_f \quad \text{(a.e.)}. \tag{3.20}$$

We then have

$$\begin{aligned} S(t,s_2)x-S(t,s_1)x &= T(t,s_2)x-T(t,s_1)x \\ &\quad -\int_{s_2}^{t} T(t,\eta)B(\eta)(\tilde{Q}_{s_2}^{-1}\tilde{B}_{s_2}^*)(\eta)[I-S(s_2,s_1)]x\,d\eta \\ &\quad +\int_{s_1}^{s_2} T(t,\eta)B(\eta)(\tilde{Q}_{s_1}^{-1}\tilde{B}_{s_1}^*)(\eta)x\,d\eta. \end{aligned} \tag{3.21}$$

From (3.18), it is clear that, as $s_2 \to s_1$ or $s_1 \to s_2$, $S(s_2,s_1)x \to x$. Then, in view of (2.3), (2.4), and (3.16), (3.21) shows that $S(t,s)x$ is continuous in s. Therefore, $S(t,s)$ is an evolution operator.

The uniform boundedness of $\|T(t,s)\|$, (3.16), and (3.18) imply that $\|S(t,s)x\|$ is uniformly bounded for each x, so that the principle of uniform boundedness implies that $\|S(t,s)\|$ is uniformly bounded for $t_0 \leqq s \leqq t \leqq t_f$. Also, (3.18) can be used to show that $S^*(\cdot,\cdot)$ has the continuity and measurability properties of $T^*(\cdot,\cdot)$.

We are now in a position to find out more about the structure of the optimal control. We rewrite (3.8) as

$$\begin{aligned} J'(t_0,x(t_0),u)v &= 2\langle Gx(t_f), \mathscr{F}_{t_0}Bv\rangle_H + 2\langle Dx_{t_0}, \mathscr{T}_{t_0}Bv\rangle_{\mathscr{H}_{t_0}} + 2\langle Qu, v\rangle_{\mathscr{U}_{t_0}} \\ &= 2\langle B^*(\mathscr{F}_{t_0}^*Gx(t_f)+\mathscr{T}_{t_0}^*Dx_{t_0})+Qu, v\rangle_{\mathscr{U}_{t_0}}, \end{aligned} \tag{3.22}$$

where $x_{t_0}(t)$ is given by (3.1) with $s=t_0$. If u is the optimal control, (3.7) and (3.22) yield

$$u=-Q^{-1}B^*(\mathscr{F}_{t_0}^*Gx(t_f)+\mathscr{T}_{t_0}^*Dx_{t_0}), \tag{3.23}$$

and, since the optimal trajectory is given by (3.17), (3.23) becomes

$$u(t)=-Q^{-1}(t)B^*(t)P(t)x(t) \quad \text{a.e.}, \tag{3.24}$$

where

$$P(t)x=T^*(t_f,t)GS(t_f,t)x+\int_t^{t_f} T^*(\eta,t)D(\eta)S(\eta,t)x\,d\eta), \quad t_0 \leqq t \leqq t_f, \qquad x \in H. \tag{3.25}$$

(Recall the definitions of $\mathscr{F}_{t_0}^*$ and $\mathscr{T}_{t_0}^*$ in (3.12) and (3.13).)

Using the properties of $T(\cdot,\cdot)$, $T^*(\cdot,\cdot)$, and $S(\cdot,\cdot)$, it is not difficult to show that $P(\cdot)$ is uniformly bounded, weakly continuous, and strongly measurable on $[t_0, t_f]$. If $T^*(\eta,t)$ is strongly continuous in t, $P(t)$ is strongly continuous.[3] Also, it is shown in Appendix A that our hypotheses on Q imply $Q^{-1} \in \mathscr{B}_\infty(t_0, t_f; U, U)$, so that $BQ^{-1}B^*P \in \mathscr{B}_\infty(t_0, t_f; H, H)$. Then (2.6), (3.1), (3.17), and (3.24) show that $S(t,s)$ is the

[3] If $T(\cdot,\cdot)$ is only weakly continuous, $P(\cdot)$ may not be even weakly continuous.

perturbed evolution operator corresponding to the perturbation of $T(t, s)$ by $-BQ^{-1}B^*P$.

We will now derive a Riccati integral equation for $P(t)$. Using (2.7), we can write (3.25) as

$$
\begin{aligned}
P(t)x = {} & T^*(t_f, t)GT(t_f, t)x - T^*(t_f, t)G\int_t^{t_f} S(t_f, \eta)B(\eta)Q^{-1}(\eta)B^*(\eta)P(\eta)T(\eta, t)x\, d\eta \\
& + \int_t^{t_f} T^*(\eta, t)D(\eta)T(\eta, t)x\, d\eta \\
& - \int_t^{t_f} T^*(\eta, t)D(\eta)\int_t^{\eta} S(\eta, \xi)B(\xi)Q^{-1}(\xi)B^*(\xi)P(\xi)T(\xi, t)x\, d\xi\, d\eta.
\end{aligned}
\tag{3.26}
$$

After using Fubini's theorem to interchange the order of integration in the third integral in (3.26) and noting that $T^*(\xi, t) = [T(\xi, \eta)T(\eta, t)]^* = T^*(\eta, t)T^*(\xi, \eta)$, we see that the first and third integrals in (3.26) become

$$
\begin{aligned}
& -\int_t^{t_f} T^*(\eta, t)T^*(t_f, \eta)GS(t_f, \eta)B(\eta)Q^{-1}(\eta)B^*(\eta)P(\eta)T(\eta, t)x\, d\eta \\
& -\int_t^{t_f} T^*(\eta, t)\int_{\eta}^{t_f} T^*(\xi, \eta)D(\xi)S(\xi, \eta)B(\eta)Q^{-1}(\eta)B^*(\eta)P(\eta)T(\eta, t)x d\xi\, d\eta \\
& = -\int_t^{t_f} T^*(\eta, t)P(\eta)B(\eta)Q^{-1}(\eta)B^*(\eta)P(\eta)T(\eta, t)x\, d\eta.
\end{aligned}
\tag{3.27}
$$

Therefore, $P(t)$ satisfies the Riccati integral equation

$$
\begin{aligned}
P(t)x = {} & T^*(t_f, t)GT(t_f, t)x \\
& + \int_t^{t_f} T^*(\eta, t)[D(\eta) - P(\eta)B(\eta)Q^{-1}(\eta)B^*(\eta)P(\eta)]T(\eta, t)x\, d\eta, \\
& \qquad t_0 \leqq t \leqq t_f, \quad x \in H.
\end{aligned}
\tag{3.28}
$$

Next, let us ask whether the solution of (3.28) is unique in $\mathcal{B}_\infty(t_0, t_f; H, H)$. The answer is yes, as we will prove by showing that any solution of (3.28) is also the unique solution of the Riccati integral equation of Curtain and Pritchard [2].[4]

For the moment, let us assume only that $P \in \mathcal{B}_\infty(t_0, t_f; H, H)$ satisfies (3.28). Define $S(\cdot\,, \cdot)$ to be the perturbed evolution operator corresponding to the perturbation of $T(\cdot\,, \cdot)$ by $-BQ^{-1}B^*P$, according to (2.6). Then, it can be shown by substitution and some manipulation using Fubini's theorem that P satisfies (3.25). Of course, the control for the trajectory defined by $x(t) = S(t, s)x(s)$ is given by (3.24). Although Datko did not derive a Riccati equation in [3] or [4], he did derive equations similar to (3.24) and (3.25) for the case when $G = 0$, and a generalization of his argument to show that $P(\cdot)$ is self-adjoint and that the optimal cost can be given in terms of $P(\cdot)$ will be helpful here. Note that, from here on, the development is based on

[4] If H is separable, one can prove uniqueness for the solution of (3.28) using Gronwall's lemma and the identity

$$P_1BQ^{-1}B^*P_1 - P_2BQ^{-1}B^*P_2 = (P_1 - P_2)BQ^{-1}B^*P_2 + P_1BQ^{-1}B^*(P_1 - P_2).$$

Separability of H is needed to guarantee measurability of $\|P_1(\cdot) - P_2(\cdot)\|$. As we will see, any solution of (3.28) is self-adjoint.

the fact that $P(\cdot)$, $S(\cdot,\cdot)$, and $u(\cdot)$ satisfy (2.6) with $C=-BQ^{-1}B^*P$, (3.24), and (3.25).

Let x and y be in H. Then from (3.25) we have

$$\langle P(t)x, y\rangle_H=\langle GS(t_f,t)x, T(t_f,t)y\rangle_H+\int_t^{t_f}\langle D(\eta)S(\eta,t)x, T(\eta,t)y\rangle_H\,d\eta. \tag{3.29}$$

Replacing $T(\cdot,\cdot)$ with $S(\cdot,\cdot)$ from (2.6) with $C=-BQ^{-1}B^*P$, we have

$$\begin{aligned}\langle P(t)x, y\rangle_H&=\langle GS(t_f,t)x, S(t_f,t)y\rangle_H\\&\quad+\int_t^{t_f}\langle GS(t_f,t)x, T(t_f,\eta)B(\eta)Q^{-1}(\eta)B^*(\eta)P(\eta)S(\eta,t)y\rangle_H\,d\eta\\&\quad+\int_t^{t_f}\langle D(\eta)S(\eta,t)x, S(\eta,t)y\rangle_H\,d\eta\\&\quad+\int_t^{t_f}\left\langle D(\eta)S(\eta,t)x, \int_t^{\eta}T(\eta,\xi)B(\xi)Q^{-1}(\xi)B^*(\xi)P(\xi)S(\xi,t)y\,d\xi\right\rangle_H d\eta.\end{aligned} \tag{3.30}$$

When the order of integration is reversed, the last integral in (3.30) becomes

$$\begin{aligned}&\int_t^{t_f}\left\langle\int_\eta^{t_f}T^*(\xi,\eta)D(\xi)S(\xi,\eta)S(\eta,t)x\,d\xi, B(\eta)Q^{-1}(\eta)B^*(\eta)P(\eta)S(\eta,t)y\right\rangle_H d\eta\\&=\int_t^{t_f}\langle[P(\eta)-T^*(t_f,\eta)GS(t_f,\eta)]S(\eta,t)x, B(\eta)Q^{-1}(\eta)B^*(\eta)P(\eta)S(\eta,t)y\rangle_H\,d\eta.\end{aligned} \tag{3.31}$$

Then, when we combine the three integrals in (3.30) we obtain

$$\begin{aligned}\langle P(t)x, y\rangle_H&=\langle GS(t_f,t)x, S(t_f,t)y\rangle_H\\&\quad+\int_t^{t_f}(\langle D(\eta)S(\eta,t)x, S(\eta,t)y\rangle_H\\&\qquad+\langle Q(\eta)Q^{-1}(\eta)B^*(\eta)P(\eta)S(\eta,t)x,\\&\qquad\quad Q^{-1}(\eta)B^*(\eta)P(\eta)S(\eta,t)y\rangle_U)\,d\eta=\langle P(t)y, x\rangle_H.\end{aligned} \tag{3.32}$$

With $y=x$, (3.32) shows that, if $P(\cdot)$ satisfies (3.25), where $S(\cdot,\cdot)$ is given by (2.6) with $C=-BQ^{-1}B^*P$, i.e., u is the linear feedback control of (3.24), then

$$J(t,x,u)=\langle P(t)x,x\rangle_H,\qquad t_0\leqq t\leqq t_f,\quad x\in H. \tag{3.33}$$

Clearly, $P(t)$ is nonnegative.

Also, from (3.32), we have

$$\begin{aligned}\langle P(t)x, x\rangle_H&=\langle S^*(t_f,t)GS(t_f,t)x, x\rangle_H\\&\quad+\left\langle\int_t^{t_f}S^*(\eta,t)[D(\eta)+P(\eta)B(\eta)Q^{-1}(\eta)B^*(\eta)P(\eta)]S(\eta,t)x\,d\eta, x\right\rangle_H,\\&\qquad\qquad t_0\leqq t\leqq t_f,\quad x\in H.\end{aligned} \tag{3.34}$$

Since $P(t)$ is self-adjoint, (3.34) implies

$$\begin{aligned}P(t)x &= S^*(t_f, t)GS(t_f, t)x\\ &\quad+\int_t^{t_f} S^*(\eta, t)[D(\eta)+P(\eta)B(\eta)Q^{-1}(\eta)B^*(\eta)P(\eta)]S(\eta, t)x\,d\eta,\\ &\qquad t_0\leqq t\leqq t_f,\quad x\in H,\end{aligned}\tag{3.35}$$

where

$$\begin{aligned}S(t, s)x &= T(t, s)x-\int_s^t T(t, \eta)B(\eta)Q^{-1}(\eta)B^*(\eta)P(\eta)S(\eta, t)x\,d\eta,\\ &\qquad t_0\leqq s\leqq t\leqq t_f,\quad x\in H.\end{aligned}\tag{3.36}$$

(3.35) is the Riccati integral equation derived by Curtain and Pritchard in [2], using a successive approximation technique. Also, in [7], Delfour and Mitter derived an equation similar to (3.35) for hereditary systems.

As was indicated earlier, one can start with (3.28) and (3.36) and derive (3.25) by substitution. Then, as we have shown, (3.25) and (3.36) imply (3.35). Also, if $P(\cdot)$ and $S(\cdot,\cdot)$ satisfy (3.35) and (3.36), and if $P(\cdot)$ is assumed self-adjoint, (3.25) and, therefore, (3.28) can be derived by substitution and manipulation. For this derivation, it is useful to recall that Theorem 2.1 says that (3.36) is equivalent to

$$\begin{aligned}S(t, s)x &= T(t, s)x-\int_s^t S(t, \eta)B(\eta)Q^{-1}(\eta)B^*(\eta)P(\eta)T(\eta, s)x\,d\eta,\\ &\qquad t_0\leqq s\leqq t\leqq t_f,\quad x\in H.\end{aligned}\tag{3.37}$$

It is important for § 4 to note that neither the derivation of (3.35) from (3.28) and (3.36) nor the derivation of (3.28) from (3.35) and (3.36) depends on G or $P(\cdot)$ being nonnegative. ($P(\cdot)$ is nonnegative if G is.) Also, recall that we do not assume the solution of (3.28) to be self-adjoint in order to derive (3.32) and (3.35).

Henceforth, let us refer to (3.28) as the "first Riccati integral equation" and to (3.35) as the "second Riccati integral equation"—not because (3.28) is derived first in this paper, but because (3.28) involves the evolution operator $T(\cdot,\cdot)$, which represents the original dynamics of the control system, while (3.35) involves the perturbed evolution operator $S(\cdot,\cdot)$, which represents the dynamics of the control system after it has been modified to implement the optimal control in closed-loop form. The following theorem gives the relationships we have found between solutions to the first Riccati integral equation and solutions to the second Riccati integral equation.

THEOREM 3.1. *Let $T(\cdot,\cdot)$, $B(\cdot)$, $D(\cdot)$, and $Q(\cdot)$ be as previously defined and let $G\in\mathscr{L}(H, H)$ be self-adjoint. If $P(\cdot)$ satisfies* (3.28), *then $P(\cdot)$ is self-adjoint, and $P(\cdot)$ and $S(\cdot,\cdot)$ satisfy* (3.35) *and* (3.36). *If $P(\cdot)$ and $S(\cdot,\cdot)$ satisfy* (3.35) *and* (3.36), *and if $P(\cdot)$ is self-adjoint, then $P(\cdot)$ satisfies* (3.28).

Of course, for the optimal control problem, G is nonnegative; however, without this restriction, Theorem 3.1 will be more useful for our subsequent analysis of control on the infinite interval.

Curtain and Pritchard proved in [2] that, as a system of equations, (3.35) and (3.36) have a unique solution $P(\cdot)$ and $S(\cdot,\cdot)$ when $G\geqq 0$ and $P(\cdot)$ is restricted to be self-adjoint, and we include their proof here because elements of it will be useful when we consider the case $t_f=\infty$, which was not considered in [2]. Let $P(\cdot)$ and $S(\cdot,\cdot)$ satisfy (3.35) and (3.36), with $P(\cdot)$ a self-adjoint element of $\mathscr{B}_\infty(t_0, t_f; H, H)$. Then the value of the cost $J(s, x(s),\cdot)$ for the trajectory given by $x(t)=S(t, s)x(s)$ and control given by $u(t)=-Q^{-1}(t)B^*(t)P(t)x(t)$ is $\langle P(s)x(s), x(s)\rangle_H$ (from (3.33)).

From (3.35) it follows that

$$\begin{aligned} P(s)x &= S^*(t,s)P(t)S(t,s)x \\ &\quad + \int_s^t S^*(\eta,s)[D(\eta)+P(\eta)B(\eta)Q^{-1}(\eta)B^*(\eta)P(\eta)]S(\eta,s)x\,d\eta, \\ &\qquad t_0 \leqq s \leqq t \leqq t_f, \quad x\in H. \end{aligned} \tag{3.38}$$

LEMMA 3.1 (Curtain and Pritchard). *Assume the hypotheses just stated for* $P(\cdot)$ *and* $S(\cdot,\cdot)$, *and, for some s between* t_0 *and* t_f, *let*

$$z(t)=S(t,s)x+\int_s^t S(t,\eta)B(\eta)\bar{u}(\eta)\,d\eta, \qquad s\leqq t\leqq t_f, \tag{3.39}$$

for some $x\in H$ *and* $\bar{u}\in\mathscr{U}_s$. *Then*

$$\begin{aligned} \langle P(s)z(s), z(s)\rangle_H &= \langle P(t)z(t), z(t)\rangle_H \\ &\quad + \int_s^t \langle [D(\eta)+P(\eta)B(\eta)Q^{-1}(\eta)B^*(\eta)P(\eta)]z(\eta), z(\eta)\rangle_H\,d\eta \\ &\quad -2\int_s^t \langle P(\eta)B(\eta)\bar{u}(\eta), z(\eta)\rangle_H\,d\eta. \end{aligned} \tag{3.40}$$

Proof. (3.40) can be verified by substitution with some rather tedious manipulation, which is facilitated by the identity

$$\begin{aligned} &\left\langle A\int_s^t S(t,\eta)B(\eta)v(\eta)\,d\eta, \int_s^t S(t,\eta)B(\eta)v(\eta)\,d\eta\right\rangle_H \\ &\quad =2\int_s^t \left\langle AS(t,\eta)B(\eta)v(\eta), S(t,\eta)\int_s^\eta S(\eta,\xi)B(\xi)v(\xi)\,d\xi\right\rangle_H d\eta, \end{aligned} \tag{3.41}$$

for $A\in\mathscr{L}(H,H)$.

For $v\in\mathscr{U}_s$, let

$$z(t)=T(t,s)x(s)+\int_s^t T(t,\eta)B(\eta)v(\eta)\,d\eta, \qquad s\leqq t\leqq t_f. \tag{3.39$'$}$$

With some manipulation, it can be shown that $z(t)$ satisfies (3.39) with

$$\bar{u}(\eta)=v(\eta)+Q^{-1}(\eta)B^*(\eta)P(\eta)z(\eta). \tag{3.42}$$

Then (3.40) shows

$$\begin{aligned} &\langle P(s)x(s), x(s)\rangle_H \\ &\quad =\langle P(t)z(t), z(t)\rangle_H \\ &\quad\quad +\int_s^t (\langle D(\eta)z(\eta), z(\eta)\rangle_H+\langle Q(\eta)v(\eta), v(\eta)\rangle_U)\,d\eta \\ &\quad\quad -\int_s^t \langle Q(\eta)\bar{u}(\eta), \bar{u}(\eta)\rangle_U)\,d\eta. \end{aligned} \tag{3.43}$$

Therefore, the unique minimum value for $J(s, x(s),\cdot)$ is $\langle P(s)x(s), x(s)\rangle_H$ if $P(\cdot)$ and $S(\cdot,\cdot)$ satisfy (3.35) and (3.36). Also, $P(s)$ is unique, since it is self-adjoint, and

$S(\cdot,\cdot)$ is then the unique solution of (3.36). The following theorem summarizes the most important results of this section.

THEOREM 3.2. *The unique control $u \in \mathcal{U}_s$ which minimizes the cost functional $J(s, x(s), \cdot)$ of* (3.2) *is the linear feedback control of* (3.24), *where $P(\cdot)$ is the unique element of $\mathcal{B}_\infty(s, t_f; H, H)$ which satisfies the first Riccati integral equation* (3.28). *The corresponding optimal trajectory is given by $x(t) = S(t, s)x(s)$, where $S(\cdot,\cdot)$ is the perturbed evolution operator corresponding to the perturbation of $T(\cdot,\cdot)$ by $-B(\cdot)Q^{-1}(\cdot)B^*(\cdot)P(\cdot)$. The minimum value of the cost functional is $\langle P(s)x(s), x(s)\rangle_H$. Also, when possible solutions of the second Riccati integral equation* (3.35) *are restricted to be self-adjoint, $P(\cdot)$ and $S(\cdot,\cdot)$ are the unique solution of the system of equations* (3.35) *and* (3.36).

It has not been necessary to define the adjoint state because, when the relationship between the control and the state is represented by the integral equation (3.1), the optimal control problem amounts to minimizing the quadratic functional of (3.6) with no constraint on the control u. If the adjoint state is desired, (3.25) indicates its definition. Recalling that the optimal state is given by $x(t) = S(t, s)x(s)$, we see from (3.25) that the appropriate adjoint state $p(t)$ is given by

$$(3.44) \qquad p(t) = T^*(t_f, t)Gx(t_f) + \int_t^{t_f} T^*(\eta, t)D(\eta)x(\eta)\, d\eta, \qquad t_0 \leqq t \leqq t_f.$$

Thus (3.25) yields the familiar relationship

$$(3.45) \qquad p(t) = P(t)x(t), \qquad t_0 \leqq t \leqq t_f.$$

Under our hypotheses, the adjoint state is required only to be weakly continuous and strongly measurable, but in applications, it usually will be strongly continuous.

The only references to an adjoint system that are important in this paper are the requirements that the homogeneous adjoint system, represented by $T^*(\cdot,\cdot)$, be strongly measurable and later, in § 5, that approximations to the homogeneous adjoint system converge strongly. We will not need the adjoint state for the analysis of control on the infinite interval to be taken up in the next section.

4. Optimal control on the infinite interval. In this section, we consider the optimal control problem of § 3 for $t_f = \infty$ and $G = 0$. We replace the inequality $t \leqq t_f$ by the strict inequality $t < \infty$ in the appropriate places, and require $\|T(t, s)\|$ to be uniformly bounded for t and s in any bounded interval. It should be emphasized that, as yet, we do not require $\|T(t, s)\|$ to be exponentially bounded. Also, we only require the bounds for $\|B(\cdot)\|$, $\|D(\cdot)\|$, and $\|Q(\cdot)\|$ (including the lower bound for $\|Q(\cdot)\|$) to be uniform almost everywhere on each bounded interval.

For a control u and the corresponding state $x(t)$ given by (3.1) with $t_0 = s$ and $x(t_0) = x$, we use the notation

$$(4.1) \qquad J_\infty(s, x, u) = \int_s^\infty (\langle D(\eta)x(\eta), x(\eta)\rangle_H + \langle Q(\eta)u(\eta), u(\eta)\rangle_U)\, d\eta;$$

and we study problems in which there is at least one u for which the cost functional of (4.1) is finite.

DEFINITION 4.1. A function u is an *admissible control for the initial time s and the initial state x*, or simply an *admissible control for s and x*, if $u(\cdot)$ is strongly measurable on (s, ∞) and $J_\infty(s, x, u)$ is finite.

It will be helpful to understand the nature of the set of admissible controls. Suppose that u_1 and u_2 are admissible controls for s and x, and let $x_1(t)$ and $x_2(t)$ be the solutions of (3.1) for u equal to u_1 and u_2, respectively. Since $L_2(s, \infty; U)$ and $L_2(s, \infty; H)$ are linear spaces, $Q^{1/2}(u_1-u_2)\in L_2(s, \infty; U)$ and $D^{1/2}(t)(x_1(t)-x_2(t))=D^{1/2}(t)\int_s^t T(t, \eta)\cdot B(\eta)(u_1(\eta)-u_2(\eta))\,d\eta \in L_2(s, \infty; H)$. Therefore, if u_{sx} is an admissible control for s and x, the set of all admissible controls for s and x is $\mathcal{U}_{ad}=u_{sx}+\mathcal{U}_{s\infty}$, where

$$(4.2)\qquad \mathcal{U}_{s\infty}=\{v\colon v(\cdot)\text{ is strongly measurable on }(s, \infty)\text{ and }J_\infty(s, 0, v)<\infty\}.$$

The optimization problem then is to find a $v\in\mathcal{U}_{s\infty}$ which minimizes $J_\infty(s, x, u_{sx}+v)$. The fact that u_{sx} is an arbitrary admissible control for s and x presents no problem for the argument here because $J_\infty(s, x, \cdot)$ is uniquely defined for each admissible control.

Since $\mathcal{U}_{s\infty}$ is a linear space, it is convex. Also, for fixed u_{sx}, our hypotheses on Q imply that $J_\infty(s, x, u_{sx}+v)$ is a strictly convex function of v. Convexity of $\mathcal{U}_{s\infty}$ and strict convexity of $J_\infty(s, x, u_{sx}+v)$ are all that is needed to guarantee that there is at most one $v\in\mathcal{U}_{s\infty}$ which minimizes $J_\infty(s, x, u_{sx}+v)$. Therefore, if an optimal control exists, it is unique. Next, we consider conditions under which an optimal control exists.

Before proceeding with the problem at hand, we should note that we actually have a much more general existence result than is needed here. It is not difficult to prove that $\mathcal{U}_{s\infty}$ is a Hilbert space when endowed with the norm $\|\cdot\|_{s\infty}^2=J_\infty(s, 0, \cdot)$. This is equivalent to saying that the mapping that takes a control $u(\cdot)$ to $D^{1/2}(\cdot)x(\cdot)$, with $x(s)=0$, is a closed linear operator from $\mathcal{U}_{sQ}=\{u\colon u(\cdot)$ is strongly measurable and $\|u\|_{sQ}^2=\int_s^\infty \langle Q(\eta), u(\eta)\rangle_U\,d\eta<\infty\}$ to $L_2(s, \infty; H)$. $\mathcal{U}_{sQ}$ is a Hilbert space because of our hypotheses on Q. Now, since $\mathcal{U}_{sQ}$ contains $\mathcal{U}_{s\infty}$ algebraically and topologically, if C is a closed, convex set in $\mathcal{U}_{sQ}$, $C\cap\mathcal{U}_{s\infty}$ is closed and convex in $\mathcal{U}_{s\infty}$. Hence, if $f(\cdot)$ is a convex functional defined on $C\cap\mathcal{U}_{s\infty}$ and $f(\cdot)$ is continuous in the $\mathcal{U}_{s\infty}$ norm, and if either $C\cap\mathcal{U}_{s\infty}$ is bounded in $\mathcal{U}_{s\infty}$ or $f(u_n)\to\infty$ as $\|u_n\|_{\mathcal{U}s\infty}\to\infty$, then $f(\cdot)$ achieves a minimum in C. (See Lions [10, pp. 6–8].) We will not appeal to this general existence result in the remainder of this paper because our subsequent proof that an optimal control exists whenever an admissible control exists for the problem here provides useful information about approximating the optimal control problem on the infinite interval with a sequence of optimal control problems on finite intervals.

We consider a sequence $\{t_n\}$, where $t_n\to\infty$ as $n\to\infty$, and investigate the sequence of problems for which $t_f=t_n$ and $G=0$. For each of these problems, we denote the solution of the Riccati integral equations (3.28) and (3.35) by $P_n(\cdot)$ and the cost functional by $J_n(s, x, \cdot)$. An important observation, which follows from the fact that $\min_{u\in\mathcal{U}s} J_n(s, x, u)=\langle P_n(s)x, x\rangle_H$, is that

$$(4.3)\qquad P_n(t)\leqq P_m(t),\qquad t_0\leqq t\leqq t_n\leqq t_m.$$

Suppose that, for some $s\geqq t_0$ and some $x\in H$,

$$(4.4)\qquad \begin{aligned}\lim_{n\to\infty}\langle P_n(s)x, x\rangle_H &= \lim_{t_n\to\infty}\int_s^{t_n}(\langle D(\eta)x_n(\eta), x_n(\eta)\rangle_H+\langle Q(\eta)u_n(\eta), u_n(\eta)\rangle_U)\,d\eta\\ &=c<\infty;\end{aligned}$$

where $u_n(\cdot)$ is the optimal control for the problem of § 3 with $t_f=t_n$ and $G=0$ and $x_n(\cdot)$ is the corresponding optimal trajectory. If we extend $x_n(\cdot)$ and $u_n(\cdot)$ to $[s, \infty)$ by setting $x_n(t)=u_n(t)=0$ for $t\geqq t_n$, (4.4) shows that the sequences $\{D^{1/2}x_n\}$ and $\{Q^{1/2}u_n\}$ are bounded sequences in $L_2(s, \infty; H)$ and $L_2(s, \infty; U)$, respectively. Therefore, there exist a subsequence $\{t_k\}$, a ψ in $L_2(s, \infty; H)$, and a $\phi\in L_2(s, \infty; U)$ such that

$$(4.5)\qquad D^{1/2}x_k\to\psi\quad\text{weakly in }L_2(s, \infty; H)$$

and

$$Q^{1/2}u_k \to \phi \quad \text{weakly in } L_2(s, \infty; U). \tag{4.6}$$

Define $u(\cdot)$ and $x(\cdot)$ by

$$u(t) = Q^{-1/2}(t)\phi(t), \qquad s \leqq t < \infty \text{ (a.e.)}, \tag{4.7}$$

and

$$x(t) = T(t, s)x + \int_s^t T(t, \eta)B(\eta)u(\eta)\, d\eta, \qquad s \leqq t < \infty. \tag{4.8}$$

Let $\bar{t} \in [s, \infty)$ and $f \in L_2(s. \bar{t}; H) \subset L_2(s, \infty; H)$ (by extension by zero outside $[s, \bar{t})$), and note that (4.6) and our hypotheses on Q imply that $u_k \to u$ weakly in $L_2(s, \bar{t}; U)$. Then, for almost all $t \in (s, \bar{t})$,

$$\begin{aligned} |\langle f(t), x_k(t) - x(t)\rangle_H \leqq & \left| \int_s^t \langle B^*(\eta)T^*(t, \eta)f(t), u_k(\eta) - u(\eta)\rangle_U d\eta \right| \\ & \to 0 \quad \text{as } k \to \infty. \end{aligned} \tag{4.9}$$

Also,

$$|\langle f(t), x_k(t) - x(t)\rangle_H| \leqq \|f(t)\|(\|x_k(t)\| + \|x(t)\|). \tag{4.10}$$

Since $\|x_k(t)\|$ and $\|x(t)\|$ are uniformly bounded in k and t for $t \in (s, \bar{t})$ and $\|f(t)\|$ is Lebesgue integrable, the dominated convergence theorem implies

$$\left| \int_s^{\bar{t}} \langle f(\eta), x_k(\eta) - x(\eta)\rangle_H d\eta \right| \to 0 \quad \text{as } k \to \infty. \tag{4.11}$$

Therefore,

$$\int_s^{\bar{t}} \langle f(\eta), D^{1/2}(\eta)(x_k(\eta) - x(\eta))\rangle_H d\eta \to 0 \quad \text{as } k \to \infty \tag{4.12}$$

for all $f \in L_2(s, \bar{t}; H)$ and $\bar{t} \in [s, \infty)$. Thus $D^{1/2}(\cdot)x(\cdot) = \psi(\cdot) \in L_2(s, \infty; H)$.

Actually, we have shown that, since the original sequence $\{D^{1/2}(\cdot)x_n(\cdot)\}$ is bounded in $L_2(s, \infty; H)$, if any subsequence $\{Q^{1/2}(\cdot)u_j(\cdot)\}$ of $\{Q^{1/2}(\cdot)u_n(\cdot)\}$ converges weakly in $L_2(s, \infty; U)$ to $Q^{1/2}(\cdot)u(\cdot)$, then the corresponding subsequence $\{D^{1/2}(\cdot)x_j(\cdot)\}$ converges weakly in $L_2(s, \infty; H)$ to $D^{1/2}(\cdot)x(\cdot)$, where $x(\cdot)$ is given by (4.8), and $\{x_k(\cdot)\}$ converges weakly in $L_2(s, \bar{t}; H)$to $x(\cdot)$, for $\bar{t} \in [s, \infty)$. Also, (4.8) shows that, if $\{u_j(\cdot)\}$ converges strongly in $L_2(s, \bar{t}; U)$, $\{x_i(\cdot)\}$ converges strongly in $L_2(s, \bar{t}; H)$.

We see then that u is an admissible control for the initial time s and initial state x. Is u optimal?

Consider the pair $(D^{1/2}(\cdot)x(\cdot), Q^{1/2}(\cdot)u(\cdot))$ as an element of the Hilbert space $\hat{H} = L_2(s, \infty; H) \times L_2(s, \infty; U)$, whose inner product is given by

$$\langle (x_1, u_1), (x_2, u_2)\rangle_{\hat{H}} = \langle x_1, x_2\rangle_{L_2(s,\infty;H)} + \langle u_1, u_2\rangle_{L_2(s,\infty;U)}. \tag{4.13}$$

Then the sequence $\{(D^{1/2}(\cdot)x_k(\cdot), Q^{1/2}(\cdot)u_k(\cdot))\}$ converges weakly in $\hat{H}$ to $(D^{1/2}(\cdot)x(\cdot), Q^{1/2}(\cdot)u(\cdot))$. It is easy to show that, if a sequence in a Hilbert space has a weak limit, the norm of that limit is not greater than the limit supremum of the sequence of norms. From (4.4), we know that $\lim_{k\to\infty}\|(D^{1/2}x_k, Q^{1/2}u_k)\|_{\hat{H}}^2 = c$. Therefore

$$J_\infty(s, x, u) = \|(D^{1/2}(\cdot)x(\cdot), Q^{1/2}(\cdot)u(\cdot))\|_{\hat{H}}^2 \leqq c. \tag{4.14}$$

If, for some admissible control u, we have strict inequality in (4.14), then there must be some k for which $J_k(s, x, u) \leqq J_\infty(s, x, u) < J_k(s, x, u_k)$. But this is impossible, since u_k is optimal for $t_f = t_k$. Therefore,

$$J_\infty(s, x, u) = c, \tag{4.15}$$

where u is given by (4.7), and we know that no admissible control can give a smaller value than c for the cost functional $J_\infty(s, x, \cdot\,)$. Since we already know that at most one admissible control can minimize the cost functional, the u of (4.7) is the unique optimal control for the initial time s and the initial state x.

Suppose now that there is at least one v which is an admissible control for s and x. Let $\{t_n\}$ be any sequence such that $t_n \to \infty$. Then, for all n,

$$J_n(s, x, u_n) = \langle P_n(s)x, x\rangle_H \leqq J_n(s, x, v) \leqq J_\infty(s, x, v), \tag{4.16}$$

where u_n is the optimal control for $t_f = t_n$ and $G = 0$. So $\langle P_n(s)x, x\rangle_H$ is a bounded sequence of real numbers and must contain a convergent subsequence. Therefore, by our previous argument, there exists a unique optimal control u and (4.16) holds with $v = u$.

Now, if the sequence $\{\langle P_n(s)x, x\rangle_H\}$ does not converge to $J_\infty(s, x, u)$, then there is a subsequence $\{\langle P_k(s)x, x\rangle_H\}$ which converges to some $c < J_\infty(s, x, u)$, and the corresponding sequence of controls $\{u_k\}$ converges in the sense of (4.6) to some $\tilde{u}$, and $J_\infty(s, x, \tilde{u}) = c$—contradicting the optimality of u. Also, in order not to contradict the uniqueness of u, $\{u_n\}$ must converge in the sense of (4.6) to u. Thus, for any sequence $\{t_n\}$ such that $t_n \to \infty$, we have

$$Q^{1/2}(\cdot\,)u_n(\cdot\,) \to Q^{1/2}(\cdot\,)u(\cdot\,) \quad \text{weakly in } L_2(s, \infty, u) \tag{4.17}$$

and

$$D^{1/2}(\cdot\,)x_n(\cdot\,) \to D^{1/2}(\cdot\,)x(\cdot\,) \quad \text{weakly in } L_2(s, \infty, H); \tag{4.18}$$

where $u(\cdot\,)$ is the optimal control for the initial time s and the initial state x, and $x(\cdot\,)$ is the corresponding optimal trajectory.

Also, note that

$$\begin{aligned} \|(D^{1/2}(\cdot\,)x_n(\cdot\,), Q^{1/2}(\cdot\,)u_n(\cdot\,))\|_{\hat{H}}^2 &= \langle P_n(s)x, x\rangle_H \\ \to J_\infty(s, x, u) &= \|(D^{1/2}(\cdot\,)x(\cdot\,), Q^{1/2}(\cdot\,)u(\cdot\,))\|_{\hat{H}}^2 \end{aligned} \tag{4.19}$$

implies strong convergence in $\hat{H}$ and, therefore, strong convergence in (4.17) and (4.18).

The most important results obtained so far in this section are summarized in the following theorem.

THEOREM 4.1. *If there is a sequence $\{t_n\}$ such that $t_n \to \infty$ and* (4.4) *is satisfied, then there is an admissible control for the initial time s and the initial state x. If there is an admissible control for the initial time s and the initial state x, then there is a unique optimal control $u(\cdot\,)$ and the corresponding optimal trajectory $x(\cdot\,)$ is given by* (4.8). *In this case, for any sequence $\{t_n\}$ such that $t_n \to \infty$, and for $s \leqq t < \infty$, we have* (*recall our hypotheses on Q*)

$$u_n(\cdot\,) \to u(\cdot\,) \quad \textit{strongly in } L_2(s, t; U), \tag{4.20}$$

$$x_n(\cdot\,) \to x(\cdot\,) \quad \textit{strongly in } L_2(s, t; H), \tag{4.21}$$

$$J_n(s, x, u_n) = \langle P_n(s)x, x\rangle \to J_\infty(s, x, u). \tag{4.22}$$

If $Q(\xi)$ is uniformly bounded away from zero for almost all $\xi \in (s, \infty)$, (4.20) *holds for $t=\infty$, and if $D(\xi)$ is uniformly bounded away from zero for almost all $\xi \in (s, \infty)$,* (4.21) *holds for $t=\infty$.*

THEOREM 4.2. *Suppose that, for some $s \geqq t_0$, there is an admissible control for s and x, for each $x \in H$. Then there is a unique nonnegative, self-adjoint operator $P_\infty(s) \in \mathscr{L}(H, H)$ such that*

$$\min_{v \in \mathscr{U}_{ad}} J_\infty(s, x, v) = \langle P_\infty(s)x, x\rangle_H, \qquad x \in H, \tag{4.23}$$

and, for any sequence $\{t_n\}$ such that $t_n \to \infty$,

$$P_n(s)x \to P_\infty(s)x \quad \textit{strongly in } H, \quad x \in H. \tag{4.24}$$

Proof. The proof is essentially the same as one given by Datko for a similar result in [3]. Let x and y be in H. By the generalized Schwarz inequality, we have $\langle P_n(s)x, y\rangle_H^2 \leqq \langle P_n(s)x, x\rangle_H \langle P_n(s)y, y\rangle_H$. Thus, by (4.22), $\sup_n |\langle P_n(s)x, y\rangle_H| < \infty$ for each pair x and y in H. Then the principle of uniform boundedness implies that $\sup_n \|P_n(s)\| < \infty$. Let $\{t_k\}$ be an increasing subsequence. Then the uniform boundedness of $\|P_k(s)\|$ and (4.3) imply the existence of a unique self-adjoint operator $P_\infty(s) \in \mathscr{L}(H, H)$ such that (4.24) holds with $n=k$. To see that (4.24) holds for the original sequence, use $\|P_n(s)x - P_\infty(s)x\| \leqq \|P_n(s)x - P_k(s)x\| + \|P_k(s)x - P_\infty(s)x\|$. By the generalized Schwarz inequality again, we have $\|(P_n(s) - P_k(s))x\|^4 \leqq \langle (P_n(s) - P_k(s))^3 x, x\rangle_H \langle (P_n(s) - P_k(s))x, x\rangle_H$, and, since $\{\|P_n(s)\|\}$ is uniformly bounded in n, (4.22) implies (4.24). (4.22) and (4.24) imply (4.23). Clearly, $P_\infty(s)$ is nonnegative.

THEOREM 4.3. *Suppose that, for some $t \geqq t_0$, there is an admissible control for t and x, for each $x \in H$. Then, for $t_0 \leqq s \leqq t$ and $x \in H$, there is an admissible control for s and x, and*

$$\langle P_\infty(s)x, x\rangle_H \leqq \langle P_\infty(t)T(t, s)x, T(t, s)x\rangle_H + \int_s^t \langle D(\eta)T(\eta, s)x, T(\eta, s)x\rangle_H \, d\eta. \tag{4.25}$$

Also, $P_\infty(\cdot)$ satisfies the Riccati integral equation

$$\begin{aligned} P_\infty(s)x &= T^*(t, s)P_\infty(t)T(t, s)x \\ &\quad + \int_s^t T^*(\eta, s)[D(\eta) - P_\infty(\eta)B(\eta)Q^{-1}(\eta)B^*(\eta)P_\infty(\eta)]T(\eta, s)x \, d\eta, \end{aligned} \tag{4.26}$$

$$t_0 \leqq s \leqq t, \quad x \in H.$$

Proof. (4.25) is obvious. To prove (4.26), we begin by noting that, since $\|T(t, s)\|$ and $\|D(t)\|$ are uniformly bounded for t and s in any finite interval, and since $P_\infty(s)$ is self-adjoint, (4.25) shows that $\|P_\infty(s)\|$ is uniformly bounded for $t_0 \leqq s \leqq t$. Let $t_n \to \infty$ and note that (4.26) holds with ∞ replaced by n, for each $n < \infty$. Now, for any s, $P_n(s) \leqq P_\infty(s)$, so $\|P_n(s)\|$ is uniformly bounded for $n < \infty$ and $t_0 \leqq s \leqq t$. Also, for $x \in H$ and almost all $\eta \in (s, t)$,

$$\begin{aligned} &P_n(\eta)B(\eta)Q^{-1}(\eta)B^*(\eta)P_n(\eta)T(\eta, s)x - P_\infty(\eta)B(\eta)Q^{-1}(\eta)B^*(\eta)P_\infty(\eta)T(\eta, s)x \\ &\quad = P_n(\eta)B(\eta)Q^{-1}(\eta)B^*(\eta)[P_n(\eta) - P_\infty(\eta)]T(\eta, s)x \\ &\quad + [P_n(\eta) - P_\infty(\eta)]B(\eta)Q^{-1}(\eta)B^*(\eta)P_\infty(\eta)T(\eta, s)x \to 0 \quad \text{as } n \to \infty (\text{by } (4.24)). \end{aligned} \tag{4.27}$$

Therefore, the dominated convergence theorem and (4.24) imply (4.26).

COROLLARY 4.1. *From Theorem* 3.1 *we know that the solution of* (4.26) *satisfies*

$$P_\infty(s)x = S_\infty^*(t, s)P_\infty(t)S_\infty(t, s)x$$

$$(4.28) \qquad + \int_s^t S_\infty^*(\eta, s)[D(\eta)+P_\infty(\eta)B(\eta)Q^{-1}(\eta)B^*(\eta)P_\infty(\eta)]S_\infty(\eta, s)x\, d\eta,$$

$$t_0 \leqq s \leqq t, \quad x \in H;$$

where

$$(4.29) \qquad S_\infty(r, s)x = T(r, s)x - \int_s^r T(r, \eta)B(\eta)Q^{-1}(\eta)B^*(\eta)P_\infty(\eta)S_\infty(\eta, s)x\, d\eta,$$

$$t_0 \leqq s \leqq r \leqq t, \quad x \in H.$$

That is to say,? $S_\infty(\cdot, \cdot)$ is the perturbed evolution operator corresponding to the perturbation of $T(\cdot, \cdot)$ by $-B(\cdot)Q^{-1}(\cdot)B^*(\cdot)P_\infty(\cdot)$.

THEOREM 4.4. *Suppose that, for each $s \geqq t_0$ and each $x \in H$, there is an admissible control for s and x. Then, for $s \geqq t_0$ and $x \in H$, the optimal control for the initial time s and the initial state x is the linear feedback control defined by*

$$(4.30) \qquad u(t) = -Q^{-1}(t)B^*(t)P_\infty(t)x(t);$$

where the optimal trajectory $x(\cdot)$ is given by

$$(4.31) \qquad x(t) = S_\infty(t, s)x, \qquad s \leqq t < \infty,$$

and the evolution operator $S_\infty(\cdot, \cdot)$ is defined by (4.29).

Proof. (4.26), (4.28), and (4.29) hold for $t_0 \leqq s \leqq t < \infty$. We have to show that the control and trajectory defined by (4.30) and (4.31) are optimal for s and x.

From (4.28), (4.30), and (4.31), we have

$$\langle P_\infty(s)x, x\rangle_H = \langle P_\infty(t)x(t), x(t)\rangle_H + \int_s^t (\langle D(\eta)x(\eta), x(\eta)\rangle_H + \langle Q(\eta)u(\eta), u(\eta)\rangle_U)\, d\eta.$$

As $t \to \infty$, the integral increases and is bounded by $\langle P_\infty(s)x, x\rangle_H < \infty$. Thus the integral converges to a finite value $J_\infty(s, x, u)$. Then (4.23) implies that $\langle P_\infty(s)x, x\rangle_H = J_\infty(s, x, u)$ and $\lim_{t\to\infty}\langle P_\infty(t)x(t), x(t)\rangle_H = 0$.

DEFINITION 4.2. $P(\cdot)$ is called a *solution of the first Riccati integral equation on the infinite interval* if $P(\cdot) \in B_\infty(t_0, t; H, H)$ for each $t \in (t_0, \infty)$, and $P(\cdot)$ satisfies (4.26) for all s and t such that $t_0 \leqq s \leqq t < \infty$; $P(\cdot)$ is called a *solution of the second Riccati integral equation on the infinite interval* if $P(\cdot) \in B_\infty(t_0, t; H, H)$ for each $t \in (t_0, \infty)$, and $P(\cdot)$ and $S(\cdot, \cdot)$ satisfy (4.28) and (4.29) for all s and t such that $t_0 \leqq s \leqq t < \infty$. (The notation $P_\infty(\cdot)$ is reserved for the $P_\infty(\cdot)$ of Theorem 4.2.)

By replacing G by $P(t)$ in the arguments of § 3, we see that any self-adjoint solution of the first Riccati integral equation on infinite interval is also a solution of the second Riccati integral equation on the infinite interval, and vice versa. Thus, we make the following definition.

DEFINITION 4.3. $P(\cdot)$ is called a *solution of the Riccati integral equations on the infinite interval* if $P(\cdot)$ is a self-adjoint solution of both the first and second Riccati integral equations on the infinite interval.

Under the hypotheses of Theorem 4.4, $P_\infty(\cdot)$ is a solution of the Riccati integral equations on the infinite interval. Next, we should inquire about uniqueness of solutions of the Riccati integral equations on the infinite interval.

Without imposing restrictions on our optimal control problem, it would be difficult, if not impossible, to obtain uniqueness results for solutions of the Riccati equations on the infinite interval. To see which restrictions correspond to problems in which uniqueness is important, consider how the Riccati integral equations might be solved for $P_\infty(\cdot)$. For some $t \geqq t_0$, $P_\infty(t)$ might be computed according to (4.24), and then (4.26) could be solved backward in time for $P_\infty(s)$. Approximate numerical computations could be based upon the results of the next section.[5] Such an approach is the only obvious possibility unless we impose further conditions on our problem, and the question of uniqueness should not be important for this approach, since each $P_n(\cdot)$ in (4.24) is the unique solution of the Riccati equations for an optimal control problem on a finite time interval. Problems in which other methods would be used for solution of the Riccati equations on the infinite interval would most likely be problems in which the system dynamics and the linear operators in the cost functional are periodic, and perhaps constant. Thus we shall consider subsequently the uniqueness question for periodic systems, but first, some useful results that do not depend on periodicity.

THEOREM 4.5. *If $P(\cdot)$ is a solution of the Riccati integral equations on the infinite interval and $P(s) \geqq 0$ for $t_0 \leqq s < \infty$, then, for each $s \geqq t_0$ and each $x \in H$, there is an admissible control for s and x, and*

$$P(s) \geqq P_\infty(s), \qquad t_0 \leqq s < \infty. \tag{4.32}$$

Proof. By hypothesis, $P(\cdot)$ and $S(\cdot,\cdot)$ satisfy (3.35) and (3.36) (or (4.28) and (4.29)) for $t_0 \leqq s \leqq t < \infty$. For $x \in H$, set $x(t) = S(t, s)x$ and $u(t) = -Q^{-1}(t)B^*(t)P(t)x(t)$. From (3.32), we have

$$\langle P(s)x, x\rangle_H = \langle P(t)x(t), x(t)\rangle_H + \int_s^t (\langle D(\eta)x(\eta), x(\eta)\rangle_H + \langle Q(\eta)u(\eta), u(\eta)\rangle_U)\, d\eta. \tag{4.33}$$

The integral in (4.33) is increasing and bounded by $\langle P(s)x, x\rangle_H$ as $t \to \infty$, so $J_\infty(s, x, u) < \infty$ and

$$\langle P(s)x, x\rangle_H \geqq J_\infty(s, x, u) \geqq \langle P_\infty(s)x, x\rangle_H. \tag{4.34}$$

THEOREM 4.6. *Suppose $T(\cdot,\cdot)$, $B(\cdot)$, $D(\cdot)$, and $Q(\cdot)$ are such that, for all $s \geqq t_0$ and $x \in H$, if u is an admissible control for s and x,*

$$\lim_{t\to\infty} \|x(t)\| = 0, \tag{4.35}$$

where $x(\cdot)$ is given by (4.8); *i.e., any admissible control drives the state to zero asymptotically. Then there is at most one uniformly bounded, nonnegative solution of the Riccati integral equations on the infinite interval.*

Proof. Let $P(\cdot)$ be such a solution and define $x(t)$ and $u(t)$ as in the proof of Theorem 4.5. The uniform boundedness of $\|P(\cdot)\|$ and (4.35) imply $\lim_{t\to\infty}\langle P(t)x(t), x(t)\rangle_H = 0$. Then (4.33) shows

$$\langle P(s)x, x\rangle_H = J_\infty(s, x, u) < \infty. \tag{4.36}$$

Let v be an admissible control for s and x, with the corresponding state $z(\cdot)$ given by (3.41). Then (3.43) holds for $x(s) = x$ and $s \leqq t \leqq \infty$, and (4.35) (with $x(t) = z(t)$), (4.36), and (3.43) show

$$J_\infty(s, x, u) = \langle P(s)x, x\rangle_H \leqq J_\infty(s, x, v). \tag{4.37}$$

[5] As discussed in § 5, finite dimensional approximations to the $P_n(\cdot)$'s would have to converge before the $P_n(\cdot)$'s converge to $P_\infty(\cdot)$, and practical algorithms might not be available for this order of convergence.

Thus u is the optimal control for s and x, and, since $P(s)$ is self-adjoint, (4.23) and (4.37) imply $P(s)=P_\infty(s)$, $t_0 \leqq s < \infty$.

The following lemma gives sufficient conditions for the hypothesis of Theorem 4.6 to hold.

LEMMA 4.1. *Suppose there exist positive constants M_2, α, and m such that*

$$\|T(t,s)\| \leqq M_2\, e^{\alpha(t-s)}, \qquad t_0 \leqq s \leqq t < \infty, \tag{4.38}$$

$$\|B(t)\| \leqq M_2, \quad \textit{a.e.} \quad \textit{in } [t_0,\infty), \tag{4.39}$$

$$D(t) \geqq m > 0 \quad \textit{a.e.} \quad \textit{in } [t_0,\infty), \tag{4.40}$$

$$Q(t) \geqq m > 0 \quad \textit{a.e.} \quad \textit{in } [t_0,\infty). \tag{4.41}$$

Then, if u is an admissible control for s and x and $x(t)$ is the corresponding trajectory, (4.35) *holds.*

Proof. (4.40) and (4.41) imply that, since u is an admissible control, $\int_s^\infty (\|x(\eta)\|^2 + \|u(\eta)\|^2)\, d\eta < \infty$. For $n \geqq 1$, choose t_n such that $\int_{t_n}^\infty (\|x(\eta)\|^2 + \|u(\eta)\|^2) d\eta < 1/n^4$. Define δ_n to be the measure of $\{t : t \geqq t_n \text{ and } \|x(t)\|^2 + \|u(t)\|^2 \geqq 1/n^2\}$. Clearly $\delta_n < 1/n^2$.

Let $t \geqq t_n + 1/n^2$. Then there must be an s between t_n and t such that $|t-s| \leqq \delta_n$ and $\|x(s)\| < 1/n$. From (3.1) we have

$$\begin{aligned} \|x(t)\| &\leqq M_2\, e^{(t-s)}(1/n) + \int_s^t M_2\, e^{\alpha(t-n)} \|B(\eta)\| \cdot \|u(\eta)\|\, d\eta \\ &\leqq M_2\, e^{\alpha\delta_n}(1/n) + M_2\, e^{\alpha\delta_n} M_2 (t-s)^{1/2} \left(\int_s^t \|u(\eta)\|^2\, d\eta \right)^{1/2} \\ &\leqq M_2\, e^{\alpha}(1/n + M_2/n^3). \end{aligned} \tag{4.42}$$

Therefore, $\lim_{t\to\infty} \|x(t)\| = 0$.

THEOREM 4.7. *If* (4.38), (4.39), *and* (4.41) *hold for some positive constants M_2 and α, and if there is a uniformly bounded, nonnegative solution of the Riccati integral equations on the infinite interval, then there are positive constants M_4 and β for which*

$$\|S_\infty(t,s)\| \leqq M_4\, e^{\beta(t-s)}, \qquad t_0 \leqq s \leqq t < \infty, \tag{4.43}$$

where $S_\infty(\cdot\,,\cdot)$ is given by (4.29).

Proof. According to Theorem (4.5), $\|P_\infty(\cdot)\|$ is uniformly bounded. Let u be the optimal control for s and x. We have

$$\begin{aligned} \|S_\infty(t,s)x\| &\leqq \|T(t,s)x\| + \int_s^t \|T(t,\eta)B(\eta)u(\eta)\|\, d\eta \\ &\leqq M_2\, e^{\alpha(t-s)}\|x\| + M_2^2\, e^{\alpha(t-s)}(t-s)^{1/2} \left(\int_s^\infty \|u(\eta)\|^2\, d\eta \right)^{1/2}. \end{aligned} \tag{4.44}$$

Since

$$\int_s^\infty \|u(\eta)\|^2\, d\eta \leqq \langle P_\infty(s)x, x \rangle \tag{4.45}$$

and $\|P_\infty(\cdot)\|$ is uniformly bounded, the existence of the M_4 and β of (4.43) follows.

The next theorem follows from (4.23), (4.31), (4.43), and Theorem 2.2.

THEOREM 4.8. *If* (4.38)–(4.41) *hold and if $P_\infty(\cdot)$ exists and is uniformly bounded on $[t_0,\infty)$, there exist positive constants M_4 and β such that*

$$\|S_\infty(t,s)\| \leqq M_4\, e^{-\beta(t-s)}, \qquad t_0 \leqq s \leqq t < \infty. \tag{4.46}$$

Now let $T(\cdot,\cdot)$, $B(\cdot)$, $D(\cdot)$, and $Q(\cdot)$ be periodic with a common period $\omega>0$; i.e., $T(t,s)=T(t+\omega,s+\omega)$ and $B(t)=B(t+\omega)$, $D(t)=D(t+\omega)$, and $Q(t)=Q(t+\omega)$ for almost all t. By replacing s and t_n by $s+\omega$ and $t_n+\omega$, respectively, in Theorem 4.2, we see that $P_\infty(s)=P_\infty(s+\omega)$, and then (4.29) shows that $S_\infty(t,s)=S_\infty(t+\omega,s+\omega)$. Thus, we can define $T(\cdot,\cdot)$, $B(\cdot)$, $D(\cdot)$, $P_\infty(\cdot)$, and $S_\infty(\cdot,\cdot)$ on the entire real line by periodic extension. Since $T(\cdot,\cdot)$ is periodic, it is exponentially bounded, and the previous results of this section yield the following theorem regarding periodic solutions of the Riccati integral equations.

THEOREM 4.9. *Let $T(\cdot,\cdot)$, $B(\cdot)$, $D(\cdot)$, and $Q(\cdot)$ have period $\omega>0$. Then there is a uniformly bounded,*[6] *nonnegative*[7] *periodic solution of the Riccati integral equations on the infinite interval if and only if, for each $s\in(-\infty,\infty)$ and each $x\in H$, there is an admissible control for s and x. Suppose that such a solution exists, and let $P_\infty(\cdot)$ and $S_\infty(\cdot,\cdot)$ be the operators of Theorems* 4.1–4.4. *Then, if $P(\cdot)$ is a nonnegative periodic solution of the Riccati integral equations on the infinite interval,* (4.32) *holds for $-\infty<s<\infty$. If the hypothesis of Theorem* 4.6 *holds, $P_\infty(\cdot)$ is the unique nonnegative, self-adjoint periodic solution of the Riccati integral equations on the infinite interval. If $D(\cdot)$ is essentially bounded away from zero, the hypothesis of Theorem* 4.6 *holds, and there are positive constants M_4 and β for which* (4.46) *holds for $-\infty<s\leqq t<\infty$.*

For the periodic problem, we can replace (4.24) by

$$\lim_{n\to\infty} P(s-n\omega)x=P_\infty(s)x, \qquad x\in H, \tag{4.47}$$

where $P(\cdot)$ is the solution of the Riccati integral equations for the problem of § 3 with $t_f=s$ and $G=0$. Then (4.3) becomes

$$P(s-n\omega)\leqq P(s-m\omega), \qquad 0\leqq n\leqq m. \tag{4.48}$$

(4.47) is much more practical for computational purposes than (4.24) because we simply have an initial value problem to be solved backward in time until the solution converges to $P_\infty(\cdot)$. The following stability result makes (4.47) even more useful.

THEOREM 4.10. *Suppose that $T(\cdot,\cdot)$, $B(\cdot)$, $D(\cdot)$, and $Q(\cdot)$ have period $\omega>0$, that there exists a nonnegative periodic solution of the Riccati integral equations on the infinite interval, and that $\lim_{t\to\infty}\|S_\infty(t,s)x\|=0$ for all s and x.*

Then, if $\hat{P}(\cdot)$ is the solution of the Riccati integral equations for the problem of § 3 *with $t_f=s$ and $G\geqq 0$, we have*

$$\lim_{n\to\infty}\hat{P}(s-n\omega)x=P_\infty(s)x, \qquad x\in H.\text{[8]} \tag{4.49}$$

Furthermore, if there exist positive constants M_4 and β such that (4.46) *holds for $-\infty<s\leqq t<\infty$, and if $G\geqq P_\infty(s)$, then*

$$P_\infty(s)\leqq \hat{P}(s-n\omega)\leqq P_\infty(s)+M_4^2\,e^{-2\beta n\omega}\|G\|, \qquad n\geqq 0. \tag{4.50}$$

Proof. Recalling the relationship between the solution of the Riccati integral equations and the minimum value of the cost functional ((3.33) and (4.23)), we see that,

[6] Any nonnegative periodic solution is uniformly bounded. Since we are assuming strong continuity for $T(\cdot,\cdot)$, which results in weak continuity for $P(\cdot)$, the principle of uniform boundedness guarantees a uniform bound for $\|P(\cdot)\|$. However, if $T(\cdot,\cdot)$ is only weakly continuous, the uniform boundedness of the solution of (3.25) on any finite interval guarantees a uniform bound for $\|P(\cdot)\|$.

[7] (4.28) shows that, if a periodic solution is nonnegative for some t, it is nonnegative for all t.

[8] Since $\hat{P}(\cdot)$ and $P_\infty(\cdot)$ are solutions of the first Riccati integral equation (3.28), for H finite dimensional and $0\leqq r\leqq\omega$, (4.49), (4.51), and Gronwall's lemma show that $\|\hat{P}(s-r-n\omega)-P_\infty(s-r)\|\to 0$ uniformly in r.

with $P(s-n\omega)$ defined as for (4.47).

$$\langle P(s-n\omega)x, x\rangle_H \leqq \langle \hat{P}(s-n\omega)x, x\rangle_H$$

$$\leqq \langle P_\infty(s-n\omega)x, x\rangle_H + \langle GS_\infty(s, s-n\omega)x, S_\infty(s, s-n\omega)x\rangle_H, \tag{4.51}$$

$$n \geqq 0, \quad x \in H.$$

Since $\|S_\infty(s, s-n\omega)x\| \to 0$ for each x, $\|S_\infty(s, s-n\omega)x\|$ and $\|\hat{P}(s-n\omega)\|$ are uniformly bounded in n. Then (4.45), the generalized Schwarz inequality, and an argument similar to the latter part of the proof of Theorem 4.2 establish (4.49), since $P_\infty(s-n\omega) = P_\infty(s)$.

To prove (4.50), denote the optimal control for the problem of Section 3 with $t_f = s$, $P(t_f) = G$, initial time $s - n\omega$, and initial state x by $\hat{u}_n(\cdot)$; and denote the corresponding optimal state by $\hat{x}_n(\cdot)$. Then, since $G \geqq P_\infty(s)$,

$$\begin{aligned}\langle \hat{P}(s-n\omega)x, x\rangle_H &\geqq \int_{s-n\omega}^{s} (\langle D(\eta)\hat{x}_n(\eta), \hat{x}_n(\eta)\rangle_H + \langle Q(\eta)\hat{u}_n(\eta), \hat{u}_n(\eta)\rangle_U)\, d\eta \\ &\quad + \langle P_\infty(s)\hat{x}_n(s), \hat{x}_n(s)\rangle_H \\ &\geqq \langle P_\infty(s-n\omega)x, x\rangle_H.\end{aligned} \tag{4.52}$$

Since $P_\infty(\cdot)$ has period ω, (4.46), (4.51), and (4.52) imply (4.50).

To end this section on a familiar note, let us now derive the Hilbert space version of the Riccati algebraic equation. Let $B(\cdot)$, $D(\cdot)$, and $Q(\cdot)$ be constant operators, and let $T(t, s) = T(t-s)$ be a strongly continuous semigroup with generator A. Since we can consider this problem to be periodic with arbitrary period ω, we see that $P_\infty(\cdot)$ is constant and that $S_\infty(t, s) = S_\infty(t-s)$ is the strongly continuous semigroup whose generator is$\cdot \hat{A} = A - BQ^{-1}B^*P_\infty$. Also, $T^*(\cdot)$ and $S^*(\cdot)$ are strongly continuous semigroups, with generators A^* and $\hat{A}^*$, respectively. Note that A and $\hat{A}$ have the same domain and that A^* and $\hat{A}^*$ have the same domain.

In (4.28) we can take the limit as $t \to \infty$. To see this, recall that, for fixed s, $\langle P_\infty(t)S_\infty(t, s)x, S_\infty(t, s)x\rangle_H \searrow 0$ as $t \to \infty$. Thus, for each x, $\|P_\infty^{1/2}(t)S_\infty(t, s)x\| \to 0$ and the principle of uniform boundedness then says that $\|P^{1/2}(\cdot)S_\infty(\cdot, s)\| = \|S_\infty^*(\cdot, s)P_\infty^{1/2}(\cdot)\|$ is uniformly bounded. Therefore, $\|S_\infty^*(t, s)P_\infty(t)S_\infty(t, s)x\| \to 0$ as $t \to \infty$. A similar argument shows that the norm of the integrand converges to zero. Note that the argument for taking the limit in (4.28) depends only on the existence of a nonnegative solution of the Riccati integral equations on the infinite interval. (See Theorem 4.5.) Then, for the time-invariant problem we are considering now,

$$P_\infty x = \int_0^\infty S_\infty^*(\eta)[D + P_\infty BQ^{-1}B^*P_\infty]S_\infty(\eta)x\, d\eta, \qquad x \in H. \tag{4.53}$$

Next, we show that P_∞ maps the domain of $\hat{A}$ into the domain of $\hat{A}^*$. For $0 \leqq t < \infty$, $h \geqq 0$, and x in the domain of $\hat{A}$,

$$\begin{aligned}&[S_\infty^*(h) - I]\int_0^t S_\infty^*(\eta)[D + P_\infty BQ^{-1}B^*P_\infty]S_\infty(\eta)x\, d\eta \\ &= \int_0^t \chi(\eta, h)S_\infty^*(\eta)[D + P_\infty BQ^{-1}B^*P_\infty]S_\infty(\eta - h)[I - S_\infty(h)]x\, d\eta \\ &\quad + \int_t^{t+h} S_\infty^*(\eta)[D + P_\infty BQ^{-1}B^*P_\infty]S_\infty(\eta - h)x\, d\eta \\ &\quad - \int_0^h S_\infty^*(\eta)[D + P_\infty BQ^{-1}B^*P_\infty]S_\infty(\eta)x\, d\eta,\end{aligned} \tag{4.54}$$

where $\chi(\eta, h)$ is the characteristic function of $\{(\eta, h): \eta \geqq h\}$. Multiplying (4.54) by $1/h$ and using dominated convergence to take the limit as $h \to 0$, we obtain

$$\begin{aligned}\hat{A}^*\int_0^t & S_\infty^*(\eta)[D+P_\infty BQ^{-1}B^*P_\infty]S_\infty(\eta)x\,d\eta \\ &= -\int_0^t S_\infty^*(\eta)[D+P_\infty BQ^{-1}B^*P_\infty]S_\infty(\eta)\hat{A}x\,d\eta \\ &\quad + S_\infty^*(t)[D+P_\infty BQ^{-1}B^*P_\infty]S_\infty(t)x-[D+P_\infty BQ^{-1}B^*P_\infty]x.\end{aligned} \tag{4.55}$$

As $t \to \infty$, the integral on the left side of (4.55) converges to $P_\infty x$, and, since the integrand in (4.53) converges to zero, the right side of (4.55) converges to $-P_\infty \hat{A}x - [D+P_\infty BQ^{-1}B^*P_\infty]x$. Therefore, since $\hat{A}^*$ is closed, $P_\infty x$ is in the domain of $\hat{A}^*$, and substituting $A-BQ^{-1}B^*P_\infty$ for $\hat{A}$, we obtain the Riccati algebraic equation

$$A^*P_\infty + P_\infty A - P_\infty BQ^{-1}B^*P_\infty + D = 0. \tag{4.56}$$

Although (4.55) is valid only for x in the domain of A, (4.56) is justified because $P_\infty BQ^{-1}B^*P_\infty - D$ is a bounded operator on H, so that $A^*P_\infty + P_\infty A$ has a bounded extension to all of H.

By imposing certain restrictions on the generator A, Lukes and Russell showed in [11] that P_∞ satisfies (4.56). They did not show that P_∞ maps all of the domain of A into the domain of A^*, and they did not obtain the minimality and uniqueness results that follow from Theorems 4.5 and 4.6 after the following definition.

DEFINITION 4.4. Let A generate a strongly continuous semigroup on H, and let the constant operators $B, D,$ and Q be as defined in our control problem. An operator $P \in \mathscr{L}(H, H)$ is called a *solution of the Riccati algebraic equation* if P maps the domain of A into the domain of A^* and satisfies (4.56). (Again, the notation P_∞ is reserved for the P_∞ of the previous theorems.)

Let P be a nonnegative, self-adjoint solution of the Riccati algebraic equation. For $-\infty < s \leqq t < \infty$ and $x \in$ domain A, we have

$$\begin{aligned}T^*(t-s)PT(t-s)x + \int_s^t & T^*(\eta-s)[D-PBQ^{-1}B^*P]T(\eta-s)x\,d\eta \\ &= T^*(t-s)PT(t-s)x - \int_s^t T^*(\eta-s)[A^*P+PA]T(\eta-s)x\,d\eta \\ &= T^*(t-s)PT(t-s)x - \int_s^t \frac{d}{d\eta}[T^*(\eta-s)PT(\eta-s)x]d\eta = Px.\end{aligned} \tag{4.57}$$

Since domain A is dense in H, (4.57) shows that P is a uniformly bounded, nonnegative solution of the Riccati integral equations on the infinite interval. Then, referring to Theorem 4.9, we obtain the concluding theorem of this section.

THEOREM 4.11. *Let the constant operators A, B, D, and Q be as previously defined, with A the generator of the strongly continuous semigroup $T(\cdot)$. There exists a nonnegative, self-adjoint solution of the Riccati algebraic equation* (4.56) *if and only if, for each $x \in H$, there is an admissible control for the initial time* 0 *and the initial state x. If P is such a solution, we have $P \geqq P_\infty$. When P_∞ exists, the optimal control $u(\cdot)$ and optimal trajectory $x(\cdot)$ for the control problem of this section are given by $u(t) = -Q^{-1}B^*P_\infty x(t)$ and $x(t) = S_\infty(t-s)x(s)$, where $S_\infty(\cdot)$ is the strongly continuous semigroup generated by $\hat{A} = A - BQ^{-1}B^*P_\infty$. If D is positive definite, P_∞ is the unique nonnegative, self-adjoint solution of the Riccati algebraic equation, and $S_\infty(\cdot)$ is uniformly exponentially stable.*

5. Approximation theory. To compute numerically the solutions to control problems of the class considered here, we need to know conditions under which the solutions to a sequence of finite dimensional optimal control problems converge to the solution to a given infinite dimensional problem. Such convergence has been proved for problems involving particular types of differential equations—see, for example, Delfour [6] and Lions [10]—and the purpose of this section is to present the convergence results which are possible and important for the very general class of systems covered by the analysis of the preceding sections.

In practice, computations for the finite dimensional approximate problems will probably be based upon ordinary differential equations, and it will be necessary to examine the particular differential equations representing the infinite dimensional systems in order to verify the hypotheses of our approximation theorems. However, it should be emphasized that the results presented here do not depend on an infinite dimensional Riccati differential equation; we refer only to solutions of the Riccati integral equations of §§ 3 and 4.

When the present paper is compared with other literature on the linear quadratic control problem, probably the most novel feature of the analysis of § 3 will be found to be the use of the composite operator $\tilde{Q}_s^{-1}\tilde{B}_s^* \in \mathscr{B}_\infty(s, t_f; H, U)$ to infer the linear feedback structure of the optimal control and to define the perturbed evolution operator $S(\cdot,\cdot)$ in terms of the operators $T(\cdot,\cdot), B(\cdot), D(\cdot), Q(\cdot)$, and G. This definition of $S(\cdot,\cdot)$ is the key to the approximation results of this section.

Referring to the optimal control problem of § 3, suppose that $\{T_i(\cdot,\cdot)\}$ is a sequence of evolution operators on H and that $\{B_i(\cdot)\}, \{D_i(\cdot)\}, \{Q_i(\cdot)\}$, and $\{G_i\}$ are sequences of operators in $\mathscr{B}_\infty(t_0, t_f; U, H), \mathscr{B}_\infty(t_0, t_f; H, H), \mathscr{B}_\infty(t_0, t_f; U, U)$, and $\mathscr{L}(H, H)$, respectively, with $D_i(\cdot), Q_i(\cdot)$, and G_i nonnegative and self-adjoint. We consider the sequence of optimal control problems corresponding to these sequences of operators, and give sufficient conditions for the sequence of solutions to converge, in an appropriate sense, to the solution of the original problem. Our hypotheses pertaining to convergence of the sequences of operators should be verifiable for standard techniques, such as the Galerkin method, for the numerical solution of partial and functional differential equations. Suppose that, for each $x \in H$ and $u \in U$,

$$T_i(t, s)x \to T(t, s)x \quad \text{strongly}, \quad t_0 \leqq s \leqq t \leqq t_f, \tag{5.1}$$

$$T_i^*(t, s)x \to T^*(t, s)x \quad \text{strongly}, \quad t_0 \leqq s \leqq t \leqq t_f, \tag{5.2}$$

$$B_i(t)u \to B(t)u \quad \text{strongly, a.e.}, \tag{5.3}$$

$$B_i^*(t)x \to B^*(t)x \quad \text{strongly, a.e.}, \tag{5.4}$$

$$D_i(t)x \to D(t)x \quad \text{strongly, a.e.}, \tag{5.5}$$

$$Q_i(t)u \to Q(t)u \quad \text{strongly, a.e.}, \tag{5.6}$$

$$G_i x \to Gx \quad \text{strongly}, \tag{5.7}$$

as $i \to \infty$. We require $\|T_i(t, s)\|, \|B_i\|_{\mathscr{B}_\infty}, \|D_i\|_{\mathscr{B}_\infty}, \|Q_i\|_{\mathscr{B}_\infty}$, and $\|G_i\|$ to be uniformly bounded in i, t, and s, and require a constant m such that, for each i, $Q_i(t) \geqq m > 0$ for almost all t. Therefore, $\|\tilde{Q}_{si}^{-1}\tilde{B}_{si}^*(\cdot)\|_{\mathscr{B}_\infty}$ is uniformly essentially bounded for $t_0 \leqq s \leqq t_f$. (5.1)–(5.7), the uniform bounds, and the dominated convergence theorem will imply the convergence we need.

From (3.3)–(3.5) and (3.10)–(3.13), we see that

$$(\tilde{Q}_{si}v)(t) \to (\tilde{Q}_s v)(t) \quad \text{strongly}, \quad v \in L_2(s, t_f; U) \tag{5.8}$$

and

$$(5.9) \qquad \tilde{B}^*_{si}(t)x \to \tilde{B}^*_s(t)x \quad \text{strongly}, \quad x \in H,$$

pointwise (a.e.) and in the L_2 sense. (Recall that $\tilde{Q}_s \in \mathcal{L}(\mathcal{U}_s, \mathcal{U}_s)$ and $\tilde{B}^*_s(\cdot) \in \mathcal{B}_\infty(s, t_f; H, U)$.) The identity

$$(5.10) \qquad \tilde{Q}_{si}^{-1} - \tilde{Q}_s^{-1} = \tilde{Q}_{si}^{-1}(\tilde{Q}_s - \tilde{Q}_{si})\tilde{Q}_s^{-1}),$$

the uniform bound for $\|\tilde{Q}_{si}\|$, and (5.8) and (5.9) imply

$$(5.11) \qquad (\tilde{Q}_{si}^{-1}\tilde{B}^*_{si})(t)x \to (\tilde{Q}_s^{-1}\tilde{B}^*_s)(t)x \quad \text{strongly}, \quad x \in H,$$

pointwise and in the L_2 sense.

For our sequence of optimal control problems, we denote the optimal controls by $u_i(\cdot)$, the sequence of perturbed evolution operators by $S_i(\cdot, \cdot)$, and the sequence of solutions of the Riccati integral equations by $P_i(\cdot)$. (3.14), (5.9), and (5.11) yield pointwise and L_2 convergence for the optimal trajectories $x_i(t) = S_i(t, s)x_i(s)$. Now that we have strong convergence for $T_i^*(\cdot, \cdot)$, G_i, $D_i(\cdot)$, and $S_i(\cdot, \cdot)$, (3.25) yields strong convergence for $P_i(\cdot)$.

Of course, uniform convergence on compact time intervals is very important, and, under reasonable hypotheses, this is what we have. The additional hypotheses are that $T(\cdot, \cdot)$ and $T^*(\cdot, \cdot)$ be jointly strongly continuous, that the operators $B(\cdot)$, $B^*(\cdot)$, $D(\cdot)$, and $Q(\cdot)$ be piecewise strongly continuous (i.e., possess a finite number of discontinuities), that the convergence in (5.1) and (5.2) be uniform for $t_0 \leqq s \leqq t \leqq t_f$, and that the convergence in (5.3)–(5.6) be uniform almost everywhere. Joint strong continuity of $T(\cdot, \cdot)$ and (3.18) imply joint strong continuity of $S(\cdot, \cdot)$, and the appropriate uniform convergence (see Theorem 5.1) can be established with the equations already cited and repeated application of the following lemma, whose proof is an easy exercise.

LEMMA 5.1. *Let X and Y be Banach spaces and let Ω be a compact subset of R^n. Let $A(\cdot)\colon \Omega \to \mathcal{L}(X, Y)$, and, for $i \geqq 1$, let $A_i(\cdot)\colon \Omega \to \mathcal{L}(X, Y)$. Suppose that $\|A_i(\xi)\|$ is uniformly bounded in i and ξ, and that, for each $x \in X$, $A_i(\xi)x$ converges to $A(\xi)x$ uniformly in ξ. Let $g(\cdot)\colon \Omega \to X$ be continuous, and suppose there is a sequence of functions $g_i(\cdot)$ which converge uniformly to $g(\cdot)$. Then $A_i(\cdot)g_i(\cdot)$ converges uniformly to $A(\cdot)g(\cdot)$.*

Our approximation results for the problem of Section 3 are summarized in the following theorem.

THEOREM 5.1. *Let* (5.1)–(5.7) *hold, along with the uniform bounds already stated. For our sequence of control problems, denote the initial states by $x_i(t_0)$ and let $x_i(t_0) \to x(t_0)$; denote the optimal controls by $u_i(\cdot)$, the optimal trajectories by $x_i(\cdot)$, and the solutions of the Riccati integral equations by $P_i(\cdot)$. For the original problem of* § 3, *denote the corresponding quantities by $x(t_0)$, $u(\cdot)$, $x(\cdot)$, and $P(\cdot)$. Then we have*

$$(5.12) \qquad u_i(t) \to u(t) \quad \textit{strongly}, \quad \textit{a.e. and in } L_2(t_0, t_f; U),$$

$$(5.13) \qquad x_i(t) \to x(t) \quad \textit{strongly, pointwise and in } L_2(t_0, t_f; H),$$

and, for $x \in H$,

$$(5.14) \qquad P_i(t)x \to P(t)x \quad \textit{strongly, pointwise and in } L_2(t_0, t_f; H).$$

If $T(\cdot, \cdot)$ is jointly strongly continuous and $B(\cdot)$, $B^*(\cdot)$, $D(\cdot)$, and $Q(\cdot)$ are piecewise strongly continuous, uniform convergence in (5.1)–(5.6) *implies uniform convergence in* (5.12)–(5.14).

In § 4, we saw that the solution of the Riccati integral equations on the infinite interval could be obtained as the limit of the solutions of Riccati integral equations on a sequence of finite time intervals. (See Theorem 4.2.) Theoretically then, we could compute the solutions of the Riccati equations on the finite intervals according to Theorem 5.1, and, with increasing lengths of the intervals, these solutions would converge to $P_\infty(\cdot)$. However, this order of convergence is not very practical for computational purposes; we need to be able to reverse the order of the limits. That is to say, we would like to know that, if $P_{\infty i}(\cdot)$ is the $P_\infty(\cdot)$ for the ith approximate optimal control problem, then $P_{\infty i}(\cdot)$ converges, in some sense, to the $P_\infty(\cdot)$ of the original problem. Based on the results of § 4, there are two sets of sufficient conditions for such convergence. Although the verification of the hypotheses of either of the following theorems may be nontrivial in particular applications, we should be able to expect at least one of these sets of hypotheses in realistic examples.

First, consider the most general problem of § 4 and assume that there exists a solution of the Riccati integral equations on the infinite interval. From Theorem 4.5, we know that, although $P_\infty(\cdot)$ may not be the unique solution, it is the minimal nonnegative solution. Thus we might expect that one way to approximate $P_\infty(\cdot)$ is with a sequence of $P_{\infty i}(\cdot)$'s that converge to $P_\infty(\cdot)$ from below. This approach is sometimes possible when the sequence of approximate problems represents the orthogonal "projection" of the original problem onto a sequence of subspaces of increasing dimension (see Lukes and Russell [11]). We have the following theorem.

THEOREM 5.2. *For the optimal control problem on the infinite interval, assume that there exists a solution of the Riccati integral equations on the infinite interval and that we have a sequence of approximate problems for which our previous hypotheses concerning convergence and uniform boundedness hold on compact time intervals. Denote by* $P_{\infty i}(\cdot)$ *the minimal solution of the Riccati integral equations on the infinite interval for the* ith *approximate problem. If*

$$(5.15) \qquad P_{\infty i}(t) \leqq P_{\infty j}(t) \leqq P_\infty(t), \qquad t_0 \leqq t,\ i \leqq j,$$

then, for $x \in H$,

$$(5.16) \qquad P_{\infty i}(t)x \to P_\infty(t)x \quad \textit{strongly}, \quad t_0 \leqq t.$$

If $T(\cdot,\cdot)$ *is jointly strongly continuous and* $B(\cdot)$, $B^*(\cdot)$, $D(\cdot)$, *and* $Q(\cdot)$ *are piecewise strongly continuous, uniform convergence on compact intervals in* (5.1)–(5.6) *implies uniform convergence on compact intervals in* (5.16).

Proof. Inequality (5.15) implies that, for $t \leqq t$, $P_{\infty i}(t)$ converges strongly to some nonnegative, self-adjoint $P(t)$. This, with (5.1)–(5.6) and the dominated convergence theorem, implies that $P(\cdot)$ is a nonnegative solution of the Riccati integral equations on the infinite interval. The second inequality in (5.15) shows that $P(\cdot) \leqq P_\infty(\cdot)$, which, with Theorem 4.5, implies $P(\cdot) = P_\infty(\cdot)$. The uniform convergence on compact intervals follows from Theorem 5.1 with $G_i = P_{\infty i}(t)$ and $G = P_\infty(t)$ for $t_0 \leqq t$.

Unfortunately, it may be quite difficult to choose an approximation scheme for which (5.15) can be shown. The final theorem pertains to periodic problems for which the feedback control system (i.e., the optimally controlled system) is uniformly exponentially stable, and does not require monotonicity of $\{P_{\infty i}(\cdot)\}$.

THEOREM 5.3. *For the* ω*-periodic optimal control problem on the infinite interval, assume that we have a sequence of* ω*-periodic approximate problems defined as above and that our previous hypotheses concerning convergence and uniform boundedness hold on compact time intervals. Let* $P_{\infty i}(\cdot)$ *be the minimal nonnegative* ω*-periodic solution of the Riccati integral equations for the* ith *approximate problem, and let* $S_{\infty i}(\cdot,\cdot)$ *be the* *corresponding perturbed evolution operator. Assume also that there is at most one nonnegative ω-periodic solution of the Riccati integral equations for the original problem. Then, if there exist positive constants M and β such that*

$$\|S_{\infty i}(t, s)\| \leqq M e^{-\beta(t-s)}, \qquad -\infty < s \leqq t < \infty, \quad i \geqq 1, \tag{5.17}$$

and, for some s and $\hat{M}$,

$$\|P_{\infty i}(s)\| \leqq \hat{M}, \qquad i \geqq 1, \tag{5.18}$$

then there exists a nonnegative ω-periodic solution $P_\infty(\cdot)$ of the Riccati integral equations for the original problem,

$$P_{\infty i}(t)x \to P_\infty(t)x, \qquad x \in H, \quad -\infty < t < \infty, \tag{5.19}$$

$$S_{\infty i}(t, s)x \to S_\infty(t, s)x, \qquad x \in H, \quad -\infty < s \leqq t < \infty, \tag{5.20}$$

and

$$\|S_\infty(t, s)\| \leqq M e^{-\beta(t-s)}, \qquad -\infty < s \leqq < \infty. \tag{5.21}$$

If there exists a positive constant m such that

$$D_i(t) \geqq m, \quad \text{a.e.}, i \geqq 1, \tag{5.22}$$

then the uniqueness hypothesis on $P_\infty(\cdot)$ holds, and (5.18) *implies* (5.17) *and hence* (5.19), (5.20), *and* (5.21). *Also, the additional hypotheses that ensure uniform convergence on compact intervals in Theorem* 5.2 *ensure uniform convergence on compact intervals in* (5.19) *and* (5.20).

Proof. First we prove that (5.17) and (5.18) imply the existence of $P_\infty(\cdot)$ and (5.19)–(5.20). Compute $P_{\infty i}(s)$ according to (4.51) of Theorem 4.10, with $G = \hat{M}I$ (I is the identity). For $\varepsilon > 0$, we can choose n according to (4.50) such that

$$\|\hat{P}_i(s - n\omega) - P_{\infty i}(s)\| \leqq \varepsilon, \qquad i \geqq 1. \tag{5.23}$$

Holding n fixed, for $x \in H$, choose $\bar{\iota}$ according to Theorem 5.1 such that

$$\|\hat{P}_i(s - n\omega)x - \hat{P}_j(s - n\omega)x\| \leqq \varepsilon\|x\|, \qquad i \geqq \bar{\iota} \text{ and } j \geqq \bar{\iota}. \tag{5.24}$$

Then we have

$$\begin{aligned} &\|P_{\infty i}(s)x - P_{\infty j}(s)x\| \\ &\leqq \|P_{\infty i}(s)x - \hat{P}_i(s - n\omega)x\| + \|\hat{P}_i(s - n\omega)x - \hat{P}_j(s - n\omega)x\| + \|\hat{P}_j(s - n\omega)x - P_{\infty j}(s)x\| \\ &\leqq 3\varepsilon\|x\|, \qquad i \geqq \bar{\iota} \text{ and } j \geqq \bar{\iota}. \quad (\bar{\iota} \text{ depends on } x.) \end{aligned} \tag{5.25}$$

Thus $P_{\infty i}(s)$ converges strongly to some nonnegative, self-adjoint $P(s)$. Then Theorem (5.1) with $G_i = P_{\infty i}(s)$ and $G = P(s)$ implies that, for $t \leqq s$, $P_{\infty i}(t)$ converges strongly to some $P(t)$, which is a nonnegative ω-periodic solution of the Riccati integral equations for the original problem. The uniqueness assumption implies that this $P(t)$ must be $P_\infty(t)$ and hence (5.19). Theorem (5.1) also implies (5.20), which implies (5.21), and, under the additional hypotheses, the uniform convergence on compact intervals.

Now suppose that (5.22) holds. Then the uniqueness of $P_\infty(\cdot)$ follows from Theorem 4.6 and Lemma 4.1. To see that (5.18) and (5.22) imply (5.17), refer to Theorem 2.2. The uniform bounds on $T_i(\cdot, \cdot)$, $Q_i(\cdot)$, and $B_i(\cdot)$ with Theorem 4.7 and its proof imply the existence of constants M_2 and α for which (2.11) holds with $T(\cdot, \cdot)$ replaced by $S_{\infty i}(\cdot, \cdot)$ for $i \geqq 1$. Also for $i \geqq 1$, (2.12) holds with $T(\cdot, \cdot)$ replaced by

$S_{\infty i}(\cdot,\cdot)$ and $M_3=\hat{M}/m$. Thus, by the discussion following Theorem 2.2, there are positive constants M and β for which (5.17) holds.

Probably the most important result here for approximations to solutions of the Riccati integral equations on the infinite interval is that, when (5.22) and the other hypotheses hold, (5.18) implies (5.19). Note that, if $P_{\infty i}(s)x$ converges (to anything) for each x in H, then the principle of uniform boundedness says that (5.18) holds. Also, since $\langle P_{\infty i}(s)x, x\rangle_H$ is the minimum value of the cost functional $J_{\infty i}(s, x, \cdot)$, if a sequence of controls u_i can be shown to exist such that $J_{\infty i}(s, x, u_i)$ is bounded in i, then (5.18) holds; for example, if there exist positive constants M and β such that (5.17) holds with $S_{\infty i}(t, s)$ replaced by $T_i(t, s)$, let $u_i=0$.

Appendix A. The definitions and properties listed here are needed to justify many of the integrals in this paper. Except for Properties A.4 and A.5, the following are standard results (see Hille and Phillips [8]) on strong measurability and Bochner integration. As in the rest of the paper, H and U are Hilbert spaces and (t_0, t_f) is an interval of the real line.

DEFINITION A.1. A function $x(\cdot)$: $(t_0, t_f)\to H$ is *strongly measurable* if $x(\cdot)$ is the limit almost everywhere of a sequence of countably valued functions. $x(\cdot)$ is *weakly measurable* if $\langle y, x(\cdot)\rangle_H$ is Lebesgue measurable for each $y\in H$.

Property A.1. If H is separable, strong measurability and weak measurability are equivalent.

Property A.2. A function $x(\cdot)$ is strongly measurable if and only if it is the uniform limit almost everywhere of a sequence of countably valued functions.

DEFINITION A.2. An operator-valued function $B(\cdot)$: $(t_0, t_f)\to\mathscr{L}(U, H)$ is called *strongly measurable* if $B(\cdot)x$ is strongly measurable for each $x\in H$. The set of all such functions $B(\cdot)$ for which $\|B(\cdot)\|$ is essentially bounded on (t_0, t_f) is denoted by $\mathscr{B}_\infty(t_0, t_f; U, H)$.

Property A.3. Property A.2 can be used to show that, under the norm $\|B(\cdot)\|_{\mathscr{B}_\infty}=$ ess sup $\|B(\cdot)\|$, $\mathscr{B}_\infty(t_0, t_f; U, H)$ is a Banach space and $\mathscr{B}_\infty(t_0, t_f; H, H)$ is a Banach algebra.

Property A.4. If $Q(\cdot)$: $(t_0, t_f)\to\mathscr{L}(U, U)$ is self-adjoint and, for some $m>0$, $Q(\cdot)\geqq m$ almost everywhere, then $Q(\cdot)\in\mathscr{B}_\infty(t_0, t_f; U, U)$ implies $Q^{-1}(\cdot)\in\mathscr{B}_\infty(t_0, t_f; U, U)$.

Proof. We may assume $\|Q(\cdot)\|_{\mathscr{B}_\infty}<1$. Then the Neumann series for $[I-(I-Q(\cdot))]^{-1}$ converges uniformly almost everywhere, i.e., in the Banach algebra $\mathscr{B}_\infty(t_0, t_f; U, U)$.

Property A.5. If $Q(\cdot)$: $(t_0, t_f)\to\mathscr{L}(U, U)$ is nonnegative and self-adjoint, then $Q(\cdot)\in\mathscr{B}_\infty(t_0, t_f; U, U)$ if and only if $Q^{1/2}(\cdot)\in\mathscr{B}_\infty(t_0, t_f; U, U)$, where $Q^{1/2}(\cdot)$ is the nonnegative, self-adjoint square root of $Q(\cdot)$.

Proof. (Only if.) We have (see Kato [9])

$$Q^{1/2}(\cdot)=(1/\pi)\int_0^\infty \lambda^{-1/2}(Q(\cdot)+\lambda)^{-1}Q(\cdot)\,d\lambda,$$

where the integral is absolutely convergent in $\mathscr{B}_\infty(t_0, t_f; U, U)$.

The Bochner integral is an extension of the Lebesgue integral to vector-valued functions. For a systematic development of the Bochner integral, see [8].

Property A.6. A function $x(\cdot)$: $(t_0, t_f)\to H$ is Bochner integrable if and only if $x(\cdot)$ is strongly measurable and

$$\int_{t_0}^{t_f}\|x(t)\|\,dt<\infty.$$

Property A.7. For $1 \leqq p < \infty$, we have the Banach space $L_p(t_0, t_f; H)$ of strongly measurable H-valued functions $x(\cdot)$ for which

$$\int_{t_0}^{t_f} \|x(t)\|^p \, dt < \infty.$$

$L_2(t_0, t_f; H)$ is a Hilbert space with the inner product

$$\langle x(\cdot), y(\cdot)\rangle_{L_2} = \int_{t_0}^{t_f} \langle x(t), y(t)\rangle_H \, dt$$

Also, $B_\infty(t_0, t_f; U, H) \subset \mathscr{L}(L_p(t_0, t_f; U), L_p(t_0, t_f; H))$.

Property A.8 (Dominated Convergence Theorem). If $\{x_n(\cdot)\} \subset L_1(t_0, t_f; H)$ converges almost everywhere to a function $x(\cdot)$ and if there exists a function $f \in L_1(t_0, t_f; R)$ such that $\|x_n(t)\| \leqq f(t)$ for all n and almost all t, then $x(\cdot) \in L_1(t_0, t_f; H)$ and

$$\lim_{n\to\infty} \int_{t_0}^{t_f} x_n(t) \, dt = \int_{t_0}^{t} x(t) \, dt.$$

Property A.9. Fubini's theorem holds for Bochner integrals, and its application in this paper is illustrated by the following example. Let $T(\cdot, \cdot)$ be an evolution operator on H, let $\phi(\cdot) \in L_1(t_0, t_f; H)$, and let $\chi(\eta, \xi)$ be the characteristic function of $\{(\eta, \xi): \eta \geqq \xi\}$. Then

$$\begin{aligned}\int_{t_0}^{t_f} T(t_f, \eta) \int_{t_0}^{\eta} \phi(\xi) \, d\xi \, d\eta &= \int_{t_0}^{t_f}\int_{t_0}^{t_f} \chi(\eta, \xi) T(t_f, \eta)\phi(\xi) \, d\xi \, d\eta \\ &\overset{\text{(Fubini)}}{=} \int_{t_0}^{t_f}\int_{t_0}^{t_f} \chi(\eta, \xi) T(t_f, \eta)\phi(\xi) \, d\eta d\xi \\ &= \int_{t_0}^{t_f}\int_{\xi}^{t_f} T(t_f, \eta)\phi(\xi) \, d\eta \, d\xi.\end{aligned}$$

Appendix B. The following is an example of an evolution operator, as defined in Definition 2.1, which is not uniformly bounded as in (2.4).

Let H be l_2, the space of all square-summable sequences of real numbers, and write $x = (x_1, x_2, \cdots, x_n, \cdots) \in l_2$. Let $t_n = 1 - 1/n$, $n \geqq 1$, and let $\Delta_n = t_{n+1} - t_n = 1/(n(n+1))$. Define $a_n(t)$ by

$$a_n(t) = \begin{cases} -3n^2(n+1), & t_n < t \leqq t_n + \frac{1}{3}\Delta_n, \\ +3n^2(n+1), & t_n + \frac{1}{3}\Delta_n < t \leqq t_n + \frac{2}{3}\Delta_n, \\ -3n^2(n+1), & t_n + \frac{2}{3}\Delta_n < t < t_{n+1}, \\ 0, & \text{otherwise}. \end{cases}$$

For $-\infty < s \leqq t < \infty$, define $T(t, s)$ by

$$T(t, s)x = y,$$

where

$$y_n = x_n \exp \int_s^t a_n(\eta) \, d\eta, \qquad n \geqq 1.$$

$T(\cdot,\cdot)$ can be shown to satisfy (2.1)–(2.3) for $-\infty < s \leqq t < \infty$. In particular, we have

$$\lim_{t\to 1} T(t,s)x = T(1,s)x$$

and

$$\lim_{s\to 1} T(1,s)x = x$$

for $x \in l_2$, where the limits are in the l_2 norm.

For $k \geqq 1$, let $x^k = (x_1^k, x_2^k, \cdots, x_n^k, \cdots)$ with $x_k^k = 1$ and $x_n^k = 0$, $n \neq k$. Then $\|x^k\|_{l_2} = 1$ and

$$\|T(t_k + \tfrac{2}{3}\Delta_k, t_k + \tfrac{1}{3}\Delta_k)x^k\| = e^k \to \infty \quad \text{as } k \to \infty.$$

REFERENCES

[1] R. F. CURTAIN, *The infinite-dimensional Riccati equation with applications to affine hereditary differential systems*, this Journal, 13 (1975), pp. 1130–1143.

[2] R. F. CURTAIN AND A. J. PRITCHARD, *The infinite-dimensional Riccati equation for systems defined by evolution operators*, this Journal, 14 (1976), pp. 951–983.

[3] R. DATKO, *A linear control problem in abstract Hilbert Space*, J. Differential Equations, 9 (1971), pp. 346–359.

[4] ———, *Unconstrained control problems with unconstrained cost*, this Journal, 11 (1973) pp. 32–52.

[5] ———, *Uniform asympototic stability of evolutionary processes in a Banach space*, SIAM J. Math. Anal., 3 (1972), pp. 428–445.

[6] 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*, Technical Report CRM-408, Centre de Recherches Mathematiques, Université de Montreal, June 1974.

[7] M. C. DELFOUR AND S. K. MITTER, *Controllability, observability and optimal feedback control of affine hereditary differential systems*, this Journal, 10 (1972), pp. 298–327.

[8] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, Colloquium Publications, vol. 31, American Mathematical Society, Providence, RI, 1957.

[9] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

[10] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.

[11] D. L. LUKES AND D. L. RUSSELL, *The quadratic criterion for distributed systems*, this Journal, 7 (1969), pp. 101–121.

[12] R. S. PHILLIPS, *Perturbation theory for semigroups of linear operators*, Trans. Amer. Math. Soc., 74 (1953), pp. 199–221.

[13] R. B. VINTER AND T. L. JOHNSON, *Optimal control of nonsymmetric hyperbolic systems in n variables on the half-space*, this Journal, 15 (1977) pp. 129–143.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 5, September 1979

0363-0129/79/1705-0001 $01.00/0

SHADOW PRICES AND DUALITY FOR A CLASS OF OPTIMAL CONTROL PROBLEMS*

J. P. AUBIN† AND F. H. CLARKE‡

Abstract. A class of optimal control problems is considered in which the cost functional is locally Lipschitz (not necessarily convex or differentiable) and the dynamics linear and/or convex. By using generalized gradients and duality methods of functional analysis, necessary conditions are obtained in which the dual variables admit interpretation as shadow prices (or rates of change of the value function). Applications are presented in three settings: infinite-horizon optimal control, optimal control of partial differential equations, and a variational problem with unilateral state constraints. A theorem is proved which characterizes the generalized gradients of integral functionals on L^p.

Introduction. We treat in this article a general class of optimal control problems in which the cost functional need only be locally Lipschitz, and in which the constraints exhibit linearity and convexity. We obtain for these hybrid problems necessary conditions with features usually associated only with the fully convex case. Chief among these is the interpretation of the multipliers as rates of change of the optimal value in the problem as the constraints undergo perturbation. In the context of mathematical programming such interpretations have figured importantly in mathematical economics (see for example [13] for a discussion), where the rates of change are in turn interpreted as "shadow prices". A further feature is the ability to guarantee the strong (Kuhn–Tucker or normal) form of the necessary conditions by means of a Slater-type constraint qualification independent of the solution to the problem. The method used here to treat these problems is novel, and employs the generalized gradient of the value function, a minimax theorem, and an abstract Green formula.

We present three applications of the abstract theory which we believe to be of independent interest; they are in settings characterized by technical difficulties. The first of these (§ 1) involves a control problem over an infinite interval, and sheds new light on the sensitive relationship between the growth rates of the cost functional and the adjoint variables, and on the role played by the size of the discount factor. In § 2 we give an example to illustrate the importance of this consideration, while § 3 gives the proof of the theorem in § 1. We have placed this application before the abstract theory in order to display the line of reasoning common to both in a more concrete, and hence more easily assimilated setting.

Section 4 recalls a stability result used in the proof, while §§ 5 and 6 develop characterizations of generalized gradients of certain functionals which are useful in interpreting the abstract conditions; these results complement other characterizations given in [9]. The abstract problem and its analysis appear in §§ 7 and 8. In § 9 an application is made to a nondifferentiable, nonconvex problem involving control of partial differential equations. While this is the first such result that we know of, it should be clear that many other similar problems could be framed within the abstract theory. The final section illustrates the use of the theory in the presence of unilateral state constraints.

We conclude by recalling for the reader's convenience the definition of the generalized gradient in the case of a locally Lipschitz function $f: X \to R$, where X is a Banach space (details appear in [7], [9]). Given v, the generalized directional derivative

* Received by the editors August 16, 1978, and in revised form January 9, 1979. This research was supported in part by National Research Council of Canada under Grant No. A 9082.

† Université de Paris IX-Dauphine 75775, Paris, France.

‡ Mathematics Department, University of British Columbia, Vancouver, B.C., Canada V6T IW5.

$f^\circ(x;v)$ is defined by

$$f^\circ(x;v)=\limsup\,[f(y+\lambda v)-f(y)]/\lambda,$$

where the upper limit is taken as y in X converges to x and λ in $(0,\infty)$ converges to 0. The generalized gradient of f at x, $\partial f(x)$, consists of all ζ in X^* such that

$$\langle v,\zeta\rangle\leqq f^\circ(x;v)\quad\text{for all } v \text{ in } X.$$

It follows that

$$f^\circ(x;v)=\max\{\langle v,\zeta\rangle:\zeta\in\partial f(x)\},$$

and that $\partial f(x)$ reduces to the derivative if f is continuously differentiable or to the subdifferential of convex analysis if f is convex.

1. An infinite-horizon optimal control problem. We are given a locally Lipschitz function $g\colon R^n\times R^m\to R$; i.e. whenever (x,u), (y,v) are restricted to a suitably chosen neighborhood of any given point in $R^n\times R^m$; there is a constant K such that

$$|g(x,u)-g(y,v)|\leqq K|(x,u)-(y,v)|.$$

Also on hand are matrices F and G, $n\times n$ and $n\times m$ respectively, a point x_0 in R^n, a positive number δ and a compact convex subset U of R^m.

The problem we consider is that of minimizing

$$\int_0^\infty e^{-\delta t}g(x(t),u(t))\,dt \tag{1.1}$$

over the (so-called) trajectory-control pairs (x,u) which satisfy

$$\begin{gathered}u(t)\in U,\quad \text{a.e. } t\geqq 0,\\ \dot x(t)=Fx(t)+Gu(t),\quad \text{a.e. } t\geqq 0,\\ x(0)=x_0,\end{gathered}\tag{1.2}$$

where $u(\cdot)$ need only be measurable and x absolutely continuous.

We posit the following *growth condition* on g: there are numbers $c, r\geqq 0$ such that, for every (x,u), for every ζ in the *generalized gradient* $\partial g(x,u)$ of g at (x,u), we have

$$|\zeta|\leqq c(1+|(x,u)|^r). \tag{1.3}$$

DEFINITION. We denote by $\alpha(s)$ the *infimum* in the above problem when the initial condition is $x(0)=s$ (rather than $x(0)=x_0$), and by $\lambda(F)$ the maximum of the real parts of the eigenvalues of F.

We now suppose given a trajectory-control pair (x,u) which solves the above problem (with the original initial condition $x(0)=x_0$).

THEOREM 1. *Suppose in the above that we have $\delta>(r+1)\lambda(F)$. Then the function $\alpha(\cdot)$ is locally Lipschitz, and if (x,u) is a solution to the problem, there exists an absolutely continuous function $p(\cdot)$ and a measurable function $(\zeta_1(\cdot),\zeta_2(\cdot))$ such that*:

$$-\dot p(t)=F^*p(t)-e^{-\delta t}\zeta_1(t)\quad \textit{a.e.}; \tag{1.4}$$

$$(\zeta_1(t),\zeta_2(t))\in\partial g(x(t),u(t))\quad \textit{a.e.}; \tag{1.5}$$

$$\begin{aligned}&\max\{\langle p(t),Gw\rangle-e^{-\delta t}\langle w,\zeta_2(t)\rangle: w\in U\}\\ &\quad=\langle p(t),Gu(t)\rangle-e^{-\delta t}\langle u(t),\zeta_2(t)\rangle\quad \textit{a.e.};\end{aligned}\tag{1.6}$$

$$(1.7)\qquad \int_0^\infty e^{(q-1)\delta t}|p(t)|^q\,dt<\infty,\qquad \int_0^\infty e^{(q-1)\delta t}|\dot p(t)|^q\,dt<\infty,$$

where $q>1$ is defined by $1/q+1/(r+1)=1$, provided $r>0$. If $r=0$, we have instead: $e^{\delta t}|p(t)|$ and $e^{\delta t}|\dot p(t)|$ are bounded;

$$(1.8)\qquad \lim_{t\to\infty} e^{(q-1)\delta t}|p(t)|^q=0\quad \textit{if } r>0;$$

if $r=0$, then we have instead: $e^{\delta t}p(t)$ tends to a finite limit as t goes to $+\infty$.

$$(1.9)\qquad -p(0)\in\partial\alpha(x_0).$$

Remark 1.10. (a) As shown in the proof, our hypotheses imply that the integral (1.1) is well-defined for all admissible pairs (x, u), so that no ambiguities result from our use of the word "solve".

(b) Except for the infinite interval of integration (called the case of an "infinite horizon" in economics) and the nondifferentiability of the cost functional, the above problem is a standard one in optimal control, and of the sort that arises often in mathematical economics. It is in this connection especially that the interpretation of $p(0)$ as a marginal cost ("shadow price") associated with perturbing the initial condition, as afforded by (1.9), is useful. The roles of the infinite horizon and the discount factor are discussed in T. C. Koopmans [13]; see also [16].

(c) It is worth noting that in the above version of Pontryagin's maximum principle, the necessary conditions are stated in as strong a form as one could hope; i.e. "normally", without the presence of a possibly vanishing multiplier, and with strong "transversality conditions" at infinity. Similar statements have been derived (incorrectly in many instances) by reasoning by analogy with the finite horizon case. It was H. Halkin [11], to our knowledge, who first pointed out the perhaps unexpected pathology that can arise due to the infinite horizon. Based on the results of this paper, one could say that the necessary conditions may be expected to hold in the strong form provided the "discount rate" δ is sufficiently large. A further moral is that in the infinite-horizon case, growth (or dual) conditions such as (1.7) on the adjoint variable p are more natural than pointwise transversality conditions such as (1.8) (which are simply consequences of (1.7)). This fact was foreshadowed in Theorem A.8 of [4], which can be obtained as a corollary of Theorem 1. Finally, it is well-known (and true) that the conclusions of the theorem are sufficient for (x, u) to be optimal if g is jointly convex in (x, u).

2. An example. Consider the control system on R^2 given by

$$\dot x_1=-x_2,$$
$$\dot x_2=-x_1+u,\qquad -1\leqq u\leqq 0,$$
$$x_1(0)=0,\qquad x_2(0)=0,$$

with the view of minimizing $\int_0^\infty e^{-\delta t}x_1(t)\,dt$. This is the case of the preceding section in which

$$F=\begin{bmatrix}0 & -1\\ -1 & 0\end{bmatrix},\qquad G=\begin{bmatrix}0\\ 1\end{bmatrix},\qquad U=[-1,0].$$

It follows that x_1 satisfies

$$x_1(0)=\dot x_1(0)=0,\qquad \ddot x_1(t)\geqq x_1(t),$$

so that, by a simple argument, $x_1(t) \geqq 0$ for all $t \geqq 0$. Consequently, the control $u \equiv 0$ and response $x \equiv 0$ are optimal for the problem.

The adjoint equation (1.4) becomes

$$-\dot{p}_1 = -p_2 - e^{-\delta t}, \qquad -\dot{p}_2 = -p_1$$

and hence

$$\ddot{p}_2 = p_2 + e^{-\delta t}. \tag{2.1}$$

The condition (1.6) yields the fact that $\dot{x}_2(t)$ must equal -1 if $p_2(t)$ is negative. Thus we must have $p_2(t) \geqq 0$ for all t. The solution of (2.1) is of the form

$$p_2(t) = ae^t + be^{-t} + e^{-\delta t}/(\delta^2 - 1)$$

It is easy to see that if $0 < \delta < 1$ and $p_2(t)$ converges to 0 as $t \to \infty$ (as must be the case if (1.8) is to hold for any $r \geqq 0$), then $p_2(t)$ is negative for large t. This contradication is not inconsistent with the theorem, since in this case $\lambda(F) = 1$. Of course, if $\delta > 1$, the conclusions of the theorem hold.

Let us note finally that for $\delta > 1$, we may take $r = 0$ in applying the theorem, and it follows that $p_2(t)$ must be $e^{-\delta t}/(\delta^2 - 1)$. The condition (1.9) now gives

$$(\delta/(\delta^2 - 1), \ -1/(\delta^2 - 1)) \in \partial\alpha(0, 0),$$

which makes evident the increasingly unstable dependence on the initial condition of the problem as δ nears 1.

3. Proof of Theorem 1. In this section, we denote the optimal pair $(\bar{x}, \bar{u})$.

Step 1. We denote by μ the measure on $[0, \infty)$ having density $e^{-\delta t}\,dt$. We set $p = 1 + r$ and we let X be the space $W_n^{1,p} \times L_m^p$, where L_k^p means $L^p([0, \infty), R^k)$ (with respect to μ) and where $W_n^{1,p}$ denotes the Sobolev space of absolutely continuous functions x such that x and $\dot{x}$ belong to L_n^p. (Recall that the norm of x in $W_n^{1,p}$ is given by $\|x\|_p + \|\dot{x}\|_p$.) We set $Y = L_n^p \times L_m^p \times R^n$, and we define the subset Z of Y via

$$Z = \{0\} \times K \times \{x_0\},$$

where K is the set of functions u in L_m^p satisfying

$$u(t) \in U \quad \text{a.e.}$$

We define $f: X \to R$ by

$$f(x, u) = \int_0^\infty e^{-\delta t} g(x(t), u(t))\,dt,$$

and $A: X \to Y$ by

$$A(x, u) = [\dot{x} - Fx - Gu, \ u, \ x(0)].$$

Theorem 2 of § 6 implies that f is locally Lipschitz on $L^p \times L^p$, and so all the more on $W_n^{1,p} \times L_m^p$. The optimal control problem of § 1 may be phrased as that of minimizing $f(x, u)$ over X subject to $A(x, u) \in Z$. We shall denote the given solution to this problem $(\bar{x}, \bar{u})$. Proposition 4.1 of § 4 will be available to us as soon as (4.1) is verified, and the reader may check that this is equivalent to the following result: (we shorten $\lambda(F)$ to λ).

LEMMA 3.1. *If $\delta > p\lambda$ and v belongs to L_n^p, then the solution x to*

$$\dot{x} = Fx + v, \qquad x(0) = x_0$$

belongs to $W_n^{1,p}$. In fact, $\|x\| \leqq \hat{c}\|v\|$ for some constant $\hat{c}$.

Proof. It suffices to show that the function

$$x(t)=e^{Ft}x_0+\int_0^t e^{F(t-s)}v(s)\,ds$$

belongs to L^p_n. Since $|e^{Ft}|$ is bounded by $e^{\lambda t}$, it is clear that $e^{Ft}x_0$ belongs to L^p_n, so we need only study the last term. Now, choosing any q between λ and δ/p, and limiting ourselves to the case $r>0$ (the case $r=0$ calls for straightforward modifications), we have

$$\begin{aligned}
&\int_0^\infty e^{-\delta t}\left|\int_0^t e^{F(t-s)}v(s)\,ds\right|^p dt\\
&\leqq \int_0^\infty e^{[\lambda p-\delta]t}\left[\int_0^t e^{s(q-\lambda)}\,e^{-sq}|v(s)|\,ds\right]^p dt\\
&\leqq \int_0^\infty e^{[\lambda p-\delta]t}\left[\int_0^t e^{s(q-\lambda)p/r}\,ds\right]^r\left[\int_0^t e^{-sqp}|v(s)|^p\,ds\right] dt \qquad \text{(Hölder)}\\
&\leqq \bar{c}\int_0^\infty e^{[\lambda p-\delta]t}\left[e^{t(q-\lambda)p}\right]\left[\int_0^t e^{-sqp}|v(s)|^p\,ds\right] dt\\
&=\bar{c}\int_0^\infty e^{-sqp}|v(s)|^p\,ds\int_s^\infty e^{t(qp-\delta)}\,dt \qquad \text{(Fubini)}\\
&=\bar{c}\int_0^\infty e^{-\delta s}|v(s)|^p\,ds<\infty. \qquad \text{Q.E.D.}
\end{aligned}$$

Proposition 4.1 now assures us that the function α defined in § 1 is Lipschitz in a neighborhood of x_0 (in fact, it says rather more about a function involving more arguments). Note that if (x, u) in X is such that $\tilde{A}(x, u)\in\{0\}\times K$, where $\tilde{A}(x, u)=[\dot{x}-Fx-Gu, u]$, then

$$f(x, u)\geqq\alpha(x(0)),$$

with equality when $(x, u)=(\bar{x}, \bar{u})$.

Thus $(\bar{x}, \bar{u})$ minimizes

$$\tilde{f}(x, u)=f(x, u)-\alpha(x(0))$$

over X subject to $\tilde{A}(x, u)\in\{0\}\times K$. By redefining $\tilde{f}$ away from $(\bar{x}, \bar{u})$, we can arrange to preserve this state of affairs while assuring that $\tilde{f}$ be locally Lipschitz on X. We may now apply Proposition 4.1 again to deduce that the function $\beta\colon L^p_n\times L^p_m\to R$ is Lipschitz in a neighborhood of $(0, 0)$, where (B is closed unit ball in R^n)

$$\beta(v, w)=\inf\{\tilde{f}(x, u)\colon \tilde{A}(x, u)\in\{0\}\times K+(v, w), x(0)\in x_0+B\}.$$

Step 2. Let us now choose any (x, u) in X and w in K. Upon observing that for any $\varepsilon>0$ sufficiently small

$$\tilde{A}((\bar{x}, \bar{u})+\varepsilon(x-\bar{x}, u-\bar{u}))\in\{0\}\times K+\varepsilon(\dot{x}-Fx-Gu, u-w),$$

$$(\bar{x}+\varepsilon(x-\bar{x}))(0)\in x_0+B,$$

we deduce

$$\tilde{f}((\bar{x}, \bar{u})+\varepsilon(x-\bar{x}, u-\bar{u}))\geqq\beta(\varepsilon(\dot{x}-Fx-Gu, u-w)).$$

After subtracting $\tilde{f}((\bar{x}, \bar{u}))=\beta(0, 0)$ from either side, dividing by ε and taking upper limits, we arrive at

$$(-\beta)^\circ(0,0;\dot{x}-Fx-Gu,u-w)+f^\circ(\bar{x},\bar{u};x-\bar{x},u-\bar{u})+(-\alpha)^\circ(x_0;x(0)-x_0)\geqq 0.$$

Recall [7], [9] that $f^\circ(x;v)$ equals $\max\{\langle\zeta,v\rangle:\zeta\in\partial f(x)\}$. Consequently,

$$\min_{x,u,w}\max_{\gamma,\zeta,v}\{\langle\gamma_1,\dot{x}-Fx-Gu\rangle+\langle\gamma_2,u-w\rangle+\langle\zeta_1,x-\bar{x}\rangle$$
$$+\langle\zeta_2,u-\bar{u}\rangle+\langle v,x(0)-x_0\rangle\}=0,$$

where the min is over all $(x, u, \dot{w})$ in $X\times K$ and the max over all $\gamma=(\gamma_1,\gamma_2)$, $\zeta=(\zeta_1,\zeta_2)$, and v lying in $\partial(-\beta)(0,0)$, $\partial f(\bar{x},\bar{u})$ and $\partial(-\alpha)(x_0)$ respectively. Because these sets are w^*-compact, the "lop-sided" minimax theorem [1] applies, and we deduce the existence of ζ, γ, v such that, for all (x, u) in X and w in K,

$$\text{(3.1)}\quad \langle\gamma_1,\dot{x}-Fx-Gu\rangle+\langle\gamma_2,u-w\rangle+\langle\zeta_1,x-\bar{x}\rangle+\langle\zeta_2,u-\bar{u}\rangle+\langle v,x(0)-x_0\rangle\geqq 0.$$

Step 3. Let us set $w=u=\bar{u}$ in (3.1). We then deduce

$$\langle\gamma_1,\dot{x}-\dot{\bar{x}}-F(x-\bar{x})\rangle+\langle\zeta_1,x-\bar{x}\rangle+\langle v,x(0)-x_0\rangle=0$$

for all x in $W_n^{1,p}$. Because $W_n^{1,p}$ is dense in L_n^p (this follows, for example, from the fact that $\mathcal{D}_n(0,\infty)$ is dense in L_n^p), it follows from Proposition 5.1, § 5, that ζ_1 belongs to $L_{n^*}^p$, the dual of L_n^p, rather than merely to the dual of $W_n^{1,p}$ (which is best avoided). Of course, ζ_2, γ_2 belong to $L_{m^*}^p$, γ_1 to $L_{n^*}^p$, and v to R^n. Thus we have, for all x in $W_n^{1,p}$,

$$\int_0^\infty e^{-\delta t}\gamma_1(t)\cdot(\dot{x}-Fx)\,dt+\int_0^\infty e^{-\delta t}\zeta_1(t)\cdot x(t)\,dt+v\cdot x(0)=0.$$

A classical, now familiar argument (Dubois–Reymond lemma) employing integration by parts derives from this the conclusion that γ_1 is absolutely continuous, and that

$$\frac{d}{dt}\{e^{-\delta t}\gamma_1(t)\}=-F^*\,e^{-\delta t}\gamma_1(t)+e^{-\delta t}\zeta_1(t),$$
$$\gamma_1(0)=v.$$

If we set $p(t)=e^{-\delta t}\gamma_1(t)$, then the above immediately yields (1.4), (1.7), (1.9), while (1.5) is a consequence of Theorem 2, § 6. Condition (1.8) is an elementary consequence of (1.7), the line of argument being that found in [4, Lemma A.7]. It remains to prove (1.6).

Step 4. Set $x=\bar{x}$, $w=\bar{u}$ in (3.1). Then

$$-\langle\gamma_1,G(u-\bar{u})\rangle+\langle\gamma_2,u-\bar{u}\rangle+\langle\zeta_2,u-\bar{u}\rangle=0$$

for all u in L_m^p. This implies

$$\text{(3.2)}\qquad e^{-\delta t}\gamma_2(t)=G^*p(t)-e^{-\delta t}\zeta_2(t)\quad\text{a.e.}$$

Now set $x=\bar{x}$, $u=\bar{u}$ in (3.1). We deduce, for every w in K,

$$\text{(3.3)}\qquad \langle\gamma_2,w-\bar{u}\rangle\leqq 0.$$

It follows that for almost each t, we have

$$\gamma_2(t)\cdot w\leqq\gamma_2(t)\cdot\bar{u}(t)\quad\text{for all } w \text{ in } U$$

(for if this were false we could, from the measurable selection theorem, find $w(\cdot)$ in K

contradicting (3.3)). Recalling (3.2), we arrive at

$$\{G^*p(t)-e^{-\delta t}\zeta_2(t)\}\cdot(w-\bar{u}(t))\leqq 0,$$

which is (1.6). Q.E.D.

4. A stability result called. Let X and Y be Banach spaces and $A: X\to Y$ a continuous linear operator. Consider, for s in a neighborhood of 0 in Y, the family of optimization problems consisting of minimizing a given function $F: X\to R$ subject to the constraints

$$x\in\Omega, \qquad Ax\in Z+s,$$

where the min is over all (x, u, w) in $X\times K$ and the max over all $\gamma=(\gamma_1, \gamma_2)$, $\zeta=(\zeta_1, \zeta_2)$, satisfying these constraints, and by $\alpha(s)$ the infimum in the above problem. Thus

$$\alpha(s)=\inf\{F(x): x\in C(s)\}.$$

The following, a special case of [3, Proposition 2], is stated here for the reader's convenience.

PROPOSITION 4.1. *Let Z and Ω be closed and convex.*

Suppose that there is a bounded subset K such that $C(s)\subset K$ when s is near 0, *and such that F is Lipschitz on K. Then, if the condition*

$$0\in\operatorname{int}\{A\Omega-Z\} \tag{4.1}$$

is satisfied, the function α is Lipschitz near 0.

5. A result on generalized gradients. Let X and Y be Banach spaces such that X is continuously imbedded in Y and such that X, as a subset of Y, is dense. Let f: $Y\to R$ be a function which is locally Lipschitz. It follows that its restriction to X is locally Lipschitz in the norm of X. We denote by ∂f_X the generalized gradient (see below) (in X^*) of this restriction, and by ∂f_Y the generalized gradient (in Y^*) of the function f defined on Y.

PROPOSITION 5.1. *Let x be a point of X. Then*

$$\partial f_X(x)\subset\partial f_Y(x),$$

in the sense that every ζ in $\partial f_X(x)$ admits a unique extension to an element of $\partial f_Y(x)$.

Proof. Recall [7], [9] the generalized directional derivative $f_Y^\circ(x; v)$, defined by

$$f_Y^\circ(x; v)=\limsup\{f(y+\lambda v)-f(y)\}/\lambda,$$

where the upper limit is taken as y converges to x in Y and λ decreases to 0 in R. The generalized gradient $\partial f_Y(x)$ consists by definition of those elements ζ in Y^* satisfying

$$f_Y^\circ(x; v)\geqq\langle\zeta, v\rangle \quad \text{for all } v \text{ in } Y.$$

It follows easily from density and from the fact that convergence in X implies convergence in Y that

$$f_X^\circ(x; v)\leqq f_Y^\circ(x; v)$$

whenever x and v lie in X. Now let ζ belong to $\partial f_X(x)$. Then (by definition)

$$\langle\zeta, v\rangle\leqq f_X^\circ(x, v) \quad \text{for all } v \text{ in } X,$$

and by the preceding, along with the fact that the function $v\to f_Y^\circ(x; v)$ is bounded on bounded subsets of Y, we deduce that the function

$$v\to\langle\zeta, v\rangle$$

mapping X, with its topology induced by Y, to R is bounded on bounded subsets. It follows that this linear function is continuous on X (with the induced topology) and hence, by a standard argument, admits a unique linear extension to the complete space Y in which X is dense. The extension, which we also label ζ, still satisfies

$$\langle \zeta, v\rangle \leqq f_Y^\circ(x; v)$$

for all v in Y, and hence belongs to $\partial f_Y(x)$ by definition. Q.E.D.

6. Generalized gradients on L^p. We are given a complete measure space (T, J, μ) with $\mu(T) < +\infty$, a separable Banach space X, and a function $g\colon T\times X \to R$. We assume that the mapping $t \to g(t, x)$ is measurable for each x, that $x \to g(t, x)$ is locally Lipschitz for each t, and we posit the existence of scalars $p \geqq 1$ and $c \geqq 0$ such that for every (t, x), for every element ζ of $\partial_x g(t, x)$ (generalized gradient in x for t fixed), the following bound holds:

$$|\zeta| \leqq c\{1+|x|^{p-1}\}. \tag{6.1}$$

Finally, we suppose that $t \to g(t, 0)$ is (finitely) integrable, and we set, for x in $L^p(T, X)$,

$$F(x) = \int_T g(t, x(t))\mu(dt). \tag{6.2}$$

As usual, p_* denotes the (possibly infinite) quantity satisfying

$$1/p + 1/p_* = 1.$$

THEOREM 2. *Under the above hypotheses, the function $F\colon L^p \to R$ given by* (6.2) *is well-defined* (*finitely*) *and locally Lipschitz* (*in fact, Lipschitz on bounded subsets of L^p*), *and we have*

$$\partial F(x) \subset \int_T \partial_x g(t, x(t))\mu(dt).$$

This means that corresponding to any element ζ of $\partial F(x)$, there is a function $\zeta(\cdot)$ in $L^{p_*}(T, X^*)$ satisfying

$$\zeta(t) \in \partial_x g(t, x(t)) \qquad \mu\text{-a.e.}$$

and such that, for every y in $L^p(T, X)$,

$$\langle \zeta, y\rangle = \int_T \langle \zeta(t), y(t)\rangle \mu(dt).$$

Proof. The growth condition (6.1) is easily seen to imply

$$|g(t, x)| \leqq |g(t, 0)| + c\{|x| + |x|^p\},$$

which, combined with the fact that $t \to g(t, x(t))$ is measurable, yields the fact that $F(x)$ is defined and finite whenever x lies in L^p. We now prove that F is locally Lipschitz. Invoking the mean value theorem for generalized gradients [14] we have

$$|F(x) - F(y)| = \left|\int_T \langle z(t), x(t) - y(t)\rangle \mu(dt)\right|,$$

where $z(t)$ belongs to $\partial_x g(t, w(t))$ and $w(t)$ lies between $x(t)$ and $y(t)$. Using (6.1) and

Hölder's inequality we have this last expression bounded above by

$$
\begin{aligned}
c\int_T \{1+(|x(t)|+|y(t)|)^{p-1}\}|x(t)-y(t)|\mu(dt)& \\
&\leqq c_1\|x-y\|_p + c\|(|x|+|y|)^{p-1}\|_{p_*}\|x-y\|_p \\
&\leqq \{c_1 + c(\|x\|_p+\|y\|_p)^{p/p_*}\|x-y\|_p\} \\
&\leqq K\|x-y\|_p, \quad \text{for some constant } K,
\end{aligned}
$$

as long as x and y remain in a bounded subset of L^p.

Now let ζ belong to $\partial F(x)$ (and hence to the dual of $L^p(T, X)$, $L^{p_*}(T, X^*)$ [5, § 2.6, Prop. 10 and No. 21]). Then for any v in L^p, $F^\circ(x; v) \geqq \langle \zeta, v\rangle$. Using Fatou's lemma (cf. [8, Lemma 3]) we may show

$$\int_T g_x^\circ(t, x(t); v(t))\mu(dt) \geqq F^\circ(x; v),$$

from which ensues

$$\int_T g_x^\circ(t, x(t); v(t))\mu(dt) \geqq \langle \zeta, v\rangle = \int_T \langle \zeta(t), v(t)\rangle \mu(dt). \tag{6.3}$$

For any $\varepsilon > 0$, define a multifunction Γ as follows:

$$
\begin{aligned}
\Gamma(t) &= \{0\} \quad \text{if } g_x^\circ(t, x(t); v) > \langle \zeta(t), v\rangle - \varepsilon \quad \text{for all } v \text{ in } X, \\
&= \{v \colon g_x^\circ(t, x(t); v) \leqq \langle \zeta(t), v\rangle - \varepsilon\} \quad \text{otherwise.}
\end{aligned}
$$

The map $(t, v) \to g_x^\circ(t, x(t); v)$ is of course continuous in v, and is seen to be measurable in t as a consequence of the fact that $g_x^\circ(t, x(t); v)$ can be expressed as the upper limit of a countable family of measurable functions (we use here the fact that X is separable).

It follows that the multifunction Γ is "measurable" and admits a measurable selection $v(t)$ [19, Thm. 4.1]. Now (6.3) implies that the set

$$\{t \colon \Gamma(t) \neq \{0\}\}$$

must have μ-measure 0, and since ε is arbitrary we deduce that for μ-almost all t, for all v in X,

$$g_x^\circ(t, x(t); v) \geqq \langle \zeta(t), v\rangle.$$

Consequently $\zeta(t)$ belongs to $\partial_x g(t, x(t))$ μ-a.e. Q.E.D.

Remark. The case $p = \infty$ is treated in [9]. As in that case, equality will hold in the theorem if $g(t, \cdot\,)$ is "regular" for μ-almost all t.

7. An abstract approach.

A. The line of reasoning used to prove Theorem 1 will work in much more general situations, as we now show by considering the following abstract optimization problem. We are given Banach spaces U, V, W, T, Z together with continuous linear operators $L\colon W \to V$, $\gamma\colon W \to T$, $C\colon U \to Z$, locally Lipschitz functions $f\colon U \to R$ and $g\colon T \to R$, and a multifunction E mapping U to V (i.e. $E(x)$ is a subset of V for x in U). W, the domain of L, is assumed to be a subset of U.

We consider the problem of minimizing

$$f(x) + g(\gamma x) \tag{7.1}$$

over the elements x of W which satisfy

$$Lx \in E(x), \tag{7.2}$$

$$\gamma x \in S, \tag{7.3}$$

$$Cx \in Y, \tag{7.4}$$

where S and Y are specified subsets of T and Z respectively. Here, L plays the role of a differential operator and γ that of a trace operator, so that (7.1) is a type of Bolza functional, (7.2) is a "differential inclusion", (7.3) a boundary condition, and (7.4) an explicit "state constraint". An implicit constraint is also incorporated by (7.2), which demands that x belong to the *domain* of E; i.e. the set of points x for which $E(x) \neq \emptyset$.

We posit the following conditions:

H1. The sets S, Y, and Gr (E) are closed and convex (Gr (E), the *graph* of E, is the set of points (x, v) such that v belongs to $E(x)$).

H2. ("trace property") The injection from W into U is continuous, γ has a continuous right inverse, and the kernel W_0 of γ is dense in U.

Remark 7.5. The type of setting we have described above is most familiar in partial differential equations. The point as far as we are concerned is, first, to avoid the dual W^* of W (typically nasty) and have intervene instead the dual U^* of U, and, second, to avoid the use of the transpose L^* of L and to replace it by the transpose $L_0^* \in \mathscr{L}(V^*, W_0^*)$ of the restriction $L_0 \in \mathscr{L}(W_0, V)$ of L to W_0. The motivation stems from the familiar case in which L is a differential operator and W is a space of functions or distributions; when W_0 is also the closure of the space of infinitely differentiable functions with compact support, then W_0^* is a subspace of distributions, and L_0^* is *also* a differential operator (in the sense of distributions) and can be computed, whereas the transpose L^* cannot in general be expressed in terms of differential operators.

When the trace property H2 holds, we can compare L^* and L_0^* by means of an *abstract Green formula* in the following way (see Aubin [2]). First we introduce the domain V_0^* of L_0^* defined by

$$V_0^* = \{p \in V^* : L_0^* p \in U^*\}, \tag{7.6}$$

where U^* is identified with a subspace of W_0^* (indeed, the transpose of the injection from W_0 into U is an injection from U^* to W_0^*). Equipped with the graph norm, V_0^* is a Banach space, and if (H2) holds then there exists a unique operator $\beta^* \in \mathscr{L}(V_0^*, T^*)$ such that

$$\langle L_0^* p, x\rangle - \langle p, Lx\rangle = \langle \beta^* p, \gamma x\rangle \quad \text{for all } x \text{ in } W, \quad \text{for all } p \text{ in } V_0^*. \tag{7.7}$$

(This will replace the integration by parts in the proof of Theorem 1.)

We associate to all (u, v, t, z) in $U \times V \times T \times Z$ the set $\Gamma(u, v, t, z)$ of points x in W satisfying

$$Lx \in E(x-u)+v, \qquad \gamma x \in S+t, \qquad Cx \in Y+z. \tag{7.8}$$

Notice that our optimization problem involves minimizing a functional over $\Gamma(0, 0, 0, 0)$. We make the following controllability assumption:

H3. There is a bounded subset Γ_0 of U such that, for every (u, v, t, z) sufficiently small, the set $\Gamma(u, v, t, z)$ is nonempty and contained in Γ_0. The functions f and g are Lipschitz on neighborhoods of Γ_0 and $\gamma(\Gamma_0)$ respectively.

We define the *value function* α on $U \times V \times T \times Z$ via

$$\alpha(u, v, t, z) = \inf \{f(x) + g(\gamma x) : x \in \Gamma(u, v, t, z)\},$$

and the *Hamiltonian* function H on $U \times V^*$ via

$$H(x, p) = \sup \{\langle p, v \rangle : v \in E(x)\}.$$

Since the graph of E is convex, H is concave with respect to x and convex with respect to p. We denote by $\partial_p H(x, p)$ the subdifferential in the sense of convex analysis [17] of the convex function $p \to H(x, p)$, and by $\partial_x H(x, p)$ the superdifferential of the concave function $x \to H(x, p)$.

THEOREM 3. *We posit* H1–H3. *If x in W minimizes* (7.1) *subject to the constraints* (7.2)–(7.4), *there exist $p \in V_0^*$, $\psi \in \partial f(x)$, $\phi \in \partial g(\gamma x)$, and $\tau \in N_Y(C)$ satisfying*

$$L_0^* p \in \partial_x H(x, p) - \psi - C^* \tau \subset \partial_x H(x, p) - \partial f(x) - C^* N_Y(Cx), \tag{7.9}$$

$$Lx \in \partial_p H(x, p), \tag{7.10}$$

$$\beta^* p \in \phi + N_S(\gamma x) \subset \partial g(\gamma x) + N_S(\gamma x). \tag{7.11}$$

Furthermore, the value function α is Lipschitz in a neighborhood of $(0, 0, 0, 0)$, *and we have*:

$$(L_0^* p + C^* \tau + \psi, \ -p, \ -\beta^* p + \phi, \ -\tau) \in \partial \alpha(0, 0, 0, 0). \tag{7.12}$$

Remark. We recognize (7.9)–(7.10) to be Hamiltonian equations and (7.11) a transversality condition. The notation $N_S(\gamma x)$, for example, refers to the normal cone to the convex set S at the point γx (i.e. the set of p in T^* such that $\langle p, s - \gamma x \rangle \leqq 0$ for all s in S).

B. The shadow price information (7.12) exists with respect to four perturbations in the problem. In some situations it may be deemed more natural or more convenient to consider instead a performance function α depending on fewer variables. For example, in the problem of § 1, we chose to mention only the interpretation of $p(0)$. One can easily imagine situations (e.g. sensitivity analysis) in which the generality of (7.12) would come into play in other ways.

As far as reduced versions of (7.12) is concerned, one cannot simply drop components to obtain them, since it is not generally true that the relation $(a, b) \in \partial f(x, y)$ implies $a \in \partial_x f(x, y)$ (this is discussed in [9, art. 14]). However, one can modify the proof of Theorem 3 to attain the desired result. We shall not attempt to consider the most general situation, but rather one that seems natural in many settings. We consider the case in which S is a singleton set $\{s_0\}$ and g is identically zero(i.e. boundary conditions reduce to $\gamma x = s_0$). We let $\tilde{\alpha}(s)$ be the infimum in the problem in which the boundary condition is $\gamma x = s$, and we maintain the hypotheses of the theorem.

COROLLARY 4. *If x is optimal for the above problem, there exist p, ψ and τ as in Theorem* 3 *such that* (7.9), (7.10) *hold. Further, $\tilde{\alpha}$ is Lipschitz near s_0, and*

$$-\beta^* p \in \partial \tilde{\alpha}(s_0). \tag{7.13}$$

C. We consider now a different version of the above problem, in which there is an explicit dependence on a control parameter, and unilateral constraints on the state variable. We introduce a Banach space Σ, closed convex subsets K of Σ and Ω of W, and linear operators F and G in $\mathscr{L}(U, V)$ and $\mathscr{L}(\Sigma, V)$ respectively. We now seek to minimize

$$f(x, \sigma) + g(\gamma x) \tag{7.14}$$

subject to

$$\begin{aligned}&x \in \Omega,\\ &Lx = Fx + G\sigma,\\ &\sigma \in K,\\ &\gamma x \in S,\\ &Cx \in Y,\end{aligned}$$

where $f\colon U \times \Sigma \to R$, and g, γ, L, S, C, Y are unchanged from Theorem 3. We denote by $\Gamma(u, s, v, t, z)$ the set of (x, σ) in $W \times \Sigma$ satisfying

$$\begin{aligned}&x \in \Omega + u,\\ &Lx = Fx + G\sigma + v,\\ &\sigma \in K + s,\\ &\gamma x \in S + t,\\ &Cx \in Y + z,\end{aligned}$$

and we define $\tilde{\alpha}$ on $U \times \Sigma \times V \times T \times X$ via

$$\alpha(u, s, v, t, z) = \inf \{f(x, \sigma) + g(\gamma x) : (x, \sigma) \in \Gamma(u, s, v, t, z)\}.$$

H4. There is a bounded subset Γ_0 of $W \times \Sigma$ such that, for every (u, s, v, t, z) sufficiently small, $\Gamma(u, s, v, t, z)$ is nonempty and contained in Γ_0. Furthermore, f is Lipschitz on a neighborhood of Γ_0.

COROLLARY 5. *Under the stated assumptions, if (x, σ) solves the above problem, there exist $p \in V_0^*$, (ψ_1, ψ_2) in $\partial f(x, \sigma)$, $\phi \in \partial g(\gamma x)$, and $\tau \in N_Y(Cx)$ satisfying*

$$-L_0^* p + F^* p \in \psi_1 + C^* \tau + N_\Omega(x), \tag{7.15}$$

$$G^* p \in \psi_2 + N_K(\sigma), \tag{7.16}$$

$$\beta^* p \in \phi + N_S(\gamma x), \tag{7.17}$$

$$(L_0^* p + C^* \tau + \psi_1, \psi_2, -p, -\beta^* p + \phi, -\tau) \in \partial\alpha(0, 0, 0, 0, 0). \tag{7.18}$$

Remark 7.19. As before, it is possible to replace (7.18) by alternate relations, such as (7.13) when the boundary conditions are simply $\gamma x = s_0$. The reader may wish to verify that Theorem 1 is a consequence of this corollary with the identifications (see § 3 for notation):

$$\begin{aligned}&W = \Omega = W_n^{1,p}, \qquad \Sigma = L_m^p, \qquad U = V = L_n^p,\\ &K = \{\sigma \in \Sigma : \sigma(t) \in U \quad \text{a.e.}\},\\ &f(x, \sigma) = \int_0^\infty e^{-\delta t} g(x(t), \sigma(t))\, dt,\\ &Lx = \dot{x}, \qquad \gamma x = x(0),\\ &Z = W = Y, \qquad C = \text{identity},\\ &L_0^* p = -\dot{p}, \qquad \beta^* p = p(0),\\ &V_0^* = \{p \in L_n^q : \dot{p} \in L_n^q\}.\end{aligned}$$

Of course, in applying the theorem to specific problems, there will be a need for appropriate characterizations of generalized gradients (such as Theorem 2 of § 6) and normal cones (such as the lemma in § 10). A further element will be the verification of the controllability hypothesis H3. In the problem of § 1, this was a result of taking δ large enough; in the applications of §§ 9 and 10, it follows from postulating "strictly feasible points", a concept akin to the Slater condition of mathematical programming.

8. Proof of the abstract necessary conditions. We shall merely sketch the proof, since the steps are identical to those in the proof of Theorem 1, albeit in a more general setting. We define, in the notation of § 4,

$$\begin{aligned} F(x) &= f(x) + g(\gamma x), \\ X &= W, \\ Y &= U \times V \times T \times Z, \\ Z &= \mathrm{Gr}\,(E) \times S \times Y, \\ Ax &= (x, Lx, \gamma x, Cx). \end{aligned}$$

Then H3 and Proposition 4.1 imply that the function α is Lipschitz in a neighborhood of 0. As in § 3, we denote the optimal solution $\bar{x}$, and we note that for any θ in (0, 1), for any x in W and (u, v, t, z) in $\mathrm{Gr}\,(E) \times S \times Y$, we have

$$\bar{x} + \theta(x - \bar{x}) \in \Gamma(\theta[x - u, Lx - v, \gamma x - t, Cx - z]).$$

Consequently,

$$\alpha(\theta[u, v, t, z]) \leqq F(\bar{x} + \theta(x - \bar{x})),$$

and taking generalized directional derivatives leads to

$$0 \leqq F^\circ(\bar{x}; x - \bar{x}) + \alpha^\circ(0, u - x, v - Lx, t - \gamma x, z - Cx).$$

Applying just as in § 3 the lop-sided minimax theorem, we deduce the existence of $(q, -p, r, -\tau) \in \partial\alpha(0, 0, 0, 0)$, $\psi \in \partial f(\bar{x})$, $\phi \in \partial g(\gamma \bar{x})$ such that, for all x in W and (u, v, t, z) in $\mathrm{Gr}\,(E) \times S \times Y$,

$$(8.1) \qquad 0 \leqq \langle q, u - x\rangle - \langle p, v - Lx\rangle + \langle r, t - \gamma x\rangle - \langle \tau, z - Cx\rangle + \langle \psi, x - \bar{x}\rangle + \langle \phi, \gamma x - \gamma \bar{x}\rangle.$$

If we set $(u, v, t, z) = (\bar{x}, L\bar{x}, \gamma\bar{x}, C\bar{x})$ in (8.1), we obtain that for every x in W,

$$(8.2) \qquad \langle q - \psi, \bar{x} - x\rangle - \langle p, L(\bar{x} - x)\rangle + \langle r - \phi, \gamma(\bar{x} - x)\rangle - \langle \tau, C(\bar{x} - x)\rangle = 0.$$

As x ranges over $\bar{x} + W_0$, this implies

$$q = L_0^* p + C^* \tau + \psi.$$

Now by definition, $(q, -p, r, -\tau)$ belongs to $\partial\alpha(0, 0, 0, 0)$ and hence to $U^* \times V^* \times T^* \times Z^*$, while by the result of § 5, ψ belongs to U^* (rather than merely to W^*); of course $C^*\tau$ belongs to U^*. These facts and the preceding equation imply that $L_0^* p$ belongs to U^*; i.e., that p belongs to V_0^*. Consequently we can use the Green formula (7.7) in (8.2) to deduce that for any x in W,

$$\langle r - \phi + \beta^* p, \gamma x\rangle = 0.$$

Since γ is surjective, this yields

$$r = \phi - \beta^* p,$$

and now (7.12) follows.

If in (8.1) we take $x=\bar{x}$, it follows immediately that

$$(-q,p)\in N_{\mathrm{Gr}(E)}(\bar{x},L\bar{x}),\qquad -r\in N_S(\gamma\bar{x}),\qquad \tau\in N_Y(C\bar{x}).$$

It follows from convex analysis that $(-q,p)$ lies in $N_{\mathrm{Gr}(E)}(\bar{x},L\bar{x})$ iff

$$q\in\partial_xH(\bar{x},p),\qquad L\bar{x}\in\partial_pH(\bar{x},p)$$

and all the conclusions of the theorem ensue. Q.E.D.

The proof of Corollary 4 uses the device introduced in § 3: we observe that $\bar{x}$ minimizes

$$f(x)=\tilde{\alpha}(\gamma x)$$

subject to

$$Lx\in E(x),\qquad Cx\in Y,\qquad \gamma x\in s_0+\delta B,$$

where $\delta>0$ is small, and we apply Theorem 3 to this new problem. The proof of Corollary 5 consists of applying Theorem 3 after the following relabelings:

$$U=U\times\Sigma,\qquad W=W\times\Sigma,\qquad V=V,$$

$$f(x,\sigma)=f(x),\qquad L(x,\sigma)=Lx,\qquad C(x,\sigma)=Cx,\qquad \gamma(x,\sigma)=\gamma x,$$

$$E(x,\sigma)=\begin{cases}Fx+G\sigma & \text{if } x\in\Omega \text{ and } \sigma\in K,\\ \phi & \text{otherwise.}\end{cases}$$

9. An example in partial differential equations. Let Ω be a bounded open subset of R^n whose boundary is a smooth differential manifold. We introduce functionals f and g defined by

$$f(x,\sigma)=\int_\Omega \phi(w,x(w),\sigma(w))\,dw$$

$$g(\xi)=\int_\Gamma \theta(w,\xi(w))\,dm(w),$$

where $dm(w)$ is a measure on Γ.

We consider the solutions $x\in H^2(\Omega)$ to the Dirichlet problem for the Laplacian:

$$\text{(9.1)}\qquad \begin{array}{ll}\text{i)} & -\Delta x+x=\sigma \qquad (\sigma \text{ ranges over } L^2(\Omega))\\ \text{ii)} & x|_\Gamma=0.\end{array}$$

We denote by $\partial/\partial n$ the normal derivative. Let Λ be a *compact* set and let c be a function $(w,\lambda)\in\Omega\times\Lambda\to c(w,\lambda)\in R$ satisfying

$$\text{(9.2)}\qquad \begin{array}{ll}\text{i)} & w\to c(w,\lambda) \text{ belongs to } L^2(\Omega) \text{ for each } \lambda\in\Lambda\\ \text{ii)} & \text{for almost all } w\in\Omega,\ \lambda\to c(w,\lambda) \text{ is continuous.}\end{array}$$

We consider the following problem:
Minimize

$$\int_\Omega \phi(w,x(w),\sigma(w))\,dw+\int_\Gamma \theta\left(w,\frac{\partial x}{\partial n}(w)\right)dm(w)$$

subject to the constraints (9.1) and

$$\text{(9.3)}\qquad \sigma \text{ belongs to the unit ball of } L^2(\Omega)$$

$$\text{(9.4)}\qquad \forall\lambda\in\Lambda,\qquad \int_\Omega c(w,\lambda)x(w)\,dw\geqq b(\lambda),$$

where $\lambda \to b(\lambda)$ is a continuous function. We posit the following controllability assumption.

There exist $\sigma_0 \in L^2(\Omega)$ such that $\|\sigma_0\|_{L^2(\Omega)} \leqq 1$ and such that the solution x_0 of the Dirichlet problem $-\Delta x_0 + x_0 = \sigma_0$ and $x_0|_\Gamma = 0$ satisfies:

$$\forall \lambda \in \Lambda, \qquad \int_\Omega c(w, \lambda) x_0(w)\, dw > b(\lambda). \tag{9.5}$$

Finally, we assume that the functions ϕ and θ satisfy

(9.6)
i) the functions $w \to \phi(w, x, \sigma)$ and $\theta(w, \xi)$ are measurable for each x, σ and ξ
ii) the functions $(x, \sigma) \to \phi(w, x, \sigma)$ and $\xi \to \theta(w, \xi)$ are locally Lipschitz for almost all w
iii) there exists a constant $c > 0$ such that

$$\partial\phi(w, x, \sigma) \subset c(1 + |x| + |\sigma|)B$$

$$\partial\theta(w, \xi) \subset c(1 + |\xi|)B$$

iv) the functions $w \to \phi(w, 0, 0)$ and $\theta(w, 0)$ are (finitely) integrable.

THEOREM 5. *We posit assumptions* (9.2), (9.5) *and* (9.6). *Let* $(\bar{x}, \bar{\sigma})$ *be an optimal solution. Then there exist* $p \in L^2(\Omega)$ *whose Laplacian* Δp *belongs to* $L^2(\Omega)$ *and functions* $\psi_1, \psi_2 \in L^2(\Omega)$ *satisfying*

$$(\psi_1(w), \psi_2(w)) \in \partial\phi(w, \bar{x}(w), \bar{\sigma}(w)) \quad \textit{for almost all } w \in \Omega$$

and a nonnegative Radon measure $\bar{\mu}$ *on* Λ *satisfying*

$$\int_\Lambda \left[\int_\Omega (c(w, \lambda)\bar{x}(w))\, dw - b(\lambda)\right] d\bar{\mu}(\lambda) = 0$$

such that

$$\text{i)} \quad -\Delta p + p + \psi_1 = \int_\Lambda c(\cdot, \lambda)\, d\bar{\mu}(\lambda)$$

$$\text{ii)} \quad -p|_\Gamma \in \partial\theta\left(w, \frac{\partial \bar{x}(w)}{\partial n}\right)$$

and

$$\int_\Omega \langle p(w) - \psi_2(w), \bar{\sigma}(w)\rangle = \|p - \psi_2\|_{L^2(\Omega)}.$$

Remark. Among the possible shadow price interpretations we could add to the above is the following: if $\alpha(\xi)$ is the infimum in the above problem when the boundary condition is $x|_\Gamma = \xi$ (instead of $x|_\Gamma = 0$), then α is Lipschitz on a neighborhood of 0 in $H^{3/2}(\Gamma)$ and

$$-\frac{\partial p}{\partial n} \in \partial\alpha(0).$$

Proof. We use Corollary 5 in the following case

$$U = V = \Sigma = L^2(\Omega), \qquad W = H^2(\Omega), \qquad \Omega = U,$$

$$T = H^{3/2}(\Gamma) \times H^{1/2}(\Gamma), \qquad F = 0, \qquad G = \text{identity}, \qquad K = \text{unit ball in } \Sigma,$$

$$\gamma \text{ is the operator defined by } \gamma x(w) = \left(x(w)|_\Gamma, \frac{\partial x(w)}{\partial n}\right).$$

The trace theorem [15] implies that γ is surjective and that Ker $\gamma = H_0^2(\Omega)$, the closure in $\mathscr{D}(\Omega)$ in $H^2(\Omega)$. We define L by $Lx = -\Delta x + x$. Hence L_0^* is defined by $L_0^* p = -\Delta p + p$ (in the sense of distributions) and its domain is the space $H^0(\Omega, \Delta)$ of functions $p \in L^2(\Omega)$ such that $\Delta p \in L^2(\Omega)$.

The Green formula can be written

$$(9.7) \qquad \langle -\Delta p + p, x\rangle - \langle p, -\Delta x + x\rangle = \int_\Gamma \frac{\partial}{\partial n} p(w) x(w)\, dm(w) - \int_\Gamma p(w) \frac{\partial}{\partial n} x(w)\, dm(w)$$

for smooth functions. By the abstract Green formula it still holds true when $x \in H^2(\Omega)$, $p \in H^\circ(\Omega, \Delta)$ since the operator β^* defined by

$$\beta^* p(w) = \left(\frac{\partial}{\partial n} p(w), -p(w)\right)$$

can be extended in a unique way to a continuous linear operator β^* from $H^\circ(\Omega, \Delta)$ to $H^{-3/2}(\Gamma) \times H^{-1/2}(\Gamma)$ in such a way that formula (9.7) holds. Finally, we choose $S = \{0\} \times H^{1/2}(\Gamma)$, and the (isoperimetric) constraints are defined by the Banach space $Z = C(\Lambda)$, the subset $Y = b + C_+(\Lambda)$ and the map C defined by $Cx(\lambda) = \int_\Omega c(w, \lambda) x(w)\, dw$.

We now show that the controllability assumption H4 of Corollary 5 is satisfied. Indeed, the operator $x \to (-\Delta x + x, x|_\Gamma)$ is an *isomorphism* from $H^2(\Omega)$ onto $L^2(\Omega) \times H^{3/2}(\Gamma)$. Let M_1 be the norm of its inverse, $M_2 = \sup_{\lambda \in \Lambda} \|c(\cdot, \lambda)\|_{L^2(\Omega)}$, $M_3 = \min_{\lambda \in \Lambda} [\int_\Omega c(w, \lambda) x_0(w)\, dw - b(\lambda)] > 0$. We choose $\gamma > 0$ such that $\gamma < M_3/(1 + 3M_1M_2)$. Now, if $\|v\|_{L^2(\Omega)}$, $\|s\|_{L^2(\Omega)}$, $\|z\|_{C(\Lambda)}$, $\|\xi\|_{H^{3/2}(\Gamma)}$ are less than γ, then the solutions $(x, \sigma) \in H^2(\Omega)$ of

i) $-\Delta x + x = \sigma + v$,

ii) $x|_\Gamma = \xi$,

iii) $\sigma \in K + s$,

iv) $Cx \geqq b + z$,

satisfy $\|\sigma\| < 1 + \gamma$ and $\|x\| < M_1(1 + 3\gamma)$. Further, such solutions exist; for take $\sigma = \sigma_0 + s$. Then the solution x to $-\Delta x + x = \sigma + v$ satisfies $\|x - x_0\| < 3\gamma M_1$, and hence

$$\begin{aligned} Cx &= Cx_0 - b + C(x - x_0) + b \\ &\geqq M_3 - M_2\|x - x_0\| + b \geqq M_3 - M_2 3\gamma M_1 + b \\ &> \gamma + b > b + z. \end{aligned}$$

This yields H4 in our present setting.

Now, assumptions (9.6) on the functions ϕ and θ imply that the functionals f and g are Lipschitz on bounded subsets of $L^2(\Omega)$ and $L^2(\Gamma)$ respectively (see § 6).

The assumptions of Corollary 5 are therefore satisfied. It remains to interpret the conclusions. There exist $p \in H^\circ(\Omega, \Delta)$, ψ_1 and $\psi_2 \in L^2(\Omega)$ such that, by Theorem 2,

$$(\psi_1(w), \psi_2(w)) \in \partial\phi(w, \bar{x}(w), \bar{\sigma}(w)) \quad \text{for almost all } w \in \Omega,$$

$\xi \in L^2(\Gamma)$ such that

$$\xi(w) \in \partial\theta\left(w, \frac{\partial \bar{x}}{\partial n}(w)\right) \quad \text{for almost all } w \in \Gamma$$

and a nonnegative Radon measure $\bar{\mu} \in C(\Lambda)^*$ (i.e. $-\tau$ of Corollary 5) such that

$$\int_\Lambda \left[\int_\Omega c(w,\lambda)\bar{x}(w)\,dw - b(\lambda)\right] d\bar{\mu}(\lambda) = 0.$$

They satisfy the following equations:

$$\text{i)}\quad -\Delta p + p + \psi_1 = \int_\Lambda c(\cdot,\lambda)\,d\bar{\mu}(\lambda),$$

$$\text{ii)}\quad -p|_\Gamma = \xi,$$

$$\text{iii)}\quad \int_\Omega \langle p(w) - \psi_2(w), \bar{\sigma}(w)\rangle = \|p - \psi_2\|_{L^2(\Omega)}.$$

Q.E.D.

10. Example—a variational problem with unilateral constraints. As a final example, we consider the following problem in the calculus of variations: to minimize

$$\int_0^1 g(x(t), \dot{x}(t))\,dt$$

over the absolutely continuous arcs $x: [0,1] \to R^n$ satisfying

$$x(t) \in \Omega, \tag{10.1}$$

$$\dot{x}(t) \in K \quad \text{a.e.}, \tag{10.2}$$

$$x(0) = x_0, \qquad x(1) = x_1, \tag{10.3}$$

where Ω and K are given closed convex subsets of R^n, x_0 and x_1 are given points in R^n, and $g: R^n \times R^n \to R$ is a locally Lipschitz function. We make the following assumptions: K is compact, and there is an arc $x_0(\cdot)$ joining x_0 to x_1 and an $\varepsilon > 0$ such that $\dot{x}_0(t) + \varepsilon B \subset K$ a.e. and such that $x_0(t)$ belongs to the *interior* of Ω for all t. We denote by $\alpha(y)$ the infimum in the above problem when the boundary conditions, instead of (10.3), are given by $x(0) = x_0$, $x(1) = y$.

THEOREM 6. *If $\bar{x}$ solves the above problem, then α is Lipschitz near x_1 and there exist an absolutely continuous function p, an element λ of L^∞, and a nonnegative Radon measure m on $[0,1]$ such that ($\int_t^1$ denotes integration over $[t,1]$):*

$$\lambda(t) \in N_\Omega(\bar{x}(t)) \quad m\text{-a.e.} \tag{10.4}$$

$$\left(\dot{p}(t), p(t) - \int_t^1 \lambda(s)m(ds)\right) \in \partial g(\bar{x}(t), \dot{\bar{x}}(t)) + \{0\} \times N_K(\dot{\bar{x}}(t)) \quad a.e. \tag{10.5}$$

$$p(1) \in \partial\alpha(x_1). \tag{10.6}$$

Remark. Certain closely related, but not strictly comparable results appear in the literature. The convex problem is treated by R. T. Rockafellar [18], who obtains necessary conditions couched in terms of vector measures and the Hamiltonian. Related cases are treated by J. Warga [20], H. Halkin [12], and F. H. Clarke [9, § 6]. In these, no relation of the form (10.6) is obtained.

Proof. We may suppose that $x_0=0$. Note that the solutions x of (10.2), (10.3) are bounded, so there is no loss of generality in assuming Ω compact, and in supposing that g satisfies the growth condition (1.3) of § 1 with $r=0$. We denote $\mathscr{A}_0^1$ the set of absolutely continuous $x:[0,1]\to R^n$ such that $x(0)=0$ and $\dot{x}$ belongs to L^1. If the norm of x is taken as $\int_0^1 |\dot{x}|\,dt$, then the dual of $\mathscr{A}_0^1$ may be identified with L^∞, with the duality pairing

$$\langle x, v\rangle = \int_0^1 \dot{x}v\,dt.$$

We shall apply Corollary 4 of § 7 with the following identifications:

$$\begin{aligned}
x &= (x, y),\\
U &= L^1\times L^1,\\
W &= \mathscr{A}_0^1\times L^1,\\
f(x, y) &= \int_0^1 g(x, y)\,dt,\\
Z &= \mathscr{A}_0^1,\\
Y &= \{z\in Z : z(t)\in\Omega\},\\
C(x, y) &= \int_0^t y(s)\,ds,\\
\gamma(x, y) &= x(1),\\
T &= R^n,\\
S &= \{x_1\},\\
L(x, y) &= \dot{x}-y,\\
V &= L^1,\\
E(x, y) &= \begin{cases}\{0\} & \text{if } y\in K_1,\\ \phi & \text{otherwise,}\end{cases}
\end{aligned}$$

where K_1 is the set of y in L^1 satisfying $y(t)\in K$, a.e. It is routine to check that $(\bar{x}, \bar{y})=(\bar{x}, \dot{\bar{x}})$ then solves the abstract problem of § 7, and to verify the hypotheses of the theorem. The controllability hypothesis follows mainly (much as in § 9) from the observation that if (u, v, t, z) is sufficiently small in $L^1\times L^1\times R^n\times Z$, then the element (x, y) of W defined by

$$y(t) = \dot{x}_0(t)+u(t)+t-\int_0^1 (u+v)\,dt,$$

$$\dot{x}(t) = y+v$$

satisfies

$$\begin{aligned}
y &\in K_1+u,\\
x(1) &= x_1+t,\\
C(x, y) &\in \Omega+z.
\end{aligned}$$

We calculate

$$\begin{aligned} V^* &= L^\infty, \\ L_0^* p &= [-\dot{p}, -p] \quad \text{(in the sense of distributions)} \\ V_0^* &= \{p \in L^\infty : \dot{p} \in L^\infty\}, \\ C^* v &= (0, v), \\ \beta^* p &= -p(1), \\ H(x, y, p) &= \begin{cases} 0 & \text{if } y \in K_1, \\ -\infty & \text{otherwise.} \end{cases} \end{aligned}$$

A straightforward interpretation of (7.9) (with the help of Theorem 2, § 6) then yields the existence of an arc p and elements τ, σ of $N_Y(C(\bar{x}, \bar{y}))$ and $N_{K_1}(\bar{y})$ such that

$$(\dot{p}, p) \in \partial g(\bar{x}, \dot{\bar{x}}) + (0, \tau + \sigma), \qquad p(1) \in \partial \alpha(x_1).$$

Now, σ is a function in L^∞ satisfying

$$\int_0^1 \sigma(t) \cdot \{k(t) - \dot{\bar{x}}(t)\}\, dt \leqq 0$$

whenever k in L^1 is such that $k(t) \in K$ a.e. It follows readily from the measurable selection theorem [19, Theorem 4.1] that for almost every t, for every k in K,

$$\sigma(t) \cdot \{k - \dot{\bar{x}}(t)\} \leqq 0,$$

i.e. $\sigma(t) \in N_K(\dot{\bar{x}}(t))$. It remains now to characterize τ. We have by definition that for any arc x satisfying $x(0) = 0$, $x(t) \in \Omega$,

$$\int_0^1 \tau(t) \cdot \{\dot{x}(t) - \dot{\bar{x}}(t)\}\, dt \leqq 0,$$

where, as explained earlier, τ is identified with an element of L^∞. The lemma below then yields an expression for τ that concludes the proof. Q.E.D.

We denote $\mathscr{A}^\infty$ the set of absolutely continuous functions $x : [0, 1] \to R^n$ with derivative $\dot{x}$ in L^∞; p is a scalar $\geqq 1$.

LEMMA 10.1. *Let Ω be a compact convex set in R^n containing 0 in its interior, and let τ belong to $L^p([0, 1]; R^n)$. If $\bar{x}$ in $\mathscr{A}^\infty$ solves the problem of maximizing*

$$\int_0^1 \tau(s) \cdot \dot{x}(s)\, ds$$

over the elements x of $\mathscr{A}^\infty$ satisfying $x(0) = 0$, $x(t) \in \Omega$ for all t, then there is a function λ in L^∞ and a nonnegative Radon measure m on $[0, 1]$ such that

$$\lambda(t) \in N_\Omega(\bar{x}(t)) \qquad m\text{-}a.e. \tag{10.7}$$

$$\tau(t) = \int_{[t,1]} \lambda(s) m(ds) \quad a.e. \tag{10.8}$$

Proof. Let h denote the support function of Ω; i.e.

$$h(p) = \max\{p \cdot w : w \in \Omega\}.$$

If S denotes the unit sphere of R^n, then x belongs to Ω iff $g(x) \leqq 0$, where

$$g(x) = \max\{p \cdot x - h(p) : p \in S\}.$$

The function g is locally Lipschitz, and it follows from [6, Theorem 2.1] that when $g(x)=0$, $\partial g(x) \subset N_\Omega(x)$. Further, because $0 \in \operatorname{int} \Omega$ we have $h(p) \geqq \delta > 0$ when $p \in S$, and this implies $0 \notin \partial g(x)$ when $g(x)=0$. The statement of the lemma implies then that $\bar{x}$ minimizes

$$\int_0^1 -\tau \cdot \dot{x}\, dt$$

subject to $x(0)=0$, $g(x(t)) \leqq 0$. We may apply [9, Thm. 4] to deduce the existence of an arc p, a function λ in L^∞ and a Radon measure m supported on the set $\{t\colon g(\bar{x}(t))=0\}$ such that

$$(\dot{p}, p) + \left(0, \int_{(0,t)} \lambda\, dm\right) = (0, -\tau), \qquad \lambda(t) \in \partial g(\bar{x}(t)) \quad m\text{-a.e.}$$

Because $x(1)$ is free, the transversality condition

$$-p(1) = \int_{(0,1]} \lambda\, dm$$

also pertains. Thus p is identically equal to this last quantity, and the result follows. Q.E.D.

REFERENCES

[1] J. P. AUBIN, *Théorème du minimax pour une classe de fonctions*, C. R. Acad. Sci. Paris Sér. A, 274 (1972), pp. 455–458.

[2] ———, *Approximation of Elliptic Boundary-Value Problems*, Wiley-Interscience, New York, 1972.

[3] J. P. AUBIN AND F. H. CLARKE, *Multiplicateurs de Lagrange en optimisation non convexe et applications*, C. R. Acad. Sci. Paris, 285 (1977), pp. 451–454.

[4] A. BENSOUSSAN, E. G. HURST JR. AND B. NASLUND, *Management Applications of Modern Control Theory*, North-Holland, American Elsevier, Amsterdam, 1974.

[5] N. BOURBAKI, *Integration*, Fascicule XXV, Chapitre VI, Hermann, Paris.

[6] F. H. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.

[7] ———, *A new approach to Lagrange multipliers*, Math. Oper. Res., 1 (1976), pp. 165–174.

[8] ———, *Multiple integrals of Lipschitz functions in the calculus of variations*, Proc. Amer. Math. Soc. 64 (1977), pp. 260–264.

[9] ———, *Generalized gradients of Lipschitz functionals*, Tech. Report 1687, Math. Res. Center, Madison, Univ. of Wisconsin 1976, Advances in Mathematics, to appear.

[10] N. CHRISTOPEIT, *Necessary optimality conditions with application to a variational problem*, this Journal, 15 (1977), pp. 683–698.

[11] H. HALKIN, *Necessary conditions for optimal control problems with infinite horizons*, Econometrica, 42 (1974), pp. 267–272.

[12] ———, *Optimization without differentiability*, Proceedings of the Conference on Optimal Control Theory, Canberra, August 1977, Springer-Verlag, New York, to appear.

[13] T. C. KOOPMANS, *Concepts of optimality and their uses*, Nobel Lectures 1975; reproduced in Amer. Econ. Rev., 67 (1977), pp. 261–274.

[14] G. LEBOURG, *Valeur moyenne pour gradient généralisé*, C. R. Acad. Sci. Paris Sér. A, 281 (1975), pp. 795–797.

[15] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes et applications*, Dunod, Paris (1968).

[16] R. PALLU DE LA BARRIÈRE, *On the cost of constraints in dynamical optimization*, Mathematical Theory of Control, A. V. Balakrishnan, L. W. Neustadt, eds., Academic Press, New York, 1967.

[17] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, N.J., 1970.

[18] ———, *State constraints in convex control problems of Bolza*, this Journal, 10 (1972), pp. 691–715.

[19] D. H. WAGNER, *Survey of measurable selection theorems*, this Journal, 15 (1977), pp. 859–903.

[20] J. WARGA, *Controllability and necessary conditions in unilateral problems without differentiability assumptions*, this Journal, 14 (1976), pp. 546–572.

[21] J. L. LIONS, *Contrôle optimal des systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.

[22] W. W. HAGER AND S. K. MITTER, *Lagrange duality theory for convex control problems*, this Journal, 14 (1976), pp. 843–856.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 5, September 1979

CONTROL OF A PSEUDO-PARABOLIC INITIAL-VALUE PROBLEM TO A TARGET FUNCTION*

L. W. WHITE†

Abstract. Let G be a bounded domain in R^n with a smooth boundary and let $Q = G \times (0, T]$. We consider the solution $y(u)$ of the pseudo-parabolic initial-value problem

$$M(x)y_t(u) + L(x)y(u) = u \quad \text{in } L^2(Q),$$
$$y(\cdot, 0; u) = 0 \quad \text{in } L^2(G)$$

to be the state corresponding to the control u. Here $M(x)$ and $L(x)$ are second order symmetric uniformly strongly elliptic operators on G. The control problem is to find a control u_0 in a given ball in $L^2(Q)$ such that, for a given Z, the trace $y(\cdot, T; u) = Z(\cdot)$ is in $L^2(G)$ and such that u_0 minimizes a certain noncoercive energy functional arising naturally from the differential equation. In this paper we give controllability results for the pseudo-parabolic initial-value problem and regularity results for u_0. Furthermore, we establish results that u_0 lies on the surface of the constraint ball in $L^2(Q)$ and that the optimal controls of similar problems that steer to balls centered at Z converge to u_0 in $L^2(Q)$ as the target radii shrink to zero. The regularity results indicate that convergence in $L^2(Q)$ is as strong as we may expect. Finally, we include a simple example to illustrate some of our results.

1. Introduction. In this paper we consider the distributed control of a system whose state is given by the solution of a pseudo-parabolic initial-value problem. The optimal control problem is to find a control u_0 that minimizes a certain noncoercive energy functional over a set of controls that are contained in a fixed ball in a Hilbert space and that steer their state functions to a given target function Z.

Problems of pseudo-parabolic type arise in many contexts in which one imposes higher order correction in the physical model. For example, in [2] a two temperature theory of thermodynamics is considered in which a quantity θ, the conductive temperature, satisfies an equation of pseudo-parabolic type. Also, in [8] it is shown that pseudo-parabolic equations arise in the study of the flow of second-order fluids. We refer to [1] for an extensive bibliography concerning pseudo-parabolic problems and their applications. Thus far, control of problems of pseudo-parabolic type have received little attention in the literature. In view of their importance in the modeling of physical systems and their close relationship to parabolic problems [7], it seems worthwhile to study control problems of this type.

The ability to steer precisely to a prescribed target function in a given time T is a controllability property. The primary difficulty in the formulation of this problem is one of showing the admissible set of controls U_{ad} to be nonempty since, given this result, the existence and uniqueness of the optimal control is standard [3]. While it is clear that there are targets Z for which the admissible set is empty, the set of attainability can be given explicitly for the pseudo-parabolic initial-value problem. This differs from the situation that occurs for analogous parabolic problems where no such explicit description can be given. In § 2 we formulate the problem and give the controllability result.

The regularity properties of the optimal control are deduced as a result of the strong endpoint condition and the representation of the solution of the pseudo-parabolic problem. This is done in § 3.

In § 4 we study the relationship of u_0 to the optimal controls in [9]. In [9] we consider this problem for targets that are balls of radius $\rho > 0$ centered at Z. We show here that the optimal controls u_ρ associated with these problems converge to u_0 in the

* Received by the editors May 24, 1978, and in revised form February 5, 1979.

† Department of Mathematics, University of Oklahoma, Norman, Oklahoma 73019.

L^2 sense as ρ goes to zero. This result depends on the property that u_0 lies on the surface of the control-constraint ball. The proof we give relies upon a more "relaxed" formulation of our problem with a more general set of admissible controls. This is done so that the set of attainability is a normal space [4] and that the Lagrange multiplier associated with the terminal constraint is a member of a space of distributions. Finally, in § 5, we present a simple example to illustrate some of our results.

2. Formulation of the problem. Let G be a bounded domain in R^n and let $Q = G\times(0, T]$ where T is fixed finite. For ease we assume that the boundary of G is of class C^∞. We shall be concerned with several function spaces. Let $L^2(Q)$ and $L^2(G)$ denote the spaces of equivalence classes of square integrable functions on Q and G. The norms on these spaces we denote by $\|\cdot\|_{0,0}$ and $\|\cdot\|_0$ and inner products by $(\cdot,\cdot)_{0,0}$ and $(\cdot,\cdot)_0$, respectively. We use the standard notation $H^k(G)$ to denote the kth order Sobolev space on G with norm $\|\cdot\|_k$ and $H_0^k(G)$ to denote the completion with respect to $\|\cdot\|_k$ of the space of infinitely differentiable functions with compact support in G. The space $H_0^1(G)\cap H^2(G)$ is a closed subspace of $H^2(G)$ and consequently is a Hilbert space with respect to $\|\cdot\|_2$. The spaces $L^2(0, T; H_0^1(G)\cap H^2(G))$ and $L^2(0, T; H_0^1(G))$ are Hilbert spaces [3], [4] with norms $\|f\|_{2,0}=(\int_0^T\|f(\cdot,t)\|_2^2\,dt)^{1/2}$ and $\|f\|_{1,0}=(\int_0^T\|f(\cdot,t)\|_1^2\,dt)^{1/2}$, respectively. We also note that $L^2(0, T; H^{-1}(G))$ is the dual of $L^2(0, T; H_0^1(G))$.

We consider the solution $y(u)=y(x, t; u)$ in $L^2(0, T; H_0^1(G)\cap H^2(G))$ of the initial-value problem

$$(1)\qquad \begin{aligned} &M(x)y_t(u)+L(x)y(u)=u \quad \text{in } L^2(Q),\\ &y(\cdot, 0; u)=0 \quad \text{in } L^2(G)\end{aligned}$$

to be the state corresponding to u. Here $M(x)$ and $L(x)$ are second order symmetric uniformly strongly elliptic operators

$$M(x)\equiv-\sum_{i,j=1}^n \frac{\partial}{\partial x_i} m_{ij}(x)\frac{\partial}{\partial x_j}+m(x)$$

and

$$L(x)\equiv-\sum_{i,j=1}^n \frac{\partial}{\partial x_i} l_{ij}(x)\frac{\partial}{\partial x_j}+l(x)$$

in G. We assume for simplicity that m_{ij}, l_{ij}, m, and l belong to $C^\infty(\bar{G})$. Hence, M is an isomorphism mapping $H_0^1(G)\cap H^2(G)$ onto $L^2(G)$ so that $M^{-1}L$ is a bounded mapping of $H_0^1(G)\cap H^2(G)$ onto itself. For each t the exponential function $t\mapsto E(t)=e^{-tM^{-1}L}$ then is a bounded linear map of $H_0^1(G)\cap H^2(G)$ onto itself. The solution of (1) may be represented by

$$(2)\qquad y(\cdot, t; u)=\int_0^t E(t-s)M^{-1}u(\cdot, s)\,ds$$

in $L^2(G)$ where the integral is in the sense of Bochner. From this formula it is clear that $y(\cdot, t; u)$ and $y_t(\cdot, t; u)$ are in $H_0^1(G)\cap H^2(G)$ for almost all t in $[0, t]$ and that $y(u)$ belongs to $L^2(Q)$. For further facts concerning the solution of (1), we refer the reader to [7].

The following observation is basic to our discussion.

PROPOSITION 1. *Let $Z\in H_0^1(G)\cap H^2(G)$. Then the control $u_z(\cdot, t)=(1/T)ME(t-T)Z(\cdot)$ has the property that $y(\cdot, T; u_z)=Z(\cdot)$ in $L^2(G)$.*

Proof. The proof follows immediately by substituting $u_z(\cdot, s)$ for $u(\cdot, s)$ in (2).

Remark 2. The essential fact here is that $M^{-1}L$ is a bounded map on $H_0^1(G) \cap H^2(G)$, so that $E(\cdot)$ is a group of operators. Consequently, $E(\cdot)$ is meaningful for negative values of t, so that the term $E(t-T)$ has meaning. This is not the case for parabolic problems where semi-groups are involved in the solution.

Define the set of attainability at time T

$$Y(T) = \{y(\cdot, T; u) : u \in L^2(Q)\}.$$

From (2) and Proposition 1, the following facts are clear.

COROLLARY 3.

$$Y(T) = H_0^1(G) \cap H^2(G).$$

COROLLARY 4. *The pseudo-parabolic problem is controllable in the sense that $Y(T)$ is dense in $L^2(G)$.*

Thus, the set

$$V(Z, \rho) = \{u \in L^2(Q) : \|y(\cdot, T; u) - Z(\cdot)\|_0 \leqq \rho\}$$

is nonempty for $\rho \geqq 0$ if $Z \in H_0^1(G) \cap H^2(G)$ and is nonempty for $\rho > 0$ if $Z \in L^2(G)$.

We now state the control problem. We assume that Z is in $H_0^1(G) \cap H^2(G)$ and is nonzero until stated otherwise.

$$\text{(3)} \qquad \begin{aligned} &\text{minimize } K(u) = (u, y(u))_{0,0} \text{ subject to } u \in L^2(Q) \\ &\|u\|_{0,0} \leqq b, \\ &\|y(\cdot, T; u) - Z(\cdot)\|_0 \leqq \rho. \end{aligned}$$

Remark 5. The control functional $K(\cdot)$ may be represented in the form of an energy functional by using (1) and integrating by parts

$$\text{(4)} \qquad \begin{aligned} K(u) = &\int_G \left[\sum_{i,j=1}^n y_{x_i}(x, T; u) m_{ij}(x) y_{x_j}(x, T; u) + m(x) y^2(x, T; u) \right] dx \\ &+ \int_Q \left[\sum_{i,j=1}^n y_{x_i}(x, t; u) l_{ij}(x) y_{x_j}(x, t; u) + l(x) y^2(x, t; u) \right] dx\, dt \end{aligned}$$

and is strictly convex, weakly lower semicontinuous, and noncoercive in u.

Remark 6. For convenience we require the parameter b to satisfy $\|u_z\|_{0,0} \leqq b$. The parameter ρ is to belong to $[0, \|Z\|_0)$. These conditions are sufficient to guarantee that the set

$$\text{(5)} \qquad U_{\text{ad}}(\rho) = \{u \in V(Z, \rho) : \|u\|_{0,0} \leqq b\}$$

is a nonempty closed bounded convex subset of $L^2(Q)$.

PROPOSITION 7. *For each ρ in $[0, \|Z\|_0)$, problem* (3) *has a unique solution that we denote by u_ρ.*

3. Regularity of the optimal control for $\rho = 0$. For this case the optimal control u_0 must satisfy $y(\cdot, T; u_0) = Z(\cdot)$ in $L^2(G)$. Hence, we may deduce regularity properties directly from the equation

$$\text{(6)} \qquad Z(\cdot) = \int_0^T E(T-s) M^{-1} u_0(\cdot, s)\, ds.$$

From (6) we see that if $Z \in H_0^1(G) \cap H^2(G)$, then $u_0 \in L^2(Q)$. Furthermore, since $M^{-1}L$

is a bounded one-to-one map of $H_0^1(G) \cap H^p(G)$ onto itself for integers $p \geqq 2$ [7], we see that if $Z \in H_0^1(G) \cap H^k(G)$ for $k \geqq 2$ then u_0 belongs to the space $L^2(0, T; H^{k-2}(G))$.

THEOREM 8. *If $Z \in H_0^1(G) \cap H^k(G)$ where $k \geqq 2$, then, for the problem in which $\rho = 0$, the optimal control u_0 belongs to $L^2(0, T; H^{k-2}(G))$. This result is in contrast to the regularity result obtained in* [9] *for the case $\rho > 0$.*

THEOREM 9. *Let $Z \in H^k(G)$ for $k \geqq 0$ and let $\rho > 0$. Then the optimal control u_ρ belongs to the space $H^1(0, T; H_0^1(G) \cap H^{k+2}(G))$.*

Remark 10. The assumption that $\rho > 0$ is essential to the proof of Theorem 9 given in [9]. This condition is shown to imply the existence of positive Lagrange multipliers associated with the inequality constraints in problem (3). The existence of these numbers allows us to express the optimal control as a linear combination of solutions to pseudo-parabolic and adjoint pseudo-parabolic problems to deduce Theorem 9. Furthermore, the existence of these numbers implies

$$\|y(\cdot, T; u_\rho) - Z(\cdot)\|_0 = \rho \quad \text{and} \quad \|u_\rho\|_{0,0} = b,$$

see [9].

4. Convergence of u_ρ to u_0.

In this section we establish the following result.

THEOREM 11. *Let $\{\rho_i\}_{i=1}^\infty$ be a sequence of positive real numbers such that $\rho_i \to 0$ as $i \to \infty$. Then the corresponding sequence of optimal controls $\{u_{\rho_i}\}_{i=1}^\infty$ is such that $u_{\rho_i} \to u_0$ strongly in $L^2(Q)$ as $i \to \infty$.*

We note first that since $\|u_{\rho_i}\|_{0,0} \leqq b$ for each $i = 1, 2, \cdots$, there is a weakly convergent subsequence, which we again denote by $\{u_{\rho_i}\}_{i=1}^\infty$, such that $u_{\rho_i} \to u$ weakly in $L^2(Q)$ as $i \to \infty$. This u has the property $\|u\|_{0,0} \leqq b$. Furthermore, since we have $\|y(\cdot, T; u_{\rho_i}) - Z(\cdot)\|_0 \leqq \rho_i$ with $\rho_i \to 0$ then $y(\cdot, T; u) = Z(\cdot)$. Hence, we see that $u \in U_{\text{ad}}(0)$. From the weak lower semicontinuity of $K(\cdot)$, we have

$$K(u) \leqq \lim K(u_{\rho_i}).$$

But from (5) it is clear that $U_{\text{ad}}(0) \subset U_{\text{ad}}(\rho)$ for $\rho > 0$. Thus, we note that $K(u_0) \geqq K(u_\rho)$ for $\rho > 0$. Therefore, we have $K(u) \leqq K(u_0)$, and since $u \in U_{\text{ad}}(0)$ and u_0 is unique, we must have $u = u_0$. The above argument holds for any sequence of ρ's converging to zero so we have quite easily shown the following.

PROPOSITION 12. *The optimal controls u_ρ converge weakly to u_0 in $L^2(Q)$ as $\rho \to 0$.*

Remark 13. We note that by using arguments above and setting

$$K(u) = (u, y(u))_{0,0} = \|u\|^2,$$

we may observe from the weak convergence in Proposition 12 that

$$\|u_{\rho_j} - u_0\|^2 = \|u_{\rho_j}\|^2 - (u_0, y(u_{\rho_j}))_{0,0} - (u_{\rho_j}, y(u_0))_{0,0} + \|u_0\|_{0,0}^2$$

converges to zero since $u_{\rho_j} \to u_0$ weakly in $L^2(Q)$ implies $y(u_{\rho_j}) \to y(u_0)$ weakly in $L^2(Q)$. However, we point out that $\|u\|$ is not coercive, that is, it is not true that $\|u\| \geqq c\|u\|_{0,0}$ for some positive constant c. This fact is illustrated in the example in § 5.

To prove $L^2(Q)$ convergence we must work with the $L^2(Q)$ norm. At this point, if we know that $\|u_0\|_{0,0} = b$ (as we do for u_ρ with $\rho > 0$), then we can show strong convergence by the well-known argument

$$\|u_{\rho_i} - u_0\|_{0,0}^2 = \|u_{\rho_i}\|_{0,0}^2 - 2(u_0, u_{\rho_i})_{0,0} + \|u_0\|_{0,0}^2 = 2(b^2 - (u_0, u_{\rho_i})_{0,0}) \to 0$$

as $i \to 0$. Hence, we seek to show that $\|u_0\|_{0,0} = b$.

In this case, however, we are not able to deduce the existence of positive Lagrange multiplier numbers. As we have mentioned previously, if $Z \in H_0^1(G) \cap H^2(G)$, then we know there is a unique solution u_0 to the problem

$$\begin{aligned}&\text{minimize } K(u) \text{ subject to } u \in L^2(Q)\\ (7)\qquad &\|u\|_{0,0} \leqq b\\ &y(\cdot, T; u) = Z(\cdot) \quad \text{in } L^2(G).\end{aligned}$$

Furthermore, if u_0 satisfies $\|u_0\|_{0,0} < b$, then, since $K(\cdot)$ is a convex functional, u_0 solves the problem

$$\begin{aligned}(8)\qquad &\text{minimize } K(u) \text{ subject to } u \in L^2(Q)\\ &y(\cdot, T; u) = Z(\cdot) \quad \text{in } L^2(G).\end{aligned}$$

Our plan is to show there is no solution to problem (8) in $L^2(Q)$.

We consider a more general problem than (8) by enlarging or "relaxing" the admissible set so that $y(\cdot, T; u)$ belongs to a normal space. Let $V = L^2(0, T; H_0^1(G))$ and $V' = L^2(0, T; H^{-1}(G))$ be the dual of V, [4]. We reformulate our problem with control space V', [6]. Hence, we have

$$\begin{aligned}(1)'\qquad &My_t(u) + Ly(u) = u \quad \text{in } V'\\ &y(\cdot, 0; u) = 0 \quad \text{in } L^2(G).\end{aligned}$$

The solution of (1)′ may again be represented in terms of a Bochner integral by

$$(2)'\qquad y(\cdot, t; u) = \int_0^t E(t-s)M^{-1}u(\cdot, s)\, ds$$

where $E(t) \equiv \exp(-tM^{-1}L)$ is now an isomorphism of $H_0^1(G)$ onto itself for each t. Hence, $y(\cdot, T; u)$ belongs to $H_0^1(G)$, and the map $u \to y(\cdot, T; u)$ is a continuous map of V' into $H_0^1(G)$. Furthermore, we have a result analogous to Proposition 1.

PROPOSITION 14. *The map $u \to y(\cdot, T; u)$ for a fixed positive T maps V' onto $H_0^1(G)$.*

The energy functional $K(\cdot)$ has the same representation as that in (4) all the derivatives being in $L^2(Q)$. The inner product representation, however, becomes

$$\begin{aligned}K(u) &= \int_0^T \langle u(\cdot, t), y(\cdot, t; u)\rangle\, dt\\ &= (\langle u, y(u)\rangle)\end{aligned}$$

where $\langle \cdot, \cdot \rangle$ denotes the $H_0^1(G) - H^{-1}(G)$ dual pairing which extends the $L^2(G)$ inner product $(\cdot, \cdot)_0$.

Recalling that $Z \in H_0^1(G) \cap H^2(G) \subset H_0^1(G)$ so that the endpoint constraint is compatible, we now consider the problem

$$\begin{aligned}(8)'\qquad &\text{minimize } K(u) \text{ subject to } u \in V'\\ &y(\cdot, T; u) = Z(\cdot) \quad \text{in } L^2(G).\end{aligned}$$

Define the sets

$$V(Z) = \{u \in L^2(Q) : y(\cdot, T; u) = Z(\cdot) \text{ in } L^2(G)\}$$

and

$$V'(Z) = \{u \in V' : y(\cdot, T; u) = Z(\cdot) \text{ in } L^2(G)\}.$$

Obviously, it is true that $V(Z) \subset V'(Z)$. Furthermore, for our problem Z in $H_0^1(G) \cap$

$H^2(G)$ implies u must be in $L^2(Q)$. Hence, we conclude that $V(Z)=V'(Z)$, and we have

$$d=\underset{u\in V(Z)}{\text{infimum}}K(u)=\underset{u\in V'(Z)}{\text{infimum}}K(u).$$

Remark 15. If there exists $u_{\text{op}}\in V(Z)$ such that $K(u_{\text{op}})=d$, then since $V(Z)=V'(Z)$, we see that u_{op} is a solution of problems (8) and (8)$'$.

We show now that problem (8)$'$ has no solution in V'. To this end, we assume the contrary that there exists a solution u_{op} to (8)$'$. We first demonstrate the existence of a Lagrange multiplier ξ in $H^{-1}(G)$ associated with the equality constraint in (8)$'$. We refer to [5] for a discussion of Lagrange multipliers and equality constraints. Define the function H from V' onto $H_0^1(G)$ by

$$H(u)\equiv y(\cdot,T;u)-Z(\cdot) \tag{9}$$

and note that by Proposition 14 the Fréchet derivative of H at u_{op}

$$H'(u_{\text{op}})(v)=y(\cdot,T;v) \tag{10}$$

maps V' onto $H_0^1(G)$ so that u_{op} is a regular point of $H(\cdot)$.

Remark 16. It is worth mentioning at this point that there are two reasons for using the problem formulated in (8)$'$: i) the above regular point property of u_{op} and ii) the Lagrange multiplier ξ is in $H^{-1}(G)$ the dual of $H_0^1(G)$. This second reason is important because $H^{-1}(G)$ is a space of distributions, and we use ξ to formulate an adjoint pseudo-parabolic problem. If we use the formulation in problem (8), then u_{op} is still a regular point when we consider that $v\to y(\cdot,T;v)$ maps onto $H_0^1(G)\cap H^2(G)$. Here, however, all that can be said is that ξ is in the dual of $H_0^1(G)\cap H^2(G)$. Since this is not a normal space, ξ is not a distribution [4], and it is difficult to give meaning to the adjoint problems. Thus, by using (8)$'$ we have not really changed the problem but we have relaxed conditions on the controls so that the constraint is in a space with a nice dual.

Defining the functional on V'

$$\Lambda(u)\equiv K(u)+\langle\xi,H(u)\rangle, \tag{11}$$

we have the identity

$$0=\Lambda'(u_{\text{op}}(v))=(\langle u_{\text{op}},y(v)\rangle)+(\langle v,y(u_{\text{op}})\rangle)+\langle\xi,y(\cdot,T;v)\rangle \tag{12}$$

for all v in V', cf. [9].

We now introduce the following adjoint initial-value problems

$$\begin{aligned}&-Mp_t(u_{\text{op}})+Lp(u_{\text{op}})=u_{\text{op}}\quad\text{in } V'\\&p(\cdot,T;u_{\text{op}})=0\quad\text{in } L^2(G)\end{aligned} \tag{13}$$

and

$$\begin{aligned}&-Mq_t+Lq=0\quad\text{in } V'\\&q(\cdot,T)=M^{-1}\xi(\cdot)\quad\text{in } H_0^1(G).\end{aligned} \tag{14}$$

Equation (12) now becomes

$$(\langle v,y(u_{\text{op}})+p(u_{\text{op}})+q\rangle)=0$$

for all v in V'. Hence, we have

$$y(u_{\text{op}})+p(u_{\text{op}})+q=0 \tag{15}$$

in V and thus in $L^2(Q)$. Differentiating (15) with respect to t and using (1)$'$, (13) and (14) we see that

$$-M^{-1}Ly(u_{\text{op}})+M^{-1}Lp(u_{\text{op}})+M^{-1}Lq=0 \tag{16}$$

in V. Furthermore, we note that

$$M^{-1}Ly(u_{\text{op}})+M^{-1}Lp(u_{\text{op}})+M^{-1}Lq=0 \tag{17}$$

in V. Subtracting (16) from (17), we have

$$M^{-1}Ly(u_{\text{op}})=0 \tag{18}$$

in V. But, $M^{-1}L$ is an isomorphism on V so $y(u_{\text{op}})=0$ in V and $u_{\text{op}}=0$ in V'. However, the element 0 is not an admissible control since Z is nonzero. Hence, we conclude that (8)′ has no solution. By Remark 15 then (8) has no solution $L^2(Q)$. We have thus proved the following lemma.

LEMMA 17. *The solution* u_0 *of problem* (7) *satisfies* $\|u_0\|_{0,0}=b$.

Thus, by our discussion at the first of this section, we have proved Theorem 11.

Remark 18. Given the regularity results stated in Theorems 8 and 9, it would seem that stronger convergence can not be expected for this problem. For example, convergence of the sequence $\{u_{p_i}\}_{i=1}^{\infty}$ in say $L^2(0, T; H^1(G))$ would imply that $u_0 \in L^2(0, T; H^1(G))$ which is not observed.

5. An example. We now present a simple example illustrating the importance of the constraint $\|u\|_{0,0} \leqq b$ to the existence of the optimal control in $L^2(Q)$. This behavior is due to the fact that $K(u)$ is not a coercive functional of u, that is, the norm induced by $K(u)$ is not equivalent to $\|\cdot\|_{0,0}$.

Consider the following initial-value problem

$$\begin{aligned} &\left(1-\frac{\partial^2}{\partial x^2}\right)y_t(x,t)-\frac{\partial^2}{\partial x^2}y(x,t)=u(x,t) \quad \text{in } (0,\pi)\times(0,T] \\ &y(x,0)=0 \quad \text{in } (0,\pi), \\ &y(0,t)=y(\pi,t)=0 \quad \text{in } (0,T] \end{aligned} \tag{19}$$

with the minimization problem

$$\begin{aligned} &\text{minimize} \quad K(u)=\int_0^T\int_0^\pi u(x,t)y(x,t;u)\,dx\,dt \\ &\text{subject to} \quad y(x,T;u)=\sin x. \end{aligned} \tag{20}$$

We introduce the Fourier series representations of y and u by setting

$$y(x,t)=\sum_{k=1}^{\infty}\eta_k(t)\sin kx \quad \text{and} \quad u(x,t)=\sum_{k=1}^{\infty}\mu_k(t)\sin kx.$$

We then obtain the sequence of initial-value problems

$$(1+k^2)\eta_k'(t)+k^2\eta_k(t)=\mu_k(t) \quad \text{in } (0,T], \qquad \eta_k(0)=0 \tag{21}$$

where

$$\eta_k(T)=\begin{cases}1 & \text{if } k=1,\\ 0 & \text{if } k\geqq 2.\end{cases} \tag{22}$$

Solving the problems (21) and imposing condition (22), we see that the admissible controls must satisfy the conditions

$$(23)\qquad \begin{aligned} &\int_0^T e^{t/2}\mu_1(t)\,dt = 2e^{T/2}\\ &\int_0^T e^{tk^2/(1+k^2)}\mu_k(t) = 0 \quad \text{for } k \geqq 2. \end{aligned}$$

Integrating (20) we obtain for the minimization problem

$$(24)\qquad \text{minimize} \int_0^T \sum_{k=1}^{\infty} \mu_k(t)\eta_k(t)\,dt$$

subject to the integral constraints in (23). Now since $K(u) \geqq 0$ and $K(u) \leqq C\|u\|_{0,0}^2$ for every u in $L^2(Q)$, we see that the functional in (24) is equal to $\sum_{k=1}^{\infty} \int_0^T \mu_k(t)\eta_k(t)\,dt$ and that $\int_0^T \mu_k(t)\eta_k(t)\,dt \geqq 0$ for every $k \geqq 1$. Thus, for our example it suffices to consider the one dimensional problem

$$(25)\qquad \begin{aligned} &\text{minimize} \int_0^T \mu_1(t)\eta_1(t)\,dt\\ &\text{subject to} \int_0^T e^{t/2}\mu_1(t)\,dt = 2\,e^{T/2} \end{aligned}$$

or entirely in terms of η_1

$$(26)\qquad \begin{aligned} &\text{minimize}\quad 1 + \int_0^T \eta_1^2(t)\,dt\\ &\text{subject to}\quad \eta_1(T) = 1. \end{aligned}$$

Now it is clear that

$$\underset{\substack{\eta_1 \in L^2(0,T)\\ \eta_1(T)=1}}{\text{infimum}} \left(1 + \int_0^T \eta_1^2(t)\,dt\right) \geqq 1$$

so that

$$\underset{y(x,T;u)=\sin x}{\text{infimum}}\ K(u) \geqq \frac{\pi}{2}.$$

We define the sequence of controls

$$u_n(x,t) = \begin{cases} 0, & 0 \leqq t < \dfrac{(n-1)T}{n},\\[2ex] \dfrac{2}{T}\,n\,e^{(T-t)/2}\sin x, & \dfrac{(n-1)T}{n} \leqq t \leqq T. \end{cases}$$

The associated sequence of states is given by

$$y(x,t;u_n) = \begin{cases} 0, & 0 \leqq t < \dfrac{(n-1)T}{n},\\[2ex] \left[n\dfrac{t}{T} + (1-n)\right] e^{(T-t)/2}\sin x, & \dfrac{(n-1)T}{n} \leqq t \leqq T. \end{cases}$$

Note that $y(x,T;u_n) = \sin x$ for each n so the sequence $\{u_n\}_{n=1}^{\infty}$ is admissible.

Furthermore, applying l'Hôpital's rule we see that

$$\lim_{n\to\infty} K(u_n) = \frac{\pi}{2}.$$

The controls however converge to an impulse function

$$u_\infty(x, t) = \begin{cases} 0, & 0 \leqq t < T, \\ +\infty, & t = T. \end{cases}$$

Although this is the zero function as far as $L^2(Q)$ and $L^2(0, T; H^{-1}(G))$ are concerned, the limit function u_∞ behaves like the Dirac delta function which certainly belongs to neither of the spaces $L^2(Q)$ or $L^2(0, T; H^{-1}(G))$.

Remark 19. It is worth pointing out that because of the specialized nature of this example, the behavior of this problem resembles that of standard finite dimensional problems [5]. Keeping those problems in mind, the temptation is to generalize further the space of controls so that there is a solution in the admissible space. Here, however, any attempt to include the Dirac delta function by enlarging the control space to $H^{-1}(Q)$ affects the regularity of the state function $y(u)$ so that the trace $y(\cdot, T; u)$ no longer has the meaning in $L^2(G)$ required by the endpoint constraint in the original problem.

Acknowledgment. I wish to thank Professor Tsuan Wu Ting for his valuable comments concerning this work.

REFERENCES

[1] R. W. CARROLL AND R. E. SHOWALTER, *Singular and Degenerate Cauchy Problems*, Academic Press, New York, 1976.

[2] P. CHEN AND M. GURTIN, *On a theory of heat conduction involving two temperatures*, Z. Angew. Math. Phys., 19 (1968), pp. 614–627.

[3] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, S. K. Mitter, translator, Springer-Verlag, New York, 1971.

[4] J. L. LIONS AND E. MAGENES, *Non-Homogenous Boundary Value Problems and Applications, I.*, P. Kenneth, translator, Springer-Verlag, New York, 1972.

[5] D. G. LUENBERGER, *Optimization by Vector Space Methods*, Wiley, New York, 1969.

[6] R. E. SHOWALTER, *The Sobolev Equation, II*, Applicable Anal. 5 (1975), pp. 81–99.

[7] R. E. SHOWALTER AND T. W. TING, *Pseudo-parabolic partial differential equations*, SIAM J. Math. Anal., 1(1970), pp. 1–26.

[8] T. W. TING, *Certain Non-steady Flows of Second-order Fluids*, Arch. Rational Mech. Anal., 14 (1963), pp. 1–26.

[9] L. W. WHITE, *Control Problems Governed by a Pseudo-parabolic Partial Differential Equation*, Trans. Amer. Math. Soc., to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 5, September 1979

0363-0129/79/1705-0003 $01.00/0

EXACT PENALTIES FOR LOCAL MINIMA*

SZYMON DOLECKI† AND STEFAN ROLEWICZ†

Abstract. We provide a sufficient condition for the exact equivalence of constrained minimization problems and the minimization of associated generalized Lagrangians with respect to a perturbing class Φ_1. Exact equivalence amounts to equality of the sets of local solutions restricted to some region. The sufficient condition is expressed in terms of certain semicontinuity properties of objective functions and constraint multifunctions; for Banach spaces it becomes local controllability. The requirement is made more specific for mathematical programming.

In this context we discuss properties of inner derivatives and approximations of multifunctions and we present a considerable extension of the Lusternik theorem.

1. Introduction. We are concerned with what we call "exact equivalence" of the minimization problem

$$f(x)\to\inf,\qquad x\in\Gamma y_0 \tag{1}$$

with the problem of unconstrained minimization of a generalized Lagrangian. An element x_0 of Γy_0 is called an R-solution of (1) if $f(x_0)\leqq f(x)$ as $x\in R\cap\Gamma_{y_0}$. We say that x_0 is a local solution, if there exists a neighborhood R of x_0 such that x_0 is an R-solution. Exact equivalence at x_0 means that all the local solutions around x_0 to both the considered problems are identical. In this sense, exact equivalence is a stronger property than that of exact penalty.

The real-valued function f in (1) is defined on a topological space X and a "constraints" multifunction Γ maps a topological space Y into subsets of X (Γ: $Y\to 2^X$). Let Φ be a class of finite real-valued functions on Y. The (generalized) Lagrange function L for (1) with respect to Φ was defined by Kurcyusz [12]

$$L(x,\varphi,y_0)=f(x)-\sup_{y\in\Gamma^{-1}x}\varphi(y)+\varphi(y_0) \tag{2}$$

where $\Gamma^{-1}x=\{y\colon x\in\Gamma y\}$ and it was pointed out [12] that most of known augmented Lagrangians (called sometimes penalty functions) are of the form (2) (see also Dolecki–Kurcyusz [5]).

In this paper we deal with the class Φ_1 on a metric space (Y,ρ):

$$\Phi_1=\{-k\rho(\cdot,z)+r;\ k>0,\ r\in\mathbb{R},\ z\in Y\}. \tag{3}$$

We first give a general sufficient condition for exact equivalence in terms of some semicontinuity properties of Γ and f (§ 4).

Section 5 is devoted to image nearly inner approximations of multifunctions in normed spaces, the notion having been introduced in [5]. In the case of equality constraints our approximation essentially generalizes the notion of continuously (in the operator norm topology) differentiable maps. As for the inequality constraints it is closely related to the Levitin–Miljutin–Osmolovskii Approximation (see Ioffe [9]).

In § 6 we discuss the form of approximations for some important special multifunctions, e.g., associated with the mathematical programming problems:

$$f(x)\to\inf\qquad G(x)=0;\ g_i(x)\leqq 0,\qquad i=1,2,\cdots,n. \tag{4}$$

In § 7 we express the sufficiency condition for exact equivalence in terms of controllability of an approximation of the multifunction and we specialize it to problems

* Received by the editors January 19, 1978, and in revised form December 20, 1978.

† Institute of Mathematics, Polish Academy of Sciences, Śniadeckich 8, Warsaw, Poland.

of type (4) obtaining an extension of the Slater condition.

This result is close in spirit to that of Ioffe [9] provides simpler verifying criteria and is sharper. Pietrzykowski [14] Howe [7] take a different point of view discussing jointly existence of solutions and "exactness". They use strong regularity assumptions and prove only exact penalty, not exact equivalence.

In [9] and in the present paper existence questions are not involved, as they are of different nature than those of equivalence, and it is natural to study them separately.

To our knowledge the first results on exact equivalence (and the concept itself) were given in [4], but it is Theorem 5 (§ 3) on localization of upper Hausdorff semicontinuity that enables us to draw rather strong conclusions from a semicontinuity theory elaborated in [3].

Results of [3] together with Theorem 5 extend the Lusternik theorem considerably.

2. Some preliminaries. Let $R\subset X$. The *primal functional* of (1) restricted to R is defined by

$$(5)\qquad \overline{f\Gamma_R}(y)=\inf_{x\in R\cap\Gamma y} f(x).$$

We shall say that two minimization problems are R-equivalent, if all their R-solutions are identical.

A constraints multifunction Γ is *of the equality type* if there is a mapping G from (a subset of) X to Y such that $\Gamma y=\{x: Gx=y\}$.

PROPOSITION 1. [5] *Let Γ be of the equality type. Let $\varphi_0\in\Phi$. The problems* (1) *and the following*

$$(6)_{\varphi_0}\qquad L(x,\varphi_0,y_0)\to\inf,\qquad x\in R,$$

are R-equivalent, if and only if φ_0 is a strict subgradient of $\overline{f\Gamma_R}$ at y_0, i.e.,

$$(7)\qquad \overline{f\Gamma_R}(y)-\varphi_0(y)>\overline{f\Gamma_R}(y_0)-\varphi_0(y_0),\qquad y\neq y_0.$$

For general multifunctions the situation is far from being so nice. However, in this respect, the class Φ_1 in (3), on metric spaces (Y,ρ), possesses exceptionally good properties.

PROPOSITION 2. ([3], compare Balder [1]) *Suppose that the sets $\Gamma^{-1}x$ are closed for $x\in R$. If $\overline{f\Gamma_R}$ is Φ_1-subdifferentiable at y_0, i.e.,*

$$(8)\qquad \overline{f\Gamma_R}(y)-\varphi_0(y)\geqq\overline{f\Gamma_R}(y_0)-\varphi_0(y_0),$$

then there is a k_0 such that for $\varphi_0(y)=-k\rho(y,y_0)$, where $k\geqq k_0$, (1) *and* $(6)_{\varphi_0}$ *are R-equivalent.*

In order to discuss the conditions of Proposition 2 we recall that if (X,π) and (Y,ρ) are metric spaces and q is a positive function on $(0,+\infty)$ then Γ is said to be *upper Hausdorff semicontinuous* (*u.H.s.c.*) at y_0 at a *rate* q if for each $r>0$

$$(9)\qquad \Gamma B(y_0,q(r))\subset B(\Gamma y_0,r),$$

where for $A\subset Y\, B(A,r)=\bigcup_{y\in A}\{z:\rho(z,y)<r\}$ and $\Gamma A=\bigcup_{y\in A}\Gamma y$

PROPOSITION 3. [4] *Assume that the sets Γy are closed, R is bounded and that the multifunction $(R\cap\Gamma)y=R\cap\Gamma y$ is u.H.s.c. at y_0 at a linear rate. Suppose that f is Lipschitz continuous on R. Then the sets $\Gamma^{-1}x$ are closed and there is a K such that $-K\rho(y,y_0)$ is a strict subgradient of $\overline{f\Gamma_R}$ at y_0.*

3. Localization of upper Hausdorff semicontinuity. In order to treat local versions of equivalence (that is, equivalence of local solutions in the sense defined later) we need a result of upper Hausdorff semicontinuity.

Observe that if Γ is u.H.s.c. at y_0 and F is a closed subset of X, then the multifunction $F \cap \Gamma$ need not be u.H.s.c. at y_0. In [3] there is an example (Example 2.11) of a multifunction Γ that is u.H.s.c. at y_0 for which all the multifunctions $\{x: \pi(x, x_0) \leqq r\} \cap \Gamma$ are not u.H.s.c. at y_0, for some $x_0 \in \Gamma y_0$ and for all $r > 0$.

Consider also the following example with a convex multifunction: X is a Hilbert space with an orthonormal basis $\{e_n\}$, and $K = \{\sum_{n=1}^{\infty} t_n e_n \in X; |\sum_{n=1}^{\infty} t_n/n| \leqq 1\}$ (Note that K is closed and convex.) Let $\Gamma t = K + te_1$ for $t \in \mathbb{R}$. Certainly Γ is u.H.s.c. at $t = 2$ but $(K \cap \Gamma)$ is no longer u.H.s.c. at 2.

It was established in Dolecki–Rolewicz [6] that the preservation of upper Hausdorff semicontinuity under intersections with closed neighborhoods characterizes upper semicontinuity (in the classical sense, see Kuratowski [11]), which is too restrictive for our purposes.

The above facts suggest serious difficulties that we can meet when confronting the "localization" of upper Hausdorff semicontinuity. Fortunately we are able to get around them.

DEFINITION 4. We say that a multifunction Γ is *mobile at y_0 at a rate q*, if for each $x_0 \in \Gamma y_0$ and each neighborhood Q of x_0 there is a neighborhood $Q_0 \subset Q$ of x_0 such that $\bar{Q}_0 \cap \Gamma$ is u.H.s.c. at y_0 at the rate q.

THEOREM 5. *Let Γ be u.H.s.c. at y_0 at a rate q. Let $c > 1$. Then Γ is mobile at y_0 at a rate $\hat{q}$, $\hat{q}(r) = q(r/c)$*

Proof. Set $Q = B(x_0, \delta)$. Let $\varepsilon = \eta_1 > \eta_2 > \cdots$ be positive numbers such that

$$2 \sum_{n=1}^{\infty} \eta_n < \delta. \tag{10}$$

Define $Q_1 = B(x_0, \varepsilon)$. Suppose that we have defined Q_n. For each $x \in Q_n$ we choose an element $y(x, n) \in \Gamma y_0$ with the property

$$\rho(x, y(x, n)) < \eta_n, \qquad \rho(x, y(x, n)) < c \operatorname{dist}(x, \Gamma y_0). \tag{11}$$

Put $Q_{n+1} = \bigcup_{x \in Q_n} B(y(x, n), \eta_{n+1})$. Note that (11) holds for some $y(x, n) \in \Gamma y_0$, because by construction for $x \in Q_n$ $\operatorname{dist}(x, \Gamma y_0) < \eta_n$. Set $Q_0 = \bigcup_{n=1}^{\infty} Q_n$.

Properties of Q_0.

(α) Q_0 is a neighborhood of x_0;

(β) $\bar{Q}_0 \subset B(x_0, \delta)$.

Indeed if $x \in \bar{Q}_0$, then for each $\xi > 0$ there exist k and $y \in Q_k$ with $\rho(x, y) < \xi$. Take $\xi < \delta - 2\sum_{n=1}^{\infty} \eta_n$ and choose a corresponding k and y. We may now pick $y_{k-1} \in Q_{k-1}$ such that $\rho(y, y_{k-1}) < \eta_k + \eta_{k-1}$ and proceeding like this we choose $y_1 \in Q_1$ and thus have $\rho(y_1, x_0) < \eta_1 = \varepsilon$. We conclude that

$$\rho(x, x_0) < \xi + 2 \sum_{n=1}^{k} \eta_n < \delta.$$

(γ) If $x \in \bar{Q}_0$, then in view of (11)

$$\operatorname{dist}(x, \Gamma y_0 \cap \bar{Q}_0) \leqq c \cdot \operatorname{dist}(x, \Gamma y_0).$$

Now, let $y \in B(y_0, q(r))$ and let $x \in \Gamma y \cap \bar{Q}_0$. By our assumptions $\operatorname{dist}(x, \Gamma y_0) < r$; thus $\operatorname{dist}(x, \Gamma y_0 \cap \bar{Q}_0) < cr$.

4. Exact equivalence.

DEFINITION 6. [4] We say that (1) and $(6)_{\varphi_0}$ are *exactly equivalent at* x_0, if there is a neighborhood Q of x_0 such that the set of all local solutions to (1) that lie in Q and the set of all local solutions to $(6)_{\varphi_0}$ that lie in Q are equal.

This concept describes a very strong interdependence between (1) and $(6)_{\varphi_0}$—much stronger than the exact penalty property. Indeed, exact equivalence when established, enables us to find the set of all the local solutions (in Q) to (1) by solving locally a single unconstrained minimization problem (of course, it may happen that the values of f at different local solutions are different).

Before presenting a result on exact equivalence we recall that Γ is termed *uniformly δ-u.H.s.c. at* (x_0, y_0) *at a rate* q, if there is a neighborhood Q of x_0 and η such that

$$Q \cap \Gamma B(y, q(r)) \subset B(\Gamma y, r), \qquad r > 0 \text{ for } y \in B(y_0, \eta). \tag{12}$$

The above property is equivalent to *uniform lower semicontinuity of* Γ *at* (x_0, y_0) in the following sense (see [3, Thm. 2.15]): If (12) holds, then there is a neighborhood Q' of x_0 and $W = B(y_0, \eta')$ and $r_0 > 0$ such that for each $x \in Q'$ and $y \in \Gamma^{-1}x \cap W$

$$\Gamma^{-1}B(x, r) \supset B(y, q(r)), \quad \text{for } r \leqq r_0. \tag{13}$$

On the other hand, (13) entails (12) with some Q'' and η''.

We shall also need the lower semicontinuity (l.s.c.) of Γ^{-1} at (y_0, x_0): for each $r > 0$ there is $s > 0$ such that $\Gamma B(y_0, r) \supset B(x_0, s)$.

Finally the metric ρ of the space Y will be supposed to have the following property: for each $0 < \alpha < 1$ for $\rho(y_0, y_1) > 0$ there is $\bar{y}$ with

$$\rho(y_1, \bar{y}) \leqq \alpha\rho(y_0, y_1) \quad \text{and} \quad \rho(\bar{y}, y_0) \leqq (1-\alpha)\rho(y_0, y_1) \tag{14}$$

(Y is almost a normed space.)

THEOREM 7. *Let* (X, π), (Y, ρ) *be metric spaces and let* ρ *satisfy* (14). *Assume* $f: X \to \mathbb{R}$ *to be a locally Lipschitz function around* x_0. *Suppose that a multifunction* Γ: $Y \to 2^X$ *satisfies the following requirements*

(a) Γ *is closed-valued*;

(b) Γ *is uniformly δ-u.H.s.c. at* (x_0, y_0) *at a linear rate* $q(r) = ar$, $r \leqq r_0$;

(c) Γ^{-1} *is l.s.c. at* (y_0, x_0).

Then there exists $\varphi_0 \in \Phi_1$ *such that* (1) *and* $(6)_{\varphi_0}$ *are exactly equivalent.*

Proof. Pick a neighborhood Q of x_0 and $W = B(y_0, \eta)$ such that

(i) f is Lipschitz continuous in Q with constant $c > 0$,

(ii) Γ satisfies (12) and (13) for $x \in Q$ and $y \in \Gamma^{-1}x \cap W$,

(iii) $\Gamma W \supset Q$.

This is possible because of (b) and (c).

For any $x \in Q$ any neighborhood $P \subset Q$ of x and each $y \in \Gamma^{-1}x \cap W$ the multifunction

$$\tilde{\Gamma}y = \Gamma y, \qquad \tilde{\Gamma}z = P \cap \Gamma z \quad \text{for } z \neq y \tag{15}$$

is u.H.s.c. at y at the rate $q(r) = ar$.

In view of Theorem 5 there is a neighborhood $R \subset P$ of x such that $R \cap \Gamma$ is u.H.s.c. at y at any rate less than ar.

Fix a number $b > a \cdot c$. By [4, Thm. 2.7, Example 2.4] the primal functional $\overline{f\Gamma}_R$ (restricted to R) satisfies

$$\overline{f\Gamma}_R(z) \geqq \overline{f\Gamma}_R(y) - b\rho(z, y), \quad \text{for } \rho(z, y) < r_0 \tag{16}$$

for some $r_0 > 0$.

Now form the Lagrange function associated with (1) setting

$$\varphi_0(z) = -m\rho(z, y_0), \tag{17}$$

where $m = 2b$.

Suppose that $\hat{x} \in Q$ is a solution to (1); thus in particular $\hat{x} \in \Gamma y_0$. We may find such a neighborhood R of $\hat{x}$ that $f(\hat{x}) = \inf_{x \in R \cap \Gamma y_0} f(x)$ and (16) holds with $z = y$, $y = y_0$ and in view of Propositions 2 and 3 $\hat{x}$ is a local solution to $(6)_{\varphi_0}$.

Suppose now that $\hat{x} \in Q$ is a local solution to $(6)_{\varphi_0}$. Consider two cases $\hat{x} \in \Gamma y_0$ and $\hat{x} \notin \Gamma y_0$. The first situation in view of (16) combined with Propositions 2 and 3 implies that $\hat{x}$ is a local solution to (1). The second case must not occur. In fact, suppose that $\hat{x} \notin \Gamma y_0$. Choose a neighborhood Q of $\hat{x}$, $Q \cap \Gamma y_0 = \phi$ and such that $\hat{x}$ is minimum of L on Q; thus by Theorem 5 and Proposition 3 there exist r and R with $B(\hat{x}, r) \subset R \subset Q$

$$\begin{aligned} f(\hat{x}) + m \cdot \inf_{y \in \Gamma^{-1}\hat{x}} \rho(y, y_0) &\leqq \inf_{x \in R} \Big(f(x) + m \cdot \inf_{y \in \Gamma^{-1}x} \rho(y, y_0)\Big) \\ &= \inf_{y \in Y} \big(\overline{f\Gamma}_R(y) + m \cdot \rho(y, y_0)\big). \end{aligned} \tag{18}$$

By (iii) the set $\Gamma^{-1}\hat{x}$ intersects $W (= B(y_0, \eta))$ and besides it is closed in view of [4, Lemma 1.12]. Therefore for each $\varepsilon > 0$ there is $\hat{y} \in Fr\Gamma^{-1}\hat{x} \cap W$ such that $\inf_{y \in \Gamma^{-1}\hat{x}} \rho(y, y_0) \geqq \rho(\hat{y}, y_0) - \varepsilon/m$. For each $y \in \Gamma^{-1}\hat{x} \cap W$ we have by (13), $\Gamma^{-1}R \supset B(y, q(r))$ where q is that of (b); thus $\Gamma^{-1}R \supset B(\Gamma^{-1}\hat{x} \cap W, q(r))$. On the other hand $y_0 \notin \Gamma^{-1}R$ that is dist $(y_0, \Gamma^{-1}\hat{x}) > q(r)$. We take $q(r)/\rho(y_0, \hat{y}) = \alpha < 1$. By (14) we may find $\bar{y} \in \Gamma^{-1}R \setminus \Gamma^{-1}\hat{x}$ such that

$$\rho(\bar{y}, \hat{y}) \leqq \alpha\rho(\hat{y}, y_0) \quad \text{and} \quad \rho(\bar{y}, y_0) \leqq (1-\alpha)\rho(y_0, \hat{y}).$$

We estimate using Proposition 3

$$\begin{aligned} &\overline{f\Gamma}_R(\bar{y}) + m\rho(\bar{y}, y_0) - \overline{f\Gamma}_R(\hat{y}) - m\rho(\hat{y}, y_0) \\ &\leqq b\rho(\bar{y}, \hat{y}) + m(\rho(\bar{y}, y_0) - \rho(\hat{y}, y_0)) \\ &\leqq b\alpha\rho(\hat{y}, y_0) - m\alpha\rho(\hat{y}, y_0) \leqq -b\alpha\rho(\hat{y}, y_0) \leqq -b\alpha \operatorname{dist}(y_0, \Gamma^{-1}\hat{x}) < 0. \end{aligned} \tag{19}$$

Putting $\varepsilon = \frac{1}{2}b\alpha$ dist $(y_0, \Gamma^{-1}\hat{x})$, (19) yields a contradiction to (18) on recalling that $\overline{f\Gamma}_R(\hat{y}) \leqq f(\hat{x})$.

5. Image nearly inner approximations (inia) of multifunctions. An inia of a multifunction Γ will be defined as a family of closed convex multifunctions verifying some continuity assumptions and approximating Γ in a certain sense. Both continuity and approximation will be of "image type".

Image continuity of a family $\{C_t\}$ of bounded linear operators in Banach spaces indexed by a topological space T is defined in [15] as the Hausdorff continuity of $t \to C_t\bar{B}(0, 1)$. Extending this notion we may consider various types of semicontinuity of a family of multifunctions $\{M_t\}$. For instance, $\{M_t\}_{t \in T}$ is termed *lower image semicontinuous* (l.i.s.c.) at t_0 if there exists $r_0 > 0$ such that for $0 < r \leqq r_0$ for each $\varepsilon > 0$ there exists a neighborhood W of t_0 such that

$$B(M_t^{-1}B(0, r), \varepsilon) \supset M_{t_0}^{-1}B(0, r).$$

Let X and Y be normed spaces and let Γ: $Y \to 2^X$. Consider a family $\{\Lambda_{(x,y)}\}$ of multifunctions defined for $(x, y) \in G(\Gamma) \cap B(x_0, \delta) \times B(y_0, \delta)$, $\Lambda_{(x,y)}$: $Y \to 2^X$, $(x, y) \in G(\Lambda_{x,y})$. Suppose that $M_{(x,y)}$, defined by $G(M_{(x,y)}) = G(\Lambda_{(x,y)}) - (x, y)$, is l.i.s.c. at

(x_0, y_0), i.e., for $r < r_0$

$$B(\Lambda_{(x,y)}^{-1}B(x,r),\varepsilon) \supset \Lambda_{(x_0,y_0)}^{-1}B(x_0,r). \tag{20}$$

DEFINITION 8. [3] A family $\{\Lambda_{(x,y)}\}$ of closed convex multifunctions is called an *image nearly inner approximation* (*inia*) *of* Γ *at* $(x_0, y_0) \in G(\Gamma)$, if $\Lambda_{(x,y)} - (x,y)$ is lower image semicontinuous[1] at (x_0, y_0) and if for each $\theta > 0$ there is $r_0 > 0$ and $\alpha > \theta$ such that

$$B(\Gamma^{-1}B(x,r),\theta r) \supset \Lambda_{(x,y)}^{-1}B(x,r) \cap B(y,\alpha r) \tag{21}$$

for $r < r_0$ and $(x,y) \in G(\Gamma) \cap B(x_0, r_0) \times B(y_0, r_0)$.

The above notion is very broad: every multifunction Γ admits a trivial inia $G(\Lambda_{(x,y)}) = \{(x,y)\}$ for $(x,y) \in G(\Gamma)$. (A multifunction Γ is called nearly convex if it possesses an inia such that there is $p > 0$ so that for each $x \in B(x_0, p)$, $\Lambda_{(x_0,y_0)}^{-1}x = \varnothing$. However, for the sake of applications we shall be concerned with the inia for which $G(\Lambda_{(x_0,y_0)})$ is "big". The main use we make of inia is the establishment of semicontinuity properties of Γ.

We say that a multifunction Δ is *locally controllable at* y_0, if there is a neighborhood W of y_0 such that

$$W \subset \Delta^{-1}X, \tag{22}$$

i.e., for all $y \in W$, Δy is not empty.

THEOREM 9. (Dolecki [3. Thm. 5.2]) *Let X, Y be Banach spaces and let $\Gamma\colon Y \to 2^X$ be closed. If there exists an inia $\{\Lambda_{(x,y)}\}$ of Γ at (x_0, y_0) with the property that $\Lambda_{(x_0,y_0)}$ is locally controllable at y_0, then for some η, Γ is δ-u.H.s.c. at (x_0, y), for $\|y - y_0\| < \eta$ at a uniform linear rate.*

Example 10. [3] Let G be a mapping from X to Y continuously differentiable about x_0 and define $\Gamma y = \{x\colon G(x) = y\}$. The family $\{\Lambda_{(x,G(x))}\}$, $\Lambda_{x,G(x)}y = \{v\colon G(x) + G'(x)(v - x) = y\}$ is an inia of Γ at $(x_0, G(x_0))$. Lower image semicontinuity follows from the continuity of $G'(\cdot)$ in the operator norm topology and the condition (13) with $\alpha = +\infty$ is a consequence of the mean value theorem (see Ioffe–Tikhomirov [10]).

Multifunctions considered in Example 10 are of the equality type.

We shall now discuss these inia of equality-type multifunctions which are of the following special form

$$\Lambda_{(x,G(x))}y = \{v\colon G(x) + \Psi(x, v - x) = y\} \tag{23}$$

where for each x, $\Psi(x,\ \cdot\ )$ is a closed linear operator.

As we shall see such an inia constitutes still a much broader notion than those of continuous derivatives of Gateaux (or Fréchet) and may concern also non differentiable mappings. The following proposition is immediate

PROPOSITION 11. *Assume that $\Gamma y = \{x\colon G(x) = y\}$ where G is a map from X to Y and that a family of multifunctions $\{\Lambda_{(x,G(x))}\}$ is given by* (23). *Λ is an inia of Γ at $(x_0, G(x_0))$, if and only if* (20) *holds and if for each $\theta > 0$ there is an $r_0 > 0$ such that for $\|x - x_0\| < r_0$ and for each h, $\|h\| < r_0$ there exists h', $\|h'\| \leqq \|h\|$ so that*

$$\|G(x + h') - G(x) - \Psi(x,h)\| \leqq \theta\|h\|. \tag{24}$$

Example 12. Let $X = Y = \mathbb{R}$ and define $G(x) = x$ for $x \leqq 0$ and $G(x) = 2x$ for $x \geqq 0$. Set $\Psi(x,h) = h$. Of course, G is not differentiable at 0 but $\Lambda_{(x,G(x))}^{-1}v = G(x) + \Psi(v - x)$ is an inia of G.

[1] The original condition is slightly different. In [3] an inia is called "inner derivative".

Note that the image character of the approximation (21) allows us to choose for a function, say $G(x)=x$, $\Psi(x, h)=-h$, and subsequently to define an inia by (23). In this case the inia is "orthogonal" to its multifunction.

Formula (20) with respect to (23) becomes: for $r \leqq r_0$ for each $\varepsilon > 0$ there is a $\delta > 0$ such that if $\|x-x_0\|<\delta$, then for each $\|h\|<r$ there exists h', $\|h'\| \leqq \|h\|$ such that

$$\|\Psi(x_0, h)-\Psi(x, h')\|<\varepsilon. \tag{25}$$

In particular this occurs, when the family of functions $\omega_h(x)=\Psi(x, h)$ is continuous at x_0 uniformly for $\|h\| \leqq r_0$.

Let us now discuss inia of multifunctions Γ of type

$$\Gamma y=\{x: g(x) \leqq y\} \tag{26}$$

where g is a real-valued function on a Banach space X, ($Y=\mathbb{R}$). We shall be concerned with inia of type

$$\Lambda_{(x,z)} y=\{v: z+\psi(x, v-x) \leqq y\} \tag{27}$$

for $z \geqq g(x)$, where ψ is a finite function $\psi(x, 0)=0$, convex l.s.c. in the second variable. Note that it is enough to check the desired properties of $\{\Lambda_{(x,z)}\}$ for $z=g(x)$.

PROPOSITION 13. *A multifunction* (27) *satisfies* (21) *with respect to the multifunction* Γ (26) *if and only if for each* $\theta>0$ *there exists* $r_0>0$ *such that for* $\|x-x_0\|<r_0$ *for* $\|h\|<r_0$ *there is an* h', $\|h'\|<\|h\|$ *so that*

$$g(x+h')-g(x)-\psi(x, h) \leqq \theta\|h\|. \tag{28}$$

Proof. Condition (21) is equivalent in our case to

$$\inf_{\|h\|<r}(g(x)+\psi(x, h)) \geqq \inf_{\|h\|<r} g(x+h)-\theta r \tag{29}$$

which is equivalent to (28).

As (28) is less stringent than (24), G from Example 12 provides an example of (26) with nondifferentiable function $g=G$ possessing an inia of type (27).

Example 14. The Levitin–Miljutin–Osmolovskii approximation φ of a function g is defined in Ioffe [9] as a function that satisfies

(i) $\varphi(x, 0)=g(x)$,
(ii) $\varphi(x, \cdot)$ is convex continuous,
(iii) $\liminf_{x \to x_0, h \to 0}\|h\|^{-1}(\varphi(x, h)-g(x+h)) \geqq 0$.

Setting $\psi(x, h)=\varphi(x, h)-g(x)$ we observe that $\psi(x, 0)=0$, $\psi(x, \cdot)$ is convex l.s.c. and (28) holds. Consequently in view of Proposition 13, the multifunction $\Lambda_{(x, g(x))} y=\{v: \varphi(x, h) \leqq y\}$, where φ is the L.M.O. approximation (of g at x_0), satisfies (21).

Lower image semicontinuity need not follow from (21) as we may see from the following example. Take $g(x)=x$, $x \in \mathbb{R}$ and define $\psi(x, h)=|h|$ for $x \neq 0$ and $\psi(0, h)=h$.

PROPOSITION 15. *If g is continuous, then* (20) *with respect to* (18) *is equivalent to the following condition: for* $r<r_0$ *for each* $\varepsilon>0$ *there is a* $\delta>0$ *such that if* $\|x-x_0\|<\delta$, *then for each* $\|h\|<r$ *there is an* h', $\|h'\| \leqq \|h\|$ *such that*

$$\psi(x_0, h) \geqq \psi(x, h')-\varepsilon. \tag{30}$$

In particular, this occurs if the functions $\omega_h(x)=\psi(x, h)$ *are upper semicontinuous at* x_0 *uniformly* $\|h\|<r$.

Example 16. Let $g = g_1 + g_2$, where g_1 is convex l.s.c. and g_2 is continuously differentiable. Then an inia of (26) may be chosen to be

$$\Lambda_{(x,g(x))}y = \{v\colon g_1(v) + g_2(x) + g_2'(x)(v-x) \leqq y\}.$$

We shall discuss the utility of a generalized derivative in constructing multifunctions of type (27). Let g be a locally Lipschitz function. Following Clarke [2] we define the directional derivative of g at x towards h as

$$g^0(x, h) = \inf_{r>0} \sup_{\substack{\|v\|<r \\ 0<t<r}} \frac{g(x+v+th) - g(x+v)}{t}.$$

Note, that $g^0(x, \cdot)$ is a continuous positively homogeneous function. It should be mentioned that $g(x) + g^0(x, h)$ need not be an L.M.O. approximation of g (Definition 13).

Example 17. Let $g\colon \mathbb{R} \to \mathbb{R}$, $g(x) = x^2 \cdot \cos(1/x)$ for $x \neq 0$ and $g(0) = 0$. Outside zero the directional derivative is just the ordinary derivative: $g^0(x, h) = g'(x)h = (2x\cos(1/x) - \sin(1/x))h$ while $g^0(0, h) = |h|$.

Consider the expression

$$g(x+h) - g(x) - g^0(x, h). \tag{31}$$

For each $\varepsilon > 0$ there exists x, $|x| < \varepsilon$ such that $g^0(x, h) = g'(x)h \leqq -(1-\varepsilon)|h|$. On the other hand exists r_0 such that if $|x| < r_0$, $|h| < r_0$ then $|g(x+h) - g(x)| < \varepsilon|h|$. Therefore (31) is greater than $-\varepsilon|h| - g'(x)h \geqq (1-2\varepsilon)|h|$ for a sequence of x tending to zero and any $|h| < r_0$. Thus (iii) of Example 14 is violated.

The following example shows that (27) built with the aid of the Clarke derivative essentially extends the area appointed by continuously differentiable functions g.

Example 18. Let p be defined on the real axis as follows $p(x) = -(x-n)$ if n is even and $|x-n| \leqq \frac{1}{2}$, and $p(x) = (x-n)$ if n is odd $|x-n| \leqq \frac{1}{2}$. Set

$$g(x) = \sum_{n=0}^{\infty} \frac{1}{2^{2n}} p(2^n x).$$

Notice that g is a Lipschitz function with constant 2 and the set on which g is not differentiable is dense. Take any $\theta > 0$, and choose n_0 such that $1/2^{n_0-1} < \theta$. Then

$$\sum_{n=0}^{n_0} \frac{1}{2^{2n}} p(2^n x) = \left(1 + \tfrac{1}{2} + \cdots + \frac{1}{2^{n_0}}\right)x$$

for $|x| < 1/2^{n_0+1}$ while the remainder $|\sum_{n=n_0+1}^{\infty} (1/2^{2n}) p(2^n x)|$ is less than $1/2^{2n_0}|x|$. Moreover for $|x| < 1/2^{n_0+2}$, $|h| < 1/2^{n_0+2}$ we have $|\sum_{n=n_0+1}^{\infty} (1/2^{2n}) \cdot (p(2^n(x+h)) - p(2^n x))| < (1/2^{n_0})|h|$. Thus $g(x+h) - g(x) - g^0(x, h) < \theta|h|$ and (27) is satisfied. g^0 satisfies (30) as well and so $\Lambda_{(x,g(x))}y = \{v\colon g(x) + g^0(x, v-x) \leqq y\}$ is a derivative of (26) at (0, 0).

In spite of its dense nondifferentiability g of Example 18 is still rather regular (it is differentiable at 0). A more sophisticated example of a densely nondifferentiable function with $f^0(x, h) = |h|$ verifying (27) follows readily from Zygmunt [16].

6. Inner derivatives of multifunctions. We shall now discuss these specific inia which "keep tight" to their multifunctions. The specification amounts to putting $h' = h$ in (24), (25), (28), (30).

DEFINITION 19. A family $\{\Lambda_{(x,y)}\}$ of closed convex multifunctions is called an *inner derivative of* Γ *at* (x_0, y_0), if for each $\theta>0$ there is an $r>0$ such that

$$B(\Gamma^{-1}x+h, \theta r)\supset \Lambda^{-1}_{(x,y)}x+h \tag{32}$$

for $\|h\|<r$ and $(x, y)\in G(\Gamma)\cap B(x_0, r)\times B(y_0, r)$, and if there is an r_0 such that for every $\varepsilon>0$ there exists a neighborhood W of (x_0, y_0) such that

$$B(\Lambda^{-1}_{(x,y)}x+h, \varepsilon)\supset \Lambda^{-1}_{(x_0,y_0)}x_0+h \tag{33}$$

for $\|h\|<r_0$ and $(x, y)\in W$.

Certainly, an inner derivative (of Γ at (x_0, y_0)) is an inia (of Γ at (x_0, y_0)). The inia of Examples 10, 14 are also inner derivatives.

The following observation will be instrumental for multifunctions of the mathematical programming type (4). Let Γ_1: $Y_1\to 2^X$ and Γ_2: $Y_2\to 2^X$. Consider Γ: $Y_1\times Y_2\to 2^X$ given by $\Gamma(y^1, y^2)=\Gamma_1 y^1\cap\Gamma_2 y^2$. Equip $Y=Y_1\times Y_2$ with the norm $\|(y^1, y^2)\|_Y=\|y^1\|_{Y_1}+\|y^2\|_{Y_2}$.

PROPOSITION 20. *Suppose that* $x_0\in\Gamma_1 y_0^1\cap\Gamma_2 y_0^2$. *Let* $\{\Lambda_{(x,y^1)}\}$ *be an inner derivative of* Γ_1 *at* (x_0, y_0^1) *and let* $\{M_{(x,y^2)}\}$ *be an inner derivative of* Γ_2 *at* (x_0, y_0^2). *Consider* $K(x, y^1, y^2)$: $Y_1\times Y_2\to 2^X$,

$$K(x, y^1, y^2)(z^1, z^2)=\Lambda_{(x,y^1)}z^1\cap M_{(x,y^2)}z^2 \tag{34}$$

(*well-defined on some* $G(\Gamma)\cap B(x_0, \delta)\times B(y_0^1, \delta)\times B(y_0^2, \delta)$).

Then $\{K_{(x,y^1,y^2)}\}$ *is an inner derivative of* Γ *at* (x_0, y_0^1, y_0^2).

Proof. Of course each $K_{(x,y^1,y^2)}$ is closed and convex. Observe that $\Gamma^{-1}x=\Gamma_1^{-1}x\times\Gamma_2^{-1}x(=\{(y^1, y^2): y^1\in\Gamma_1 x, y^2\in\Gamma_2 x\})$ and $K^{-1}_{(x,y^1,y^2)}v=\Lambda^{-1}_{(x,y^1)}v\times\Lambda^{-1}_{(x,y^2)}v$. Now, it is enough to notice that in Y, $B(A_1, r)\times B(A_2, r)\subset B(A_1\times A_2, 2r)$ for arbitrary $A_1\subset Y_1$, $A_2\subset Y_2$ and $r>0$.

7. Finale. Combined, Theorems 7 and 9 yield the following theorem on exact equivalence.

THEOREM 21. *Let* X, Y *be Banach spaces. Let* Γ: $Y\to 2^X$ *be a closed multifunction. Assume* f: $X\to\mathbb{R}$ *to be a locally Lipschitz function around* x_0. *Suppose that* Γ^{-1} *is l.s.c. at* (y_0, x_0) *and that there exists an inia* $\{\Lambda_{x,y}\}$ *such that* $\Lambda_{(x_0,y_0)}$ *is locally controllable at* y_0.

Then there exists $\varphi_0\in\Phi_1$, $\varphi_0(y)=-k\|y-y_0\|$, *such that* (1) *and* $(6)_{\varphi_0}$ *are exactly equivalent.*

We consider now a problem of mathematical programming (4).

We define Y to be $Y_0\times\mathbb{R}^n$ where Y_0 is a Banach space and we set $\|(y_0, y_1, \cdots, y_n)\|=\|y_0\|+\sum_{i=1}^n|y_i|$. Here G: $X\to Y_0$.

If we assume that G and g_i, $i=1, \cdots, n$, are continuous then

$$\Gamma y=\{x\colon G(x)=y_0, g_i(x)\leqq y_i, i=1, \cdots, n\} \tag{35}$$

the constraint multifunction is closed and Γ^{-1} is l.s.c. If we take $\varphi(y)=-k\|y\|$, then the associated Lagrange functional (2) becomes

$$L(x, k, 0)=f(x)+k\Big(\|G(x)\|+\sum_{i=1}^n\max(g_i(x), 0)\Big). \tag*{$(36)_k$}$$

Let an inner derivative of G be of type (23) where Ψ: $X\times X\to Y_0$. Let inner derivatives of g_i be of type (27) where ψ_i: $X\times X\to\mathbb{R}$. Then we may define an inner derivative of (35) by

$$\begin{aligned}\Lambda_{(x,G(x),g_i(x))}y=\{x+h: G(x)+\Psi(x, h)=y_0,\\ g_i(x)+\psi_i(x, h)\leqq y_i, i=1, \cdots, n\}.\end{aligned} \tag{37}$$

Let x_0 fulfil the constraint of (4). We denote by I the set of all numbers for which $x_0 \in \mathrm{Fr}\{x: g_i(x) \leqq 0\}$ where Fr denotes the topological boundary (I is smaller than the usual set of active constraints indices).

THEOREM 22. *Let f be locally Lipschitz continuous about x_0 and let Ψ and ψ_i, $i = 1, \cdots, n$, be as defined above. Assume that there is an $\varepsilon > 0$ such that for any $\|(y_0, \cdots, y_n)\| < \varepsilon$ there exists $h \in X$ so that $\Psi(x_0, h) = y_0$, $\psi_i(x_0, h) \leqq y_i$, $i \in I$.*

Then there is $k_0 > 0$ such that (4) *and* $(36)_{k_0}$ *are exactly equivalent.*

Proof. If $j \notin I$, then g_i is no longer stringent for the local minimization around x_0 and may be removed. The second condition amounts to the local controllability of $\Lambda_{(x_0,y)}$, where Λ is given by (37). All the assumptions of Theorem 21 are satisfied and thus the proof is complete.

Note that if $\Psi(x_0, \cdot)$ is a closed linear operator and $\psi_i(x_0, \cdot)$ are continuous, then our controllability condition of Theorem 22 is equivalent to the Slater condition: $\Psi(x_0, X) = Y$ and there is an $h \in X$ such that $\Psi(x_0, h) = 0$, $\psi_i(x_0, h) < 0$ for each i; this is weaker and more general than that of Pietrzykowski [14].

Remark 23. Combined Theorems 5 and 9 extend considerably the Lusternik theorem [11] which claims that if $G'(x_0)X = Y$, then $\Gamma y = G^{-1}y$ is δ-u.H.s.c. at y_0 linearly. In the case of equality constraints our theorems do not require differentiability of G; they give uniform results for a neighborhood of y_0, and finally localize upper Hausdorff semicontinuity.

Remark 24. From the uniform character of our considerations it follows that exact equivalence is stable in y. Moreover for y_1 close to y_0 we may use φ_1 "close" to φ_0. This gives a sidelight on problems of the stability of Lagrange multipliers.

Note. After the revised version of this paper had been completed we became acquainted with the paper of Han and Mangasarian [17], many results of which follow from ours, but some others present new interesting developments, e.g. second order conditions for exactness, the estimates for the constant k_0 (in Theorem 22) and its dual characterizations.

Acknowledgment. We wish to express our appreciation to Professor A. D. Ioffe for his highly valuable remarks that helped to improve the paper.

REFERENCES

[1] E. J. BALDER, *An extension of duality-stability relations to nonconvex optimization problems*, SIAM J. Control, 15 (1977), pp. 329–343.

[2] F. H. CLARKE, *A new approach to Lagrange multipliers*, Math. Operations Res., 1 (1976), pp. 165–174.

[3] S. DOLECKI, *Semicontinuity in constrained optimization, I*, Control Cyber., 7 (1978) No. 2, pp. 5–16, 7 (1978) No. 3, pp. 17–26.

[4] ———, *Semicontinuity in constrained optimization, II*, Ibid., to appear.

[5] S. DOLECKI AND S. KURCYUSZ, *On Φ-convexity in extremal problems*, this Journal, 16 (1978), pp. 277–300.

[6] S. DOLECKI AND S. ROLEWICZ, *Metric characterizations of the upper semicontinuity*, J. Math. Anal. Appl., to appear.

[7] S. M. HOWE, *New conditions for exactness of a simple penalty function*, this Journal, 11 (1973), pp. 378–381.

[8] A. D. IOFFE, *Regular points of Lipschitz mappings*, to appear.

[9] ———, *Necessary and sufficient conditions for a local minimum, 2, Conditions of Levitin–Miljutin–Osmolovskii type*, this Journal, 17 (1979), pp. 251–265.

[10] A. D. IOFFE AND A. M. TIKHOMIROV, *Theory of Extremal Problems*, Nauka, Moscow, 1974. (In Russian.)

[11] K. KURATOWSKI, *Topology*, vol. I, II, Academic Press and Polish Scientific Publishers, London-New York-Warsaw, 1966.

[12] S. KURCYUSZ, *Some remarks on generalized Lagrangians*, Proc. 7th IFIP Conf. (Nice 1975).

[13] L. W. NEUSTADT, *A general theory of extremals*, J. Comp. Syst. Sci., 3 (1969), pp. 57–92.

[14] T. PIETRZYKOWSKI, *Erratum*; *An exact potential method for constrained maxima*, SIAM J. Numer. Anal., 8 (1971), p. 481.

[15] S. ROLEWICZ, *Funktionalanalysis und Steuerungstheorie*, Springer-Verlag, Berlin-Heidelberg-New York, 1976.

[16] W. ZYGMUNT, *On the full solution of the paratingent equation*, Ann. Univ. M. Curie–Skłodowska, 268 (1972), pp. 103–108.

[17] S.-P. HAN AND O. L. MANGASARIAN, *Exact penalty functions in nonlinear programming*, Math. Programming, to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 5, September 1979

RATES OF CONVERGENCE FOR STOCHASTIC APPROXIMATION TYPE ALGORITHMS*

HAROLD J. KUSHNER† AND HAI HUANG‡

Abstract. We consider the general form of the stochastic approximation algorithm $X_{n+1} = X_n + a_n h(X_n, \xi_n)$, where h is not necessarily additive in ξ_n. Such algorithms occur frequently in applications to adaptive control and identification problems, where $\{\xi_n\}$ is usually obtained from measurements of the input and output, and is almost always complicated enough that the more classical assumptions on the noise fail to hold. Let $a_n = A/(n+1)^\alpha$, $0<\alpha\leqq 1$, and let $X_n \to \theta$ w.p. 1. Define $U_n = (n+1)^{\alpha/2}(X_n - \theta)$. Then, loosely speaking, it is shown that the sequence of suitable continuous parameter interpolations of the sequence of "tails" of $\{U_n\}$ converges weakly to a Gaussian diffusion. From this we can get the asymptotic variance of U_n as well as other information. The assumptions on $\{\xi_n\}$ and $h(\cdot,\cdot)$ are quite reasonable from the point of view of applications.

1. Introduction. Rates of convergence for stochastic approximation problems were given in [1], [2], [3], [4], the latter two references getting better results via weak convergence methods, for both constrained and unconstrained systems.

A form of stochastic approximation algorithm which is of increasing importance is the following. Let $\{a_n\}$ denote a sequence of positive real numbers with $\sum_n a_n = \infty$, h a suitable function and $\{\xi_n\}$ a sequence of random variables. Define the sequence of R^r-valued random variables $\{X_n\}$ by

$$X_{n+1} = X_n + a_n h(X_n, \xi_n). \tag{1.1}$$

In [1]–[4], the function h was essentially additive in ξ_n, as is usually the case in classical Kiefer–Wolfowitz and Robbins–Munro type stochastic approximation algorithms. Of course, if $\{\xi_n\}$ is a sequence of independent random variables, then $h(X_n, \xi_n)$ can be written in the form $E[h(X_n, \xi_n)|X_n] + \psi_n$, where $\psi_n = h(X_n, \xi_n) - E[h(X_n, \xi_n)|X_n]$ is a member of an orthogonal sequence, and we are back to the classical case. In the applications that we have in mind the $\{\xi_n\}$ can be rather general processes.

The more general form (1.1) arises in applications to problems in the recursive identification of the parameters of linear systems, or in the so-called self-tuning regulators or in other applications of adaptive systems [5], [6]. Such applications are the motivation for this work. Often X_n is an estimate of the vector system parameter and ξ_n is a random vector which is related to the measured inputs and outputs of the system. The rate of convergence problem for such situations has not been dealt with, and somewhat different methods are required.

In this paper we develop rate of convergence results for (1.1) under quite reasonable conditions. Owing to the way in which (1.1) arises in applications, the $\{\xi_n\}$ is rarely a sequence of independent random variables, and $E(h(X_n, \xi_n)|\xi_0, \cdots, \xi_{n-1})$ is rarely a function only of X_{n-1}. Thus classical rate of convergence methods (as in [1], [2]) cannot be used directly. We use some of the ideas in [3], [4], but adapted to our case, and under weaker conditions on the noise sequences.

The problem is formulated and some assumptions given in § 2. Weak convergence of a sequence of normalized $\{X_n\}$ is given in § 3, and the general rate result appears in § 4.

* Received by the editors November 27, 1978.

† Brown University, Divisions of Engineering and Applied Mathematics (Center for Dynamical Systems), Providence, Rhode Island 02912. The work of this author was supported by AFOSR AF-AFOSR 76-3063, NSF-Eng-73-03846-A-03, ONR N0014-76-C-0279-P0002.

‡ Brown University, Division of Applied Mathematics (Center for Dynamical Systems), Providence, Rhode Island 02912. The work of this author was supported by NSF-Eng-73-03846-A-03.

2. Terminology and problem formulation. For $\alpha \in (0, 1]$ and A a matrix, set $a_n = A/(n+1)^{\alpha}$. Since we are concerned with rates of convergence, we assume convergence (see [4] for a detailed discussion of the convergence both w.p. 1 and weakly). In particular, we suppose that there is a $\theta \in R^r$ such that $X_n \to \theta$ w.p. 1. Set $U_n = (n+1)^{\alpha/2}(X_n - \theta)$, $\Delta t_n = (n+1)^{-\alpha}$, $h_n = h(\theta, \xi_n)$ and $\bar{h}_n = (n+2/n+1)^{\alpha/2} h_n$. Let $h(\cdot, \xi)$ be continuously differentiable for each ξ, with the gradient $h_x(\cdot, \cdot)$ being Borel-measurable.

There is a function $O(\cdot)$ such that with H_n defined by (2.1), (2.2) holds. (See [3, eq. (5.2)] for a related calculation for the case where h is additive in ξ.)

$$
\begin{aligned}
H_n = {} & A h_x(\theta, \xi_n) + \frac{\alpha}{2(n+1)^{1-\alpha}} I + O\left(\frac{1}{n+1}\right) I \\
(2.1) \qquad & + A\left(\frac{n+2}{n+1}\right)^{\alpha/2} \int_0^1 [h_x(\theta + t(X_n - \theta), \xi_n) - h_x(\theta, \xi_n)]\, dt \\
& + A\left[\left(\frac{n+2}{n+1}\right)^{\alpha/2} - 1\right] h_x(\theta, \xi_n),
\end{aligned}
$$

$$(2.2)^1 \qquad U_{n+1} = (I + \Delta t_n H_n) U_n + A\sqrt{\Delta t_n}\, \bar{h}_n.$$

For future use *define* $\delta W_n = \sqrt{\Delta t_n}\, h_n$, $\delta \bar{W}_n = \sqrt{\Delta t_n}\, \bar{h}_n$.

Lemmas 1 and 2 contain some preparatory results concerning the iteration (2.2), and tightness of $\{U_n\}$ (i.e., $\sup_n P(|U_n| \geqq N) \to 0$ as $N \to \infty$) is proved in Theorem 1.

Next, following the general approach of [3], a sequence of processes $\{U^N(\cdot)\}$ is defined as follows. Let $t_n = \sum_{i=0}^{n-1} \Delta t_i$, $t_0 = 0$ and define $m(t) = \max\{k : t_k \leqq t\}$. Set $U^N(0) = U_N$ and $U^N(t) = U_{N+n}$ in $[t_{N+n} - t_N, t_{N+n+1} - t_N)$. Thus $U^N(\cdot)$ is a process whose paths are piecewise constant and in $D^r[0, \infty)$, the space of R^r-valued functions which are right continuous on $[0, \infty)$ and have left-hand limits on $(0, \infty)$. Since it will be important for us to go back and forth between the $\{U_n\}$ and $\{U^N(\cdot)\}$ sequences, the functions $m(\cdot)$ and t_n will be used quite frequently, occasionally (and regrettably) causing some complicated notation.

Owing to the scale factor $a_n = A\Delta t_n$, the interpolation $U^N(\cdot)$ is quite natural for this problem. In Theorem 2 it will be shown that $\{U^N(\cdot)\}$ is tight in $D^r[0, \infty)$ and converges weakly to the stationary linear Gaussian diffusion (4.1). As is common in applications of weak convergence theory, if a sequence of measures $\{u_n\}$ is tight and converges weakly to u (all on R^r or $D^r[0, \infty)$), and u_n and u are induced by processes $X^n(\cdot)$ and $X(\cdot)$, resp. (with paths in R^r or $D^r[0, \infty)$), then we abuse terminology and say that $\{X^n\}$ is tight and converges weakly to X. This weak convergence gives us the basic rate of convergence result. Some advantages of our approach are discussed in [3]. It yields the convergence in distribution (to a normally distributed random variable, the stationary distribution of (4.1)) of $\{U_n\}$, but also more, since it gives information on the correlation structure of the *process* $\{U_{N+n}, n \geqq 0\}$ for large N.

Remark on weak convergence. Billingsley [7] is the most comprehensive reference. The space $D[0, T]$ is discussed in [7, §§ 14 and 15]. A brief summary of relevant facts is given in [4, Chap. 2]. $D^r[0, \infty)$ is endowed with the usual [7, § 14] Skorokhod topology, with which it is a complete separable metric space. Convergence in $D^r[0, \infty)$ occurs if, for some sequence $T \to \infty$, it occurs (for the truncated functions) in each $D^r[0, T]$.

[1] From (2.1) we can guess that if $\alpha = 1$ (resp. $\alpha < 1$) the "effective" component of H_n is $(Ah_x(\theta, \xi_n) + I/2)$ $(Ah_x(\theta, \xi_n)$, resp.).

Assumptions. (A1)–(A5) will be used throughout the paper.

(A1) $X_n \to \theta$ w.p. 1.

(A2) *$h(\cdot,\cdot)$ is a Borel function, continuously differentiable in its first argument for each value of the second, and the gradient $h_x(\cdot,\cdot)$ is Borel. Also $Eh(\theta,\xi_n)\equiv 0$ and*

$$\int_0^1 [h_x(\theta+t(X_n-\theta),\xi_n)-h_x(\theta,\xi_n)]\,dt\to 0 \quad \text{w.p. } 1$$

as $n\to\infty$. (Certainly this is true if the ξ_n are bounded and $h_x(\cdot,\cdot)$ is continuous.)

(A3a) *There is a matrix H such that for some* (*hence each*) $T>0$ *and each* $\varepsilon>0$

$$\lim_{n\to\infty} P\left\{\sup_{j\geqq n}\max_{0\leqq t\leqq T}\left|\sum_{i=m(jT)}^{m(jT+t)-1}\Delta t_i(h_x(\theta,\xi_i)-H)\right|\geqq\varepsilon\right\}=0.$$

(A3b) *There is a constant τ such that for each $\varepsilon>0$ and $T>0$,*

$$\lim_{n\to\infty} P\left\{\sup_{j\geqq n}\max_{0\leqq t\leqq T}\left|\sum_{i=m(jT)}^{m(jT+t)-1}\Delta t_i(|h_x(\theta,\xi_i)|-\tau)\right|\geqq\varepsilon\right\}=0,$$

where $|x|=(x'x)^{1/2}$ and $|M|=\sup_{|x|=1}|Mx|$ if M is a matrix.

Remark on (A3a) *and* (A3b). Conditions of type (A3a), (A3b) were used extensively in the monograph [4], and as shown in that reference are rather weak and quite natural for the problem. See, for example, the several cases discussed in [4, Chap. 2.2]. The conditions are commonly satisfied by the noise processes which appear in the usual applications to the identification problem. We mention *only* the following three cases for (A3a): (a) $\sum a_n^2<\infty$ and $\{h_x(\theta,\xi_n)-Eh_x(\theta,\xi_n)\}$ a martingale; (b) $h_x(\theta,\xi_n)-Eh_x(\theta,\xi_n)=\sum_{i=0}^{\infty} b_i\psi_{n-i}$, for a broad class of $\{b_i\}$, $\{\psi_i\}$ where $\{\psi_i\}$ are independent and identically distributed; (c) $\{\xi_n\}$ stationary, (A5) holds for h_x replacing h and $\sum a_i^2\,(\log_2 i)^2<\infty$ holds.

In order to illustrate our terminology and get some additional insight into (A3), let us define a process $\eta(t)$ as follows: $\eta(0)=0$, and $\eta(t)=\sum_{i=0}^{n-1}\Delta t_i(h_x(\theta,\xi_i)-H)$ on $[t_n,t_{n+1})$. Then

$$\eta(t)=\sum_{i=0}^{m(t)-1}\Delta t_i(h_x(\theta,\xi_i)-H).$$

Condition (A3a) implies that the variation of the "increasing compressed interpolation" $\eta(t)$ over an arbitrary interval $(\alpha,\alpha+T)$ goes to zero w.p. 1 as $\alpha\to\infty$.

(A4) *If $\alpha=1$, set $\bar H=AH+I/2$, and if $\alpha<1$, set $\bar H=AH$. The eigenvalues of $\bar H$ have negative real parts.*

(A5) *Define R_{mk} by $R_{mk}=Eh'(\theta,\xi_m)h(\theta,\xi_k)$. Then* $\sup_m\sum_{k=0}^{\infty}|R_{mk}|<\infty$. *Also* $\sup_m E|h_x(\theta,\xi_m)|^2<\infty$.

3. Tightness of $\{U_n\}$. In order to simplify the presentation of the chain of calculations, we present them partially in a sequence of lemmas. Among other things, we wish to show that the H_n and $\bar h_n$ in (2.2) can be replaced by $\bar H$ and h_n, respectively. Apart from differences due to the greater generality of the noise here, the main differences between the treatment of (1.1) and the past work where h was assumed additive in ξ are due to the randomness of the H_n. To deal with them, we exploit the "averaging" or "smoothing" conditions (A3) and the stability condition (A4). *We use K to denote a constant whose value may change from usage to usage.*

Henceforth $\{\varepsilon_k\}$ denotes a sequence of positive real numbers such that $\sum_k\varepsilon_k<\infty$. Let $\{M_k\}$ be a sequence of integers tending to ∞ as $k\to\infty$, and define the measurable sets (in the sample space) A_k, B_k and C_k by (note that $j\varepsilon_k\geqq t_{M_k}$ and $m(j\varepsilon_k)\geqq M_k$ are equivalent statements)

$$A_k = \left\{ \sup_{j\varepsilon_k \geqq t_{M_k}} \max_{0\leqq t\leqq \varepsilon_k} \left| \sum_{i=m(j\varepsilon_k)}^{m(j\varepsilon_k+t)-1} \Delta t_i (Ah_x(\theta,\xi_i) - AH)\right| \geqq \varepsilon_k^2 \right\},$$

$$B_k = \left\{ \sup_{j\varepsilon_k \geqq t_{M_k}} \max_{0\leqq t\leqq \varepsilon_k} \left| \sum_{i=m(j\varepsilon_k)}^{m(j\varepsilon_k+t)-1} \Delta t_i (|h_x(\theta,\xi_i)| - \tau)\right| \geqq \varepsilon_k^2 \right\},$$

$$C_k = \left\{ \sup_{j\geqq M_k} \left| \int_0^1 [h_x(\theta + t(X_j-\theta), \xi_j) - h_x(\theta,\xi_j)]\, dt \right| \geqq \varepsilon_k^2 \right\}.$$

Set $D_k = \bigcup_{i=k}^{\infty} (A_i \cup B_i \cup C_i)$: Choose M_k such that $P\{A_k\}+P\{B_k\}+P\{C_k\}\leqq \varepsilon_k$ and $\Delta t_i \leqq \varepsilon_k^2$, $i \geqq M_k$. Such a choice is possible by (A3). Then $P\{D_k\} \equiv \mu_k \to 0$ as $k \to \infty$. Consequently for $\omega \notin D_k$ and $i \geqq M_k$, (A3) implies that the individual terms in the sums in (A3) satisfy

$$|\Delta t_i(Ah_x(\theta,\xi_i) - AH)| \leqq 4\varepsilon_k^2,$$
$$|\Delta t_i(|h_x(\theta,\xi_i)| - \tau)| \leqq 4\varepsilon_k^2.$$

From the definitions of M_k and D_k we immediately get the following lemma.

LEMMA 1. *Under* (A1)–(A3), *there is a constant K such that for each k and $\omega \notin D_k$ and $j \geqq M_k$,*

$$\sum_{i=m(j\varepsilon_k)}^{m(j\varepsilon_k+\varepsilon_k)-1} \Delta t_i |H_i| \leqq K\varepsilon_k,$$

$$\left| \sum_{i=m(j\varepsilon_k)}^{m(j\varepsilon_k+t)-1} \Delta t_i (H_i - \bar{H}) \right| \leqq K\varepsilon_k^2, \qquad t \leqq \varepsilon_k.$$

We now proceed to put the iteration (2.2) into a more convenient form. Define C_n^N by $C_{N+1}^N = I$ and for $n \leqq N$, $C_n^N = \prod_{j=n}^N (I + \Delta t_j H_j) \equiv (I + \Delta t_N H_N) \cdots (I + \Delta t_n H_n)$.

LEMMA 2. *Assume* (A1) *to* (A3). *Then on a set whose probability is arbitrarily close to* 1

$$C_{m(t_N+s)}^{m(t_N+t+s)} \to \exp \bar{H} t \tag{3.1}$$

as $N \to \infty$, uniformly on bounded t-intervals. Also, there is a real K such that for each k and each $N \geqq M_k$ and $\omega \notin D_k$ and $t \leqq \varepsilon_k$

$$C_{m(t_N+s)}^{m(t_N+t+s)} = [I + \bar{H} t + \sigma], \tag{3.2}$$

where $|\sigma| \leqq K\varepsilon_k^2$.

Proof. Equation (3.1) follows directly from (3.2) and we only prove (3.2) for $t \leqq \varepsilon_k$ and $s = 0$. For $M \geqq m$ we have

$$C_m^M = \prod_m^M (I + \Delta t_i H_i) = I + \sum_{i=m}^M \Delta t_i H_i + \sum_{i_2=m}^M \sum_{i_1>i_2}^M \Delta t_{i_1}\Delta t_{i_2} H_{i_1} H_{i_2} + \cdots$$
$$+ \Delta t_M \cdots \Delta t_m H_M \cdots H_m.$$

$$\left| C_m^M - \left(I + \sum_{i=m}^M \Delta t_i H_i\right)\right|$$

$$\leqq \sum_{i_2=m}^M \sum_{i_1>i_2}^M \Delta t_{i_1}\Delta t_{i_2} |H_{i_1}|\,|H_{i_2}| + \cdots + \Delta t_M \cdots \Delta t_m |H_M| \cdots |H_m| \tag{3.3}$$

$$\leqq \frac{1}{2}\left(\sum_{i=m}^M \Delta t_i |H_i| \right)^2 + \cdots.$$

Now using Lemma 1 to upper bound the right side of (3.3) and to estimate $\sum_{i=m}^{M} \Delta t_i H_i$ yields (3.2). Q.E.D.

We require one more preparatory setup. For any M, m and vector z_0 define

$$z_1 = \prod_{i=m}^{M} (I + \Delta t_i H_i) z_0 = C_m^M z_0,$$

where $t_{M+1} - t_m \leqq \varepsilon_k$ and $m \geqq M_k$. Let P denote the unique (under (A4)) symmetric positive definite matrix such that $\bar{H}'P + P\bar{H} = -I$; $x'Px$ is a Lyapunov function for the differential equation $\dot{x} = \bar{H}x$, which is asymptotically stable under (A4). Define $|x|_P = (x'Px)^{1/2}$, and let u denote a *positive* constant such that $u|x|_P^2 \leqq |x|^2$. By Lemma 2, if $m \geqq M_k$ and $(t_{M+1} - t_m) \leqq \varepsilon_k$ and $\omega \notin D_k$, we have

$$z_1 = [I + (t_{M+1} - t_m)\bar{H} + \sigma] z_0$$

where $|\sigma| \leqq K\varepsilon_k^2$ and (under (A4) and using $\bar{H}'P + P\bar{H} = -I$)

$$\begin{aligned} z_1'Pz_1 = z_0'Pz_0 - (t_{M+1} - t_m)|z_0|^2 \\ + z_0'[P\sigma + \sigma'P + \sigma'P\sigma + (t_{M+1} - t_m)(\bar{H}'P\sigma + \sigma'P\bar{H}) \\ + (t_{M+1} - t_m)^2 \bar{H}'P\bar{H}] z_0 \end{aligned}$$

from which we get (for some real K)

$$\begin{aligned} |z_1|_P^2 &\leqq (1 - u(t_{M+1} - t_m) + K\varepsilon_k^2)|z_0|_P^2 \\ &\leqq \exp[-u(t_{M+1} - t_m) + K\varepsilon_k^2]|z_0|_P^2. \end{aligned} \tag{3.4}$$

Thus $|C_m^M|_P \leqq \exp[-u(t_{M+1} - t_m) + K\varepsilon_k^2]$. We are now ready for the first theorem.

THEOREM 1. *Under* (A1) *to* (A5), $\{U_n\}$ *is tight on* R^r.

Proof. By iterating (2.2) we get

$$U_{N+n+1} = C_N^{N+n} U_N + \sum_{l=0}^{n} C_{N+l+1}^{N+n} A \delta \bar{W}_{n+l}. \tag{3.5}$$

Define

$$\begin{aligned} \bar{W}_j^m &= \delta \bar{W}_j + \cdots + \delta \bar{W}_m, \\ W_j^m &= \delta W_j + \cdots + \delta W_m. \end{aligned}$$

Then a summation by parts of (3.5) yields

$$U_{N+n+1} = C_N^{N+n} U_N + C_{N+1}^{N+n} A \bar{W}_N^{N+n} - \sum_{l=1}^{n} C_{N+l+1}^{N+n} H_{N+l} A \bar{W}_{N+l}^{N+n} \Delta t_{N+l}. \tag{3.6}$$

The estimate (3.4) will now be used heavily. By dividing the interval $[t_N, t_{N+n+1}]$ into subintervals of length ε_k (except for the last subinterval, which is $\leqq \varepsilon_k$) and using (3.4), we get that there is a sequence of real numbers $\delta_k \to 0$ such that if $\omega \notin D_k$ and $N \geqq M_k$, then

$$\begin{aligned} |U_{N+n+1}|_P \leqq (1 + \delta_k) \exp\left[-\frac{u}{2}(t_{N+n+1} - t_N)\right] \cdot |u_N|_P \\ + (1 + \delta_k) \exp\left[-\frac{u}{2}(t_{N+n+1} - t_N)\right] |A\bar{W}_N^{N+n}|_P \\ + (1 + \delta_k) \sum_{l=1}^{n} \exp\left[-\frac{u}{2}(t_{N+n+1} - t_{N+l})\right] \cdot \Delta t_{N+l} |H_{N+l} A \bar{W}_{N+l}^{N+n}|_P. \end{aligned} \tag{3.7}$$

Henceforth, *purely for notational convenience*, we suppose that the $\delta \bar{W}_i$ are scalar-valued. In general, we need only work with one component at a time anyway. Proceeding, let us next evaluate $E|W_m^M|^2$:

$$
\begin{aligned}
E|W_m^M|^2 &= E\sum_{i,j=m}^{M} \sqrt{\Delta t_i}\sqrt{\Delta t_j}\, h_i h_j \\
&\leqq 2\sum_{i=m}^{M} \sqrt{\Delta t_i} \sum_{j\geqq i}^{M} \sqrt{\Delta t_j}\, |Eh_i h_j| \\
&\leqq 2\sum_{i=m}^{M} \sqrt{\Delta t_i} \sum_{j\geqq i} \sqrt{\Delta t_j}\, |R_{ij}| \\
&\leqq 2K \sum_{i=m}^{M} \Delta t_i = 2K(t_{M+1}-t_m),
\end{aligned} \tag{3.8}
$$

where the last inequality follows by the first half of (A5). With perhaps a different K, the same inequality holds for $E|\bar{W}_m^M|^2$. By this estimate and the second half of (A5), there is a constant K_k such that for $N \geqq M_k$

$$E|H_{N+l} A \bar{W}_{N+l}^{N+n}|_p I_{\{\omega \notin D_k\}} \leqq K_k (t_{N+n+1} - t_{N+l})^{1/2}. \tag{3.9}$$

Inequality (3.7) holds with probability $1 - P\{D_k\} = \rho_k \to 1$. Let us modify the $\{U_i, H_i, i \geqq M_k\}$ on D_k in a way such that (3.7) holds for all n and (3.9) holds without the indicator function and where K_k does not depend on k. Let $\{U_i^k, H_i^k\}$ denote the altered sequence. Then (3.7) and (3.9) together imply that $\sup_{i \geqq M_k} E|U_i^k|^2 < \infty$. Thus the sequence $\{U_i, i < M_k;\ U_i^k, i \geqq M_k\}$ is tight on R^r. Since k is arbitrary and $\rho_k \to 1$ as $k \to \infty$, this implies that the original $\{U_i\}$ sequence is tight. Q.E.D.

4. Weak convergence of $\{U^N(\cdot)\}$ and the rate of convergence. In this section, we show that $\{U^N(\cdot)\}$ converges weakly in $D^r[0,\infty)$ to the stationary solution to the Gauss–Markov diffusion

$$dU = \bar{H} U\, dt + A R^{1/2}\, dB, \tag{4.1}$$

where $B(\cdot)$ is a standard Wiener process and $R^{1/2}$ is a square root of the matrix R in (A6) below. In particular, this implies that $(X_n - \theta)(n+1)^{\alpha/2}$ converges in distribution to a normal random variable with mean 0 and covariance

$$\int_0^\infty (\exp \bar{H}t) A R A' (\exp \bar{H}'t)\, dt.$$

We will require the following additional assumptions.

(A6) $\{h_j\}$ *is a stationary sequence, and* $E|h_j|^6 < \infty$. *Define* $R(i) = Eh_j h'_{j+i}$. *Then* $R \equiv \sum_{-\infty}^{\infty} R(i)$ *is bounded by* (A8).

Let $\mathcal{B}_j = \mathcal{B}(h_l, l \leqq j)$ and let E_j denote the expectation conditional on $\mathcal{B}_j$.

(A7) *Define* $\rho_1(i)$ *by*

$$\rho_1(i) = \sup_{j,l \geqq 0} E^{1/2} |E_j h_{j+i} h'_{j+i+l} - R(l)|^2.$$

Then $\sum_i \rho_1^{1/2}(i) < \infty$.

The $\sup_j$ above and $\sup_k$ below are redundant if we assume that the $\{h_j\}$ process started at $j = -\infty$, and choose the sample space appropriately.

(A8) *Define* $\rho_2(i)$ *by* $\rho_2(i) = \sup_k E^{1/2} |E_k h_{k+i}|^2$. *Then* $\sum_i \rho_2^{1/2}(i) < \infty$.

We now give some examples of (A7) and (A8). First suppose that $\{h_j\}$ is a stationary and bounded ϕ-mixing process in the sense of [7, p. 166], with of course $Eh_j \equiv 0$. Let K denote an arbitrary constant. By [8, Lemma 1], $|E_j h_{j+k}| \leqq K\phi_k$ and $|E_j h_{j+k} h'_{j+k+l} - R(l)| \leqq K\phi_k$. Thus $\rho_1(i) \leqq K\phi_i$, $\rho_2(i) \leqq K\phi_i$. If $\sum_l \phi_l^{1/2} < \infty$, then (A7) and (A8) hold. However, if the h_j are bounded and ϕ-mixing, then a slightly different proof of Theorem 2 can be given, requiring only $\sum_l \phi_l^{1/2} < \infty$.

An example of (A6) *to* (A8). Let Q denote a matrix whose eigenvalues are *inside* the unit circle, let $\{\psi_n\}$ denote a sequence of independent and identically distributed Gaussian random variables and define ξ_n, $\infty > n > -\infty$, by $\xi_{n+1} = Q\xi_n + \psi_n$. Then $\{\xi_n\}$ is a stationary sequence. Let $Eh(\theta, \xi_j) \equiv Eh_j = 0$ and suppose that $\bar{h}(\cdot) = h(\theta, \cdot)$ satisfies a uniform Lipschitz condition, with constant K_1. Let $\mathscr{G}_j$ measure ψ_i, $i < j$, and $E_k f = E_{\mathscr{G}_k} f$.

Let us evaluate $E|E_k \bar{h}(\xi_{k+i})|$. Let $\{\tilde{\psi}_i\}$ denote a sequence with the same distribution as $\{\psi_i\}$, but independent of it. We have

$$\xi_{k+i} = Q^i \xi_k + \sum_{l=0}^{i-1} Q^l \psi_{k+i-l-1}$$

which has the same distribution as

$$\sum_{l=0}^{\infty} Q^l \tilde{\psi}_l - \sum_{l=i}^{\infty} Q^l \tilde{\psi}_l + Q^i \xi_k.$$

Using the fact that the first term above has the same distribution as ξ_m has for any m, together with the Lipschitz condition, yields

$$\left| E\left[\bar{h}\left(\text{first term} - \sum_{l=1}^{\infty} Q^l \tilde{\psi}_l + Q^i \xi_k\right) - E\bar{h}(\text{first term}) | \xi_k \right] \right| \leqq K_1 E \left| \sum_{l=i}^{\infty} Q^l \tilde{\psi}_l \right| + K_1 |Q^i \xi_k|,$$

from which (A8) follows. A similar (and omitted) calculation yields (A7).

THEOREM 2. *Under* (A1)–(A8), $\{U^N(\cdot)\}$ *converges weakly to the stationary solution to* (4.1).

Proof. Part 1. Define the "approximation to a Wiener process" $W^N(\cdot)$ by

$$W^N(t) = W_N^{m(t_N+t)-1} = \sum_{i=N}^{m(t_N+t)-1} \sqrt{\Delta t_i}\, h_i,$$

with a similar definition for $\bar{W}^N(\cdot)$ (but using $\delta \bar{W}_i$ in lieu of δW_i). We will show that $\{W^N(\cdot)\}$ is tight in $D^r[0, \infty)$ and converges to a Wiener process with covariance matrix Rt. It easily follows from this that the same result must hold for $\{\bar{W}^N(\cdot)\}$, since $(n+2/n+1)^{\alpha/2} = 1 + O(1/n)$ implies that $\{|W^N(\cdot) - \bar{W}^N(\cdot)|\}$ tends weakly to the zero process.

First we prove *tightness* of $\{W^N(\cdot)\}$. For notational convenience only, we assume that the h_j are scalar-valued in this part of the proof. Otherwise, we would work with one component at a time anyway, so there is no loss of generality.

Let $l \geqq k \geqq j \geqq i$. We have

$$(4.2) \qquad |Eh_i h_j h_k h_l| \leqq |Eh_i h_j h_k h_l - Eh_i h_j Eh_k h_l| + |Eh_i h_j|\, |Eh_k h_l|.$$

The first term on the right satisfies (use (A7))

$$|Eh_i h_j (E_j h_k h_l - Eh_k h_l)| \leqq E^{1/2} |h_i h_j|^2 E^{1/2} |E_j h_k h_l - Eh_k h_l|^2 \leqq K\rho_1(k-j).$$

By (A8), the first term on the right of (4.2) is bounded above by

$$\begin{aligned}|Eh_ih_jh_kE_kh_l|+|Eh_ih_jEh_kE_kh_l| &\leqq E^{1/2}|h_ih_jh_k|^2E^{1/2}|E_kh_l|^2\\ &\quad+|Eh_ih_j|E^{1/2}h_k^2E^{1/2}|E_kh_l|^2\\ &\leqq K\rho_2(l-k).\end{aligned}$$

Thus

$$|Eh_ih_jh_kh_l|\leqq K\rho_1^{1/2}(k-j)\rho_2^{1/2}(l-k)+|R(j-i)|\,|R(l-k)|. \tag{4.3}$$

Using these bounds, we get

$$\begin{aligned}E|W^N(t+s)-W^N(t)|^4&=E\left|\sum_{i=m(t_N+t)}^{m(t_N+t+s)-1}\sqrt{\Delta t_i}\,h_i\right|^4\\ &\leqq K\sum_{i\leqq j\leqq k\leqq l}(\Delta t_i\;\Delta t_j\;\Delta t_k\;\Delta t_l)^{1/2}|Eh_ih_jh_kh_l|\end{aligned}$$

(summation between $m(t_N+t)$ and $m(t_N+t+s)-1$; at each use of K it may have a different value)

$$\begin{aligned}&\leqq K\sum_{i\leqq j\leqq k\leqq l}(\Delta t_i\;\Delta t_j\;\Delta t_k\;\Delta t_l)^{1/2}[\rho_1^{1/2}(k-j)\rho_2^{1/2}(l-k)\\ &\qquad+|R(j-i)|\cdot|R(l-k)|],\end{aligned}$$

(sum over l and use $\Delta t_k\geqq\Delta t_l$)

$$\leqq K\sum_{i\leqq j\leqq k}(\Delta t_i\;\Delta t_j)^{1/2}\,\Delta t_k[\rho_1^{1/2}(k-j)+|R(j-i)|]$$

(sum over j and use $\Delta t_i\geqq\Delta t_j$)

$$\leqq K\sum_{i\leqq k}\Delta t_i\;\Delta t_k\leqq Ks^2 \tag{4.4}$$

where the last inequality holds if t_N+t+s and t_N+t take values in the set $\{t_i\}$.

If (4.4) holds for all t, s, N, then [7, Thms. 15.5 and 12.3] implies that $\{W^N(\cdot)\}$ is tight in $D^r[0,\infty)$ and that all processes which are weak limits have continuous paths w.p. 1. But, since $\Delta t_n\to0$ and the paths are piecewise constant, it is enough that (4.4) hold for t_N+t+s and t_N+t in the $\{t_i\}$ set. Thus $\{W^N(\cdot)\}$ is tight and all limit processes have continuous paths w.p. 1.

Part 2. Now, the h_i are treated as vectors rather than scalars. Let N index a weakly convergent subsequence of $\{W^N(\cdot)\}$ and denote the (continuous w.p. 1) weak limit by $W(\cdot)$. Note that (4.4) implies that $\{|W^N(\cdot)|^2\}$ is uniformly integrable. Let $s_i\leqq t\leqq t+s$ and q be arbitrary. Let $g(\cdot)$ denote a bounded continuous function of $W^N(s_i)$, $i\leqq q$, and let E_t^N denote expectation conditioned on $\{h_j,\,j\leqq m(t_N+t)-1\}$. Then

$$Eg(W^N(s_i),\,i\leqq q)[W^N(t+s)-W^N(t)]=Eg(W^N(s_i),\,i\leqq q)E_t^N\sum_{i=m(t_N+t)}^{i=m(t_N+t+s)-1}\sqrt{\Delta t_i}\,h_i$$

goes to zero as $N\to\infty$ by (A8). This together with the uniform integrability and weak convergence imply that $Eg(W(s_i),\;i\leqq q)[W(t+s)-W(t)]=0$ for all q, bounded continuous g and $s_i\leqq t\leqq t+s$. Thus $W(\cdot)$ is a continuous martingale. To compute its quadratic variation, repeat the above argument with $[W^N(t+s)-W^N(t)][W^N(t+s)-$

$W^N(t)]'$ replacing $[W^N(t+s)-W^N(t)]$. Using (A6), the weak convergence and uniform integrability yields

$$Eg(W^N(s_i), i \leqq q)[W^N(t+s)-W^N(t)][W^N(t+s)-W^N(t)]' \to Eg(W(s_i), i \leqq q)Rs.$$

Then the arbitrariness of g and $s_i \leqq t \leqq t+s$ yield that the quadratic variation (at s) is Rs. Thus $W(\cdot)$ is a Wiener process with covariance Rs, as asserted. This result does not depend on the chosen convergent subsequence.

Part 3. Define the function $C^n(t, t+s) = C_{m(t_N+t)}^{m(t_N+t+s)-1}$. Define a function H^N with values $H_t^N = H_{N+n}$ in $[t_{N+n}-t_N, t_{n+N+1}-t_N)$, not to be confused with $H^N(\cdot)$ below.

Then for $t \in \{t_{N+i}-t_N, i \geqq 0\}$, and modulo a factor for each term which goes to zero uniformly in t w.p. 1 as $N \to \infty$, the sum (3.6) can be written in the integral form (since the integrand is constant over Δt_i intervals)

$$\begin{aligned} U^N(t) &= C^N(0,t)U^N(0) + C^N(0,t)A\bar{W}^N(t) \\ &\quad - \int_0^t C^N(s,t)H_s^N A[\bar{W}^N(t) - \bar{W}^N(s)]\,ds, \end{aligned} \tag{4.5}$$

for $t > 0$. Between the $\{t_i\}$, the integral in (4.5) is just a linear interpolation instead of a piecewise constant interpolation of the sum in (3.6), and we may work with it instead. Define $H^N(\cdot)$ by $N^N(t) = \sum_{i=m(t_N)}^{m(t_N+t)-1} H_i \, \Delta t_i$. By (A3), $\{H^N(\cdot)\}$ is tight in $D^r[0,\infty)$ and all limits are the *constant* process with value $\bar{H}t$ at t. Note that $\{C^N(0,t)\}$ is tight on $D^q[0,\infty)$ for an appropriate integer q, since it converges to $\exp \bar{H}t$ uniformly on bounded intervals w.p. 1.

We now have essentially all the limits that are required. If H_s^N converged to the constant $\bar{H}$ w.p. 1 as $N \to \infty$, then the weak convergence of $\bar{W}^N(\cdot)$ and convergence of $C^N(s,t)$ would imply that (4.5) holds with all functions replaced by their limits (and a weakly convergent subsequence of $\{U^N(0)\}$ taken). Since H_s^N does not usually converge in the above sense, a slightly indirect method must be used to allow us to make the replacements suggested above. It is convenient to have all the random functions defined on the same space and to work with w.p. 1 rather than with weak convergence. To do this we apply the imbedding technique of Skorokhod [9, Thm. 3.1.1]. The family $\{U^N(0), H^N(\cdot), \bar{W}^N(\cdot), C^N(0,\cdot)\} = \{\Phi^N(\cdot)\}$ is tight in the appropriate space $R^r \times D^{2q+r}[0,\infty) \equiv \mathscr{D}$ and all limit functions are continuous w.p. 1. Extract a convergent subsequence, index it by N, and denote the limit by $(U(0), \bar{H}(\cdot), W(\cdot), C(0,\cdot)) \equiv \Phi(\cdot)$. By the Skorokhod imbedding method [9, Thm. 3.1.1], there exists a probability space $(\tilde{\Omega}, \tilde{P}, \tilde{B})$ with random processes $\{\tilde{U}^N(0), \tilde{H}^N(\cdot), \tilde{W}^N(\cdot), \tilde{C}^N(0,\cdot)\} \equiv \{\tilde{\Phi}^N(\cdot)\}$ and $(\tilde{U}(0), \tilde{\bar{H}}(\cdot), \tilde{W}(\cdot), \tilde{C}(0,\cdot)) \equiv \tilde{\Phi}(\cdot)$ defined on it, where $\tilde{\Phi}^N(\cdot)$ (resp., $\tilde{\Phi}(\cdot)$) has the same distribution as $\Phi^N(\cdot)$ (resp., $\Phi(\cdot)$), all the processes in $\tilde{\Phi}(\cdot)$ have continuous paths and $\tilde{\Phi}^N(\cdot) \to \tilde{\Phi}(\cdot)$ w.p. 1 in the topology of $\mathscr{D}$. Since the limit processes are continuous, this means uniform convergence on bounded intervals. From $\tilde{H}^N(\cdot)$, we can recover the random variables $\tilde{H}_{N+i}$, $i \geqq 0$, from which it was constructed, since $\tilde{H}^N(\cdot)$ is also piecewise constant w.p. 1. Also $\{\tilde{H}_{N+i}, i \geqq 0\}$ has the same distribution as has $\{H_{N+i}, i \geqq 0\}$.

We work with the imbedded processes, but drop the tilde affix. Now, return to (4.5) and, via the imbedding, suppose that all weak convergences are w.p. 1 in the above-cited topology. The first two terms of (4.5) converge to $(\exp \bar{H}t)U(0)$ and $(\exp \bar{H}t)W(t)$, respectively. Note that $C^N(s,t) = C^N(0,t)[C^N(0,s)]^{-1}$ also converges w.p. 1 uniformly on bounded sets to $\exp \bar{H}(t-s)$. We next write the integral in (4.5) in a more convenient way.

Let $\Delta > 0$, and let $M = \max\{i: i\Delta \leqq t\}$. We have

$$\sum_{i=0}^{M-1} \left| \int_{i\Delta}^{i\Delta+\Delta} \{C^N(s,t)H_s^N A[W^N(t) - W^N(s)] - C(i\Delta, t)H_s^N A[W(t) - W(i\Delta)]\} \, ds \right|$$
$$+ \left| \int_{M\Delta}^{t} \{C^N(s,t)H_s^N A[W^N(t) - W^N(s)] - C(M, t)H_s^N A[W(t) - W(M\Delta)]\} \, ds \right|$$
$$\leqq \sum_{i=0}^{M-1} \sup_{i\Delta \leqq s \leqq i\Delta+\Delta} [|C^N(s,t) - C(i\Delta, t)| + |W^N(s) - W(i\Delta)| + |W(t) - W^N(t)|]$$
$$\cdot [|W^N(t) - W^N(s)| + |C(i\Delta, t)|] \int_{i\Delta}^{i\Delta+\Delta} |H_s^N| \, ds |A| \text{ plus a similar expression for the end term.}$$

By the w.p. 1 uniform convergence (on bounded intervals) and continuity of the limit functions and the estimate (A3b), the limit of the above expression goes to zero uniformly on bounded t sets, w.p. 1, as $N \to \infty$ and then $\Delta \to 0$.

Thus, we need only examine the limits of

$$\sum_{i=0}^{M-1} \int_{i\Delta}^{i\Delta+\Delta} C(i\Delta, t)H_s^N A[W(t) - W(i\Delta)] \, ds$$
$$(4.6) \qquad + \int_{M\Delta}^{t} C(M\Delta, t)H_s^N A[W(t) - W(M\Delta)] \, ds.$$

But, by (A3a), (4.6) converges to the same expression with $\bar{H}$ replacing H_s^N, uniformly on bounded intervals, w.p. 1 as $N \to \infty$. By the above calculations we can write the limit of the third term in (4.5) as

$$(4.7) \qquad -\int_0^t C(s,t)\bar{H}A[W(t) - W(s)] \, ds$$

for the imbedded, hence the original processes. Thus $U^N(t)$ (the imbedded process) converges to

$$(4.8) \qquad U(t) \equiv C(0,t)U(0) + C(0,t)AW(t) + (4.7)$$

uniformly on finite intervals, w.p. 1. Consequently the original $U^N(\cdot)$ converges weakly to the process (4.8). But (4.8) is the unique solution to (4.1) with initial condition $U(0)$. The form is independent of the selected convergent subsequences. Also, via an integration by parts,

$$(4.9) \qquad U(t) = C(0,t)U(0) + \int_0^t C(s,t)A \, dW_s.$$

We need only show that $U(0)$ is the "stationary" initial condition. This can be easily shown in the following manner. The set of all possible $U(0)$ is tight because $\{U_n\}$ is. Also the weak limits of $\{U^N(\cdot)\}$ are also weak limits of the restrictions to $[T, \infty)$ of the weak limits of (the functions are left-shifted by T) $\{U^{m(t_N - T)}(\cdot)\}$ on $D^r[0, \infty)$, since $U^{m(t_N-T)}(T) = U_N$. But the latter limits are of the form (4.9) also. The restriction to $[T, \infty)$ involves simply replacing t by $T + t$ in (4.9). From this, the tightness of possible $U(0)$, the arbitrariness of T and the fact that $C(0,t) = \exp \bar{H}t \to 0$ as $t \to \infty$, we get that $U(0)$ must be the "stationary" initial condition. Q.E.D.

Note. In a work to appear, only $X_n^P \to \theta$ is used, conditions for this are given and $a_n = a$ (small) and nonstationary data are treated.

REFERENCES

[1] J. SACKS, *Asymptotic distribution of stochastic approximation procedures*, Ann. Math. Statist., 29 (1958), pp. 273–405.

[2] V. FABIAN, *On asymptotic normality in stochastic approximation*, Ibid., 39 (1968), pp. 1327–1332.

[3] H. J. KUSHNER, *Rates of convergence for sequential Monte-Carlo optimization methods*, this Journal, 16 (1978), pp. 150–168.

[4] H. J. KUSHNER AND D. S. CLARK, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Applied Math. Sci. Series no. 26 (1978), Springer-Verlag, Berlin.

[5] L. LJUNG, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automatic Control, AC–22 (1977), pp. 551–575.

[6] ———, *On positive real transfer functions and the convergence of some recursive schemes*, Ibid., AC–22 (1977), pp. 539–550.

[7] P. BILLINGSLEY, *Convergence of Probability Measures*, Wiley, New York, 1968.

[8] G. C. PAPANICOLAOU AND W. KOHLER, *Asymptotic theory of mixing stochastic processes*, Comm. Pure Appl. Math., 27 (1974), pp. 641–668.

[9] A. V. SKOROKHOD, *Limit theorems for stochastic processes*, Theory Probability Appl., 1 (1956), pp. 262–290.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 5, September 1979

0363-0129/79/1705-0005$01.00/0

A NEW CLASS OF AUGMENTED LAGRANGIANS IN NONLINEAR PROGRAMMING*

G. DI PILLO† AND L. GRIPPO‡

Abstract. In this paper a new class of augmented Lagrangians is introduced, for solving equality constrained problems via unconstrained minimization techniques. It is proved that a solution of the constrained problem and the corresponding values of the Lagrange multipliers can be found by performing a single unconstrained minimization of the augmented Lagrangian. In particular, in the linear quadratic case, the solution is obtained by minimizing a quadratic function. Numerical examples are reported.

1. Introduction. During recent years, a number of research works in the area of nonlinear programming have been devoted to the study of methods for solving constrained problems of the form:

$$\text{minimize } f(x) \quad \text{subject to } g(x)=0,$$

via unconstrained minimization techniques.

The most recent results are concerned with the "method of multipliers", which was independently introduced in 1968 by Hestenes [1] and Powell [2]. We refer, e.g. to [3]–[6] for an introduction to this method and for an exposition of related refinements and extensions.

As it is known, the method of multipliers provides the solution of the constrained problem via the solution of a sequence of unconstrained problems of the form:

$$\min_x f(x)+[\lambda, g(x)]+c\|g(x)\|^2,$$

where $c>0$ is a penalty coefficient and λ is an approximation of the Lagrange multiplier.

The relevant feature of the method is that, under suitable assumptions, the solution of the constrained problem is obtained by recursively updating λ, without the need to increase c to infinity. Thus the ill-conditioning associated with the usual penalty methods can be avoided.

The main drawback of the multiplier method is that, in principle, it requires an infinite sequence of unconstrained minimization problems to be solved.

To overcome this difficulty, a further development of the method was proposed by Fletcher [7], [8], who introduced in the augmented Lagrangian a multiplier vector continuously dependent on x. In this way a single minimization is required, as opposed to a sequence of minimizations required in the multiplier method. A related algorithm was proposed and analyzed by Mukai and Polak [9]. These methods, however, require a matrix inversion at each function evaluation and this may limit somewhat their applicability.

A different possibility was considered by Wierzbicki [10], who devised several algorithms for locating directly the saddle-point of the augmented Lagrangian by simultaneous updating of x and λ and without resorting to matrix inversions.

* Received by the editors April 5, 1978 and in revised form January 17, 1979. This work was supported in part by the Consiglio Nazionale delle Ricerche.

† Istituto di Automatica—Università di Roma, Via Eudossiana, 18-00184 Roma, Italy.

‡ Centro di Studio dei Sistemi di Controllo e Calcolo Automatici del CNR, Via Eudossiana, 18-00184 Roma, Italy.

In this paper, we propose a different approach based on the consideration of a new class of augmented Lagrangians obtained by adding to the augmented Lagrangian of Hestenes a penalty term on the first order necessary condition $\nabla_x f+[\partial g/\partial x]^T\lambda=0$.

This leads to a function of the form:

$$S(x,\lambda;c)=f(x)+[\lambda,g(x)]+c\|g(x)\|^2+\left\|M(x)\left(\nabla_x f(x)+\frac{\partial g(x)^T}{\partial x}\lambda\right)\right\|^2,$$

where $M(x)$ is an appropriate weighting matrix.

It is shown that, under suitable hypotheses, a local solution of the constrained problem and the corresponding values of the Lagrange multipliers can be found by performing a single local unconstrained minimization of $S(x,\lambda;c)$ with respect to both x and λ, for finite values of the penalty coefficient and without requiring matrix inversions. In particular, in the linear quadratic case there exists a value of c for which $S(x,\lambda;c)$ is a positive definite quadratic function of (x,λ), so that the solution of a quadratic programming problem with equality constraints can be obtained in a finite number of iterations of a conjugate direction algorithm.

The proposed method has been tested on several problems, with quite satisfactory results. Although the primary emphasis of this paper is not on numerical aspects, we report here a set of numerical examples showing that the method appears to be promising.

2. Problem formulation. We consider the following minimization problem:

Problem P.

$$\text{minimize } f(x), \qquad x\in R^n \quad \text{subject to } g(x)=0 \tag{1}$$

where $f\colon R^n\to R^1$ and $g\colon R^n\to R^m$, with $m\leqq n$. We assume, unless otherwise stated, that the functions f and g are three times continuously differentiable on R^n.

The Lagrangian $L\colon R^n\times R^m\to R^1$ for problem P is defined by

$$L(x,\lambda)=f(x)+[\lambda,g(x)],$$

where $[\cdot,\cdot]$ denotes the Euclidean scalar product.

We introduce the following augmented Lagrangian function

$$S(x,\lambda;c)=f(x)+[\lambda,g(x)]+c\|g(x)\|^2+\left\|M(x)\left(\nabla f(x)+\frac{\partial g(x)^T}{\partial x}\lambda\right)\right\|^2 \tag{2}$$

where $c>0$, $M(x)$ is a $(p\times n)$ matrix with twice continuously differentiable elements and $m\leqq p\leqq n$.

To simplify notation, we shall denote by $\nabla_x L(x,\lambda)$ the gradient and by $\nabla_x^2 L(x,\lambda)$ the Hessian of $L(x,\lambda)$ with respect to x, i.e.:

$$\nabla_x L(x,\lambda)\triangleq\nabla f(x)+\frac{\partial g(x)^T}{\partial x}\lambda,$$

$$\nabla_x^2 L(x,\lambda)\triangleq\nabla^2 f(x)+\sum_{i=1}^m\lambda_i\nabla^2 g_i(x).$$

3. Preliminary results. In the sequel we shall make use of the following properties, which establish the relationships between stationary points of $L(x,\lambda)$ and stationary points of $S(x,\lambda;c)$, under the assumption that f, g are two times continuously differentiable and that $M(x)$ is a continuously differentiable matrix.

PROPOSITION 1. *Let $(\bar{x}, \bar{\lambda})$ be a stationary point for $L(x, \lambda)$, then $(\bar{x}, \bar{\lambda})$ is a stationary point for $S(x, \lambda; c)$.*

Proof. Employing a dyadic expansion for $M(x)$, that is

$$M(x) = \sum_{j=1}^{p} e_j m_j(x),$$

where e_j is the jth column of the $(p \times p)$ identity matrix and $m_j(x)$ is the jth row of $M(x)$, we obtain the following expressions for the components of the gradient of $S(x, \lambda; c)$ on $R^n \times R^m$:

$$\begin{aligned}\nabla_x S(x, \lambda; c) = \nabla_x L(x, \lambda) + 2c \frac{\partial g(x)^T}{\partial x} g(x) + 2\nabla_x^2 L(x, \lambda) M^T(x) M(x) \nabla_x L(x, \lambda) \\ + 2\left[\sum_{j=1}^{p} \left(\frac{\partial m_j^T(x)}{\partial x}\right)^T \nabla_x L(x, \lambda)\, e_j^T\right] M(x) \nabla_x L(x, \lambda),\end{aligned} \tag{3}$$

$$\nabla_\lambda S(x, \lambda; c) = g(x) + 2 \frac{\partial g(x)}{\partial x} M^T(x) M(x) \nabla_x L(x, \lambda). \tag{4}$$

Therefore, $\nabla L(\bar{x}, \bar{\lambda}) = 0$ and $g(\bar{x}) = 0$ imply $\nabla_x S(\bar{x}, \bar{\lambda}; c) = 0$, $\nabla_\lambda S(\bar{x}, \bar{\lambda}; c) = 0$. □

PROPOSITION 2. *Let $(\bar{x}, \bar{\lambda})$ be a stationary point for $S(x, \lambda; c)$ and assume that $g(\bar{x}) = 0$ and that $M(\bar{x})\ [\partial g(\bar{x})/\partial x]^T$ is an $(m \times m)$ nonsingular matrix. Then $(\bar{x}, \bar{\lambda})$ is a stationary point for $L(x, \lambda)$.*

Proof. Under the hypotheses stated, $\nabla_\lambda S(\bar{x}, \bar{\lambda}; c) = 0$ and $g(\bar{x}) = 0$ imply $M(\bar{x}) \nabla_x L(\bar{x}, \bar{\lambda}) = 0$ so that from $\nabla_x S(\bar{x}, \bar{\lambda}; c) = 0$ we get $\nabla_x L(\bar{x}, \bar{\lambda}) = 0$. □

PROPOSITION 3. *Let $X \times L$ be a compact subset of $R^n \times R^m$ and assume that $M(x)\ [\partial g(x)/\partial x]^T$ is an $(m \times m)$ nonsingular matrix for any $x \in X$. Then, there exists a $\bar{c} > 0$ such that for all $c \geqq \bar{c}$, if $(\bar{x}, \bar{\lambda}) \in X \times L$ is a stationary point of $S(x, \lambda; c)$, $(\bar{x}, \bar{\lambda})$ is also a stationary point of $L(x, \lambda)$.*

Proof. Let $(\bar{x}, \bar{\lambda}) \in X \times L$ be a stationary point of $S(x, \lambda; c)$. Then, by (4), $\nabla_\lambda S(\bar{x}, \bar{\lambda}; c) = 0$ implies

$$M(\bar{x}) \nabla_x L(\bar{x}, \bar{\lambda}) = -\frac{1}{2}\left[\frac{\partial g(\bar{x})}{\partial x} M^T(\bar{x})\right]^{-1} g(\bar{x}).$$

Therefore, making use of (3) and recalling that $\nabla_x S(\bar{x}, \bar{\lambda}; c) = 0$, we have:

$$\begin{aligned}0 = M(\bar{x}) \nabla_x S(\bar{x}, \bar{\lambda}; c) = \Bigg[-\frac{1}{2}\left[\frac{\partial g(\bar{x})}{\partial x} M^T(\bar{x})\right]^{-1} + 2c M(\bar{x}) \frac{\partial g(\bar{x})^T}{\partial x} \\ - M(\bar{x}) \left(\nabla_x^2 L(\bar{x}, \bar{\lambda}) M^T(\bar{x}) + \sum_{j=1}^{m} \left[\frac{\partial m_j^T(\bar{x})}{\partial x}\right]^T \nabla_x L(\bar{x}, \bar{\lambda})\, e_j^T\right) \left[\frac{\partial g(\bar{x})}{\partial x} M^T(\bar{x})\right]^{-1}\Bigg] g(\bar{x}).\end{aligned}$$

Hence, by the continuity assumptions and the compactness of $X \times L$, there exists a $\bar{c} > 0$ such that for all $c \geqq \bar{c}$ and any $(\bar{x}, \bar{\lambda}) \in X \times L$ the matrix multiplying $g(\bar{x})$ is nonsingular, so that for $c \geqq \bar{c}$, $g(\bar{x}) = 0$. Then, the proof can be completed as in Proposition 2. □

4. Local optimality results. In order to establish a relationship between local minimum points of (1) and local unconstrained minimum points of S in $R^n \times R^m$ we need a known result on pairs of quadratic forms.

LEMMA 1. *Suppose that P and Q are quadratic forms with the property that $P(y) \leqq 0$ and $Q(y) \leqq 0$ only if $y = 0$. If one of them is nonnegative, then there is a number $c > 0$ such* *that*

$$P(y)+cQ(y)>0,$$

for all $y \neq 0$.

Proof. See [6, p. 113]. □

Then we can prove the following:

THEOREM 1. *Let* $(\bar{x}, \bar{\lambda})$ *be a stationary point for* $L(x, \lambda)$ *and assume that*

(i) $M(\bar{x})\ [\partial g(\bar{x})/\partial x]^T$ *has full rank*

(ii) $\bar{x}$ *is a local minimum point for Problem* P, *satisfying the second order sufficiency condition*:

$$[x, \nabla_x^2 L(\bar{x}, \bar{\lambda})x]>0 \qquad \forall x: x \neq 0,\ \frac{\partial g(\bar{x})}{\partial x}x=0.$$

Then there exists a $c^*>0$ *such that for any* $c \geqq c^*$, $(\bar{x}, \bar{\lambda})$ *is an isolated local minimum point for* $S(x, \lambda; c)$.

Proof. By Proposition 1, $(\bar{x}, \bar{\lambda})$ is a stationary point for $S(x, \lambda; c)$. Consider the Hessian matrix of $S(x, \lambda; c)$ evaluated at $(\bar{x}, \bar{\lambda})$. Since $\nabla_x L(\bar{x}, \bar{\lambda})=0$ and $g(\bar{x})=0$ we have:

$$\nabla_x^2 S(\bar{x}, \bar{\lambda}; c)=\nabla_x^2 L(\bar{x}, \bar{\lambda})+2c\frac{\partial g(\bar{x})^T}{\partial x}\frac{\partial g(\bar{x})}{\partial x}+2\nabla_x^2 L(\bar{x}, \bar{\lambda})M^T(\bar{x})M(\bar{x})\nabla_x^2 L(\bar{x}, \bar{\lambda}),$$

$$\nabla_\lambda^2 S(\bar{x}, \bar{\lambda}; c)=2\frac{\partial g(\bar{x})}{\partial x}M^T(\bar{x})M(\bar{x})\frac{\partial g(\bar{x})^T}{\partial x},$$

$$\nabla_{x\lambda}^2 S(\bar{x}, \bar{\lambda}; c)=\frac{\partial g(\bar{x})^T}{\partial x}+2\nabla_x^2 L(\bar{x}, \bar{\lambda})M^T(\bar{x})M(\bar{x})\frac{\partial g(\bar{x})^T}{\partial x}.$$

Introduce now the quadratic forms in (x, λ):

$$P(x, \lambda)=[x, \nabla_x^2 L(\bar{x}, \bar{\lambda})x]+2\left\|M(\bar{x})\nabla_x^2 L(\bar{x}, \bar{\lambda})x+M(\bar{x})\frac{\partial g(\bar{x})^T}{\partial x}\lambda\right\|^2+2\left[x, \frac{\partial g(\bar{x})^T}{\partial x}\lambda\right],$$

$$Q(x, \lambda)=2\left\|\frac{\partial g(\bar{x})}{\partial x}x\right\|^2.$$

It can be easily verified that

$$[x^T \quad \lambda^T]\begin{bmatrix}\nabla_x^2 S(\bar{x}, \bar{\lambda}; c) & \nabla_{x\lambda}^2 S(\bar{x}, \bar{\lambda}; c)\\ \nabla_{\lambda x}^2 S(\bar{x}, \bar{\lambda}; c) & \nabla_\lambda^2 S(\bar{x}, \bar{\lambda}; c)\end{bmatrix}\begin{bmatrix}x\\ \lambda\end{bmatrix}=P(x, \lambda)+cQ(x, \lambda),$$

so that $(\bar{x}, \bar{\lambda})$ is an isolated local minimum of $S(x, \lambda; c)$ if $P(x, \lambda)+cQ(x, \lambda)$ is positive definite.

We observe now that the assumptions of Lemma 1 are satisfied. In fact $Q(x, \lambda) \geqq 0$; moreover $Q(x, \lambda)=0$ implies $[\partial g(\bar{x})/\partial x]x=0$ so that, taking into account assumption (ii), $P(x, \lambda) \leqq 0$ implies $x=0$ and $M(\bar{x})\ [\partial g(\bar{x})/\partial x]^T\lambda=0$. Finally, by (i) the last equality gives $\lambda=0$.

Then, by Lemma 1, there exists a value $c^*>0$ such that $P(x, \lambda)+c^*Q(x, \lambda)>0$ $\forall (x, \lambda) \neq 0$ and being $Q(x, \lambda) \geqq 0$ the same is true for any $c \geqq c^*$. □

A converse result is given in the following theorem.

THEOREM 2. *Let* $(\bar{x}, \bar{\lambda})$ *be a local minimum point for* $S(x, \lambda; c)$ *and assume that*:

(i) $g(\bar{x})=0$,

(ii) $p=m$, $M(\bar{x})[\partial g(\bar{x})/\partial x]^T$ *is nonsingular*,

(iii) *the Hessian matrix* $\nabla^2 S(\bar{x}, \bar{\lambda}; c)$ *is positive definite*.

Then $\bar{x}$ is an isolated local minimum point for Problem P.

Proof. By Proposition 2, $(\bar{x}, \bar{\lambda})$ is a stationary point for $L(x, \lambda)$; therefore, since $\nabla^2 S(\bar{x}, \bar{\lambda}; c)$ is positive definite, we have:

$$[x, \nabla_x^2 L(\bar{x}, \bar{\lambda})x] + 2\left[x, \frac{\partial g(\bar{x})^T}{\partial x}\lambda\right] + 2c\left\|\frac{\partial g(\bar{x})}{\partial x}x\right\|^2$$

$$+2\left\|M(\bar{x})\nabla_x^2 L(\bar{x}, \bar{\lambda})x + M(\bar{x})\frac{\partial g(\bar{x})^T}{\partial x}\lambda\right\|^2 > 0 \qquad \forall (x, \lambda) \neq 0. \tag{5}$$

Now let $x \neq 0$ be such that $[\partial g(\bar{x})/\partial x]x = 0$ and take:

$$\lambda = -\left[M(\bar{x})\frac{\partial g(\bar{x})^T}{\partial x}\right]^{-1} M(\bar{x})\nabla_x^2 L(\bar{x}, \bar{\lambda})x.$$

Then we obtain from (5) that $(\bar{x}, \bar{\lambda})$ satisfies the second order sufficiency condition:

$$[x, \nabla_x^2 L(\bar{x}, \bar{\lambda})x] > 0 \qquad \forall x : x \neq 0, \frac{\partial g(\bar{x})}{\partial x}x = 0$$

so that $\bar{x}$ is an isolated local minimum point for Problem P. □

We note that a local result can also be stated under assumptions weaker than those employed in Theorem 2.

THEOREM 3. *Let f, g be two times continuously differentiable and let $(\bar{x}, \bar{\lambda})$ be a local minimum point for $S(x, \lambda; c)$. Assume that*

(i) $g(\bar{x}) = 0$;

(ii) *$M(x)$ is a continuously differentiable $(m \times n)$ matrix such that $M(\bar{x})[\partial g(\bar{x})/\partial x]^T$ is nonsingular.*

Then $\bar{x}$ is a local minimum point for problem P.

Proof. By Proposition 2, $(\bar{x}, \bar{\lambda})$ is a stationary point for $L(x, \lambda)$. This implies:

$$S(\bar{x}, \bar{\lambda}; c) = f(\bar{x}).$$

Therefore, since $(\bar{x}, \bar{\lambda})$ is a local minimum point for $S(x, \lambda; c)$, there exist neighborhoods Ω, Λ of $\bar{x}$, $\bar{\lambda}$, such that

$$f(\bar{x}) \leqq S(x, \lambda; c) \qquad \forall\, x \in \Omega, \quad \lambda \in \Lambda.$$

This yields

$$f(\bar{x}) \leqq f(x) + \left\|M(x)\nabla f(x) + M(x)\frac{\partial g(x)^T}{\partial x}\lambda\right\|^2$$

$$\forall\, x \in \Omega \cap \{x : g(x) = 0\} \tag{6}$$

$$\forall\, \lambda \in \Lambda.$$

On the other hand, by the continuity assumptions, there exists a neighborhood Ω' of $\bar{x}$, $\Omega' \subseteq \Omega$, such that

$$\lambda = -\left[M(x)\frac{\partial g(x)^T}{\partial x}\right]^{-1} M(x)\nabla f(x) \in \Lambda, \qquad \forall\, x \in \Omega'. \tag{7}$$

By combining (6), (7) it can be concluded

$$f(\bar{x}) \leqq f(x) \qquad \forall\, x \in \Omega' \cap \{x : g(x) = 0\}. \qquad \square$$

Under the assumptions stated in Proposition 3 it is also possible to ensure that a local minimum point for $S(x, \lambda; c)$ is an admissible point for Problem P.

Therefore, we obtain the following theorem:

THEOREM 4. *Let f, g be two times continuously differentiable and let $X \times L$ be a compact subset of $R^n \times R^m$. Assume that $M(x)$ is a continuously differentiable $(m \times n)$ matrix such that $M(x)[\partial g(x)/\partial x]^T$ is nonsingular for any $x \in X$. Then, there exists a $\bar{c} > 0$ such that for all $c \geqq \bar{c}$, if $(\bar{x}, \bar{\lambda}) \in X \times L$ is a local minimum point of $S(x, \lambda; c)$, $\bar{x}$ is a local minimum point for problem* P.

Proof. The proof follows from Proposition 3 and Theorem 3. □

5. Choice of $M(x)$. In the preceding theorems an important role is played by the matrix $M(x)$. We indicate here some possible choices for $M(x)$ which ensure, under suitable assumptions on the originary problem, that the hypotheses made on $M(\bar{x})$ $[\partial g(\bar{x})/\partial x]^T$ are satsified.

In particular, assume that $[\partial g(\bar{x})/\partial x]^T$ has full rank; then

(a) the choice:

$$M(x) = \mu \frac{\partial g(x)}{\partial x}, \qquad \mu > 0,$$

satisfies the hypotheses of Theorems 1, 2, 3, 4;

(b) the same is true for any choice of $M(x)$ such that $M(x)[\partial g(x)/\partial x]^T$ is an $m \times m$ invertible submatrix of $[\partial g(x)/\partial x]^T$;

(c) the choice:

$$M(x) = \mu I, \qquad \mu > 0,$$

satisfies the hypotheses of Theorem 1.

An important special case of (b) is when the vector x can be split into two subvectors, x_1, x_2, $x_1 \in R^m$, $x_2 \in R^{n-m}$, such that $\partial g(x)/\partial x_1$ is nonsingular for any x. This happens when there exists a set of independent or "decision" variables x_2 and a set of dependent or "state" variables x_1. In such a case a convenient choice for $M(x)$ could be:

$$M(x) = \mu[I_m \mid 0].$$

6. Global optimality results. The results given in § 4 are local in character. A global result is obtained when Problem P has a unique global solution $\bar{x}$ on a compact set X and $\bar{x}$ is an interior point of X:

THEOREM 5. *Let $(\bar{x}, \bar{\lambda})$ be a stationary point for $L(x, \lambda)$ and assume that:*

(i) *assumptions* (i) *and* (ii) *of Theorem* 1 *are satisfied*

(ii) *$\bar{x}$ is the unique global minimum point of Problem* P *on a compact set $X \subseteq R^n$ and $\bar{x} \in \operatorname{int}(X)$.*

Then, for every compact set $L \subseteq R^m$ such that $\bar{\lambda} \in \operatorname{int}(L)$, there exists a $c^*(L) > 0$ such that for any $c \geqq c^*(L)$, $(\bar{x}, \bar{\lambda})$ is the unique global minimum point of $S(x, \lambda; c)$ on $X \times L$.

Proof. Let $L \subseteq R^m$ be a compact set such that $\bar{\lambda} \in \operatorname{int}(L)$. By Theorem 1, there exists a $c_1^* > 0$ such that for $c \geqq c_1^*$, $(\bar{x}, \bar{\lambda})$ affords a local isolated minimum to $S(x, \lambda; c)$. Therefore, since $\bar{x} \in \operatorname{int}(X)$ and $\bar{\lambda} \in \operatorname{int}(L)$ there exist, for $c \geqq c_1^*$, spherical neighborhoods $\Omega(\bar{x}, \varepsilon_c)$, $\Lambda(\bar{\lambda}, \varepsilon_c)$ of $\bar{x}$ and $\bar{\lambda}$ such that $\Omega(\bar{x}, \varepsilon_c) \subseteq X$, $\Lambda(\bar{\lambda}, \varepsilon_c) \subseteq L$ and

$$S(x, \lambda; c) > S(\bar{x}, \bar{\lambda}; c) \qquad \forall (x, \lambda) \in \Omega(\bar{x}, \varepsilon_c) \times \Lambda(\bar{\lambda}, \varepsilon_c), \quad (x, \lambda) \neq (\bar{x}, \bar{\lambda}).$$

Assume now that the conclusion of the theorem is false. Then, for any integer $k \geqq c_1^*$ there exists $(x_k, \lambda_k) \in X \times L$ such that:

$$S(x_k, \lambda_k; k) \leqq S(\bar{x}, \bar{\lambda}; k) = f(\bar{x}).$$

Moreover, it can be easily verified that for $k \geqq c_1^*$,

$$\Omega(\bar{x}, \varepsilon_k) \times \Lambda(\bar{\lambda}, \varepsilon_k) \supseteq \Omega(\bar{x}, \varepsilon_{c_1^*}) \times \Lambda(\bar{\lambda}, \varepsilon_{c_1^*})$$

so that either $\|x_k - \bar{x}\| \geqq \varepsilon_{c_1^*}$ or $\|x_k - \bar{x}\| < \varepsilon_{c_1^*}$ and $\|\lambda_k - \bar{\lambda}\| \geqq \varepsilon_{c_1^*}$. Now, since $X \times L$ is compact, the sequence $\{(x_k, \lambda_k)\}$ admits a convergent subsequence $\{(x_{k_j}, \lambda_{k_j})\}$ for $x_{k_j} \to \hat{x} \in X$, $\lambda_{k_j} \to \hat{\lambda} \in L$ and

$$S(x_{k_j}, \lambda_{k_j}; k_j) \leqq f(\bar{x}).$$

It follows:

$$\limsup_{j \to \infty} S(x_{k_j}, \lambda_{k_j}; k_j) \leqq f(\bar{x}),$$

from which we obtain

$$f(\hat{x}) + [\hat{\lambda}, g(\hat{x})] + \limsup_{j \to \infty} k_j \|g(x_{k_j})\|^2 + \left\| M(\hat{x}) \nabla f(\hat{x}) + M(\hat{x}) \frac{\partial g(\hat{x})^T}{\partial x} \hat{\lambda} \right\|^2 \leqq f(\bar{x}).$$

This implies

$$g(\hat{x}) = 0, \qquad f(\hat{x}) \leqq f(\bar{x}).$$

Then, by assumption (ii), we have $\hat{x} = \bar{x}$ and, by (i) of Theorem 1, $\hat{\lambda} = \bar{\lambda}$. Therefore we get a contradiction either with $\|\hat{x} - \bar{x}\| \geqq \varepsilon_{c_1^*}$ or with $\|\hat{\lambda} - \bar{\lambda}\| \geqq \varepsilon_{c_1^*}$. It can be concluded that there exists a value $c^*(L)$ such that for $c \geqq c^*(L)$, $(\bar{x}, \bar{\lambda})$ is the unique global minimum point of $S(x, \lambda; c)$ on $X \times L$. □

A converse result can easily be stated if it is assumed that any global minimum point of Problem P on X is a stationary point of the Lagrangian $L(x, \lambda)$:

THEOREM 6. *Let f, g be differentiable, let $X \times L$ be a given subset of $R^n \times R^m$ and let $(\bar{x}, \bar{\lambda})$ be a global minimum point for $S(x, \lambda; c)$ on $X \times L$. Assume that*

(i) $g(\bar{x}) = 0$

(ii) *for any global minimum point $\hat{x}$ of Problem* P *on X there exists a multiplier $\hat{\lambda} \in L$ such that $\nabla_x L(\hat{x}, \hat{\lambda}) = 0$.*

Then $\bar{x}$ is a global minimum point of Problem P *on X.*

Proof. By (i) we obtain:

$$S(\bar{x}, \bar{\lambda}; c) = f(\bar{x}) + \left\| M(\bar{x}) \nabla f(\bar{x}) + M(\bar{x}) \frac{\partial g(\bar{x})^T}{\partial x} \bar{\lambda} \right\|^2 \leqq S(x, \lambda; c) \quad \forall (x, \lambda) \in X \times L.$$

Therefore we obtain, in particular,

$$f(\bar{x}) \leqq f(x) \qquad \forall (x, \lambda) \in X \times L : g(x) = 0, \quad \nabla f(x) + \frac{\partial g(x)^T}{\partial x} \lambda = 0.$$

This implies by (ii) that $\bar{x}$ is a global minimum point for Problem P on X. □

Making use of the results given in Proposition 3 and in Theorem 6, we can also state the following:

THEOREM 7. *Let f, g be two times continuously differentiable and let $X \times L$ be a compact subset of $R^n \times R^m$. Assume that $M(x)$ is a continuously differentiable $(m \times n)$ matrix such that $M(x)[\partial g(x)/\partial g]^T$ is nonsingular for any $x \in X$; assume further that* (ii) *of Theorem 6 holds. Then, there exists a $c^* > 0$ such that for all $c \geqq c^*$, if $(\bar{x}, \bar{\lambda}) \in \operatorname{int}(X \times L)$ is a global minimum point of $S(x, \lambda; c)$ on $X \times L$, $\bar{x}$ is a global minimum point for Problem* P *on X.* □

An important special case in which a global property holds is that of quadratic problems with linear equality constraints:

Problem QP.

$$\text{minimize } f(x)=[x, Ax]+[a, x]$$

subject to:

$$Bx=b,$$

where:

(i) $[x, Ax]>0 \ \forall x : x\neq 0, Bx=0$

(ii) B has full rank.

In this case we can take for $M(x)$ any constant matrix M such that MB^T has rank m.

We have the following:

THEOREM 8.*Under the assumptions stated for Problem* QP, *there exists a* $c^*>0$ *such that for* $c\geqq c^*$ *the function* $S(x, \lambda; c)$ *is a positive definite quadratic function whose global minimum is the unique solution of Problem* QP.

Proof. By Proposition 1, the optimal solution of Problem QP is a stationary point of $S(x, \lambda; c)$ for any c.

On the other hand, noting that the second order homogeneous part of $S(x, \lambda; c)$ is given by $P(x, \lambda)+cQ(x, \lambda)$ where

$$P(x, \lambda)=[x, Ax]+\|M(2Ax+B^T\lambda)\|^2+[x, B^T\lambda],$$

$$Q(x, \lambda)=\|Bx\|^2$$

and making use of Lemma 1 it can be proved, as in Theorem 1, that there exists a $c^*>0$ such that for $c\geqq c^*$ the quadratic function $S(x, \lambda; c)$ is positive definite. □

7. Numerical examples. In order to evaluate the theory, several numerical examples were explored.

We report here the results obtained for the same set of test problems considered in [11].

The unconstrained minimization of $S(x, \lambda; c)$ with respect to x and λ was performed by the Fletcher–Reeves conjugate gradient method assuming as starting point

$$x_i=2, \qquad i=1, \cdots, n,$$

$$\lambda_i=0, \qquad i=1, \cdots, m$$

and taking for the penalty coefficient the value $c=10$ whenever it worked; only in Example 6 it was necessary to increase c, and the value $c=100$ was used.

For each example we indicate the matrix $M(x)$ employed, and the number N of iterations needed to reach a local minimum point (x^*, λ^*) with the given significant figures.

Example 1.

Minimize

$$f(x)=(x_1-x_2)^2+(x_2+x_3-2)^2+(x_4-1)^2+(x_5-1)^2$$

subject to

$$x_1+3x_2=0,$$

$$x_3+x_4-2x_5=0,$$

$$x_2-x_5=0,$$

$M(x)=I$
$x^*=(-0.7674, 0.2558, 0.6279, -0.1162, 0.2558)$
$\lambda^*=(2.0465, 2.2325, -5.9534)$
$N=8.$

Example 2.
Minimize

$$f(x)=(x_1-1)^2+(x_1-x_2)^2+(x_2-x_3)^4$$

subject to:

$$x_1(1+x_2^2)+x_3^4-4-3\sqrt{2}=0,$$

$M(x)=I$
$x^*=(1.1048, 1.1966, 1.5352)$
$\lambda^*=-0.1072\times 10^{-1}$
$N=32.$

Example 3.
Minimize

$$f(x)=(x_1-1)^2+(x_1-x_2)^2+(x_3-1)^2+(x_4-1)^4+(x_5-1)^6$$

subject to:

$$x_1^2x_4+\sin(x_4-x_5)-2\sqrt{2}=0,$$

$$x_2+x_3^4x_4^2-8-\sqrt{2}=0,$$

$M(x)=I$
$x^*=(1.1661, 1.1821, 1.3802, 1.5060, 0.6109)$
$\lambda^*=(-0.8553\times 10^{-1}, -0.3187\times 10^{-1})$
$N=189.$

Example 4.
Minimize

$$f(x)=(x_1-1)^2+(x_1-x_2)^2+(x_2-x_3)^2+(x_3-x_4)^4+(x_4-x_5)^4$$

subject to:

$$x_1+x_2^2+x_3^3-2-3\sqrt{2}=0,$$

$$x_2-x_3^2+x_4+2-2\sqrt{2}=0,$$

$$x_1x_5-2=0,$$

$M(x)=\partial g(x)/\partial x$
$x^*=(1.1911, 1.3626, 1.4728, 1.6350, 1.6790)$
$\lambda^*=(-0.3882\times 10^{-1}, -0.1674\times 10^{-1}, -0.2898\times 10^{-3})$
$N=80.$

Example 5.

Minimize

$$f(x)=0.01(x_1-1)^2+(x_2-x_1^2)^2$$

subject to:

$$x_1+x_3^2+1=0,$$

$M(x)=[1 \quad 0 \quad 0]$
$x^*=(-1.0000,\ 1.0000,\ 0.2294\times 10^{-5})$
$\lambda^*=0.3999\times 10^{-1}$
$N=52.$

Example 6.

Minimize

$$f(x)=-x_1$$

subject to:

$$x_2-x_1^3-x_3^2=0,$$

$$x_1^2-x_2-x_4^2=0,$$

$M(x)=10^3\, I$
$x^*=(1.0000,\ 1.0000,\ 0.0000,\ 0.0000)$
$\lambda^*=(-1.0000,\ 1.0000)$
$N=90.$

Example 7.

Minimize

$$f(x)=\log(1+x_1^2)-x_2$$

subject to:

$$(1+x_1^2)^2+x_2^2-4=0$$

$M(x)=\partial g(x)/\partial x$
$x^*=(0.0000,\ 1.7320)$
$\lambda^*=0.2867$
$N=15.$

8. Concluding remarks. From a theoretical standpoint, the method proposed in this paper combines several advantages of existing techniques for the solution of constrained problems via unconstrained minimization. On the other hand, possible disadvantages are the increase in dimensionality of the minimization problem, the presence of first order derivatives in the augmented Lagrangian and the fact that $S(x, \lambda; c)$ may be unbounded with respect to λ. This latter difficulty, however, can be overcome in many instances by employing suitable transformations. Another point where attention is needed is the threshold value of the penalty coefficient. Actually it happens that in the convex case the threshold value c^* for $S(x, \lambda; c)$ is larger than the threshold value of the penalty coefficient in the method of multipliers, where c can be given, in principle, any positive value.

The implementation of a computing procedure which makes the best use of the results given here will be considered in the future. There seems to be particular interest in the extension to the proposed method of the results on the automatic selection of the penalty parameter already established for algorithms based on other Lagrangians [9], [12]. Moreover, further investigations will be devoted to the extension of the results considered here to inequality constrained problems and to optimal control problems.

Acknowledgments. The authors would like to thank the Control Theory Center of the University of Warwick, G.B., where much of this work was carried out.

REFERENCES

[1] M. R. HESTENES, *Multipliers and gradient methods*, Computing Methods in Optimization Problems, vol. 2, L. A. Zadeh, L. W. Neustadt, A. V. Balakrishnan, eds., Academic Press, New York, 1969, pp. 143–163.

[2] M. J. D. POWELL, *A method for nonlinear constraints in minimization problems*, Optimization, R. Fletcher, ed., Academic Press, New York, 1969, pp. 283–298.

[3] R. T. ROCKAFELLAR, *Penalty methods and augmented Lagrangians in nonlinear programming*, 5th IFIP Conference on Optimization Techniques, Part I, R. Conti, A. Ruberti, eds., Springer–Verlag, Berlin, 1973, pp. 418–425.

[4] D. P. BERTSEKAS, *Multiplier methods: a survey*, Automatica, 12 (1976), pp. 133–145.

[5] D. A. PIERRE AND M. J. LOWE, *Mathematical Programming via Augmented Lagrangians: an Introduction with Computer Programs*, Addison-Wesley, Reading, MA, 1975.

[6] M. R. HESTENES, *Optimization Theory. The Finite Dimensional Case*, John Wiley, New York, 1975.

[7] R. FLETCHER, *An exact penalty function for nonlinear programming with inequalities*, Math. Programming, 5 (1973), pp. 129–150.

[8] ———, *Methods related to Lagrangian functions*, Numerical Methods for Constrained Optimization, P. E. Gill, W. Murray, eds., Academic Press, New York, 1974, pp. 219–239.

[9] H. MUKAI AND E. POLAK, *A quadratically convergent primal-dual algorithm with global convergence properties*. Memo No. ERL-M455, College of Engineering, Univ. of California, Berkeley, June 1974.

[10] A. P. WIERZBICKI, *A primal-dual large scale optimization method based on augmented Lagrange functions and interaction shift prediction*, Ricerche di Automatica, 7 (1976), pp. 34–58.

[11] A. MIELE, P. E. MOSELEY, A. V. LEVY AND G. M. COGGINS, *On the method of multipliers for mathematical programming problems*, J. Optimization Theory Appl., 10 (1972), pp. 1–33.

[12] E. POLAK, *On the stabilization of locally convergent algorithms for optimization and root finding*, Automatica, 12 (1976), pp. 337–342.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 5, September 1979

A BANG-BANG THEOREM WITH BOUNDS ON THE NUMBER OF SWITCHINGS*

HÉCTOR J. SUSSMANN†

Abstract. For systems of the form $\dot{x}=f(x)+ug(x)$, with f and g analytic, and $-1\leqq u\leqq 1$, we prove a bang-bang theorem with a priori bounds on the number of switchings, provided that the following condition is satisfied: in a neighborhood of every point x, it is possible to express, for each j, the vector field $[g, (\operatorname{ad} f)^j(g)]$ as a linear combination of the $(\operatorname{ad} f)^i(g)$, $i\leqq j+1$, in such a way that the coefficient of $(\operatorname{ad} f)^{j+1}(g)$ in this expression is bounded in absolute value by a constant $c<1$.

1. Introduction. In this paper we will prove that certain systems with a scalar control satisfy a bang-bang property with bounds on the number of switchings. Precisely, we consider systems of the form

$$\dot{x}=f(x)+ug(x), \qquad |u|\leqq 1. \tag{1}$$

The variable x is supposed to belong to an analytic manifold M, and f and g are analytic vector fields on M. We say that such a system satisfies the *bang-bang property with bounds on the number of switchings* if the following holds: (BBBNS) For every compact subset K of M, and every time $T>0$, there exists a positive integer N such that, whenever γ is a time-optimal trajectory of (1) which is entirely contained in K and goes from a point p of K to a point q of K, then there is a time-optimal trajectory from p to q which is bang-bang with at most N switchings.

Remark. (BBBNS) does *not* say that every time-optimal trajectory is bang-bang. All it does say is that, if p can be transferred to q in a time-optimal way via some trajectory γ, then this transfer can also be effected by a bang-bang trajectory, which need not be the same as γ. Moreover, (BBBNS) also asserts that, as long as γ is restricted to a compact set, then there is an a priori bound on the number of switchings for the corresponding bang-bang trajectory.

Under certain conditions to be specified below, it will be shown that the system (1) satisfies the BBBNS condition. The importance of this result is that it implies that when our conditions are satisfied, the time-optimal problem for the system (1) has a solution in feedback form which is a "regular synthesis" in a sense which is a slight modification of that of Boltyanski [1] (cf. Sussmann [5] for details).

In [2], Krener studied systems of the form

$$\dot{x}=\sum_{i=1}^{k} u_iX_i(x), \qquad 0\leqq u_i, \quad \sum u_i=1. \tag{2}$$

He proved [2, Thm. 3] a bang-bang theorem under certain conditions. For the case $k=2$, and for analytic systems, our hypotheses are essentially those of Krener's. However our conclusion is much stronger. Krener shows that every piecewise smooth control can be replaced by a bang-bang one. However, the usual compactness argument that proves the existence of time-optimal controls only enables us to get such controls to be measurable. Hence Krener's result would not apply to conclude that the time-optimal problem with fixed endpoints p, q has a bang-bang solution.

The argument utilized here does not seem to generalize to problems of the form (2) with $k>2$, under hypotheses similar to Krener's.

* Received by the editors August 8, 1978, and in revised form January 8, 1979.

† Mathematics Department, Rutgers University, New Brunswick, New Jersey 08903.

Technically, the main point of the paper is Lemma 3, which generalizes a well known result on the number of zeros of a solution of an ordinary differential equation. Lemma 3 is stated in a separate section (§ 4). Since the proof is quite lengthy, we have included, at the beginning of § 4, a sketch of its main ideas.

2. Statement of the theorem. A system of the form (1) will be called *analytic* if the state space M is an analytic manifold and the vector fields f and g are analytic.

Our crucial hypothesis is condition (Δ), which we now state. Let $[X, Y]$ denote the Lie bracket of the vector fields X and Y. Let ad X be the operator which assigns to each vector field Y the vector field $[X, Y]$. Let $p \in M$, and let $m > 0$ be an integer. We say that $(\Delta_{p,m})$ holds if there is a neighborhood U of p such that

$$[g, (\operatorname{ad} f)^m(g)] = \sum_{i=0}^{m} \alpha_i (\operatorname{ad} f)^i(g) + \beta (\operatorname{ad} f)^{m+1}(g), \tag{3}$$

where the α_i and β are analytic functions on U, and $|\beta(x)| < 1$ for all $x \in U$. We say that condition (Δ) holds if $(\Delta_{p,m})$ is satisfied for all p, m.

Our main result is as follows:

THEOREM. *If the system* (1) *is analytic, and satisfies property* (Δ), *then* (1) *satisfies the BBBNS condition.*

3. Proof of the theorem. Given a function $u = u(t)$ defined on an interval I, a *trajectory* for u is a curve $\gamma: I \to M$ that is absolutely continuous and satisfies

$$\dot\gamma(t) = f(\gamma(t)) + u(t) g(\gamma(t))$$

for almost every $t \in I$.

An *admissible pair* for the system (1) is a pair (γ, u), where u is a measurable function defined on an interval I, with values in $[-1, 1]$, and $\gamma: I \to M$ is a trajectory for u.

We shall use T_xM, T_x^*M to denote, respectively, the tangent and cotangent space to M at x.

Suppose that (γ, u) is an admissible pair for the system (1). An *adjoint solution* for (γ, u) is a continuous map $t \to \lambda(t)$, defined for all t in the domain of definition of γ, and such that $\lambda(t) \in T^*_{\gamma(t)}M$ for all t, and that λ satisfies the *adjoint equation*. This equation can be written in local coordinates as follows:

$$\frac{d\lambda_i}{dt} = -\sum_j \lambda_j \left(\frac{\partial f_j}{\partial x_i} + u(t) \frac{\partial g_j}{\partial x_i} \right), \tag{4}$$

if, relative to the coordinates $x_1, \dots, x_n$, the vector fields f, g and the cotangent vector λ are given by

$$f = \sum_i f_i \frac{\partial}{\partial x_i}, \qquad g = \sum g_i \frac{\partial}{\partial x_i}, \qquad \lambda = \sum \lambda_i \, dx_i.$$

The invariance of the adjoint equation under changes of coordinates can be checked directly, or it can be established from the following equivalent characterization, which is clearly invariant.

Suppose that $s \to \delta(s)$ is a C^1 curve with $\delta(0) = \gamma(t_0)$. Let $t \to \gamma_s(t)$ denote the solution of $\dot x = f + ug$ for which $\gamma_s(t_0) = \delta(s)$. Let $v(t)$ be the tangent vector to the curve $s \to \gamma_s(t)$ at $s = 0$. Any map $t \to v(t) \in T_{\gamma(t)}M$ which is obtained in this way is called a *variational vector field along* (γ, u). If v is a tangent vector at $\gamma(t_0)$, then there is a unique variational vector field $V(t)$ such that $V(t_0) = v$. Let $V(t) = F(t, t_0)v$. Then $F(t, t_0)$ is linear. If $\lambda(t) \in T^*_{\gamma(t)}M$ for every t, then λ is an adjoint solution if and only if

$$\lambda(t) \circ F(t, t_0) = \lambda(t_0)$$

for all t, t_0.

From the invariant characterization of adjoint solutions, a trivial but important consequence follows. Suppose that we are given a family of linear subspaces $Q(t)$ of $T_{\gamma(t)}M$, which is *invariant*, in the sense that

$$F(t, t_0)Q(t_0) = Q(t) \tag{5}$$

for all t, t_0. Then, *if λ is an adjoint solution, and if $\lambda(t)$ is nontrivial on $Q(t)$ for some t, it follows that $\lambda(t)$ is nontrivial on $Q(t)$ for every t.* In particular, this conclusion holds if the $Q(t)$ are obtained from an *invariant family of submanifolds*, as follows. Suppose that, for each t in the domain of (γ, u), we have a smooth manifold $S(t)$, such that $\gamma(t) \in S(t)$ and that, whenever $t \to \delta(t)$ is a solution of $\dot{x} = f + ug$, for which $\delta(t_0) \in S(t_0)$, it follows that $\delta(t) \in S(t)$ for all t. Then it is clear that the spaces $Q(t) = T_{\gamma(t)}S(t)$ constitute an invariant family.

The Pontryagin maximum principle asserts that, if (γ, u) is time-optimal, then there exists a nontrivial adjoint solution $t \to \lambda(t)$ which satisfies

$$\min_v H(\lambda(t), \gamma(t), v) = H(\lambda(t), \gamma(t), u(t)) \tag{6}$$

for almost all t, where

$$H(\lambda, x, v) = \langle \lambda, f(x) + vg(x) \rangle. \tag{7}$$

Suppose that S is a submanifold of M such that f and g are tangent to S and γ is contained in S. Then we can consider the system Σ' obtained from our original system Σ by restricting the state space to S. The pair (γ, u) is clearly time-optimal for Σ' as well, so we can apply the maximum principle to Σ', and conclude that there is a nontrivial adjoint solution $t \to \mu(t) \in T^*_{\gamma(t)}S$ that satisfies (6). But then, if we choose in an arbitrary way a t_0 and a linear functional $\bar{\lambda}$ on $T^*_{\gamma(t_0)}M$ whose restriction to $T_{\gamma(t_0)}S$ is $\mu(t_0)$, and if we let $\lambda(t)$ be the adjoint solution (for Σ) with $\lambda(t_0) = \bar{\lambda}$, it is easy to see that $\mu(t)$ is the restriction of $\lambda(t)$ to $T_{\gamma(t)}S$. Hence $\lambda(t)$ is nontrivial on $T_{\gamma(t)}S$ for all t.

Now let L denote the Lie algebra of vector fields generated by f and g. Since L is a Lie algebra of analytic vector fields, it follows from Nagano's theorem (cf. Nagano [3]) that M is partitioned into submanifolds S—*the maximal integral manifolds of L*—such that, for each $x \in S$,

$$T_xS = L(x) = \{X(x) : X \in L\}. \tag{8}$$

It is clear that, if (γ, u) is an admissible pair, then γ is entirely contained in one maximal integral manifold S of L. The vector fields f and g are tangent to S, so we can apply the preceding remark to this case, and conclude:

LEMMA 1. *If (γ, u) is time-optimal, then there exists an adjoint solution $t \to \lambda(t)$ for (γ, u) which satisfies condition* (6), *and is such that $\lambda(t)$ is nontrivial on $L(\gamma(t))$ for all t.*

We would like to strengthen Lemma 1 by substituting for $L(\gamma(t))$ the subspace $L_0(\gamma(t))$ defined as follows: L_0 is the ideal generated by g of the Lie algebra L. For all x in M:

$$L_0(x) = \{X(x) : X \in L_0\}.$$

Then L_0 is a subset of L, so that $L_0(x)$ is a subspace of $L(x)$ for every x in M.

Let us call the pair (γ, u) *strongly extremal* if there exists an adjoint solution λ that satisfies (6) and is such that, in addition, $\lambda(t)$ is nontrivial on $L_0(\gamma(t))$ for all t. Then it is not true in general that every time-optimal pair (γ, u) connecting two points p, q is necessarily strongly extremal. But a weaker property still holds, namely, that every time-optimal pair (γ, u) connecting two points p, q can be replaced by a concatenation of strong extremals and constant-control trajectories. To make this precise, let us consider 4-tuples (K, K', T, u_0), where K and K' are compact sets such that $K \subseteq K'$, $T > 0$ is a number, and u_0 is a control value (i.e. a number such that $|u_0| \leqq 1$). We shall say that (K, K', T, u_0) has the *strong extremal replacement property* (henceforth abbreviated as SERP) with N steps (N being a positive integer), if the following holds:

(SERP_N). Whenever (γ, u) is a time-optimal pair, defined on some interval $[0, T']$, with $T' \leqq T$, such that $\gamma(t) \in K$ for $0 \leqq t \leqq T'$, then it follows that there exists some other pair $(\tilde{\gamma}, \tilde{u})$, defined on the same interval $[0, T']$, such that (i) $\tilde{\gamma}(0) = \gamma(0)$, (ii) $\tilde{\gamma}(T') = \gamma(T')$, (iii) $\tilde{\gamma}(t) \in K'$ for all t, and (iv) $(\tilde{\gamma}, \tilde{u})$ is the concatenation of at most N pairs, each of whom is either strongly extremal or constant-control, the value of the control being u_0.

LEMMA 2. *Let K, K' be compact subsets of M, with $K \subseteq \mathrm{int}(K')$. Let T, u_0 be arbitrary, with $T > 0$, $|u_0| \leqq 1$. Then there exists an $N > 0$ such that (K, K', T, u_0) has the* SERP *with N steps.*

Proof. It is easy to see that, if (K, K', T_i, u_0) have the SERP with N_i steps, for $i = 1, 2$, then $(K, K', T_1 + T_2, u_0)$ has the SERP with $N_1 + N_2$ steps. Hence it is sufficient to prove that, if K, K', u_0 are given, with K, K' compact such that $K \subseteq \mathrm{int}(K')$, and $|u_0| \leqq 1$, then there is a $T > 0$ such that (K, K', T, u_0) has the SERP with two steps.

Let us choose another compact set K'' such that $K' \subseteq \mathrm{int}(K'')$. Because K, K', K'' are compact subsets of $\mathrm{int}(K')$, $\mathrm{int}(K'')$, M, respectively, there exist a $T > 0$ with the property that, whenever a control $t \to u(t)$, $a \leqq t \leqq b$, is defined on some interval $[a, b]$ whose length $b - a$ is not greater than T, and γ is a trajectory corresponding to the control u, and defined on a maximal interval $I \subseteq [a, b]$, then:

(a) if $\gamma(t) \in K''$ for some t then $I = [a, b]$,

(b) if $\gamma(t) \in K'$ for some t then γ is entirely contained in K'',

(c) if $\gamma(t) \in K$ for some t then γ is entirely contained in K'.

We now show that this choice of T is correct, i.e. that (K, K', T, u_0) has the SERP with two steps. So, let (γ, u) be a time-optimal pair, defined on a time interval $[0, T']$, with $T' \leqq T$, such that $\gamma(t) \in K$ for $0 \leqq t \leqq T'$. Let $p = \gamma(0)$, $q = \gamma(T')$. We will prove that there is a pair $(\tilde{\gamma}, \tilde{u})$, also defined on $[0, T']$, such that $\tilde{\gamma}(0) = p$, $\tilde{\gamma}(T') = q$, that $(\tilde{\gamma}, \tilde{u})$ is either strongly extremal, or the concatenation of a strong extremal and a $(\hat{\gamma}, \hat{u})$ with $\hat{u} = \text{constant} = u_0$, and that $\tilde{\gamma}$ is entirely contained in K'.

Now let us use some results from Sussmann-Jurdjevic [4]. For each point x in M, let $S(x)$, $S_0(x)$ denote the maximal integral manifolds through x of L, L_0 respectively. Then, for every x, the dimension of $S_0(y)$ remains constant as y varies over all points in $S(x)$. This dimension is either equal to $\dim S(x)$, or to $\dim S(x) - 1$. If a trajectory of the system (1) goes through a point x, then this trajectory is necessarily contained in $S(x)$. (These three facts are proved in [4].)

In particular, let $k = \dim S(p)$, $k' = \dim S_0(p)$. Then $k = k'$ or $k = k' + 1$. Moreover, the curve γ is entirely contained in $S(p)$. Suppose that $k = k'$. Then $L_0(\gamma(t)) = L(\gamma(t))$ for all t. Lemma 1 implies that (γ, u) is itself strongly extremal. So we can take $\tilde{\gamma} = \gamma$, $\tilde{u} = u$. This disposes of the case $k = k'$.

Now let us assume that $k' = k - 1$. We define a time-dependent vector field h on $\mathrm{int}(K'')$, for times t such that $|t| \leqq T$, as follows. For x in M, let $t \to G(t, x)$ denote the integral curve of the vector field $X = f + u_0 g$ which passes through x when $t = 0$. Then, for x in K'', $G(t, x)$ is well defined for $-T \leqq t \leqq T$, because of condition (a) of the choice of T. So, for $|t| \leqq T$, the map $G(t, \cdot)$ is well defined on int (K''). For $x \in \operatorname{int}(K'')$, $|t| \leqq T$, put

$$h(t, x) = G(-t, \cdot)_* g(G(t, x)).$$

(Here $G(t, \cdot)_*$ is the differential of the map $G(t, \cdot)$.)

As shown in [4], the maps $G(t, \cdot)$ take integral manifolds of L_0 to integral manifolds of L_0. Hence $G(t, \cdot)_*$ will map, for every x, $L_0(x)$ to $L_0(G(t, x))$. In particular, $h(t, x)$ belongs to $L_0(x)$ for every x in int (K'') and every t such that $|t| \leqq T$. From this it follows by a standard argument, that, if $t \to w(t)$ is a bounded measurable function, and if $t \to x(t)$ is a solution of $\dot{x}(t) = w(t)h(t, x(t))$, for t in some interval $a \leqq t \leqq b$ of length not greater than T, then $x(t)$ belongs to $S_0(x(a))$ for all t. (Proof: let L_0'' denote the set of restrictions to int (K'') of the vector fields in L_0, and let $S_0''(x)$ denote the maximal integral manifold of L_0'' through x, so that the $S_0''(x)$, as x varies over int (K'') constitute a partition of int (K''). Since the equation $\dot{x} = w(t)h(t, x)$ makes perfect sense when restricted to integral submanifolds of L_0'', and has local existence and uniqueness of solutions, it follows that, whenever a solution of this equation is defined on some time interval, then, for every integral submanifold of L_0'', the set of times for which the solution belongs to the given submanifold is open. Hence the solution is entirely contained in the integral submanifold $S_0''(x)$, for some x. But $S_0''(x)$ is also a connected integral manifold of L_0, although not necessarily a maximal one. In any case, $S_0''(x)$ is a subset of $S_0(x)$, so the given solution is contained in $S_0(x)$ as well.)

Now let $v: [0, T''] \to [-1, 1]$ be an arbitrary admissible control, with $T'' \leqq T$. Let δ_v be the trajectory of our system (1) that corresponds to the control v and the initial condition $\delta_v(0) = p$. Let η_v be the solution of

$$\dot{x}(t) = (v(t) - u_0)h(t, x(t)), \tag{9}$$

also with the initial condition $\eta_v(0) = p$. Then $\eta_v(t) \in S_0(p)$ for all t in $[0, T'']$, as shown before. Let $\delta_v'(t) = G(t, \eta_v(t))$. Then an easy calculation shows that

$$\begin{aligned}\dot{\delta}_v'(t) &= (f + u_0 g)(\delta_v'(t)) + G(t, \cdot)_*(\dot{\eta}_v(t)) \\ &= (f + u_0 g)(\delta_v'(t)) + (v(t) - u_0)g(\delta_v'(t)) \\ &= f(\delta_v'(t)) + v(t)g(\delta_v'(t)).\end{aligned}$$

Moreover, $\delta_v'(0) = p = \delta_v(0)$. So $\delta_v'(t) = \delta_v(t)$ for all t. Therefore, we have proved that

$$\delta_v(t) = G(t, \eta_v(t)) \quad \text{for } 0 \leqq t \leqq T''. \tag{9$'$}$$

Equation (9) defines a time-varying control system on int (K''), which we shall name $\tilde{\Sigma}$. Also, we can restrict equation (9) to the submanifold $S_0''(p)$ (defined above). Let us use $\tilde{\Sigma}''$ to refer to this restriction. Finally, let Σ denote our original system (1).

Equation (9′) establishes a correspondence between trajectories of Σ that start at p, and trajectories of $\tilde{\Sigma}''$ that start at p. In particular, corresponding to our trajectory γ (i.e. δ_u) there is a trajectory η_u, which is entirely contained in $S_0''(p)$, and steers p to $q' = G(-T', q)$ in time T'. By compactness, there exists, for the system $\tilde{\Sigma}$, a control $\bar{v}$ which steers p to q' time-optimally in some time $\bar{T}$. (Note: all this depends heavily on the fact that T is small enough so that (a), (b) and (c) hold. The fact that η_v is a trajectory of $\tilde{\Sigma}''$ for every v is a consequence of (b) and (c). Indeed, (c) implies that $\delta_v(t) \in K'$ for all t. Then (b) implies that $\eta_v(t) \in K''$, because, for each t, the curve $s \to G(s, \eta_v(t))$ is a trajectory of (1) which goes through a point of K' when $s = t$, and therefore is contained in K'' for $0 \leqq s \leqq t$ so that, in particular, $G(0, \eta_v(t))$ is in K'', i.e. $\eta_v(t) \in K''$. The compactness argument also depends on our choice of T. The crucial part is the choice, which is possible by compactness, of a weakly convergent sequence of controls v_n which steer p into q', for the system $\tilde{\Sigma}$, in times T_n that converge to the optimal time $\bar{T}$. If $\bar{v}$ is the weak limit of the v_n, then one must show that $\bar{v}$ actually steers p to q' in time $\bar{T}$. This is trivial, by the continuous dependence of the solutions of (9) on the controls, provided that it is known that the solution of (9) for the control $\bar{v}$ and the initial condition p is defined for all t in $[0, \bar{T}]$. But this follows from our choice of T, since this choice implies that $\eta_v(t)$ is well defined for all t, as long as v is an admissible control with domain $[0, a]$, $a \leqq T$.)

Let $q'' = \delta_{\bar{v}}(\bar{T}) = G(\bar{T}, q')$. Then, for the system Σ, $\bar{v}$ steers p to q'' in time $\bar{T}$. If we let $\tilde{v}:[0, T'] \to [-1, 1]$ be the control whose value at t is $\bar{v}(t)$ for $0 \leqq t \leqq \bar{T}$, and u_0 for $\bar{T} < t \leqq T'$, then $\tilde{v}$ steers p to $G(T' - \bar{T}, G(\bar{T}, q'))$, i.e. to $G(T', q')$, i.e. to q. The corresponding trajectory $\delta_{\tilde{v}}$ is the concatenation of $\delta_{\bar{v}}$ and of a trajectory that corresponds to the constant control u_0. So our conclusion will be proved if we show that the pair $(\delta_{\bar{v}}, \bar{v})$ is strongly extremal.

The pair $(\eta_{\bar{v}}, \bar{v})$ is time-optimal for the system $\tilde{\Sigma}''$. Hence, by the maximum principle, there exists a map $t \to \mu(t)$, defined for $0 \leqq t \leqq \bar{T}$, such that $\mu(t) \in L_0(\eta_{\bar{v}}(t))^*$ and $\mu(t) \neq 0$ for all t, that μ is an adjoint solution for $\tilde{\Sigma}''$, and that, for almost every t:

$$\text{(10)} \qquad \min_w H'(t, \mu(t), \eta_{\bar{v}}(t), w) = H'(t, \mu(t), \eta_{\bar{v}}(t), \bar{v}(t)),$$

where $H'(t, \nu, x, w) = (w - u_0)\langle \nu, h(t, x)\rangle$.

Now define $\rho(t) \in L_0(\delta_{\bar{v}}(t))^*$ by "pulling back" $\mu(t)$ via $G(t, \cdot)$, i.e. let

$$\rho(t) = G(-t, \cdot)^* \mu(t),$$

where $G(-t, \cdot)^* : L_0(\eta_{\bar{v}}(t))^* \to L_0(\delta_{\bar{v}}(t))^*$ is the dual map of $G(-t, \cdot)_* : L_0(\delta_{\bar{v}}(t)) \to L_0(\eta_{\bar{v}}(t))$.

Then $\rho(t) \neq 0$ for all t. If we let $H''(\nu, x, w) = (w - u_0)\langle \nu, g(x)\rangle$, then it follows from (9″) that, for almost all t,

$$\min_w H''(\rho(t), \delta_{\bar{v}}(t), w) = H''(\rho(t), \delta_{\bar{v}}(t), \bar{v}(t)).$$

It is easy to see that, if $t \to A(t)$ is a variational vector field for the system $\tilde{\Sigma}''$ along $(\eta_{\bar{v}}, \bar{v})$, and if we let $B(t) = G(t, \cdot)_* A(t)$, then B is a variational vector field for Σ along $(\delta_{\bar{v}}, \bar{v})$. Since μ is an adjoint solution for $\tilde{\Sigma}''$ along $(\eta_{\bar{v}}, \bar{v})$, it follows that $\langle \mu(t), A(t)\rangle$ is constant. Therefore $\langle \rho(t), B(t)\rangle = \text{constant}$.

Now choose, in an arbitrary fashion, a linear functional $\sigma(0) \in T_p^* M$ whose restriction to $L_0(p)$ is $\rho(0)$. Let $t \to \sigma(t)$ be the adjoint solution for Σ along $(\delta_{\bar{v}}, \bar{v})$ whose value at $t = 0$ is $\sigma(0)$. Then, if A, B are as above, the function $t \to \langle \sigma(t), B(t)\rangle$ is also constant. Since $\langle \sigma(0), B(0)\rangle = \langle \rho(0), B(0)\rangle$, we conclude that $\langle \sigma(t), B(t)\rangle = \langle \rho(t), B(t)\rangle$ for all t, and for all variational vector fields along $(\delta_{\bar{v}}, \bar{v})$ that are obtainable from an $A(t)$ in the manner described above. If V is any tangent vector at $\delta_{\bar{v}}(t)$, such that $v \in L_0(\delta_{\bar{v}}(t))$, then there exists a variational vector field $s \to B(s)$, of the desired form, such that $B(t) = V$. Hence $\langle \rho(t), V\rangle = \langle \sigma(t), V\rangle$. So the restriction of $\sigma(t)$ to $L_0(\delta_{\bar{v}}(t))$ coincides with $\rho(t)$. This shows, in particular, that $\sigma(t)$ is nontrivial on $L_0(\delta_{\bar{v}}(t))$ for all t. Also, we have

$$\min_w (w - u_0)\langle \sigma(t), g(\delta_{\bar{v}}(t))\rangle = (\bar{v}(t) - u_0)\langle \sigma(t), g(\delta_{\bar{v}}(t))\rangle$$

for almost all t.

If we add $\langle \sigma(t), (f + u_0 g)(\delta_{\bar{v}}(t))\rangle$ to both sides of the preceding equality, we see that the adjoint solution μ satisfies (6). So $(\delta_{\bar{v}}, \bar{v})$ is indeed strongly extremal. Q.E.D.

Remark. Let us say that (K, T, u_0) has the SERP with N steps if (K, K', T, u_0) has the SERP with N steps for some compact K' that contains K. Then what we have proved in Lemma 2 is that, whenever K, T, u_0 are given, there is some N such that (K, T, u_0) has the SERP with N steps. However, the only reason why it was necessary to allow for the possibility that N may be large was that the trajectories of the system (1) may fail to be everywhere defined. If the trajectories are everywhere defined (i.e. if, for every bounded measurable $u:[a, b]\rightarrow[-1, 1]$, every t_0 in $[a, b]$, and every x_0 in M there is a solution of (1) defined in $[a, b]$, whose value at t_0 is x_0), then it is easy to see that every (K, T, u_0) has the SERP with two steps. Indeed, in this case there is no need to worry about the compact sets K', K'', the vector field $h(t, x)$ is defined for all t and all x in M, and the same proof we have given of Lemma 2 establishes the stronger conclusion.

For an example where the SERP with two steps fails, consider the system $\dot{x} = 1 + u$, $\dot{y} = 1 - u$, $|u| \leqq 1$, defined on the strip $M = \{(x, y) \in \mathbb{R}^2 : x - 1 \leqq y \leqq x + 1\}$. Then every control is time-optimal, but a strongly extremal control is necessarily bang-bang. The point $(0, 0)$ can be transferred to $(1, 1)$ in time 1, and every control that effects this transfer is time-optimal. There are bang-bang controls with three switchings that do the job, but no control with two switchings. If the strip M seems too artificial as the choice of the state space for a control system, one can use the fact that M is diffeomorphic to the whole plane to transform the preceding example into one that is defined on $\mathbb{R}^2$.

4. A technical lemma. We now state and prove the main technical lemma of this paper. The reader who so wishes may limit himself to reading the statement of Lemma 3, and then may proceed directly to § 5.

The purpose of Lemma 3 is to extend a well known result on the number of zeros of a nontrivial solution of an ordinary differential equation. Consider the equation

$$\varphi^{(N)} + \alpha_1\varphi^{(N-1)} + \cdots + \alpha_N\varphi = 0,$$

where $\alpha_1, \ldots, \alpha_N$ are bounded, measurable real-valued functions of t. Then it is well known that there exists a $T > 0$ such that, if φ is any nontrivial solution, then φ has at most $N - 1$ zeros on any interval of length $\leqq T$. Moreover, the number T can be taken to be independent of the particular functions $\alpha_1, \ldots, \alpha_N$ as long as the α_i are bounded by a fixed constant A.

This result can be reformulated as follows. Let $\varphi_1 = \varphi$, $\varphi_2 = \varphi'$, $\ldots$, $\varphi_N = \varphi^{(N-1)}$. Let $M(t)$ be the N by N matrix function:

$$\mathrm{M} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ & & & \cdots & & \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -\alpha_N & -\alpha_{N-1} & -\alpha_{N-2} & \cdots & -\alpha_2 & -\alpha_1 \end{bmatrix}.$$

Then the vector $F(t)$ whose components are $\varphi_1(t), \ldots, \varphi_N(t)$ satisfies $\dot{F}(t) = M(t)F(t)$. The result we have stated says that, for every N and every $A > 0$, there is a $T > 0$ such that, if M is an N by N matrix *of the particular type specified above* and if the components α_i are bounded by A, then, if F is a nontrivial solution of $\dot{F} = MF$, it follows that, on every interval of length $\leqq T$, the first component of F has at most $N - 1$ zeros.

Lemma 3 extends this result to more general matrix functions $M(t) = \{m_{ij}(t)\}$. We want to allow all the entries on or below the main diagonal to be nonzero. Moreover, instead of requiring that $m_{i,i+1}$ be identically equal to one, we want to allow the $m_{i,i+1}$ to be functions of t, that are positive and bounded away from zero, and bounded above by a constant. It turns out that the result stated above holds for this more general situation as well, except for the fact that the choice of T now depends not only on N and on the upper bound A for the absolute values of the components of the matrix M, but also on the lower bound B for the functions $m_{i,i+1}$. Precisely, we shall prove:

LEMMA 3. *Let A, B be real numbers such that $A > B > 0$. Let N be a positive integer. Then there exists a positive real number $T = T(N, A, B)$ with the following property:*

Whenever $[a, b]$ is an interval of length $b - a \leqq T$, and the α_{ij} $(i = 1, \dots, N$, $j = 1, \dots, i)$, $\beta_i (i = 1, \dots, N-1)$ are measurable real-valued functions on $[a, b]$ such that $|\alpha_{ij}(t)| \leqq A$ and $B \leqq \beta_i(t) \leqq A$ for all t in $[a, b]$ then, if $\varphi_1, \dots, \varphi_N$ are absolutely continuous functions on $[a, b]$ that satisfy the system of equations

$$\begin{aligned} \dot\varphi_1 &= \alpha_{11}\varphi_1 + \beta_1\varphi_2, \\ \dot\varphi_2 &= \alpha_{21}\varphi_1 + \alpha_{22}\varphi_2 + \beta_2\varphi_3, \\ &\vdots \\ \dot\varphi_i &= \alpha_{i1}\varphi_1 + \alpha_{i2}\varphi_2 + \cdots + \alpha_{ii}\varphi_i + \beta_i\varphi_{i+1}, \\ &\vdots \\ \dot\varphi_N &= \alpha_{N1}\varphi_1 + \alpha_{N2}\varphi_2 + \cdots + \alpha_{NN}\varphi_N \end{aligned}$$

and if φ_1 does not vanish identically on $[a, b]$, then φ_1 has at most $N-1$ zeros on $[a, b]$.

Outline of the proof. We try to mimick the standard proof of the classical result quoted above. First, let us recall how that proof goes. Suppose that the α_i are continuous, that φ is a solution of $\varphi^{(N)} + \alpha_1\varphi^{(N-1)} + \cdots + \alpha_N\varphi = 0$, and that φ has at least N zeros on some interval $[a, b]$ of length $T = b - a$. We shall show that, if T is sufficiently small, this can only happen if φ vanishes identically. Indeed, the functions φ, $\varphi', \dots, \varphi^{(N-1)}$ are of class C^1. Hence, between any two zeros of one of these functions, there must be a zero of its derivative. If φ has N zeros on $[a, b]$, then φ' must have $N-1$ zeros, φ'' must have $N-2$ zeros, etc. In general, $\varphi^{(i)}$ will have at least $N-i$ zeros. So, for $i = 0, \dots, N-1$, $\varphi^{(i)}$ has at least one zero on $[a, b]$. On the other hand, Gronwall's inequality gives a bound $\|F(t)\| \leqq \|F(a)\| e^{C(t-a)}$, where C is some constant that depends only on N and A. So, if $T \leqq 1$, we get the bound $\|F(t)\| \leqq K\|F(a)\|$, where $K = K(N, A)$. If φ is nontrivial then we can multiply it by a constant, and obtain a nontrivial φ for which $\|F(a)\| = 1$. Then we have $\|F(t)\| \leqq K$ for all t. Since $\dot F = MF$, we can conclude that all the $\varphi^{(i)}$ are Lipschitz with a fixed constant J. Since each $\varphi^{(i)}$ has a zero in $[a, b]$, it follows that $|\varphi^{(i)}(t)| \leqq JT$ for all t in $[a, b]$ and all $i = 0, \dots, N-1$. In particular, $\|F(a)\|$ is bounded by a fixed constant times T. But, if T is small enough, this contradicts $\|F(a)\| = 1$.

To prove Lemma 3 along similar lines, we assume that φ_1 has N zeros on the interval $I = [a, b]$, and that the vector $F(a)$ with components $\varphi_1(a), \dots, \varphi_N(a)$ is normalized so that $\|F(a)\| = 1$. Exactly as before, we get a bound $\|F(t)\| \leqq$ constant, and then we conclude that all the φ_i are Lipschitzian with a fixed constant C, as long as $b - a \leqq T \leqq 1$. We then try to show that each φ_i must have at least one zero somewhere on I.

Unfortunately, we do not know how to prove that each φ_i must have a zero. For instance, if we take two different zeros p, q of φ_1, and if we try to find a zero of φ_2 in between, then the obvious candidate for the location of such a zero should be a point $t_{p,q}$ where $|\varphi_1|$ has a local maximum. However, we cannot conclude that $\varphi_2(t_{p,q}) = 0$ for at least two reasons, namely, (a) that φ_1 is only known to be absolutely continuous, and not known to be C^1, so that $\dot\varphi_1(t_{p,q})$ need not even exist, and (b) that even if $\dot\varphi_1(t_{p,q})$ did exist, its value would be $(\alpha_{11}\varphi_1 + \beta_1\varphi_2)(t_{p,q})$, so it is this number, rather than $\varphi_2(t_{p,q})$, that has to vanish.

In order to overcome this difficulty, we observe that it is not really necessary to prove that $\varphi_2(t_{p,q})$ is equal to zero, but only that it is small. Let us use the symbol

$O(h(T))$, if h is some function with positive values, to denote a quantity which is bounded by some fixed constant times $h(T)$. If we could prove that, for $i=1, \ldots, N$, there are points s_i such that $\varphi_i(s_i)$ is $O(T)$, then we would conclude that the $\varphi_i(t)$ are $O(T)$, since each φ_i is Lipschitzian with a fixed constant, and I has length T. In particular, it would follow that $\|F(a)\|=O(T)$. Hence, by taking T small enough, we would get a contradiction with $\|F(a)\|=1$.

To prove that $\varphi_2(t_{p,q})$ is $O(T)$, we use the equation

$$\dot\varphi_1=\alpha_{11}\varphi_1+\beta_1\varphi_2.$$

We claim that $\varphi_2(t_{p,q}) \leqq AB'CT$, where $B'=1/B$. Indeed, if $\varphi_2(t_{p,q})$ were $>AB'CT$, then we would get, by the continuity of φ_2, that $\varphi_2(t)>AB'CT$ for t in some interval J containing $t_{p,q}$ in its interior. But then $(\alpha_{11}\varphi_1+\beta_1\varphi_2)(t)$ would be $>-ACT+BAB'CT$, i.e. $\dot\varphi_1(t)$ would be >0 throughout J. (Recall that φ_1 is bounded by CT, because φ_1 is Lipschitzian with constant C, and has a zero in I.) But, if $\dot\varphi_1>0$ throughout J, then φ_1 is strictly increasing on J, and this contradicts the fact that $t_{p,q}$ was a local maximum of $|\varphi_1|$. A similar contradiction arises if $\varphi_2(t_{p,q})<-AB'CT$. So $|\varphi_2(t_{p,q})| \leqq AB'CT$, i.e. $|\varphi_2(t_{p,q})|$ is $O(T)$.

The argument of the preceding paragraph can be generalized into an observation that we shall refer to as (Obs), and that will be stated more precisely in the proof. Informally, (Obs) says that, if $\varphi_1, \ldots, \varphi_k$ are known to be $O(h(T))$ for some function h, and if t is a local maximum of $|\varphi_k|$, then $|\varphi_{k+1}(t)|$ is $O(h(T))$. The proof is exactly as above: using $\dot\varphi_k=\alpha_{k1}\varphi_1+\cdots+\alpha_{kk}\varphi_k+\beta_k\varphi_{k+1}$, we conclude that, if φ_{k+1} were not $O(h(T))$, then the term $\beta_k\varphi_{k+1}$ would dominate the sum in a neighborhood of t (since β_k is positive and bounded away from zero), and then $\dot\varphi_k$ would have constant sign near t, contradicting the fact that t is a local maximum of $|\varphi_k|$.

Having shown that $|\varphi_2(t_{p,q})|$ is $O(T)$ for the $N-1$ distinct points $t_{p,q}$ that can be obtained from the N pairs (p, q) of consecutive zeros of members of a set Z of N zeros of φ_1, one can continue the proof in a similar fashion. Call $Z_1=Z$, and let Z_2 be the set of the $N-1$ points $t_{p,q}$. One now constructs a set Z_3 by taking for every pair (p, q) of consecutive points of Z_2, a point $t^2_{p,q}$ in the interior of the interval $[p, q]$ where $|\varphi_2|$ has a local maximum. Since we already know that φ_1 and φ_2 are $O(T)$, using (Obs) we can conclude that $\varphi_3(t^2_{p,q})$ is $O(T)$ for each $t^2_{p,q}$ in Z_3. Hence φ_3 is $O(T)$, since φ_3 is Lipschitzian. The argument proceeds by induction. One constructs, for each k, a set Z_k of $N+1-k$ points that are local maxima of $|\varphi_{k-1}|$, and one proves inductively, using (Obs), that φ_k must be $O(T)$ on Z_k, and hence that $\varphi_k(t)$ is $O(T)$ for all t in I. The induction can continue all the way to $k=N$. Indeed, each Z_k consists of exactly $N+1-k$ points, so Z_k is nonempty for $k=1, \ldots, N$. Having proved that φ_k is $O(T)$ for $k=1, \ldots, N$ it then follows that $\|F(a)\|$ is $O(T)$, and this contradicts $\|F(a)\|=1$, if T is small enough.

There is however, a complication. For the induction to be possible, one needs to know that, for each pair p, q of consecutive points of Z_k, the function $|\varphi_k|$ has a local maximum at some *interior* point of the interval $[p, q]$. If, on some such interval $[p, q]$, the maximum of $|\varphi_k|$ were attained at p, or at q, then the argument would not go through, because (i) the maximum so obtained might not be a local maximum of $|\varphi_k|$ in I, and therefore (Obs) would not be applicable to bound φ_{k+1} at such a point, and (ii) even if the points we got are indeed local maxima, there might be too few of them. (Example: let p, q, r be three consecutive points of Z_k, and suppose that $|\varphi_k|$ increases from p to q, and then decreases from q to r. Then the point t where $|\varphi_k|$ reaches its maximum on $[p, q]$ is the same as the t' where $|\varphi_k|$ reaches its maximum on $[q, r]$. In fact, $t=t'=q$. But then the two pairs p, q and q, r of consecutive points of Z_k give rise to only one point in

Z_{k+1}. Then card $(Z_{k+1}) \leqq$ card $(Z_k) - 2$. Since we always have card $(Z_{j+1}) \leqq$ card $(Z_j) - 1$, it would then follow that Z_N is empty. Naturally, if Z_N were empty, the fact that $\varphi_N(t)$ is $O(T)$ for t in Z_N would not enable us to conclude that $\varphi_N(t)$ is $O(T)$ for all t, and the whole proof would break down.)

So we must prove that, at each step of our induction (henceforth referred to as the *main induction*), one can be sure that, for each pair of consecutive points p, q of Z_k, the function $|\varphi_k|$ does not attain its maximum over the interval $[p, q]$ at one of the endpoints. This we prove by contradiction. Assume that, at some step $k \to k+1$ of the main induction, there are consecutive p, q in Z_k such that $|\varphi_k(t)| \leqq \max(|\varphi_k(p)|, |\varphi_k(q)|)$, for all t in $[p, q]$. We show that, if T is small enough, this implies that F must vanish identically. For this it suffices to prove that φ_1 must vanish identically on $[p, q]$. (Indeed, if $\varphi_1 \equiv 0$ on $[p, q]$, we get from $\dot{\varphi}_1 = \alpha_{11}\varphi_1 + \beta_1\varphi_2$, $\beta_1 \neq 0$, that $\varphi_2 \equiv 0$ on $[p, q]$. Then we get in the same way the conclusion that $\varphi_3 \equiv 0$ on $[p, q]$, that $\varphi_4 \equiv 0$ on $[p, q]$, etc. So F vanishes identically on $[p, q]$, and therefore $F \equiv 0$ on I.) To prove that φ_1 vanishes identically on $[p, q]$, we show that φ_1 is $O(T^m)$ on $[p, q]$ for every m. Naturally, this does not suffice to conclude that $\varphi_1 \equiv 0$, but the actual proof, carried out below, will not only give $\varphi_1 = O(T^m)$, but also the values of the constants H_m such that $|\varphi_1(t)| \leqq H_m T^m$ for t in $[p, q]$. From these values, it is seen that the H_m grow geometrically, so that, if T is small enough, $\mathrm{H}_m T^m \to 0$ as $m \to \infty$, and therefore $\varphi_1(t) = 0$ for $p \leqq t \leqq q$.

To prove that φ_1 is $O(T^m)$ for all m, we will actually prove that for every m, and every $i = 1, \ldots, k$, φ_i is $O(T^{k+1+m-i})$. This will be proved by induction on m. This induction, which is carried out within each step $k \to k+1$ of the main induction, will be referred to as the *subsidiary induction*. For the induction to work, it is not enough to consider the interval $J_k = [p, q]$. One must work with a different interval J_i for each i. Starting with $J_k = [p, q]$, one constructs the J_i backwards. Each J_i is the interval between two—not necessarily consecutive—points p_i, q_i of Z_i. Having defined p_i and q_i for a given i, there are unique pairs (p_{i-1}, r_i) and (s_i, q_{i-1}) of consecutive points of Z_{i-1} such that $p_i \in [p_{i-1}, r_i]$ and $q_i \in [s_i, q_{i-1}]$. This defines the points p_{i-1}, q_{i-1}, and therefore the interval J_{i-1}. The point of this is that, every time one has a bound for φ_i on J_{i+1}, the bound extends to the larger interval J_i. This follows from the fact, that on $[p_i, r_{i+1}]$, $|\varphi_i|$ is maximized at p_{i+1} so that, if $|\varphi_i(t)| \leqq K$ for t in J_{i+1} then $|\varphi_i(t)| \leqq |\varphi_i(p_{i+1})| \leqq K$ for t in $[p_i, r_{i+1}]$. Similarly, $|\varphi_i(t)| \leqq K$ for t in $[s_{i+1}, q_i]$, so $|\varphi_i(t)| \leqq K$ throughout J_i.

The subsidiary induction will prove that φ_i is $O(T^{k+1+m-i})$ on J_i for all $i = 1, \ldots, k$, and for all m, from the hypothesis that the maximum of $|\varphi_k|$ on J_k is reached at one of the endpoints. The inductive step of this induction on m goes as follows. If all the φ_i are $O(T^{k+m+1-i})$ on J_i then, in particular, φ_k is $O(T^{m+1})$ on J_k and φ_{k-1} is $O(T^{m+2})$ on J_{k-1}. Moreover, all the φ_i for $i \leqq k-1$ are also $O(T^{m+2})$. Since $|\varphi_{k-1}|$ has a local maximum at p_k and at q_k, it follows from (Obs) that $\varphi_k(p_k)$ and $\varphi_k(q_k)$ are $O(T^{m+2})$. Since the maximum of $|\varphi_k|$ on J_k is reached on the boundary, we conclude that $\varphi_k(t)$ is $O(T^{m+2})$ on J_k, thereby improving the original bound $\varphi_k = O(T^{m+1})$. The interval J_k contains a point r of Z_{k-1}. So r is a local maximum of $|\varphi_{k-2}|$. Since $\varphi_1, \varphi_2, \ldots, \varphi_{k-2}$ are $O(T^{m+3})$, it follows from (Obs) that $\varphi_{k-1}(r)$ is $O(T^{m+3})$. On the other hand, $\varphi_1, \ldots, \varphi_{k-1}$ are $O(T^{m+2})$, and we have just shown that φ_k is $O(T^{m+2})$ as well. Then, from $\dot{\varphi}_{k-1} = \alpha_{k-1,1}\varphi_1 + \cdots + \alpha_{k-1,k-1}\varphi_{k-1} + \beta_{k-1}\varphi_k$, we conclude that $\dot{\varphi}_{k-1}$ is $O(T^{m+2})$ on J_k, so φ_{k-1} is Lipschitz with a constant that is $O(T^{m+2})$. Since there is one point in J_k where φ_{k-1} is $O(T^{m+3})$, it follows that φ_{k-1} is $O(T^{m+3})$ on J_k. Since bounds for φ_{k-1} on J_k extend to J_{k-1}, we conclude that φ_{k-1} is $O(T^{m+3})$ on J_{k-1}, which improves upon the bound $\varphi_{k-1} = O(T^{m+2})$ which was part of the inductive hypothesis of the subsidiary induction. Proceeding in the same way, one increases by one the exponent in the bound for φ_{k-2}, then for φ_{k-3}, and so on, until φ_1 is reached. When this happens, the inductive step of the subsidiary induction is complete. Notice that, for this induction to work, one has to prove the bounds by grouping the exponents together as indicated, i.e. by going from $\varphi_k = O(T^{m+1})$, $\varphi_{k-1} = O(T^{m+2}), \ldots, \varphi_1 = O(T^{m+k})$ to the same bounds with the exponents raised by one. The simpler idea of proving by induction on m that $\varphi_i = O(T^m)$ for $i = 1, \ldots, k$, is easily seen not to work. In fact, if we know that $\varphi_1, \varphi_2, \ldots, \varphi_{k-1}, \varphi_k$, are $O(T^m)$ on J_{k-1}, then the fact that $|\varphi_{k-1}|$ has a local maximum at p_k and at q_k only enables us to conclude, via (Obs), that $\varphi_k(p_k)$ and $\varphi_k(q_k)$ are $O(T^m)$, which we already knew anyhow, and does not improve upon the bound on φ_k.

One problem still remains, however. For the proof of the bounds $\varphi_i = O(T^{k+1+m-i})$, it is not sufficient to be able to carry out the inductive step $m \to m+1$. We must be able to start the induction by knowing, e.g. the case $m = 0$, i.e. that φ_i is $O(T^{k+1-i})$ on J_i for $i = 1, \ldots, k$. In order to do this, we will actually modify the proposition to be proved in the main induction. Instead of just proving that $\varphi_1, \ldots, \varphi_k$ are $O(T)$, we will prove that φ_1 is $O(T^k)$, φ_2 is $O(T^{k-1}), \ldots, \varphi_k$ is $O(T)$. The inductive step $k \to k+1$ therefore contains as part of the inductive hypothesis precisely the initial step $m = 0$ of the subsidiary induction. Hence there is no question now that the subsidiary induction can be carried out, and that the points $t^k_{p,q}$ of Z_{k+1} can be constructed as interior points of the intervals between consecutive points of Z_k. But now, in order to carry out the step $k \to k+1$ of the main induction, it is not enough to prove that φ_{k+1} is $O(T)$. One must go back to the estimates for $\varphi_1, \varphi_2, \ldots, \varphi_k$, and improve them, by replacing the bounds $\varphi_i = O(T^{k+1-i})$ by the stronger bounds $\varphi_i = O(T^{k+2-i})$. This, however, is easy. The inductive hypothesis of the main induction implies, in particular, that $\varphi_1, \varphi_2, \ldots, \varphi_{k-1}$ are $O(T^2)$. Using (Obs), we find that, at the local maxima of $|\varphi_{k-1}|$, φ_k is $O(T^2)$. On the other hand, $\varphi_1, \ldots, \varphi_k$ are known to be $O(T)$ by the inductive hypothesis, and φ_{k+1} is proved to be $O(T)$. Since $\dot{\varphi}_k$ is a linear combination of $\varphi_1, \ldots, \varphi_{k+1}$, it follows that φ_k is Lipschitz with a constant that is $O(T)$. Since φ_k is $O(T^2)$ on the nonempty set Z_k, it follows that φ_k is $O(T^2)$, which gives the desired improvement on the bound for φ_k. The proof that φ_{k-1} is $O(T^3)$, that φ_{k-2} is $O(T^4)$, etc., is done in the same way.

We now give the details of the proof outlined above. The main point that needs care is the explicit computation of the constants involved.

Proof of Lemma 3. We begin by specifying the choice of T. Let $\bar{A} = \max(A, 1)$. Let

$$\zeta = N(N+1)\bar{A}^N\left(1+\frac{\bar{A}}{B}\right),$$

$$\eta = N\bar{A}^N\left(N+\frac{\bar{A}}{B}\right),$$

$$\rho = 1 + N\bar{A}^N,$$

$$\lambda = \bar{A}^N, \qquad \mu = \frac{NA}{B}.$$

We let T be such that

$$\zeta T < 1, \tag{11}$$

and

$$2N^{3/2}(\eta + N\lambda)(\eta + \rho u)^N A T e^{NAT} \leqq 1. \tag{12}$$

We prove our result by contradiction. Assume that the conclusion is not true. Then there exist:

(a) an interval $[a, b]$, of length $b-a = T' \leqq T(N, A, B)$;

(b) measurable real-value functions α_{ij} $(i = 1, \dots, N, j = 1, \dots, i)$ and β_i $(i = 1, \dots, N-1)$ such that $|\alpha_{ij}(t)| \leqq A$, and $B \leqq \beta_i(t) \leqq A$ for all $t \in [a, b]$, and all i, j;

(c) absolutely continuous functions $\varphi_i : [a, b] \to \mathbb{R}$, of which at least one does not vanish identically on $[a, b]$, and that satisfy the differential equations (10);

(d) a subset Z of $[a, b]$ that contains exactly N points, and is such that $\varphi_1(t) = 0$ for $t \in Z$.

The vector-valued function $F = (\varphi_1, \dots, \varphi_N)$ is a solution of the linear homogeneous system of ordinary differential equations

$$\dot{F}(t) = M(t)F(t), \tag{13}$$

where we have written F as a column, and where the matrix M is given by

$$M = \begin{bmatrix} \alpha_{11} & \beta_1 & 0 \; \dots\dots\dots\dots & 0 \\ \alpha_{21} & \alpha_{22} & \beta_2 \, 0 \; \dots\dots\dots & 0 \\ \dots\dots\dots\dots & \dots\dots\dots\dots & \dots\dots\dots\dots & \dots\dots \\ \dots\dots\dots\dots & \dots\dots\dots\dots & \dots\dots\dots\dots & \dots\dots \\ \alpha_{N-1,1} & \alpha_{N-1,2} & \dots . \alpha_{N-1,N-1} & \beta_{N-1} \\ \alpha_{N,1} & \alpha_{N,2} & \dots\dots\dots\dots & \alpha_{NN} \end{bmatrix}. \tag{14}$$

The condition that not all the φ_i vanish identically implies that $\|F(a)\| \neq 0$ (where $\|\cdot\|$ denotes the Euclidean norm). Conditions (a)–(d) do not change if all the φ_i are multiplied by the same nonzero constant, so we can assume that $\|F(a)\| = 1$. The operator norm of M satisfies

$$\|M(t)\| \leqq NA.$$

Hence, Gronwall's inequality gives

$$\|F(t)\| \leqq e^{NA(t-a)}, \qquad t \in [a, b].$$

Therefore $\|\dot{F}(t)\| \|M(t)\| \, \|F(t)\| \leqq C$ where

$$C = NA\, e^{NAT}. \tag{15}$$

Hence $\|F(t) - F(\tau)\| \leqq C|t - \tau|$ for every t, τ in $[a, b]$. Therefore all the functions φ_j satisfy a Lipschitz condition with constant C.

We now construct a finite sequence $Z_1, \dots, Z_r$ of finite subsets of $[a, b]$. We let $Z_1 = Z$. Having defined $Z_1, \dots, Z_k$, we define Z_{k+1} as follows. For each pair (p, q) of consecutive points of Z_k, find a point $t^k_{p,q} \in [p, q]$ such that

$$|\varphi_k(t^k_{p,q})| = \max\{|\varphi_k(t)| : p \leqq t \leqq q\}. \tag{16}$$

(It is clear that such a point $t^k_{p,q}$ exists, since φ_k is continuous. Of course, $t^k_{p,q}$ need not be unique, but we can make the definition completely unambiguous by stipulating, for instance, that $t^k_{p,q}$ is the leftmost point of the set of t in $[p, q]$ where $|\varphi_k|$ reaches its maximum value.)

We then define Z_{k+1} to be the set of all the points $t^k_{p,q}$. If ν_k is the number of elements of Z_k, it is clear that $\nu_{k+1} \leqq \nu_k - 1$, since there are exactly $\nu_k - 1$ pairs of consecutive points of Z_k. We let r be the first k for which $\nu_k = 1$. Then Z_{r+1} is empty. Because $\nu_1 = N$, we see that $r \leqq N$. We will show later that $\nu_{k+1} = \nu_k - 1$, so that $r = N$ and $\nu_k = N + 1 - k$ for $k = 1, \dots, N$.

We shall use I_j, to denote the interval whose endpoints are the smallest and the largest elements of Z_j.

We define constants $D_1, \ldots, D_N, E_1, \ldots, E_N$ as follows:

$$D_1 = 0, \qquad E_1 = C, \tag{17}$$

$$E_{k+1} = \eta E_k + \rho D_k + \lambda C, \tag{18}$$

$$D_{k+1} = \mu E_{k+1} \quad \text{for } k \geqq 1. \tag{19}$$

We prove, by induction on k, that the following four facts are true for each k such that $1 \leqq k \leqq N$.

(i_k) For every $j = 1, \ldots, k-1$, and every pair p, q of consecutive points of Z_j, the point $t^j_{p,q}$ is interior to the interval $[p, q]$.

(ii_k) $\nu_j = N + 1 - j$ for $j = 1, \ldots, k$.

(iii_k) $|\varphi_j(p)| \leqq D_k T^{k+2-j}$ for $j = 1, \ldots, k$ and $p \in Z_j$.

(iv_k) $|\varphi_j(t)| \leqq E_k T^{k+1-j}$ for $j = 1, \ldots, k$ and $t \in [a, b]$.

Notice that the inductive step does not just involve proving an estimate for $\varphi_{k+1}(p)$, $\varphi_{k+1}(t)$, using the estimates for $\varphi_j(p)$, $\varphi_j(t)$ for $j \leqq k$. More than that is required. From $|\varphi_j(p)| \leqq D_k T^{k+2-j}$ and $|\varphi_j(t)| \leqq E_k T^{k+1-j}$ for $j \leqq k$, $p \in Z_j$, $t \in I_j$ we must not only prove a new estimate for $\varphi_{k+1}(p)$, $\varphi_{k+1}(t)$, but also go back to the preceding estimates and sharpen them, by substituting $D_{k+1} T^{k+3-j}$ and $E_{k+1} T^{k+2-j}$ for $D_k T^{k+2-j}$, $E_k T^{k+1-j}$, respectively.

We first prove (i_1), (ii_1), (iii_1), (iv_1). The first condition holds vacuously. The second one is trivial, since we know that Z_1 has N elements. The third one is also trivial, since φ_1 vanishes on Z_1. Finally, to prove (iv_1), recall that φ_1 is Lipschitz with constant C. Since the interval $[a, b]$ has length $T' \leqq T$ and contains a zero of φ_1, we can conclude that

$$|\varphi_1(t)| \leqq CT \quad \text{for } t \in [a, b]. \tag{20}$$

Hence (iv_1) holds.

Before we proceed to the induction step, we make the following observation: (OBS) if $|\varphi_j|$ has a local maximum at an interior point t of $[a, b]$, then

$$|\varphi_{j+1}(t)| \leqq \frac{A}{B} \sum_{i=1}^{j} |\varphi_i(t)|. \tag{21}$$

Let us prove (OBS). Assume that (21) were false. Then the inequality

$$|\varphi_{j+1}(\tau)| > \frac{A}{B} \sum_{i=1}^{j} |\varphi_i(\tau)| \tag{22}$$

must hold for all τ in an interval $(t-h, t+h)$, because the φ_i are continuous. Hence

$$\begin{aligned} |\beta_j(\tau)\varphi_{j+1}(\tau)| &\geqq B|\varphi_{j+1}(\tau)| \\ &> A \sum_{i=1}^{j} |\varphi_i(\tau)| \\ &\geqq \left| \sum_{i=1}^{j} \alpha_{ji}(\tau)\varphi_i(\tau) \right|. \end{aligned}$$

If we let $f = \beta_j \varphi_{j+1}$, $g = \sum \alpha_{ji}\varphi_i$, we have $\dot{\varphi}_j = f + g$. The inequality (22) implies, in particular that φ_{j+1} never vanishes on $(t-h, t+h)$. Since φ_{j+1} is continuous, it must have a constant sign on $(t-h, t+h)$. Since β_j is positive, the function f also has constant sign on $(t-h, t+h)$. Since $|f(\tau)|>|g(\tau)|$ for all τ, we conclude that $\dot{\varphi}_j(\tau)$ is either strictly positive for all τ, or strictly negative for all τ. Since φ_j is absolute continuous, we conclude that φ_j is strictly monotonic on $(t-h, t+h)$. This contradicts the fact that $|\varphi_j|$ has a local maximum at t. Hence (21) holds. Therefore, (OBS) is true.

We now carry out the induction step. We shall first do the $1 \Rightarrow 2$ case, and then the general $k \Rightarrow k+1$ case. So, let us assume that $N \geqq 2$, and let us prove (i_2), (ii_2), (iii_2) and (iv_2).

To prove (i_2), we must show that, if p and q are two different zeros of φ_1, then φ_1 does not vanish identically on the interval between p and q. To see this, assume that φ_1 vanishes identically on some interval I. Then the equation $\dot{\varphi}_1 = \alpha_{11}\varphi_1 + \beta_1\varphi_2$ implies that φ_2 vanishes identically on I. But then it follows from $\dot{\varphi}_2 = \alpha_{21}\varphi_1 + \alpha_{22}\varphi_2 + \beta_2\varphi_3$ that $\varphi_3 \equiv 0$ on I. Proceeding in this way, we see that all the φ_j vanish identically on I. Hence the φ_j vanish identically on $[a, b]$, since they are solutions of a homogeneous linear system. But this contradicts the assumption that not all the φ_j vanish throughout $[a, b]$.

Having proved (i_2), it is now clear that $\nu_2 = N-1$. Since we already know that $\nu_1 = N$, conclusion (ii_2) follows. Now we must prove (iii_2). Since $|\varphi_1(p)| = 0$ for $p \in Z_1$, it is clear that $|\varphi_1(p)| \leqq D_2T^3$ for such p. Now let $s \in Z_2$. Then $s = t^2_{p,q}$ for some p, q in Z_1 such that $p<q$. Since we have already proved (i_2), we know that s is an interior point of $[p, q]$. Hence $|\varphi_1|$ has a local maximum at s. By (OBS),

$$|\varphi_2(s)| \leqq \frac{A}{B}|\varphi_1(s)|.$$

On the other hand, we know that $|\varphi_1(t)| \leqq CT$ for all $t \in [a, b]$. In particular, $|\varphi_1(s)| \leqq CT$, and then $|\varphi_2(s)| \leqq (AC/B)T$. The preceding inequality is true for every $s \in Z_2$. Since $Z_2 \neq \varnothing$ (because $N \geqq 2$), and since φ_2 is also Lipschitz with constant C, we conclude that $|\varphi_2(t)| \leqq C(1+A/B)T$ for $t \in [a, b]$. But then, for $t \in [a, b]$,

$$\begin{aligned} |\dot{\varphi}_1(t)| &= |\alpha_{11}(t)\varphi_1(t) + \beta_1(t)\varphi_2(t)| \\ &\leqq ACT + AC\left(1+\frac{A}{B}\right)T \\ &= AC\left(2+\frac{A}{B}\right)T. \end{aligned}$$

So φ_1 is Lipschitz with constant $AC(2+A/B)T$. Since $[a, b]$ contains a zero of φ_1, and has length T', it follows that $|\varphi_1(t)| \leqq AC(2+A/B)T^2$. But then, if $s \in Z_2$,

$$|\varphi_2(s)| \leqq \frac{A}{B}|\varphi_1(s)| \leqq \frac{A^2C}{B}\left(2+\frac{A}{B}\right)T^2 = D_2T^2.$$

Hence (iii_2) holds. Finally, we observe that, in the course of proving (iii_2), we showed that

$$|\varphi_1(t)| \leqq AC\left(2+\frac{A}{B}\right)T^2$$

and that

$$|\varphi_2(t)| \leqq C\left(1+\frac{A}{B}\right)T \quad \text{for all } t \in [a, b].$$

Since $AC(2+A/B) \leqq E_2$ and $C(1+A/B) \leqq E_2$, conclusion (iv_2) follows. Hence, the desired result has been established for $k=2$.

We now consider the general case. We assume that (i_k), (ii_k), (iii_k), (iv_k) hold, and that $k<N$ (for, if $k=N$, the induction ends). Because of (ii_k), the set Z_k contains at least two points, so that Z_{k+1} is nonempty.

Proof that (i_k), (ii_k), (iii_k), (iv_k)$\Rightarrow$(i_{k+1}). Assume that (i_{k+1}) is false. Since (i_k) is known to hold, (i_{k+1}) can only fail to hold if there are consecutive points p, q in Z_k such that $|\varphi_k(t)|$, $t\in[p,q]$ reaches its maximum value for $t=p$ or $t=q$.

Let us agree to call any closed interval whose endpoints are in Z_j a *j-interval.* If $1<j\leqq k$, and if $J=[\sigma,\tau]$ is a j-interval then we know—because (i_k) holds—that there exist unique pairs (σ_1,σ_2), (τ_1,τ_2) of consecutive points of Z_{j-1} such that $\sigma\in[\sigma_1,\sigma_2]$ and $\tau\in[\tau_1,\tau_2]$. Let $J'=[\sigma_1,\tau_2]$. Then J' is a $(j-1)$-interval, which we shall call *the* $(j-1)$*-interval associated with J.* It is clear that $J\subseteq J'$ and that, if $J\cap Z_j$ has m points, then $J'\cap Z_{j-1}$ has $m+1$ points. Moreover, since the maximum of $|\varphi_{j-1}|$ on $[\sigma_1,\sigma_2]$ is attained at σ, and the maximum on $[\tau_1,\tau_2]$ is reached at τ, it is clear that

$$\max\{|\varphi_{j-1}(t)|:t\in J'\}=\max\{|\varphi_{j-1}(t)|:t\in J\}. \tag{23}$$

We now define a sequence $J_1,\dots,J_k$, such that each J_j is a j-interval, by backwards recursion. We start with $J_k=[p,q]$. Having defined the j-interval J_j, with $j>1$, we let J_{j-1} be the $(j-1)$-interval associated with J_j. The J_j clearly satisfy

$$J_1\supset J_2\supset\cdots\supset J_k$$

and

$$|\varphi_j(t)|\leqq\sup\{|\varphi_j(\tau)|:\tau\in J_{j+1}\}\quad\text{for } t\in J_j,\ j<k.$$

We now prove, by induction on m, that the following two estimates are satisfied for every nonnegative integer m.

$$(\dagger_m)\qquad |\varphi_j(s)|\leqq H_mT^{k+m+2-j}$$

for $s\in J_j\cap Z_j$, $1\leqq j\leqq k$.

$$(\dagger\dagger_m)\qquad |\varphi_j(t)|\leqq K_mT^{k+m+1-j}$$

for $t\in J_j$, $1\leqq j\leqq k$.

The constants H_m, K_m, are defined by

$$H_0=D_k,\qquad K_0=E_k,$$
$$K_{m+1}=(k+1)\bar{A}^kH_m+k\bar{A}^kK_m,$$
$$H_{m+1}=\frac{kA}{B}K_{m+1}.$$

Since $H_0=D_k$, $K_0=E_k$, the estimates ($\dagger_0$) and ($\dagger\dagger_0$) are simply restatements of (iii_k) (iv_k), which are being assumed. If ($\dagger_m$) and ($\dagger\dagger_m$) hold, let us prove that ($\dagger_{m+1}$) and ($\dagger\dagger_{m+1}$) are true.

Hypothesis ($\dagger_m$) implies, in particular, that $|\varphi_k(p)|$ and $|\varphi_k(q)|$ are bounded by H_mT^{m+2}. Since we are assuming that $|\varphi_k|$ attains its maximum value on the boundary of the interval $[p,q]$, it follows that

$$(\dagger\dagger_{m,0})\qquad |\varphi_k(t)|\leqq H_mT^{m+2},\qquad t\in J_k.$$

If $j<k$, then all the functions $|\varphi_j|$ are bounded by $K_mT^{k+m+1-j}$ on J_j and hence, in particular, on J_k. Since $j<k$, and $T=T(N,A,B)\leqq 1$, we have $K_mT^{k+m+1-j}\leqq K_mT^{m+2}$.

Hence $|\varphi_j(t)| \leqq K_m T^{m+2}$ for $t \in J_k$. But then

$$\begin{aligned}|\dot{\varphi}_{k-1}| &= \left|\sum_{i=1}^{k-1} \alpha_{ij}\varphi_j + \beta_k\varphi_k\right| \\ &\leqq [(k-1)AK_m + AH_m]T^{m+2}\end{aligned}$$

on the interval J_k. On the other hand, J_k contains one point τ of $J_{k-1} \cap Z_{k-1}$. For that point, estimate $(\dagger_m)$ implies that $|\varphi_{k-1}(\tau)| \leqq H_m T^{m+3}$.

Hence, since J_k has length $\leqq T$, we have the bound

$$(\dagger\dagger_{m,1}) \qquad |\varphi_{k-1}(t)| \leqq H_{m,1} T^{m+3}$$

for $t \in J_k$, where

$$H_{m.1} = (k-1)AK_m + AH_m + H_m.$$

Since $|\varphi_{k-1}(t)| \leqq \max\{|\varphi_{k-1}(s)| : s \in J_k\}$ for $t \in J_{k-1}$, we see that $(\dagger\dagger_{m,1})$ holds for all $t \in J_{k-1}$. Now suppose that we have proved the estimate

$$(\dagger\dagger_{m,\mu}) \qquad |\varphi_{k-\mu}(t)| \leqq H_{m,\mu} T^{m+\mu+2}$$

for $T \in J_{k-\mu}$, where the constants $H_{m,\mu}$ are defined by

$$\begin{aligned}H_{m,0} &= H_m, \\ H_{m,\mu+1} &= (k-1-\mu)AK_m + AH_{m,\mu} + H_m.\end{aligned}$$

We can then prove $(\dagger\dagger_{m,\mu+1})$, if $\mu < k-1$. Indeed, if we let $j = k-\mu-1$, the estimate $(\dagger\dagger_{m,\mu})$ asserts that $|\varphi_{j+1}| \leqq H_{m,\mu} T^{m+\mu+2}$ on J_{j+1}. Also, by $(\dagger\dagger_m)$, we have $|\varphi_i| \leqq K_m T^{k+m+1-i}$ on J_i. If $i \leqq j$, then $|\varphi_i|$ is bounded by $K_m T^{m+\mu+2}$ on J_i, and hence on J_{j+1} (because $i \leqq k-\mu-1$—so that $k+m+1-i \geqq m+\mu+2$—, and $T \leqq 1$).

Hence, on J_{j+1}:

$$\begin{aligned}|\dot{\varphi}_j| &= \left|\sum_{i=1}^{j} \alpha_{ji}\varphi_i + \beta_j\varphi_{j+1}\right| \\ &\leqq (jAK_m + AH_{m,\mu})T^{m+\mu+2}.\end{aligned}$$

Moreover, J_{j+1} contains one point τ which belongs to $J_j \cap Z_j$. Hence, by $(\dagger_m)$

$$|\varphi_j(\tau)| \leqq H_m T^{m+\mu+3}.$$

Therefore, for $t \in J_{j+1}$,

$$\begin{aligned}|\varphi_{k-(\mu+1)}(t)| = |\varphi_j(t)| &\leqq (jAK_m + AH_{m,\mu} + H_m)T^{m+\mu+3} \\ &= H_{m,\mu+1} T^{m+(\mu+1)+2}.\end{aligned}$$

This is exactly $(\dagger\dagger_{m,\mu+1})$, except for the fact that we have only proved it for $t \in J_{k-\mu}$. But then equality (23) shows that $(\dagger\dagger_{m,\mu+1})$ holds on $J_{k-\mu-1}$. This completes the proof, by induction, that the $(\dagger\dagger_{m,\mu})$ hold for all $\mu = 0, \ldots, k-1$. Now, it is easy to see that

$$\begin{aligned}H_{m,\mu} &= \left(\sum_{i=0}^{\mu} A^i\right)H_m + \left[\sum_{i=1}^{\mu} (k-i)A^{\mu+1-i}\right]K_m \\ &\leqq (k+1)\bar{A}^k H_m + k^2\bar{A}^k K_m \\ &= K_{m+1}.\end{aligned}$$

Hence we have proved (putting $j = k - \mu$) that

$$|\varphi_j(t)| \leqq K_{m+1} T^{k+m+2-j}$$

for $j = 1, \ldots, k$, and $t \in J_j$. But this is precisely $(\dagger\dagger_{m+1})$. To complete the induction, we now prove $(\dagger_{m+1})$. It is clear that $(\dagger_{m+1})$ holds for $j = 1$, since φ_1 vanishes on Z_1. For $1 < j \leqq k$, we use (OBS). If $\tau \in J_j \cap Z_j$, then τ is a local maximum of $|\varphi_{j-1}|$, because we are assuming that (i_k) holds. But then

$$|\varphi_j(\tau)| \leqq \frac{A}{B} \sum_{i=1}^{j-1} |\varphi_i(\tau)|.$$

Since we have already proved $(\dagger\dagger_{m+1})$, we have

$$|\varphi_i(\tau)| \leqq K_{m+1} T^{k+m+2-i}$$

for $i = 1, \ldots, j-1$. But, for such i, $T^{k+m+2-i} \leqq T^{k+m+3-j}$, because $T \leqq 1$. Then, for $\tau \in J_j \cap Z_j$, $j = 1, \ldots, k$,

$$\begin{aligned} |\varphi_j(\tau)| &\leqq (j-1)\frac{A}{B} K_{m+1} T^{k+(m+1)+2-j} \\ &\leqq \frac{kA}{B} K_{m+1} T^{k+(m+1)+2-j} \\ &= H_m T^{k+(m+1)+2-j}. \end{aligned}$$

So $(\dagger_{m+1})$ holds. This completes the proof that $(\dagger_m)$, $(\dagger\dagger_m)$ hold for all m. Now, if $m > 0$, we have

$$\begin{aligned} K_{m+1} &= (k+1)\bar{A}^k H_m + k^2 \bar{A}^k K_m \\ &= \left(\frac{k(k+1)\bar{A}^{k+1}}{B} + k^2 \bar{A}^k\right) K_m. \end{aligned}$$

Hence, for $m > 0$

$$K_m = \left(\frac{k(k+1)\bar{A}^{k+1}}{B} + k^2 \bar{A}^k\right)^m K_1.$$

Therefore

$$K_m \leqq \zeta^m K_1,$$

where $\zeta = N(N+1)\bar{A}^N(1 + \bar{A}/B)$.

Now apply $(\dagger\dagger_m)$ with $j = 1$. We find

$$|\varphi_1(t)| \leqq K_1 T^k (\zeta T)^m$$

for every $t \in J_1$, $m > 0$. Since $\zeta T < 1$, this implies that $\varphi_1(t) = 0$ for $t \in J_1$. So φ_1 vanishes identically on J_1. But this is a contradiction, because we have already shown (while proving (i_0)) that φ_1 cannot vanish identically on a nonempty open interval. This contradiction proves that (i_{k+1}) holds.

End of the induction step. Having proved (i_{k+1}), it is clear that (ii_{k+1}) follows. Let now $p \in Z_{k+1}$. Then we know that the function $|\varphi_k|$ has a local maximum at p. Hence, by (OBS)

$$|\varphi_{k+1}(p)| \leqq \frac{A}{B} \sum_{i=1}^{k} |\varphi_i(p)|.$$

But $|\varphi_i(p)| \leqq E_k T^{k+1-i}$ for $i = 1, \ldots, k$, since $p \in I_i$ for such i, and we are assuming that (iv$_k$) holds. Also, for $i = 1, \ldots, k$ we have $T^{k+1-i} \leqq T$, since $T \leqq 1$. Then

$$|\varphi_{k+1}(p)| \leqq \frac{kA}{B} E_k T.$$

Now Z_{k+1} is nonempty, since $k < N$ and we already know that $\nu_{k+1} = N - k$, because (ii$_{k+1}$) holds. Hence there is some point of $[a, b]$ where $|\varphi_{k+1}|$ is bounded by $(kA/B)E_k T$. But φ_{k+1} is Lipschitz on $[a, b]$ with constant C, and $b - a = T'$. Hence

$$|\varphi_{k+1}(t)| \leqq \left(C + \frac{kAE_k}{B}\right)T, \qquad t \in [a, b].$$

On the other hand, (iv$_k$) implies that $|\varphi_j|$ is bounded by $E_k T$ for $j \leqq k$. Hence

$$\begin{aligned} |\dot{\varphi}_k| &= \left|\sum_{i=1}^{k} \alpha_{ki}\varphi_i + \beta_k \varphi_{k+1}\right| \\ &\leqq kAE_k T + B\left(C + \frac{kAE_k}{B}\right)T \\ &= (BC + 2kAE_k)T. \end{aligned}$$

Also, there is some point $p \in Z_k$. For that point we have the inequality $|\varphi_k(p)| \leqq D_k T^2$. Hence

$$|\varphi_k(t)| \leqq (D_k + BC + 2kAE_k)T^2, \qquad t \in [a, b].$$

Assume that we have proved the inequalities

$$|\varphi_j(t)| \leqq E_{k,j} T^{k+2-j}$$

for all $t \in [a, b]$ and all $j = i+1, \ldots, k+1$, where the $E_{k,j}$ are given by

$$\begin{aligned} E_{k,k+1} &= C + \frac{kAE_k}{B}, \\ E_{k,i} &= iAE_k + AE_{k,i+1} + D_k. \end{aligned}$$

Then it follows easily that $|\varphi_i(t)| \leqq E_{k,i} T^{k+2-i}$ for $t \in [a, b]$. Indeed, the bounds $|\varphi_m(t)| \leqq E_k T^{k+1-m}$ hold for $m = 1, \ldots, i$ because of (iv$_k$). Also, for such m, $T^{k+1-m} \leqq T^{k+1-i}$. Finally, we have $|\varphi_{i+1}(t)| \leqq E_{k,i+1} T^{k+2-(i+1)} = E_{k,i+1} T^{k+1-i}$. Hence

$$\begin{aligned} |\dot{\varphi}_i| &= \left|\sum_{m=1}^{i} \alpha_{im}\varphi_m + \beta_i \varphi_{i+1}\right| \\ &\leqq (iAE_k + AE_{k,i+1}) T^{k+1-i}. \end{aligned}$$

But, if we pick a point $p \in Z_i$, we have

$$|\varphi_i(p)| \leqq D_k T^{k+2-i}.$$

Hence, if $t \in [a, b]$,

$$\begin{aligned} |\varphi_i(t)| &\leqq (iAE_k + AE_{k,i+1} + D_k) T^{k+2-i}. \\ &= E_{k,i} T^{k+2-i}. \end{aligned}$$

On the other hand, it is easy to see that

$$E_{k,i}=A^{k+1-i}C+\frac{k}{B}A^{k+2-i}E_k+\left(\sum_{j=0}^{k-1}A^j\right)D_k+\left(\sum_{j=i}^{k}jA^{j-i+1}\right)E_k.$$

Therefore

$$\begin{aligned}E_{k,i}&\leqq \bar{A}^N\left(C+\frac{N\bar{A}}{B}E_k\right)+(1+N\bar{A}^N)D_k+N^2\bar{A}^NE_k.\\&=\eta E_k+\rho D_k+\bar{A}^NC\\&=E_{k+1}.\end{aligned}$$

So we have proved that

$$|\varphi_i(t)|\leqq E_{k+1}T^{k+2-i}$$

for $i=1,\ldots,k+1$, i.e. we have proved estimate (iv_{k+1}).

To conclude the induction, we must prove (iii_{k+1}). Let $1<j\leqq k$, and let $p\in Z_j$. Then p belongs to the interior of the interval between two consecutive points q_1, q_2 of Z_{j-1}, and $|\varphi_{j-1}|$ has a local maximum at p. Hence

$$|\varphi_j(p)|\leqq\frac{A}{B}\sum_{i=1}^{j-1}|\varphi_i(p)|.$$

Now $|\varphi_i(p)|\leqq E_{k+1}T^{k+2-i}$. For $i\leqq j-1$ we have $k+2-i\geqq k+3-j$. Then $T^{k+2-i}\leqq T^{k+3-j}$, because $T\leqq 1$. Therefore

$$\begin{aligned}|\varphi_j(p)|&\leqq\frac{NA}{B}E_{k+1}T^{k+3-j}\\&=D_{k+1}T^{k+3-j}.\end{aligned}$$

This completes the proof of (iii_{k+1}). Hence the induction step is complete, and we have show that (i_k), (ii_k), (iii_k), (iv_k) hold for all $k\leqq N$. In particular, this shows that Z_j has exactly $N+1-j$ elements.

Now, for each j, we can apply estimate (iv_k) with $k=j$, and conclude that

$$|\varphi_j(t)|\leqq E_jT\quad\text{for } t\in[a,b].$$

On the other hand, an easy computation shows that

$$E_2=(\eta+\lambda)C$$

and that, for $k\geqq 2$:

$$\begin{aligned}E_k&=\eta(\eta+\rho\mu)^{k-2}C+\lambda\left[\sum_{i=0}^{k-2}(\eta+\rho\mu)^i\right]C\\&\leqq(\eta+N\lambda)(\eta+\rho\mu)^NC.\end{aligned}$$

Therefore the Euclidean norm of the vector $F(t)$ satisfies

$$\begin{aligned}\|F(t)\|&\leqq\sqrt{N}(\eta+N\lambda)(\eta+\rho\mu)^NCT\\&\leqq\tfrac{1}{2}\end{aligned}$$

for all t in $[a,b]$. But this contradicts the fact that $\|F(a)\|=1$. The proof of Lemma 3 is therefore complete.

5. End of the proof. We need another Lemma, which is a direct consequence of Lemma 3.

LEMMA 4. *Assume that the system $\dot{x} = f(x) + ug(x)$ is analytic and satisfies condition (Δ). Then every strong extremal is bang-bang. Moreover, for every compact set $K \subseteq M$ and every $T > 0$ there exists a positive integer $N(K, T)$ such that, if (γ, u) is a strong extremal defined on a time interval of length T, and γ is contained in K, then u has at most $N(K, T)$ switchings.*

Proof. Let $p \in M$. Let R_p denote the ring of germs of analytic functions at p, and let V_p be the R_p-module of germs at p of analytic vector fields. For any analytic vector field X in a neighborhood of p, let X_p denote the germ of X at p. Let W_p be the R_p-submodule of V_p generated by the germs $(\operatorname{ad} f)^i(g)_p$, $i = 0, 1, \ldots$. Then W_p is finitely generated, because R_p is noetherian and V_p is finitely generated. Hence there is a positive integer m such that

$$(\operatorname{ad} f)^{m+1}(g)_p = \sum_{j=0}^{m} a_j (\operatorname{ad} f)^j (g)_p$$

for some germs $a_0, \ldots, a_m \in R_p$. Therefore there is a neighborhood U_p of p where there exist analytic real-valued functions $b_0, \ldots, b_m$, such that

$$(\operatorname{ad} f)^{m+1}(g)(x) = \sum_{j=0}^{m} b_j(x)(\operatorname{ad} f)^j(g)(x) \tag{24}$$

for all $x \in U_p$.

Because of condition (Δ), there exists a neighborhood U'_p of p where the identities

$$[g, (\operatorname{ad} f)^j(g)] = \sum_{i=0}^{j} \delta_{ij} (\operatorname{ad} f)^i (g) + \eta_j (\operatorname{ad} f)^{j+1}(g) \tag{25}$$

are valid for $j = 0, \ldots, m$, the δ_{ij}, η_j being analytic functions on U'_p such that $|\eta_j(x)| < 1$ for all $x \in U'_p$.

By shrinking U_p, if necessary, we can assume that the closure of U_p is compact and contained in U'_p. Then the functions δ_{ij} are bounded in absolute value on U_p, by a constant E, and the $|\eta_j|$ are bounded by a constant $E' < 1$. By further shrinking U_p, if needed, we can assume that the b_j are bounded by a constant E''.

Now let (γ, u) be an arbitrary admissible pair, defined on an interval of length $T' \leqq T$. Suppose that (γ, u) is a strong extremal, and that γ is contained in U_p. We will show that (γ, u) is bang-bang, and that it has at most N switchings, where the number N does not depend on (γ, u).

Let (γ, u) be defined on the interval $I = [a, a + T']$. Let $t \to \lambda(t)$ be an adjoint solution which is nontrivial on $L_0(\gamma(t))$ and satisfies $H(\lambda(t), \gamma(t), u(t)) \leqq H(\lambda(t), \gamma(t), v)$ for all $v \in [-1, 1]$, and almost all $t \in I$. Define functions $\varphi_j : I \to R$ by

$$\varphi_j(t) = \langle \lambda(t), (\operatorname{ad} f)^{j-1}(g)(\gamma(t)) \rangle$$

for $j = 1, 2, \ldots$.

Then the φ_j are absolutely continuous on I, and the time derivative of φ_j is given by

$$\dot{\varphi}_j(t) = \varphi_{j+1}(t) + u(t) \langle \lambda(t), [g, (\operatorname{ad} f)^{j-1}(g)](\gamma(t)) \rangle.$$

For $j = 1, \ldots, m + 1$, we can write

$$[g, (\operatorname{ad} f)^{j-1}(g)] = \sum_{i=0}^{j-1} \delta_{ij-1} (\operatorname{ad} f)^i (g) + \eta_{j-1} (\operatorname{ad} f)^j (g)$$

so that

$$(26)\qquad \dot{\varphi}_j(t)=\varphi_{j+1}(t)+\sum_{i=0}^{j-1}u(t)\delta_{ij-1}(\gamma(t))\varphi_i(t)+u(t)\eta_{j-1}(\gamma(t))\varphi_{j+1}(t).$$

For $j=1,\ldots,m$ (but *not* for $j=m+1$), put

$$\alpha_{ji}(t)=u(t)\delta_{i-1,j-1}(\gamma(t)),$$

$$\beta_j(t)=1-u(t)\eta_{j-1}(\gamma(t)).$$

Then the equations

$$\dot{\varphi}_j=\sum_{i=1}^{j}\alpha_{ji}\varphi_i+\beta_j\varphi_{j+1}$$

hold for $j=1,\ldots,m$.

For $j=m+1$, equation (26) contains φ_{m+2} on the right side, but we can express φ_{m+2} in terms of $\varphi_1,\ldots,\varphi_{m+1}$ using equation (24). The final result is

$$\dot{\varphi}_{m+1}=\sum_{i=1}^{m+1}\alpha_{m+1,i}\varphi_i.$$

where

$$\alpha_{m+1,i}(t)=b_{i-1}(\gamma(t))+u(t)(\delta_{im}(\gamma(t))+\eta_m(\gamma(t))b_{i-1}(\gamma(t))).$$

It is clear that

$$|\alpha_{ji}(t)|\leqq E$$

for $j=1,\ldots,m$, $i=1,\ldots,j$, and that

$$1-E'\leqq\beta_j(t)\leqq 1+E'$$

for $j=1,\ldots,m$.

Also

$$|\alpha_{m+1,i}(t)|\leqq E''+E+E'E''.$$

So, if we let A be the largest of $1+E'$ and $E''+E+E'E''$, and $B=1-E'$, we see that the $|\alpha_{ji}|$ are bounded by A, and that

$$0<B\leqq\beta_j(t)\leqq A$$

for all t.

The constants A, B do not depend on the particular choice of our strong extremal, but only on the constants E, E', E'', which depend only on the neighborhood U_p. The hypotheses of Lemma 3 are satisfied, with $N=m+1$. Let T_0 be the time $T(N,A,B)$ whose existence is assured by Lemma 3. Then either (a) the function φ_1 has at most m zeros on any subinterval of I of length $\leqq T_0$, or (b) all the φ_i, $i=1,\ldots,m+1$ vanish identically on I. Assume (a) holds. Let ν be the smallest integer such that $\nu T_0\geqq T$. Then φ_1 has at most νm zeros on I. In any interval J between consecutive zeros of φ_1, we have $\langle\lambda(t),g(\gamma(t))\rangle=\varphi_1(t)\neq 0$. Therefore, the condition that $H(\lambda(t),\gamma(t),v)$ is minimized for $\nu=u(t)$ implies that u is constant on J, and has the value $+1$ if $\varphi_1(t)<0$, -1 if $\varphi_1(t)>0$. So (γ,u) is bang-bang with at most νm switchings.

We must now exclude case (b). Assume that $\varphi_1,\ldots,\varphi_{m+1}$ vanish identically. Let Q be the set of all vector fields $X\in L$ whose restriction to U_p can be written as a linear combination of g, $(\operatorname{ad}f)(g),\ldots,(\operatorname{ad}f)^m(g)$ with analytic coefficients. If $X\in Q$ then, on

U_p, we have

$$X=\sum_{i=0}^{m} h_i(\operatorname{ad} f)^i(g). \tag{27}$$

Then

$$[f,X]=\sum_{i=0}^{m}(fh_i)\cdot(\operatorname{ad} f)^i(g)+\sum_{i=0}^{m} h_i(\operatorname{ad} f)^{i+1}(g).$$

By (24), $(\operatorname{ad} f)^{m+1}(g)$ can be written as a linear combination of g, $(\operatorname{ad} f)(g), \dots$, $(\operatorname{ad} f)^m(g)$. So we see that $[f,X]\in Q$. A similar reasoning, using (25) instead of (24), shows that $[g,X]\in Q$. Hence Q is an ideal of L. Since $g\in Q$, it follows that $L_0\subseteq Q$. Now, if $X\in Q$ has an expression of the form (27), then

$$\langle\lambda(t), X(\gamma(t))\rangle=\sum_{i=0}^{m} h_i(\gamma(t))\varphi_{i+1}(t)=0$$

since we are assuming that $\varphi_1,\dots,\varphi_{m+1}$ vanish identically. Since $L_0\subseteq Q$, $\lambda(t)=0$ on $L_0(\gamma(t))$. But λ was chosen so that $\lambda(t)$ is nontrivial on $L_0(\gamma(t))$, so we have reached a contradiction. Hence the $\varphi_1,\dots,\varphi_{m+1}$ cannot all vanish identically, and possibility (b) is eliminated.

Now let K be an arbitrary compact set, and let $T>0$. We cover K with a finite number $U_{p_1},\dots,U_{p_n}$ of neighborhoods of the form U_p as constructed above. With respect to some Riemannian metric on M, the vectors $f(x)$, $g(x)$ are bounded for $x\in K$. Hence there is a constant C such that, if γ is a trajectory of $x=f+ug$, and γ is contained in K, then $d(\gamma(t_1),\gamma(t_2))\leqq C(t_1-t_2)$ for all t_1, t_2 in the domain of γ. (Here d is the distance.) Let $\delta>0$ be a Lebesgue number for the covering $\{U_{p_1},\dots,U_{p_n}\}$ of K. Let $\varepsilon=\delta/C$. Let $N_1,\dots,N_n$ be such that, whenever (γ,u) is a strong extremal contained in U_{p_1}, and defined on an interval of length $\leqq\varepsilon$, then (γ,u) is bang-bang with at most N_i switchings. Let $N=\max(N_i,\dots,N_n)$. Let μ be such that $\mu\varepsilon\geqq T$.

Then, if (γ,u) is a strong extremal contained in K, and defined on an interval I of length $\leqq T$, we can partition I into at most μ intervals I_j of length $\leqq\varepsilon$. The restriction (γ_j,u_j) of (γ,u) to I_j is also a strong extremal, and $\gamma_j(I_j)$ has diameter $\leqq C\varepsilon=\delta$. Hence γ_j is entirely contained in one of the U_{p_i}. Hence (γ_j,u_j) is bang-bang with at most μ_N switchings. If we let $N(K,T)=\mu N$, the proof is complete.

We are now ready to end the proof of our main theorem. Choose $u_0=1$. By Lemma 2, there is a compact $K'\supseteq K$, and an integer N_1, such that (K,K',T,u_0) has the strong extremal replacement property with N_1 steps. By Lemma 4, there is an integer N_2 such that every strong extremal contained in K', and defined over a time interval of length $\leqq T$, is bang-bang with at most N_2-1 switchings. If p can be steered time-optimally to q in time T by a trajectory in K, then p can be steered time-optimally to q by a trajectory in K' that is the concatenation of at most N_1 pieces, each of whom is either a strong extremal or constant bang-bang. Each of the strong extremals is the concatenation of at most N_2 constant bang-bang trajectories. Hence p can be steered to q by a concatenation of at most N_1N_2 constant bang-bang controls, i.e. by a bang-bang control with at most N_1N_2-1 switchings. Q.E.D.

Acknowledgment. The author is indebted to an anonymous referee for suggesting an improvement to the proof of Lemma 2.

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] A. J. KRENER, *A generalization of Chow's theorem and the bang-bang theorem to nonlinear control problems*, this Journal, 12 (1974), pp. 43–51.

[3] T. NAGANO, *Linear differential systems with singularities and an application to transitive Lie algebras*, J. Math. Soc. Japan, 18 (1966), pp. 398–404.

[4] H. J. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.

[5] J. J. SUSSMANN, *Analytic stratifications and optimal feedback control*, Proc. 1978 International Congress of Mathematicians (Helsinki), to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 5, September 1979

A BOUND ON THE BOUNDARY INPUT MAP FOR PARABOLIC EQUATIONS WITH APPLICATION TO TIME OPTIMAL CONTROL*

DON WASHBURN†

Abstract. A semigroup formulation of boundary input problems for systems governed by parabolic partial differential equations is presented. A useful bound on the operator kernel of the input map is established under very general conditions. This bound is used to study the input map and the results used to examine the time optimal boundary control problem.

1. Introduction. A semigroup approach to modeling boundary input problems for linear partial differential equations was first presented by Fattorini in [7]. This approach was extended by Balakrishnan in [1] and [2][1] and used to solve the linear quadratic regulator problem for parabolic equations with boundary control in [3]. Key to this approach is the existence of a useful bound on the growth rate of the operator kernel of the boundary input map. Such a bound was first conjectured by Balakrishnan based on an examination of special cases. It was studied in some detail by the author in [14] and [15]. This paper, after establishing notation and setting in § 2, presents the best possible bound in an abstract sense in § 3. Here it is shown that the bound depends on the Greens map and the choice of spaces for the problem. In § 4, we show that a large class of practical problems is covered by the abstraction in 3. Section 4 then uses the input map plus the bound to briefly study the time optimal boundary control problem.

2. Canonical example. Let Ω be a bounded open subset of $\mathbb{R}^n$ with boundary Γ. We consider a system governed by the heat equation on Ω with input applied on Γ, i.e.,

$$\frac{\partial f}{\partial t} = \Delta f \text{ in } \Omega; \qquad f = u \text{ on } \Gamma; \qquad f = 0 \text{ when } t = 0. \tag{2.1}$$

Let G denote the *Greens Map* for the problem defined by $Gv = g$ where:

$$\Delta g = 0 \text{ in } \Omega; \qquad g = v \text{ on } \Gamma.$$

We require Ω to be such that $G: L^2(\Gamma) \to L^2(\Omega)$ continuously; a mild restriction which in particular allows corners [11, p. 250].

Let A be the closed restriction of the Laplacian Δ defined by

$$Ag = \Delta g \text{ for } g \in \mathcal{D}(A) = H_0^1(\Omega) \cap \mathcal{D}(\Delta)$$

where

$$\mathcal{D}(\Delta) = \{u/u, \Delta u \in L^2(\Omega)\}$$

and $H_0^1(\Omega)$ is the 1st order Sobolev space of functions vanishing on Γ. Then A generates a strongly continuous semigroup in $L^2(\Omega)$, denote it by $S(t)$. Recall that $S(t)$ is analytic and compact.

Initially we restrict u to be in $C_0^\infty(0, T, ; C(\Gamma))$, where the sub zero denotes compact support in $(0, T)$. In this case $Gu \in C_0^\infty(0, T, ; C(\Omega))$ and we reformulate (2.1)

* Received by the editors June 27, 1978, and in revised form January 29, 1979.

† Department of Mathematical Sciences, United States Air Force Academy, Colorado. Now at the Air Force Weapons Laboratory, Kirtland Air Force Base, New Mexico 87115.

[1] R. Triggiani has recently used a similar cosine operator approach to study hyperbolic boundary input problems [13].

as follows:

$$\frac{\partial(f-Gu)}{\partial t}=\Delta(f-Gu)-\frac{\partial Gu}{\partial t} \quad \text{in } \Omega$$

$$f-Gu=0 \text{ on } \Gamma; \qquad f-Gu=0 \text{ at } t=0.$$

If we set $w=f-Gu$ then we have the familiar Banach space formulation:

$$\frac{dw}{dt}=Aw-\frac{dGu}{dt}; \qquad w(0)=0; \qquad w\in\mathcal{D}(A)$$

where the derivatives are in the $L^2(\Omega)$ norm. Notice that the boundary conditions are now contained in the statement $w\in\mathcal{D}(A)$. This problem has the well known solution given by:

$$w(t)=-\int_0^t S(t-\tau)\frac{d}{d\tau}(Gu)(\tau)\,d\tau$$

or

$$f(t)=-\int_0^t S(t-\tau)\frac{d}{d\tau}(Gu)(\tau)\,d\tau+(Gu)(t).$$

This is essentially the solution of Fattorini [7]. Under the assumptions on u, the right hand side may be integrated by parts to obtain:

$$f(t)=-\int_0^t AS(t-\tau)(Gu)(\tau)\,d\tau. \tag{2.2}$$

The solution (2.2) was obtained by assuming $u\in C_0^\infty$ $(0, T; C(\Gamma))$, an awkward restriction for most control problems. Fortunately, we can show that the map

$$u\to f$$

given by (2.2) is bounded when considered as a map

$$L^2(0, T; L^2(\Gamma))\to L^2(0, T; L^2(\Omega)).$$

Therefore, we may extend our notion of solution to the case where

$$u\in L^2(0, T; L^2(\Gamma)).$$

For arbitrary $u\in L^2(0, T; L^2(\Gamma))$ f given by (2.2) is now a "generalized" solution of (2.1) and is no longer defined pointwise, i.e.,

$$f(t)\in L^2(\Omega) \quad \text{only a.e. } t.$$

Note that a.e. t is in fact the best we can do since there exist example where $f(t)\notin L^2(\Omega)$ for some values of t. See example in [17, p. 202].

This is untenable for time optimal and final value problems and thus we must restrict u to lie in a smaller set. In this regard Balakrishnan [1] established that

$$|AS(t)G|=O(t^{-3/4}) \qquad (O(\cdot)\text{ is the standard Landau symbol for } t\to 0) \tag{2.3}$$

for the case where Ω is a square by using eigenfunction expansions and known growth rates on the eigenvalues of Δ.

Such a bound on the "kernel" of the operator in (2.2) allows one to establish that if $u\in L^p(0, T; L^2(\Gamma))$ for $p>4$ then $f(t)\in L^2(\Omega)$ $\forall t$, and permits a detailed analysis of the operator in (2.2).

In the next section we will present general results concerning a bound such as (2.3) and study the properties of maps such as (2.2).

3. Abstraction and a study of the boundary input map.

3.1. Generality of $|AS(t)G| = O(t^{-\theta})$. Let V and E be Hilbert spaces (corresponds to $L^2(\Gamma)$ and $L^2(\Omega)$ respectively). Let $G\colon V \to E$ be bounded (corresponds to the *Greens map*). Let $A\colon \mathcal{D}(A) \subset E \to E$ be the infinitesimal generator of an analytic semigroup, denoted $S(t)$, defined on E.

Question. What is the behavior of $|AS(t)G|$ for small t?

Notice that if G is onto E then the best behavior to hope for is that $|AS(t)G| = O(t^{-1})$, which follows from the analyticity of the semigroup. However, if G maps into an appropriate subspace of E, then we can hope for better behavior on $|AS(t)G|$. To study this proboem we need to review intermediate spaces between two Hilbert spaces.

Recall that if X and E are two Hilbert spaces with X densely and continuously embedded in E, then there exists a positive, self-adjoint operator, Λ, unbounded in E with domain X and range E. The definition of the fractional powers of such operators, Λ^θ, is well known [6, p. 1196]. We may, therefore, define intermediate spaces between X and E, denote $[X, E]_\theta$, by $[X, E]_\theta = \mathcal{D}(\Lambda^{1-\theta})$, for $\theta \in [0, 1]$. The spaces $[X, E]_\theta$ are independent of the choice of Λ and are thus well defined subspaces of E. When X is the domain of the generator of a bounded strongly continuous semigroup in E, we have the following alternate characterization of these spaces.

THEOREM A. *Let A be the generator of an equi-bounded strongly continuous semigroup $S(t)$ defined in a Hilbert space E. Then for $\theta \in (0, 1)$ we have*

$$f \in [\mathcal{D}(A), E]_{1-\theta} \quad \textit{if and only if} \quad \int_0^\infty \left|\frac{S(t)f - f}{t^\theta}\right|^2 \frac{dt}{t} < \infty.$$

The proof of this familiar result may be found in [9, p. 48]. In order to prove our basic theorem we will also need the following nontrivial result which is an immediate consequence of a theorem of Butzer and Berens [4, p. 195].

THEOREM B. *If $S(t)$ is an equi-bounded strongly continuous semigroup in E and $\theta \in (0, 1)$ then*

$$\int_0^\infty \left|\frac{S(t)f - f}{t^\theta}\right|^2 \frac{dt}{t} < \infty \quad \textit{implies} \quad |S(t)f - f| \leqq M_f t^\theta.$$

The basic theorem concerning $AS(t)$ is Theorem 1 below.

THEOREM 1. *Let $S(t)$ be a strongly continuous analytic semigroup in E with generator A. Then*

(a) $|AS(t)f| = O(t^{\theta-1})$ *for some* $\theta \in (0, 1)$ *implies* $f \in [\mathcal{D}(A), E]_{1-\theta_0}$, $\theta_0 \in (0, \theta)$;

(b) $f \in [\mathcal{D}(A), E]_{1-\theta}$ *for some* $\theta \in (0, 1)$ *implies* $|AS(t)f| = O(t^{\theta-1})$.

Proof. Initially assume that $S(t)$ is equi-bounded and thus Theorems A and B apply.

Proof of (a). Assume $|AS(t)f| = O(t^{\theta-1})$. Then given $\varepsilon > 0$ there exists C such that $|AS(t)f| < Ct^{\theta-1}$ for $t \in (0, \varepsilon)$. Since

$$S(t)f - f = \int_0^t AS(s)f\, ds$$

we have $|S(t)f - f| \leqq C \int_0^t s^{\theta-1}\, ds = (C/\theta)t^\theta$ for $t \in (0, \varepsilon)$; thus

$$\int_0^\varepsilon \left|\frac{S(t)f - f}{t^{\theta_0}}\right|^2 \frac{dt}{t} = \frac{C^2}{\theta^2} \int_0^\varepsilon \frac{t^{2\theta}}{t^{2\theta_0}} \frac{dt}{t} < \infty$$

and since $S(t)$ is equi-bounded $\int_0^\infty |(S(t)f-f)/t^{\theta_0}|^2\,dt/t<\infty$. Theorem A then implies $f\in[\mathscr{D}(A),E]_{1-\theta_0}$ and (a) is proved.

Proof of (b). Let $f\in[\mathscr{D}(A),E]_{1-\theta}$ then $\int_0^\infty |(S(t)f-f)/t^{\theta}|^2\,dt/t<\infty$ follows by Theorem A. Thus, $\int_0^t |(S(t)f-f)/t^{\theta}|^2\,dt/t<\infty$ and Theorem B implies that $|S(t)f-f|<Ct^{\theta}$. Note that $S(t)=\sum_{j=0}^n S(2^jt)(I-S(2^jt))+S(2^{n+1}t)$ which follows by induction using the semigroup property of $S(t)$. Thus,

$$|AS(t)f|\leqq\sum_{j=0}^{n}|AS(2^jt)|\,|I-S(2^jt)f|+|AS(2^{n+1}t)|\,|f|.$$

Using Theorem B and the fact that analyticity implies $|AS(t)|\leqq C/t$, we have

$$|AS(t)f|\leqq\sum_{j=0}^{n}\left(\frac{C}{2^jt}\right)(M_f(2^jt)^{\theta})+\frac{C|f|}{2^{n+1}t}.$$

Therefore,

$$t^{1-\theta}|AS(t)f|\leqq C\left(\sum_{j=0}^{n}\frac{M_f}{(2^j)^{1-\theta}}+\frac{|f|}{2^{n+1}t^{\theta}}\right)$$

which holds for all n. It follows that

$$t^{1-\theta}|AS(t)f|\leqq K_f\quad\text{or}\quad|AS(t)f|\leqq K_ft^{1-\theta}$$

and (b) is proven.

Removal of equi-bounded restriction. Note that if B is the infinitesimal generator of an unbounded semigroup $T(t)$ with $|T(t)|\leqq Ce^{\omega t}$, then $A=B-\omega I$ is the generator of the equi-bounded semigroup $S(t)=e^{-\omega t}T(t)$ [16, p. 232]. Suppose that $|BT(t)f|=O(t^{\theta-1})$, then

$$|AS(t)f|=|(B-\omega I)\,e^{-\omega t}T(t)f|\leqq|BT(t)f|\,e^{-\omega t}+\omega\,e^{-\omega t}|T(t)f|\leqq Mt^{\theta-1}+K.$$

Therefore,

$$|AS(t)f|=O(t^{\theta-1})\quad\text{or}\quad f\in[\mathscr{D}(A),E]_{1-\theta_0}=[\mathscr{D}(B),E]_{1-\theta_0}.$$

Thus, (a) holds for nonequi-bounded semigroups. Now let $f\in[\mathscr{D}(B),E]_{1-\theta}=[\mathscr{D}(A),E]_{1-\theta}$. Then $|AS(t)f|=O(t^{\theta-1})$. Therefore,

$$|BT(t)f|=|(A+\omega I)\,e^{\omega t}S(t)f|\leqq|AS(t)f|\,e^{\omega t}+|\omega\,e^{\omega t}S(t)f|=O(t^{\theta-1})$$

and (b) holds for nonequi-bounded semigroups. Q.E.D.

THEOREM 2. *Let A, $S(t)$ be as in Theorem* 1, *and let G be a bounded map from V into E.*

(a) *If $|AS(t)G|\leqq K(t^{\theta-1})$ for some $0<\theta<1$, then G: $V\to[\mathscr{D}(A),E]_{1-\theta_0}$, $0<\theta_0<\theta$;*

(b) *If G: $V\to[\mathscr{D}(A),E]_{1-\theta}$ for $0<\theta<1$, $|AS(t)G|\leqq Mt^{\theta-1}$.*

Proof of (a). $|AS(t)Gx|\leqq Kt^{\theta-1}|x|=O(t^{\theta-1})$ for all $Gx\in V$. Thus, by Theorem 1 we have $Gx\in[\mathscr{D}(A),E]_{1-\theta}$ $\forall x\in V$ and (a) is proven.

Proof of (b). If G: $V\to[\mathscr{D}(A),E]_{1-\theta}$, then $Gx\in[\mathscr{D}(A),E]_{1-\theta}$ $\forall x\in V$ and $|AS(t)Gx|=O(t^{\theta-1})$ $\forall x\in V$ follows by Theorem 1. Thus, for $0<t<1$ there exists M_x such that $|AS(t)Gx|\leqq M_x(t^{\theta-1})$ or $|t^{1-\theta}AS(t)Gx|\leqq M_x$ and by the uniform boundedness principle there exists an M independent of x such that $|t^{1-\theta}AS(t)G|\leqq M$. Equivalently, $|AS(t)G|\leqq Mt^{\theta-1}$ and (b) is proven. Q.E.D.

Note. The result in (a) is the best possible in the sense that there exists examples where $|AS(t)G|=O(t^{\theta-1})$ but $GV\not\subset[\mathscr{D}(A),E]_{1-\theta}$. See Appendix A.

3.2. A look at the input map. We will use the following notation throughout this chapter: If H is a Hilbert space and $1 \leqq p < \infty$ then $W^P(H) = L^P((0T), H) =$ the space of strongly measurable H valued functions on $(0, T)$ such that

$$|u|_{W^P(H)} = |u(\cdot)|_{W^P(H)} = \int_0^t |u(t)|_H^P \, dt^{1/P} < \infty$$

when $1 \leqq P < \infty$, and when $P = \infty$

$$|u|_{W^\infty(H)} = |u(\cdot)|_{W^\infty(H)} = \operatorname*{ess\,sup}_{t \in (0,\, T)} |u(t)|_H < \infty.$$

Further, if $u \in L^P(0, T; V) \triangleq W^P(V)$. Define L by $Lu = f$ where $f(t) = \int_0^t AS(t-\tau)(Gu)(\tau)\, d\tau$. Define L_T by $L_T u = (Lu)(T) = \int_0^T AS(T-\tau)(Gu)(\tau)\, d\tau$ when the above makes sense.

THEOREM 3. *If $|AS(t)G| = O(t^{\theta-1})$, for some $0 < \theta < 1$, then L_T: $W^P(V) \to E$ continuously for all P such that $1/\theta < P \leqq \infty$.*

Proof.

$$\begin{aligned} |L_T v| &\leqq \int_0^T |AS(T-\tau)Gv(\tau)|\, d\tau \\ &\leqq M \int_0^T (T-\tau)^{\theta-1} |v(\tau)|\, d\tau \\ &\leqq M\left(\int_0^T ((T-\tau)^{\theta-1})^q\, d\tau\right)^{1/q} \left(\int_0^T |v(\tau)|^P\, d\tau\right)^{1/P} \end{aligned}$$

where $1/P + 1/q = 1$. Evaluating the first integral we have

$$\int_0^T ((T-\tau)^{\theta-1})^q\, d\tau = C(T-\tau)^{(\theta-1)q+1}|_0^T = CT^{(\theta-1)q+1}$$

which is finite if and only if $(\theta-1)q + 1 > 0$ or $q < 1/(1-\theta)$. Therefore, if the second integral is finite for $P > 1/\theta$, then $|L_T v| < \infty$. For the case $P = \infty$, we have

$$|L_T v| \leqq M \int_0^T (T-\tau)^{\theta-1}\, d\tau |v(\cdot)|_{W^\infty(V)} \leqq C|v(\cdot)|_{W^\infty(V)}. \qquad \text{Q.E.D.}$$

Note. Since θ is typically $\simeq \frac{1}{4}$, we have $1/\theta \simeq 4$ in the above theorem. We have already pointed out that L: $W^2(V) \to W^2(E)$ continuously. We further point out that L: $W^P(V) \to W^\infty(E)$ continuously when $1/\theta < P \leqq \infty$.

The following theorem establishes the compactness of the map L_T: $W^\infty(V) \to E$. This fact, one of the distinguishing features of problems with boundary input, causes considerable difficulties in the study of the time optimal control problem since it implies that the set of attainable states has empty interior. Therefore, one only has a weakened version of the separation theorems available.

THEOREM 4. *Let A, $S(t)$, G be as above, with $|AS(t)G| = O(t^{\theta-1})$. If G is compact, then L_T: $W^P(V) \to E$ is compact $\forall P > 1/\theta$.*

Proof. Since for us range of G is always separable, we give a proof using this fact. In this case, G compact implies it is the uniform limit of a sequence of operators, G_n, with finite dimensional range. We first show that L_ε defined by

$$L_\varepsilon u = \int_0^{T-\varepsilon} AS(T-\tau)Gu(\tau)\, d\tau$$

is compact by showing it is the uniform limit of a sequence of operators L_ε^n defined by

$$L_\varepsilon^n u = \int_0^{T-\varepsilon} AS(T-\tau)G_n u(\tau)\,d\tau$$

which obviously have finite dimensional range and are, therefore, compact. Then we show that L_T is the uniform limit of L_ε as $\varepsilon \to 0$, therefore, L_T is also compact. In what follows we assume $u \in W^P(V)$ for $P > 1/\theta$. First L_ε:

$$\begin{aligned}|(L_\varepsilon - L_\varepsilon^n)u| &\leqq \int_0^{T-\varepsilon} |AS(T-\tau)|\,|G-G_n|\,|u(\tau)|\,d\tau \\ &\leqq \int_0^{T-\varepsilon} \frac{K}{|T-\tau|}|u(\tau)|\,d\tau|G-G_n| \\ &\leqq \frac{K}{\varepsilon}|u|_{W^P(V)}|G-G_n| \\ &\leqq C|u|_{W^P(V)}|G-G_n| \to 0.\end{aligned}$$

Thus, $|L_\varepsilon - L_\varepsilon^n| \to 0$ as $n \to \infty$ and L_ε is compact. Now L_T:

$$\begin{aligned}|(L_T - L_\varepsilon)u| &\leqq \int_{T-\varepsilon}^{T} |AS(T-\tau)G|\,|u(\tau)|\,d\tau \\ &\leqq \int_{T-\varepsilon}^{T} K(T-\tau)^{\theta-1}|u(\tau)|\,d\tau \\ &\leqq \left(\int_{T-\varepsilon}^{T} (K(T-\tau)^{\theta-1})^q\,d\tau\right)^{1/q}\left(\int_{T-\varepsilon}^{T} |u(\tau)|^P\,d\tau\right)^{1/P} \\ &\leqq C(T-\tau)^{(\theta-1)q+1}|_{T-\varepsilon}^{T}|u|_{W^P(V)}.\end{aligned}$$

Since $P > 1/\theta$ and $1/P + 1/q = 1$, $(\theta-1)q+1>0$ which implies

$$|(L_T - L_\varepsilon)u| \leqq C\varepsilon^{(\theta-1)q+1}|u|_{W^P(V)} \to 0.$$

Thus, $|L_T - L_\varepsilon| \to 0$ as $\varepsilon \to 0$ and L_T is compact. Q.E.D.

The following two lemmas characterize L_T', the dual of L_T, which is required in the study of the time optimal control problem in § 4.

We assume throughout that $|AS(t)G| \leqq M/t^{1-\theta}$ and recall from Theorem 3 that

$$L_T\colon W^\infty(V) \to E$$

continuously, where L_T is given by

$$L_T\colon v(\cdot) \to \int_0^T AS(T-\tau)v(\tau)\,d\tau.$$

Therefore

$$L_T'\colon E \to (W^\infty(V))' \qquad \text{(Yosida [16, p. 195])}$$

continuously. (Note $E' = E$.)

LEMMA 1. *$L_T'\colon E \to W^1(V) \subset (W^\infty(V))'$ is given by $(L_T'y)(t) = (AS(t-t)G)^*y$.*

Proof. L_T' is defined by

$$(L_T v, y)_{E\times E'} = \langle v, L_T'y\rangle_{W^\infty(V)\times(W^\infty(V))'}$$

where we take $(\cdot,\cdot)_{E\times E'} = (\cdot,\cdot)_E =$ the inner product in E. We let the pairing

$\langle \cdot, \cdot \rangle_{W^\infty(V)\times(W^\infty(V))'}$ be the extension of

$$\langle v, u\rangle = \int_0^T (v(\tau), u(\tau))_V \, d\tau$$

defined on $W^\infty(V)\times W^1(V)$.

Therefore

$$\begin{aligned}(L_T v, y)_E &= \int_0^T (AS(T-\tau)Gv(\tau), y)_E \, d\tau \\ &= \int_0^T (v(\tau), (AS(T-\tau)G)^* y)_V \, d\tau \\ &= \langle v, (AS(T-\cdot)G)^* y\rangle \\ &= \langle v, (AS(T-\cdot)G)^* y\rangle_{W^\infty(V)\times(W^\infty(V))'}\end{aligned}$$

for all $y \in E$ since $(AS(T-\cdot)G)^* y \in W^1(V)$ for all $y \in E$. From this we see that

$$(L'_T y)(t) = (AS(T-t)G)^* y \quad \forall y \in E$$

and that L'_T actually is continuous into $W^1(V)$. Q.E.D.

LEMMA 2. *If $S(t)$ is analytic in a sector containing $t>0$ then $L'_T y$ is analytic in a sector containing $t<T$.*

Proof. Let $S(t)$ be analytic in t for $t \in \Delta_\phi$, where:

$$\Delta_\phi = \{\text{open sector in the complex plane containing } t>0\}.$$

Then

$$AS(t) \text{ is analytic in } \Delta_\phi,$$

i.e.,

$$(AS(t)x, y) = (x, (AS(t))^* y)_E \quad \text{is analytic in } \Delta_\phi.$$

In particular then

$$(Gv, (AS(t))^* y)_E = (AS(t)Gv, y)_E = (v, (AS(t)G)^* y)_V$$

is analytic in Δ_ϕ. Therefore, both $(AS(t)G)^*$ and $AS(t)G$ are analytic in Δ_ϕ.

It follows that $L'_T = (AS(T-\cdot)G)^*$ is analytic in $T-\Delta_\phi$, a sector containing $t<T$. Q.E.D.

4. Classes of problems covered. In order to show the broad applicability of the results of the preceding section, we discuss a class of evolution equations in which the map G satisfies the conditions of Theorem 2.

To this end we require Ω to be *smooth* by which we mean it is a bounded open subset of $\mathbb{R}^n$ with an infinitely differentiable boundary, Γ, and such that Ω lies totally on one side of Γ. The Sobolev spaces $H^s(\Omega)$, $H^s(\Gamma)$, etc., are standard and as described in [9, p. 1]. We require τ to be a 2nd order uniformly strongly elliptic operator with real $C^\infty(\bar{\Omega})$ coefficients, i.e., τ has the form:

$$\tau(x, D)u = \sum_{|\alpha|\leqq 2} a_\alpha(x) D^\alpha u, \qquad a_\alpha(\cdot)(\text{real}) \in C^\infty(\bar{\Omega}) \tag{4.1}$$

and is such that there exists a $\beta>0$ such that $\forall x \in \bar{\Omega}$ we have

$$\gamma \sum_{|\alpha|=2} a_\alpha(x)\xi^\alpha \geqq \beta \xi^\alpha, \quad \forall \xi \in \mathbb{R}^N \tag{4.2}$$

where γ is fixed at + or -1.

If we now define an operator A by

$$(4.3)\qquad Au=\gamma\tau u \quad \text{for } u\in\mathcal{D}(A)\triangleq H_0^1(\Omega)\cap\mathcal{D}(\tau)$$

where: $\mathcal{D}(\tau)=\{u/u,\ \tau u\in L^2(\Omega)\}$, then A is the infinitesimal generator of a strongly continuous semigroup $S(t)$, $t\geqq 0$ in $L^2(\Omega)$, [6, p. 1,767].

We further have that $S(t)$ is analytic in t. If Ω is *smooth* then Theorem 17.2 [8, p. 67] implies that if $u\in H_0^1(\Omega)$ and $\tau u\in L^2(\Omega)$ then $u\in H^2(\Omega)$. It follows that

$$\mathcal{D}(A)=H_0^1(\Omega)\cap\mathcal{D}(\tau)=H_0^1(\Omega)\cap H^2(\Omega).$$

We state the following specialization of a major theorem of Lions–Magenes.

THEOREM C. *Let $\tau=\tau(x,D)$ be given by* (4.1) *and satisfy* (4.2). *Let Ω be* smooth. *If $v\in H^{s-1/2}(\Gamma)$ for arbitrary real s then there exists a unique solution of*

$$(4.4)\qquad \tau u=0 \quad \textit{in } \Omega,$$

$$(4.5)\qquad u=v \quad \textit{on } \Gamma.$$

Further, $u\in H^s(\Omega)$ and

$$(4.6)\qquad |u|_{L^2(\Omega)}\leqq C|u|_{H^s(\Omega)}\leqq K|v|_{H^{s-1/2}(\Gamma)}.$$

The solution u is, of course, infinitely differentiable in Ω and therefore satisfies (4.4) in the classical sense. (4.5) is satisfied in the sense of traces.

Remark. This theorem is contained in the final result of Lions–Magenes, [9, p. 188] on elliptic boundary value problems. We should point out that the result in Lions–Magenes covers a much larger class of problems than we consider here and therefore this discussion could be extended. For instance, we could consider more general and higher order operators with a wide range of boundary conditions.

Proceeding, we have the following theorem:

THEOREM 5. *Let G be defined by $u=Gv$ where u is the solution to* (4.4) *and* (4.5). *Then*

$$G: L^2(\Gamma)\to[\mathcal{D}(A),L^2(\Omega)]_{1-\theta},\qquad 0<\theta<\tfrac{1}{4}.$$

Proof. To prove the theorem we will establish the following lemma which also has independent interest.

LEMMA 3. *If $0<\theta<1$ and $2\theta\neq$ integer $+\frac{1}{2}$, then*

$$\text{(i)}\qquad H_0^{2\theta}(\Omega)\subset[\mathcal{D}(A),L^2(\Omega)]_{1-\theta}\subset H^{2\theta}(\Omega).$$

If $0<2\theta<\frac{1}{2}$ $(\theta<\frac{1}{4})$, then

$$\text{(ii)}\qquad H_0^{2\theta}(\Omega)=[\mathcal{D}(A),L^2(\Omega)]_{1-\theta}=H^{2\theta}(\Omega).$$

If $2\theta=$ integer $+\frac{1}{2}$, then

$$\text{(iii)}\qquad [\mathcal{D}(A),L^2(\Omega)]_{1-\theta}\subset H^{2\theta}(\Omega).$$

Proof of lemma. We recall $\mathcal{D}(A)=H_0^1(\Omega)\cap H^2(\Omega)$ and note that

$$H_0^2\subset H_0^1\cap H^2\subset H^2$$

(we omit reference to Ω) with continuous injection. Therefore, we have

$$[H_0^2,L^2]_{1-\theta}\subset[H_0^1\cap H^2,L^2]_{1-\theta}\subset[H^2,L^2]_{1-\theta}$$

for all $0<\theta<1$. However (see [9, p. 40])

$$[H^2,L^2]_{1-\theta}=H^{2\theta},\qquad 0<\theta<1,$$

and

$$[H_0^2, L^2]_{1-\theta} = H_0^{2\theta}, \qquad 0<\theta<1, \quad 2\theta \neq \text{integer} + \tfrac{1}{2}.$$

Therefore, (i) and (iii) are proven. Further, if $s \in (0, \frac{1}{2})$ then $H_0^s = H^s$[9, p. 55]. Therefore,

$$H_0^{2\theta} = H^{2\theta}, \qquad 0<\theta<\tfrac{1}{4},$$

and (ii) is proven. Q.E.D.

Proof of theorem. From Theorem C we have

$$G: L^2(\Gamma) \to H^{1/2}(\Omega).$$

Now $H^{1/2}(\Omega) \subset H^{2\theta}(\Omega)$, $0<\theta<\frac{1}{4}$, see [9, p. 55]. Therefore,

$$H^{1/2}(\Omega) \subset H^{2\theta}(\Omega) = [\mathscr{D}(A), L^2(\Omega)]_{1-\theta}, \qquad 0<\theta<\tfrac{1}{4},$$

by Lemma 3, and the theorem is proven. Q.E.D.

This theorem shows that for A, $S(t)$ and G as described above we have

$$|AS(t)G| = O(t^{-3/4}).$$

Thus

$$f(t) = -\int_0^t AS(t-\tau)Gv(\tau)\, d\tau$$

is a generalized solution of

$$\frac{\partial f}{\partial t} = \tau f; \qquad f|_\Gamma = v; \qquad f(0) = 0$$

and, what is important for time optimal control we have $f(t) \in L^2(\Omega)$ $\forall t$ provided $v \in W^P$ $(L^2(\Gamma))$ for some $P>4$.

Remarks. A. We have already mentioned that when Ω was the square in $\mathbb{R}^2$ and $\tau = \Delta$ we have $|AS(t)G| = O(t^{-3/4})$ and that the preceding results hold for at least some regions with corners. In this regard we have the following theorem which is proven in Appendix B.

THEOREM. *If $\tau = \Delta$ and $\Omega \subset \mathbb{R}^n$ is any cylinder with a C^∞ base then* $|AS(t)G| = O(t^{-3/4})$.

Note. It is believed that such a result holds for much more general regions (cone condition) and a large class of operators.

B. The restriction of V and E to Hilbert spaces is artificial and in fact the results go through if they are merely Banach spaces; thus it should be possible to get similar results when the boundary functions are required to be spatially continuous, i.e., $v \in L^\infty(0, T; C^\infty(\Gamma))$, etc., although this has not been investigated.

C. The restriction in § 4 to 2nd order operators was artificial. The full power of the theorem of Lions–Magenes on the map G for arbitrary order operators could be used to obtain results for higher order equations.

D. For additional discussion and examples of the viewpoint presented here, see [1], [2], [3], [14].

5. A brief look at the time optimal control problem. In § 5.1 we use the map and bound developed in § 3 to prove the existence of a time optimal boundary control in a general setting. In § 5.2 we discuss a partial characterization of the control, i.e., we show that it is the weak star limit of a sequence of bang-bang controls.

Let A, $S(t)$, G, E, V, and L_T be as previously defined and assume that $|AS(t)G| = O(t^{\theta-1})$ for some $0<\theta<1$. Further denote $U = \{u/u \in W^\infty(V), |u| \leqq 1\}$. Let x be a given $W^\infty(V)$ reachable point in E, i.e., assume $\exists u \in U$ and $T>0$ such that $L_T u = x_0$. Questions of reachability/controllability for similar problems have been studied by Fattorini, Triggiani and others.

The time optimal boundary control problem then consists of finding $T_0 > 0$ and $u_0 \in U$ such that $L_{T_0} u_0 = x_0$ and T_0 is such that $T_0 \leqq T \ \forall T$ such that there exists $u \in U$ with $L_T u = x_0$. Heuristically we wish to drive the internal state of a system governed by a diffusion equation from the zero state to some predetermined final state, x_0, in minimum time by appropriately choosing the boundary control function from a predetermined admissible set in $W^\infty(V)$. Here we call T_0 the optimal time and u_0 an optimal control.

5.1. Existence of optimal control. In this part we will show that if x_0 is reachable, then the optimal control always exists.

We first recall the following standard material: If $u(\cdot)$ in $W^1(G)$ and $v(\cdot)$ in $W^\infty(G)$, where G is a Hilbert space, then

$$u(\cdot) \to \langle v, u\rangle \triangleq \int_0^T (v(\tau), u(\tau))_G \, d\tau$$

is a continuous linear functional on $W^1(G)$. In the case that G is separable all of the linear functionals on $W^1(G)$ arise in this way (see Dieudonné [5]). Therefore, since $W^1(G)$ is the predual of $W^\infty(G)$, we have

$$v(\cdot) \to \langle v, u\rangle, \qquad u(\cdot) \in W^1(G)$$

defines a continuous linear functional on $W^\infty(G)$. Not all of the continuous linear functionals on $W^\infty(G)$ arise in this way, but the ones that do serve to define the weak star topology on $W^\infty(G)$. If $v_n(\cdot) \in W^\infty(G)$ we say $v_n(\cdot)$ converges weak star to $v(\cdot)$ iff

$$\langle v_n - v, u\rangle \to 0 \quad \forall u(\cdot) \in W^1(G).$$

The well-known theorem of Alaoglu (Dunford–Schwartz [6, p. 424]) states that the closed unit sphere in a Banach space is weak star compact. For our purposes this implies that given

$$v_n(\cdot) \in U = \text{closed unit ball in } W^\infty(V)$$

there exists a subsequence $v_{n_i}(\cdot)$ which converges weak star to an element $v(\cdot) \in U$. If V is separable, which we will henceforth assume, this is equivalent to

$$\int_0^T (v_{n_i}(\tau) - v(\tau), u(\tau))_V \, d\tau \to 0 \quad \forall u(\cdot) \in W^1(V).$$

We are now ready to establish the main result of this section: the existence of an optimal control.

THEOREM 6. *If x_0 is reachable, then $\exists v_0 \in U$ and $T_0 > 0$ such that $L_{T_0} v_0 = x_0$, and $T_0 \leqq T$ for all T such that $L_T v = x_0$ for some $v \in U$.*

Proof. Let $T_0 = \inf C$ where $C = \{T/L_T v = x_0$ for some $v \in U\}$. The hypothesis shows that C is nonempty and since it is bounded below by zero, T_0 exists.

Let $T_n \in C$, $T_n \downarrow T_0 > 0$, and let $v_n \in U$ be such that $L_{T_n} v_n = x_0$.

Since U is weak star compact we may assume that v_n converge weak star to $v_0 \in U$. We may also assume $v_n(t) = 0$ for $t > T_n$, $n = 0, 1, \cdots$.

We will show $L_{T_n}v_n \to L_{T_0}v_0$ weakly in E, and therefore $L_{T_0}v_0 = x_0$.

$$\begin{aligned}
&|(L_{T_n}v_n - L_{T_0}v_0, y)| \\
&= \left|\int_0^{T_n} (AS(T_n-\tau)Gv_n(\tau), y)\,d\tau - \int_0^{T_0} (AS(T_0-\tau)Gv_0(\tau), y)\,d\tau\right| \\
&\leqq \left|\int_{T_0}^{T_n} (AS(T_n-\tau)Gv_n(\tau), y)\,d\tau\right| \\
&\qquad + \left|\int_0^{T_0} (AS(T_n-\tau)Gv_n(\tau) - AS(T_0-\tau)Gv_n(\tau), y)\,d\tau\right| \\
&\qquad + \left|\int_0^{T_0} ((AS(T_0-\tau)G)(v_n(\tau)-v_0(\tau)), y)\,d\tau\right|.
\end{aligned}$$

The first integral goes to zero as $n \to \infty$ since the integrand is in $L^1(0, T)$.

The second integral is

$$\begin{aligned}
&\leqq \int_0^{T_0} |((AS(T_n-\tau)G - AS(T_0-\tau)G)v_n(\tau), y)|\,d\tau \\
&\leqq \int_{T_n-T_0}^{T_0} |((AS(\gamma)G)^* - (AS(T_0-T_n+\gamma)G)^*)y|\,d\gamma \\
&\leqq \int_{T_n-T_0}^{T_0} \left(\frac{M}{\gamma^{\theta-1}} + \frac{M}{(T_0-T_n+\gamma)^{\theta-1}}\right) d\gamma |y| \to 0
\end{aligned}$$

as $n \to \infty$ (i.e., $T_n \to T_0$), since the integrand is in $L^1(0, T)$.

The third integral $\to 0$ as $n \to \infty$, since $v_n(\cdot) \to v_0(\cdot)$ weak star and $(AS(T-\tau)G)^*y \in W^1(V)$. We have established then that

$$L_{T_n}v_n \to L_{T_0}v \quad \text{weakly in } E.$$

Since $L_{T_n}v_n = x_0$ we have $L_{T_0}v = x_0$. Q.E.D.

5.2. Partial characterization of the optimal control. We now seek more information on the optimal control v_0 and for this we use supporting hyperplane theory. It turns out that L_TU is compact and convex and therefore has no interior points. Thus, we can only characterize a sequence of controls, $v_n(\cdot)$, which converge weak star to an optimal control $v_0(\cdot)$. In at least one case ($A \sim \Delta$) we can show that $v_n(\cdot)$ has unit norm ($|v_n(t)|_V = 1$) except on a countable set.

Throughout this section we assume that G and thus L_T is compact. We begin with:

LEMMA 4. *L_TU is compact and convex in E.*

Proof. L_TU is obviously convex since U is. It is compact because U is closed and bounded and L_T is compact. Q.E.D.

Recall that if $C \subset E$ (Hilbert space) is closed convex, we call $x \in C$ a *support point* for C if there exists a $y \in E$ such that

$$(x, y) = \sup_{\zeta \in C} (\zeta, y).$$

We have the following theorem regarding the support points of a convex set.

THEOREM (support). *In a Hilbert space the set of support points of a convex set is dense in the set of its boundary points.*

Proof. For the proof see Balakrishnan [1]. Q.E.D.

In our case it follows from Lemma 4 that every point of $L_{T_0}U$ is a boundary point of $L_{T_0}U$. In particular then, x_0 is a boundary point of $L_{T_0}U$. However, it follows from the support theorem that there exists a sequence, $x_n \in L_{T_0}U$, of support points with the property that $x_n \to x_0$ in E. Therefore, there exists $y_n \in E$, $|y_n| = 1$, such that

$$(x_n, y_n) = \sup_{\zeta \in L_{T_0}U} (\zeta, y_n).$$

Let $v_n \in U$ be such that $L_{T_0}v_n = x_n$. We assume at the outset that the v_n converge weak star to $v_1 \in W(V)$. We have

$$(L_{T_0}v_n, y_n) = \sup_{v \in U} (L_{T_0}v, y_n),$$

or

$$\int_0^{T_0} (AS(T-\tau)Gv_n(\tau), y_n)\,d\tau = \sup_{v \in U} \int_0^{T_0} (AS(T-\tau)Gv(\tau), y_n)\,d\tau$$

and via the standard argument we have

$$(AS(T-t)Gv_n(t), y_n) = \sup_{v \in U} (AS(T-t)Gv(t), y_n)$$

for a.e. $t \in (0, T)$. We therefore have

$$(v_n(t), (L'_{T_0}y_n)(t)) = \sup_{v \in U} (v(t), (L'_{T_0}y_n)(t))$$

a.e. $t \in (0, T)$. It follows that

$$v_n(t) = \frac{(L'_{T_0}y_n)(t)}{|(L'_{T_0}y_n)(t)|_V}$$

for those t such that $|(L'_{T_0}y_n)(t)|_V \neq 0$. If we assume $S(t)$ is analytic, then $(L'_{T_0}y_n)(t)$ is analytic (Lemma 2). In this case $|(L_{T_0}y_n)(t)|_V$ vanishes at, at most, a finite number of points in $[0, T-\varepsilon]$ or else is identically zero. We can now state the following theorem on characterization of the optimal control.

THEOREM 7. *If T_0 is the optimal time for x_0 then there exists an optimal control $v_0 \in U$ and a subsequence $v_n \in U$ converging weak star to v_0 such that*

$$x_n = L_{T_0}v_n \to x_0 = L_{T_0}v_0 \quad \text{in } E.$$

There exists $y_n \in E$ (we take $|y_n| = 1$) such that

$$(v_n(t), (L'_{T_0}y_n)(t)) = \sup_{v \in U} (v(t), (L'_{T_0}y_n)(t)) \quad \text{a.e. } t.$$

For those t such that $|(L'_{T_0}y_n)(t)|_V \neq 0$, $v_n(t)$ has the form

$$v_n(t) = \frac{(L'_{T_0}y_n)(t)}{|(L'_{T_0}y_n)(t)|_V}.$$

If $S(t)$ is analytic, then either $(L'_{T_0}y_n)(t)$ is identically zero or vanishes at, at most, a finite number of points in any compact of $[0, T)$.

Proof. It only remains to establish that v_n actually converge weak star to an optimal control. Recall that the v_n were assumed to converge weak star to $v_1 \in U$. Therefore

$$(L_{T_0}v_n, y) = \langle v_n, L'_{T_0}y\rangle \to \langle v_1, L'_{T_0}y\rangle = (L_{T_0}v_1, y)$$

for all $y \in H$. But $L_{T_0}v_n \to x_0$ in H. Therefore $L_{T_0}v_1 = x_0$ by the uniqueness of weak limits and v_1 is an optimal control, which completes the proof. Q.E.D.

The following theorem gives an example where $L'_{T_0}y_n$ of Theorem 7 cannot be identically zero and therefore $L'_{T_0}y_n$ vanishes at, at most, a finite number of points in any compact of $[0, T)$. The v_n of Theorem 7 are then such that $|v_n(t)| = 1$ except on a finite number of points in compacts of $[0, T)$.

THEOREM 8. *If $A \sim \Delta$ (the Laplacian) with $\mathscr{D}(A) = H_0^1(\Omega) \cap \mathscr{D}(\Delta)$, then for $y \in E$, $y \neq 0$, $(L'_T y)(t)$ does not vanish on any t interval.*

Proof. Green's theorem for the Laplacian implies that

$$G^*v = \frac{\partial}{\partial\eta} R(0; A)v|_\Gamma \quad (\text{here } R(0, A) = (-A)^{-1}).$$

Since A and $S(t)$ are self-adjoint, we have

$$(L'_T y_n)(t) = (AS(T-t)G)^*y = G^*AS(T-t)y = \frac{\partial}{\partial\eta} S(t)y|_\Gamma.$$

Therefore $(L'_T y_n)(t)$ is zero on a t interval iff

$$\frac{\partial}{\partial\eta} S(t)y|_\Gamma = 0 \tag{5.1}$$

on a t interval. But $S(t)y$ is just the solution of the heat equation for initial data y. That is, $u(t) = S(t)y$ solves:

$$\frac{\partial u}{\partial t} = \Delta u \in \Omega \times (0, T),$$

$$u(t)|_\Gamma = 0; \qquad u(0) = y.$$

If (5.1) holds for $t \in (t_1, t_2)$, $t_2 > t_1 > 0$, then u satisfies

$$\frac{\partial u}{\partial t} = \Delta u \qquad \text{in } \Omega \times (t_1, t_2),$$

$$u(t)|_\Gamma = 0 \qquad \text{in } \Gamma \times (t_1, t_2),$$

$$\left.\frac{\partial u(t)}{\partial \eta}\right|_\Gamma = 0 \qquad \text{in } \Gamma \times (t_1, t_2).$$

This implies that (see [12, p. 340]).

$$u(t) \equiv 0 \quad \text{in } \Omega \times (t_1, t_2),$$

which of course implies that

$$u(t) \equiv 0 \quad \text{in } \Omega \times (0, T).$$

Therefore, $y = 0$, which is a contradiction. Q.E.D.

Appendix A. Theorem 2(a) is best in a sense. In this appendix we show that the result of Theorem 2(a) cannot be improved by exhibiting an example where $|AS(t)G| \leqq Mt^{\theta-1}$ but G does not map V into $\mathscr{D}[(A), E]_{1-\theta}$ for $\theta = \frac{1}{4}$ but only for $\theta < \frac{1}{4}$. The example we consider is the 1-dimensional problem on a bounded interval of the real line.

Consider

$$\frac{d^2u}{dt^2}=\Delta u \quad \text{for } x\in(0,\pi) \text{ where } u\sim u(t)\sim u(t,x)$$

$$u(t)|_\Gamma=\begin{bmatrix}u(t,0)\\u(t,\pi)\end{bmatrix}=\begin{bmatrix}v_1(t)\\v_2(t)\end{bmatrix} \quad \text{given in } L^2(0,T)\times L^2(0,T).$$

Let

$$A\simeq\Delta \qquad \left(\Delta=\frac{d^2}{dx^2}\right)$$

on

$$\mathscr{D}(A)=\{f\in L^2(0,\pi)|f,f' \text{ are A.C.}; f''\in L^2(0,\pi); f(0)=f(\pi)=0\}$$

The eigenset of A is then given by

$$\theta_n(x)=\sqrt{\frac{2}{\pi}}\sin(nx); \qquad \lambda_n=-n^2; \quad n=1,2,\cdots.$$

which is a complete orthonormal set in $L^2(0,\pi)$. The semigroup generated by A has the simple representation

$$S(t)\theta_n=e^{\lambda_n t}\theta_n$$

or

$$S(t)\left(\sqrt{\frac{2}{\pi}}\sin nx\right)=\sqrt{\frac{2}{\pi}}e^{-n^2t}\sin(nx).$$

The map $G\colon \mathbb{R}^2\to L^2(0,\pi)$ is defined by

$$\frac{d^2}{dx^2}\left(G\begin{bmatrix}a\\b\end{bmatrix}\right)=0 \quad \text{in } (0,\pi)$$

$$G\begin{bmatrix}a\\b\end{bmatrix}\bigg|_{x=0}=a; \qquad G\begin{bmatrix}a\\b\end{bmatrix}\bigg|_{x=\pi}=b.$$

Therefore,

$$G\begin{bmatrix}a\\b\end{bmatrix}=\frac{b-a}{\pi}x+a.$$

We therefore have the following representation of $AS(t)G$ in terms of the eigenfunctions of A.

$$\begin{aligned}AS(t)G\begin{bmatrix}a\\b\end{bmatrix}&=\sum\left(AS(t)G\begin{bmatrix}a\\b\end{bmatrix},\theta_n\right)\theta_n(x)\\&=\sum\left(G\begin{bmatrix}a\\b\end{bmatrix},S(t)A\theta_n\right)\theta_n(x)\\&=-\sum n^2e^{-n^2t}\left(G\begin{bmatrix}a\\b\end{bmatrix},\theta_n\right)\theta_n(x)\\&=-\sum n^2e^{-n^2t}\int_0^\pi\left(\frac{b-a}{\pi}s+a\right)\sin(ns)\,ds\theta_n(x)\sqrt{\frac{2}{\pi}}\\&=-\sum n\,e^{-n^2t}(a-b\cos(n\pi))\theta_n(x)\sqrt{\frac{2}{\pi}}.\end{aligned}$$

The bound on $AS(t)G$ is calculated as follows:

$$\left|AS(t)G\begin{bmatrix}a\\b\end{bmatrix}\right|^2 = \sum n^2 e^{-2n^2t}(a+b\cos(n\pi))^2\frac{2}{\pi}$$

$$\leqq\left(2\sum n^2 e^{-2n^2t}\right)(a^2+b^2)\leqq\frac{M^2}{t^{3/2}}\left\|\begin{bmatrix}a\\b\end{bmatrix}\right\|^2_{\mathbb{R}^2}$$

where we have used[2]

$$\sum n^2 e^{-2n^2t}\sim\int_0^\infty n^2 e^{-2n^2t}\,dn\sim\frac{1}{t^{3/2}}\int_0^\infty x^{1/2}e^{-x}\,dx\sim\frac{1}{t^{3/2}}.$$

Therefore,

$$\left|AS(t)G\begin{bmatrix}a\\b\end{bmatrix}\right|\leqq\frac{M}{t^{3/4}}\left\|\begin{bmatrix}a\\b\end{bmatrix}\right\|_{\mathbb{R}^2}.$$

Thus,

$$|AS(t)G|\leqq\frac{M}{t^{3/4}}.$$

We now show that G does not map $\mathbb{R}^2$ into $[\mathscr{D}(A), L^2(0,\pi)]_{3/4}$. Since $-A$ is positive self-adjoint with $\mathscr{D}(-A)=\mathscr{D}(A)$ and $\mathbb{R}(-A)$ (range of $-A$) is $L^2(0,\pi)$, we have

$$[\mathscr{D}(A), L^2(0,\pi)]_{1-\theta}=\mathscr{D}((-A)^\theta).$$

Further, we have the simple representation of $-A$.

$$(-Af)(x)=\sqrt{\frac{2}{\pi}}\sum_n n^2(f,\sin(nx))\sin(nx).$$

Therefore, $(-A)^\theta$ has the representation

$$((-A)^\theta f)(x)=\frac{2}{\pi}\sum(n^2)^\theta(f,\sin(nx))\sin(nx)$$

and

$$f\in\mathscr{D}((-A)^\theta)\quad\text{iff}\quad\sum(n^2)^{2\theta}(f,\sin(nx))^2<\infty.$$

Thus,

$$f\in[\mathscr{D}(A), L^2(0,\pi)]_{1-\theta}\quad\text{iff}\quad\sum(n^2)^{2\theta}(f,\sin(nx))^2<\infty.$$

We have already seen that $|AS(t)G|=M/t^{1-1/4}$ and we now show that G: $V=R^2\to[\mathscr{D}(A), L^2(0,\pi)]_{1-\theta}\ \forall\theta<\frac{1}{4}$, but not for $\theta\geqq\frac{1}{4}$. This is equivalent to showing

$$\sum_n(n^2)^{2\theta}\left(G\begin{bmatrix}a\\b\end{bmatrix},\sin(nx)\right)^2<\infty\quad\forall\begin{bmatrix}a\\b\end{bmatrix}\in\mathbb{R}^2$$

and all $\theta<\frac{1}{4}$, but not if $\theta\geqq\frac{1}{4}$. Now

$$\sum_n(n^2)^{2\theta}\left(G\begin{bmatrix}a\\b\end{bmatrix},\sin(nx)\right)^2=\sum n^{4\theta-2}(a-b\cos(n\pi))^2$$

[2] The symbol $\sim$ is used throughout to mean "is of the same order of magnitude as."

which is finite for all $\begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^2$ if $4\theta - 2 < -1$, i.e. $\theta < \frac{1}{4}$. However, for $\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ the r.h.s. is infinite when $4\theta - 2 \geqq -1$, i.e. $\theta \geqq \frac{1}{4}$. Therefore,

$$G\mathbb{R}^2 \subset \mathscr{D}((-A)^{\theta_0}) = [\mathscr{D}(A), L^2(0, \pi)]_{\theta_0}$$

for $\theta_0 < \theta = \frac{1}{4}$ but not when $\theta_0 \geqq \theta = \frac{1}{4}$.

Appendix B. Extension to cylinders with C^∞ bases. Here we let Ω be a cylinder with a C^∞ base and show that the map G, defined as in Theorem 5, with $\tau = \Delta$, satisfies

$$Gv \in [\mathscr{D}(A), L^2(\Omega)]_{1-\alpha} \quad \text{for } 0 < \alpha < \tfrac{1}{4}$$

for all $v \in L^2(\Gamma)$.

We require of the following cast of characters. Let:

$\Omega' \subset R^{N-1}$ be smooth with boundary Γ';

$\Omega = \Omega' \times (0, \pi)$ with boundary Γ;

$\Gamma_L, \Gamma_T, \Gamma_B$ be respectively the lateral, top, and bottom boundary surfaces of Ω, i.e.,

$$\Gamma_L = \Gamma' \times (0, \pi); \qquad \Gamma_{B(T)} = \{x = (x_1, \cdots, x_N) \in \Omega | x_N = O(\pi)\};$$

Δ be the Laplacian in R^N, i.e., $\Delta = \sum_{i=1}^N \partial^2/\partial x_i^2$;

Δ' be the Laplacian in R^{N-1}, i.e., $\Delta' = \sum_{i=1}^{N-1} \partial^2/\partial x_i^2$;

A be defined by

$$Au = \Delta u \quad \text{for } u \in H_0^1(\Omega) \cap \mathscr{D}(\Delta)$$

$$\mathscr{D}(\Delta) = \{u/u, \Delta u \in L^2(\Omega)\};$$

A' be defined similarly with Δ', Ω';

G be defined by $u = Gv$ where

$$\Delta u = 0 \text{ in } \Omega, \qquad u|_\Gamma = v \in L^2(\Gamma);$$

G' be defined by $u' = G'v'$ where:

$$\Delta u' = 0 \text{ in } \Omega', \qquad u'|_{\Gamma'} = v' \in L^2(\Gamma');$$

$-\lambda_{mn}$, θ_{mn} denote the eigenset of A, i.e.,

$$\Delta\theta_{mn} = -\lambda_{mn}\theta_{mn}, \qquad \theta_{mn}|_\Gamma = 0$$

and we know they form a complete orthonormal set in $L^2(\Omega)$;

$-\mu_n$, ψ_n denote the eigenset of A', i.e.,

$$\Delta'\psi_n = -\mu_n\psi_n, \qquad \psi_n|_{\Gamma'} = 0$$

and such functions form a complete orthonormal set in $L^2(\Omega')$.

From Theorem 5 we have

$$G'L^2(\Gamma') \subset [\mathscr{D}(A'), L^2(\Omega')]_{1-\alpha} = \mathscr{D}((-A')^\alpha)$$

for $0 < \alpha < \frac{1}{4}$. This is equivalent to

(B.1) $$\sum_n \mu_n^{2\alpha} (G'v', \psi_n)^2_{L^2(\Omega')} < \infty.$$

We wish to show that

(B.2) $$GL^2(\Gamma) \subset [\mathscr{D}(A), L^2(\Omega)]_{1-\beta} = \mathscr{D}((-A)^\beta)$$

for $0<\beta<\frac{1}{4}$ which is equivalent to

$$\text{(B.3)} \qquad \sum_{m,n} \lambda_{mn}^{2\beta}(Gv,\, \theta_{mn})^2_{L^2(\Omega)}<\infty \quad \forall v \in L^2(\Gamma).$$

We now seek a representation of λ_{mn}, $\theta_{mn}(x)$ by separating variables. We will let

$$x=(x_1, \cdots, x_N)=(x', x_N).$$

We seek eigenvectors of Δ of the form

$$\theta(x)=\psi(x')\Phi(x_N);$$

thus

$$\Delta\psi\Phi=\Phi\Delta'\psi+\psi\frac{d^2}{dx_N^2}\Phi=-\lambda\psi\Phi, \qquad \psi\phi|_\Gamma=0$$

from which we get

$$\psi^{-1}\Delta'\psi=-\lambda-\Phi^{-1}\Phi''=-\mu$$

which gives the separated equations

$$\text{(i)} \quad \Delta'\psi(x')=-\mu\psi(x'); \qquad \psi(x')|_{\Gamma'}=0;$$

$$\text{(ii)} \quad \frac{d^2}{dx_N^2}\Phi=(\mu-\lambda)\Phi; \qquad \Phi(x_N)|_{x_N=0}^{x_N=\pi}=0.$$

The solution of (i) is the eigenset of A', i.e., μ_n, $\psi_n(x')$. The solutions of (ii) are given by

$$\Phi_m(x_N)=\sin\,(mx_N)\sqrt{\frac{2}{\pi}}; \qquad \mu-\lambda=-m^2, \qquad m=1, 2, \cdots.$$

Thus the eigenset of A is given by

$$\Phi_{mn}(x', x_N)=\psi_n(x')\Phi_m(x_N); \qquad \text{where: } \Phi_m(x_N)=\sin\,(mx)\sqrt{\frac{2}{\pi}};$$

$$\lambda_{mn}=m^2+\mu_n.$$

Recall that $\mu_n \sim n^C$ [10, p. 324] where $C=2/(N-1)$ ($N-1$ is dimension of space here).

We seek to establish (B.2). We first assume $v=0$ on the top and bottom faces of Γ. (Recall G^* from the Theorem 8 proof.) Then

$$\begin{aligned}(Gv,\, \theta_{mn})&=\frac{1}{\lambda_{mn}}\left(v, \frac{\partial}{\partial\eta}\theta_{mn}|_\Gamma\right)_{L^2(\Gamma)}\\ &=\frac{1}{\lambda_{mn}}\left(v, \Phi_m\frac{\partial}{\partial\eta}\psi_n|_{\Gamma_L}\right)_{L^2(\Gamma_L)}\\ &=\frac{1}{\lambda_{mn}}\int_0^\pi \Phi_m(x_N)\left(v(\cdot, x_N), \frac{\partial}{\partial\eta}\psi_n|_{\Gamma'}\right)_{L^2(\Gamma')}dx_N\\ &=\frac{\mu_n}{\lambda_{mn}}\int_0^\pi \Phi_m(x_N)(G'v(\cdot, x_N), \psi_n)_{L^2(\Omega')}\,dx_N\\ &=\frac{\mu_n}{\lambda_{mn}}I_{mn}\end{aligned}$$

where

$$I_{mn}=\int_0^\pi \Phi_m(x_N)(G'v(\cdot, x_N), \psi_n)_{L^2(\Omega')}\,dx_N.$$

Now

$$\begin{aligned}\sum_{mn} \lambda_{mn}^{2\beta}(Gv, \theta_{mn})^2 &= \sum_{mn} \frac{\lambda_{mn}^{2\beta}\mu_n^2}{\lambda_{mn}^2} I_{mn}^2 \\ &= \sum_{mn} \frac{\mu_n^{2-2\alpha}}{\lambda_{mn}^{2-2\beta}} \mu_n^{2\alpha} I_{mn}^2 \\ &\sim \sum_{mn} \frac{n^{c(2-2\alpha)}}{(m^2+n^c)^{2-2\beta}} \mu_n^{2\alpha} I_{mn}^2.\end{aligned}$$

We calculate $\max_n n^{c(2-2\alpha)}/(m^2+n^c)^{2-2\beta}$ by setting $(d/dn)n^{c(2-2\alpha)}/(m^2+n^c)^{2-2\beta}=0$ which yields $n^c = km^2$, $(k = \text{const.})$.

Therefore

$$\max_n \frac{n^{c(2-2\alpha)}}{(m^2+n^c)^{2-2\beta}} = k' m^{2(\beta-\alpha)} \leqq k'$$

if $\beta < \alpha$. We have then, for $\beta < \alpha$, that

$$\text{(B.4)} \qquad \sum_{mn} \lambda_{mn}^{2\beta}(Gv, \theta_{mn})^2 \leqq K'' \sum_{mn} \mu_n^{2\alpha} I_{mn}^2.$$

Now

$$\begin{aligned}\sum_{mn} \mu_n^{2\alpha} I_{mn}^2 &= \sum_{mn} \mu_n^{2\alpha}\left(\int_0^\pi \Phi_m(x_N)(G'v(\cdot, x_N), \psi_n)_{L^2(\Omega')}\, dx_N\right)^2 \\ &= \sum_{mn}\left(\int_0^\pi \Phi_m(x_N)\mu_n^\alpha(G'v(\cdot, x_N), \psi_n)_{L^2(\Omega')}\, dx_N\right)^2 \\ &= \sum_{mn}\left(\int_0^\pi \Phi_m(x_N)((-A')^\alpha G'v(\cdot, x_N), \psi_n)_{L^2(\Omega')}\, dx_N\right)^2 \\ &= \sum_{mn}\left(\int_0^\pi \int_{\Omega'} (-A')^\alpha G'v(x', x_N)\theta_{mn}(x', x_N)\, dx'\, dx_N\right)^2 \\ &= \sum_{mn} ((-A')^\alpha G'v, \theta_{mn})^2.\end{aligned}$$

Now $(-A')^\alpha G'\colon L^2(\Gamma') \to L^2(\Omega')$ continuously, i.e.,

$$\int_{\Omega'} |(-A')^\alpha G'v'(x', x_N)|^2\, dx' \leqq C \int_{\Gamma'} |v'(x; x_N)|^2\, d\sigma_{x'}$$

which follows from (B.1) and the closed graph theorem. Integrating both sides with respect to x_N then shows that

$$(-A')^\alpha G'\colon L^2(\Gamma) \to L^2(\Omega) \text{ continuously.}$$

Therefore

$$\sum_{mn} \mu_n^{2\alpha} I_{mn}^2 \leqq C'|v|_{L^2(\Gamma)}^2 < \infty,$$

and it follows from (B.4) that

$$\sum_{mn} \lambda_{mn}^{2\beta}(Gv, \theta_{mn})^2 \leqq C''|v|_{L^2(\Gamma)}^2 < \infty.$$

Then (B.2), (B.3) imply, for $v = 0$ on Γ_T, Γ_B, that

$$Gv \in [\mathscr{D}(A), L^2(\Omega)]_{1-\beta}$$

for all $\beta < \alpha < \frac{1}{4}$. However α is arbitrary, $\alpha < \frac{1}{4}$; therefore the result holds for $\beta < \frac{1}{4}$.

We now suppose that $v = 0$ on Γ_L, Γ_T, i.e., $v \neq 0$ only on Γ_B, the bottom face of the cylinder. We note that ψ_n form a complete orthonormal set on this surface and set

$$v_n = (v, \psi_n)_{L^2(\Gamma_B)}.$$

We have

$$(Gv, \theta_{mn}) = \frac{1}{\lambda_{mn}}\left(v, \frac{\partial \theta_{mn}}{\partial \eta}\bigg|_{\Gamma}\right)_{L^2(\Gamma)} = \frac{1}{\lambda_{mn}}\left(v, \psi_n \frac{\partial \Phi_m}{\partial \eta}\bigg|_{\Gamma_B}\right)_{L^2(\Gamma_B)}$$

$$= \frac{m}{\lambda_{mn}}(v, \psi_n)_{L^2(\Gamma_B)} = \frac{m}{\lambda_{mn}} v_n.$$

We again wish to establish (B.3).

$$\sum_{mn} \lambda_{mn}^{2\beta}(Gv, \theta_{mn})^2 = \sum_{mn} \frac{m^2 v_n^2}{\lambda_{mn}^{2-2\beta}} = \sum_{mn} \frac{m^2 v_n^2}{(m^2 + \mu_n)^{2-2\beta}}$$

$$\sim \sum_{mn} \frac{m^2 v_n^2}{(m^2 + n^c)^{2-2\beta}}$$

$$\sim \sum_{n} v_n^2 \int_0^\infty \frac{m^2}{(m^2 + n^c)^{2-2\beta}}\, dm$$

$$\sim \sum_{n} \frac{v_n^2}{n^{(1/2-2\beta)c}} \int_0^\infty \frac{x^2\, dx}{(x^2 + 1)^{2-2\beta}}.$$

The integral is finite iff $\beta < \frac{1}{4}$. In this case we also have $1/n^{(1/2-2\beta)c} < 1$ and therefore

$$\sum_{mn} \lambda_{mn}^{2\beta}(Gv, \theta_{mn})^2 \leqq C \sum_{n} v_n^2 = C|v|_{L^2(\Gamma)}.$$

Equation (B.3) is established for $v \in L^2(\Gamma_B)$. The demonstration is the same for $v \in L^2(\Gamma_T)$ and we are finished.

Apparently, the method we have used here can be extended to self-adjoint operators other than Δ; however, it should be apparent that the technical difficulties would be considerable. It is undoubtedly true that the result we have established holds for a wide range of operators (nonself-adjoint) on a much larger class of regions; however this fact still awaits proof.

Acknowledgment. The author wishes to thank his thesis advisor Professor A. V. Balakrishnan for his guidance and Professor Roberto Triggiani for his constructive comments including a simplification of the proof of Theorem 4. This paper is based on part of a Ph.D. thesis at UCLA.

REFERENCES

[1] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer-Verlag, New York, 1976.

[2] ———, *Filtering and control problems for partial differential equations*, Proceedings of the 2nd Kingston Conference on Differential Games and Control Theory (University of Rhode Island), Marcel Dekker Inc., New York, 1976.

[3] ———, *Boundary Control of Parabolic Equations: L-Q-R Theory*, Proceedings of the Conference on Theory of Non Linear Operators (September 1977, Berlin), to be published by Academie-Verlag, Berlin, 1978.

[4] P. L. BUTZER AND H. BERENS, *Semigroups of Operators and Approximation*, Springer-Verlag, New York, 1967.

[5] J. DIEUDONNÉ, *Sur le Théorème de Lebesgue–Nikodym, Part V*, Canad. J. Math., 3 (1951), pp. 129–139.

[6] N. DUNFORD AND J. SCHWARTZ, *Linear Operators*, Interscience, New York, 1963.

[7] H. O. FATTORINI, *Boundary control systems*, this Journal, 6 (1968), pp. 349–385.

[8] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.

[9] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. 1. Springer-Verlag, New York, 1967.

[10] S. G. MIKHLIN, *Mathematical Physics, an Advanced Course*, North-Holland, Amsterdam, 1970.

[11] J. NECAS, *Les Methodes Directes en Théorie des Equations Elliptiques*, Ed. Acad. Tchecoslavaque des Sciences, Prague, 1967.

[12] M. H. PROTTER, *Properties of solutions of parabolic equations and inequalities*, Canad. J. Math., 13 (1961), pp. 331–345.

[13] R. TRIGGIANI, *A cosine operator approach to modelling boundary input hyperbolic systems*, Proceedings 8th IFIP Conference on Optimization (University of Würzburg, West Germany, September 1977).

[14] D. WASHBURN, *A semi-group approach to time optimal boundary control of diffusion processes*, Ph.D. dissertation, 1974, UCLA.

[15] ———, *A Semi-Group Theoretic Approach to Modeling of Boundary Input Problems*, Proceedings of IFIP Working Conference (University of Rome), Lecture Notes in Control and Information Science, Springer-Verlag, 1977.

[16] K. YOSIDA, *Functional Analysis*, 3rd ed., Springer-Verlag, New York, 1971.

[17] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 6, November 1979

0363-0129/79/1706-0001$01.00/0

MODIFICATIONS IN OPTIMIZATION AND SADDLE POINT PROBLEMS*

ISAK BEHAR†

Abstract. A general optimization problem of finding Inf $\{f(x) \mid x \in A\}$ is modified with a sequence of functions (H_n), $n \in \mathbb{N}$ to yield the problems Inf $\{f(x)+H_n(x) \mid x \in A\}$. The properties of convergence are studied in two general classes of modifications, These results are applied to modify saddle point problems to obtain dual modification methods for optimization problems.

1. Introduction. Let E be a locally convex topological vector space. We denote by $CL(E)$ the space of lower semicontinuous functions defined on E with values on $\bar{\mathbb{R}}$ (the extended real line) and by $\Gamma_0(E)$ the space of functions in $CL(E)$ which are convex, not identically equal to $+\infty$, and never taking on the value $-\infty$. The topological dual of E will be denoted by E^*. The value of a linear functional $y \in E^*$ for $x \in E$ will be denoted by $\langle y, x\rangle$. A point $y_1 \in E^*$ will be called a subgradient of a function $f \in \Gamma_0(E)$ at a point $x_1 \in E$ if

$$f(x) \geqq f(x_1)+\langle y, x-x_1\rangle \quad \text{for all } x \in E.$$

The set of all subgradients of f at x_1 will be called the subdifferential of f at x_1 and will be denoted by $\partial f(x_1)$. (See [7].) A real valued function f defined on a closed set $A \subset E$ will be called lower (resp. upper) compact on A if for all $\lambda \in R$, the sets $S_\lambda = \{x \in A \mid f(x) \leqq \lambda\}$ (resp. $S_\lambda = \{x \in A \mid f(x) \geqq \lambda\}$) are compact.

For any function $f: E \to R$ we note by f^* and f^{**}, the conjugate and the second conjugate functions of f defined by

$$f^*(y) = \operatorname{Sup}\{\langle x, y\rangle - f(x) \mid x \in E\}$$

$$f^{**}(x) = \operatorname{Sup}\{\langle x, y\rangle - f^*(y) \mid y \in E^*\}$$

where E^* denotes the topological dual of E.

We shall study the general problem

$$(1.1) \qquad (\mathrm{P}) \qquad \alpha = \operatorname{Inf}\{f(x) \mid x \in A\}, \qquad \Omega = \{x \in A \mid f(x) = \alpha\}$$

where $f \in CL(E)$ and A is a nonempty closed subset of E. Consider a sequence of function (f_n) and the problems

$$(1.2) \qquad (\mathrm{P}_n) \qquad \alpha_n = \operatorname{Inf}\{f_n(x) \mid x \in A\}, \qquad \Omega_n = \{x \in A \mid f_n(x) = \alpha_n\}.$$

DEFINITION 1.1. Any sequence (x_n) such that for each $n \in \mathbb{N}$ $x_n \in \Omega_n$ will be called a sequence of solutions of the family $\{\mathrm{P}_n\}$.

DEFINITION 1.2. The sequence of problems (P_n) is called convergent to (P) if $\Omega_n \neq \phi$ for each $n \in \mathbb{N}$, $\lim_{n\to+\infty} \alpha_n = \alpha$, and every sequence of solutions of the family $\{\mathrm{P}_n\}$ has at least one cluster point and all the cluster points are solutions of (P).

DEFINITION 1.3. When the function f_n has the form $f_n = f + H_n$, where $H_n \in CL(E)$, the problem (P_n) will be called the H_n-modification of (P).

The nature of the function H_n will determine the type of modification. Two types of modifications will be studied:

Zero limit modifications
Modification by translations (or simply, translations).

* Received by the editors March 7, 1978, and in revised form December 29, 1978.

† Department of Mathematics, Bogazici University, Istanbul, Turkey. Visiting Northeastern University, Boston, Massachusetts 02115.

2. Zero limit modifications. Consider a sequence of functions $H_n \in CL(E)$ satisfying the following assumptions:

1. For every $n \in \mathbb{N}$ there exists $M_n \in \mathbb{R}$ such that
$$H_n(x) \geqq M_n \quad \text{for all } x \in A,$$
$$\lim_{n\to+\infty} M_n = 0.$$

2.$\mathscr{A}$. 2. For every $n \in \mathbb{N}$ there exists $r_n > 0$ such that $\lim_{n\to+\infty} r_n = +\infty$.
The set of functions $\{h_n \in CL(E) \mid h_n = r_n H_n\}$ is equi-continuous with respect to A.
The function h, defined by $h(x) = \lim_{n\to+\infty} h_n(x)$, is finite for all $x \in A$.

3. The family of functions $\{h_n \mid n \in N\}$ is uniformly lower compact on A (i.e., for all $\lambda \in \mathbb{R}$ there exists a compact $K_\lambda \subset A$ such that the sets $\{x \in A \mid h_n(x) \leqq \lambda\}$ are included in K_λ for all $n \in \mathbb{N}$).

DEFINITION 2.1. If a sequence (H_n) satisfies the assumptions 2.$\mathscr{A}$, then the sequence of problems (P_n) defined by (1.2) is called a zero *limit modification* of the problem (P).

THEOREM 2.1. *If α is finite and* (P_n) *is a zero limit modification of* (P) *then*

1. $\Omega_n \neq \phi$ *for all* $n \in \mathbb{N}$.
2. $\lim_{n\to\infty} \alpha_n = \alpha$.
3. *If (x_n) is a sequence of solutions then* $\lim_{n\to\infty} f(x_n) = \alpha$ *and* (x_n) *has cluster points if and only if* $\Omega \neq \phi$. *In which case* (P_n) *converges to* (P) (*Definition* 1.2).
4. *If $\bar{x}$ is a cluster point of a sequence of solutions then*
$$h(\bar{x}) = \mathrm{Min}\,\{h(z) \mid z \in \Omega\}$$
(*where h is the function defined in* 2.$\mathscr{A}$.2).

Proof. 1. Because of our assumptions, the function $f + H_n$ is lower compact and bounded below on A. Therefore $\Omega_n \neq \phi$.

2. By definition of M_n
$$f(x) + M_n \leqq f(x) + H_n(x) \quad \text{for all } x \in A, \qquad n \in \mathbb{N};$$
hence
$$\alpha + M_n \leqq f(x) + H_n(x) \quad \text{for all } x \in A, \qquad n \in \mathbb{N};$$
then
$$\alpha \leqq \liminf \alpha_n \leqq \limsup \alpha_n \leqq f(x) \quad \text{for all } x \in A;$$
therefore $\lim_{n\to+\infty} \alpha_n = \alpha$.

3. We have
$$\alpha + M_n \leqq f(x_n) + M_n \leqq f(x_n) + H_n(x_n) = \alpha_n.$$
Since $\lim_{n\to+\infty} \alpha_n = \alpha$ and $\lim_{n\to+\infty} M_n = 0$, we conclude that $\lim_{n\to+\infty} f(x_n) = \alpha$. If $\bar{x}$ is a cluster point of (x_n), the lower semicontinuity of f implies $f(\bar{x}) \leqq \lim_{n\to+\infty} f(x_n) = \alpha$, which is possible only if $\bar{x} \in \Omega$. Conversely if $\Omega \neq \phi$ let $u \in \Omega_n$ and $z \in \Omega$. We have
$$\alpha_n = f(u) + \frac{1}{r_n}(u) \leqq f(x) + \frac{1}{r_n} h_n(z).$$
Since $f(z) \leqq f(u)$ and $r_n > 0$, we get

(2.1) $$h_n(u) \leqq h_n(z) \quad \text{for all } u \in \Omega_n, \qquad z \in \Omega_n, \quad n \in \mathbb{N}.$$

Since the sequence $h_n(z)$ is convergent (to $h(z)$) there exists $\lambda \in \mathbb{N}$ such that

$$h_n(u) \leqq \lambda \quad \text{for all } u \in \Omega_n, \quad n \in \mathbb{N}.$$

The uniform lower compactness of the functions h_n implies that there exists a compact $K \subset A$ such that $\Omega_n \subset K$ for all $n \in \mathbb{N}$, which implies in turn, that every sequence of solutions has at least one cluster point. Since $\lim_{n \to +\infty} f(x_n) = \alpha$, such a cluster point belongs to Ω.

4. By (2.1) we have

$$(2.2) \qquad h_n(x_n) \leqq h_n(z) \quad \text{for all } z \in \Omega, \quad n \in \mathbb{N}.$$

Thus if $\bar{x}$ is a cluster point of (x_n), the equicontinuity of the family $\{h_n\}$ implies

$$(2.3) \qquad h(\bar{x}) \leqq h(z) \quad \text{for all } z \in \Omega.$$

Remark 2.1. The equicontinuity of the family $\{h_n\}$ is used only to pass from (2.2) to (2.3). In fact, the assumptions 2.$\mathscr{A}$ may be simplified by considering a sequence $H_n = (1/r_n)h$ where $h \in CL(E)$ is lower compact on A. In this case the lower semicontinuity of h is enough to pass from (2.2) to (2.3).

Remark 2.2 (rate of convergence). If $M_n \geqq 0$ and $z \in \Omega$ then

$$\alpha + M_n \leqq \alpha_n \leqq f(z) + \frac{1}{r_n} h_n(z);$$

hence

$$0 \leqq M_n \leqq \alpha_n - \alpha \leqq \frac{1}{r_n} h_n(z).$$

Therefore α_n converges to α at least as fast as $H_n(z) = (1/r_n)h_n(z)$ converges to $0 \cdot h(z) = 0$.

Remark 2.3 (convex case). Suppose that the problem (P) is convex (i.e., $f \in \Gamma_0(E)$ and A is a nonempty closed convex subset of E). In this case Ω is convex. We can choose the sequence (H_n) such that the functions h_n and h are strictly convex. (This can be done, for example, by choosing H_n as in Remark 2.1 with h strictly convex.) In this case every problem (P_n) has a unique solution, so there is a unique sequence of solutions (x_n). On the other hand, any cluster point of (x_n) is also the unique solution of the problem

$$\operatorname{Inf} \{h(z) \mid z \in \Omega\}.$$

Therefore the sequence (x_n) is convergent.

3. Convex-concave saddle point problems. Let E and F be locally convex topological vector spaces and A and B nonempty closed convex subsets of E and F respectively. We consider a function $L: E \times F \to \bar{R}$ satisfying:

1. $L(\cdot, v) \in \Gamma_0(E)$ for all $v \in B$,
   $-L(x, \cdot) \in \Gamma_0(F)$ for all $x \in A$.
2. There exists an $a_0 \in A$ and $b_0 \in B$ such that the
3.$\mathscr{L}$. $\operatorname{Inf} \{L(x, b_0) \mid x \in A\} > -\infty$ and $\sup \{L(a_0, v) \mid v \in B\} < +\infty$.
3. There exists $b \in B$ such that $L(\cdot, b)$ is lower compact on A.
4. There exists $a \in A$ such that $L(a, \cdot)$ is upper compact on B.

Consider the problems:

$$(3.1) \qquad (\mathrm{P}) \quad \alpha = \operatorname{Inf} \{f(x) \mid x \in E\}, \qquad \Omega_P = \{x \in A \mid f(x) = \alpha\},$$

$$(3.2) \qquad (\mathrm{D}) \quad \beta = \operatorname{Sup} \{g(v) \mid v \in F\}, \qquad \Omega_D = \{v \in B \mid g(v) = \beta\},$$

where

$$(3.3)\qquad f(x)=\begin{cases}\operatorname*{Sup}_{v\in B} L(x,v), & \text{if } x\in A,\\ +\infty, & \text{if } x\notin A,\end{cases}$$

$$(3.4)\qquad g(v)=\begin{cases}\operatorname*{Inf}_{x\in A} L(x,v), & \text{if } v\in B,\\ -\infty, & \text{if } v\notin B,\end{cases}$$

when $\alpha=\beta$, α is called the saddle value of the triplet (L, A, B). We consider the problem

(S) *find the saddle value and the saddle points* $(\bar{x}, \bar{v})$, *of the triplet* (L, A, B). (i.e., $L(\bar{x}, v)\leqq L(\bar{x}, \bar{v})\leqq L(x, \bar{v})$ for all $x\in A$, $v\in B$).

Problems (P) and (D) are called respectively, the *primal* and *dual* problems associated with the problem (S). It is well known that if the triplet (L, A, B) has a saddle point $(\bar{x}, \bar{v})$ then $\alpha=\beta=L(\bar{x}, \bar{v})$ and $\bar{x}\in\Omega_P$, $\bar{v}\in\Omega_D$. Conversely, if $\Omega_P\times\Omega_D\neq\phi$ and $\alpha=\beta$, then any element of $\Omega_P\times\Omega_D$ is a saddle point of (L, A, B).

The fact that $3.\mathscr{L}$ guarantees the existence of saddle points of (L, A, B) can be easily obtained by the results of conjugate functions and duality theory elaborated by J. J. Moreau [6], R. T. Rockafellar [13] and P. J. Laurent [6] etc. However, because of the important role of assumptions $3.\mathscr{L}$ in our paper, we give here a proof using the results of the works mentioned above.

PROPOSITION 3.1. *If the conditions* 1, 2 *and* 3 *of* $3.\mathscr{L}$ *are satisfied then the triplet* (L, A, B) *has a saddle value (i.e.,* $\alpha=\beta$*) and* $\Omega_P=\phi$. *If all the conditions of* $3.\mathscr{L}$ *are satisfied then the triplet* (L, A, B) *has at least one saddle point.*

Proof. Let E^* be the topological dual of E and denote by $w(E^*, E)$ and $\tau(E^*, E)$ the weakest and the strongest topologies of E^* for which the dual of E^* can be identified with E. We define

$$\hat{L}(x,v)=\begin{cases}L(x,v), & \text{if } x\in A \text{ and } v\in B,\\ +\infty, & \text{if } x\notin A \text{ and } v\in B,\\ -\infty, & \text{if } v\notin B,\end{cases}$$

and denote $\hat{L}_v=\hat{L}(\cdot, v)$, $\hat{L}_x=\hat{L}(x,\cdot)$. Then the function $\hat{L}_x\colon F\to\bar{R}$ is concave and upper semicontinuous for all $x\in A$, and the function $\hat{L}_v\colon E\to\bar{R}$ is convex, lower semicontinuous and $(\hat{L}_v)^{**}=\hat{L}_v$ for all $v\in B$.

We define a function $\psi\colon E^*\times F\to\bar{R}$ with

$$\psi(y,v)=\operatorname*{Sup}_{x\in E}((x,y)-\hat{L}(x,v))=(\hat{L}_v)^*(y).$$

Then ψ is convex and lower semicontinuous and the function

$$k(y)=\operatorname*{Inf}_{v\in F}\psi(y,v)$$

is convex (see [3, Thm. 1]). It is easy to see that

$$k(0)=-\operatorname*{Sup}_{v}\operatorname*{Inf}_{x}\hat{L}(x,v)=-\beta,$$

$$k^*(x)=\operatorname*{Sup}_{v}\hat{L}v(x),$$

$$k^{**}(0)=-\operatorname*{Inf}_{x}\operatorname*{Sup}_{v}\hat{L}(x,v)=-\alpha.$$

Condition $3.\mathscr{L}.2$ implies that both α and β are finite. On the other hand, since the function $\hat{L}_b \colon E \to \bar{R}$ is lower compact, applying J. J. Moreau's theorem about the duality between the τ- continuity of a function and the weak lower compactness of its conjugate (see[6]), we deduce that the function $\psi(\cdot\,, b)$ is $\tau(E^*, E)$-continuous at $0 \in E^*$. Therefore the function k is bounded above on a $\tau(E^*, E)$-neighborhood of $0 \in E^*$ and hence $\tau(E^*, E)$-continuous at $0 \in E^*$ (see [3, Thm. 6.2.7]). We have then $k(0) = k^{**}(0)$ which, with the above relations, imply $\alpha = \beta$. Condition 3 of $3.\mathscr{L}$ implies that f is lower compact. Therefore $\Omega_P \neq \phi$. If condition 4 of $3.\mathscr{L}$ is also satisfied then the function g is upper compact and $\Omega_D \neq \phi$. Thus $\Omega_P \times \Omega_D \neq \phi$ and any element of $\Omega_P \times \Omega_D$ is a saddle point of (L, A, B).

4. Partial zero limit modifications. Consider a sequence of functions $H_n \in \Gamma_0(F)$ satisfying

$4.\mathscr{A}.$ 1. assumptions $2.\mathscr{A}$ with respect to the nonempty closed convex subset B of F,
2. the functions $h_n = (1/r_n)H_n$ and $h = \lim_{n\to\infty} h_n$ are strictly convex and the problem

(S_n) *find the saddle value and the saddle points of the triplet* (L_n, A, B) *where* $L_n(x, v) = L(x, v) - H_n(v)$.

Remark 4.1. By this modification of the problem (S), the problems (P) and (D) are also modified. The objective functions of the primal problem (P_n) and dual problem (D_n) associated with (S_n) are

$$f_n(x) = \begin{cases} \operatorname{Sup}\{L_n(x, v) \mid v \in B\}, & \text{if } x \in A, \\ +\infty, & \text{if } x \notin A, \end{cases} \tag{4.1}$$

$$\begin{aligned} g(v) &= \begin{cases} \operatorname{Inf}\{L_n(x, v) \mid xA\}, & \text{if } v \in B, \\ -\infty & \text{if } v \notin B, \end{cases} \\ &= g(v) - H_n(v), \end{aligned} \tag{4.2}$$

where $g(v)$ is defined as in (3.4). Therefore the problem (D_n) is a zero limit modification of (D) and all the results of § 2 can be applied to the family of problems (D_n). However, the relationship between (P_n) and (P) cannot be explained by zero limit modification. This remark will lead us to the consideration of dual modifications. (See § 5.)

THEOREM 4.1. *If the triplet* (L, A, B) *satisfies conditions* 1, 2, *and* 3 *of* $3.\mathscr{L}$ *and the sequence* H_n *satisfies assumptions* $4.\mathscr{A}$ *then the triplet* (L, A, B) *has a saddle value* α *and*

1. *For every* $n \in \mathbb{N}$, *the triplet* (L_n, A, B) *has at least one saddle point.*
2. *If* α_n *is the saddle value of* (L_n, A, B), *then* $\lim_{n\to+\infty} \alpha_n = \alpha$.
3. *Let* (x_n, v_n) *be a sequence such that* (x_n, v_n) *is a saddle point of* (L_n, A, B); *then* (x_n, v_n) *has cluster points if and only if* $\Omega_D \neq \phi$ *and all the cluster points are saddle points of* (L, A, B). *Furthermore, in that case, the sequence* (v_n) *actually converges and its limit minimizes the function* h *on* Ω_D.

Proof. The existence of a saddle value α follows from Proposition 3.1.

1. Let $a_0 \in A$ be as in $3.\mathscr{L}.2$; then the function $L(a_0, \cdot) - H_n$ is upper compact. Therefore, all the conditions of $3.\mathscr{L}$ are satisfied for (L_n, A, B).

2. Since α_n is the value of the problem (D_n), Remark 4.1 and Theorem 2.1 imply that α_n converges to the value α of the problem (D).

3. First we will prove that the sequence (x_n) has cluster points and all its cluster points are solutions of (P). Let $b \in B$ be as in $3.\mathscr{L}.3$. We have

$$L(x_n, b) = L_n(x_n, b) + H_n(b).$$

Since (x_n, v_n) is a saddle point of (L_n, A, B), we have

$$L_n(x_n, b) \leqq L_n(x_n, v_n) = \alpha_n.$$

Thus

$$L(x_n, b) \leqq \alpha_n + H_n(b).$$

The lower compactness of $L(\cdot, b)$ and the boundedness of $\alpha_n + H_n(b)$ imply that the sequence (x_n) is included in a compact subset of A, and therefore has cluster points. Let u be an arbitrary element of B. We have

$$(4.3) \qquad L(x_n, u) = L_n(x_n, u) + H_n(u) \leqq L_n(x_n, v_n) + H_n(u) = \alpha_n + H_n(u).$$

Since α_n converges to α and $H_n(u)$ converges to 0, for every $\varepsilon > 0$, there exists $m \in \mathbb{N}$ such that

$$(4.4) \qquad \begin{aligned} L(x_n, u) \leqq \alpha_n + H_n(u) \leqq \alpha + \varepsilon \quad &\text{for all } n > m,\\ x_n \in T_{\alpha,\varepsilon} = \{x \in A \mid L(x, u) \leqq \alpha + \varepsilon\} \quad &\text{for all } n > m. \end{aligned}$$

Since $L(\cdot, u)$ is lower semicontinuous on A, it follows that $T_{\alpha,\varepsilon}$ is closed and any cluster point of (x_n) belongs to $T_{\alpha,\varepsilon}$. That is $L(\bar{x}, u) \leqq \alpha + \varepsilon$; hence $L(\bar{x}, u) \leqq \alpha$. Since $u \in B$ was arbitrary, we conclude that

$$f(\bar{x}) = \operatorname*{Sup}_{u \in B} L(\bar{x}, u) \leqq \alpha.$$

This is possible only if $f(\bar{x}) = \alpha$. Therefore $\bar{x} \in \Omega_P$. On the other hand, Theorem 2.1, Remarks 2.3 and 4.1 imply that the sequence (v_n) is convergent if and only if $\Omega_D \neq \phi$, and if $\bar{v} = \lim_{n \to \infty} v_n$ then $\bar{v} \in \Omega_D$ and

$$h(\bar{v}) \leqq h(w) \quad \text{for all } w \in \Omega_D.$$

Remark 4.2. The rate of convergence of α_n to α is proportional to the rate of $\lim_{n \to +\infty} 1/r_n = 0$, as it can be seen by applying Remark 2.2 to the problems (D_n) and (D).

5. Dual modifications. In Remark 4.1 we pointed out that although the problem (D_n) is a zero limit modification of (D), the problem (P_n) is not obtained by the same type of modification. On the other hand Theorem 4.1 shows that the sequence of problems (P_n) converges to the problem (P).

DEFINITION 5.1. The modification (P_n) of (P) obtained by means of the partial zero limit modification (S_n) of (S) will be called a *dual zero limit modification* of P.

With the above definition, Theorem 4.1 can be summarized as follows:

A sequence of dual zero limit modifications is convergent.

Example 5.1. Consider the convex programming problem

$$(\mathrm{P}) \qquad \alpha = \operatorname{Inf}\{f_0(x) \mid x \in A \text{ and } q_i(x) \leqq 0 i = 1, 2, \cdots, m\},$$

where $f_0 \in \Gamma_0(E)$, $q_i \in \Gamma_0(E)$ $i = 1, \cdots, m$.

The above problem (P) is the primal problem associated with the triplet $(L, A, \mathbb{R}^m_+)$, where

$$L(x, v) = f_0(x) + \sum_{i=1}^{m} v_i q_i(x), \quad x \in A, \quad v = (v_1, v_2, \cdots, v_m) \in \mathbb{R}^m_+.$$

If α is finite, f_0 is lower compact on A and there exists $v^0 \in R_+^m$ such that

$$\operatorname*{Inf}_{x\in A}\left(f_0(x)+\sum_{i=1}^m v_i^0 q_i(x)\right)>-\infty,$$

then by virtue of Proposition 3.1, the triplet $(L, A, \mathbb{R}_+^m)$ has a saddle value and $\Omega_P \neq \phi$.

Consider a sequence of functions $H_n \in \Gamma_0(\mathbb{R}^m)$ having the form $H_n(v_1, v_2, \cdots, v_m) = \sum_{i=1}^m H_{n,i}(v_i) = (1/r_n)\sum_{i=1}^m h_{n,i}(v_i)$ where

5.$\mathscr{A}$.
1. $H_{n,i}(v_i) = +\infty$ for $v_i < 0$ and $H_{n,i}(v_i) \geqq 0$ for $v_i \geqq 0$,
2. $H_{n,i}(0) = 0$,
3. the sequences $(H_{n,i})$ $(i = 1, \cdots, m)$ satisfy the assumptions 4.$\mathscr{A}$ with respect to the set $\mathbb{R}_+$.

We define the triplet $(L_n, A, \mathbb{R}_+^m)$ with $L_n(x, v) = L(x, v) - H_n(v)$. The objective function of the primal problem (P_n) associated with $(L_n, A, \mathbb{R}_+^m)$ has the form

$$f_n(x) = f_0(x) + \sum_{i=1}^m Q_{n,i}(x),$$

where

$$Q_{n,i}(x) = \operatorname*{Sup}_{v_i \geqq 0}\,(v_i q_i(x) - H_{n,i}(v_i)) \begin{cases} =0 & \text{if } q_i(x) \leqq 0, \\ \geqq 0 & \text{if } q_i(x) > 0. \end{cases} \tag{5.1}$$

Indeed, if $q_i(x) \leqq 0$ since $H_{n,i}(v_i) \leqq 0$ for all $v_i \leqq 0$, we have $v_i q_i(x) - H_{n,i}(v_i) \leqq 0$ for all $v_i \geqq 0$, and zero for $v_i = 0$. On the other hand if $q_i(x) > 0$ then $Q_{n,i}(x) \geqq 0$ and $\lim_{n\to+\infty} Q_{n,i}(x) = +\infty$. To prove that, let us fix $k_i > 0$. Since by assumption $\lim_{n\to+\infty} H_{n,i}(v_i) = \lim_{n\to+\infty} (1/r_n)h_{n,i}(v_i) = 0 \cdot h_i(v_i) = 0$, for any $\varepsilon > 0$ there exists $n_0 \in \mathbb{N}$ such that $0 \leqq H_{n,i}(k_i) \leqq \varepsilon$ for all $n > n_0$. Then

$$Q_{n,i}(x) \geqq k_i q_i(x) - H_{n,i}(k_i) \geqq k_i q_i(x) - \varepsilon \quad \text{for all } n > n_0.$$

Since $q_i(x) > 0$ and k_i was arbitrary, that shows $\lim_{n\to+\infty} Q_{n,i}(x) = +\infty$ for each x such that $q_i(x) > 0$. Therefore the function

$$Q_n = \sum_{i=1}^m Q_{n,i}$$

is an exterior penalty function. If $H_{n,i}(v_i) = (1/(n \cdot s))v_i^s$ for $v_i > 0$ we have

$$f_n(x) = f_0(x) + \frac{s-1}{s} n^{1/(s-1)} \sum_{i=1}^m [q_i(x)_+]^{s/(s-1)}.$$

For $s = 2$, we obtain the classical penalty function $Q_{n,i}(x) = (n/2)[q_i(x)_+]^2$. Thus, the exterior penalty methods can appear as dual zero limit modifications. However, it is not true that all exterior penalty functions can be so viewed. For example the function

$$Q_n(x) = n \sum_{i=1}^m q_i(x)_+$$

satisfies the requirements for an exterior penalty function (see [2]). The zero limit modification corresponding to this exterior penalty function is

$$H_n(v) = \sum_{i=1}^m \delta_n(v_i),$$

where

$$\delta_n(v_i) = \begin{cases} 0, & \text{if } 0 \leqq v_i \leqq n, \\ +\infty, & \text{otherwise.} \end{cases}$$

Such a sequence (H_n) obviously does not satisfy all the Assumptions $5.\mathscr{A}$. Actually, only conditions 1, 2 of $5.\mathscr{A}$ and the fact that $\lim_{n\to+\infty} H_{n,i}(v_i)=0$, are enough to show that $Q_{n,i}$ is an exterior penalty function, even without the convexity assumptions on f and $q_i (i=1,\cdots,m)$. However, the results of Theorem 4.1 are not applicable in the nonconvex case, where there is usually a duality gap $(\beta\neq\alpha)$.

6. Total zero limit modifications in convex-concave saddle point problems. Consider a function $h: E\times F\to\bar{\mathbb{R}}$, where E and F are locally convex topological vector spaces such that

$6.\mathscr{A}$.
1. $h(\cdot, v)$ is strictly convex, lower semicontinuous, lower compact and bounded below on A for all $v\in B$,
2. $h(x,\cdot)$ is strictly concave, upper semicontinuous, upper compact and bounded above on B for all $x\in A$,

where $A\subset E$ and $B\subset F$ are nonempty convex closed sets.

Consider a triplet (L, A, B) satisfying $3.\mathscr{L}$ and the problem

(Z_n) *find the value and the saddle points of the triplet* (L_n, A, B),

where $L_n(x,v)=L(x,v)+(1/r_n)h(x,v)$ and (r_n) is a positive sequence such that $\lim_{n\to+\infty} r_n=+\infty$.

THEOREM 6.1. *Under the assumptions* $3.\mathscr{L}$ *and* $6.\mathscr{A}$ *we have*:

1. *For all* $n\in\mathbb{R}$, *the problem* (Zn) *has a unique saddle point* $(x_n, v_n)\in A\times B$.
2. *The value* α_n *of* (L_n, A, B) *converges to the value* α *of* (L, A, B) *and the sequence* (x_n, v_n) *converges to one of the saddle points of* (L, A, B) *which is the* (*unique*) *saddle point of the triplet* (h, Ω_P, Ω_D) (*where* Ω_P *and* Ω_D *are the solution sets of the primal and dual problems associated with the triplet* (L, A, B)).

Proof 1. One easily verifies that the triplet (L_n, A, B) satisfies the assumptions $3.\mathscr{L}$. 2. We have

$$L(x_n, v)+(1/r_n)h(x_n, v)\leqq\alpha_n\leqq L(x, v_n)+(1/r_n)h(x, v_n) \tag{6.1}$$

for all $x\in A$ and $v\in B$.

If $(\tilde{x}, \tilde{v})$ is a saddle point of (L, A, B), we have

$$L(\tilde{x}, v_n)\leqq\alpha\leqq L(x_n, \tilde{v})\quad\text{for all } n\in\mathbb{N}. \tag{6.2}$$

Thus (6.1) (with $x=\tilde{x}$, $v=\tilde{v}$) and (6.2) imply

$$\alpha+(1/r_n)h(x_n, \tilde{v})\leqq\alpha_n\leqq\alpha+(1/r_n)h(\tilde{x}, v_n). \tag{6.3}$$

On the other hand, the assumptions $6.\mathscr{A}$. imply

$$h_{\sup}(x)=\operatorname{Sup}\{h(x,v)\mid v\in B\}<+\infty\quad\text{for all } x\in A, \tag{6.4}$$

$$h_{\inf}(v)=\operatorname{Inf}\{h(x,v)\mid x\in A\}>-\infty\quad\text{for all } v\in B. \tag{6.5}$$

Thus from (6.3) we deduce

$$\alpha+(1/r_n)h_{\inf}(\tilde{v})\leqq\alpha_n\leqq\alpha+(1/r_n)h_{\sup}(\tilde{x}) \tag{6.6}$$

which implies $\lim_{n\to+\infty}\alpha_n=\alpha$.

The relation (6.3) implies

$$-\infty<h_{\inf}(\tilde{v})\leqq h(x_n, \tilde{v})\leqq h(\tilde{x}, v_n)\leqq h_{\sup}(\tilde{x})<+\infty \tag{6.7}$$

for all $\tilde{x}\in\Omega_P$, $\tilde{v}\in\Omega_D$ and $n\in\mathbb{N}$. Therefore the lower compactness of $h(\cdot, \tilde{v})$ and the upper compactness of $h(\tilde{x}, \cdot)$ imply that each of the sequences (x_n) and (v_n) and therefore the sequence (x_n, v_n) is included in a compact subset of $A\times B$, which guarantees the existence of at least one cluster point in $A\times B$.

The relations (6.1), (6.4) and (6.5) imply

$$L(x_n, v)+(1/r_n)h_{\inf}(v)\leqq \alpha_n \leqq L(x, v_n)+(1/r_n)h_{\sup}(x)$$

for all $x\in A$ and $v\in B$. The lower (resp. upper) semicontinuity of $L(\,\cdot\,, v)$ (resp. $L(x, \cdot\,)$) and the fact that $r_n \to +\infty$ imply then that for any cluster point $(\bar{x}, \bar{v})$ of (x_n, v_n), we have

$$L(\bar{x}, v)\leqq \alpha \leqq L(x, \bar{v}) \quad \text{for all } x\in A \text{ and } v\in B.$$

Therefore $(\bar{x}, \bar{v})$ is a saddle point of (L, A, B).

Let us finally show that $(\bar{x}, \bar{v})$ is a saddle point of the triplet (h, Ω_P, Ω_D). The relation (6.3) implies

$$h(x_n, \tilde{v})\leqq h(\tilde{x}, v_n) \quad \text{for all } \tilde{x}\in\Omega_P,\ \tilde{v}\in\Omega_D \text{ and } n\in\mathbb{N}.$$

The lower (resp. upper) semicontinuity of $h(\,\cdot\,, \tilde{v})$ (resp. $h(\tilde{x}, \cdot\,)$) implies then

$$h(\bar{x}, \tilde{v})\leqq (\tilde{x}, \bar{v}) \quad \text{for all } \tilde{x}\in\Omega_P,\quad \tilde{v}\in\Omega_D$$

which shows that $(\bar{x}, \bar{v})$ is a saddle point of the triplet (h, Ω_P, Ω_D). This fact is true for every cluster point of the sequence (x_n, v_n), but the strict convexity-strict concavity of h and the assumptions $6.\mathscr{A}$ imply that such a saddle point is unique. Thus the sequence (x_n, v_n) has a unique cluster point. Therefore it is convergent.

Remark 6.1. The relation (6.6) shows that the rate of convergence of α_n to α is proportional to the rate of $\lim_{n\to+\infty} 1/r_n = 0$. Thus, while the partial and total zero limit modifications have the same rate of convergence (see Remark 4.2), the total zero limit modifications have two important advantages over the partial zero limit modifications. In partial modifications, only the sequence (v_n) is convergent (and the sequence (x_n) has only cluster points). Whereas, in total modifications, both sequences (x_n) and (v_n) are convergent. The other advantage is that here the function $L_n(x, v)$ is strictly convex in x and strictly concave in v, while in partial zero limit modifications, the function $L_n(x, v)$ is only strictly concave in v.

7. Translations in convex optimization: Consider a function $f\in\Gamma_0(E)$ (where E is a locally convex, Hausdorff vector space) and the problem

$$\text{(P)} \qquad \alpha + \operatorname{Inf}\{f(x)\mid x\in E\},\ \Omega=\{x\in E\mid f(x)=\alpha\}.$$

Let $H\in\Gamma_0(E)$ be a function satisfying

$7.\mathscr{A}$.
1. $H(x)\neq+\infty$ and H is continuous for all $x\in E$,
2. $H(x)\geqq 0$ and $H(x)=0$ if and only if $x=0$,
3. $\partial H(0)=\{0\}$,
4. H is strictly convex.

It is easy to see that if α is finite and either f or H is lower compact on E then the problem

$$\operatorname{Inf}\{f(x)+H(x-z)\mid x\in E\} \tag{7.1}$$

has a unique solution x_z for any given $z\in E$. Thus we establish the correspondence

$$U\colon z\in E\to x_z\in E. \tag{7.2}$$

PROPOSITION 7.1. *If α is finite and H satisfies $7.\mathscr{A}$ then the two following statements are equivalent*

1. $\bar{x} \in \Omega$
2. $\bar{x} = U(\bar{x})$.

Proof. $1 \Rightarrow 2$ is obvious since H is always nonnegative. $2 \Rightarrow 1$. We set

$$H_{-z}(x) + H(x-z) \quad \text{for all } x \in E.$$

By assumption, we have

$$f(\bar{x}) \leqq f(x) + H(x-\bar{x}) \quad \text{for all } x \in E.$$

Thus $0 \in \partial(f + H_{-\bar{x}})(x)$. Since H is continuous, we can apply the additivity of subdifferentials (see [8]). We conclude that $0 \in \partial f(\bar{x}) + \partial H(\bar{x} - \bar{x})$. Since $\partial H(0) = \{0\}$, we have $0 \in \partial f(\bar{x})$. Q.E.D.

The mapping U can be expressed in terms of the subdifferentials of f and H. We note by $\partial^{-1}H$ the inverse of the multifunction $\partial H \colon E \to E^*$.

LEMMA 7.1. *Under the assumptions $7.\mathscr{A}$., $\partial^{-1}H$ is single valued.*

Proof. Suppose that for $x_1 \neq x_2$ we have $\{x_1, x_2\} \subset \partial^{-1}H(y_0)$ for some $y_0 \in E^*$. That means $y_0 \in \partial H(x_1) \cap \partial H(x_2)$. Then the strict convexity of H implies that

$$H(x_2) > H(x_1) + \langle y_0, x_2 - x_1 \rangle,$$

$$H(x_1) > H(x_2) + \langle y_0, x_1 - x_2 \rangle$$

which is a contradiction.

PROPOSITION 7.2. *The operator $(I - \partial^{-1}H(-\partial f))^{-1} \colon E \to E$ is single valued and represents the mapping U defined in* (7.2).

Proof. Consider $x_0 \in (I - \partial^{-1}H(-\partial f))^{-1}(z_0)$. Then we have $z_0 \in x_0 - \partial^{-1}H(-\partial f(x_0))$ and $x_0 - z_0 \in \partial^{-1}H(-\partial f(x_0))$, which is equivalent to $(-\partial f(x_0) \cap \partial H(x_0 - z_0)) \neq \phi$. Then there exists $y_0 \in -\partial f(x_0)$ and $y_0 \in \partial H(x_0 - z_0)$. Therefore

$$0 \in \partial f(x_0) + \partial H(x_0 - z_0) = \partial(f + H_{-z_0})(x_0)$$

and hence $x_0 = U(z_0)$. Q.E.D.

Proposition 7.1 establishes the equivalence between the problem (P) and the problem of finding the fixed points of U. Frequently the fixed point problems are studied assuming some contraction properties. The assumptions we have made on H are too general to imply that U is a contraction. The case where H is a Hilbert space and $H(x) = (1/2)\|x\|^2$ has been studied extensively. In this case the function U was named prox_f by Moreau [5] who studied the problem of decomposition of an element z by $z = \operatorname{prox}_f(z) + \operatorname{prox}_{f^*}(z)$. The results of Moreau have since been generalized by Wexler [14].

The proximal function prox_f is proved in [5] to be pseudo contracting i.e.

$$\|\operatorname{prox}_f(x) - \operatorname{prox}_f(x')\|^2 \leqq \|x - x'\|^2 - \|\operatorname{prox}_f(x) - x - (\operatorname{prox}_f(x') - x')\|^2.$$

Martinet [4] showed later that for any pseudo contracting mapping V, the sequence $x_{n+1} = V(x_n)$ (for any initial x_0) converges weakly to a fixed point of V. More general results are obtained by Rockafellar [10] who introduced the proximal point algorithm to find an element $\bar{x}$ such that $0 \in T(\bar{x})$ where T is a maximal monotone multifunction defined in a Hilbert space. The proximal point algorithm consists of defining a sequence x_n such that $x_{n+1} \approx (I + c_k T)^{-1}(x_n)$ where c_k is a certain positive sequence. The case where $T = \partial f$ (for $f \in \Gamma_0(E)$) corresponds to

$$x_{n+1} \approx \operatorname{Arg\,Min}\{f(x) + (1/2c_k)\|x - x_n\|^2 \mid x \in E\}.$$

Under some reasonable criteria of approximation (of $\dot{x}_{n+1}$ to $(I+c_kT)^{-1}(x_n)$), the sequence (x_n) converges weakly to a point $\bar{x}$ such that $0 \in T(\bar{x})$ (to a solution of problem (P) when $T=\partial f$).

In what follows we are going to find a solution of problem (P) by using a sequence $x_{n+1}=U(x_n)$ where U is the mapping defined in 7.2 and the initial point x_0 is an arbitrary point in dom f. We note

$$(\mathrm{P}_n) \qquad \alpha_n = \operatorname{Inf}\{f(x)+H(x-x_{n-1}) \mid x \in E\}.$$

DEFINITION 7.1. The problem (P_n) defined above is called a translation of problem (P).

THEOREM 7.1. *If α is finite, $f \in \Gamma_0(E)$ is lower compact and H satisfies the assumption 7.$\mathscr{A}$, then for each $n \in \mathbb{N}$, the problem (P_n) has a unique solution x_n and the problems (P_n) converge to* (P). (*Therefore any cluster point of the sequence (x_n) is a fixed point of U.*)

Proof. Under the assumptions made, the function

$$x \to f(x)+H(x-x_{n-1})$$

is lower compact bounded below by α, and strictly convex. Therefore problem (P_n) has a unique solution. On the other hand, we have

$$\alpha_{n+1}=f(x_{n+1})+H(x_{n+1}-x_n) \leqq f(x)+H(x-x_n) \quad \text{for all } x \in E.$$

For $x=x_n$, we get

$$f(x_{n+1})+H(x_{n+1}-x_n) \leqq f(x_n) \quad \text{for all } n \in \mathbb{N}.$$

Since H is nonnegative, we deduce that the sequence $(f(x_n))$ is nonincreasing and bounded from below by α. Therefore $f(x_n)$ converges to some $\beta \geqq \alpha$, and $\lim_{n\to+\infty} H(x_{n+1}-x_n)=0$. Therefore $\lim_{n\to+\infty}\alpha_n=\beta$. We will prove that $\beta=\alpha$.

Since the sequence $f(x_n)$ is decreasing, we have $x_n \in S_0=\{x \in E \mid f(x) \leqq f(x_0)\}$. S_0 is compact by assumption. Therefore the sequence (x_n) has at least one cluster point. On the other hand, the convergence of α_n to β implies that for every $\varepsilon>0$, there exists $M \in \mathbb{N}$ such that, for all $n>M$, we have

$$\beta-\varepsilon \leqq \alpha_n \leqq f(x)+H(x-x_{n-1}) \quad \text{for all } x \in E.$$

Hence

$$(7.3) \qquad \beta-\varepsilon \leqq \operatorname{Inf}\{f(x)+H(x-x_{n-1}) \mid n>M\} \quad \text{for all } x \in E.$$

If $\bar{x}$ is a cluster point of the sequence (x_n), the continuity of H and (7.3) imply that

$$\beta-\varepsilon \leqq f(x)+H(x-\bar{x}) \quad \text{for all } x \in E \text{ and } \varepsilon>0.$$

Hence

$$(7.4) \qquad \beta \leqq f(x)+H(x-\bar{x}) \quad \text{for all } x \in E.$$

In particular

$$\beta \leqq f(\bar{x}).$$

On the other hand, the lower semicontinuity of f implies that $f(\bar{x}) \leqq \beta$. Therefore

$$(7.5) \qquad f(\bar{x}) \leqq f(x)+H(x-\bar{x}) \quad \text{for all } x \in E.$$

Thus $\bar{x}=U(\bar{x})$ and, by Proposition 7.1, we have $\beta=\alpha=f(\bar{x})$. Q.E.D.

8. Translations in Banach spaces. Let E be a Banach case. In this case, the assumptions of Theorem 3.1 (continuity of H and lower compactness of f for the same topology) may be too restrictive. We now assume that H is continuous in the norm topology and f is lower compact for the weak topology $\sigma(E, E^*)$ (where E^* is the topological dual of E). This requires some modifications of the assumptions 7.$\mathscr{A}$.

Consider a function $H \in \Gamma_0(E)$ such that:

1. $H(x) \neq +\infty$, and H is continuous in the norm topology for all $x \in E$.
2. $H(x) \geqq 0$ for all $x \in E$, and $H(x) = 0$ if and only if $x = 0$.
3. For all $x \in E$, $\partial H(x)$ is a singleton whose only element we denote by $H'(x)$.
4. $H'(0) = 0$, and the mapping H': $E \to E^*$ is continuous at $x = 0$ for the norm topologies of E and E^*.
5. For any sequence (w_n) of elements of E, $\lim_{n\to+\infty} H(w_n) = 0$ implies $\lim_{n\to+\infty} \|w_n\| = 0$.
6. H is strictly convex.

8.$\mathscr{A}$.

THEOREM 8.1. *If α is finite, f is weakly lower compact, and H satisfies the assumptions* 8.$\mathscr{A}$, *then the problem* (P_n) *defined in* (7.4) *converges to* (P).

Proof. As in the proof of Theorem 4.1 we observe that the sequence $(f(x_n))$ is decreasing and there exists $\beta \geqq \alpha$ such that

$$\lim_{n\to+\infty} f(x_n) = \lim_{n\to+\infty} \alpha_n = \beta, \tag{8.1}$$

$$\lim_{n\to+\infty} H(x_{n+1} - x_n) = 0. \tag{8.2}$$

The last relation and 5.$\mathscr{A}$.5 imply

$$\lim_{n\to+\infty} \|x_{n+1} - x_n\| = 0. \tag{8.3}$$

By definition of the subdifferential we have

$$0 \in \partial(f + H_{-x_n})(x_{n+1}). \tag{8.4}$$

The additivity of the subdifferential remains valid (continuity of H), and we have

$$0 \in \partial f(x_{n+1}) + H'(x_{n+1} - x_n). \tag{8.5}$$

Thus there exists a point $y_{n+1} = -H'(x_{n+1} - x_n) \cap \partial f(x_{n+1})$. The relation (8.3) and the assumption 8.$\mathscr{A}$.4 imply therefore that

$$\lim_{n\to+\infty} \|y_{n+1}\|_{E^*} = 0. \tag{8.6}$$

On the other hand, the sequence (x_n) is contained in $S_0 = \{x \in E \mid f(x) \leqq f(x_0)\}$ which (provided $x_0 \in \operatorname{dom} f$) is, by assumption, weakly compact and therefore bounded. We have

$$f(x) \geqq f(x_{n+1}) + \langle y_{n+1}, x - x_{n+1} \rangle \quad \text{for all } x \in E. \tag{8.7}$$

Since the sequence (x_n) is bounded (8.6) and (8.7) imply that $f(x) \geqq \beta$ for all $x \in E$, which is possible only if $\beta \leqq \alpha$. Thus $\beta = \alpha$. The lower semicontinuity of f and the fact that α is minimal together with (8.1) imply that any cluster point of the sequence (x_n) belongs to Ω.

We could not obtain the convergence of (x_n) when H is any function satisfying 8.$\mathscr{A}$. We also had to assume the weak lower compactness of f to guarantee the existence of cluster points of (x_n). The existence and uniqueness of the cluster points (i.e., the convergence of x_n) obtained in [4] and [11] seems to be due to the nonexpansiveness of the function prox_f. The assumptions $8.\mathscr{A}$ do not seem enough to imply the pseudo contraction or the nonexpansiveness of our function U.

Corollary 8.1 below is a direct generalization of Proposition 8 in [10].

COROLLARY 8.1. *If, in addition to the assumptions of Theorem* 8.1, *we have one of the following*:

a) $\Omega=\{\bar{x}\}$ *and* $0\in \operatorname{Int}(\partial f(\bar{x}))$ (*for the topology of the norm of* E^*),

b) f *is polyhedral*,

then there exists $n_0\in\mathbb{N}$ *such that* $x_n\in\Omega$ *and* $x_n=x_{n_0+1}$ *for all* $n>n_0$.

Proof. If we have a), then, like in the proof of [10, Thm. 3] we obtain the existence of an $\varepsilon>0$ such that $\{\bar{x}\}=\partial^{-1}f(y)$ for all y such that $\|y\|<\varepsilon$. On the other hand since $\lim_{n\to\infty}H'(x_{n+1}-x_n)=0$, there exists $n_0\in\mathbb{N}$ such that $\|y_{n+1}\|=\|H'(x_{n+1}-x_n)\|<\varepsilon$ for $n>n_0$. Since by (8.5) we have $x_{n+1}\in\partial^{-1}f(y_{n+1})$, we conclude that $x_{n+1}=\bar{x}$.

If we have b), we can assume without loss of generality that $\alpha=0$. We define a function $k:E\to\bar{\mathbb{R}}$

$$k(x)=\begin{cases}0 & \text{if } x\in\Omega,\\ +\infty, & x\notin\Omega.\end{cases}$$

It is shown in [10], that there exists a neighborhood T of $0\in E^*$ such that $f^*(y)=k^*(y)$ for all $y\in T$. By (8.6) there exists $n_0\in\mathbb{N}$ such that $y_{n+1}\in T$ for all $n\geqq n_0$ and since $y_{n+1}\in\partial f(x_{n+1})$, we have $x_{n+1}\in\partial k^*(y_{n+1})$, which implies

$$0\in\partial k(x_{n+1})+H'(x_{n+1}-x_0).$$

Therefore $k(x_{n+1})$ is finite and $x_{n+1}\in\Omega$. Hence $U(x_{n+1})=x_{n+1}$ and the sequence x_n is constant for $n>n_0$.

9. Uniform convexity.

DEFINITION 9.1. We say that a *function* $q\in\Gamma_0(E)$ (where E is a Banach space) is uniformly convex, if there exists a nondecreasing function $\delta_q\colon\mathbb{R}_+\to\mathbb{R}_+$ such that

$$\delta_q(r)=0 \quad \text{if and only if } r=0,$$

$$q\left(\frac{a+b}{2}\right)\leqq\frac{q(a)+q(b)}{2}-\delta_q(\|a-b\|)\quad\text{for all } a,\, b\in E.$$

The Propositions 9.1 and 9.2 are proved in [1].

PROPOSITION 9.1. *If* $q\in\Gamma_0(E)$ *is uniformly convex then the series*

$$\Delta_q(r)=\sum_{n=0}^{\infty}2^n\delta_q(r/2^n) \tag{9.1}$$

is convergent for all $r\geqq 0$, *and if* $y_0\in\partial q(x_0)$ *then*

$$q(x)\geqq q(x_0)+\langle y_0, x-x_0\rangle+\sum_{k=0}^{n}2^k\delta_q\left(\frac{\|x-x_0\|}{2^k}\right) \tag{9.2}$$

for all $x\in E$, $n\in\mathbb{N}$.

PROPOSITION 9.2. *If* $q\in\Gamma_0(E)$, *then the function* δ_q *can be chosen such that*

$$\delta_q(2r)\geqq 4\delta_q(r).$$

Remark 9.1. The assumptions $8.\mathscr{A}.5$ and $8.\mathscr{A}.6$ are satisfied if H is uniformly convex.

Remark 9.2. If, in addition to the assumptions of Theorem 8.1, the function f is uniformly convex, then the sequence of solutions (x_n) converges strongly to the unique solution of (P). Indeed, if $\bar{x}$ is the solution of (P), then (9.2) implies that

$$f(x_n) \geqq f(\bar{x}) + 2\delta_f(\|x_n - \bar{x}\|).$$

Since $\lim_{n\to\infty} f(x_n) = f(\bar{x})$, we have $\lim_{n\to\infty} \delta_f(\|x_n - \bar{x}\|) = 0$ which implies $\lim_{n\to\infty} \|x_n - \bar{x}\| = 0$.

10. Partial translations. Consider the triplet (L, A, B) satisfying the assumptions $3.\mathscr{L}$ where F is a Banach space and the upper compactness of $L(a, \cdot)$ is with respect to the weak topology $\sigma(F, F^*)$. Let $H \in \Gamma_0(F)$ be a function satisfying $8.\mathscr{A}$. Under these assumptions, for any $w \in B$ the triplet (L_w, A, B), where $L_w(x, v) = L(x, v) - H(v - w)$, satisfies $3.\mathscr{L}$ and hence admits saddle points. We can thus construct the problems

(S_n) *find the saddle value and saddle points of the triplet* (L_n, A, B)

where $L_n(x, v) = L(x, v) - H(v - v_{n-1})$ and v_{n-1} is the unique solution of the dual problem (D_{n-1}) associated with (L_{n-1}, A, B). (v_0 can be chosen such that $v_0 \in \operatorname{dom} g$, where g is defined as in (3.4).) This way we obtain a partial modification of the problem (S) defined in § 3. The objective function of the dual problem (D_n) is

$$g_n(v) = \operatorname{Inf}\{L_n(x, v) \mid x \in A\} = g(v) - H(v - v_{n-1}).$$

That corresponds to a translation of problem (D) defined in (3.2).

Assumptions $3.\mathscr{L}$ imply the weak upper compactness of the function g. Therefore, all the assumptions of Theorem 8.1 are satisfied. Then if α_n is the saddle value of the triplet (L_n, A, B), α_n converges to the saddle value α of the triplet (L, A, B) and the problem (D_n) converges to the problem (D).

THEOREM 10.1. *Let (L, A, B) and $H \in \Gamma_0(F)$ satisfy assumptions $3.\mathscr{L}$ and $8.\mathscr{A}$, respectively. Then any sequence $(x_n, v_n) \in A \times B$ such that (x_n, v_n) is a saddle point of (L_n, A, B) has one of the following properties*:

1. *There exists $m \in \mathbb{N}$ such that $v_m = v_{m+1}$ and then (x_{m+1}, v_{m+1}) is a saddle point of (L, A, B).*
2. *The sequence (x_n, v_n) has at least one cluster point in the product topology of E and $\sigma(F, F^*)$ and all the limit points are saddle points of (L, A, B).*

Proof. 1. If $v_{m+1} = v_m$, we have

$$L(x_{m+1}, v) - H(v - v_m) \leqq L(x_{m+1}, v_m) \leqq L(x, v_m) \quad \text{for all } x \in A, \quad v \in B. \tag{10.1}$$

The first part of the inequality shows that v_m is a fixed point of the mapping $U_{m+1}: B \to B$ such that, for $w \in B$, $U_{m+1}(w)$ is the unique solution of the problem

$$\operatorname{Sup}\{L(x_{m+1}, v) - H(v - w) \mid v \in B\}.$$

Thus, by virtue of Proposition 7.1, we have

$$L(x_{m+1}, v) \leqq L(x_{m+1}, v_m) \quad \text{for all } v \in B. \tag{10.2}$$

Relations (10.1) and (10.2) imply (x_{m+1}, v_m) is a saddle point of (L, A, B).

2. Problem (D_n) associated with (L_n, A, B) is a translation of the problem (D). Therefore by virtue of Theorem 8.1, the sequence (v_n) has weak cluster points and all these cluster points are solutions of (D).

If (x_n, v_n) is any saddle point of (L_n, A, B) then

$$L(x_n, v) - H(v - v_{n-1}) \leqq L(x_n, v_n) - H(v_n - v_{n-1}) \leqq L(x, v_n) - H(v_n - v_{n-1}) \tag{10.3}$$

for all $x \in A$, $v \in B$. Let us set $\tilde{L}(x_n, v) = -L(x_n, v)$. Then

$$\tilde{L}(x_n, v) + H(v - v_{n-1}) \geqq \tilde{L}(x_n, v_n) + H(v_n - v_{n-1}) \quad \text{for all } v \in B,$$

which implies $0 \in \partial(\tilde{L}(x_n, \cdot)+H_{-v_{n-1}})(v_n)$ (where $H_{-v_{n-1}}$ is defined by $H_{-v_{n-1}}(v)= H(v-v_{n-1})$). By the additivity of subdifferentials we deduce that there exists $u_n \in F^*$ such that $u_n=-H'(v_n-v_{n-1})$, $u_n \in \partial_2\tilde{L}(x_n, v_n)$ (where ∂_2 is the subdifferential with respect to the second variable). Then

$$\tilde{L}(x_n, v) \geqq \tilde{L}(x_n, v_n)+\langle u_n, v-v_n\rangle \quad \text{for all } v \in F$$

or

$$L(x_n, v) \leqq L(x_n, v_n)-\langle u_n, v-v_n\rangle \quad \text{for all } v \in F. \tag{10.4}$$

On the other hand, since the assumptions of Theorem 8.1 are satisfied for the problem (D_n) we have $\lim_{n\to+\infty}\|v_n-v_{n-1}\|=0$, then

$$\lim_{n\to+\infty}\|u_n\|=0. \tag{10.5}$$

By (10.3) we have $L(x_n, v_n)=\operatorname{Inf}\{L(x, v_n) \mid x \in A\}=g(v_n)$. Theorem 8.1 implies that the sequence (v_n) is bounded and $\lim_{n\to+\infty} g(v_n)=\alpha$. Thus in view of (10.5) we conclude that $\lim_{n\to+\infty}\langle u_n, w-v_n\rangle=0$ for all $w \in F$. Hence for all $\varepsilon>0$ and $w \in F$ there exists $m_{w,\varepsilon} \in \mathbb{N}$ such that

$$L(x_n, v_n)-\langle u_n, w-v_n\rangle \leqq \alpha+\varepsilon \quad \text{for all } n>m_{w,\varepsilon}.$$

Thus by (10.4), for all $\varepsilon>0$, $w \in F$ there exists $m_{w,\varepsilon} \in \mathbb{N}$ such that

$$x_n \in K_{w,\varepsilon}=\{x \in A \mid L(x, w) \leqq \alpha+\varepsilon\} \quad \text{for all } n>m_{w,\varepsilon}. \tag{10.6}$$

The relation (10.6) may be written for all $w \in F$. For $w=b$ (see assumptions 3.$\mathscr{L}$) $K_{b,\varepsilon}$ is compact. Thus the sequence (x_n) has cluster points. The sets $K_{w,\varepsilon}$ are closed for every $w \in B$; therefore any cluster point $\bar{x}$ of (x_n) belongs to $K_{w,\varepsilon}$ for all $w \in B$ and $\varepsilon>0$. That means,

$$L(\bar{x}, v) \leqq \alpha+\varepsilon \quad \text{for all } v \in B \text{ and } \varepsilon>0,$$

which implies

$$L(\bar{x}, v) \leqq \alpha \quad \text{for all } v \in B.$$

On the other hand, if $\bar{v}$ is a weak cluster point of (v_n), we have

$$\alpha=g(\bar{v}) \leqq L(x, \bar{v}) \quad \text{for all } x \in A.$$

Therefore

$$L(\bar{x}, v) \leqq L(\bar{x}, \bar{v}) \leqq L(x, \bar{v}) \quad \text{for all } x \in A, \quad v \in B.$$

COROLLARY 10.1. *If, in addition to the assumptions of Theorem* 10.1, *we have one of the following*:

a) $\Omega_D=\{\bar{v}\}$ *with* $0 \in \operatorname{Int}(\partial(-g(\bar{v})))$ *for the norm topology,*

b) g *is polyhedral,*

then for any sequence (x_n, v_n) *such that* (x_n, v_n) *is a saddle point of* (L_n, A, B) *there exists* $n_0 \in \mathbb{N}$ *such that for all* $n>n_0$, (x_n, v_n) *is a saddle point of* (L, A, B).

Proof. Corollary 8.1 implies the existence of an $n_0 \in \mathbb{N}$ such that $v_n \in \Omega_D$ and $v_n=v_{n_0+1}$ for all $n>n_0$; then by Theorem 10.1 we conclude that (x_{n_0+1}, v_{n_0+1}) is a saddle point of (L, A, B).

11. Dual translations. Although (D_n) is obtained by translation from (D), the relation between (P_n) and (P) has a different nature. This act motivates the following

DEFINITION 11.1. The problem (P_n) obtained from (P) by means of the partial translation (S_n) of (S) will be called a *dual translation* of (P).

Example 11.1. Consider the triplet $(L, A, \mathbb{R}^m_+)$ defined in Example 5.1 and the function $H(v) = (1/(2K)) \sum_{i=1}^m v_i^{2K}$ where $K \in \mathbb{N}$. Let (S) and (S_n) be the saddle point problems of $(L, A, \mathbb{R}^m_+)$ and $(L_n, A, \mathbb{R}^m_+)$ respectively. $(L_n(x, v) = L(x, v) - (1/(2K)) \sum_{i=1}^m (v_i - v_{i,n-1})^{2K})$. The dual translation of the problem (P) is

$$(\mathrm{P}_n) \qquad \alpha_n = \mathrm{Inf}\, \{f_n(x) \mid x \in A\},$$

where $f_n(x) = \mathrm{Sup}\, \{L_n(x, v) \mid v \in \mathbb{R}^m_+\} = f_0(x) + \sum_{i=1}^m \psi(q_i(x), v_{i,n-1})$ and

$$\psi(q_i(x) \quad v_{i,n-1}) = \begin{cases} q_i(x) v_{i,n-1} + \dfrac{2K-1}{2K} (q_i(x))^{2K/(2K-1)}, & \text{if } (q_i(x) + (v_{i,n-1})^{2K-1}) \geqq 0, \\ -\dfrac{1}{2K} (v_{i,n-1})^{2K}, & \text{otherwise.} \end{cases}$$

Also $f_n(x) = l_K(x, v_{n-1}, 1)$ where $l_K : E \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ is defined as

$$l_K(x, v, c) = \mathop{\mathrm{Sup}}_{w \in \mathbb{R}^m_+} \left(f_0(x) + \sum_{i=1}^m w_i q_i(x) - \frac{c}{2K} \|w - v\|^{2K} \right).$$

For $K = 1$, l_1 is the well know augmented Lagrangian introduced in [12] and elaborated in [11], [15] among several other works. When $K = 1$ our method corresponds to the method of multipliers (see [11]) which can be resumed as the proximal point algorithm applied to the dual problem (D).

REFERENCES

[1] I. BEHAR, *Procèdes de Regularisation de Problèmes d'optimisation et de Mini-Max*, Thèse, 14 février, 1974, Université Scientifique et Médical de Grenoble, France.

[2] A. V. FIACCO AND G. P. MCCORMICK. *Nonlinear programming: Sequential unconstrained Minimization Techniques*, J. Wiley, New York, 1968.

[3] P. J. LAURENT, *Approximation et Optimisation*, Hermann, Paris, 1972.

[4] B. MARTINET, *Algorithmes pour la Résolution des Problèmes d'optimisation et de Mini-Max*, Thèse le 24 avril, 1972, Université Scientifique et Médical de Grenoble, France.

[5] J. J. MOREAU, *Proximité et dualité dans un espace Hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.

[6] ———, *Sur la polarité d'une fonctionelle semi continue supérieurement*, C. R. Acad. Sci. Paris, 258 (1964) pp. 1128–1130.

[7] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[8] ———, *Extension of Fenchel's duality theorem for convex functions*, Duke Math. J., 33 (1966), pp. 81–89.

[9] ———, *Level sets and continuity of conjugate functions*, Trans. Amer. Math. Soc., 123 (1966), pp. 46–63.

[10] ———, *Monotone operators and the proximal point algorithm*, this Journal, 14 (1976), pp. 877–898.

[11] ———, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res., 1 (1976), pp. 97–116.

[12] ———, *New applications of duality in optimization*, Proc. of the 4th Conf. on Probability, Brasov, Roumania, 1971.

[13] ———, *Conjugate Duality and Optimization*, Conf. Series in Applied Math., Society for Industrial and Applied Mathematics, Philadelphia, 1974.

[14] D. WEXLER, *Prox-mappings associated with a pair of Legendre conjugate functions*, Rev. Française d'Augo, Inform. et Rech. Oper., 7 (1973), pp. 39–65.

[15] A. P. WIERZBICKI AND S. KURCYUSZ, *Projections on a cone, penalty functionals and duality theory for problems with inequality constraints in Hilbert space*, this Journal, 15 (1977), pp. 25–56.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 6, November 1979

0363-0129/79/1706-0002$01.00/0

SUFFICIENT CONDITIONS FOR KUHN-TUCKER VECTORS IN CONVEX PROGRAMMING*

P. LEVINE† AND J. CH. POMEROL†

Abstract. In this paper, we give new sufficient conditions for the existence of a Kuhn-Tucker vector for convex programs. These conditions generalize all the previously known ones.

1. Introduction. The aim of this paper is to give efficient conditions for the existence of a Kuhn-Tucker vector for convex programs in Banach spaces. Three conditions have already been introduced for convex programs, see e.g. Rockafellar [14], [16], [17] and Robinson [13]. A fourth, in the framework of continuous programming, can be found in Grinold [6], [7].

In this paper we shall present sufficient conditions which are easy to handle and which generalize all the previously known conditions.

Our approach will be the following: in § 3, by perturbing the objective function of the program, we prove that the existence of Kuhn-Tucker vectors is related to the closedness of certain convex sets. This fact will lead us to study three types of sufficient conditions (§ 4). In § 5, we study the relationships between our conditions and the known ones. Section 6 is devoted to continuous linear programming for which it is possible to weaken the sufficient conditions of Grinold [6].

2. Statement of the problem. We are concerned with the following convex program (π):

$$\text{Minimize} \quad f(x)$$

subject to

$$Ax - a \in -Q, \qquad x \in P.$$

Here P and Q are closed convex sets, A is a continuous linear map from X into U (where X and U are two real locally convex vector spaces), $a \in U$ and the functional f, from X into $\bar{\mathbb{R}}$ (the extended real line), is convex, lower semicontinuous (lsc) and proper.

We are considering this kind of program to permit comparison with known results. However it can be easily shown that there is no loss of generality under this form.

We shall denote by inf (π) the infimum of $f(x)$ under the constraints $Ax - a \in -Q$, $x \in P$ and we shall assume that inf $(\pi) < +\infty$. Let Y(resp. V) denote the topological dual of U (resp. X). The polar set of a subset $C \subset U$ is defined by $C^\circ = \{y \in Y \mid \forall u \in C \langle u,y\rangle \geqq -1\}$ and the indicator function of C is denoted by ψ_C. In what follows f^* denotes the conjugate of a functional f.

One can associate with (π) the Lagrangian functional K defined by

$$K(x,y) = \begin{cases} f(x) + \langle y, a - Ax\rangle - \psi_Q^*(y) & \text{if } x \in P, \\ +\infty & \text{otherwise.} \end{cases}$$

We are now ready to define a Kuhn-Tucker vector for the program (π):
$\bar{y} \in Y$ is a *Kuhn-Tucker vector* for (π) if:

$$\inf(\pi) = \inf_{x \in X} K(x, \bar{y}).$$

* Received by the editors December 29, 1976, and in revised form November 16, 1978.

† Université P. et M. Curie, 75230 Paris CEDEX 5, France.

The aim of this paper is to derive sufficient conditions for the existence of a Kuhn–Tucker vector for (π). This problem has been treated by many authors and several sufficient conditions have already been proposed. Let us list the sharpest ones known:

(C_1) *P, Q and f are polyhedral in finite dimensional Euclidean spaces;* see e.g. Rockafellar [16, Thm. 29.2].

(C_2) $\forall u \in U, \exists \varepsilon > 0$ *such that* $\varepsilon u \in A(P \cap \operatorname{dom} f) + Q - a$, i.e. $0 \in$ core $(A(P \cap \operatorname{dom} f) + Q - a)$.

This condition appears in Rockafellar [17, Thm. 6], and [18, Thm, 18, c] when U is a Banach space and V a Banach space for a topology compatible with the pairing $\langle X, V\rangle$. When X and U are Banach spaces, this condition is stated by Robinson [13, Cor. 1].

Notice also that (C_2) is implied by (see for instance Rockafellar [14] and [18, pp. 47–48] the well-known condition:

There exists $\bar{x} \in P \cap \operatorname{dom} f$ such that:

$$A\bar{x} \in a - \operatorname{int}(Q).$$

(int (Q) denotes the interior of Q in U.)

A dual condition (C_3), which implies (C_2) is given by Rockafellar [18, Thm 8, e]. This condition will be recalled in § 5.

However these conditions are never fulfilled for some type of programs having Kuhn–Tucker vectors. Let us give a typical example. We consider programs such as:

$$\begin{aligned} &\text{Maximize} && \int_0^1 p(t)x(t)\,dt \\ &\text{subject to} && A(t)x(t) \leqq a(t) \\ &&& x(t) = (x_1(t), \cdots, x_n(t)) \geqq 0 \quad \text{and} \quad x_i \in L^1[0, 1] \end{aligned}$$

where $L^1[0, 1]$ is defined by the Lebesgue measure on [0, 1] and the constraints must be satisfied almost everywhere. The n-vector $p(t)$ and the m-vector $a(t)$ are given such that for every i, $a_i(t)$ and $p_i(t)$ belong to L^∞ [0, 1]. Moreover $a(t)$ is positive, the $(m \times n)$ matrix $A(t)$ takes on a finite number of values, and possesses at least one positive row on a nonzero measurable subset of [0, 1].

The fact that this program has a Kuhn–Tucker vector will be proved in § 6. But neither of the conditions (C_1) and (C_2) can ve applied in the present case. It is clear that (C_1) does not apply. To verify (C_2) we need for every $u \in L^1[0,1]$ a positive ε such that $\varepsilon u(t) \geqq A(t)x(t) - a(t)$ is satisfied with a positive $x(t)$. But this is impossible when u is not bounded from below on the measurable subset where $A(t)$ has a positive row.

The main purpose of this paper is therefore to give sufficient conditions for the existence of Kuhn–Tucker vectors, more general than (C_1) and (C_2) which are fulfilled at least by programs such as the previous ones.

3. Abstract conditions for the existence of Kuhn–Tucker vectors. Let us introduce some notation. The projection from $V \times Y \times \mathbb{R}$ onto $V \times \mathbb{R}$ is denoted by Pr (i.e. Pr $(v, y, t) = (v, t)$). The adjoint mapping of a given linear map is denoted by A^*. The lower semicontinuous hull of a functional f is denoted by lsc f, whereas its upper semicontinuous hull is usc f. In what follows "neighborhood" is written for "closed convex neighborhood."

We define now the following subset of $V \times Y \times \mathbb{R}$:

$$\xi = \{(v, y, t) | \langle a, y\rangle - \psi_Q^*(y) - (f + \psi_P)^*(A^*y + v) \geqq t\}.$$

To motivate the introduction of this set, it is convenient to use the concept of perturbation [15]; see for instance Rockafellar [18] from whom we borrow the notation. Thus let us re-express (π) as:

Minimize $F(x, 0)$ on X

where $F(x, u)=\begin{cases} f(x) & \text{when } x \in P \text{ and } Ax-a+u \in -Q, \\ +\infty & \text{otherwise.}\end{cases}$

We calculate:

$$\begin{aligned} G(y, v) &= \inf_{x,u} (\langle u, y\rangle - \langle x, v\rangle + F(x, u)) \\ &= \inf_{\substack{x\in P \\ q \in Q}} (-\langle x, v\rangle + \langle -q-Ax+a, y\rangle - f(x)) \\ &= \langle a, y\rangle - \sup_{q\in Q} \langle q, y\rangle - \sup_{x \in P} (\langle x, v+A^*y\rangle - f(x)) \\ &= \langle a, y\rangle - \psi_Q^*(y) - (f+\psi_P)^*(v+A^*y). \end{aligned}$$

Note that ξ is just equal to $\{(v, y, t) | G(y, v) \geqq t\}$.

In order to relate the set ξ to the program π we shall define the family of perturbed programs (π_v), $v \in V$, as follows:

$$\begin{aligned} &(\pi_v) \text{ minimize } f(x) - \langle x, v\rangle \\ &Ax - a \in -Q, \qquad x \in P. \end{aligned}$$

Finally we denote by δ_v the line $\delta_v = \{(v, t) | t \in \mathbb{R}\}$ and by $\gamma(v)$ the functional $\sup_{y\in Y} G(y, v)$.

Let us recall the following equivalence:

PROPOSITION 3.1. *The following two assertions are equivalent*:

(i) (π_{v_0}) *has a Kuhn–Tucker vector*;

(ii) $\mathrm{Pr}\,(\xi) \cap \delta_{v_0} = \overline{\mathrm{Pr}(\xi)} \cap \delta_{v_0}$.

Proof. Replacing $F(x, u)$ by $\tilde{F}(x, u) = F(x, u) - \langle x, v_0\rangle$ we obtain in the above formula $\tilde{G}(y, v) = G(y, v_0 + v)$. Let us consider $\tilde{\gamma}(0) = \gamma(v_0)$. Applying to $\tilde{\gamma}$ Rockafellar's results [18, Thm. 16a, and Thm. 15d] it follows that (π_{v_0}) has a Kuhn–Tucker vector $\bar{y}$ if and only if $\mathrm{cl}\,\gamma(v_0) = \gamma(v_0) = G(\bar{y}, v_0)$. The assumption $\inf(\pi) < +\infty$ implies $\inf(\pi_{v_0}) < +\infty$. Thus we are not in Rockafellar's exceptional case [18, Thm. 7]; hence for every $v \in V$ we have $\mathrm{cl}\,\gamma(v) = \mathrm{usc}\,\gamma(v) < +\infty$. Then it is easy to see that $(v, \mathrm{usc}\,\gamma(v)) \in \overline{\mathrm{Pr}(\xi)}$ whenever $\mathrm{usc}\,\gamma(v)$ is finite. It follows that $\overline{\mathrm{Pr}\,(\xi)} = \mathrm{epi}\,\mathrm{usc}\,\gamma$. In other words we have $\mathrm{usc}\,\gamma(v) = \sup\{t | (v, t) \in \overline{\mathrm{Pr}\,(\xi)}\}$, the supremum being a maximum when $\mathrm{usc}\,\gamma(v)$ is finite. Thus the equivalence holds when $\mathrm{usc}\,\gamma(v_0)$ is finite, and it is obvious when $\mathrm{usc}\,\gamma(v_0) = -\infty$, both sets in (ii) being empty. Q.E.D.

Then from Proposition 3.1 come two other results.

PROPOSITION 3.2. *The two following assertions are equivalent*:

(i) $\forall v \in V$, (π_n) *has a Kuhn–Tucker vector*;

(ii) $\mathrm{Pr}\,(\xi)$ *is weak*-closed.*

Proof. Let (v, t) be an element of $\overline{\mathrm{Pr}\,(\xi)}$; then $(v, t) \in \delta_v$. From Proposition 3.1 it follows that $(v, t) \in \mathrm{Pr}(\xi)$. Conversely if $\mathrm{Pr}\,(\xi)$ is weak*-closed, it follows that for every $v \in V$, $\mathrm{Pr}\,(\xi) \cap \delta_n = \overline{\mathrm{Pr}\,(\xi)} \cap \delta_v$ which, at the view of Proposition 3.1, completes the proof.

PROPOSITION 3.3. *If there exist* $t_0 \in [-\infty, +\infty[$ *satisfying* $\gamma(0) \leqq t_0 \leqq \operatorname{usc} \gamma(0)$ *and a neighborhood* $(N \times M)$ *of* $(0, t_0)$ *such that* $\operatorname{Pr}(\xi) \cap (N \times M)$ *is weak*-closed, then* (π) *has a Kuhn–Tucker vector.*

Proof. It suffices to prove that $\operatorname{Pr}(\xi) \cap \delta_0 = \overline{\operatorname{Pr}(\xi)} \cap \delta_0$ (Proposition 3.1). Assume that $\gamma(0) < \operatorname{usc} \gamma(0)$. As $(0, \operatorname{usc} \gamma(0)) \in \operatorname{Pr}(\xi)$ which is a convex set, it follows that $(0, t_0) \in \overline{\operatorname{Pr}(\xi)}$. Taking an arbitrary 0-neighborhood $(N_0 \times M_0)$ such that $N_0 \subset N$ and $t_0 + M_0 \subset M$, we have $[(0, t_0) + (N_0 \times M_0)] \cap \operatorname{Pr}(\xi) \neq \emptyset$. Thus $(0, t_0) \in \overline{\operatorname{Pr}(\xi) \cap (N \times M)}$ which is equal to $\operatorname{Pr}(\xi) \cap (N \times M)$. It follows that $\gamma(0) \geqq t_0$ and there exists y such that $G(y, 0) \geqq t_0$. If t_0 was such that $\gamma(0) < t_0$ it is absurd. If $t_0 = \gamma(0)$ we can replace t_0 by $t'_0 \in \operatorname{int}(N)$ satisfying $t'_0 > \gamma(0)$, which is also absurd. Q.E.D.

Remark 3.1. It has been known for a long time that, in Proposition 3.2. (ii) implies (i); see for instance Kretschmer [10] in linear programming, and Dieter [2] in convex programming. To our knowledge the part (i)$\Rightarrow$(ii) appears firstly for linear programming in Pomerol (3rd cycle dissertation, Université de Paris 6, 1973) and for convex programming in Lévine ("Stabilité, sous-convergence et applications C-fermées" Université de Paris 6, D.P., 1973), then independently in Krabs [9] for linear programs and Gwinner [8] for convex programs.

Let us now deduce from these three propositions three types of sufficient conditions for the existence of Kuhn–Tucker vectors.

4. Sufficient conditions for the existence of Kuhn–Tucker vectors.

First we derive from Proposition 3.1 a sufficient condition for the existence of Kuhn–Tucker vectors which holds at point $v = 0$.

THEOREM 4.1. *Assume that either X is a Banach space or V is normed for a topology compatible with the pairing. Then the following condition implies the existence of a Kuhn–Tucker vector for* (π).

(C_4) $\gamma(0)$ *is finite and there exist a neighborhood M of $\gamma(0)$, a real $k > 0$ and a weak*-compact set B in Y such that: If $(v, t) \in \operatorname{Pr}(\xi)$, $t \in M$, $\|v\| \leqq k$ then there exists $y \in B$ satisfying $G(y, v) \geqq t$.*

Proof. Since $\gamma(0)$ is finite there exists a sequence t_n which converges to $\gamma(0)$ and $G(0, y_n) \geqq t_n$. By (C_4) y_n can be chosen in B. Thus y_n weakly converges to y_0 and $(0, y_0, \gamma(0)) \in \xi$ since ξ is weak*-closed. It follows that $G(0), y_0) = \gamma(0)$ and $(0, \gamma(0)) \in \operatorname{Pr}(\xi)$. Assume now that $\gamma(0) < \operatorname{usc} \gamma(0)$. Let us consider $t_0 \in \operatorname{int}(M)$ and $\gamma(0) < t_0 \leqq \operatorname{usc} \gamma(0)$. Since $(0, t_0)$ belongs to $\overline{\operatorname{Pr}(\xi)}$ there exists a generalized sequence $(v_\alpha, t_\alpha) \in \operatorname{Pr}(\xi)$ which converges to $(0, t_0)$. Thus this sequence is strongly bounded ([1 Chap. IV § 3 No. 2, Prop. 2] when X is a Banach space). Then there exists $k_0 > 0$ such that $\|v_\alpha\| \leqq k_0$. The set $\operatorname{Pr}(\xi)$ being convex $(v'_\alpha, t'_\alpha) = (1-\lambda)(0, \gamma(0)) + \lambda(v_\alpha, t_\alpha)$ belongs to $\operatorname{Pr}(\xi)$ whenever $\lambda \in [0, 1]$. Setting $\lambda = \min(1, k/k_0)$ one obtains $\|v'_\alpha\| \leqq k$. Taking t_α in M we have $t'_\alpha \in M$ because $\gamma(0) \leqq t'_\alpha \leqq t_\alpha$. In view of ($C_4$) there exists $y'_\alpha \in B$ such that $(v'_\alpha, y'_\alpha, t'_\alpha) \in \xi$. The set B being weak*-compact there exists a subsequence of $(v'_\alpha, y'_\alpha, t'_\alpha)$ which converges to $(0, y_0, (1-\lambda)\gamma(0) + \lambda t_0) \in \xi$, implying that $G(0, y_0) \geqq (1-\lambda)\gamma(0) + \lambda t_0 > \gamma(0)$ which is absurd.

We derive now from Proposition 3.2 a sufficient condition for the existence of Kuhn–Tucker vectors which holds for every $v \in V$ (even if $\inf(\pi_v) = -\infty$).

THEOREM 4.2. *Assume that either X is a Banach space or V is normed for a topology compatible with the pairing. Then the following condition implies the existence of a Kuhn–Tucker vector for* (π_v), $\forall v \in V$.

(C_5) $\forall k > 0, \exists B_k$ *a weak*-compact subset of Y such that if $(v, t) \in \operatorname{Pr}(\xi)$, $\|v\| \leqq k$, $|t| \leqq k$, then there exists $y \in B_k$ satisfying $G(y, v) \geqq t$.*

Proof. Let us show that (C_5) implies the closedness of $\operatorname{Pr}(\xi)$. Let (v, t) be an element of $\overline{\operatorname{Pr}(\xi)}$. Then there exists a generalized sequence (v_α, t_α) in $\operatorname{Pr}(\xi)$ which converges to

(v, t). Therefore this sequence is strongly bounded and there exists $k>0$ such that $\|v_\alpha\| \leqq k$, $|t_\alpha| \leqq k$. By (C_5) there exists a sequence (y_α) with $y_\alpha \in B_k$ and $(y_\alpha, v_\alpha, t_\alpha) \in \xi$. The set B_k being weak*-compact we can extract a subsequence of $(y_\alpha, v_\alpha, t_\alpha)$ which weakly converges to $(y, v, t) \in \xi$, which proves the closedness of $\Pr(\xi)$.

Remark 4.1. Theorem 4.1 and its proof may be regarded as conjointly elaborated with R. T. Rockafellar to whom we are greatly indebted.

It is not obvious to deduce Theorem 4.2 from Theorem 4.1. Actually when $\gamma(0) = -\infty$, the condition of Theorem 4.1 does not apply whereas Theorem 4.2 still holds and implies that $\varphi(0) = -\infty$. The original proof of Theorem 4.2 used the concept of C-closed mapping (linear transformation which maps a closed convex set onto a closed one). For these mappings the closedness of $\Pr(\xi)$ directly follows from the closedness of ξ (see [11]).

When U is a Banach space the weak*-compact set B of Theorems 4.1 and 4.2 can be expressed by a set such that $\{y \mid \|y\| \leqq k'\}$. In this case, we deduce from Theorem 4.2 the following sufficient condition which is easy to handle in most examples [12].

(C_7) *There exist two positive numbers α and β such that: For every $(v, t) \in \Pr(\xi)$ one can find $y \in Y$ satisfying $G(y, v) \geqq t$ and $\|y\| \leqq \alpha \max(\|v\|, |t|) + \beta$.*

Remark 4.2. Let us mention that (C_7) has been known for many years. Condition (C_7) is given by Lévine ("Stabilité, sous-convergence et applications C-fermées", Université de Paris 6 D.P., 1973). Independently, in linear programming, Evers [3, Prop. 6.15] and Tröltzsch [19, Satz 1] use (C_7). Evers has also obtained (C_7) for the convex program:

$$\text{Minimize } q(x) \quad \text{subject to} \quad G(x) \leqq a \text{ and } x \geqq 0.$$

(q and G are convex weak*-continuous functions and X a reflexive space, [4, Thm. 21].)

A strong version of (C_7) which consists of replacing "one can find y" by "for every y" has already been given by Eisner and Olsen [5, Thm. 5.1).

Finally we obtain from Proposition 3.3 a condition for the existence of a Kuhn–Tucker vector which holds in locally convex vector spaces and does not ask that $\gamma(0)$ be finite.

THEOREM 4.3. *The following condition implies the existence of a Kuhn–Tucker vector for (π).*

(C_6) *There exist $t_0 \in [-\infty, +\infty[$ satisfying $\gamma(0) \leqq t_0 \leqq \operatorname{usc} \gamma(0)$, a neighborhood M of t_0, a 0-neighborhood N in V for the Mackey topology and a weak*-compact set B in Y such that:*

If $(v, t) \in \Pr(\xi) \cap N \times M$ then there exists $y \in B$ satisfying $G(y, v) \geqq t$.

Proof. Let $(\bar v, \bar t)$ be an element of $\overline{\Pr(\xi) \cap N \times M}$. There exists a generalized sequence (v_α, t_α) converging to $(\bar v, \bar t)$ with $(v_\alpha, t_\alpha) \in \Pr(\xi) \cap N \times M$. By (C_6) there exists $y_\alpha \in B$ such that $(v_\alpha, y_\alpha, t_\alpha) \in \xi$. The set B being weak*-compact there exists a subsequence of $(v_\alpha, y_\alpha, t_\alpha)$ which converges to $(\bar v, \bar y, \bar t) \in \xi$. Thus $G(\bar y, \bar v) \geqq \bar t$ and $(\bar v, \bar t) \in \Pr(\xi) \cap N \times M$, which completes the proof.

In the sequel, we denote by (CS_4) and (CS_5) the strong versions of Conditions (C_4) and (C_5) defined as follows:

(CS_4) *$\gamma(0)$ is finite, there exist a neighborhood M of $\gamma(0)$, and a real $k>0$ such that $\{y \mid G(y, v) \geqq t, \|v\| \leqq k, t \in M\}$ is weak*-compact.*

(CS_5) *$\forall k>0$ the set $\{y \mid G(y, v) \geqq t, \|v\| \leqq k, |t| \leqq k\}$ is weak*-compact.*

5. Comparison with known results. (a) We have shown that in finite dimensional Euclidean spaces (C_1) implies (C_7), [11, Thm. 5.1].

(b) In locally convex spaces Rockafellar has given the following condition [18, Thm. 18.e], which we denote (C_3).

(C3) *There exists a* 0-*neighborhood* N *for the Mackey topology* $\mathcal{T}(V, X)$ *and a real number* β, $\beta < \gamma(0)$, *such that the set*

$$\{y \mid \exists v \in N,\ G(y, v) \geqq \beta\} \text{ is equicontinuous.}$$

Since every closed equicontinuous subset is weak*-compact in Y, it is clear that (C_3) implies (C_6).

THEOREM 5.1. *The following assertions are equivalent*:

(i) $\varphi(0)$ *is finite and* φ *is bounded above on a* 0-*neighborhood*;

(ii) (C_3) *is satisfied.*

Proof. Rockafellar has shown that (ii) implies (i) [18, Thm. 18(e)]. Assume (i); then φ is continuous at zero and the level sets $\{y \mid G(y, 0) \geqq \beta\}$ are equicontinuous [18, Theorem 10]. We consider $\gamma_u(v) = \sup_{y \in Y} (G(y, v) - \langle u, y\rangle)$; we have usc $\gamma_u(0) = \varphi(u)$ whenever $u \in$ core dom φ[18, Thm. 7′], observing that we are not in the exceptional case since $\varphi(u) < +\infty$. Thus if $\mathcal{V}$ denotes a 0-neighborhood basis for a compatible topology one has usc $\gamma_u(0) = \inf_{\tilde{V} \in \mathcal{V}} \sup_{v \in \tilde{V}} \gamma_u(v)$; therefore there exists $\bar{V} \in \mathcal{V}$ such that $\sup_{v \in \bar{V}} \gamma_u(0) \leqq$ usc $\gamma_u(0) + \alpha\, (\alpha > 0)$. Introducing $g_{\bar{V}}(y) = sup_{v \in \bar{V}} G(y, v)$ which is a concave functional [18 p. 42], we have $g^*_{\bar{V}}(u) = -\sup_{y \in Y} \sup_{v \in \bar{V}} (G(y, v) - \langle u, y\rangle)$. It follows that:

$$(1) \qquad -g^*_{\bar{V}}(u) \leqq \varphi(u) + \alpha \quad \forall u \in \text{core dom } \varphi.$$

One has $\varphi(0) = \text{lsc}\, \varphi(0) = \gamma(0)$ [18, Thm. 7] which are finite. Thus there exists y_0 such that $G(y_0, 0) > -\infty$. Then since the level sets are equicontinuous, one has $\sup_{y \in Y} G(y, 0) = \gamma_0(0) = G(\bar{y}, 0) < +\infty$ [18, Thm. 9]; hence $\gamma_0(0)$ is finite. If $v \in \bar{V}$, we always have

$$(2) \qquad \gamma_u(v) \leqq \sup_{v \in \bar{V}} \gamma_u(v) = -g^*_{\bar{V}}(u).$$

Finally $\gamma_0(0) \leqq -g^*_{\bar{V}}(0) \leqq \varphi(0) + \alpha$; hence since $g^*_{\bar{V}}(0)$ is finite it is proper and usc $g_{\bar{V}} = g^{**}_{\bar{V}}$ [18, p. 43].

Assume now that there exists $\bar{v} \in V$ such that $G(\bar{y}, \bar{v}) \geqq \beta$ with $\beta < \gamma(0)$. Then $g_{\bar{V}}(\bar{y}) \geqq \beta$; therefore one has usc $g_{\bar{V}}(\bar{y}) \geqq \beta$. From (1) $-g^*_{\bar{V}}(u)$ is continuous at zero and the level sets $\{y \mid g^{**}_{\bar{V}}(y) \geqq \beta\}$ are equicontinuous. Recalling that $g^{**}_{\bar{V}} = \text{usc}\, g_{\bar{V}}$, we conclude that $\{y \mid \text{usc}\, g_{\bar{V}}(y) \geqq \beta\}$ is equicontinuous, and $\bar{y}$ belongs to this set. Q.E.D.

To obtain a corollary of this theorem, we introduce the functional $\varphi_v(u) = \inf_{x \in X} (F(x, u) - \langle x, v\rangle)$. Thus $\varphi_v(0)$ is the value of the program (π_v), and φ_0 is identical to φ.

THEOREM 5.1′. *The following assertions are equivalent*:

(i) $0 \in$ core dom φ_0.

(ii) *There exist* $v_0 \in V$, *a neighborhood* V_0 *of* v_0 *in the Mackey topology, and* $\beta \in \mathbb{R}$, $\beta < \gamma(v_0)$ *such that*:

$$\{y \mid \exists v \in V_0 \text{ satisfying } G(y, v) \geqq \beta\} \text{ is bounded.}$$

Proof. When $\gamma(0)$ is finite, taking $v_0 = 0$ our result follows from the proof of Theorem 5.1 where we have replaced "φ_0 bounded on a 0-neighborhood" by "$0 \in$ core dom φ_0" and "equicontinuous" by "bounded". Suppose now that $\forall y \in Y G(y, 0) = -\infty$. Then there exists (y_0, v_0) such that $G(y_0, v_0) > -\infty$; otherwise G should be identically equal to $-\infty$ and $F = -\infty$, which is absurd because f is proper. We can now consider $\tilde{G}(y, v) = G(y, v_0 + v)$; we have $\tilde{G}(y_0, 0) = G(y_0, v_0) > -\infty$. We observe that $\tilde{F}(x, u) = F(x, u) - \langle v_0, x\rangle$ and that $0 \in$ core dom φ_0 is equivalent to $0 \in$ core dom $\tilde{\varphi}_0$ (since $\tilde{\varphi}_0 = \varphi_{v_0}$ and dom φ_0 = dom φ_{v_0}). It is possible now to apply the previous proof to $\tilde{G}$ which completes the proof.

(c) In normed spaces we can replace neighborhoods by balls.

THEOREM 5.2. *Assume that either X is a Banach space or V is normed in a topology compatible with the pairing, then the following assertions are equivalent*:

(i) *$\varphi(0)$ is finite and φ is bounded above on a 0-neighborhood.*

(ii) *There exist $k>0$ and $\beta<\gamma(0)$ such that the set $\{y \mid \exists v$ satisfying $\|v\| \leqq k$ and $G(y, v) \geqq \beta\}$ is equicontinuous.*

Proof. When V is normed in a compatible topology, the result follows directly from Theorem 5.1. Assume now that X is a Banach space. To prove that (i) implies (ii) we observe that in the proof of Theorem 5.1 we can choose a basis $\mathscr{V}$ of 0-neighborhood for the weak*-topology. Thus $\bar V$ being a weak*-neighborhood it contains a ball of radius k:

$$\Big(\{v \mid \|v\| \leqq k\} \subset \bar V = \{v \mid |\langle v, x_i\rangle| \leqq \varepsilon, i = 1, \cdots, n\},\ k \leqq \frac{\varepsilon}{\sup_i \|x_i\|}\Big).$$

Replacing $\bar V$ by the ball $\{v \mid \|v\| \leqq k\}$ we can continue as previously the proof.

To see that the converse implication holds we have to show that Rockafellar's proof of Theorem 18(e), [18], is still valid when X is a Banach space and N is replaced by $\{v \mid \|v\| \leqq k\}$. It suffices to prove that $\lim_{v\to 0}\sup \gamma(v) \leqq \sup_{\|v\|\leqq k} \gamma(v)$ (i.e. that (7.14) still holds in Rockafellar's proof). Assume that $a = \sup_{\|v\|\leqq k}\gamma(v) < +\infty$; otherwise it is obvious. If $\lim_{v\to 0}\sup \gamma(v) > a$, then there exists a generalized sequence v_α which weakly converges to 0 and satisfies $\gamma(v_\alpha) > a$. This sequence is weak*-compact and its convex hull also, because V is quasi-complete [1 Cor. 2, Thm. 1, Chap. IV, § 2, N^{0}2]. Thus this sequence is equicontinuous and strongly bounded [1, Chap. III, § 3, Prop. 7]. Moreover $\gamma(0)$ is finite. As in the proof of Theorem 4.1 we consider the sequence $v'_\alpha = \lambda v_\alpha$ and we obtain an absurdity, which completes the proof.

It should be possible to give a similar result for $0 \in \operatorname{core} \operatorname{dom} \varphi_0$ using Theorem 5.1′. To give a more convenient form to the condition (ii) of Theorem 5.2 when X or V are normed, we need the following proposition:

PROPOSITION 5.1. *The following assertions are equivalent when V is normed* (*not necessarily for a compatible topology*):

(i) *$\forall \beta \in \mathbb{R}$, $\forall k>0$ the set $\{y \mid \exists v$ satisfying $\|v\| \leqq k$ and $G(y, v) \geqq \beta\}$ is equicontinuous* (*resp. weak*-compact, bounded*).

(ii) *$\exists \beta_0 \in \mathbb{R}$, $\exists k_0 > 0$ such that there exists (y_0, v_0) satisfying $\|v_0\| < k_0$, $G(y_0, v_0) > \beta_0$ and the set $\{y \mid \exists v$ satisfying $\|v\| \leqq k_0$ and $(G(y, v) \geqq \beta_0\}$ is equicontinuous* (*resp. weak*-compact, bounded*).

Proof. It is clear that (i)$\Rightarrow$(ii) when G is not identically equal to $-\infty$. Assuming (ii), there exist β_0 and k_0 such that $B = \{y \mid \exists v$ satisfying $\|v\| \leqq k_0$ and $G(y, v) \geqq \beta_0\}$ is equicontinuous. Let $\|\bar v\| \leqq k$ such that there exists $\bar y$ with $G(\bar y, \bar v) \geqq \beta$. We set $G(y_0, v_0) = \beta_0 + \varepsilon$, $(\varepsilon > 0)$, and $\|v_0\| = k_0 - \varepsilon'(\varepsilon' > 0)$. One has $G(t\bar y + (1-t)y_0,\ t\bar v + (1-t)v_0) \geqq tG(\bar y, \bar v) + (1-t)G(y_0, v_0)$. Taking $t_0 \leqq \min(1, \varepsilon[\max(0, \beta_0 - \beta) + \varepsilon]^{-1})$ we obtain $G(t_0\bar y + (1-t_0)y_0,\ t_0\bar v + (1-t_0)v_0) \geqq \beta_0$. Then taking $t_1 = \min(1, t_0, \varepsilon'[\max(k - k_0, 0) + \varepsilon']^{-1})$ we have $\|t_1\bar v + (1-t_1)v_0\| \leqq k_0$. Thus $t_1\bar y + (1-t_1)y_0 \in B$ which is equicontinuous. It follows that whenever y belongs to $\{y \mid \exists v$ satisfying $\|v\| \leqq k$ and $G(y, v) \geqq \beta\}$ it also belongs to $y_0 + (B - y_0)t_1^{-1}$. Since "equicontinuous", "weak*-compact" and "bounded" sets are stable by translation and multiplication by a scalar the proof is completed.

COROLLARY 5.1. *With the assumptions of Theorem 5.2 the following assertions are equivalent*:

(i) *There exists $v_0 \in V$ such that $\varphi_{v_0}(0)$ is finite and bounded above on a 0-neighborhood.*

(ii) $\forall\beta\in\mathbb{R}, \forall k>0$ *the set* $\{y \mid \exists v$ *satisfying* $\|v\|\leqq k$ *and* $G(y,v)\geqq\beta\}$ *is equicontinuous.*

Proof. Applying Theorem 5.2 to $\tilde{G}$ we obtain (i)$\Rightarrow$(ii) by Proposition 5.1. The converse is true at any point v_0 where $\gamma(v_0)>-\infty$. Q.E.D.

We introduce now the following condition:

(C_8) $\forall\beta\in\mathbb{R}, \forall k>0$ *the set* $\{y \mid \exists v$ *satisfying* $\|v\|\leqq k$ *and* $G(y, v)\geqq\beta\}$ *is bounded.*

And we get:

COROLLARY 5.1′. *With the assumptions of Theorem* 5.2, *the following assertions are equivalent*:

(i) $0\in$ core dom φ_0.

(*ii*) (C_8) *is satisfied.*

Proof. The proof follows from Proposition 5.1 and Theorem 5.1′.

PROPOSITION 5.2. *With the assumptions of Theorem* 5.2 *one has*: (C_3)$\Rightarrow$(CS_4)$\Rightarrow$(CS_5)$\Rightarrow$(C_8)$\Leftrightarrow$(C_2). *Moreover if* $\gamma(0)$ *is finite* (CS_5)$\Leftrightarrow$(CS_4); *finally when* $\gamma(0)$ *is finite and U is a Banach space, these five conditions are equivalent.*

Proof. The first assertion is obvious because equicontinuous implies weak*-compact which implies bounded, and $0\in$ core dom φ_0 is the abstract version of (C_2).

When $\gamma(0)$ is finite "(CS_5) is equivalent to (CS_4)" results from Proposition 5.1. Finally in the dual of a Banach space bounded and weak*-compact sets are equicontinuous. Q.E.D.

Finally, while neither of the three conditions (C_1), (C_2), (C_3) apply to the programs described in § 2, let us give an example of this type where (C_2) holds. At the same time, this will prove that our conditions (here (C_5)) may be satisfied in cases where the strong versions (here (CS_5)) are not satisfied.

We consider the following program ($\bar{\pi}$):

$$\begin{aligned}
\text{Minimize}\quad & \int_0^1 -(t^2x_1(t)-x_2(t))\,dt\\
\text{subject to}\quad & tx_1(t)-x_2(t)\leqq 1,\\
& g(t)x_2(t)\leqq 0,\\
& x_1(t)\geqq 0, \qquad x_1\in L^1[0,1],\\
& x_2(t)\geqq 0, \qquad x_2\in L^1[0,1].
\end{aligned}$$

$L^1[0, 1]$ is defined by the Lebesgue measure on [0, 1], g is the function defined by:

$$g(t)=\begin{cases}0 & \text{if } t\in[0,\frac{1}{2}],\\ 1 & \text{otherwise}\end{cases}$$

and the constraints must be satisfied almost everywhere.

Let us prove that ($\bar{\pi}$) satisfies (C_7) which implies that (C_5) is satisfied. Applied to ($\bar{\pi}$) it is equivalent to the following: there exist positive real numbers α and β, such that: if one can find y_1, y_2, u_1 and u_2 in $L^\infty[0, 1]$ and a real number θ satisfying:

(1) $\theta\geqq\int_0^1 u_1(t)\,dt$,

(2) $tu_1(t)\geqq y_1(t)$,

(3) $-u_1(t)+u_2(t)g(t)\geqq y_2(t)$,

(4) $u_1(t)\geqq 0$,

(5) $u_2(t)\geqq 0$

then there exist $\bar{u}_1$ and $\bar{u}_2$ in $L^\infty[0, 1]$ satisfying together with y_1, y_2 and θ the constraints (1)–(5) and:

(6) $\|(\bar{u}_1^2+\bar{u}_1^2)^{1/2}\|_\infty \leqq \alpha \max(\|(y_1^2+y_2^2)^{1/2}\|_\infty, |\theta|)+\beta$.
If y_1, y_2, u_1, u_2 and θ satisfy the constraints (1)–(5)

$$0 \leqq u_1(t) \leqq -y_2(t) \quad \text{on } [0, \tfrac{1}{2}].$$

Thus, we can set

$$\bar{u}_1(t)=\begin{cases} u_1(t) & \text{on } [0, \frac{1}{2}], \\ y_1^+(t)t^{-1} & \text{on }]\frac{1}{2}, 1] \end{cases}$$

(where $y_1^+(t)=\max(0, y_1(t))$), and

$$\bar{u}_2(t)=\begin{cases} 0 & \text{on } [0, \frac{1}{2}], \\ (y_2(t)+y_1^+(t)t^{-1})^+ & \text{on}]\frac{1}{2}, 1]. \end{cases}$$

Consequently (6) is satisfied with $\alpha=4$ and $\beta=0$, since almost everywhere:

$$\bar{u}_1(t)^2+\bar{u}_2(t)^2 \leqq 8(|y_1(t)|+|y_2(t)|)^2 \leqq 16(y_1(t)^2+y_2(t)^2).$$

Let us notice that it is obvious that the vectors u satisfying the constraints (1)–(5) do not define a bounded set. Thus (CS$_5$) fails to be satisfied and consequently (C$_2$) and (C$_3$) do not hold.

However, tightening the inequalities we have found a choice of u which is bounded.

This example is a special case of a wide class of programs which satisfy (C$_7$), as will be seen in the next section.

6. Continuous linear programs. Let us consider the following continuous program (π):

$$\text{Minimize} \int_0^T f(t)x(t)\,dt$$

$$\text{subject to} \quad B(t)x(t)-\int_0^t K(t, s)x(s)\,ds \leqq a(t), \quad x(t) \geqq 0.$$

For each $t \in [0, T]$, $B(t)$ is a $m \times n$ matrix, $f(t)$ an n-vector, $a(t)$ an m-vector and $K(t, s)$ an $m \times n$ matrix which is zero if $s > t$. The functions B, K and f are all bounded and Lebesgue measurable, a is Lebesgue measurable and $\int_0^T |a(t)|\,dt$ is finite (i.e. $a \in L_m^1[0, T]$). T is finite, the constraints must be satisfied almost everywhere and the solution x is required to be in $L_n^1[0, T]$.

Now, let us consider the following assumptions on (π)

(A$_1$) (i) $\forall t\ {}^tB(t)u \geqq 0$ and $u \geqq 0$ imply $\langle u, a(t)\rangle \geqq 0$;

(ii) $\forall t\ {}^tB(t)u \geqq 0$ and $u \geqq 0$ imply ${}^tK(t, s)u \geqq 0\ \forall_s \leqq t$.

Let us denote by $H[B(t), d]$ the convex hull of the extreme points of $\{v \mid v \in \mathbb{R}^m v \geqq 0\ {}^tB(t)v \geqq d\}$.

(A$_2$) There exists $p \geqq 0$ such that:

$$\forall t \in [0, T], \forall d \in \mathbb{R}^n,\ v \in H[B(t), d] \quad \text{implies} \quad \|v\| \leqq p\|d\|.$$

($\|\cdot\|$ is the Euclidean norm.)

PROPOSITION 6.1. *Under Assumptions* (A$_1$) *and* (A$_2$) *the program* (π) *satisfies* (C$_7$). *Thus* (π) *has a Kuhn–Tucker vector if* $\inf(\pi) < +\infty$.

Proof. We shall prove the following: there exists a positive real number α such that if (y, v, θ) satisfy

(1) $y \in L_n^\infty[0, T]$, $y(t) \geqq 0$ a.e.,

(2) $v \in L_m^\infty[0, T]$, $v(t) \geqq 0$ a.e.,

(3) ${}^tB(t)v(t) - \int_t^T {}^tK(s, t)v(s)\,ds \geqq y(t)$ a.e.,

(4) $\theta \geqq \int_0^T a(t)v(t)\,dt$,

then there exists $\bar{v} \in L_m^\infty$ satisfying together with y and θ (2), (3), (4) and

(5) $\|\bar{v}\|_\infty \leqq \alpha \|y\|_\infty$.

Let us remark that, if v satisfies (2) and (3) one can find $\bar{v}$ and $\tilde{v}$ such that (if one sets $d(t) = y(t) + \int_t^T {}^tK(s, t)v(s)\,ds$), $v = \bar{v} + \tilde{v}$ with $\bar{v}(t) \in H[B(t), d(t)]$ and $\tilde{v}(t) \in \{w \mid {}^tB(t)w \geqq 0, w \geqq 0\} \forall t \in [0, T]$, Grinold [6, Cor. 9, p. 39]. It follows from (A_1) that $\bar{v}$ also satisfies (2), (3) and (4). Moreover Grinold shows that $\|\bar{v}\|_\infty \leqq \alpha \|y\|_\infty$ which completes the proof.

Remark 6.1. The assumption (A_1) is the algebraic assumption II of Grinold [6], while (A_2) is his boundedness condition (ii). So, we need only the "dual assumptions" of Grinold [6] and [7], to prove that (π) has a Kuhn–Tucker vector.

Programs defined in § 2 obviously satisfy (A_1). Grinold has shown that they satisfy (A_2), [7, Prop. A2, p. 96]. Thus, such programs have a Kuhn–Tucker vector.

Moreover, it is easy to check that the program $(\bar{\pi})$ of § 5, which satisfies (A_1) and (A_2) does not satisfy the boundedness condition (i) of Grinold [6]. Thus, (A_1) and (A_2) are strictly weaker than Grinold's conditions.

7. Tables of results.

In these tables we use the following abbreviations: K.T. for Kuhn–Tucker vector, and * means: "when $\gamma(0)$ is finite".

7.1. In locally convex vector spaces

$$\begin{tikzcd}
 & (0 \in \text{core dom } \varphi_0) = (C_2) & \\
(C_3) \arrow[ur, Rightarrow] \arrow[r, Rightarrow] & (C_6) \arrow[r, Rightarrow] & \textit{K.T. for } (\pi)
\end{tikzcd}$$

7.2. When either X is a Banach space or V is normed in a compatible topology

$$\begin{tikzcd}
 & & (C_8) \arrow[r, Leftrightarrow] & (C_2) & & & \\
 & (CS_5) \arrow[ur, Rightarrow] \arrow[rrr, Rightarrow] \arrow[d, Rightarrow, shift left, "*"] & & & (C_5) \arrow[rr, Rightarrow] \arrow[dr, Rightarrow, "*"] & & \forall v \in V, \textit{K.T. for } (\pi_v) \\
(C_3) \arrow[ur, Rightarrow] \arrow[r, Rightarrow] \arrow[drrrr, Rightarrow] & (CS_4) \arrow[u, Rightarrow, shift left] \arrow[rrrr, Rightarrow] & & & & (C_4) \arrow[r, Rightarrow] & \textit{K.T. for } (\pi) \\
 & & & & (C_6) \arrow[ur, Rightarrow, "*"] \arrow[r, Rightarrow] & \textit{K.T. for } (\pi) &
\end{tikzcd}$$

7.3. When either X is a Banach space or V is normed in a compatible topology and U is a Banach space.

$$\begin{tikzcd}
 & (C_1) \arrow[rr, Rightarrow, "\textit{X and U finite}", "\textit{dimensional}"'] & & (C_7) \arrow[dr, Rightarrow] & & & \\
 & (CS_5) \arrow[r, Leftrightarrow] \arrow[d, Rightarrow, "*"] & (C_8) \arrow[r, Leftrightarrow] & (C_2) \arrow[r, Rightarrow] & (C_5) \arrow[rr, Rightarrow] \arrow[dr, Rightarrow, "*"] & & \forall v \in V, \textit{K.T. for } (\pi_v) \\
(C_3) \arrow[ur, Rightarrow] \arrow[r, Leftrightarrow] \arrow[drrrr, Rightarrow] & (CS_4) \arrow[rrrr, Rightarrow] & & & & (C_4) \arrow[r, Rightarrow] & \textit{K.T. for } (\pi) \\
 & & & & (C_6) \arrow[ur, Rightarrow, "*"] \arrow[r, Rightarrow] & \textit{K.T. for } (\pi) &
\end{tikzcd}$$

Acknowledgments. We wish to thank Professor R. T. Rockafellar who has made numerous helpful suggestions (not to mention improvements of our English) on the earlier versions of this paper. In particular, thanks to Rockafellar's stimulating comments we have introduced local conditions. We are also indebted to J. Gwinner for pointing out reference [5].

This work was partly done while the first author was with the CORE.

REFERENCES

[1] N. BOURBAKI, *Espaces vectoriels topologiques*, Chaps. III, IV et V, Hermann, Paris, 1967.

[2] U. DIETER, *Optimierungsaufgaben in topologischen Vektorräumen I: Dualitäts-theorie*, Z. Wahrscheinlichkeitstheorie Verw. Gebiete, 5 (1966), pp. 89–117.

[3] J. J. M. EVERS, *Linear Programming Over an Infinite Horizon*, Tilburg University Press, the Netherlands, 1973.

[4] ———, *Optimization in normed vector spaces with applications to optimal economic growth theory*, Tilburg Institute of Economics Research Memorandum 50, 1974.

[5] M. J. EISNER AND P. OLSEN, *Duality for stochastic programming interpreted as L.P. in L_P-space*, SIAM J. Appl. Math., 28 (1975), pp. 779–792.

[6] R. C. GRINOLD, *Continuous programming. Part one: linear objectives*, J. Math. Anal. Appl., 28 (1969), pp. 32–51.

[7] ———, *Symmetric duality for continuous linear programs*, SIAM J. Appl. Math., 19 (1970), pp. 84–97.

[8] J. GWINNER, *Closed images of convex multivalued mappings in linear topological spaces with applications*, J. Math. Appl., 60 (1977), pp. 75–86.

[9] W. KRABS, *Optimierung und Approximation*, Teubner, Stuttgart, 1975.

[10] K. S. KRETSCHMER, *Programmes in paired spaces*, Canad. J. Math., 13 (1961), pp. 221–238.

[11] P. LEVINE AND J. CH. POMEROL, *C-closed mappings and Kuhn–Tucker vectors in convex programming*, CORE D.P. 7620, Université Catholique de Louvain, Louvain, 1976.

[12] J. CH. MAILLARD, *Approximation des programmes convexes et dualité*, 3rd cycle thesis, Université P. et M. Curie, Paris, 1975.

[13] S. M. ROBINSON, *Regularity and stability for convex multivalued functions*, Math. of Operations Res. 1 (1976), pp. 130–143.

[14] R. T. ROCKAFELLAR, *Duality and stability in extremum problems involving convex functions*, Pacific J. Math., 21(1967), pp. 167–187.

[15] ———, *Duality in nonlinear programming*, Mathematics of the Decision Sciences Part 1, Lectures in Applied Mathematics 11, American Mathematical Society, 1968, pp. 401–422.

[16] ———, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[17] ———, *Saddle-points and convex analysis, in differential games and related topics*, H. W. Kuhn and G. P. Szegö, eds., North-Holland, Amsterdam, 1971.

[18] ———, *Conjugate duality and optimization*, CBMS/NSF regional conference series in applied mathematics 16, Society for Industrial and Applied Mathematics, Philadelphia, 1974.

[19] F. TRÖLTZSCH, *Existenz-und Dualitätsaussagen für lineare Optimierungsprobleme in Reflexiven Banach-Räumen*, Math. Operationsforsch. Statist., 6 (1975), pp. 901–912.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 6, November 1979

0363-0129/79/1706-0003$01.00/0

THE PRINCESS AND MONSTER DIFFERENTIAL GAME*

CARL H. FITZGERALD†

Abstract. In this game of R. Isaacs, the players are a hider and a searcher. The payoff is the time until capture. The hider is assumed to be mobile within a bounded set $\mathcal{D}$. The searcher has an arbitrarily small detection radius. There is incomplete information in that neither player knows the present or past location of the other. For various sets $\mathcal{D}$, the value of the game is demonstrated by the presentation of ε-optimal strategies. Specifically, the game is solved in case $\mathcal{D}$ is convex or is the finite union of convex sets in $\mathbb{R}^n$ with $n \geqq 2$. Also the game is discussed in the case where $\mathcal{D}$ is a network. The results settle two conjectures of S. Gal.

1. Introduction. In this paper we study a search game with a mobile hider and incomplete information. The original problem was posed by Rufus Isaacs [3] with the name, "The Princess and the Monster." It can be formulated as follows: The princess and the monster are in a totally dark room (of any shape). Each knows his own location and previous path. The monster is searching for the princess. His maximum speed is one and is known to both. Capture takes place when the monster is within a known detection distance r of the princess. The princess' maximum speed is s where s is positive and known to both. The payoff is the expected time to capture.

The princess and the monster are not given the other's location. The possibility that information might be inferred makes the analysis more difficult.

To motivate the problem, the reader might substitute "submarine" for princess and "destroyer" for monster. Also the reader could compare these results with the considerable literature on search problems with immobile hider or hider with a random motion of given distribution. The princess and the monster problem is the worst possible case of a hider with some type of known random motion.

Various discrete versions and special network problems have been studied (cf. [1], [3]). It is only recently that something close to a general solution has been found [1]. When the "dark room" is a rectangular set in $\mathbb{R}^n$ or a combination of two rectangles, S. Gal has solved the problem for arbitrarily small detection radius r. Also, he gave a strategy for the princess in any bounded convex set in $\mathbb{R}^n$ and analyzed it. Gal made the following two conjectures in that paper:

Conjecture A. Gal's search strategy for the monster can be extended to convex sets, and the expected time until capture is the length of time for the monster to cover an area (or volume) equal to the area (or volume) of the set $\mathcal{D}$. Also, it was suggested [2] that it might work for finite unions of convex sets.

Conjecture B. Suppose the "dark room" is a network and that capture takes place if the princess and the monster are at the same point. The expected time until capture is less than or equal to twice the time necessary for the monster to travel a distance equal to the full length of the network.

These two conjectures appeared to be closely related. In Gal's paper, the results of a careful analysis of certain networks were used to solve the problem in the case where $\mathcal{D}$ is a rectangle. Also, the conjectures appeared to be motivated by similar intuitive arguments.

The analyses given in this paper complement each other. Intuition is shown to be correct in one case, and wrong in the other. Section 4 is a discussion of a modified

* Received by the editors May 31, 1978 and in revised form December 18, 1978.

† Department of Mathematics, University of California, San Diego, La Jolla, California 92093. This research was supported in part by the National Science Foundation under Grant NSF-C MCS77-03497.

princess and monster game. The new problem suggests the pertinent difference between the domain in the first conjecture and the network in the second.

2. Searching in domains in $\mathbb{R}^n$. Consider a closed, bounded, convex set $\mathcal{D}$ in $\mathbb{R}^2$. Suppose the interior of $\mathcal{D}$ has positive area. The strategy Gal gave for the princess is depicted by the dotted line in Fig. 1. For any $\varepsilon > 0$, the strategy can be stated as follows: a) The princess should start at a point of $\mathcal{D}$ chosen at random with uniform probability over $\mathcal{D}$. b) She should wait there the length of time it would take the monster to search ε of the area of $\mathcal{D}$, then pick a point of $\mathcal{D}$ at random with uniform probability over $\mathcal{D}$, and move along a straight line to that point with maximum speed s. c) Step b should be repeated indefinitely.

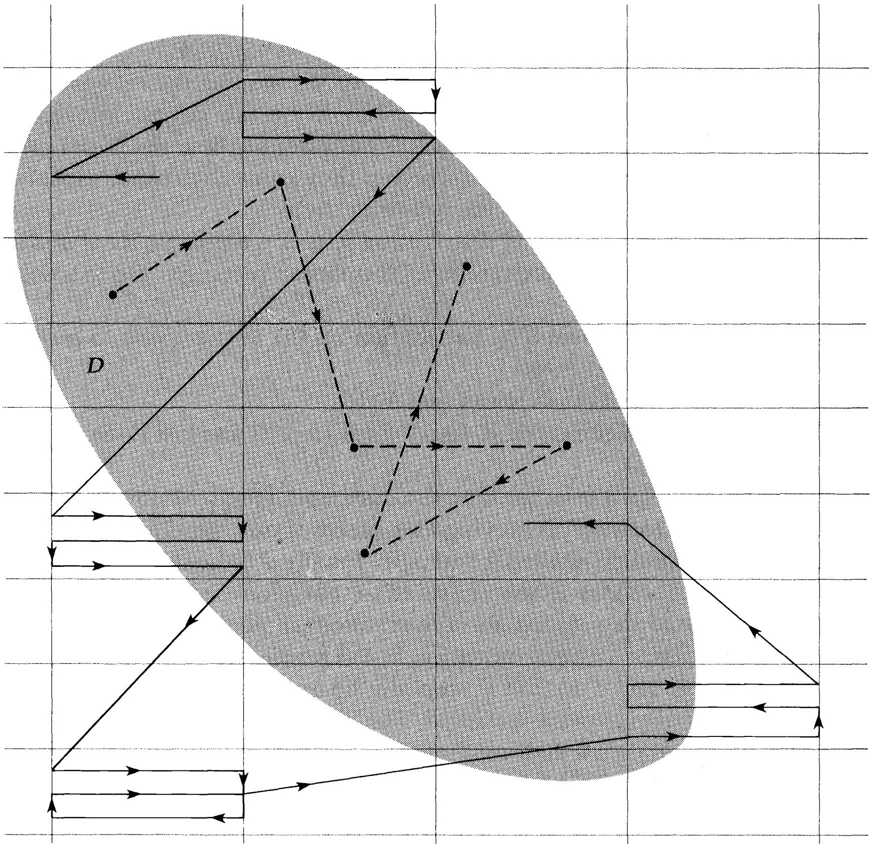

FIG. 1

For sufficiently small detection radius r, this strategy guarantees the princess an expected time to capture which is greater than the time necessary for the monster to cover an area of the size of $\mathcal{D}$ times $(1-\varepsilon)$. Since the monster has a speed of one and searches area at the rate $2r$, the expected time until capture is greater than $(1-\varepsilon)$ (area of $\mathcal{D}) \div (2r)$.

For the case in which $\mathcal{D}$ is a rectangle, the analysis of the game was completed by giving an ε-optimal strategy for the monster. Conjecture A is just that the strategy can be extended to be ε-optimal for more general domains and sufficiently small detection radius. A slightly different strategy is presented here. An analysis shows it is ε-optimal when $\mathcal{D}$ is convex and the detection radius is sufficiently small. The strategy requires the monster to move outside of $\mathscr{D}$. Subsequently it is shown that the strategy can be modified to avoid such excursions. Generalization to finite unions of convex sets, to higher dimensions, and varying detection radius are discussed. We now present the strategy for the case in which $\mathscr{D}$ is convex in $\mathbb{R}^2$. The strategy is depicted by the solid line in Fig. 1.

For the moment, we shall assume that the monster can move out of $\mathscr{D}$. Let A denote the area of $\mathscr{D}$, and let L denote the diameter of $\mathscr{D}$. Consider tilings of the plane with congruent, horizontal rectangles. For ε given such that $1>\varepsilon>0$ and $L>\varepsilon$, we require the lengths of the rectangles to be $d=\varepsilon A/(8^2L)$. The heights of the rectangles are chosen to be $h=\varepsilon d/8$. We also fix the tiling and let $\mathscr{R}$ be the collection of rectangles containing points of $\mathscr{D}$ or rectangles which are adjacent to such rectangles. Every point in a rectangle of $\mathscr{R}$ is within $3d$ of $\mathscr{D}$. Using the estimate that $A \leqq L^2$, one can easily show that the total area of the rectangles of $\mathscr{R}$ is less than $(1+\varepsilon/8)A$.

A strategy for the monster is now presented. We assume the recognition radius r satisfies the inequality $0<r\leqq \varepsilon^5A^2/(8^8L^3)$. Let K be the largest integer such that $2rK<(\varepsilon/8)h$, thus, $K\geqq[(8^4L^2)/(3\varepsilon^2A)]$. It is not essential how the initial locations of the princess and monster are chosen or whether they know each other's locations. Our strategy for the monster consists of the following steps:

(i) Use a uniform probability distribution to pick a rectangle from $\mathscr{R}$ at random. With equal probability pick a vertical side of the chosen rectangle. Go to a point of that side.

(ii) Use a uniform probability distribution on the vertical side to pick a level at random. Go straight to that level.

(iii) Go directly across the chosen rectangle.

(iv) If step (ii) has been done K times since step (i) was last done, go to step (i). Otherwise go to step (ii).

Let Δ equal $2L$ which is larger than the maximum length of time that the monster might need to complete step (i) moving at a speed of one. Similarly let δ equal $\varepsilon d/8$ which is the maximum time needed for step (ii). Finally d is the time to do step (iii). The time spent in doing each step is now specified so that at any instant the monster will be doing the same type of step regardless of the random choices. This specification is done to simplify later analysis. To keep moving as fast as possible, he must spend exactly Δ on step (i), δ on step (ii), and d on (iii). In step (iii) the monster will always be moving with maximum speed of one across a rectangle. The length of time for one full cycle of the strategy is $\Delta+K(\delta+d)$. Call that number T.

In each cycle of length T, the time spent on step (iii) is more than $(1-\varepsilon/4)T$. Note that $K\delta=(\varepsilon/8)Kd$. Also $Kd\geqq(8^4L^2/(4\varepsilon^2A))\varepsilon A/(8^2L)=(8/\varepsilon)(2L)$, and thus $(\varepsilon/8)Kd\geqq 2L=\Delta$. Hence $Kd\geqq K\delta+Kd-(\varepsilon/8)Kd\geqq\Delta+K\delta+Kd-(\varepsilon/4)Kd\geqq T-(\varepsilon/4)T$.

We now examine the area of the rectangle expected to be covered during this time spent on step (iii) in each cycle of length T. There are at most K strips of width $2r$ and length d. The total of the widths is at most $K\,2r\leqq(\varepsilon/8)h$ by the definition of K. Since the levels are chosen at random there may be some overlapping. But, since at most $\varepsilon/8$ of the height of the rectangle is covered, the expected area covered is $\geqq(1-\varepsilon/15)K\,2r\,d$. This inequality also allows for the possibility that some of the area searched is outside the rectangle and thus is not to be counted.

To demonstrate the effectiveness of this strategy, we will first analyze the case in which the princess is stationary. Other strategies of the princess will then be compared with the stationary strategy.

Suppose the princess is farther from the boundary of a rectangle than the detection radius. To find the princess during step (iii), the monster must have chosen the correct rectangle and a height within r of the princess' height. The chance of doing this within some full cycle of the strategy is $p =$ [area expected to be covered in step (iii) in one full cycle] $\div$ [the total area of the rectangles]. An upper bound on the expected time until capture is now determined. Expected time is less than

$$\begin{aligned}\sum_{k=1}^{\infty} kTp(1-p)^{k-1} &= T\frac{p}{1-p}\sum_{k=1}^{\infty} k(1-p)^k\\ &= T\frac{p}{1-p}\left[\sum_{k=1}^{\infty}(1-p)^k\right]\left[\sum_{k=0}^{\infty}(1-p)^k\right]\\ &= T\frac{p}{1-p}\,\frac{1-p}{1-(1-p)}\,\frac{1}{1-(1-p)}=\frac{T}{p}.\end{aligned}$$

Since $Kd \geqq (1-\varepsilon/4)T$, $T \leqq Kd \div (1-\varepsilon/4)$. Hence the expected time until capture is less than $[Kd \div (1-\varepsilon/4)]$ [the total area of the rectangles] $\div$ [the expected area covered in step (iii) in one full cycle]

$$\begin{aligned}&< (1-\varepsilon/4)^{-1}(1-\varepsilon/15)^{-1}[Kd \div (K2rd)] \quad \text{[the total area of the rectangles]},\\ &< (1-\varepsilon/3)^{-1}(1+\varepsilon/8)A \div 2r,\\ &\leqq (1+\varepsilon/2)A \div 2r,\\ &= (1+\varepsilon/2) \quad \text{[the time for monster to cover an area equal that of } \mathscr{D}].\end{aligned}$$

If the princess is close to a boundary of a rectangle, the possibility of her being captured while the monster is in an adjacent rectangle must be taken into account. Clearly, the final result is the same.

Can the princess find a strategy against the monster that will increase the expected time until capture to at least $(1+\varepsilon)$ [the time for the monster to cover an area equal to that of $\mathscr{D}$]? Suppose such a strategy exists. It can depend only on information which the princess obtains by *not* being captured. Hence the intended path for several full cycles can be chosen before starting on the assumption that there is no capture.

Consider the strategy during the first period of length T, that is, the first full cycle of the monster's search; then from T to $2T$, etc. During at least some of those intervals, the princess' strategy must be sufficiently effective that if she is free at the beginning, her chance of capture during the interval is less than $(1-\varepsilon/2)p$, where p is the probability of capture when she is stationary.

Suppose $[0, T]$ is such an interval of time with an effective defense. If the princess has not been captured up to time t for $0 \leqq t \leqq T$, what is the probability that the monster and princess are in the same rectangle and the monster chose a particular initial direction of search of that rectangle?

P(monster chose a given rectangle and initial direction and the princess is caught during the interval 0 to t)
$+P$(monster chose a given rectangle and initial direction and the princess is not caught during the interval 0 to t)
$= P$(monster chose a given rectangle and direction)
$=$ [2 times the number of rectangles]$^{-1}$.

Since P (capture in interval 0 to T) $\leqq (1-\varepsilon/2)(\varepsilon/8)$ [area of a rectangle] $\div$ [total area of rectangles], clearly P(monster chose a given rectangle and initial direction and the princess is not caught during time 0 to t) $> (1-\varepsilon/4)$ [2 times the number of rectangles]$^{-1}$. Thus, if the princess is relatively safe from 0 to T, she has essentially no information concerning which rectangle the monster is in.

Consider the subcycles involving one step of type (ii) and one step of type (iii). Each subcycle has a period of $\delta + d$. Trace the intended path of the princess to be followed as long as no capture takes place.

Suppose she starts and finishes in the same vertical column of rectangles. By the intermediate value theorem, there exists a time τ_R at which she will have the same horizontal coordinate as the monster if he happens to pick a rectangle in that column and is going to the right. Assume that at such a moment $t = \tau_R$ she is farther from the boundary of the rectangle which she is in than the detection radius r. What is the probability of capture? It has been shown that the probability of the monster picking the particular rectangle which the princess is in at $t = \tau_R$ and picking the appropriate initial direction is greater than $(1-\varepsilon/4) \div$[two times the number of rectangles]. Suppose the monster picks that rectangle and initial direction and that the princess has not been caught before the particular subcycle. The probability that the monster will pick a level within r of the princess is $2r \div h$. There is the same amount of risk of capture with the monster moving to the left, there being a time τ_L analogous to τ_R.

Hence, if the monster has not caught the princess before the subcycle, the chance of capture during the subcycle is at least $2\{(1-\varepsilon/4) \div$[two times the number of rectangles]$\}(2r \div h) = (1-\varepsilon/4)(2r/h) \div$[the number of rectangles]. The assumption that the princess is away from the boundary is not essential. If she is close at the time $t = \tau$, the possibility of capture with the monster in an adjacent rectangle must be taken into account. Notice that the princess can avoid being caught at $t = \tau_R$ or $t = \tau_L$, but the analysis has shown that in doing so she must run as much risk during the subcycle as if she were stationary times $(1-\varepsilon/4) = (1-\varepsilon/4)p$.

A similar analysis can be carried out if the princess ends in a different column of rectangles from the one which she starts in. If it is to the left, there are at least two times, τ_{R_1} and τ_{R_2}, when she will have the same horizontal component as the monster if he chooses a rectangle in the same column which she is in at $t = \tau_{R_1}$ or $t = \tau_{R_2}$ and the appropriate initial direction. And similarly if she ends in a column to the right. Thus, the total risk will be as much as calculated in the preceding, that is, at least as much as if she were stationary times $(1-\varepsilon/4)$.

Since this estimate holds for each subcycle, the chance of capture during a full cycle of the monster's search is at least $(1-\varepsilon/4)p$. This contradicts the hypothesis that her strategy was particularly safe during the period considered. Thus the monster's strategy has been shown to be ε-optimal for sufficiently small recognition radius.

If one objects to the monster being allowed to move outside of the convex set $\mathscr{D}$, the following modification is easily made. It is depicted in Fig. 2. Consider a vertical interval I of length $2r$ centered on the monster. If in step (iii) the monster is to start or end out of the set $\mathscr{D}$, then the monster should modify steps (ii) and (iii) as follows. The monster should determine which end of I is in $\mathscr{D}$ the longest and go to the point at which that end is first in $\mathscr{D}$ and follow it as long as it is in $\mathscr{D}$ during the original step (iii). For rectangles along the upper boundary of $\mathscr{D}$, the trajectory is shifted down; for rectangles along the lower boundary of $\mathscr{D}$ it is shifted up. When the instruction is ambiguous, the monster should not shift the trajectory up or down, but should follow the original trajectory as long as it is in $\mathscr{D}$. If step (i) would require the monster to exit the set, he should consider the first subcycle for which the preceding modification allows the monster to move within $\mathscr{D}$. If every subcycle stays out of $\mathscr{D}$, then the next full cycle should be considered. The convexity of $\mathscr{D}$ shows that the modification can be carried out with the new trajectory consisting of straight line intervals and the appropriate timing can be maintained. Note that it suffices to count only captures taking place during step (iii) with the monster and princess having the same horizontal component. Hence it is clear that the effectiveness of the search has not been decreased.

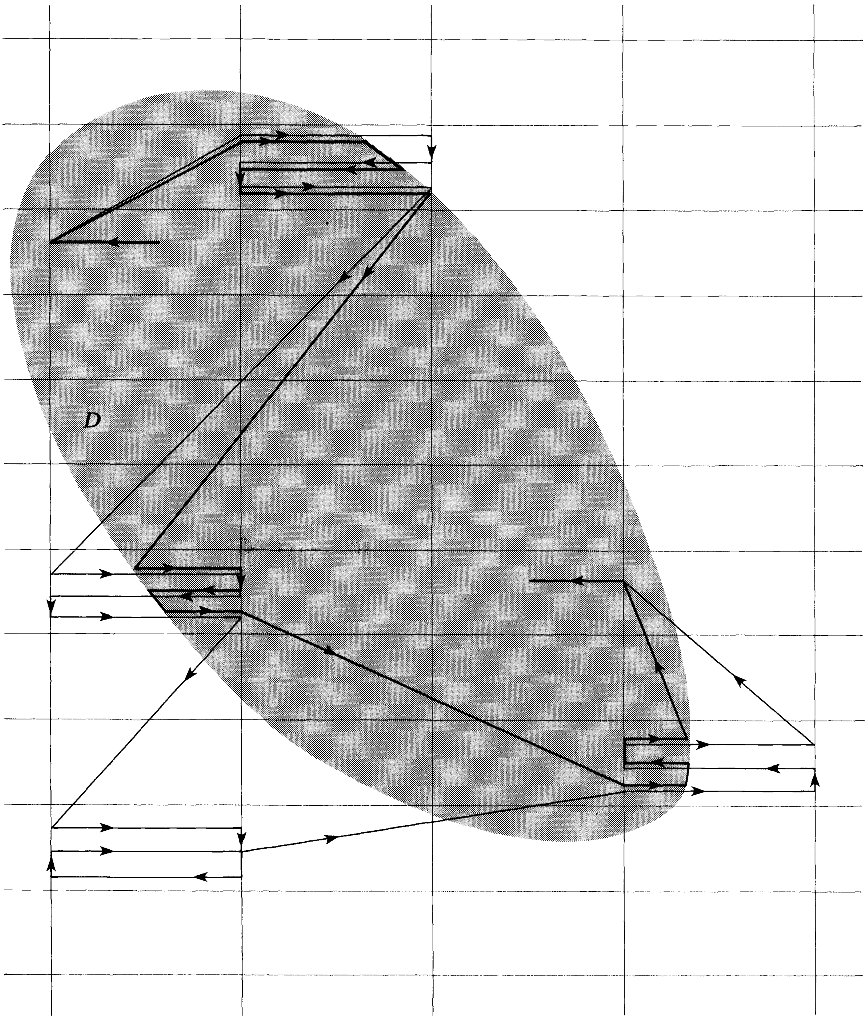

FIG. 2

Various generalizations of the search game are easily solved in a similar fashion. In particular, if $\mathscr{D}$ is a bounded convex set in $\mathbb{R}^n$ with $n>2$ and if $\mathscr{D}$ has a nonempty interior, the modifications in the strategies are small. No essential change need be made for the princess. For the monster, divide the space into congruent rectangular prisms. In step (ii) pick points at random in the ends of the prisms.

If $\mathscr{D}$ is a finite union of bounded convex sets, the strategies must be adjusted. Suppose $\mathscr{D}$ is a connected set and $\mathscr{D} = C_1 \cup C_2 \cup \cdots \cup C_N$ where each C_k is convex and has interior with positive area (volume). The princess should still wait until ε of the area (volume) could have been covered; she should then move to a point chosen with uniform probability over $\mathscr{D}$. It is important that in moving between the points she should not go past a corner too often or otherwise make some region a good search area for the monster. Figure 3 depicts a path sufficiently random. Suppose the princess wishes to go from P to Q. There exists a minimal sequence of convex sets $C_{k_1}, C_{k_2}, \cdots, C_{k_n}$ contained in $\mathscr{D}$ such that $P \in C_{k_1}$ and $Q \in C_{k_n}$ and $C_{k_m} \cap C_{k_{m+1}} \neq \phi$ for $m = 1, \cdots, N-1$. Pick P_1 at random in $C_{k_1} \cap C_{k_2}, P_2$ in $C_{k_2} \cap C_{k_3}$, etc. Then $\overline{PP_1}, \overline{P_1P_2}, \cdots, \overline{P_{M-1}Q}$ is a route for the princess from P to Q. For a fixed set $\mathscr{D}$, and for sufficiently small detection radius, the moves are made infrequently enough and are distributed evenly enough that knowing how and when the transitions will be made would be of little help to the monster.

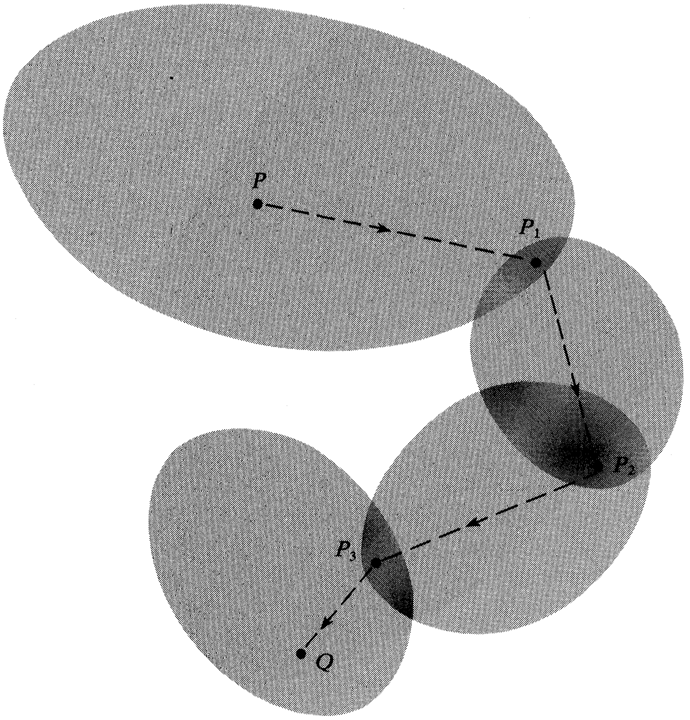

FIG. 3

The monster's strategy must be adjusted also. We are supposing that $\mathscr{D}$ is the union of N convex sets. Follow the same tiling procedure as before, except give weight N to those rectangles within the detection radius of points which are not in $\mathscr{D}$. This counting should be used in figuring the total area in the rectangles and in determining the probability that the monster picks a particular rectangle in step (i). Figure 2 depicts the modification of the search of the rectangles needed to keep the monster within $\mathscr{D}$ if $\mathscr{D}$ is one convex set. When $\mathscr{D}$ is the union of N such sets, the same procedure is followed for each convex set for those rectangles in the convex set and within the detection radius of the boundary of $\mathscr{D}$. The multiple counting allows for step (i) to include a random choice of convex set. The monster should carry out the transitions between areas as the princess does.

The monster's strategy still is ε-optimal for sufficiently small detection radius. The multiple covering of a small area does not lower the effectiveness of the search elsewhere. The expected time until capture is essentially the length of time to search out an area (volume) equal that of $\mathscr{D}$. Hence Gal's Conjecture A has been proved. The result is now summarized.

THEOREM. *Let C be a closed, bounded, convex set in $\mathbb{R}^n$ for $n \geqq 2$. Suppose C has interior with positive n-dimensional content. Suppose $\mathcal{D}$ is a finite union of such convex sets, and the interior of $\mathcal{D}$ is connected. Consider a mobile hider and searcher within $\mathcal{D}$. The positive, maximum speed of the hider is known to both participants. The maximum speed of the searcher is taken to be one and is known to both. Each player knows its own past and present location. There is a detection radius r such that if the hider and searcher are within r of each other, capture takes place. Let the payoff of the game be the time until capture. Let $g = 2r$ if $n = 2$, $g = \pi r^2$ if $n = 3$, etc. Then given $\varepsilon > 0$, for sufficiently small r, the value of the game to within a factor of $(1 \pm \varepsilon)$ is the area (volume) of $\mathcal{D}$ divided by g.*

The proof of the theorem suggests a generalization. Let $\mathcal{D}$ be a set as described in the theorem. If the small detection radius varies from one part of $\mathcal{D}$ to another, the idealized search problem can still be solved. Let r_k be the detection radius in a convex set $\tilde{C}_k$ for $k = 1, 2, \cdots, M$. Suppose these sets $\tilde{C}_k$ represent $\mathcal{D}$ in the sense that $\tilde{C}_{k_1} \cap \tilde{C}_{k_2}$ has no interior for $k_1 \neq k_2$ and $\mathcal{D} = C_1 \cup \cdots \cup C_M$. When the princess picks a point at random, she should give a relative weight to each area C_k of $1/r_k$ times its area. Similarly, the monster should give a relative weight to the rectangles in C_k of $1/r_k$ per rectangle. Fixing the ratios of the detection radii, the theorem can be extended as follows. Given $\varepsilon > 0$, for sufficiently small detection radii $\{r_k\}$, the value of the game is between $(1+\varepsilon)$ times and $(1-\varepsilon)$ times the time it would take the monster to search out an area the size of C_1 using detection radius r_1 plus C_2 using detection radius r_2, etc.

How essential is the oft repeated condition "for sufficiently small detection radius"? If $\mathcal{D}$ were a rectangle with width a small fraction of the detection radius, obviously the expected time until capture would be many times the time necessary for the monster to search out an area equal that of $\mathcal{D}$. Suppose the geometry of $\mathcal{D}$ is such that the monster can search $\mathcal{D}$ without being closer to the boundary than his detection radius. Even then the expected time until capture may be many times the time necessary to search an area the size of $\mathcal{D}$. An example of such a $\mathcal{D}$ is shown in the next section provided it is understood that any network can be approximated in $\mathbb{R}^3$ by a finite union of cylindrical tubes ending with hemispheres. The radii of the tubes and hemispheres should be equal to the detection radius.

3. A network counterexample.

Gal's second conjecture is incorrect. Assume the princess' maximum speed is at least one. Then, for any number K, there exists a network for which the expected time until capture is greater than K times the length of the network.

Consider first the network indicated in Fig. 4. There are N points in a small, symmetric, circular arrangement. Every pair of these points is joined by a line, thus forming a complete graph. Also from each point there is a line of length L going away from the N points; such a line will be called a ray of length L. The total length of the complete graph is much less than L. The full figure will be called a blossom.

A possible strategy for the princess is now indicated. She could stay far away from the N points at the end of a ray until a small fraction η of the total length of the network could be traversed by the monster; then she could move to the other end of the ray, pick a ray at random, go directly to it and down it to its far end, and wait, etc. Choose N so that $\eta N \gg 1$. Consider the risk the princess takes in relocating. She moves at least as fast as the monster. Consequently, even if the monster knew when the trip started, he could blockade at most one ray of length L from her exit and one ray from her entry. For a typical choice of edge from the complete graph, there are N congruent edges. The monster could prevent entry for two of these and exit for two. Thus the probability of

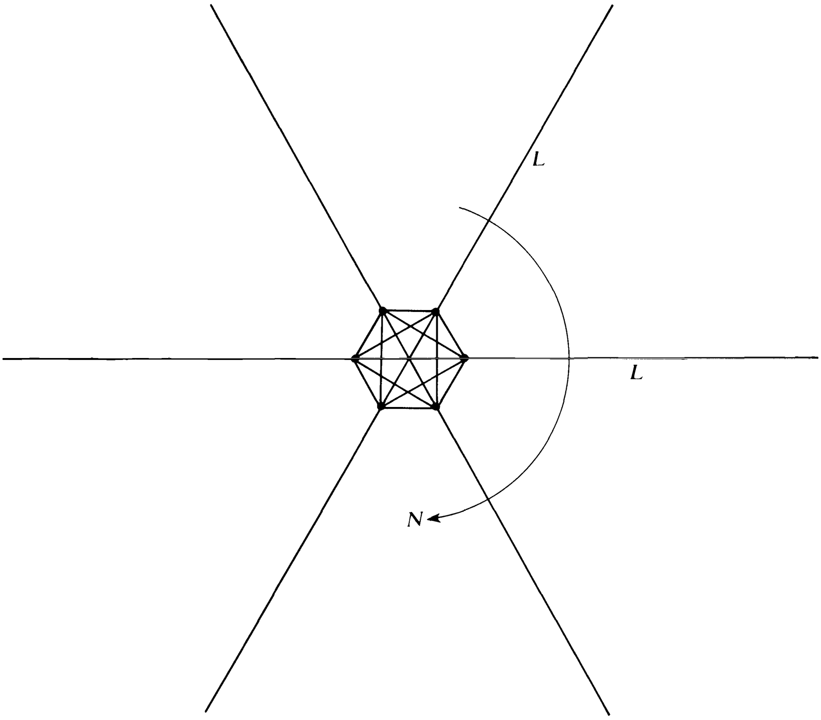

FIG. 4

capture is less than $6/N$. This risk is taken so seldom it can be ignored. The monster is far more likely to find the princess while she is waiting.

The monster should select a ray, go down it to the end and back; pick a ray at random, etc. During each such step he goes a distance of $2L$ plus a small transit distance and has a chance of finding her of slightly more than $1/N$. At each step her chance of not having been found is multiplied by another factor of $(1-1/N)$. After k steps the probability of her not having been captured is $(1-1/N)^k$. The expected time until capture is approximately

$$\begin{aligned}\sum_{k=0}^{\infty} 2Lk\left(1-\frac{1}{N}\right)^k \frac{1}{N} &\leq \frac{2L}{N}\sum_{k=0}^{\infty} k\left(1-\frac{1}{N}\right)^k \\ &= \frac{2L}{N}\left[\sum_{k=0}^{\infty}\left(1-\frac{1}{N}\right)^k\right]\left[\sum_{k=1}^{\infty}\left(1-\frac{1}{N}\right)^k\right] \\ &= \frac{2L}{N}\frac{1}{1-(1-1/N)}\frac{1-1/N}{1-(1-1/N)} = \frac{2L}{N}NN\left(1-\frac{1}{N}\right) \\ &= 2LN(1-1/N).\end{aligned}$$

which is approximately twice the length of the network. Gal [1] alluded to such a network and conjectured that it is as hard for the monster to search as any network could be.

Note that the probability the princess is still free after the monster has had time to travel almost twice the length of the network is $p_0 \doteq (1-1/N)^N \doteq e^{-1}$. The only type of edges not searched an average of twice are those joining pairs of the N points.

Now consider the network sketched in Fig. 5. Again N points are joined by lines in all possible ways. The total length of those lines is to be much less than L. To each point a ray is joined. At the other end of each ray is a blossom as in Fig. 4. The length of the N rays is $\alpha_1 NL$ where α_1 is to be specified later.

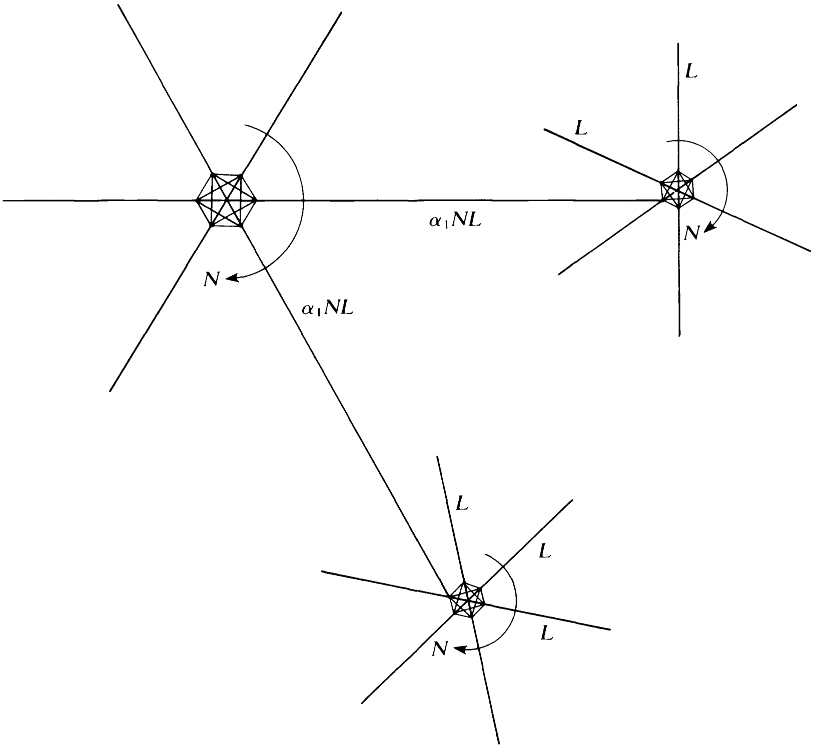

FIG. 5

A strategy for the princess involves a number η such that $1 \gg \eta > 0$ and $N\eta \gg 1$. She could move within a blossom as previously described. She could change blossoms every time the monster would have had time to search out several of them, for example $\eta N^2 L$ would be an appropriate length of time between moves between blossoms.

How should the monster search against such a strategy? If a_1 were very small, he should go from one ray of length L to another in the same blossom until the probability the princess is in that blossom has declined slightly; then he should go to another blossom. Thus he would never search a length equal to $2NL$, the distance to fully search a blossom.

If α_1 were ten, he should search out the rays in a blossom until the chance of her being in the blossom is rather small. Certainly he should search it out more than one time. For any value of α_1 he should search until the chance of finding her in the next ray of length L is as high per unit distance as the average chance per unit distance of finding her in traveling back and forth on the ray of length $\alpha_1 NL$ plus the searching of the blossom.

There is a value of α_1 such that the optimal strategy for the monster against the stated strategy of the princess is to search out each blossom a distance equal to $2NL$. Pick that value for α_1. As N tends to infinity, the value of α_1 tends to a positive, finite limit. Thus α_1 can be regarded as independent of N.

In searching a blossom which the princess is in, the probability the monster will find her is $(1-p_0)$. Since the princess rarely changes blossoms, in order to find the princess, the monster must pick the correct blossom. The chance of finding her in searching one random blossom once is $(1-p_0)/N$. The chance of the princess not being found is $1-(1-p_0)/N$. After searching N random blossoms, the chance the princess is free is $p_1 \doteq (1-(1-p_0)/N)^N \doteq e^{-(1-p_0)} = e^{-1}e^{p_0} \doteq e^{-1}\, e^{1/e}$. Thus, the expectation of the princess still being free after the monster has searched the network a distance equal to almost twice its length is $p_1 \doteq e^{-1}\, e^{1/e} > 1/e \doteq p_0$. The only types of edges which are not covered an average of twice are those which join pairs of nearby points.

Consider a network as in Fig. 6. Regardless of the value of α_2, an optimal search of such a network will involve searching each blossom once, covering each ray of length L an average of twice, before moving away from the blossom by using a ray of length $\alpha_1 NL$. A choice of α_2 can be made so that in an optimal search the N rays of length $\alpha_1 NL$ are covered an average of twice, once in each direction, between traversing a ray of length $\alpha_2 N^2 L$. As N tends to infinity, the chosen value of α_2 tends to a positive, finite limit. Hence α_2 can be regarded as independent of N.

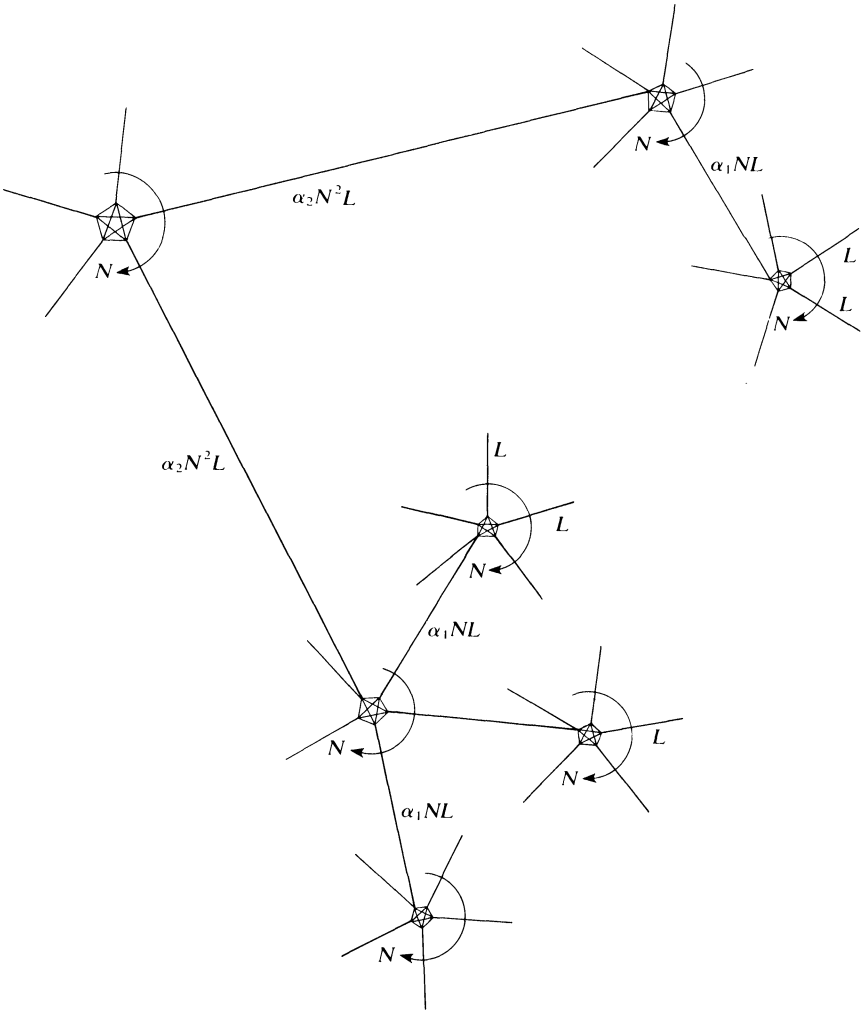

FIG. 6

After this network is searched sufficiently long that each ray of length L and $\alpha_1 NL$ and $\alpha_2 N^2 L$ is covered an average of twice, there is a probability p_2 that the princess has not been found. Clearly $p_2 \doteq (1-(1-p_1)/N)^N \doteq e^{-1} e^{p_1}$.

Networks can be built up in this way. Suppose a network is to have a large fixed number M of stages as described in the preceding. For large N, the numbers α_1 through α_M closely approximate their limits. The number η must be chosen so that $1 \gg \eta > 0$. For N sufficiently large, $\eta N \gg 1$ and the risk of the princess being captured while relocating is small. For sufficiently large N, the approximations $p_0 \doteq (1-1/N)^N \doteq e^{-1}$ and $p_{m+1} \doteq (1-(1-p_m)/N)^N \doteq e^{-1} e^{p_m}$ for $m = 0, 1, \cdots, M-1$ are accurate. Hence it has been shown that, for any fixed M, there is a network for which the approximations $p_0 = e^{-1}$ and $p_{m+1} = e^{-1} e^{p_m}$ for $m = 0, 1, \cdots, M-1$ are arbitrarily accurate. In that sense, these equations can be regarded as exact.

It is not hard to show that $\lim_{m\to\infty} p_m = 1$. Let $f(x) = e^{-1} e^x$. Then $f'(x) > 0$ and $\{p_m\}$ is monotonic. Since $p_0 < p_1$, $\{p_m\}$ is increasing. If p_{m+1} were greater than 1, then $e^{-1} e^{p_m} > 1$ and p_m would be greater than 1. Since $p_0 < 1$, all $p_m < 1$. Hence there exists a limit p. By continuity of $f(x)$ and x, $e^{-1} e^p = p$. If p were less than one, then $f(p) - p = f(1) - 1$ and $(f(1) - f(p))/1 - p = 1 = f'(\xi)$ for some ξ with $p < \xi < 1$. But $f'(x) = e^{-1} e^x = 1$ only for $x = 1$. Thus $\lim_{m\to\infty} p_m = 1$.

We can now complete the final step in the counterexample. Given a large number K there exists a sufficiently large positive integer M that $(p_M)^K > e^{-1}$. Thus for a network of M stages the expected time before capture would be K times the time to search it out once, that is, almost $2K$ times the total length of the network.

4. Can a knight help? A story with a princess and a monster should have a chance for a happy ending. Imagine that a well-prepared knight is in the network with the princess and the monster. If he can find the monster before the monster captures the princess, the ending will be pleasant.

We will suppose that when the knight is searching along an edge of the network he has a rather slow maximum speed of v in comparison to the monster's maximum speed of one. The knight will have the special ability to jump from any point in one of the groups of N points to any point in another group of N points. The time it takes to make such a jump will be assumed to be the same as it takes him to move once along an edge of length L, that is L/v. Can a network of the type described in § 3 be found in which the knight can rescue the princess?

A candidate for such a network is constructed depending on a parameter ε. Let ε be given such that $1 \gg \varepsilon > 0$. Pick a positive integer M such that $1 > (p_M)^{2/v} > 1 - \varepsilon/2$. Consider a network of this M number of stages. Pick N so large that $\alpha_1 NL, \alpha_2 N^2 L, \cdots, \alpha_M N^M L$ are all much greater than L and large enough that $p_{m+1} = e^{-1} e^{p_m}$ is a good approximation for $m = 1, 2, \cdots, M-1$.

Consider the following search strategy for the knight. He should consider all the edges of the network, those of length L, of length $\alpha_1 NL$, of length $\alpha_2 N^2 L, \cdots$, of length $\alpha_M N^M L$, and the complete graphs on N points. Denote this collection by $\mathscr{E}$ and call its elements "generalized edges." He should pick at random a generalized edge. He should select a point on the generalized edge in one of the groups of N points and go to that point. The knight should move through that generalized edge in both directions. Then he should start again by picking at random a generalized edge, etc.

Consider the expected results in the time it takes the monster to go through his search of the network $2/v$ times. Note that this period is essentially the same as the time for the monster to cover a distance equal to the full length of the network $4/v$ times since the complete graphs on N points are a small part of the network. In the same network period, the knight will be able to cover each generalized edge an average of one time in each direction and will have the same amount of time for transitions, that is, for jumping between generalized edges. The chance of the princess escaping from the monster during this period is $p_m^{2/v} > 1-\varepsilon/2$ without taking into account the possibility that the knight has found the monster. The chance of the knight finding the monster in this period can be estimated as follows. If the monster has not been found, the chance of discovery during the search of the next generalized edges is at least one divided by,the number of generalized edges. Let n be the number of generalized edges. The chance of the monster avoiding the knight during this period is less than $(1-1/n)^n \doteq e^{-1}$ regardless of what strategy the monster adopts. Thus, the chance that the knight finds the monster is greater than $1-1/e > 1/2$. Hence the probability that the princess will ultimately be saved is greater than $1-\varepsilon$. The knight can save the princess because his jumping ability makes him a more effective searcher than the monster.

A similar problem can be posed for a bounded, closed, convex set $\mathscr{D}$ in the plane. The knight is to have the same, very small, detection radius r as the monster. Suppose the maximum speed v of the knight is small in comparison to the monster's maximum speed of one. The knight has the ability to jump from any point in $\mathscr{D}$ to any other taking a time equal to the transit time for a distance $4r$. During the jump, the knight cannot find the monster. It is left to the reader to show that the outcome is not pleasant.

5. Summary. The princess and monster problems has been solved for a wide class of domains $\mathscr{D}$ for sufficiently small detection radius. The value of the game has been shown to be approximately the time necessary for the monster to cover an area (volume) equal to that of the domain $\mathscr{D}$. The results extend work of Gal and prove one of his conjectures.

A network version of the problem was also considered. The results were contrary to a conjecture of Gal. It was shown that the expected time to capture may be arbitrarily many times the period required for the monster to search the length of the network.

A modified form of the princess and the monster problem was discussed. The observations suggested the difference between a domain and a network. The transition times for the monster in a domain are small in comparison to the time necessary to perform a local search. In a network, it may be a lengthy process to reach a different part of the network.

Acknowledgments. Pleasant conversations are gratefully recalled with a princess (Miss Gayle Mayfield) and the following colleagues: Professors R. Bartels, R. Byrd, T. Corwin, S. Gal, R. Horn, R. Isaacs, and A. Karr.

REFERENCES

[1] S. GAL, *Search games with mobile and immobile hider*, this Journal, 17 (1979), pp. 99–122.

[2] ———, Oral communication.

[3] R. ISAACS, *Differential Games, A Mathematical Theory with Applications to Warfare and Pursuit, Control and Optimization*, Reprint, Robert E. Krieger, Huntington, NY, 1975.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 6, November 1979

0363-0129/79/1706-0004$01.00/0

WIENER-HOPF METHODS FOR OPEN-LOOP UNSTABLE DISTRIBUTED SYSTEMS*

JON H. DAVIS†

Abstract. A Wiener–Hopf based solution to the linear regulator problem is presented for a class of distributed systems with a finite-dimensional unstable subspace. The results provide an integral representation of the optimal feedback gains in terms of the system description and associated spectral factorization.

1. Introduction. In this paper we consider a Wiener–Hopf approach to least-squares control problems for a class of open-loop unstable distributed systems. Our aim is to derive an integral representation of the optimal feedback gains dual to the result given for distributed filters in [1].

Since we consider open-loop unstable systems, the results obtained do not follow directly from those of [1]; in fact it appears technically inconvenient to attempt an argument strictly "dual" to that used in [1] even for the stable open-loop case.

Spectral factorization methods have been applied to related problems in [3]. Our approach is to make sufficiently strong assumptions on the system model so that the classical spectral factorization results of Gohberg and Krein are applicable. This produces (as expected) a bounded linear functional as the feedback operator. These results are a generalization of the results of [5] for finite-dimensional systems, and similarly require use of "transfer function" data (as well as geometrical knowledge of the "unstable mode subspace").

The method avoids consideration of distributed Riccati equations, and for this reason appears to be useful as a computational approach in certain problems [5].

2. Problem formulation. We consider the "linear regulator problem" for a distributed system of the form

$$\frac{dx}{dt}=Ax+Bu(t), \qquad y=Cx, \qquad t\geqq 0. \tag{1}$$

Since our approach is through Wiener–Hopf methods, we assume that the control space is finite dimensional, so that the input mapping B in (1) represents a bounded linear transformation from C^m to a separable Hilbert space H_0.

In view of results on stabilizability of systems with finite-dimensional controls [6], [7], it is natural to assume that the open-loop system splits as the direct sum of a finite dimensional subspace (the unstable subspace) and a complementary subspace on which the semigroup $\{S_t\}$ is exponentially stable. That is, we assume that the singularities of the resolvent $R(\lambda; A)$ in the right-half plane consist of a finite number of poles of finite order, that the spectral projections associated with these poles have finite dimensional range, and that the system is exponentially stable on the complementary subspace. See Appendix A.

The output mapping C is assumed bounded from H_0 to a separable Hilbert space H_1. (It is probably only necessary to assume that C is A-bounded and to make suitable assumptions on the operator valued function $CR(s; A)$ in order to obtain these results.)

* Received by the editors June 12, 1978, and in revised form February 6, 1979.

† Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, Canada K7L 3N6. This research was supported in part by the C.M.C./N.R.C. Workshop on Control Theory and Optimization, Centre de Recherches Mathématique, Université de Montréal, May 1978.

Consider the Hilbert space $\mathscr{H}_1 = L_2((0, \infty); H_1)$ (equivalence classes of strongly measurable, H_1 valued, square integrable functions on $(0, \infty)$) and the linear mapping $T: H_0 \oplus L_2((0, \infty); C^m) \to \mathscr{H}_1$, defined by the relation

$$T\begin{bmatrix} x \\ u(\cdot) \end{bmatrix}(t) = C\left[S_t x + \int_0^t S_{t-\tau} Bu(\tau)\, d\tau\right] \tag{2}$$

for $\begin{bmatrix} x \\ u(\cdot) \end{bmatrix} \in \mathscr{D}(T)$, the domain of T.

In the case that the system is open-loop unstable, T as defined by (2) fails to map into $\mathscr{H}_1$ unless the domain of T is suitably restricted.

The appropriate restriction on the domain is naturally determined in the case that the subsystem associated with the unstable subspace is minimal; that is, the system model

$$\frac{dx}{dt} = Ax + Bu, \qquad y = Cx$$

is both stabilizable and detectable.

We define the domain of T by

$$\mathscr{D}(T) = \left\{\begin{bmatrix} x \\ u(\cdot) \end{bmatrix} \middle| T\begin{bmatrix} x \\ u(\cdot) \end{bmatrix} \in \mathscr{H}_1\right\}. \tag{3}$$

With our assumption on the right-half plane singularities of the resolvent of A, it is easy to give a "frequency domain" interpretation of $\mathscr{D}(T)$. In fact, $\begin{bmatrix} x \\ u(\cdot) \end{bmatrix} \in \mathscr{D}(T)$ if and only if

$$CR(s; A)x + CR(s; A)B\hat{u}(s)$$

($\hat{u}(\cdot)$ is the Fourier–Laplace transform of $u(\cdot)$) belongs to the H_1-valued Hardy space $H^2(\pi^+; H_1)$.

Using the assumption of finite dimensionability of the unstable subspace, it is shown in Appendix A that $\mathscr{D}(T)$ is determined by orthogonality conditions of the form

$$\mathscr{D}(T) = \left\{\begin{bmatrix} x \\ u(\cdot) \end{bmatrix} \middle| \begin{bmatrix} x \\ u(\cdot) \end{bmatrix} \perp \begin{bmatrix} \psi_j \\ B_u^* e^{-A_u^* t} \psi_j \end{bmatrix}, \right. \tag{4}$$
$$\left. \{\psi_j\}_{j=1}^N \text{ a basis for the unstable subspace.}\right\}$$

(Here B_u and A_u are the input map and semigroup generator associated with the unstable subsystem.)

These computations identify $\mathscr{D}(T)$ as a closed subspace of finite codimension in $H_0 \oplus L_2((0, \infty); C^m)$; T acting on $\mathscr{D}(T)$ is everywhere defined and bounded on this Hilbert space.

We equip $\mathscr{D}(T)$ with the graph norm:

$$\left\|\begin{bmatrix} x \\ u(\cdot) \end{bmatrix}\right\|_{\mathscr{D}}^2 = \left\|\begin{bmatrix} x \\ u(\cdot) \end{bmatrix}\right\|_{H_0 \oplus L_2}^2 + \left\|T\begin{bmatrix} x \\ u(\cdot) \end{bmatrix}\right\|_{\mathscr{H}_1}^2. \tag{5}$$

The linear regulator problem is then identified as a standard minimum norm problem for the $\mathscr{D}$-norm:

$$\text{minimize} \left\|\begin{bmatrix} x \\ u(\cdot) \end{bmatrix}\right\|_{\mathscr{D}}^2$$

subject to the linear constraint

$$L\begin{bmatrix} x \\ u(\cdot) \end{bmatrix} = x_0, \tag{6}$$

where L is the projection onto the first component of the element $\begin{bmatrix} x \\ u(\cdot) \end{bmatrix} \in \mathscr{D}(T)$.

Since the range of the projection L is the entire state space, a unique minimizing pair exists, characterized by

$$\begin{bmatrix} x_0 \\ u_{\text{opt}}(\cdot) \end{bmatrix} = L^* z, \tag{7}$$

where z is any solution of $LL^* z = x_0$ [9].

It now remains to identify a Wiener–Hopf equation for the optimal control, and to synthesize the optimal control feedback form.

3. Derivation and solution of the Wiener–Hopf equation. For the solution of the minimum norm problem proved above, it is necessary to compute L^*, the adjoint of the projection L relative to the $\mathscr{D}$-inner product.

If we let P denote the orthogonal projection of $H_0 \oplus L_2((0, \infty); C^m)$ onto $\mathscr{D}(T)$, then it is easy to verify that L^*: $H_0 \to \mathscr{D}(T)$ is given by

$$L^* y = (I + (TP)^* TP)^{-1} P\begin{bmatrix} y \\ 0 \end{bmatrix}. \tag{8}$$

This follows since

$$\begin{aligned}
\left\langle \begin{bmatrix} x \\ u \end{bmatrix}, L^* y \right\rangle_{\mathscr{D}} &= \left\langle \begin{bmatrix} x \\ u \end{bmatrix}, (I + (TP)^* TP)(I + (TP)^* TP)^{-1} P\begin{bmatrix} y \\ 0 \end{bmatrix} \right\rangle_{H_0 \oplus L_2} \\
&= \left\langle \begin{bmatrix} x \\ u \end{bmatrix}, P\begin{bmatrix} y \\ 0 \end{bmatrix} \right\rangle_{H_0 \oplus L_2} \\
&= \left\langle P\begin{bmatrix} x \\ u \end{bmatrix}, \begin{bmatrix} y \\ 0 \end{bmatrix} \right\rangle_{H_0 \oplus L_2} \\
&= \left\langle \begin{bmatrix} x \\ u \end{bmatrix}, \begin{bmatrix} y \\ 0 \end{bmatrix} \right\rangle_{H_0 \oplus L_2} \\
&= \langle x, y \rangle_{H_0} \\
&= \left\langle L\begin{bmatrix} x \\ u \end{bmatrix}, y \right\rangle_{H_0}
\end{aligned} \tag{9}$$

Using this representation of the adjoint in the determining equation for the optimal control, we find that the equation

$$LL^* z = x_0$$

reduces to the obvious statement that the first component of the optimal pair is the initial state. The optimal pair is characterized by

$$\begin{bmatrix} x_0 \\ u_{\text{opt}} \end{bmatrix} = L^* z. \tag{10}$$

That is,

$$(I+(TP)^*TP)\begin{bmatrix} x_0 \\ u_{\text{opt}} \end{bmatrix} = P\begin{bmatrix} z \\ 0 \end{bmatrix} \tag{11}$$

for some $z \in H_0$.

From the derivation it might be expected that it is necessary to solve the above for z. However, this is not the case, as only u_{opt} is unknown and (see below) the domain condition is sufficient to calculate the effect of z on u_{opt}.

In order to obtain a Wiener–Hopf equation for the optimal control, it is useful to obtain a "frequency domain representation" of the bounded linear operator TP.

Define the projection P_+:

$$H_0 \oplus L_2((-\infty, \infty); C^m) \to H_0 \oplus L_2((0, \infty); C^m)$$

in the usual fashion.

$$P_+\begin{bmatrix} x_0 \\ u(\cdot) \end{bmatrix}(t) = \begin{bmatrix} x_0 \\ \begin{cases} 0, & t<0, \\ u(t), & t \geqq 0, \end{cases} \end{bmatrix}. \tag{12}$$

The operator $\widehat{TPP}^+$: $H_0 \oplus L_2((-\infty, \infty); C^m) \to \mathscr{H}_1$ then has a simple representation in the Fourier transform domain:

$$\widehat{TPP}_+ \colon H_0 \oplus \hat{L}_2((-\infty, \infty); C^m) \to \hat{\mathscr{H}}_1,$$

$$\widehat{TPP}_+\begin{bmatrix} x_0 \\ U(\omega) \end{bmatrix} = [CR(i\omega; A) \quad CR(i\omega; A)B]\hat{P}\hat{P}_+\begin{bmatrix} x_0 \\ U(\omega) \end{bmatrix}.$$

This follows since $\mathscr{D}(T)$ consists exactly of those pairs for which the transform representation has square integrable boundary values and extends analytically over the right-half plane.

We consider first

$$(I+(TPP_+)^*(TPP_+))\begin{bmatrix} x_0 \\ u_{\text{opt}} \end{bmatrix} = P_+P\begin{bmatrix} z \\ 0 \end{bmatrix}, \tag{13}$$

and then pass to the Fourier transform representation

$$(I+(\widehat{TPP}_+)^*(\widehat{TPP}_+))\begin{bmatrix} x_0 \\ U_{\text{opt}}(\omega) \end{bmatrix} = \widehat{P_+P}\begin{bmatrix} z \\ 0 \end{bmatrix}. \tag{14}$$

We refer to the above as the Wiener–Hopf equation for the optimal control. In the case of a *stable open-loop* the "second equation" in (14) reduces to the usual Wiener–Hopf equation for the optimal control function. That is

$$\hat{U}(\omega) + \hat{P}_+[G^*(i\omega)G(i\omega)\hat{U}(\omega)] = -\hat{P}_+[G^*(i\omega)CR(i\omega; A)x_0],$$

where

$$G(i\omega) = CR(i\omega; A)B,$$

$$G^*(i\omega) = B^*[R(i\omega; A)]^*C^*.$$

This equation is solved by the usual Wiener–Hopf technique, which may be loosely described as a process of dropping the projection $\hat{P}_+$ from the equation, and making subsequent allowance for the (unknown) element of the nullspace of $\hat{P}_+$ introduced by this process.

This general procedure is also applicable to the Wiener–Hopf equations (13), (14). In this case the presence of both the half-line projection P_+ and the domain projection P both interfere with the production of a "whole-line" convolution equation involving the optimal control function. Unknown elements of the nullspace of each projection must be added to the equation as the projections are dropped.

Since the projection P is onto a subspace of finite codimension (under our hypotheses of codimension equal to the dimension of the unstable subspace), P may be visualized as a Gram–Schmidt procedure carried out with respect to the (linearly independent by the stabilizability hypothesis) vectors given in (4).

Dropping P from the left side of (13) therefore introduces an unknown term of the form

$$\sum_{j=1}^{N} \gamma_j \begin{bmatrix} \psi_j \\ B_u^* e^{-A_u^* t}\psi_j \end{bmatrix}$$

into the equation. A term which may be compactly represented in the form

$$B_u^*(i\omega + A_u^*)^{-1}\alpha = \mathscr{F}\{B_u^* e^{-A_u^* t}\alpha\},$$

where α is an unknown element of the unstable subspace, appears in the equation for the optimal control after P is dropped from the left side of (14).

The "forcing term" $P\begin{bmatrix} z \\ 0 \end{bmatrix}$ contributes a term of the same form, which we absorb into the above. (This verifies the transient nature of our interest in the element z arising from the minimum norm formulation.)

The consequence of this procedure is that we obtain the relation

(15) $$[I + G^*(i\omega)G(i\omega)]U_{\text{opt}}(\omega) = -G^*(i\omega)CR(i\omega; A)x_0 - B_u^*(i\omega + A_u^*)^{-1}\alpha + \underline{\beta}(\omega),$$

where $\underline{\beta}(\cdot)$ is the Fourier transform of an element of $L_2((-\infty, 0); C^m)$, and α remains to be determined from the condition that the optimal pair belongs to $\mathscr{D}(T)$. (It will be shown below that this condition uniquely determines α under our hypotheses.)

Equation (15) is solved by the spectral factorization technique.

In order to obtain the "feedback form" of the optimal control, it is useful to invoke the $\hat{L}_1$ spectral factorization theory of [4]. It is therefore necessary to verify that the (matrix) elements of the expression G^*G of (15) are Fourier transforms of integrable functions. In fact, our assumptions on the original model (1) are sufficiently strong to guarantee this.

A typical "matrix element" of G^*G has the form

$$b_i^*[R(i\omega; A)]^*C^*CR(i\omega; A)b_j,$$

where b_k is the image under B of a standard basis vector in C^m (i.e., a vector in H_0).

Assume first that A generates an exponentially stable semigroup. Then the functions $y_k: (-\infty, \infty) \to H_1$

$$y_k(t) = \begin{cases} CS_t b_k, & t \geqq 0, \\ 0, & t < 0, \end{cases}$$

are piecewise continuous H_1-valued functions, integrable in norm. Define (for each t) $y_i^*(t)$ as the canonical element of the dual of H_1 defined by $y_i(t)$. It then follows that the scalar valued function

(16) $$h_{ij}(t) \triangleq \int_{-\infty}^{\infty} y_i^*(\tau - t)y_j(\tau)\, d\tau$$

belongs to $L_1(-\infty, \infty)$. Since the Fourier transform of $h_{ij}(\cdot)$ is the desired matrix element, the conclusion follows.

In the case that the system has a finite dimensional unstable subspace, a similar argument applies. It is necessary only to interpret terms in the resolvent corresponding to the unstable subspace as Fourier transforms of integrable vector-valued functions of negative support. The functions $\{y_k(\cdot)\}$ may then be appropriately defined, and the conclusion then follows as above.

An argument parallel to that used above leads to the conclusion that the term

$$G^*CR(i\omega; A)x_0$$

appearing on the right side of (15) is the Fourier transform of an $L_2((-\infty, \infty); C^m)$ function, so that the usual Wiener–Hopf methods are available to complete the solution of (15). We factorize the coefficient matrix in the form

$$(I+G^*G)(\omega)=F^-(i\omega)F^+(i\omega), \tag{17}$$

where $F^+(i\omega)$, $[F^+(i\omega)]^{-1}$ (resp., $F^-(i\omega)$, $[F^-(i\omega)]^{-1}$) differ from an identity matrix by the Fourier transform of a causal (resp., anti-causal) L_1 convolution operator [4]. We recall that with this normalization at infinity, the factors are unique and related (for real ω) by the adjoint mapping. (Writing the arguments of the spectral factors as "$i\omega$" facilitates later Laplace transform computations.)

Using this factorization, the Fourier transform of the optimal control is given as

$$U_{\text{opt}}(\omega)=-[F^+(i\omega)]^{-1}\hat{P}_+\{[F^-(i\omega)]^{-1}[G^*(i\omega)CR(i\omega; A)x_0+B_u^*(i\omega+A_u^*)^{-1}\alpha]\}. \tag{18}$$

Applying the domain condition via Parseval's theorem (see Appendix A) leads to the finite-dimensional system of equations for the vector α,

$$\begin{aligned}\frac{1}{2\pi}\int_{-\infty}^{\infty}(-i\omega+A_u)^{-1}B_u[F^+(i\omega)]^{-1}\hat{P}_+\{[F^-(i\omega)]^{-1}B_u^*(i\omega+A_u^*)^{-1}\}\,d\omega\cdot\alpha\\ =Px_0-\frac{1}{2\pi}\int_{-\infty}^{\infty}(-i\omega+A_u)^{-1}B_u[F^+(i\omega)]^{-1}\\ \cdot\hat{P}_+\{[F^-(i\omega)]^{-1}G^*(i\omega)CR(i\omega; A)x_0\}\,d\omega.\end{aligned} \tag{19}$$

This system has the form

$$m\cdot\alpha=\beta(x_0)=\begin{bmatrix}\beta_1(x_0)\\ \vdots\\ \beta_N(x_0)\end{bmatrix}, \tag{20}$$

where each of the $\beta_i(\cdot)$ represents a bounded linear function on H_0.

The coefficient operator in (20) is self-adjoint; in fact, it is essentially a "Gram operator" closely related to the usual controllability (matrix) operator.

Consider the linear mapping Φ: H_u (the unstable subspace) → $\hat{L}_2((0, \infty); C^m)$

$$\Phi\alpha=\hat{P}_+\{[F^-(i\omega)]^{-1}B_u^*(i\omega+A_u^*)^{-1}\alpha\}.$$

Then Φ is one-to-one, since the condition

$$\hat{P}_+\{[F^-(i\omega)]^{-1}B_u^*(i\omega+A_u^*)^{-1}\alpha\}=0$$

implies that for some transform $\underline{\beta}(\cdot)$ of a function of negative support

$$B_u^*(i\omega + A_u^*)^{-1}\alpha = F^-(i\omega)\cdot\underline{\beta}(\omega),$$

since the left and right sides of this equality are transforms of functions of opposite support, we have

$$B_u^*(i\omega + A_u^*)^{-1}\alpha = 0.$$

Since the unstable subsystem is controllable by hypothesis, this forces

$$\alpha \equiv 0,$$

so that Φ is one-to-one. Since

$$m = \Phi^*\Phi \tag{21}$$

we conclude that m is invertible. These considerations uniquely determine the open-loop form of the optimal control. We collect these preliminary results as Lemma 1.

LEMMA 1. *Consider the linear regulator problem*

$$\min_{\left[\begin{smallmatrix} x_0 \\ u(\cdot)\end{smallmatrix}\right] \in \mathscr{D}(T)} \int_0^\infty \|u\|^2 + \|y\|^2 \, dt$$

subject to the constraints

$$x(t) = S_t x_0 + \int_0^t S_{t-\tau} B u(\tau)\, d\tau,$$

$$y(t) = Cx(t)$$

and with assumptions on the system model as given in Appendix A. *Then the unique optimal control is given in* (*Fourier-transformed*) *open loop form by*

$$U_{\text{opt}}(\omega) = -[F^+(i\omega)]^{-1}\hat{P}_+\{[F^-(i\omega)]^{-1}\{G^*(i\omega)CR(i\omega; A)x_0 + B_u^*(i\omega + A_u^*)^{-1}\alpha\}\},$$

where F^+, F^- *are the associated spectral factors* (17), *and the vector* α *is uniquely determined by* (20), (21).

4. Feedback synthesis of the optimal control. The previous sections have derived a form of open-loop control for a distributed linear regulator problem. It is expected, of course, that the optimal system is governed by a linear system of the form

$$\frac{dx}{dt} = \{A - B[B^*K]\}x \tag{22}$$

with $[B^*K]$ (the "feedback gains") some bounded linear mapping $H_0 \to C^m$ (hence essentially consisting of m continuous linear functionals on H_0).

The problem that remains is to derive an "integral representation" of the optimal feedback gains (see [5], [1]) and the conclusion that the optimal trajectory is governed by (22) (or a mild form thereof). In [1], a dual result was obtained by simply verifying that the hypothesized gain meets the requirements of the problem. In the present problem, this approach appears inconvenient due to algebraic complications which arise in the open-loop unstable case.

The approach taken below is based on the following intuitive argument. If the optimal trajectory is associated with a solution of an evolution equation (22), then the Laplace transform of the optimal trajectory must represent the resolvent of the corresponding semigroup generator. Since conditions are known determining operator-valued functions which are resolvents of semigroup generators [10, Thm. 5.8.3], these should determine the optimal closed loop dynamics.

We denote by $\hat{u}_{\text{opt}}(s)$ the evaluation of the Laplace transform of the optimal control function at a point s, $\operatorname{Re}(s)>0$. That is,

$$(23)\ \hat{u}_{\text{opt}}(s)\equiv[F^+(s)]^{-1}P_+\{F^-[(i\omega)]^{-1}\{G^*(i\omega)CR(i\omega;A)x_0+B_u^*(i\omega+A_u^*)^{-1}\alpha(x_0)\}\}(s).$$

(We use the notation $P_+\{\cdot\}(s)$ to denote the Laplace transform of the corresponding function of time defined on a half-line. This is naturally computable in terms of a Cauchy-type integral (cf. [1]).) It is clear that for each fixed s, $\operatorname{Re}(s)>0$, the above defines a bounded linear map $H_0\rightarrow C^m$: $x_0\rightarrow\hat{u}_{\text{opt}}(s)(x_0)$.

Now define an operator-valued function by

$$L(s)x_0=R(s;A)(x_0+B\hat{u}_{\text{opt}}(s)(x_0)). \tag{24}$$

From our assumptions on A and the fact that the pair $\begin{bmatrix} x_0 \\ u_{\text{opt}} \end{bmatrix}$ belongs to $\mathscr{D}(T)$, it follows that $L(s)$ is analytic for $\operatorname{Re}(s)>0$.

LEMMA 2. *The operator-valued function $L(\cdot)$ defined above satisfies the first resolvent equation, that is,*

$$L(\lambda)L(s)=\frac{1}{s-\lambda}[L(\lambda)-L(s)] \tag{25}$$

for $\operatorname{Re}\lambda,\ s>0,\ \lambda\neq s$.

Proof. Denote by $\hat{u}_{\text{opt}}(\lambda;\tau)$ the Laplace transform (with respect to λ) of the optimal control left-shifted by $\tau\geqq 0$.

$$\hat{u}_{\text{opt}}(\lambda;\tau)=\int_0^\infty e^{-\lambda t}u_{\text{opt}}(t+\tau)\,dt,\qquad \tau\geqq 0,\qquad \operatorname{Re}(\lambda)>0. \tag{26}$$

From the time invariance of the system and the "principle of optimality", it follows that

$$\hat{u}_{\text{opt}}(\lambda;\tau)=\hat{u}_{\text{opt}}(\lambda)(x(\tau)), \tag{27}$$

where $x(\tau)\in H_0$ is the value of the optimal state of time τ.

Compute now the Laplace transform of the above with respect to the τ-variable. The right side becomes

$$\hat{u}_{\text{opt}}(\lambda)(L(s)x_0),$$

since the Laplace transform commutes with the bounded linear functionals represented by $\hat{u}_{\text{opt}}(\lambda)$. The left term may be computed either directly from the definition (26), or by identifying the result as the representation of the resolvent of the left-shift semigroup on Fourier–Laplace transforms of $L_2((0,\infty);C^m)$. This results in

$$\frac{\hat{u}_{\text{opt}}(s)(x_0)-\hat{u}_{\text{opt}}(\lambda)(x_0)}{\lambda-s}=\hat{u}_{\text{opt}}(\lambda)(L(s)x_0). \tag{28}$$

If we denote by $\hat{x}(s)$,

$$\hat{x}(s)=R(s;A)(x_0+B\hat{u}_{\text{opt}}(s)(x_0)), \tag{29}$$

then we obtain successively

$$\begin{aligned} L(l)L(s)x_0 &= R(\lambda\,; A)(\hat{x}(s) + B\hat{u}_{\text{opt}}(\lambda)(\hat{x}(s)) \\ &= R(\lambda\,; A)\Big(R(s; A)(x_0 + B\hat{u}_{\text{opt}}(s)(x_0)) \\ &\qquad + B\left[\frac{\hat{u}_{\text{opt}}(s)(x_0) - \hat{u}_{\text{opt}}(\lambda)(x_0)}{\lambda - s}\right]\Big) \\ &= \frac{[R(\lambda\,; A) - R(s; A)]}{s - \lambda}(x_0 + B\hat{u}_{\text{opt}}(s)(x_0)) \\ &\quad - \frac{R(\lambda\,; A)}{s - \lambda}(B\hat{u}_{\text{opt}}(s)(x_0) - B\hat{u}_{\text{opt}}(\lambda)(x_0)) \\ &= \frac{1}{s - \lambda}[L(\lambda) - L(s)](x_0), \end{aligned}$$

so that the operator valued function $L(\cdot)$ satisfies the first resolvent equation.

LEMMA 3. *The operator-valued function $L(\cdot)$ defined above is in fact the resolvent of a closed operator $\tilde{A}$ which has the form of a finite-dimensional perturbation of the semigroup generator A. That is,*

$$\tilde{A} = A - B[B^*K],$$

where $[B^*K]$ is a bounded linear map: $H_0 \to C^m$.

Proof. Since $L(\cdot)$ satisfies the first resolvent equation, it follows from [10, p. 183] that it is sufficient to show that $L(\lambda_0)$ has an inverse (not necessarily bounded) for some λ_0 in the domain of $L(\cdot)$. Consider, then, the linear equation in H_0

$$L(\lambda_0)x_0 = R(\lambda_0; A)(x_0 + B\hat{u}_{\text{opt}}(\lambda_0)(x_0)) = w \tag{30}$$

for $w \in \mathcal{D}(A)$ and $\lambda_0 \notin \sigma(A)$. Then

$$x_0 + B\hat{u}_{\text{opt}}(\lambda_0)(x_0) = (\lambda_0 - A)w. \tag{31}$$

Assuming (without loss of generality) that B is "full rank", recall that

$$\pi_B = B(B^*B)^{-1}B^* \tag{32}$$

is the projection onto $\mathcal{R}(B)$, and define

$$\pi_B^\perp = I - \pi_B \tag{33}$$

as the complementary projection. Then

$$\pi_B^\perp x_0 = \pi_B^\perp(\lambda_0 - A)w \tag{34}$$

and

$$\pi_B x_0 + B\hat{u}_{\text{opt}}(\lambda_0)(\pi_B x_0) = \pi_B(\lambda_0 - A)w - B\hat{u}_{\text{opt}}(\lambda_0)(\pi_B^\perp(\lambda_0 - A)w). \tag{35}$$

Multiplying by $(B^*B)^{-1}B^*$ results in

$$v + \hat{u}_{\text{opt}}(\lambda_0)(Bv) = (B^*B)^{-1}B^*(\lambda_0 - A)w - \hat{u}_{\text{opt}}(\lambda_0)\pi_B^\perp(\lambda_0 - A)w, \tag{36}$$

where $v = (B^*B)^{-1}B^*x_0$. Solving the above equation for v (the coefficient matrix is invertible for Re (λ_0) sufficiently large, for example) and using

$$x_0 = \pi_B x_0 + \pi_B^\perp x_0 = Bv + \pi_B^\perp(\lambda_0 - A)w$$

gives

$$x_0=(\lambda_0-A)w-B(I+\hat{u}_{\text{opt}}(\lambda_0)B)^{-1}\hat{u}_{\text{opt}}(\lambda_0)((\lambda_0-A)w). \tag{37}$$

($\hat{u}_{\text{opt}}(\lambda_0)B$ here represents the matrix of the linear map $C^m \to C^m$: $v \to \hat{u}_{\text{opt}}(\lambda_0)(Bv)$.) This represents the (unbounded densely defined) inverse

$$x_0=[L(\lambda_0)]^{-1}w, \qquad w \in \mathscr{D}(A),$$

so that (by [10, Thm. 5.8.3]) $L(\lambda)$ is the resolvent of

$$\begin{aligned}\tilde{A}&=(\lambda_0-[L(\lambda_0)]^{-1}),\\ \tilde{A}w&=Aw+B(I+\hat{u}_{\text{opt}}(\lambda_0)B)^{-1}\hat{u}_{\text{opt}}(\lambda_0)((\lambda_0-A)w).\end{aligned} \tag{38}$$

That is,

$$\tilde{A}w=Aw-B[B^*K]w, \tag{39}$$

where the linear mapping $[B^*K]$: $H_0 \to C^m$ is defined as the extension (by continuity) of

$$-(I+\hat{u}_{\text{opt}}(\lambda_0)B)^{-1}\hat{u}_{\text{opt}}(\lambda_0)((\lambda_0-A)w).$$

(At first appearance, the above is only densely defined and unbounded; it is shown in Appendix B that the explicit form of $\hat{u}_{\text{opt}}(\lambda_0)(\cdot)$ guarantees boundedness.)

LEMMA 4. *The operator $A-B[B^*K]$ defined above generates an exponentially stable semigroup of class C_0 in H_0.*

Proof. Consider for each $x_0 \in H_0$ the vector-valued function

$$R(s;A-B[B^*K])x_0=L(s)x_0. \tag{40}$$

It follows from the definition of $\mathscr{D}(T)$ in the derivation of the optimal control that

$$L(s)x_0 \in H^2(\pi^+;H_0),$$

and hence that (for each x_0)

$$\int_0^\infty \|\tilde{S}_t x_0\|^2\,dt<\infty,$$

where $\{\tilde{S}_t\}$ is the semigroup generated by $A-B[B^*K]$. By the results of Datko [11], it follows that $\{\tilde{S}_t\}$ is exponentially stable.

5. Computation of the optimal gains.

It was shown above that the optimal control takes the expected form

$$u_{\text{opt}}(t)(x_0)=-[B^*K]x_{\text{opt}}(t), \tag{41}$$

where the optimal trajectory is given by

$$x_{\text{opt}}(t)=S_{A-B[B^*K]}(t)\cdot x_0. \tag{42}$$

From these it follows that $[B^*K]$ may be evaluated as

$$[B^*K]x_0=\lim_{t\to 0^+} u_{\text{opt}}(t)(x_0). \tag{43}$$

The open-loop optimal control is given in Fourier-transform form by Lemma 1:

$$\begin{aligned}U_{\text{opt}}(\omega)=&-[F^+(i\omega)]^{-1}\\ &\cdot \hat{P}_+\{[F^-(i\omega)]^{-1}\{G^*(i\omega)CR(i\omega;A)x_0+B_u^*(i\omega+A_u^*)^{-1}\alpha(x_0)\}\}.\end{aligned}$$

It follows from the consideration of § 3 above, that

$$[F^{-}(i\omega)]^{-1}\{G^*(i\omega)CR(i\omega;A)x_0\}$$

is the Fourier transform of function in $L_1 \cap L_2 \cap C_0(-\infty, \infty)$. Further,

$$[[F^{-}(i\omega)]^{-1}-I]B_u^*(i\omega+A_u^*)^{-1}\alpha(x_0)$$

has the same properties, while we have

$$\mathscr{F}^{-1}\{B_u^*(i\omega+A_u^*)^{-1}\}=\begin{cases} B_u^* e^{-A_u^* t}, & t>0, \\ 0, & t<0. \end{cases} \tag{44}$$

These observations lead directly to the following theorem.

THEOREM. *The optimal feedback gains for the open-loop unstable linear regulator problem described above are given by*

$$[B^*K]x_0=\frac{1}{2\pi}\int_{-\infty}^{\infty}[F^{-}(i\omega)]^{-1}G^*(i\omega)CR(i\omega;A)x_0\,d\omega$$
$$+\frac{1}{2\pi}\int_{-\infty}^{\infty}\{[F^{-}(i\omega)]^{-1}-I\}B_u^*(i\omega+A_u^*)^{-1}\alpha(x_0)\,d\omega+B_u^*\alpha(x_0).$$

Proof. Since $(F^+)^{-1}$ as an operator represents an identity plus a causal L_1 convolution, it suffices to compute the limit at 0 of the right-continuous function defined as

$$\mathscr{F}^{-1}\{[F^{-}(i\omega)]^{-1}\{G^*(i\omega)CR(i\omega;A)x_0+B_u^*(i\omega+A_u^*)^{-1}\alpha(x_0)\}\}. \tag{45}$$

From considerations similar to those of § 3, it follows that

$$[F^{-}(i\omega)]^{-1}\{G^*(i\omega)CR(i\omega;A)x_0\}$$

belongs to $L_1((-\infty,\infty), d\omega; C^m)$.

From an argument dual to that in [1], it follows that

$$F^+-I=P_+\{G^*G(F^-)^{-1}\}, \tag{46}$$

so that $(F^+-I)\in(L_1\cap L_2)\hat{}$. Since (similarly)

$$P^-((F^+)^{-1}G^*G)=(F^-)^{-1}-I, \tag{47}$$

so that $[F^-(i\omega)]^{-1}-I$ represents a matrix of transforms of functions in $L_1 \cap L_2$. This shows that

$$[[F^{-}(i\omega)]^{-1}-I]B_u^*(i\omega+A_u^*)^{-1}\alpha(x_0)$$

also belongs to $L_1((-\infty,\infty), d\omega; C^m)$. The desired result now follows from Fourier inversion and the direct evaluation of

$$\lim_{t\to 0^+}\mathscr{F}^{-1}\{B_u^*(i\omega+A_u^*)^{-1}\alpha(x_0)\}=B_u^*\alpha(x_0). \tag{48}$$

5.1. Some computational considerations. As described in [5], results of the sort presented above lead to effective computational methods in the case of systems for which the resolvent is compact. Use of an eigenfunction basis for evaluation of the optimal gains leads to evaluation of the spectral factors at points inside the half plane of analyticity. This procedure is inherently more stable in the numerical sense than the evaluation of the boundary values of the factors which may appear to be required for the gain evaluation.

These remarks also apply in the case of the open-loop unstable system considered above. In this case the "unstable poles" of the resolvent also contribute to the residue calculations.

Evaluation of the terms $\alpha(x_0)$ in the optimal gain formula also proceeds on the basis of a residue calculation. This computation requires the "Gram operator" associated with

$$\hat{P}_+\{[F^-(i\omega)]^{-1}B_u^*(i\omega+A_u^*)^{-1}\}.$$

By direct calculation in the case that the eigenvalues of A_u are simple, so that the resolvent may be represented as

$$(i\omega+A_u^*)^{-1}=\sum_{j=1}^{N}\frac{1}{i\omega+\bar{\lambda}_j}\psi_j\varphi_j^*,\quad \operatorname{Re}(\bar{\lambda}_j)>0, \tag{49}$$

the function required for the Gram matrix calculation may be obtained in closed form.

The Laplace transform of the positive projection of the function

$$\mathscr{F}^{-1}\{[F^-(i\omega)]^{-1}B_u^*(i\omega+A_u^*)^{-1}\}$$

may be represented in the form

$$\frac{1}{2\pi i}\int_{-\infty}^{\infty}\frac{1}{s-i\omega}\{[F^-(i\omega)]^{-1}B_u^*(i\omega+A_u^*)^{-1}\}\,d\omega.$$

Using the fact that $[F^-(\cdot)]^{-1}$ (by definition of the spectral factor) extends analytic and uniformly bounded over the left-half plane, the above integral may be evaluated by residues to give

$$\sum_{j=1}^{N}[F^-(-\bar{\lambda}_j)]^{-1}\frac{B_u^*\psi_j\varphi_j^*}{s+\bar{\lambda}_j}.$$

Passing to the boundary value to obtain the Fourier transform representation gives the result

$$\hat{P}_+\{[F^-(i\omega)]^{-1}B_u^*(i\omega+A_u^*)^{-1}\}=\sum_{j=1}^{N}[F^-(-\bar{\lambda}_j)]^{-1}\frac{B_u^*\psi_j\varphi_j^*}{i\omega+\bar{\lambda}_j}.$$

Using this result, the Gram operator required in § 3 above may be explicitly evaluated in terms of the structure of the unstable subsystem and point evaluations of the spectral factor $F^-(\cdot)$.

It is clear that analogous "explicit" calculations are possible in the case of higher multiplicity for the unstable subsystem. The results in this case also require point evaluations of derivatives of the spectral factor (in the region of analyticity of the factor).

6. Conclusions. A Wiener–Hopf derivation of the optimal control law for a class of open-loop unstable distributed systems has been given.

The derivation leads to a closed form representation of the optimal feedback operator, which may be evaluated in terms of the associated spectral factors. For certain problems (for example, those exhibiting spatial symmetries in the control actions) this provides direct computational methods [5].

It may be possible to generalize these results to systems containing more general unstable subsystems than the finite dimensional case considered above. In view of condition (A.5), systems for which the (time reversed) unstable subsystem is exactly reachable in infinite time seem obvious candidates. The computational prospects of such an extension do not appear hopeful.

The cases of A-bounded output mappings and boundary controls may also be considered. In these cases it appears difficult to retain use of the Gohberg–Krein factorization results; this loss substantially complicates the arguments, as properties of the factorization are repeatedly used in the derivations above.

Appendix A. Description of the domain conditions. The basic system description is

$$\frac{dx}{dt}=Ax+Bu(t), \qquad y=Cx \tag{A.1}$$

in differential form; more convenient is the corresponding "mild solution" formulation

$$x(t)=S_tx_0+\int_0^t \dot{S}_{t-\tau}Bu(\tau)\,d\tau, \qquad y=Cx, \tag{A.2}$$

where $\{S_t\}$ denotes the semigroup generated by A.

We assume that the singularities of the resolvent of A in the open right-half plane consist of a finite number of poles which may be separated from the remainder of $\sigma(A)$ by a rectifiable simple contour C (or union of such contours).

As in [6], we construct the projection P

$$P=\frac{1}{2\pi i}\int_C R(\lambda\,;A)\,d\lambda \tag{A.3}$$

and decompose the system into a strictly unstable and complementary subspace:

$$\begin{aligned} x_u(t)&=S_u(t)x_u(0)+\int_0^t S_u(t-\tau)B_uu(\tau)\,d\tau,\\ x_s(t)&=S_s(t)x_s(0)+\int_0^t S_s(t-\tau)B_su(\tau)\,d\tau, \end{aligned} \tag{A.4}$$

where $B_u=PB$, $B_s=(I-P)B$.

We assume that:

(i) $H_u \triangleq PH_0$ is a finite dimensional subspace;

(ii) if C_u represents the restriction of the bounded linear operator C to H_u, that $[A_u, B_u, C_u]$ is a minimal realization;

(iii) the complementary semigroup $\{S_s(t)\}$ is exponentially stable; $\|S_s(t)\|\leqq Me^{-\delta t}$, for some $\delta>0$.

With these assumptions, $\mathscr{D}(T)$ is determined as those pairs $\begin{bmatrix} x_0\\ x(\cdot)\end{bmatrix}$ for which (with $\hat{u}$ denoting the Laplace transform of $u(\cdot)$)

$$R(s\,;A_u)\cdot(Px_0+B_u\hat{u}(s))$$

is analytic at each of the poles of the resolvent $R(s;A_u)$.

This condition may be translated into an orthogonality condition by following a procedure suggested by that used in [4] to determine the range of certain Fredholm operators associated with Wiener–Hopf equations.

Expand

$$R(s\,;A_u)\cdot(Px_0+B_u\hat{u}(s))$$

in a Laurent series about each right-half plane pole of the resolvent. Since the Taylor series coefficients in the expansions

$$\hat{u}(s) = \sum_{n=0}^{\infty} u_n(s-\lambda)^n$$

may be identified with

$$u_n = \int_0^{\infty} \frac{(-t)^n}{n!} e^{-\lambda t} u(t)\, dt,$$

the condition that the above expression be analytic at each of the right-half plane poles may be rewritten in the time domain.

Using the Jordan canonical form for the finite dimensional unstable subsystem (see [8], for example), one verifies that the required condition is equivalent to

$$\text{(A.5)} \qquad Px_0 + \int_0^{\infty} e^{-A_u t} B_u u(t)\, dt = 0.$$

A verbal description of (A.5) is simply that the net equivalent initial excitation on the unstable subspace should vanish.

Selecting a basis $\{\psi i\}_{i=1}^{N}$ for H_u allows one to express the condition (A.5) in terms of $N = \dim \mathrm{H_u}$ orthogonality conditions

$$\text{(A.6)} \qquad \psi_i^* x_0 + \int_0^{\infty} (B_u^* e^{-A_u^* t}\psi_i)^* u(t)\, dt = 0, \qquad i = 1, \cdots, N.$$

That is,

$$\begin{bmatrix} x_0 \\ u(\cdot) \end{bmatrix} \perp \begin{bmatrix} \psi_i \\ B_u^* e^{-A_u^*(\cdot)}\psi_i \end{bmatrix}, \qquad i = 1, \cdots, N.$$

Using Parseval's theorem, (A.5) may also be written in the form

$$\text{(A.7)} \qquad Px_0 + \frac{1}{2\pi}\int_{-\infty}^{\infty} (-i\omega + A_u)^{-1} B_u \hat{u}(i\omega)\, d\omega = 0$$

or

$$\text{(A.8)} \qquad \psi_j^* x_0 + \frac{1}{2\pi}\int_{-\infty}^{\infty} \psi_j^*(-i\omega + A_u)^{-1} B_u \hat{u}(i\omega)\, d\omega = 0, \qquad j = 1, \cdots, N.$$

These computations identify $\mathscr{D}(T)$ as a subspace of finite co-dimension in $H_0 \oplus L_2((0,\infty); C^m)$. The linear mapping T restricted to $\mathscr{D}(T)$ is bounded, and the norm may be estimated by the usual frequency-domain methods. (T acts on $\mathscr{D}(T)$ according to the usual frequency-domain multiplication; the domain restriction ensures exactly that the resulting Fourier transform is the boundary value of a function in the Hardy space $H^2(\pi^+; H_1)$).

Appendix B. Boundedness of the optimal gain. From Lemma 3 it follows that the optimal gain operator, $[B^*K]$ is densely defined by

$$\text{(B.1)} \qquad [B^*K]w = -(I + \hat{u}_{\mathrm{opt}}(\lambda_0)B)^{-1}\hat{u}_{\mathrm{opt}}(\lambda_0)((\lambda_0 - A)w)$$

for $w \in \mathscr{D}(A)$.

Now from Lemma 1,

$$\text{(B.2)} \qquad \begin{aligned} \hat{u}_{\mathrm{opt}}(\lambda_0)(x_0) = -[F^+(\lambda_0)]^{-1}P_+\{[F^-(i\omega)]^{-1}\{G^*(i\omega)CR(i\omega; A)x_0 \\ + B_u^*(i\omega + A_u)^{-1}\alpha(x_0)\}\}(\lambda_0), \end{aligned}$$

so that to verify the existence of a bounded extension of $[B^*K]$ it suffices to consider the mappings

$$w \to P_+\{[F^-(i\omega)]^{-1}G^*(i\omega)CR(i\omega; A)(\lambda_0 - A)w\}(\lambda_0)$$

and

$$w \to \alpha((\lambda_0 - A)w).$$

The explicit form of the first of these is (for Re $(\lambda_0) > 0$)

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} [F^-(i\omega)]^{-1}G^*(i\omega)CR(i\omega; A)(\lambda_0 - A)w\frac{1}{\lambda_0 - i\omega}\,d\omega;$$

the integral is seen to be convergent for $w \in \mathcal{D}(A)$ by considerations similar to those used in the derivation of Lemma 1 (i.e., Parseval's theorem).

Using the fact that for $w \in \mathcal{D}(A)$

$$\text{(B.3)} \qquad R(i\omega; A)(\lambda_0 - A)w = [(\lambda_0 - i\omega)R(i\omega; A) + I]w,$$

the above becomes

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} [F^-(i\omega)]^{-1}G^*(i\omega)CR(i\omega; A)w\,d\omega + P_+\{[F^-(i\omega)]^{-1}G^*(i\omega)Cw\}(\lambda_0).$$

The above integrals are easily seen to be strongly convergent, defining a bounded linear map $H_0 \to C^m$.

Since we have

$$\alpha((\lambda_0 - A)w) = m^{-1}\beta((\lambda_0 - A)w),$$

consider

$$\beta((\lambda_0 - A)w) = P(\lambda_0 - A_0)w - \frac{1}{2\pi}\int_{-\infty}^{\infty} (-i\omega + A^{\cdot}_u)^{-1}B_u[F^+(i\omega)]^{-1}$$
$$\cdot \hat{P}_+\{[F^-(i\omega)]^{-1}G^*(i\omega)CR(i\omega; A)(\lambda_0 - A)w\}\,d\omega.$$

Considerations closely related to those immediately above lead to the conclusion that the integral expression above defines a bounded mapping as required. With regard to the term

$$P(\lambda_0 - A)w,$$

recall that $(\lambda_0 - A)$ commutes with the projection P (on $\mathcal{D}(A)$). Since the restriction of A to the unstable subspace is bounded, this completes the proof that $[B^*K]$ extends to a bounded linear mapping $H_0 \to C^m$.

REFERENCES

[1] J. H. DAVIS, *A distributed filter derivation without Riccati equations*, this Journal, 16 (1978), pp. 584–592.

[2] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.

[3] J. W. HELTON, *A spectral factorization approach to the distributed regular problem; the algebraic Riccati equation*, this Journal, 14 (1976), pp. 639–661.

[4] I. C. GOHBERG AND M. G. KREIN, *Systems of integral equations with kernel depending on the difference of the arguments*, Trans. Amer. Math. Soc., 14 (1960), pp. 217–287.

[5] J. H. DAVIS AND B. M. BARRY, *A distributed model for stress control in multiple locomotive trains*, J. Appl. Math. Optimization, 3 (1977), pp. 163–190.

[6] R. TRIGGIANI, *On the stabilizability problem in Banach space*, J. Math. Anal. Appl., 52 (1975), pp. 383–403.

[7] A. FEINTUCH AND M. ROSENFELD, *On pole assignment for a class of infinite dimensional systems*, this Journal, 16 (1978), pp. 270–276.

[8] C. T. CHEN, *Introduction to Linear System Theory*, Holt, Reinhart and Winston, New York, 1970.

[9] D. G. LUENBERGER, *Optimization by Vector, Space Methods*, John Wiley, New York, 1969.

[10] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semi-Groups*, American Mathematical Society Colloquium Series, Vol. 31, Providence, RI, 1957.

[11] R. DATKO, *Extending a theorem of A. M. Lyapunov to Hilbert space*, J. Math. Anal. Appl., 32 (1970), pp. 610–616.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 6, November 1979

JUMP-DIFFUSION APPROXIMATIONS FOR ORDINARY DIFFERENTIAL EQUATIONS WITH WIDE-BAND RANDOM RIGHT HAND SIDES*

HAROLD J. KUSHNER†

Abstract. Let $y(\cdot)$ be a stationary mixing process and $J^\varepsilon(\cdot)$ an approximation to a random impulsive process. Kurtz's (1975) results on approximation of a general semigroup by a Markov semigroup are used to prove (weak and a similar type of) convergence of the solutions to (1.1) and (1.2) to jumping diffusions. Previous results are generalized in various ways. The case of unbounded $y(\cdot)$ is also treated as is the combined jump-diffusion case. Also, a limit theorem for an integral with respect to "approximate white noise" in terms of an Itô integral is given. The method has the advantages of generality and relative ease of use.

1. Introduction. In [1], Kurtz gave some fairly general semigroup methods for showing convergence of a sequence of non-Markov process to a Markov process, either in the sense of weak convergence or in the sense of convergence of finite dimensional distributions. Let $y(\cdot)$ denote (a Euclidean space valued) right continuous strong mixing stationary process. For each $\varepsilon > 0$, define $y^\varepsilon(t) = y(t/\varepsilon^2)$, and for suitable F, G, define the process $x^\varepsilon(\cdot)$ by

$$(1.1)\qquad \dot{x}_t^\varepsilon = \frac{F(x_t^\varepsilon, y_t^\varepsilon)}{\varepsilon} + G(x_t^\varepsilon, y_t^\varepsilon), \qquad x_0^\varepsilon = x_0 \in R^m,\ y(t) \in R^{m'}.$$

Khazminskii [2], Papanicolaou [3], Papanicolaou and Kohler [4] and Blankenship and Papanicolaou [5] have all treated the problem of weak convergence of $x^\varepsilon(\cdot)$ to a diffusion. The problem is, of course, closely related to the original problem of Wong and Zakai [6]. In this paper, Kurtz's results, (together with a technique exploited in references [3] and [5]) will be used to get similar types of results under conditions which are weaker.[1] The method of proof has the great advantage of being quite straightforward and easy to use for both the diffusion and jump-diffusion cases.

We also treat limits of systems of the type

$$(1.2)\qquad \dot{x}_t^\varepsilon = \frac{F(x_t^\varepsilon, y_t^\varepsilon)}{\varepsilon} + G(x_t^\varepsilon, y_t^\varepsilon) + \sum_i H_i(x_t^\varepsilon) J_{i,t}^\varepsilon,$$

where $\int_0^t J_{i,s}^\varepsilon\, ds$ is an approximation to a pure jump process, and we obtain a limit which is a jumping diffusion.

Sections 2 and 3 recapitulate Kurtz's method of proving convergence of finite dimensional distributions and tightness, resp. The results are recapitulated partly for the sake of self-containment, and partly to state the precise form in which they are to be used. Section 4 states the assumptions used in Section 5, which gives the result for limits of (1.1) when $y(\cdot)$ is bounded, a restriction also used in the past references. Theorem 3 gives a result which is useful in approximating stochastic integrals with respect to a Wiener process by ordinary integrals. Such results are needed for identification and

* Received by the editor October 2, 1978, and in revised form January 30, 1979.

† Divisions of Applied Mathematics and Engineering, Brown University, Providence, Rhode Island 02912. This research was partially supported by the Air Force Office of Scientific Research under AF-AFOSR 76-3063, the National Science Foundation under NSF-Eng73-03846-A-03 and by the Office of Naval Research under N0014-76-C-0279-P0002.

[1] References [2] and [3] allowed F and G to depend also on t. At the expense of extra detail, this case could be handled by our method.

related problems (see Balakrishnan [8], [9]). A result on convergence of finite dimensional distributions for unbounded $y(\cdot)$ is given in Section 6, and tightness for unbounded $y(\cdot)$ is proved in Section 7. Section 8 deals with the relatively simple case where the input is an approximation to a random impulse process, and (1.2), an approximation to a jumping diffusion, is treated in Section 9. The result in Sections 6 and 7 cover the much used case where $y(\cdot)$ is a Gaussian diffusion.

A method similar to that of Section 5 is outlined in Section 4, [3], for the problem of showing convergence of finite dimensional distributions for the bounded $y(\cdot)$ case. The results there are not in a particularly usable form, and actually require more smoothness of F and G than needed here since partial differential equation methods are ultimately used there. Here, we do not need to solve or even to approximate solutions of partial differential equations but merely to check the action of certain operators on smooth functions. In Reference [14] (where the terminology $\hat{C}$ and $\hat{C}_0$ have different meanings), a similar method is used for a different noise structure, the total noise effects on the system being effectively the cumulative results of the effects of a sequence of "small" random variables.

2. Convergence of finite dimensional distributions.[2] Let $(\Omega, P, \mathscr{F})$ denote a probability space, $\{\mathscr{F}_t\}$ a nondecreasing sequence of sub σ-algebras of $\mathscr{F}$, $\mathscr{L}$ the space of progressively measurable real valued processes f on $[0,\infty)$, adapted to $\{\mathscr{F}_t\}$ and such that $\sup_t E|f(t)|<\infty$. Let f_n and f be in $\mathscr{L}$. Define the limit "p-lim" by $p\text{-}\lim f_n = f$ iff $\sup_n \sup_t E|f_n(t)|<\infty$ and $E|f_n(t)-f(t)|\to 0$ for each t. For each $s>0$, define the operator $\mathscr{T}(s)\colon \mathscr{L}\to\mathscr{L}$ by $\mathscr{T}(s)f=$ function in $\mathscr{L}$ whose value at t is the random variable $E_{\mathscr{F}_t} f(t+s)$. There is a version which is progressively measurable [1, Appendix] and we always assume that this is the one which is used. The $\mathscr{T}(s)$, $s\geqq 0$, are a semigroup of linear operators on $\mathscr{L}$. Let $\hat{\mathscr{L}}_0$ denote the subspace of $\mathscr{L}$ of p-right continuous functions. If the limit $p\text{-}\lim_{s\to 0}[1/s(\mathscr{T}(s)f-f)]$ exists and is in $\hat{\mathscr{L}}_0$, we call it $\hat{A}f$ and say that $f\in\mathscr{D}(\hat{A})$. The operators $\mathscr{T}(s)$ and $\hat{A}$ are analogous to the semigroup and weak infinitesimal operator of a Markov process. Among the properties to be used later is ([1], equation (1.9))

$$\mathscr{T}(s)f-f=\int_0^s \mathscr{T}(\tau)\hat{A}f\,d\tau, \qquad f\in\mathscr{D}(\hat{A}), \tag{2.1a}$$

or, equivalently

$$E_{\mathscr{F}_t}f(t+s)-f(t)=E_{\mathscr{F}_t}\int_0^s \hat{A}f(t+\tau)\,d\tau, \text{ for each } t\geqq 0. \tag{2.1b}$$

Equation (2.1b) also holds for uniformly bounded random times s. If, for some process $Z^\varepsilon(\cdot)$, $\mathscr{F}_t=\sigma(Z_s^\varepsilon, s\leqq t)$, we may write[3] $\mathscr{F}_t^\varepsilon$, T_t^ε and $\hat{A}^\varepsilon$ for $\mathscr{F}_t$, $\mathscr{T}(t)$ and $\hat{A}$, resp. Let $\hat{C}$ and $\hat{C}^i$ denote the spaces of real valued functions on R^m which vanish at ∞ and which are continuous and which have continuous partial derivatives up to order i (and which also vanish at infinity), resp. Let $\hat{C}_0$ and $\hat{C}_0^i$ denote the sets of these functions which have compact support.

The following Theorem (a specialization of [1], Theorem 3.11) is our main tool for dealing with (1.1). Henceforth, unless otherwise mentioned, $\varepsilon\to 0$ replaces $n\to\infty$ in p-lim.

[2] From [1], with slightly altered terminology. Sometimes we write f_t and sometimes $f(t)$ for the value of a process f at time t.

[3] The σ-algebras will often be completed, but the same notation will be used.

THEOREM 1. *Let $Z^\varepsilon(\cdot)=x^\varepsilon(\cdot)$, $y^\varepsilon(\cdot)$, $\varepsilon>0$, denote a sequence of $R^{m+m'}$ valued right continuous processes, $x(\cdot)$ a (R^m-valued) Markov process with semigroup T_t mapping $\hat{C}$ into $\hat{C}$ and which is strongly continuous (sup norm) on $\hat{C}$. For some $\lambda>0$ and dense set D in $\hat{C}$, let* Range $(\lambda-A|_D)$ *be dense in $\hat{C}$ (sup norm, A = infinitesimal operator of $x(\cdot)$). Suppose that, for each $f\in D$, there is a sequence $\{f^\varepsilon\}$ of progressively measurable functions adapted to $\{\mathcal{F}_t^\varepsilon\}$ and such that*

$$(2.2)\qquad p\text{-lim}\,[f^\varepsilon-f(x^\varepsilon(\cdot))]=0,$$

$$(2.3)\qquad p\text{-lim}\,[\hat{A}^\varepsilon f^\varepsilon-Af(x^\varepsilon(\cdot))]=0.$$

Then, if $x_0^\varepsilon\to x_0$ weakly, the finite dimensional distributions of $x^\varepsilon(\cdot)$ converge to those of $x(\cdot)$.

Equations (2.2) and (2.3) are equivalent to (the limits are taken for each t as $\varepsilon\to 0$)

$$(2.2)'\qquad \sup_{\varepsilon,t} E|f^3(t)-f(x^\varepsilon(t))|<\infty,\qquad E|f^3(t)-f(x^\varepsilon(t))|\to 0,$$

$$(2.3)'\qquad \sup_{\varepsilon,t} E|\hat{A}^\varepsilon f^\varepsilon(t)-Af(x^\varepsilon(t))|<\infty,\qquad E|\hat{A}^\varepsilon f^\varepsilon(t)-Af(x^\varepsilon(t))|\to 0.$$

3. Tightness. Let $y(\cdot)$, $y^\varepsilon(\cdot)$, $x^\varepsilon(\cdot)$ denote the functions in the model (1.1) or (1.2). *Let $\mathcal{F}_t$ and $\mathcal{F}_t^\varepsilon$ denote the (completed) $\sigma(y_s;\, s\leqq t)$ and $\sigma(y_s^\varepsilon,\, s\leqq t)$. Write E_t and E_t^ε for $E_{\mathcal{F}_t}$ and $E_{\mathcal{F}_t^\varepsilon}$, resp.*

Again, we describe results from [1]. Let $D^m[0,\infty)$ denote the space of R^m valued functions on $[0,\infty)$ which are right continuous on $[0,\infty)$ and have left hand limits on $(0,\infty)$. Note that $x^\varepsilon(\cdot)\in D^m[0,\infty)$ w.p.1. Suppose that the finite dimensional distributions of $x^\varepsilon(\cdot)$ converge to those of a process $x(\cdot)$, where $x(\cdot)$ has paths in $D^m[0,\infty)$ w.p. 1. Then, as noted in [1, bottom of page 628], $\{x^\varepsilon(\cdot)\}$ is tight in $D^m[0,\infty)$ if $\{f(x^\varepsilon(\cdot))\}$ is tight in $D[0,\infty)$ for each $f\in\hat{C}_0$ ($\hat{C}_0$ is used there, but it can be replaced by[4] $\hat{C}_0^3$.) It follows from [1, Theorem 4.20], that $\{f(x^\varepsilon(\cdot))\}$ is tight in $D[0,\infty)$ if $x_0^\varepsilon\to x_0$ weakly and if, for each real $T>0$, there is a random variable $\gamma_\varepsilon(\delta)$ such that

$$(3.1)\qquad E_t^\varepsilon\gamma_\varepsilon(\delta)\geqq E_t^\varepsilon \min\{1,[f(x_{t+u}^\varepsilon)-f(x_t^\varepsilon)]^2\},$$

for all $0\leqq t\leqq T$, $0\leqq u\leqq\delta\leqq 1$, and

$$(3.2)\qquad \lim_{\delta\to 0}\overline{\lim_{\varepsilon\to 0}}\, E\gamma_\varepsilon(\delta)=0.$$

In [1, p. 629], Kurtz suggests a method of getting the $\gamma_\varepsilon(\delta)$. This method is developed in Lemma 1 and is used in the sequel. The f^ε below will be obtained in the same manner as we will obtain the f^ε of Theorem 1. We have ($\|f\|=\sup_x|f(x)|$)

$$(3.3)\qquad E_t^\varepsilon[f(x_{t+u}^\varepsilon)-f(x_t^\varepsilon)]^2\leqq 2\|f\|\,|E_t^\varepsilon f(x_{t+u}^\varepsilon)-f(x_t^\varepsilon)|+|E_t^\varepsilon f^2(x_{t+u}^\varepsilon)-f^2(x_t^\varepsilon)|.$$

LEMMA 1. *Let $f\in\hat{C}_0^3$, and let there be a sequence $\{f^\varepsilon\}$ in $\mathcal{L}$, where $(f^\varepsilon)^i\in\mathcal{D}(\hat{A}^\varepsilon)$, $i=1,2$, and such that, for each real $T>0$ there is a random variable M_ε such that*

$$(3.4)\qquad \sup_{t\leqq T}|f^\varepsilon(t)-f(x_t^\varepsilon)|\to 0\quad w.p.1,\ as\ \varepsilon\to 0,$$

$$(3.5)\qquad \sup_{t\leqq T}|\hat{A}^\varepsilon(f^\varepsilon(t))^i|\leqq M_\varepsilon,\quad w.p.1,\ \mathrm{i}=1,2,$$

and $\sup_\varepsilon P\{M_\varepsilon\geqq N\}\to 0$ *as $N\to\infty$. Then $\{f(x^\varepsilon(\cdot))\}$ is tight in $D[0,\infty)$.*

[4] Or by any set of functions dense in $\hat{C}_0$ in the sup norm.

Proof. By (2.1)

$$(3.6)\quad E_t^\varepsilon (f^\varepsilon(t+u))^i - (f^\varepsilon(t))^i = E_t^\varepsilon \int_0^u \hat{A}^\varepsilon (f^\varepsilon(t+\tau))^i \, d\tau = E_t^\varepsilon \int_t^{t+u} \hat{A}^\varepsilon (f^\varepsilon(v))^i \, dv.$$

If $\sup_\varepsilon EM_\varepsilon < \infty$, then (3.3), (3.5) and (3.6) yield the existence of a $\gamma_\varepsilon(\delta)$ of the desired form (for each interval $[0, T]$) for the sequence of processes $\{f^\varepsilon(\cdot)\}$. Then, by (3.4), we can get the $\gamma_\varepsilon(\cdot)$ of the desired form for the sequence $\{f(x^\varepsilon(\cdot))\}$ which is, consequently, tight.

In the general case, when $EM_\varepsilon \to \infty$, a truncation argument can be used. For each $\delta > 0$, define $T_\delta = \min\{t: |\hat{A}^\varepsilon f^\varepsilon(t)| \geqq 1/\delta$ or $|\hat{A}^\varepsilon (f^\varepsilon(t))^2| \geqq 1/\delta\}$. In (3.6) replace the $(t+u)$ and t on the left and right sides by $(t+u) \cap T_\delta^\varepsilon$ and $t \cap T_\delta^\varepsilon$, resp. A repetition of the argument of the previous paragraph yields tightness of $\{f(x^\varepsilon(\cdot \cap T_\delta^\varepsilon))\}$, hence of $\{f(x^\varepsilon(\cdot))\}$, since $\mathrm{P}\{T_\delta^\varepsilon \to \infty\} = 1$ by hypotheses. Q.E.D.

4. Assumptions for Model (1.1); bounded $y(\cdot)$.

(A1) *$F(\cdot,\cdot)$ and $G(\cdot,\cdot)$ are continuous, and the first x-partial and first and second x-partial derivatives of G and F, resp., are continuous.*

(A2) *There is a constant M such that*

$$|F(x, y)| + |G(x, y)| \leqq M(1 + |x|).$$

(A1) and (A2) assure the global existence of solutions to (1.1).

(A3) *$y(\cdot)$ is a right continuous, bounded stationary process and $EF(x, y_s) = 0$, each x.*

There is a measurable function $\rho(\cdot)$ such that

$$\sup_{B_i, t} |P(B_2|B_1) - P(B_2)| \leqq \rho(\tau),$$

where $B_1 \in \sigma(y_u, u \leqq t)$, $B_2 \in \sigma(y_u, u \geqq t+\tau)$. Let

$$(4.1)\qquad \int_0^\infty \rho^{1/2}(t)\, dt < \infty.$$

Define the operator A on $\hat{C}^2$ by (the subscript x denotes gradient)

$$(4.2)\qquad Af(x) = EG'(x, y_s) f_x(x) + \int_0^\infty d\tau\, EF'(x, y_s)(F'(x, y_{s+\tau}) f_x(x))_x.$$

$$\equiv \sum_i b_i(x) \frac{\partial}{\partial x_i} f(x) + \frac{1}{2} \sum_{i,j} a_{ij}(x) \frac{\partial^2 f(x)}{\partial x_i \partial x_j}.$$

By (A1) and (A3), the integrals on the right exist. In fact, the improper Lebesgue integral is absolutely convergent (i.e. $\int_0^T |E(\cdot)|\, d\tau$ converges) uniformly in x.[5] Furthermore, if $f \in \hat{C}_0^3$ then $Af(x)$ is continuously differentiable in x, and the gradient of $Af(x)$ is the function obtained by simply replacing the argument of E in (4.2) by its x-gradient. If this is done, then the improper Lebesgue integral still is absolutely convergent, uniformly in x. In a paper in preparation, martingale methods are used to obtain similar results without conditions such as (A4).

(A4) *A is the restriction to $\hat{C}_0^2$ of the strong infinitesimal operator of a strong Markov conservative (no finite escape time) diffusion process $x(\cdot)$, with semigroup T_t. T_t maps $\hat{C}$ into $\hat{C}$ and is strongly continuous on $\hat{C}$.*

[5] This follows readily from the strong mixing. It is also a consequence of Billingsley [7], p. 170, by using $EF(x, y_s) \equiv 0$, (4.1) and the fact that the functions have bounded support (let $r \to 1$, $s \to \infty$ in [7], equation (20.23), with a proper identification of ξ, η there). We will use this and similar facts frequently in the sequel.

(A5) *$\{\mathscr{F}_t\}$ is right continuous. That is, $\mathscr{F}_t = \bigcap_{\delta>0}\mathscr{F}_{t+\delta}$, each $t \geqq 0$.*

(A6) *The set $(\lambda - A)\hat{C}_0^3 \equiv \{g\colon g = (\lambda - A)f,\ f \in \hat{C}_0^3\}$ is dense in $\hat{C}$ for some $\lambda > 0$.*

Remark on (A5). Let $f \in \hat{C}_0$ and let $\bar{f}$ denote either $Ff,\cdot Gf$ of any of the g_i or f_i introduced below. Condition (A5) is introduced only because we want $E^\varepsilon_{t+s}\bar{f}(x^\varepsilon_{t+s}, t+s+u)$ to converge to $E^\varepsilon_t\bar{f}(x^\varepsilon_t, t+u)$ in probability as $s\downarrow 0$. Many of the calculations in Theorems 2 and 4 involve this type of right continuity together with uniform integrability.

Remark on (A6). Some condition such as (A6) is required for use of Theorem 1. Let A_c denote the strong infinitesimal operator of T_t acting on $\hat{C}$. Then (A6) is equivalent to A_c being the closure of the operator A (of (4.2) acting on $\hat{C}_0^2$, or on $\hat{C}_0^3$, since $\hat{C}_0^3$ is dense in $\hat{C}_0^2$ in the norm $\|f\|_2 = \sup_x (|f(x)| + |f_x(x)| + |f_{xx}(x)|))$. This condition does not seem to be particularly restrictive. We only remark that it holds in the two extreme cases: (1) the $b_i(\cdot)$ and $a_{ij}(\cdot)$ in (4.2) are bounded, satisfy a uniform Hölder condition and $a(\cdot)$ is uniformly positive definite; (2) where T_t maps $\hat{C}^2$ into $\hat{C}^2$ and is strongly continuous on $\hat{C}^2$ with respect to the norm $\|f\|_2$ defined above. The same remarks were made by Kurtz [1, p. 632].

In case (2), we can actually consider T_t as acting on the Banach space $\hat{C}^2$ with norm $\|f\|_2$, and where $f(x)$ and its first and second derivatives go to zero as $|x| \to \infty$, and modify (A4) accordingly. In this case, the closure of the operator A (the domain of A is $\hat{C}_0^2$ here) is just the strong infinitesimal operator (of T_t) acting on its domain in $\hat{C}^2$. Suppose that there is a matrix valued $\sigma(\cdot)$ such that $\sigma(x)\sigma'(x) = a(x)$, that (A4) holds (as modified above) and that $b_i(\cdot)$, $\sigma_{ij}(\cdot)$ are locally Lipschitz. Then (see remarks below on bounding), it is enough to prove Theorem 2 under the additional condition that $b_i(\cdot)$, $\sigma_{ij}(\cdot)$ are bounded and, for arbitrary N, arbitrarily smooth out of the sphere S_N of center 0 and radius N. Assume that $b_i(\cdot)$ and $\sigma_{ij}(\cdot)$ have continuous first and second derivatives. Then by the above remark on bounding, we can assume that the coefficients are bounded on R^m. Then (Gikhman and Skorokhod [13, Chap. 8.4], case (2) above holds. The conditions imposed are weaker than those in [4] when F and G do not depend on t.

Remark on (A6) *and time dependent coefficients.* Let us elaborate the case of the last paragraph when the coefficients in A are time dependent. In particular, let $a(\cdot,\cdot) = \sigma(\cdot,\cdot)\sigma'(\cdot,\cdot)$, where $\sigma(\cdot,\cdot)$ and $b(\cdot,\cdot)$ are bounded and continuous, together with their first and second partial x-derivatives and first t-derivative. Theorem 2 can be proved under conditions close to (A1)–(A6). We only replace A by the operator $(A + \partial/\partial t)$ acting on its domain in $\hat{\mathscr{C}}$, where $\hat{\mathscr{C}}$ is defined to be the closure under uniform convergence of $\hat{\mathscr{C}}_0$, the set of bounded continuous functions on $[0,\infty)\times R^r$ with compact support. Here, $\hat{\mathscr{C}}$ replaces $\hat{C}$. For $u(\cdot,\cdot) \in \hat{\mathscr{C}}_0^{1,2}$, define $f(\cdot,\cdot)$ by

$$f(t,x) = \int_0^\infty E_{t,x}u(t+s, x_{t+s})\, e^{-\lambda s}\, ds.$$

Then $(A + \partial/\partial t - \lambda)f = -u$ and $|f(x,t)|$ (and its first and second x-derivatives and first t-derivative) goes to zero as $|x| \to \infty$, uniformly in t. For each $\delta > 0$, there is a $f^\delta(\cdot,\cdot) \in \hat{\mathscr{C}}_0^{1,2}$ such that $|(A + \partial/\partial t - \lambda)(f - f^\delta)| \leqq \delta$, which yields (A6) in this case.

Remark on bounding the coefficients. Suppose $a(x) = \sigma(x)\sigma'(x)$, where $b_i(\cdot)$ and $\sigma_{ij}(\cdot)$ are locally Lipschitz continuous and $x(\cdot)$, the diffusion with generator A, is conservative. Define an N-truncation as follows. Let $b_i^N(\cdot)$, $\sigma_{ij}^N(\cdot)$ equal $b_i(\cdot)$, $\sigma_{ij}(\cdot)$, resp., in S_N, be bounded on R^m, have bounded derivatives of any desired order in the complement of S_{2N} and be at least as smooth in $S_{2N} - S_N$ as $b_i(\cdot)$, $\sigma_{ij}(\cdot)$ are. Then the Itô process $x^N(\cdot)$ with coefficients $b_i^N(\cdot)$, $\sigma_{ij}^N(\cdot)$ is called an N-truncation of $x(\cdot)$ if the

$b^N(\cdot)$ and $\sigma^N(\cdot)$ can be obtained by a modification of F and G in $R^m - S_N$. N-truncations always exist since we can multiply F and G by suitable smooth real valued functions $m_f(x)$ and $m_g(x)$, resp., which equal unity in S_N. We only need prove Theorem 2 and verify (A4) and (A6) for some N-truncation for each N.

The proof of the assertion will be omitted. It is essentially a note that the parts of $x^N(\cdot)$ and $x(\cdot)$ up to the first escape from S_N are equal, and that the probability of escape from S_N on an interval $[0, T]$ goes to zero as $N \to \infty$, for each fixed x_0. Here, we suppose that $x(\cdot)$ are defined with respect to the same Wiener process.

5. Proof of weak convergence; bounded $y(\cdot)$ and (1.1). The main job in using Theorem 1 is to get f^ε when f is given. To do this, we use an idea exploited for a similar purpose in § 3 of [5] and in § 4 of [3]. We look for functions of the form[6] $f^\varepsilon(x, t) = f(x) + \varepsilon f_1^\varepsilon(x, t) + \varepsilon^2 f_2^\varepsilon(x, t)$, and define $f^\varepsilon(x_t^\varepsilon, t) = f^\varepsilon(t)$. Define operators $\hat{A}_x^\varepsilon$ and $\hat{A}_y^\varepsilon$ as follows. Let $g(x, t)$ be smooth as a function of x and such that $g(x, t)$ is a function of y_s^ε, $s \leqq t$, for each x. At $x = x_t^\varepsilon$, $y = y_t^\varepsilon$, let

$$\hat{A}_x^\varepsilon g(x, t) = g_x'(x, t)\left[\frac{F(x, y)}{\varepsilon} + G(x, y)\right]; \tag{5.1}$$

i.e., $\hat{A}_x^\varepsilon$ is $\hat{A}^\varepsilon$, but acting on $g(x_t^\varepsilon, t)$ considered only a function of its first argument. Let $\hat{A}_y^\varepsilon g(x, t)$ be $\hat{A}^\varepsilon g(x, t)$, but where g is considered to be a function of its second argument only. Assume for the moment that $\hat{A}^\varepsilon = \hat{A}_x^\varepsilon + \hat{A}_y^\varepsilon$. Then, in order to use Theorem 1, we apply $\hat{A}^\varepsilon$ to $f^\varepsilon(x, t)$, and insist that

$$\begin{aligned} &[f_x(x) + \varepsilon f_{1,x}^\varepsilon(x, t) + \varepsilon^2 f_{2,x}^\varepsilon(x, t)]'\left[\frac{F(x, t)}{\varepsilon} + G(x, t)\right] \\ &\quad + \hat{A}_y^\varepsilon(\varepsilon f_1^\varepsilon(x, t) + \varepsilon^2 f_2^\varepsilon(x, t)) = Af(x, t) + O(\varepsilon), \end{aligned} \tag{5.1$'$}$$

where equation (5.1)′ must determine both the operator A and equations yielding the $f_i^\varepsilon(x, t)$. In Theorem 2, we merely write down formulas for the $f_i^\varepsilon(x, t)$ and verify the conditions of Theorem 1.

THEOREM 2. *Under* (A1) *to* (A6), $x^\varepsilon(\cdot)$ *converges weakly in* $D^m[0, \infty)$ *to the diffusion* $x(\cdot)$, *with initial condition* x_0.

Proof. First (Parts 1 to 3) we prove convergence of finite dimensional distributions, using Theorem 1. In Part 4 tightness is proved, via Lemma 1. Henceforth, f is a fixed element of $\hat{C}_0^3$. Since $\hat{C}_0^3$ is dense in $\hat{C}_0^2$ in the norm $\|f\|_2$, it is enough to work with $\hat{C}_0^3$.

Part 1. $f_1^\varepsilon(x, t)$ is defined to be a solution (suggested by (5.1)′) to $\hat{A}_y^\varepsilon f_1^\varepsilon(x, t) = -g_1(x, y_t^\varepsilon) \equiv -F'(x, y_t^\varepsilon) f_x(x)$. More precisely, define $f_1^\varepsilon(t) = f_1^\varepsilon(x_t^\varepsilon, t)$, where

$$\begin{aligned} f_1^\varepsilon(x, t) &= \int_0^\infty E_t^\varepsilon g_1\left(x, y\left(\frac{t}{\varepsilon^2} + s\right)\right) ds \\ &= \frac{1}{\varepsilon^2} \int_0^\infty E_t^\varepsilon g_1(x, y_{t+s}^\varepsilon)\, ds \end{aligned} \tag{5.2}$$

(both forms will be used). The improper Lebesgue integral exists and is bounded and absolutely convergent, uniformly in ω, x and in t in bounded sets, by the strong mixing (A3), and the facts that $EF(x, y_s) \equiv 0$ and that g_1 has compact x-support. Furthermore, there are versions of $f_1^\varepsilon(x, t)$ and $f_1^\varepsilon(x_t^\varepsilon, t)$ which are progressively measurable.

[6] For each x and t, $f_i^\varepsilon(x, t)$ will be a function of y_s^ε, $s \leqq t$. The discussion in this paragraph is purely formal.

We next show that $f_1^\varepsilon(t) \in \mathscr{D}(\hat{A}^\varepsilon)$. We have

$$\hat{A}^\varepsilon f_1^\varepsilon(t) = p-\lim_{\delta\to 0} [E_t^\varepsilon f_1^\varepsilon(x_{t+\delta}^\varepsilon, t+\delta) - f_1^\varepsilon(x_t^\varepsilon, t)]/\delta$$

$$(5.3) \qquad = p-\lim_{\delta\to 0} [E_t^\varepsilon \{f_1^\varepsilon(x_{t+\delta}^\varepsilon, t+\delta) - f_1^\varepsilon(x_t^\varepsilon, t+\delta)\}]/\delta$$

$$+ p-\lim_{\delta\to 0} [E_t^\varepsilon f_1^\varepsilon(x_t^\varepsilon, t+\delta) - f_1^\varepsilon(x_t^\varepsilon, t)]/\delta$$

if the limits exist and are in $\hat{\mathscr{L}}_0$. It is easy to verify that the second limit exists, is in $\hat{\mathscr{L}}_0$ and equals $-g_1(x_t^\varepsilon, y_t^\varepsilon)/\varepsilon^2$. Now, $f_1^\varepsilon(x, t)$ is differentiable in x. Indeed,

$$f_{1,x}^\varepsilon(x, t) = \frac{1}{\varepsilon^2}\int_0^\infty E_t^\varepsilon g_{1,x}(x, y_{t+s}^\varepsilon)\, ds,$$

since $\int_0^T |E_t^\varepsilon g_{1,x}(y_{t+s}^\varepsilon)|\, ds$ converges uniformly in x, and in t in bounded sets, as $T \to \infty$. This fact together with the representation

$$(5.4) \qquad \begin{aligned} &E_t^\varepsilon[f_1^\varepsilon(x_{t+\delta}^\varepsilon, t+\delta) - f_1^\varepsilon(x_t^\varepsilon, t+\delta)]/\delta \\ &= \frac{1}{\delta}\int_0^\delta E_t^\varepsilon[f_{1,x}^\varepsilon(x_{t+u}^\varepsilon, t+\delta)]'\left[\frac{F(x_{t+u}^\varepsilon, y_{t+u}^\varepsilon)}{\varepsilon} + G(x_{t+u}^\varepsilon, y_{t+u}^\varepsilon)\right] du \end{aligned}$$

and the facts that $f_{1,x}^\varepsilon$ is right continuous in the mean at t, and that the integrand is zero out of a bounded x, y range, can be used to show that the first limit on the right side of (5.3) exists, is in $\hat{\mathscr{L}}_0$ and equals ($x = x_t^\varepsilon$, $y = y_t^\varepsilon$)

$$(5.5) \qquad [f_{1,x}^\varepsilon(x, t)]'[F(x, y)/\varepsilon + G(x, y)] = \hat{A}_x^\varepsilon f_1^\varepsilon(x, t).$$

Part 2. With $x = x_t^\varepsilon$, we will define $f_2^\varepsilon(t) = f_2^\varepsilon(x, t)$, where $f_2^\varepsilon(x, t)$ is the formal solution to

$$(5.6) \qquad \begin{aligned} \hat{A}_y^\varepsilon f_2^\varepsilon(x, t) &= -g_2(x, t) \\ &\equiv -[F'(x, y_t^\varepsilon) f_{1,x}^\varepsilon(x, t) + G'(x, y_t^\varepsilon) f_x(x) - Af(x)], \end{aligned}$$

where A is defined in (4.2). More precisely, *define* f_2 by

$$(5.7) \qquad f_2^\varepsilon(x, t) = \frac{1}{\varepsilon^2}\int_0^\infty E_t^\varepsilon g_2(x, t+s)\, ds.$$

There are versions of $f_2^\varepsilon(x, t)$ and $f_2^\varepsilon(x_t^\varepsilon, t)$ which are progressively measurable. *We now ignore the G terms*, for the difficulty lies with the $Ff_{1,x}^\varepsilon$ term. The improper Lebesgue integral $\int |E_t^\varepsilon g_2(x, t+s)|\, ds$ of (5.7) converges absolutely, uniformly in ω, x and in t in any bounded set, by the strong mixing property and the definition of A. (Indeed, this is the reason for the choice of A.) To see this, note that (5.7) (without the G-terms) equals

$$(5.8) \qquad \frac{1}{\varepsilon^2}\int_0^\infty ds\, \{E_t^\varepsilon F'(x, y_{t+s}^\varepsilon) f_{1,x}^\varepsilon(x, t+s) - EF'(x, y_{t+s}^\varepsilon) f_{1,x}^\varepsilon(x, t+s)\},$$

where the average value of the coefficient of ds is zero (by stationarity and the definition of A), and use the strong mixing condition. In fact, using the change of variables $s/\varepsilon^2 \to s'$, it can be seen from (5.8) and the strong mixing and compact x-support of f_1^ε that $|f_2^\varepsilon(t)|$ is bounded w.p.1, uniformly in x and ω and in bounded t intervals. Furthermore, $g_2(x, t)$ is continuously differentiable in x. In fact, the convergence assertion in the sentence above (5.4) also holds for g_2 replacing g_1 and

$$f_{2,x}(x,t)=\frac{1}{\varepsilon^2}\int_0^\infty E_t^\varepsilon g_{2,x}(x,t+s)\,ds.$$

Expression (5.3) holds *with* f_2^ε *replacing* f_1^ε if the limits exist and are in $\hat{\mathscr{L}}_0$. Again, we readily verify that the second limit in (5.3) exists, is in $\hat{\mathscr{L}}_0$ and equals $-g_2(x,t)/\varepsilon^2(x=x_t^\varepsilon)$. An argument similar to that leading to (5.5) yields that the first limit in (5.3) exists, is in $\hat{\mathscr{L}}_0$ and equals (5.5) with f_2^ε replacing f_1^ε.

Part 3. Now, we apply Theorem 1. Since

$$\sup_{t,\varepsilon>0} E(|f_1^\varepsilon(t)|+|f_2^\varepsilon(t)|)<\infty,$$

we have

$$p\text{-lim}\,[f^\varepsilon-f(x^\varepsilon(\cdot))]=0.$$

Now, calculate $\hat{A}^\varepsilon f^\varepsilon$. By Parts 1 and 2, with $x=x_t^\varepsilon$, $y=y_t^\varepsilon$,

$$\begin{aligned}\hat{A}^\varepsilon f^\varepsilon(x,t)&=\hat{A}^\varepsilon f(x)+\varepsilon\hat{A}^\varepsilon f_1^\varepsilon(x,t)+\varepsilon^2\hat{A}^\varepsilon f_2^\varepsilon(x,t)\\&=[F(x,y)/\varepsilon+G(x,y)]'f_x(x)\\&\quad+\varepsilon[-F'(x,y)f_x(x)/\varepsilon^2+(F(x,y)/\varepsilon+G(x,y))'f_{1,x}^\varepsilon(x,t)]\\&\quad+\varepsilon^2\Big[-\frac{1}{\varepsilon^2}\{F'(x,y)f_{1,x}^\varepsilon(x,t)+G'(x,y)f_x(x)-Af(x)\}\\&\qquad\qquad+(F(x,y)/\varepsilon+G(x,y))'f_{2,x}^\varepsilon(x,t))\Big]\\&=Af(x)+\varepsilon[G'(x,y)f_{1,x}^\varepsilon(x,t)+F'(x,y)f_{2,x}^\varepsilon(x,t)]\\&\quad+\varepsilon^2G'(x,y)f_{2,x}^\varepsilon(x,t).\end{aligned}\tag{5.9}$$

We now readily verify that $p\text{-lim}\,[\hat{A}^\varepsilon f^\varepsilon-Af(x^\varepsilon(\cdot))]=0$. Since $x_0^\varepsilon=x_0$, Theorem 1 implies that the finite dimensional distributions converge.

Part 4. *Tightness.* For tightness, we use Lemma 1. Each $f\in\hat{C}_0$ can be approximated uniformly arbitrarily closely by an $f\in\hat{C}_0^3$. Thus, by the lemma and discussion preceding it, we only need prove that $\{f(x^\varepsilon(\cdot))\}$ is tight for each $f\in\hat{C}_0^3$. Let $f\in\hat{C}_0^3$ and construct $f_1^\varepsilon, f_2^\varepsilon$ exactly as the $f_1^\varepsilon, f_2^\varepsilon$ were constructed above, and define $f^\varepsilon(t)$ as above. Note that $[(f^\varepsilon(t+\delta))^2-f^\varepsilon(t)]^2/\delta=[f^\varepsilon(t+\delta)+f^\varepsilon(t)][f^\varepsilon(t+\delta)-f^\varepsilon(t)]/\delta$. Now use the facts that f^ε is uniformly bounded and in $\mathscr{D}(\hat{A}^\varepsilon)$ and the right continuity properties of f^ε (as implied by (A3), (A5)) to get that $(f^\varepsilon)^2\in\mathscr{D}(\hat{A}^\varepsilon)$ and $\hat{A}^\varepsilon(f^\varepsilon(t))^2=2f^\varepsilon(t)\hat{A}^\varepsilon f^\varepsilon(t)$. There is a *constant* M such that w.p.1

$$\sup_{\varepsilon>0,\omega,t}|\hat{A}^\varepsilon(f^\varepsilon(t))^i|\leqq M,\qquad i=1,2,$$

$$\sup_t|f^\varepsilon(t)-f(x^\varepsilon(t))|\to 0,\quad\text{as }\varepsilon\to 0.$$

Thus, Lemma 1 implies that $\{f(x^\varepsilon(\cdot))\}$ is tight in $D[0,\infty)$, for each $f\in\hat{C}_0^3$; hence $\{x^\varepsilon(\cdot)\}$ is tight in $D^m[0,\infty)$. The tightness, together with the convergence of finite dimensional distributions implies weak convergence. Q.E.D.

An approximation to an integral. In problems where changes of measure via Girsanov-like transformations are involved, such as occur in some identification problems (Balakrishnan [8], [9]), we need to get limits of integrals such as $z_t^\varepsilon=\int_0^t q'(x_s^\varepsilon)(y_s^\varepsilon/\varepsilon)\,ds$, as $\varepsilon\to 0$, where $q(\cdot)$ is some given function.

Let $\hat{C}_0^i$ denote the real valued functions on $R^{m+m'+1}$ with compact support, whose ith partial derivatives are continuous. Let $Y_t^\varepsilon = \int_0^t (y_s^\varepsilon/\varepsilon)\, ds$ and $u^\varepsilon = (x^\varepsilon, Y^\varepsilon, z^\varepsilon)$. Then u_t^ε is $R^{m+m'+1}$ valued and

$$\dot{u}^\varepsilon = \tilde{G}(x^\varepsilon, y^\varepsilon) + \frac{1}{\varepsilon}\tilde{F}(x^\varepsilon, y^\varepsilon),$$

where

$$\tilde{G}(x, y) = (G(x, y), 0, 0),$$
$$\tilde{F}(x, y) = (F(x, y), y, q'(x)y).$$

The remarks below $(A6)$ *all pertain here also.*

THEOREM 3. *Let* $q(\cdot)$ *satisfy the conditions on* $F(\cdot)$ *in* (A1)–(A2). Assume (A1) *to* (A6) *where* (A4) *and* (A6) *hold for the process* $u(\cdot)$ *and operator* $\bar{A}$ *defined on* $\hat{C}_0^2$ *by*

$$\bar{A}f(u) = E\tilde{G}'(x, y_0)f_u(u) + \int_0^\infty d\tau\, E\tilde{F}'(x, y_0)(\tilde{F}'(x, y_\tau)f_u(u))_u. \tag{5.10}$$

Then $u^\varepsilon(\cdot)$ *converges weakly in* $D^{m+m'+1}[0, \infty)$ *to* $u(\cdot) = (x(\cdot),\ Y(\cdot),\ z(\cdot))$, $u_0 = (x_0, 0, 0)$, *a diffusion with generator* $\bar{A}$ *on* $\hat{C}_0^2$. *The process* Y_t *is a Brownian motion with covariance* $R = \int_{-\infty}^\infty Ey_t y_0'\, d\tau$. *The limit* $z(\cdot)$ *has the Itô representation* (*the expectation is over* $y(\cdot)$) *in terms of the limits* $x(\cdot)$, $Y(\cdot)$

$$dz = q'(x)\, dY + \left[\int_0^\infty EF'(x, y_0)(q'(x)y_\tau)_x\, d\tau\right] dt. \tag{5.11}$$

(The last term in (5.11) is the so-called correction term of the limiting integral approximation.)

The proof of weak convergence of $u^\varepsilon(\cdot)$ to $u(\cdot)$ is simply an application of Theorem 2. Once the weak convergence is known, then the representation (5.11) is not hard to get, and we omit the details.

6. Unbounded $y(\cdot)$ and (1.1). Our approach here will be only a little different from that in § 5. In order to avoid conditions which look overly complicated, we specialize $F(x, y)$ and $G(x, y)$ to $F(x)y$ and $G(x)$, resp.

Assumptions. In this section, the convergence of finite dimensional distribution is proved, and tightness is treated in the next section. Owing to the unboundedness of $y(\cdot)$, it is convenient to artificially bound the F and G. We do this by dealing with a sequence of approximations to the original processes. The operator A is still defined by (4.2).

(B1) $F(\cdot)$ *and its first and second order partial derivatives are continuous.*

(B2) $G(\cdot)$ *and its first order partial derivatives are continuous.*

(B3) $\{\mathcal{F}_t\}$ (*again completed*) *is right continuous, and so is the stationary process* $y(\cdot)$, *w.p.*1. (*See the remark concerning* (A5).)

Define

$$v_t^\varepsilon = \frac{1}{\varepsilon^2}\int_0^\infty E_t^\varepsilon y_{t+s}^\varepsilon\, ds,\ v_t = \int_0^\infty E_t y_{t+s}\, ds.$$

(B4) *For some* $\rho > 0$, $\sup_t E(\int_0^\infty |E_t y_{t+s}|\, ds)^{2+\rho} < \infty$.

Thus, v_t^ε and v_t are well-defined.

(B5) $E|y_t|^{2+\rho} < \infty$, *some* $\rho > 0$.

(B6) $\sup_t E(\int_0^\infty ds\, |E_t y_{t+s} v_{t+s}' - Ey_{t+s} v_{t+s}'|)^{2+\rho} < \infty$, *some* $\rho > 0$.

Note that $Ey_{t+s}v'_{t+s} = Ey_{t+s}\int_0^\infty E_{t+s}y'_{t+s+u}\,du = \int_0^\infty Ey_ty'_{t+u}\,du$ and does not depend on t or s, and is well-defined by (B4), (B5).

Owing to the special form of F and G, there are locally Lipschitz $b(\cdot)$ and $\sigma(\cdot)$ such that (see (4.2)) $\sigma(\cdot)\sigma'(\cdot) = a(\cdot)$. See also the remarks on bounding below (A6). x^N and $x^{\varepsilon,N}(\cdot)$ denote the N-truncations of $x(\cdot)$ and $x^\varepsilon(\cdot)$.

(B7) *For a sequence $N \to \infty$, there are N-truncations which satisfy* (A4) *and* (A6).

Remark on assumptions (B3)–(B6). Let $w(\cdot)$ denote a vector Brownian motion, Q a matrix with eigenvalues in the open left half plane and let D and G be matrices. Define processes $Y(\cdot)$ and $y(\cdot)$ by

$$\begin{aligned} dY &= QY\,dt + D\,dw, \\ y(\cdot) &= GY(\cdot). \end{aligned} \tag{6.1}$$

Then (B3) to (B6) hold. In this case, we can let $\mathcal{F}_t^\varepsilon$ and $\mathcal{F}_t$ measure Y_s, $s \leqq t/\varepsilon^2$, and Y_s, $s \leqq t$, resp., in all the foregoing. Then v_t is proportional to Y_t and $|E_ty_{t+s}| \leqq |Y_t|c_1e^{-c_2s}$, where the c_i are positive constants.

Theorem 4 deals with the convergence of finite dimensional distributions.

THEOREM 4. *Under* (B1) *to* (B7) *and the first sentence of* (A4), *the finite dimensional distributions of $x^\varepsilon(\cdot)$ converge to those of $x(\cdot)$ with initial condition x_0, as $\varepsilon \to 0$.*

Proof. If the finite dimensional distributions of $x^{\varepsilon,N}(\cdot)$ converge to those of $x^N(\cdot)$ (initial condition x_0) as $\varepsilon \to 0$ for a sequence $N \to \infty$, then the conservative and the strong Markov properties of (A4), (B7) yield the theorem. So we only need prove convergence for a fixed N. Consequently, we may assume that F and G are bounded and drop the affixes N.

The details are very similar to those of Theorem 2, and we only make a few remarks. As before, define f_1^ε and f_2^ε by

$$f_1^\varepsilon(t) = f_1^\varepsilon(x_t^\varepsilon, t), \qquad f_2^\varepsilon(t) = f_2^\varepsilon(x_t^\varepsilon, t),$$

where $f_i^\varepsilon(x, t)$ is defined as in Theorem 2. These functions are no longer bounded, but still $\sup_{t,\varepsilon>0} E|f_i^\varepsilon(t)| < \infty$. It is rather straightforward to verify (in fact, easier than in Theorem 2 owing to the special form of $F(x, y)$ and $G(x, y)$ here) via (B1) to (B6), that $f_i^\varepsilon \in \mathcal{D}(\hat{A}^\varepsilon)$ and[7] take the same values as in Theorem 2. Furthermore, the expectations of the absolute values of the coefficients of ε and ε^2 on the far right side of (5.9) are bounded, uniformly in ε. Also

$$\sup_{t,\varepsilon>0} E\{|\hat{A}^\varepsilon f_i^\varepsilon(t)| + |f_i^\varepsilon(t)|\} < \infty.$$

Thus,

$$p\text{-}\lim\,[f^\varepsilon - f(x^\varepsilon(\cdot))] = 0 \quad \text{and} \quad p\text{-}\lim\,[\hat{A}^\varepsilon f^\varepsilon - Af(x^\varepsilon(\cdot))] = 0,$$

from which the theorem follows, by Theorem 1. Q.E.D.

7. Tightness; unbounded $y(\cdot)$. Owing to the unboundedness, it is more difficult to prove tightness via the method of Lemma 1. To avoid (what are at the moment) awkward conditions, we suppose that $y(\cdot)$ satisfies (6.1). Then the f^ε can be explicitly evaluated and the proof is easy.

THEOREM 5. *Assume* (B1), (B2), (B7), *the first sentence of* (A4) *and that $y(\cdot)$ satisfies* (6.1). *Then $\{x^\varepsilon(\cdot)\}$ is tight and converges weakly to $x(\cdot)$ as $\varepsilon \to 0$.*

[7] For example, to show that the expression for $\hat{A}^\varepsilon f_i^\varepsilon(t)$ is in $\hat{\mathcal{L}}_0$, we note that the compact support of f and (B4) to (B6) imply uniform integrability of the expression. This, together with (B3) yields p-right continuity.

Remark. The tightness argument only uses the compact support of f, (B1)–(B2) and (6.1).

Proof. The method and notation of Lemma 1 and Theorem 2, Part 4, will be used here. We need to show that, for each T and each $f \in \hat{C}_0^3$

(i) $\lim_{\varepsilon\to 0} \sup_{t\leqq T} |f(x_t^\varepsilon) - f^\varepsilon(t)| = 0$ w.p.1;

(ii) $(f^\varepsilon)^2 \in \mathscr{D}(\hat{A}^\varepsilon)$;

(iii) $\overline{\lim}_{\varepsilon\to 0} \sup_{t\leqq T} |\hat{A}^\varepsilon(f^\varepsilon(t))^j| < \infty$, $j = 1, 2$, w.p.1.

In our case, there is a matrix C_0 such that

$$f_1^\varepsilon(x, t) = [F(x)C_0 Y(t/\varepsilon^2)]' f_x(x),$$

$$g_2^\varepsilon(x, t) = \{F(x)y(t/\varepsilon^2)\}'\{[F(x)C_0 Y(t/\varepsilon^2)]' f_x(x)\}_x + G'(x) f_x(x) - Af(x).$$

$f_2^\varepsilon(x, t)$ is a quadratic form in the components of $Y(t/\varepsilon^2)$ where the coefficients are bounded differentiable functions of x with compact support. Also, as is readily verifiable, $(f^\varepsilon(t))^2 \in \mathscr{D}(\hat{A}^\varepsilon)$, and $\hat{A}^\varepsilon(f^\varepsilon(t))^j$, $j = 1, 2$, have terms in powers of the components of $Y(t/\varepsilon^2)$ up to $2j + 1$. Thus, to verify (i) and (iii) it is enough to verify that for each $T > 0$

$$\lim_{\varepsilon\to 0} \sup_{t\leqq T} \varepsilon |Y(t/\varepsilon^2)|^3 = 0 \quad \text{w.p.1.} \tag{7.1}$$

Equation (7.1) holds since, for each $\alpha > 0$, the Gaussianess, stationarity, and special form (6.1) imply that there are finite w.p.1 ω-functions C_1 and C_2 such that $|Y(t)| \leqq C_1 t^\alpha + C_2$ for all t, w.p.1. Q.E.D.

8. An approximate jump case; (1.2) with no $y^\varepsilon(\cdot)$ term. Since the classical papers of Wong and Zakai [6], the problem of using Itô or other types of equations to approximately model processes which are the solutions to ordinary differential equations has received much attention; e.g., [3]–[5] and §§ 5–7 above. Alternative approaches have been taken by McShane [10] and Sussmann [11] who sought either a theory of integration or a differential equation and a topology on the input functions so that the output is a continuous function of the input. The differential equations were of the so-called Stratonovich form, which, in fact, are precisely Itô equations with appropriate dynamic terms.

Little has been done when the input is an approximation to an impulse (its integral is an approximation to a pure jump process). Marcus [12] has done some work along McShane's "belated integral" point of view. This work [12] has some interesting aspects, but also a number of shortcomings. The dynamics are rather special, being (in part) analytic functions. This is a disadvantage in any approximation theory, where robustness is a key word. Some heavy lie algebra machinery was used, and the form of the results tended to obscure the basic simplicity of the problem. Also, a very particular impulse approximation was used (piecewise constant). In this section, we take a simple minded but inherently natural and robust approach, using pathwise approximations and limits. The limits being either ordinary or stochastic differential equations with impulsive or jump inputs.

Let $\bar{N}_i(ds \times dy)$, $i = 1, \cdots, k$, denote a sequence of scalar valued random measures and define $N_i(t) = \iint_0^t \alpha \bar{N}_i(ds \times d\alpha)$, where $N_i(\cdot)$ is taken to be right continuous. The range of the jumps of $N_i(\cdot)$ is a bounded set R_i. Each $N_i(\cdot)$ is assumed to have a finite number of jumps on each bounded interval w.p.1, and the probability is zero that different $N_i(\cdot)$ have simultaneous jumps.

In this section, we deal with the equation

$$\dot{x}_t^\varepsilon = G(x_t^\varepsilon) + \sum_{i=1}^{k} H_i(x_t^\varepsilon) J_i^\varepsilon(t), \tag{8.1}$$

where the input $J_i^\varepsilon(t)$ is an "approximation" to the impulse $\dot{N}_i(t)$. Figure 1 illustrates some ways in which an actual integrated input $\int_0^t J_i^\varepsilon(s)\,ds$ might approximate an ideal integrated impulsive input $N_i(t)$. In the figure a jump of Y occurs at $t = t_0$. With approximation (1), $J_i(t) = Y/\varepsilon$ on $[t_0, t_0+\varepsilon]$. Define $K_i(x, y)$ to be such that $x + K_i(x, Y)$ is the solution to $\dot{x} = H_i(x)Y$ at $t = 1$, with $x_0 = x = x(0)$.

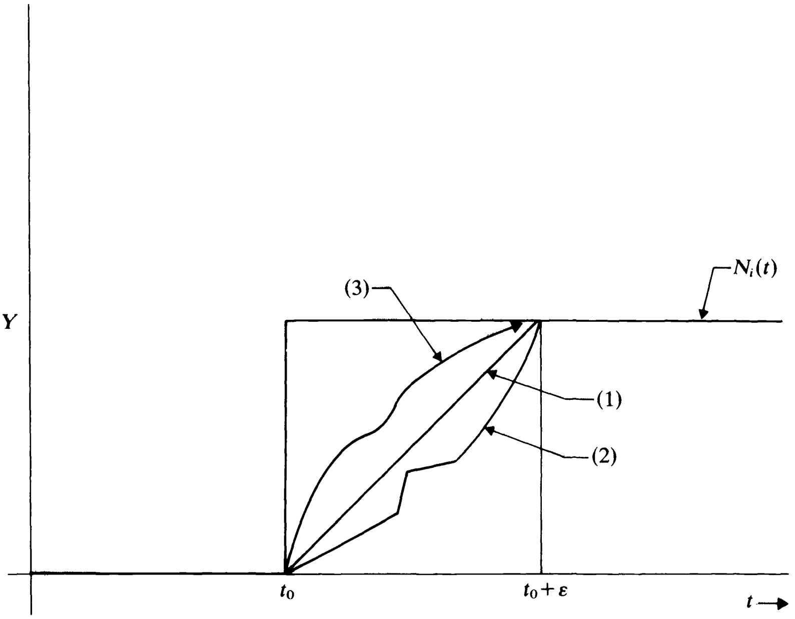

FIG. 1. *Illustration of* 3 *possibilities for* $\int_0^t J_i^\varepsilon(s)\,ds$.

It is convenient to start with the vector of *ideal* integrated inputs $N(\cdot) = \{N_i(\cdot), i \leqq k\}$ and to get the *actual* inputs J_t^ε from this, as indicated in Figure 1. When the parameter is ε, we work only with paths for which the interjump times of the vector $\{N_i(\cdot), i = 1, \cdots, k\}$ are $\geqq \varepsilon$. Obviously, this involves neglecting a set of paths whose probability goes to zero as $\varepsilon \to 0$, and it has no effect on the limiting process. Thus, for our "limit" results, only the case of one input need be treated, and unless noted otherwise set $k = 1$ and drop the indices i on J, N, H and K. Let $N(\cdot)$ jump Y_j at $t = \sigma_j$, its jth jump time. Define $p_t^\varepsilon = J_t^\varepsilon / Y_j$ on $[\sigma_j, \sigma_j + \varepsilon]$, and equal to zero out of $\bigcup_j [\sigma_j, \sigma_j + \varepsilon]$ and define $P_t^\varepsilon = \int_0^t p_s^\varepsilon\,ds$. Thus $P_{\sigma_j+\varepsilon}^\varepsilon - P_{\sigma_j}^\varepsilon = 1$. Now (8.1) can be rewritten in the form

$$\dot{x}_t^\varepsilon = \begin{cases} G(x_t^\varepsilon) + H(x_t^\varepsilon) Y_j p_t^\varepsilon & \text{on } [\sigma_j, \sigma_j+\varepsilon], \\ G(x_t^\varepsilon) & \text{otherwise.} \end{cases} \tag{8.2}$$

The value of $p^\varepsilon(\cdot)$ can depend on the jump time and size and on the state prior to the jump (and on the index i in case (8.1)).

Assumptions.

(C1) *The* G *and* H_i *are continuous.*

(C2) *There is one and only one solution to* $\dot{x} = G(x)$, *for each* x_0 *in* R^m; *for each* $T < \infty$, *the solution is bounded on* $[0, T]$ *uniformly in bounded* x_0 *sets.*

(C3) *For each i, there is one and only one solution on* $[0, 1]$ *to* $\dot{w} = H_i(w)Y$ *for each* $Y \in R_i$ *and* $w_0 \in R^m$. *This solution is bounded, uniformly on bounded* (w_0, Y) *sets.*

(C4) *There are real numbers* $m_\varepsilon > 0$, $M_\varepsilon < \infty$, *such that* $m_\varepsilon \leqq p^\varepsilon(t) \leqq M_\varepsilon$ *on the* $[\sigma_j, \sigma_j + \varepsilon]$ *intervals and* $p^\varepsilon(\cdot)$ *is continuous.*

(C5) *Let* $0 < \mu^\varepsilon(t)$, *where* $\mu^\varepsilon(\cdot)$ *is bounded and continuous on* $[0, 1]$ *and* $\int_0^1 \mu^\varepsilon(s)\, ds \leqq \varepsilon$. *Define* $w^\varepsilon(\cdot)$ *by* $\dot{w}^\varepsilon = G(w)\mu^\varepsilon(t) + H_i(w)Y$. *For each i, let* $w^\varepsilon(\cdot)$ *exist and be bounded on* $[0, 1]$ *uniformly on bounded* (Y, w_0^ε) *sets and in* $\{\varepsilon, \mu^\varepsilon(\cdot), \varepsilon \leqq \varepsilon_0\}$ *for some* $\varepsilon_0 > 0$.

Remarks. (C2) implies that $x(\cdot)$ is continuously dependent on x_0, and (C3) implies that (for each i) $w(\cdot)$ is continuously dependent on w_0, Y. (C2), (C3) and (C5) are partially redundant, but it seemed easier to state the assumptions in this way.

Let $\dot{w} = H_i(w)Y$ and $w_0^\varepsilon \to w_0$. Then (C5) and (C3) imply that $|w^\varepsilon(t) - w(t)| \to 0$ as $\varepsilon \to 0$ uniformly on $[0, 1]$ and on bounded (Y, w_0) sets, and in $\{\mu^\varepsilon(\cdot)\}$.

THEOREM 6. *Assume* (C1) *to* (C5). *Let* $x(\cdot)$ *be defined by*

$$x_t = x_0 + \int_0^t G(x_s)\, ds + \sum_i \int_{R_i} K_i(x_{s^-}, \alpha)\bar{N}_i(d\alpha \times ds). \tag{8.3}$$

Let $x_0^\varepsilon \equiv x_0$. *Then for each* $T < \infty$,

$$\sup_{t \in T_\varepsilon} |x_t^\varepsilon - x_t| \to 0 \quad as\ \varepsilon \to 0,\ w.p.1,$$

where $T_\varepsilon = [0, T] - \bigcup_j [\sigma_j, \sigma_j + \varepsilon]$.

Remarks. Equation (8.3) is the correct limit equation—the analogue of the Wong–Zakai or Stratonovich equation for the modeling of the output of a system with approximate jump inputs. The sequence $\{x^\varepsilon(\cdot)\}$ does not converge to $x(\cdot)$ in the Skorokhod topology since $x(\cdot)$ is discontinuous and the $x^\varepsilon(\cdot)$ are continuous.

A main feature of the theorem is the *robustness* of the result; *the limit does not depend on the precise form of the approximations* $J_i^\varepsilon(\cdot)$. Obviously, the interpolation need not be over only an ε-interval.

Proof. We need treat only *one* jump and *one* H_i term, owing to the assumptions on $N_i(\cdot)$ and on the continuity with respect to parameters implied by (C2), (C3) and (C5). So return to (8.2) with σ_j set equal to zero.

We change the time scale. Define a monotone increasing function on $[0, \varepsilon]$, $\tau(t) = \tau$ by $d\tau/dt = p_t^\varepsilon$ or $\tau(t) = \int_0^t p_s^\varepsilon\, ds$. Thus, $\tau(\varepsilon) = 1$, and the inverse $t(\tau)$ exists by (C4). Define $z^\varepsilon(\tau) = x^\varepsilon(t(\tau))$. Then

$$\frac{dz^\varepsilon(\tau)}{d\tau} = G(z^\varepsilon(\tau))\mu^\varepsilon(\tau) + H(z^\varepsilon(\tau))Y, \qquad \tau \leqq 1, \tag{8.4}$$

where

$$\mu^\varepsilon(\tau) = [p^\varepsilon(t(\tau))]^{-1} = \frac{dt(\tau)}{d\tau},$$

$$z^\varepsilon(0) = x^\varepsilon(0).$$

Now $\int_0^1 \mu^\varepsilon(\tau)\, d\tau = \varepsilon$ and $\mu^\varepsilon(\cdot)$ satisfies the conditions in (C5). Let $x^\varepsilon(0) \to x(0)$ as $\varepsilon \to 0$. Then $x^\varepsilon(\varepsilon) \to K(x(0), Y) + x(0)$ as $\varepsilon \to 0$. This, together with the continuity conditions (C2), (C3), (C5) and a concatenation of the argument, implies the theorem. Q.E.D.

9. The jump-diffusion case (1.2); bounded $y(\cdot)$. We now return to the full model (1.2). Owing to the nonweak convergence of $x^\varepsilon(\cdot)\to x(\cdot)$ in the pure jump case, (see the remark below Theorem 6), the combined jump-diffusion case will be treated by a "piecing together" argument. Here $\bar{N}_i(\cdot)$, $i=1,\cdots,k$, are independent Poisson random measures with rates $\lambda_i>0$, jump distributions $D_i(\cdot)$ (with bounded support R_i) and are *independent* of $y(\cdot)$. With A defined by (4.2), define the operator A_J on $\hat{C}_0^2$ by

$$A_Jf(x)=Af(x)+\sum_i \lambda_i\int[f(x+K_i(x,\alpha))-f(x)]D_i(d\alpha). \tag{9.1}$$

Let $\bar{x}(\cdot)$ denote the jump-diffusion process whose infinitesimal operator on $\hat{C}_0^2$ is A_J. Except for the nonconvergence problem at the jump points, $x^\varepsilon(\cdot)$ will essentially converge weakly to $\bar{x}(\cdot)$.

Set $\sigma_0=0$ and $\sigma_q=q$th jump of the vector valued process $N(\cdot)=\{N_i(\cdot), i\leqq k\}$. Let the qth jump be a jump of $N_{l_q}(\cdot)$ and have value Y_q. Define x_0^ε to be the solution of (1.1) with initial condition x_0 and let $x_q^\varepsilon(\cdot)(q\geqq 1)$ be the solution to (1.1) with initial condition $x^\varepsilon(\sigma_q+\varepsilon)$, and where $y^\varepsilon(\sigma_q+\varepsilon+\cdot)$ is used in lieu of $y^\varepsilon(\cdot)$. Let $x_q(\cdot)$ denote the diffusion of §§ 4 and 5, with initial condition x_0 if $q=0$, and $x_q(0)=x_{q-1}(\sigma_q-\sigma_{q-1})+K_{l_q}(x_{q-1}(\sigma_q-\sigma_{q-1}), Y_q)$ in general. We will need either (D1) or (D2) to replace (C5). Let $A_i(Y)$ denote the operator on $\hat{C}_0^1$ functions which is defined by $(H_i(w)Y)'(\partial/\partial w)$.

(D1) *$\mu^\varepsilon(s)/\varepsilon$ is bounded in $s\leqq 1$ and ε. The functions F, G and H_i are bounded.*

(D2) *$\mu^\varepsilon(s)/\varepsilon$ is bounded in $s\leqq 1$ and ε. Each $A_i(Y)(Y\in R_i)$ is the strong infinitesimal operator of a conservative Markov semigroup mapping* (*and strongly continuous on*) *$\hat{C}$ into $\hat{C}$. Also $(\lambda-A_i(Y))\hat{C}_0^2$ is dense in $\hat{C}$ for each $Y\in R_i$ and i, and some $\lambda>0$* (*which can depend on Y and i*).

THEOREM 7. *Assume* (A1) *to* (A6), (C1), (C4) *and either* (D1) *and* (C3) *or* (D2). *Then for each N, $\{x_q^\varepsilon(\cdot), q\leqq N\}$ converges to $\{x_q(\cdot), q\leqq N\}$ weakly in $D^{mN}[0,\infty)$.*

Note. The remarks below (A6) apply here also.

Proof. Owing to the independence of $N(\cdot)$ and $y(\cdot)$ and right continuity of $y(\cdot)$, $y(\sigma_j+\varepsilon+\cdot)$ has the properties of $y(\cdot)$. Due to this independence, the independent increments property of $N(\cdot)$ and the uniqueness and strong Markov property of the $x(\cdot)$ of §§ 4 and 5 we can use a "piecing together" method based on the following assertion: Let a component $N_{l_1}(\cdot)$ jump Y at $t=\sigma$, with no other jumps on $[\sigma-\varepsilon,\sigma+\varepsilon]$, and let $x^\varepsilon(\sigma)\equiv\tilde{x}^\varepsilon(\sigma)\to\tilde{x}(\sigma)$ weakly as $\varepsilon\to 0$ and define $\tilde{x}^\varepsilon(\cdot)$ for $t\in(\sigma,\sigma+\varepsilon]$ by

$$\dot{\tilde{x}}_s^\varepsilon=G(\tilde{x}_s^\varepsilon, y_s^\varepsilon)+\frac{1}{\varepsilon}F(\tilde{x}_s^\varepsilon, y_s^\varepsilon)+H_{l_1}(\tilde{x}_s^\varepsilon)p_s^\varepsilon Y. \tag{9.2}$$

Then (to be proved) $\tilde{x}^\varepsilon(\sigma+\varepsilon)$ converges weakly to $\tilde{x}(\sigma)+K_{l_1}(\tilde{x}(\sigma), Y)$, as $\varepsilon\to 0$. We will prove only the assertion, and the proof uses a combination of the ideas in Theorems 2 and 6. For notational simplicity, let $\sigma=0$ and drop the index l_1.

As in Theorem 6, change the time scale by defining $\tau(\cdot)$ on $[0,\varepsilon]$ and $w^\varepsilon(\cdot)$ by $d\tau(t)/dt=p^\varepsilon(t)$, $w^\varepsilon(\tau)=\tilde{x}^\varepsilon(t(\tau))$, where $t(\cdot)$ is the inverse of $\tau(\cdot)$. Then

$$\begin{aligned}\frac{dw^\varepsilon(\tau)}{d\tau}&=G(w^\varepsilon(\tau), y^\varepsilon(t(\tau)))\mu^\varepsilon(\tau)+\frac{\mu^\varepsilon(\tau)}{\varepsilon}F(w^\varepsilon(\tau), y^\varepsilon(t(\tau)))\\&\quad+H(w^\varepsilon(\tau))Y, \qquad w^\varepsilon(0)=\tilde{x}^\varepsilon(0),\end{aligned} \tag{9.3}$$

where $\mu^\varepsilon(\tau)=[p^\varepsilon(t(\tau))]^{-1}=dt(\tau)/d\tau$. We need only show that, for fixed Y, $w^\varepsilon(1)$ converges weakly to $\tilde{x}(0)+K(\tilde{x}(0), Y)$, the value at $t=1$ of the solution $w(\cdot)$ to $\dot{w}=H(w)Y$, $w(0)=\tilde{x}(0)$, $\sigma=0$. We will actually prove the stronger result of weak convergence of $w^\varepsilon(\cdot)$ to $w(\cdot)$ in $D^m[0, 1]$, for each fixed nonrandom Y.

First, the proof under (D2) will be given. Using the method of Theorem 2, let $f \in \hat{C}_0^2$ and set $\mu^\varepsilon(\tau) = 0$ and $t(\tau) = t(1) = \varepsilon$ for $\tau \geqq 1$ and define $f_1(w, \tau)$ by $(\tau \leqq 1)$

$$f_1^\varepsilon(w, \tau) = \frac{1}{\varepsilon^2} \int_0^\infty E_{t(\tau)}^\varepsilon F'(w, y^\varepsilon(t(\tau + s)))\mu^\varepsilon(\tau + s) f_w(w)\, ds.$$

In the definition of T_τ^ε and $\hat{A}^\varepsilon$, use $F_{t(\tau)}^\varepsilon$ just as F_τ^ε was used in §§ 3 to 5. Set $f^\varepsilon(\tau) = f(w^\varepsilon(\tau)) + \varepsilon f_1^\varepsilon(w^\varepsilon(\tau), \tau)$. Then, it can readily be shown that $f(w^\varepsilon(\tau))$ and $f_1^\varepsilon(w^\varepsilon(\tau), \tau)$ are in the domain of $\hat{A}^\varepsilon$ and that

$$\begin{aligned}\hat{A}^\varepsilon f^\varepsilon(\tau) = f'_w(w^\varepsilon(\tau))[&G(w^\varepsilon(\tau), y^\varepsilon(t(\tau)))\mu^\varepsilon(\tau)\\ &+F(w^\varepsilon(\tau), y^\varepsilon(t(\tau)))\frac{\mu^\varepsilon(\tau)}{\varepsilon} + H(w^\varepsilon(\tau))Y]\\ &-F'(w^\varepsilon(\tau), y^\varepsilon(t(\tau)))f_w(w^\varepsilon(\tau))\frac{\mu^\varepsilon(\tau)}{\varepsilon} + \varepsilon f_{1,w}^{\varepsilon'}(w^\varepsilon(\tau), \tau)\dot{w}^\varepsilon(\tau),\end{aligned}$$

which equals $A(Y)f(w^\varepsilon(\tau)) + O(\varepsilon)$ (we dropped the l_1 subscript on $A(Y)$). This yields convergence of the finite dimensional distributions of $w^\varepsilon(\cdot)$ to those of $w(\cdot)$, as in Theorem 2. Tightness is also proved in the same way as done in Theorem 2, completing the proof under (D2). Note that the function f_2^ε, which we used in Theorem 2, is not needed here.

Now, we prove the assertion under (D1) and (C3). In this case $|\dot{w}^\varepsilon(\tau)|$ is bounded uniformly in $\varepsilon > 0$ and $\tau \leqq 1$ and $w^\varepsilon(0) \to \tilde{x}(0)$ weakly. Thus $\{w^\varepsilon(\cdot)\}$ is tight in $C^m[0, 1]$ and so is the function with values

$$\int_0^t \left[\frac{\mu^\varepsilon(\tau)}{\varepsilon} F(w^\varepsilon(\tau), y^\varepsilon(t(\tau))) + G(w^\varepsilon(\tau), y^\varepsilon(t(\tau)))\mu^\varepsilon(\tau)\right] d\tau.$$

Consequently, drawing a convergent subsequence and indexing it also by ε we have that $w^\varepsilon(\cdot)$ converges weakly to a continuous process $\bar{w}(\cdot)$. By using a Skorokhod imbedding technique, we can assume for our purposes that the convergence is w.p.1, uniformly on bounded intervals and write

$$(9.4) \qquad \bar{w}(v) = \tilde{x}(0) + \int_0^v H(\bar{w}(s))Y\, ds + \lim_\varepsilon \int_0^v \frac{\mu^\varepsilon(\tau)}{\varepsilon} F(w^\varepsilon(\tau), y^\varepsilon(t(\tau)))\, d\tau.$$

We wish to show that the limit in (9.4) is zero w.p.1 for each v. If so, then since it is continuous w.p.1, it must be identically zero w.p.1. Then, by the uniqueness (C3), $\bar{w}(\cdot) = w(\cdot)$ and the proof will be concluded.

The limit equals

$$\lim_\varepsilon \int_0^v \frac{\mu^\varepsilon(\tau)}{\varepsilon} F(\bar{w}(\tau), y^\varepsilon(t(\tau)))\, d\tau$$

by the continuity of F and boundedness of $y(\cdot)$. Let $\alpha > 0$. Define $\bar{w}^\alpha(t) = \bar{w}(m\alpha)$ on $[m\alpha, m\alpha + \alpha)$ for each integer m. The difference between the last limit and $\lim_\varepsilon \int_0^v (\mu^\varepsilon(\tau)/\varepsilon)F(\bar{w}^\alpha(t), y^\varepsilon(t(\tau)))\, d\tau$ goes to zero as $\alpha \to 0$. Thus, we need only show that $\lim_\varepsilon \int_{m\alpha}^{m\alpha+\alpha} (\mu^\varepsilon(\tau)/\varepsilon)F(\bar{w}(m\alpha), y^\varepsilon(t(\tau)))\, d\tau$ is zero w.p.1 for each α. Now, by changing the time scale back to the original one, the last limit equals $\lim_\varepsilon \int_{t(m\alpha)}^{t(m\alpha+\alpha)} (F(\bar{w}(m\alpha), y^\varepsilon(u))/\varepsilon)\, du$. Since $t(n\alpha) \to 0$ for each n as $\varepsilon \to 0$, the results of Theorem 2 imply that this last limit is zero w.p.1. Q.E.D.

REFERENCES

[1] T. G. KURTZ, *Semigroups of conditional shifts and approximation of Markov processes*, Ann. Prob., 4 (1975), pp. 618–642.

[2] R. Z. KHASMINSKII, *A limit theorem for solutions of differential equations with random right hand sides*, Theory of Prob. and Appl., 11 (1966), pp. 390–406.

[3] G. C. PAPANICOLAOU, *Some probabilistic problems and methods in singular perturbations*, Rocky Mountain J. Math., 6 (1976), pp. 653–674.

[4] G. C. PAPANICOLAOU AND W. KOHLER, *Asymptotic theory of mixing stochastic ordinary differential equations*, Comm. Pure and Appl. Math., 27 (1974), pp. 641–668.

[5] G. BLANKENSHIP AND G. C. PAPANICOLAOU, *Stability and control of stochastic systems with wide-band noise disturbances*, SIAM J. Appl. Math., 34 (1978), pp. 437–476.

[6] E. WONG AND M. ZAKAI, *On the relationship between ordinary and stochastic differential equations*, Internat. J. Engin. Science, 3 (1965), pp. 213–229.

[7] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley and Sons, New York, 1968.

[8] A. V. BALAKRISHNAN, *Identification of systems subject to random state disturbance*, Report UCLA Engin. 7348, 1973.

[9] ———, *On the approximation of Itô integrals using band-limited processes*, Report UCLA Engin. 7342, 1973.

[10] E. J. MCSHANE, *Stochastic Calculus and Stochastic Models*, Academic Press, New York, 1974.

[11] H. J. SUSMANN, *On the gap between deterministic and ordinary differential equations*, Ann Prob., 6 (1978), pp. 19–41.

[12] S. I. MARCUS, *Modelling and analysis of stochastic differential equations driven by point processes*, IEEE Trans. Information Theory, IT-24 (1978), pp. 164–172.

[13] I. I. GIKHMAN AND A. V. SKOROKHOD, *Introduction to the Theory of Random Processes*, Saunders, Philadelphia, 1965.

[14] H. J. KUSHNER, *Approximations of solutions to differential equations with random inputs by diffusion processes*, January 1979, Bonn Conference on Stochastic Control; proceedings to be published in Lecture Notes in Mathematics, Springer-Verlag, Berlin.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 6, November 1979

0363-0129/79/1706-0006 $01.00/0

NONLINEAR PERTURBATION OF LINEAR PROGRAMS*

O. L. MANGASARIAN† AND R R. MEYER†

Abstract. The objective function of any solvable linear program can be perturbed by a differentiable, convex or Lipschitz continuous function in such a way that (a) a solution of the original linear program is also a Karush–Kuhn–Tucker point, local or global solution of the perturbed program, or (b) each global solution of the perturbed problem is also a solution of the linear program.

We are concerned here with the linear program

$$\text{(1)} \qquad \text{Minimize } px \quad \text{subject to} \quad Ax \geqq b,$$

where p and b are given vectors in R^n and R^m respectively and A is a given $m \times n$ real matrix. We shall assume throughout this work that this problem has a nonempty optimal solution set $\bar{S} \subset S = \{x | Ax \geqq b\}$. We shall be interested in the perturbed problem $P(\varepsilon)$ definied as follows:

$$\text{(2)} \qquad \text{Minimize } px + \varepsilon f(x) \quad \text{subject to} \quad Ax \geqq b,$$

where f: $R^n \rightarrow R$ and ε is a nonnegative real number. For convenience we define the optimal solution set of (2) as $\bar{S}(\varepsilon)$. Note that $\bar{S}(0) = \bar{S}$. Perturbed problems such as (2) are considered in [3], [4]. In [3] it was shown that if (1) has a *unique* solution $\bar{x}$ and f is a differentiable function at $\bar{x}$, then there exists a positive $\bar{\varepsilon}$ such that for all ε in $[0, \bar{\varepsilon}]$, $\bar{x}$ satisfies the Karush–Kuhn–Tucker conditions [1], [2] for the perturbed problem (2). By considering a specific perturbation $f(x) = \frac{1}{2}x^Tx$ in [4] an iterative technique is proposed for solving linear programming problems. In this work we show that, under suitable conditions, given by f there exists a positive number $\bar{\varepsilon}$ such that some solution of the linear program is a Karush–Kuhn–Tucker point or a local or global solution of the perturbed problem (2) for ε in the interval $[0, \bar{\varepsilon}]$. In Theorem 1 we show that if f is differentiable and has a bounded level set on $\bar{S}$ then there exists a Karush–Kuhn–Tucker point of the perturbed problem (2) which also solves the original linear program (1). In Theorem 2 we indicate how the same type of perturbation applies to a nonlinear programming problem. The rest of the paper is again devoted to the perturbed linear program. In Theorem 3 we show that if f satisfies a local Lipschitz or local convexity property then there exists a solution of the linear program (1) which is a local solution of the perturbed problem (2). Among other things Theorem 4 globalizes the result of Theorem 3 and shows that for sufficiently small $\varepsilon \geqq 0$ the set of optimal solutions of the perturbed problem is actually a subset of the solutions of (1). Corollary 1 deals with the case when the linear program (1) has a unique solution, while Corollary 2 treats the case when the perturbation function f is strictly convex on R^n. We begin with the first result.

THEOREM 1. *Let f be a function from R^n into R which is differentiable on the nonempty solution set $\bar{S}$ of* (1). *Let either the level set $L = \{x | x \in \bar{S}, f(x) \leqq \beta\}$ be nonempty and bounded for some real number β, or let $\bar{\theta}$ be the minimum value of* (1) *and let the nonlinear program*

$$\text{(3)} \qquad \text{Minimize } f(x) \quad \text{subject to} \quad Ax \geqq b, \quad px \leqq \bar{\theta}$$

have a Karush–Kuhn–Tucker point. Then there exists an $\bar{x}$ in R^n and an $\bar{\varepsilon} > 0$ such that

* Received by the editors November 8, 1978 and in revised form March 25, 1979.

† Computer Sciences and Industrial Engineering Departments, University of Wisconsin—Madison, Madison, Wisconsin 53706. This research was supported in part by National Science Foundation Grant MCS74-20584 A02.

for each ε in $[0, \bar{\varepsilon}]$ there exists a $\bar{u}(\varepsilon)$ in R^m such that $(\bar{x}, \bar{u}(\varepsilon))$ is a Karush–Kuhn–Tucker point of the perturbed problem (2), *and $\bar{x}$ is also a solution of the linear program* (1). *If in addition f is convex or pseudoconvex at $\bar{x}$, then $\bar{x}$ solves the perturbed problem* (2) *for ε in $[0, \bar{\varepsilon}]$.*

Proof. By explicit assumption or by the boundedness of L, problem (3) has a Karush–Kuhn–Tucker point $(\bar{x}, \bar{v}, \bar{\gamma})$ in R^{n+m+1} which satisfies

$$\begin{aligned} \nabla f(\bar{x}) - A^T\bar{v} + \bar{\gamma}p &= 0,\\ A\bar{x} &\geqq b,\\ p\bar{x} &= \bar{\theta},\\ \bar{v}(A\bar{x}-b) &= 0,\\ \bar{v}, \bar{\gamma} &\geqq 0. \end{aligned} \tag{4}$$

Since $\bar{x}$ is also a solution of the linear program (1), there exists a $\bar{w}$ in R^m such that

$$-A^T\bar{w} + p = 0, \quad A\bar{x} \geqq b, \quad \bar{w}(A\bar{x}-b) = 0, \quad \bar{w} \geqq 0. \tag{5}$$

Case 1: $\bar{\gamma} = 0$. From (4) and (5) we have that for any $\varepsilon \geqq 0$

$$\begin{aligned} \varepsilon\nabla f(\bar{x}) - A^T(\bar{w}+\varepsilon\bar{v}) + p &= 0,\\ A\bar{x} &\geqq b,\\ (\bar{w}+\varepsilon\bar{v})(A\bar{x}-b) &= 0,\\ \bar{w}+\varepsilon\bar{v} &\geqq 0. \end{aligned}$$

Hence $(\bar{x}, \bar{w}+\varepsilon\bar{v})$ is a Karush–Kuhn–Tucker point of (2) for any $\varepsilon \geqq 0$.

Case 2: $\bar{\gamma} > 0$. When $\bar{\gamma} > 0$, it follows from (4) that $(\bar{x}, \bar{u} = \bar{v}/\bar{\gamma})$ is a Karush–Kuhn–Tucker point of (2) with $\varepsilon = \bar{\varepsilon} = 1/\bar{\gamma}$. From (4) and (5) we have for $\bar{\gamma} > 0$ and $\lambda \in [0, 1]$ that

$$\begin{aligned} \frac{\lambda}{\bar{\gamma}}\nabla f(\bar{x}) - A^T\left((1-\lambda)\bar{w} + \lambda\frac{\bar{v}}{\bar{\gamma}}\right) + p &= 0,\\ A\bar{x} &\geqq b,\\ \left((1-\lambda)\bar{w} + \lambda\frac{\bar{v}}{\bar{\gamma}}\right)(A\bar{x}-b) &= 0,\\ (1-\lambda)\bar{w} + \lambda\frac{\bar{v}}{\bar{\gamma}} &\geqq 0. \end{aligned}$$

Hence $(\bar{x}, (1-\lambda)\bar{w} + \lambda(\bar{v}/\bar{\gamma}))$ is a Karush–Kuhn–Tucker point for (2) for $\varepsilon = \lambda\bar{\varepsilon} = \lambda/\bar{\gamma}$ and $\lambda \in [0, 1]$.

The last statement of the theorem follows from the standard sufficiency theory of nonlinear programming [2, Thm. 10.1.2]. □

We can apply the same proof technique above to a considerably more general problem than (1), namely to the nonlinear programming problem:

$$\text{Minimize } \theta(x) \quad \text{subject to} \quad g(x) \leqq 0, \quad h(x) = 0, \tag{6}$$

where θ, g and h are functions from R^n into R, R^m and R^k respectively. However because of a constraint qualification restriction the results apply to a narrow class outside linear programs. Hence we shall merely state the result and omit the proof which is quite similar to the proof of Theorem 1. We shall again associate with (6) a perturbed problem, namely for some $\varepsilon \geqq 0$

$$\text{(7)} \qquad \text{Minimize } \theta(x)+\varepsilon f(x) \quad \text{subject to} \quad g(x) \leqq 0, \quad h(x)=0,$$

where f is from R^n into R. We shall assume that (6) has a local solution at $\tilde{x}$ with minimum value of $\bar{\theta}=\theta(\tilde{x})$ and that B is the open ball with center $\tilde{x}$ such that $\theta(\tilde{x}) \leqq \theta(x)$ for all x in B satisfying the constraints $g(x) \leqq 0$ and $h(x)=0$. We further admit the possibility of the nonuniqueness of $\tilde{x}$ and define $\tilde{S}=\{x \mid \theta(x)=\bar{\theta},\ g(x) \leqq 0,\ h(x)=0,\ x \in B\}$.

THEOREM 2. *Let* (6) *have a nonempty set* $\tilde{S}$ *of local optimal solutions satisfying a constraint qualification. Let* θ, g, h *and* f *be differentiable on* $\tilde{S}$ *and let the nonlinear program*

$$\text{(8)} \qquad \textit{Minimize } f(x) \quad \textit{subject to} \quad \begin{aligned} &g(x) \leqq 0,\\ &h(x)=0,\\ &\theta(x) \leqq \bar{\theta}, \quad x \in B \end{aligned}$$

have a Karush–Kuhn–Tucker point $(\bar{x}, \bar{v}, \bar{s}, \bar{\gamma})$ *in* $R^{n+m+k+1}$ *Then there exists an* $\bar{\varepsilon}>0$ *such that for each* ε *in* $[0, \bar{\varepsilon}]$ *there exists a* $(\bar{u}, \bar{r}):[0, \bar{\varepsilon}] \rightarrow R^{m+k}$ *such that* $(\bar{x}, \bar{u}(\varepsilon), \bar{r}(\varepsilon))$ *is a Karush–Kuhn–Tucker point for the perturbed problem* (7), *and* $\bar{x}$ *is also a local solution of the nonlinear program* (6). *In fact it is possible to take* $\bar{\varepsilon}=1/\bar{\gamma}$ *when* $\bar{\gamma}>0$ *and* $\bar{\varepsilon}$ *as any positive number when* $\bar{\gamma}=0$.

The main cause of the restrictive nature of this theorem outside linear programming is that in order for (8) to have a Karush–Kuhn–Tucker point its constraints must in general satisfy a constraint qualification. This is difficult when g, h and θ are nonlinear because of the constraint $\theta(x) \leqq \bar{\theta}$. However when h is linear and θ and g are pseudoconcave or concave at $\bar{x}$ then a constraint qualification is automatically satisfied [2, Thm. 11.3.6]. This is a somewhat restrictive extension which does however include the case when (6) is a linear program.

The rest of the paper is devoted exclusively to the perturbation (2) of the linear program (1). We will first show that, under appropriate assumptions, some element $\bar{x}$ of the solution set $\bar{S}$ of (1) will be a local (global) solution of $P(\varepsilon)$ for all sufficiently small $\varepsilon \geqq 0$. We will then show that under slightly stronger assumptions, each global solution of $P(\varepsilon)$ for sufficiently small $\varepsilon \geqq 0$ is also a solution of (1). We begin by assuming that $\min_{x \in \bar{S}} f(x)$ has a local (global) solution $\bar{x}$, so that there exists an open ball B with center $\bar{x}$ such that $\bar{x} \in \bar{S} \cap B$ is optimal for the problem

$$\text{(9)} \qquad \text{Minimize } f(x) \quad \text{subject to} \quad x \in \bar{S} \cap B.$$

The proof of the subsequent results depends crucially on establishing a minimum rate of increase of px in certain directions that lead "away" from $\bar{S}$. These directions are related to *projections* of points in S on $\bar{S}$. The projection of a point x on $\bar{S}$ is denoted by $\mu(x)$ with $\mu(x) \in \bar{S}$ and

$$\|\mu(x)-x\|=\min_{\mu \in \bar{S}} \|\mu-x\|,$$

where $\|\cdot\|$ denotes the ∞ norm *throughout this paper* unless otherwise subscripted. We state now the key result which gives the desired lower bound on $p(x-\mu(x))$ and give the proof in the Appendix.

LEMMA 1. *There exists an* $\alpha>0$ *such that*

$$p(x-\mu(x)) \geqq \alpha\|x-\mu(x)\| \quad \textit{for all } x \in S.$$

We shall also need the following Lipschitz property on the perturbation function f. There exist positive numbers δ and K such that

$$f(\mu(x))-f(x)\leqq K\|x-\mu(x)\| \quad \text{for } x\in S \text{ and } \|x-\mu(x)\|\leqq \delta. \tag{10}$$

Note that it follows from the definition of $\mu(x)$ that $\|x-\mu(x)\|\leqq \delta$ whenever $\|x-\bar{x}\|\leqq \delta$. With the above concepts we establish our next principal result.

THEOREM 3. *Let $\bar{x}$ be a local solution of* $\min_{x\in\bar{S}} f(x)$. *Then, for sufficiently small $\varepsilon \geqq 0$, $\bar{x}$ is both a global solution of the linear program* (1) *and a local solution of the perturbed problem* (2) *provided that either of the two following conditions holds*:

(a) *The Lipschitz property* (10) *holds.*

(b) *f is convex on some open set containing $\bar{x}$.*

Proof. (a) Let (10) hold and let $B=B(\bar{x},\bar{\delta})=\{x \mid \|x-\bar{x}\|<\bar{\delta}\}$ where $\bar{\delta}$ is chosen such that $0<\bar{\delta}\leqq\delta$ and $\bar{x}$ is an optimal solution of (9). Note that if $x\in B(\bar{x},\bar{\delta})$, then

$$\|x-\mu(x)\|\leqq\|x-\bar{x}\|<\bar{\delta}\leqq\delta.$$

Hence by part (b) of Lemma A.3 of the Appendix we have upon noting the equality $p\bar{x}=p\mu(x)$ that

$$\varepsilon f(\bar{x})+p\bar{x}\leqq\varepsilon f(x)+px \quad \text{for } x\in S\cap B\left(\bar{x},\frac{\bar{\delta}}{2}\right)$$
$$\text{and } \varepsilon\in\left[0,\frac{\alpha}{K}\right].$$

Hence $\bar{x}$ solves (2) for $\varepsilon\in[0,\alpha/K]$ with the added constraint that $x\in B(\bar{x},\bar{\delta}/2)$.

(b) Let f be convex on $B(\bar{x},r)$ for some $r>0$. By Theorem 10.4 of [5] f is Lipschitzian on any open ball $B(\bar{x},\delta)$ with $\delta<r$, and again we have that

$$\|x-\mu(x)\|\leqq\|x-\bar{x}\|<\delta \quad \text{for } x\in B(\bar{x},\delta).$$

Hence because f is Lipschitzian on $B(\bar{x},\delta)$, the first inequality of (10) holds for $x\in B(\bar{x},\delta)$, and because f is convex on $B(\bar{x},\delta)$, $\bar{x}$ is an optimal solution of (9) with $B=B(\bar{x},\delta)$. Again by part (b) of Lemma A.3 of the Appendix we have that

$$\varepsilon f(\bar{x})+p\bar{x}\leqq\varepsilon f(x)+px \quad \text{for } x\in S\cap B\left(\bar{x},\frac{\delta}{2}\right)$$
$$\text{and } \varepsilon\in\left[0,\frac{\alpha}{K}\right].$$

Hence $\bar{x}$ solves (2) for $\varepsilon\in[0,\alpha/K]$ with the added constraint that $x\in B(\bar{x},\delta/2)$. □

Example 1. To illustrate the need for the Lipschitz property (10), let $x\in R^1$, let $S=\{x\geqq 0\}$, $p=1$, $f(x)=-x^{1/2}$. Note that f is continuous on S, but does *not* have the Lipschitz property in a neighborhood of $\bar{S}=\{0\}$. (Note also that f is convex on S, but cannot be extended to a finite convex function on R^1.) In this case it is easily verified that $\bar{S}(\varepsilon)=\varepsilon^2/4$ for $\varepsilon\geqq 0$ and thus $\bar{S}(\varepsilon)$ never includes $\{0\}$ for any positive ε.

Note that in Theorem 3, the Lipschitz property (10) is needed *only* for those $x\in S$ that lie in some open neighborhood of $\bar{x}$, since only such points are involved in the statement of the theorem and its proof. On the other hand by using the full strength of (10) and under slightly stronger assumptions than those of Theorem 3 we can show that each global solution of the perturbed problem (2) for sufficiently small $\varepsilon\geqq 0$ is also a solution of the linear program (1). In particular we have the following.

THEOREM 4. *Let $\bar{x} \in \bar{S}$ be a solution of* $\min_{x \in \bar{S}} f(x)$ *and let* $px + \varepsilon^* f(x)$ *be bounded from below on* S *for some* $\varepsilon^* > 0$. *Then* $\bar{x} \in \bar{S}(\varepsilon) \subset \bar{S}$ *for sufficiently small* $\varepsilon \geqq 0$ *provided that any of the following conditions holds*:

(a) *The Lipschitz property* (10) *holds.*

(b) f *is convex on some open convex set containing* S.

(c) f *has continuous first partial derivatives on some open set containing* $\bar{S}$ *and* $\bar{S}$ *is compact.*

Proof. We will first establish that $\bar{x} \in \bar{S}(\varepsilon) \subset \bar{S}$ for sufficiently small $\varepsilon \geqq 0$ under hypothesis (a) by showing that for sufficiently small $\varepsilon \geqq 0$

$$p\bar{x} + \varepsilon f(\bar{x}) < px + \varepsilon f(x) \quad \text{for } x \in S \backslash \bar{S} \tag{11}$$

and

$$p\bar{x} + \varepsilon f(\bar{x}) \leqq px + \varepsilon f(x) \quad \text{for } x \in \bar{S}. \tag{12}$$

Inequality (12) holds because $\bar{x}$ minimizes f on $\bar{S}$. To establish (11), let $x \in S \backslash \bar{S}$; thus $x \neq \mu(x)$, and consider the two following cases.

Case 1: $0 < \|\mu(x) - x\| \leqq \delta$. The strict inequality (11) follows from part (a) of Lemma A.3 of the Appendix for $\varepsilon \in [0, \alpha/K)$ upon noting that $p\bar{x} = p\mu(x)$.

Case 2: $\|\mu(x) - x\| > \delta$. Let ν be such that $px + \varepsilon^* f(x) \geqq \nu$ for $x \in S$, so that $f(x) \geqq \nu/\varepsilon^* - px/\varepsilon^*$. By defining

$$q = -p/\varepsilon^* \quad \text{and} \quad \rho = -\nu/\varepsilon^* + f(\bar{x}) + p\bar{x}/\varepsilon^*$$

we have that

$$f(\bar{x}) - f(x) \leqq q(\mu(x) - x) + \rho \quad \text{for } x \in S.$$

Because $\|\mu(x) - x\| > \delta$ it follows upon using the Hölder inequality that

$$\varepsilon(f(\bar{x}) - f(x))/\|x - \mu(x)\| \leqq \varepsilon\|q\|_1 + \varepsilon\rho/\delta \quad \text{for } x \in S$$

and consequently for ε small enough, that is $\varepsilon \in [0, \alpha/(\|q\|_1 + \rho/\delta))$, the right hand side of the last inequality is less than α. Thus, for such ε

$$\begin{aligned} \varepsilon(f(\bar{x}) - f(x)) &< \alpha\|x - \mu(x)\| \\ &\leqq px - p\bar{x} \qquad \text{(by Lemma 1).} \end{aligned}$$

This establishes (11) for this second case also.

Now note that hypothesis (c) implies (a) and that hypothesis (b) also implies (a) in the case that $\bar{S}$ is compact [5, Thm. 10.4], so that the proof will be completed by showing that the result holds under hypothesis (b) even when $\bar{S}$ is not compact. Let

$$T = \{x \mid \|x - \bar{x}\| \leqq k\},$$

where k is some positive number, let $S' = S \cap T$, and let $\bar{S}' = \bar{S} \cap T$. Note that S' is a compact polyhedral set and that $\bar{S}'$ is the set of optimal solutions of $\min_{x \in S'} px$, so that the preceding arguments imply that there exists an $\varepsilon' > 0$ such that $\bar{x} \in \bar{S}'(\varepsilon) \subset \bar{S}'$ for $\varepsilon \in [0, \varepsilon']$ where $\bar{S}'(\varepsilon)$ denotes the solution set of $\min_{x \in S \cap T} px + \varepsilon f(x)$. Now suppose that for some $\varepsilon \in [0, \varepsilon']$, $\bar{S}(\varepsilon)$ contains a point $\tilde{x} \notin \bar{S}$. By the convexity of $px + \varepsilon f(x)$, $\bar{x} \in \bar{S}'(\varepsilon)$ implies that $\bar{x} \in \bar{S}(\varepsilon)$ (since a local solution of $P(\varepsilon)$ must also be a global solution), and consequently by the convexity of $\bar{S}(\varepsilon)$ it follows that

$$x(\lambda) = (1 - \lambda)\bar{x} + \lambda\tilde{x} \in \bar{S}(\varepsilon) \quad \text{for all } \lambda \in [0, 1].$$

However, for $\lambda \in (0, 1]$ we have that $x(\lambda) \notin \bar{S}$ and hence $x(\lambda) \notin \bar{S}'$. But for sufficiently small $\lambda > 0$, $x(\lambda) \in \bar{S}'(\varepsilon) \subset \bar{S}'$, which is a contradiction. Thus $\bar{S}(\varepsilon) \subset \bar{S}$ for $\varepsilon \in [0, \varepsilon']$, and since $\bar{x} \in \bar{S}(\varepsilon)$ the theorem is established under hypothesis (b). □

In the terminology of point-to-set mappings the result $\bar{S}(\varepsilon) \subset \bar{S}$ of Theorem 4 for $\varepsilon \geqq 0$ sufficiently small implies that the mapping $\bar{S}(\varepsilon)$ is *upper semi-continuous at* 0 in a strong sense. (Note that if $px + \varepsilon f(x)$ is *not* bounded from below for any $\varepsilon > 0$, then the inclusion $\bar{S}(\varepsilon) \subset \bar{S}$ holds trivially, since $\bar{S}(\varepsilon) = \phi$ for all $\varepsilon > 0$.)

To see that the compactness of $\bar{S}$ is necessary in hypothesis (c) of Theorem 4 we give below an example in which the conclusion of Theorem 4 fails when the compactness assumption of part (c) is dropped.

Example 2. Let $x \in R^2$, $p = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, $S = \{(x_1, x_2) | 1 \leqq x_1, 0 \leqq x_2 \leqq 1\}$ and $f(x) = -x_1x_2 + x_1^3x_2^2$. Note that on S, $px + \varepsilon f(x) \geqq \varepsilon(-x_1x_2 + (x_1x_2)^2) \geqq -\varepsilon/4$. Moreover,

$$\bar{S} = \{(x_1, x_2) | x_1 \geqq 1, x_2 = 0\}$$

and $f(x) = 0$ on $\bar{S}$, so that $px + \varepsilon f(x) = 0$ on $\bar{S}$ for all $\varepsilon \geqq 0$. However for $0 < \varepsilon \leqq 2$, $x_1 = 2/\varepsilon$, $x_2 = \varepsilon^2/16$, we have that $(x_1, x_2) \in S$, $px + \varepsilon f(x) = -\varepsilon^2/32 < 0$, and hence no solution of $\min_{x \in \bar{S}} f(x)$ can be in $\bar{S}(\varepsilon)$, the solution set of $\min_{x \in S} px + \varepsilon f(x)$. It can also be shown that $\bar{S}(\varepsilon)$ is nonempty for all $\varepsilon > 0$ so that $\bar{S}(\varepsilon)$ is not contained in $\bar{S}$.

Note that in the case that the linear program (1) has a *unique* solution, many of the results above may be simplified. In particular, Theorems 1 and 4 yield the following. (See also Remark 4 in [3].)

COROLLARY 1. *Let $\bar{S}$ consist of a single point $\bar{x}$. If f is differentiable at $\bar{x}$, then $\bar{x}$ is a Karush–Kuhn–Tucker point of* (2) *for all sufficiently small $\varepsilon \geqq 0$. If, in addition, $\bar{S}(\varepsilon^*) \neq \emptyset$ for some $\varepsilon^* > 0$, then $\bar{S}(\varepsilon) = \{\bar{x}\}$ for all sufficiently small $\varepsilon \geqq 0$.*

Proof. The first conclusion follows directly from Theorem 1. The second follows from the fact that $\bar{S} = \{\bar{x}\}$ implies that $\mu(x) = \bar{x}$ for all $x \in S$, so that the Lipschitz property (10) holds as a consequence of differentiability of f at $\bar{x}$. This part of the corollary then follows from Theorem 4. □

A similar result also holds without assuming uniqueness in (1) if a strict convexity property is assumed instead.

COROLLARY 2. *If f is strictly convex on some open set containing S and if $\bar{x}$ is the solution of* $\min_{x \in \bar{S}} f(x)$, *then $\bar{S}(\varepsilon) = \{\bar{x}\}$ for all sufficiently small $\varepsilon > 0$.*

Proof. The proof follows from Theorem 3 and the fact that, for $\varepsilon > 0$, $px + \varepsilon f(x)$ is strictly convex and therefore assumes its minimum at not more than one point in S. □

Appendix.

LEMMA A.1. *There exists an $\alpha > 0$ such that*

$$p(x - \mu(x)) \geqq \alpha \|x - \mu(x)\| \quad \textit{for all } x \in S.$$

Proof. Obviously the lemma holds trivially when $x \in \bar{S}$ or equivalently when $x = \mu(x)$. Suppose now that $x \in S \backslash \bar{S}$ and let e be a vector of ones in R^n. Then

$$0 < \|x - \mu(x)\| = \underset{\delta, \mu}{\text{Minimum}} \{\delta | -\delta e \leqq \mu - x \leqq \delta e, A\mu \geqq b, p\mu \leqq \bar{\theta}\}$$

$$= \underset{y, v, \zeta, w}{\text{Maximum}} \{x(y - v) - \bar{\theta}\zeta + bw | y - v - p\zeta + A^T w = 0,$$

$$ey + ev = 1, \quad y, v, \zeta, w \geqq 0\}$$

(by linear programming duality)

$$\begin{aligned}
&= \underset{y,v,\zeta,w}{\text{Maximum}} \{\zeta(px-\bar{\theta})+w(b-Ax) \mid y-v-p\zeta+A^Tw=0, \\
&\qquad\qquad\qquad\qquad ey+ev=1, \quad y,\, v,\, \zeta,\, w \geqq 0\}. \\
&= \zeta(x)(px-p\mu(x))+w(x)(b-Ax) \\
&\quad \text{(since } \bar{\theta}=p\mu(x) \text{ and } (y(x),\, v(x),\, \zeta(x),\, w(x)) \text{ is a} \\
&\qquad\qquad \text{solution to the maximum problem)} \\
&\leqq \zeta(x)p(x-\mu(x)) \\
&\quad \text{(since } w(x) \geqq 0, \text{ and } b-Ax \leqq 0).
\end{aligned} \tag{A.1}$$

Thus $\zeta(x)>0$ for $x \in S \backslash \bar{S}$, and in addition

$$1 \leqq \zeta(x)\frac{p(x-\mu(x))}{\|x-\mu(x)\|} \quad \text{for } x \in S\backslash \bar{S}. \tag{A.2}$$

But since $\zeta(x)$ may be chosen as a component of a solution vertex of the linear program (A.1) and since the feasible region of (A.1) is independent of x and has a finite number of vertices, $\zeta(x)$ for $x \in S \backslash \bar{S}$ may be bounded as follows

$$\begin{aligned}
\zeta(x) \leqq \frac{1}{\alpha} :=& \text{ maximum } \{\zeta \mid (y,\, v,\, \zeta,\, w) \text{ is a vertex of } y-v-p\zeta+A^Tw=0, \\
&\qquad\qquad\qquad\qquad ey+ev=1, \quad y,\, v,\, \zeta,\, w \geqq 0\}.
\end{aligned}$$

This bound on $\zeta(x)$ together with (A.2) establishes the lemma. □

LEMMA A.2. *If $\bar{x} \in \bar{S}$ and $x \in R^n$ then*

$$\|\mu(x)-\bar{x}\| \leqq 2\|x-\bar{x}\|.$$

Proof.

$$\begin{aligned}
\|\mu(x)-\bar{x}\| &\leqq \|\mu(x)-x\|+\|x-\bar{x}\| \\
&\leqq \|\bar{x}-x\|+\|x-\bar{x}\| \qquad \text{(since } \mu(x) \text{ is the projection of } x \text{ on } \bar{S}) \\
&= 2\|x-\bar{x}\|.
\end{aligned}$$

□

LEMMA A.3. *Let the Lipschitz condition* (10) *hold, let $\bar{x} \in \bar{S} \cap B$ be a solution of* $\min_{x \in \bar{S} \cap B} f(x)$ *with the ball $B=B(\bar{x}, \bar{\delta})$ for some $\bar{\delta}>0$. Then for $\|x-\mu(x)\| \leqq \delta$ and $x \in S \cap B(\bar{x}, \bar{\delta}/2)$*

(a) $\varepsilon(f(\bar{x})-f(x))<p(x-\mu(x))$ *for $x \neq \mu(x)$ and $\varepsilon \in [0, \alpha/K)$*

and

(b) $\varepsilon(f(\bar{x})-f(x)) \leqq p(x-\mu(x))$ *for $\varepsilon \in [0, \alpha/K]$.*

Proof. Let $x \in S \cap B(\bar{x}, \bar{\delta}/2)$ and $\|x-\mu(x)\| \leqq \delta$; then

$$\begin{aligned}
\varepsilon(f(\bar{x})-f(x)) &\leqq \varepsilon(f(\mu(x))-f(x)) && \text{(since by Lemma A.2 } \mu(x) \in \bar{S} \cap B(\bar{x}, \bar{\delta})) \\
&\leqq \varepsilon K\|\mu(x)-x\| && \text{(by (10) and } \|x-\mu(x)\| \leqq \delta), \\
&< \alpha\|\mu(x)-x\| && \text{(for } \varepsilon \in [0, \alpha/K) \text{ and } x \neq \mu(x)), \\
&\leqq p(x-\mu(x)) && \text{(by Lemma A.1).}
\end{aligned}$$

This establishes part (a) of the lemma. Part (b) follows by changing the strict inequality in the above string of inequalities to an inequality for the case of $\varepsilon \in [0, \alpha/K]$. □

REFERENCES

[1] W. KARUSH, *Minima of functions of several variables with inequalities as side conditions*, Master of Science Dissertation, Dept. of Mathematics, Univ. of Chicago, December 1939.

[2] O. L. MANGASARIAN, *Nonlinear Programming*, McGraw-Hill, New York, 1969.

[3] ———, *Uniqueness of solution in linear programming*, Linear Algebra and Appl., 25 (1979), pp. 151–162.

[4] ———, *Iterative solution of linear programs*, Computer Sciences Tech. Rep. 327, Univ. of Wisconsin, Madison, in preparation.

[5] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 6, November 1979

0363-0129/79/1706-0007 $01.00/0

NULL CONTROLLABILITY OF NONLINEAR FUNCTIONAL DIFFERENTIAL EQUATIONS*

ROBERT G. UNDERWOOD† AND DONALD F. YOUNG‡

Abstract. For various types of linear and nonlinear functional differential equations null controllability (local or global) is established. Previously it was believed that for a broad class of nonlinear equations local null controllability was implied by the null controllability of the linear approximation. Such an implication does not hold in this generality, as is demonstrated by a counterexample in the present paper. However, under certain conditions on the system structure, one is able to establish this type of relationship. For equations of certain other forms (both linear and nonlinear), null controllability can be deduced from the null controllability of related equations with the delays removed. The systems considered include those not only with delays in the state but also with delays in the control.

1. Introduction. In this paper we shall study techniques for establishing null controllability of linear and nonlinear functional differential equations. The linear case has been actively studied, as in [1], [2], [7], [9], [13]. However, the nonlinear case has received little attention in the existing literature. In [12] Weiss considered the null controllability of nonlinear systems in terms of their linear approximation, in a manner analogous to the standard approach used for ordinary differential equations [8]. Unfortunately, Weiss' nonlinear results are incorrect.

In the proof of Theorem 5 in [12], Weiss introduces a parameter $\xi \in R^n$ into his control. This approach is essentially the same as the one used for ordinary control systems (see [8, p. 366]). There, the control u in the nonlinear system is parameterized by the point ξ steered by u to the origin for the approximating linear system. Using the implicit function theorem, one then shows that the mapping which associates ξ with the corresponding initial condition of the nonlinear system is an open mapping in a neighborhood of the origin. For functional differential equations, however, this argument will not work, because it would involve establishing a linear homeomorphism from the parameter space R^n onto the space of initial conditions, which is infinite dimensional. It is precisely such an error that Weiss makes in proving Theorem 5 of [12].

The infinite dimensionality of the state space is what makes investigating null controllability of nonlinear functional differential equations difficult. In order to circumvent these difficulties in this paper, the problem is reduced to one involving *ordinary* differential equations. We consider two situations in which such a reduction can be of use. One of these is a situation in which the *difference* between a nonlinear control system and its linear approximation is equivalent to one or several ordinary differential equations. (Throughout this paper, "linear approximation" will always mean the linear approximation about the point $x=0$, $u=0$.) We apply to this difference (the solution of which we label z) a parametrization argument in the spirit of [8]. The parameters ψ are the initial conditions for the linear approximation. We shall show that the mapping $\psi \to z_{t_0}$ has its Fréchet derivative equal to zero. This will imply that the mapping $\psi \to x_{t_0}$ is an open mapping (where x is the solution of the nonlinear control system). Critical to these arguments is the ability to "backout" the z equation from the origin. In the second situation which we consider, a control system involving functional differential equations can be converted into an ordinary differential control system by

* Received by the editors April 16, 1979.

† Department of Mathematics, Colorado School of Mines, Golden, Colorado 80401.

‡ Department of Mathematics, Agnes Scott College, Decatur, Georgia 30030.

making a "cascading" substitution. What is meant by this will become clear in § 3. Also in § 3 we consider certain situations in which lags appear in the control.

A counterexample to Theorem 5 in [12] is the system

$$\begin{aligned}\dot{x}_1(t)&=x_2(t)+x_1(t-1)^2+x_2(t-1)^2,\\ \dot{x}_2(t)&=u(t).\end{aligned}\tag{1.1}$$

The linear approximation is

$$\begin{aligned}\dot{x}_1(t)&=x_2(t),\\ \dot{x}_2(t)&=u(t)\end{aligned}\tag{1.2}$$

which is easily seen to be null controllable. (Controllability results used throughout this paper for ordinary differential equations can be found in [8].) However, (1.1) is not locally null controllable to the zero function in $C([-1, 0], R^2)$ on any interval. To see this, suppose $[t_0, t_1]$ is an interval of length greater than one, suppose $\phi \in C([-1, 0], R^2)$ and suppose $u \in L^2([t_0, t_1], R)$ steers $x_{t_0}=\phi$ to $x_{t_1}=0$ for the system (1.1). Let x be the corresponding trajectory. Then $x_1(t)=x_2(t)=0$ for $t_1-1\leqq t\leqq t_1$, and so for $t_1-2\leqq t\leqq t_1-1$ the first of the two equations gives $x_1(t)^2+x_2(t)^2=0$. Therefore, $x_1(t)=x_2(t)=0$ for $t_1-2\leqq t\leqq t_1-1$. Continuing in this way, we conclude that $x(t)=0$ for $t_0-1\leqq t\leqq t_1$. In particular, $\phi=0$. Thus, the only function that can be steered to $0\in C([-1, 0], R^2)$ at time t_1 for (1.1) is the zero function.

For function space controllability, results obtained must depend on the particular state space chosen (e.g. [1], [2], [9]). In our state space, $C([-r, 0], R^n)$, one could expect only denseness results when considering complete controllability (see [1, p. 611]). However, we deal in this paper only with null controllability; for nonlinear systems this is the appropriate approach (cf. [8, § 6.1]).

We shall consider null controllability only on a specified time interval $[t_0, t_1]$, rather than on arbitrary intervals of length greater than the lag, as in [1]. Since most of our results depend so heavily on the length of the interval, there is no real advantage to considering arbitrary intervals here.

1.1. Notation. Suppose p and q are positive integers and $\mathscr{J}$ is any interval of the real line R. Denote by R^p the space of real p-tuples with the usual Euclidean norm. The norm in R^p will be denoted by $|\cdot|$, and the norm in any other Banach space will be denoted by $\|\cdot\|$. The space of real $p\times q$ matrices with the operator norm will be denoted by $\mathcal{M}_{pq}$. Vectors in R^p will be identified with $p\times 1$ matrices. However, occasionally for compactness such vectors will be written as rows when no ambiguity will result. The space of bounded linear transformations from X to Y will be denoted by $\mathscr{B}(X, Y)$, where X and Y are Banach spaces, and if $X=Y$ we shall write $\mathscr{B}(X)$ for $\mathscr{B}(X, X)$. The identity operator on any Banach space will be denoted by I. The usual Lebesgue space of square-integrable (equivalence classes of) functions from $\mathscr{J}$ to R^p will be denoted by $L^2(\mathscr{J}, R^p)$. Here, and throughout this paper, any statements involving measures are understood to refer to Lebesgue measure. If X is a metric space, $C(\mathscr{J}, X)$ represents the set of continuous functions from $\mathscr{J}$ to X. If $\mathscr{J}$ is compact and X is a Banach space, then $C(\mathscr{J}, X)$ is a Banach space with the supremum norm. If $r>0$ and $[t-r, t]\subset\mathscr{J}$ for some real number t and if $x\in C(\mathscr{J}, R^p)$, then x_t denotes the element of $C([-r, 0], R^p)$ given by $x_t(\theta)=x(t+\theta)$, $-r\leqq\theta\leqq 0$. (It will always be clear from context whether such a subscript is being used as just noted or is being used to indicate the component of a vector.) All vector spaces will be over the field of real numbers.

If f is any continuous function from a subset of a Banach space U to a Banach space V and if f has a Fréchet derivative at $x\in U$, then this derivative will be denoted by

$Df(x)$. If U is the Cartesian product of several Banach spaces, then the partial derivative with respect to the ith variable will be denoted by $D_if(x)$.

Throughout the remainder of this paper, r will be a positive real number; n and m will be positive integers; and t_0 and t_1 will be real numbers, with $t_1-t_0>r$. For any $t\in[t_0-r, t_1]$, $\max\{t_0, t\}$ will be denoted by $\alpha(t)$. Denote the interval $[t_0, t_1]$ by J and the Banach space $C([-r, 0], R^n)$ by X. The space of admissible controls will always be $L^2(J, R^m)$.

2. Techniques using the implicit function theorem. Let $N\subset R^n$ be an open convex set containing the origin. In this section we shall consider control systems of the form

$$\dot{x}(t)=f(t, x_t; u(t))+L(t, x_t)+C(t)u(t) \tag{2.1}$$

where $C: J\to \mathcal{M}_{nm}$ is square-integrable and where $f: J\times C([-r, 0], N)\times R^m\to R^n$ and $L: J\times X\to R^n$ satisfy certain assumptions to be described. The technical assumptions needed on the nonlinear term f are the following:

(A1) $f(t, \cdot, \cdot)$ is continuously differentiable for each t.

(A2) $f(\cdot, \phi, w)$ is measurable for all ϕ and w.

(A3) For each compact set $K\subset N$ there exists an integrable function $M_1: J\to[0, \infty)$ and square-integrable functions $M_i: J\to[0, \infty)$, $i=2, 3$, such that

$$\|D_2f(t, \phi, w)\|\leqq M_1(t)+M_2(t)|w|, \tag{2.2}$$

$$\|D_3f(t, \phi, w)\|\leqq M_3(t) \tag{2.3}$$

for all t and w and all $\phi\in C([-r, 0], K)$.

We also assume that $x_t=0$, $u=0$ is a critical point of the system (2.1):

(A4) $f(t, 0, 0)=0$ for all t.

The technical assumptions on the linear term L are as follows:

(A5) $L(t, \cdot)\in\mathcal{B}(X, R^n)$ for each t.

(A6) $L(\cdot, \phi)$ is measurable for each ϕ.

(A7) There exists an integrable function $M_4: J\to[0, \infty)$ such that

$$\|L(t, \cdot)\|\leqq M_4(t) \tag{2.4}$$

for all t.

Assumption (A8) below gives an important special relationship between f and L. This relationship is utilized in the "backing out" argument (for the interpretation of this phrase, cf. [8, pp. 4–11]) involved in the proof of Theorem 2.7, when the implicit function theorem is applied to the difference between system (2.1) and its linear approximation. It makes possible the use of functions f specialized enough for the backing out process while providing greater generality by the inclusion of L.

By the Riesz representation theorem, there exists a unique function $\eta: J\times[-r, 0]\to\mathcal{M}_{nn}$ such that for each $t\in J$, $\eta(t, \cdot)$ is of bounded variation on $[-r, 0]$, is left-continuous on $(-r, 0)$ and satisfies $\eta(t, 0)=0$, and such that $L(t, \phi)=\int_{-r}^{0}[d_\theta\eta(t, \theta)]\phi(\theta)$ for all $t\in J$, $\phi\in X$. With $\eta(t, \theta)$ as described, our assumption relating f and L is the following:

(A8) For each t and θ, the range of f is contained in the null space of $\eta(t, \theta)$.

Throughout this section, assumptions (A1)–(A8) will always be in force regarding f and L in (2.1). For emphasis, we shall reiterate them in the statements of the theorems.

2.1. The basic approach as applied to one general class of null controllable systems. If $F: J\times C([-r, 0], N)\times R^m\to R^n$ satisfies (A1)–(A4), then for the control system

$$\dot{x}(t)=F(t, x_t, u(t)), \tag{2.5}$$

(A1)–(A8) hold with $L=0$, $C=0$ and $f=F$. On the other hand, if we denote the right-hand side of (2.1) by $F(t, x_t, u(t))$, then it is easily seen that F satisfies (A1)–(A4). Thus, without further assumptions on the structure of the right-hand side of (2.1), there would be no point in considering (2.1) separately from (2.5). It is precisely because we wish to make sure further assumptions that we consider (2.1) in the form as given.

With additional requirements on the structure of the right-hand side of (2.1), (2.1) will properly be a special case of (2.5). We now make a few observations and give some definitions regarding the general system (2.5), where F is assumed to satisfy (A1)–(A4).

A solution of (2.5) on $[t_0-r, \bar{t}]$, where $\bar{t}\in(t_0, t_1]$, is by definition an absolutely continuous function $x\colon [t_0-r, \bar{t}]\to R^n$ such that (2.5) holds almost everywhere on J. Thus, if x is such a solution, then $F(t, x_t, u(t))$ must be integrable. In order to justify the definitions of certain integral operators to be given later, we shall establish this integrability under the more general assumption that x is *any* continuous mapping from $[t_0-r, t_1]$ to N and that $u\in L^2(J, R^m)$. Under this assumption, there exists a compact, convex set $K\subset N$ containing the origin such that $x(t)\in K$, $t_0-r\leqq t\leqq t_1$. For this K, let M_1 and M_3 be as in (A3) (as regards F). Then for all $t\in J$, $\phi\in C([-r, 0], K)$, $w\in R^m$, we have

$$|F(t, \phi, w)|\leqq M_1(t)\|\phi\|+M_3(t)|w|. \tag{2.6}$$

This is easily established using (2.2), (2.3) and the mean value theorem (which requires the convexity of K). It follows from (2.6) that $F(t, x_t, u(t))$ is an integrable function of t. (The required measurability is easily established using standard results in [10].)

For any $u\in L^2(J, R^m)$, $\phi\in C([-r, 0], N)$ and $\bar{t}\in(t_0, t_1]$, there exists at most one solution of (2.5) on $[t_0-r, \bar{t}]$ satisfying the initial condition $x_{t_0}=\phi$. This can be proved by using (2.2) and standard Gronwall type arguments.

If for some $\bar{t}\in(t_0, t_1]$ and some $u\in L^2(J, R^m)$, (2.5) has a solution x on $[t_0-r, \bar{t}]$, then we say that x is a *trajectory* on $[t_0-r, \bar{t}]$ corresponding to the control u. If x also satisfies the initial condition $x_{t_0}=\phi$, then we say that x is the trajectory corresponding to u and ϕ. If x is the trajectory on $[t_0-r, \bar{t}]$ corresponding to u and ϕ and if $x_{\bar{t}}=\psi$, then we say that *u steers ϕ to ψ* at time $\bar{t}$. We shall say that (2.5) is (*globally*) *null controllable* on $[t_0, \bar{t}]$ if for every $\phi\in X$ there exists a control u which steers ϕ to $0\in X$ at time $\bar{t}$. We shall say that (2.5) is *locally null controllable* on $[t_0, \bar{t}]$ if such a u exists for each ϕ in some open neighborhood of the origin in X.

Let $A(t, \cdot)=D_2f(t, 0, 0)$ and $B(t)=D_3f(t, 0, 0)$. Then the linear approximation to (2.1) is

$$\dot{x}(t)=A(t, x_t)+L(t, x_t)+(B(t)+C(t))u(t) \tag{2.7}$$

which itself is of the form (2.5), with $F(t, \phi, w)=A(t, \phi)+L(t, \phi)+(B(t)+C(t))w$ satisfying (A1)–(A4). (Note that the measurability of $A(t, \phi)$ in t and the measurability of $B(t)$ follow easily from the measurability of $f(t, \phi, w)$ in t and the definition of Fréchet derivative.) The following theorem gives one criterion on the structure of (2.1) which makes it possible to deduce the local null controllability of (2.1) from the null controllability of (2.7). The integer $j\in[1, n]$ is fixed. For any $x\in R^n$, we shall let Π_1x denote the projection of x onto its first j components and let Π_2x denote the projection of x onto its last $n-j$ components, and we shall write $x=(x^1, x^2)$, where $x^1=\Pi_1x$ and $x^2=\Pi_2x$. Of course if $j=n$, then Π_2x is vacaous and in the notation $x=(x^1, x^2)$ the x^2 entry is vacuous. Define $\tilde{\Pi}_1\colon x\to C([-r, 0], R^j)$ and $\tilde{\Pi}_2\colon X\to C([-r, 0], R^{n-j})$ by $(\tilde{\Pi}_i\phi)(\theta)=\Pi_i\phi(\theta)$, $i=1, 2$.

THEOREM 2.1. *Let* (A1)–(A8) *hold. Suppose* $t_1-t_0>3r$, *and suppose there exist functions* $f_1\colon J\times R^j\times C([-r, 0], R^{n-j})\times R^m\to R^j$ *and* $f_2\colon J\times R^{n-j}\times R^m\to R^{n-j}$ *such* *that*

$$f(t, \phi, w)=(f_1(t, \Pi_1\phi(0), \tilde{\Pi}_2\phi, w), f_2(t, \Pi_2\phi(0), w)) \tag{2.8}$$

for all (t, ϕ, w) *in the domain of* f. *Then the null controllability of* (2.7) *on* $[t_0, t_1-2r]$ *implies the local null controllability of* (2.1) *on* $[t_0, t_1]$.

This theorem will be proven as a corollary to Theorem 2.7. Note that under the hypotheses of Theorem 2.1, (2.1) takes the form

$$\begin{aligned}\dot{x}^1(t)&=f_1(t, x^1(t), x_t^2, u(t))+L_1(t, x_t)+C_1(t)u(t),\\ \dot{x}^2(t)&=f_2(t, x^2(t), u(t))+L_2(t, x_t)+C_2(t)u(t),\end{aligned} \tag{2.9}$$

where $L(t, \phi)=(L_1(t, \phi), L_2(t, \phi))$ and $C(t)w=(C_1(t)w, C_2(t)w)$ for all $t\in J$, $\phi\in X$, $w\in R^m$.

The simplest case of (2.9) is when $j=n$ and f is independent of w, so that $f(t, \phi, w)=f_1(t, \phi(0))$ (suppressing the third and fourth arguments of f_1 in (2.8)). In this situation, (2.9) takes the form

$$\dot{x}(t)=f_1(t, x(t))+L(t, x_t)+C(t)u(t). \tag{2.10}$$

Besides Theorem 2.1, two other theorems of a similar nature will be proven as corollaries to Theorem 2.7, and corollaries relating to further special cases of (2.1) are possible. Theorem 2.9 relates specifically to system (2.10).

The reason for choosing to give special attention to system (2.9) in Theorem 2.1 is that (2.9) retains some of the simplicity of (2.10) while at the same time illustrating some of the more complex situations that can arise. Theorem 2.7 is the main result of the paper in so far as it provides a general theoretical viewpoint, but it is cumbersome to apply. In terms of practical applications its corollaries, Theorems 2.1, 2.8 and 2.9, are of primary interest. Before proceeding further with the general case of Theorem 2.7, we shall discuss system (2.10) in some detail, since (2.10) illustrates well the reasons for the particular form of system (2.1) and the significance of the hypotheses in Theorem 2.7.

The linear approximation to (2.10) with y in place of the state variable is

$$\dot{y}(t)=A(t)y(t)+L(t, y_t)+C(t)u(t), \tag{2.11}$$

where $A(t)=D_2f_1(t, 0)$. As applied to (2.10), the central idea of the proof of Theorem 2.7 is to subtract from (2.10) its linear approximation (2.11) and then to drop the L term, giving the ordinary differential equation

$$\dot{z}(t)=f_1(t, y(t)+z(t))-A(t)y(t), \tag{2.12}$$

where $z=x-y$. Suppose a control u steers y_{t_0} to $0\in X$ at time t_1 for system (2.11). Substituting the u and y of (2.11) into (2.12) and solving the "final-value problem" consisting of (2.12) with the initial (or in this case, final) condition $z(t_1)=0$ gives the function $z(t)$ which satisfies the condition $z_{t_1}=0$, due to the fact that the origin in z-space is a critical point of (2.12) for $t_1-r\leqq t\leqq t_1$. (Note that $y_{t_1}=0$.) For $\|y_{t_0}\|$ and $\|u\|$ sufficiently small, the function $z(t)$ will exist for $t_0\leqq t\leqq t_1$. With $z(t)$ existing on this interval, and with $z_{t_0}(\theta)$ set equal to $z(t_0)$ for $-r\leqq\theta\leqq 0$, we can conclude from (A8) that $L(t, z_t)=0$ for $t_0\leqq t\leqq t_1$. Adding the term $L(t, z_t)$ to the right-hand side of (2.12), adding (2.11) to the result and setting $x=y+z$ then gives (2.10). Thus, the control u which steers $y_{t_0}=\psi$ to $0\in X$ at time t_1 for system (2.11) (we shall show that this control can be chosen in a bounded linear way depending on ψ) also steers $\phi=\psi+z_{t_0}$ to $0\in X$ at time t_1 for system (2.10). By an argument using the implicit function theorem, we are able to show that if the functions ψ cover a neighborhood of the origin in X, then so do the functions ϕ. Hence, the null controllability of (2.11) on $[t_0, t_1]$ implies the local null controllability of (2.10) on $[t_0, t_1]$. The details of this argument are contained (somewhat more abstractly) in the proof of Theorem 2.7. Note that for the special case (2.10) we only require null controllability of the linear approximation on $[t_0, t_1]$ instead of on $[t_0, t_1-2r]$ as is required in Theorem 2.1.

We have seen that the essence of proving the local null controllability of (2.10) is to "back out of the origin" the difference between (2.10) and (2.11), using the ordinary differential equation (2.12). Here assumption (A8) is necessary in order to drop the term $L(t, z_t)$. On the other hand, if the term $L(t, x_t)$ were not present in (2.10) (in which case we would have $L=0$ and (A8) would automatically be satisfied) then (2.10) would simply be an ordinary differential equation and the well-known results from [8] would be applicable. Now the reason for the particular form in which (2.1) is written, the significance of assumption (A8) and the need for some type of special structure of the function f in (2.1) should be clear.

In applications of Theorem 2.7 to more complex situations, the "backing-out" system might not be just a simple ordinary differential equation. For example, the "backing-out" system for (2.9) analogous to (2.12) is (2.35).

We now show how the techniques we have discussed can be applied to two examples.

Example 1. Consider the system

$$\begin{aligned} \dot{x}_1(t)&=x_1(t)^2+x_2(t)+x_2(t-1)^2,\\ \dot{x}_2(t)&=x_3(t-1)+x_2(t)^2+\sin u(t),\\ \dot{x}_3(t)&=x_3(t)+x_3(t-1)+u(t). \end{aligned} \tag{2.13}$$

This system may be written in the form of (2.1), with

$$L(t,\phi)=\begin{bmatrix}0\\ \phi_3(-1)\\ \phi_3(0)+\phi_3(-1)\end{bmatrix},\qquad f(t,\phi,w)=\begin{bmatrix}\phi_1(0)^2+\phi_2(0)+\phi_2(-1)^2\\ \phi_2(0)^2+\sin w\\ 0\end{bmatrix},$$

$$C(t)=\begin{bmatrix}0\\0\\1\end{bmatrix}.$$

It is readily verified that (A1)–(A8) are satisfied. Assuming we have taken our interval $[t_0, t_1]$ to be of length greater than 3, the hypotheses of Theorem 2.1 are satisfied, with

$$\begin{aligned} f_1&=\phi_1(0)^2+\phi_2(0)+\phi_2(-1)^2,\\ f_2&=\begin{bmatrix}\phi_2(0)^2+\sin w\\ 0\end{bmatrix}. \end{aligned}$$

The linear approximation to (2.13) is

$$\begin{aligned} \dot{x}_1(t)&=x_2(t),\\ \dot{x}_2(t)&=x_3(t-1)+u(t),\\ \dot{x}_3(t)&=x_3(t)+x_3(t-1)+u(t). \end{aligned}$$

This system is null controllable on $[t_0, \bar{t}]$ for any $\bar{t}>t_0+1$, by [1, Corollary 3.3]. Thus, we can conclude from Theorem 2.1 that (2.13) is locally null controllable on $[t_0, t_1]$ for any $t_1>t_0+3$.

Example 2. Consider the system

$$\dot{x}_1(t)=x_1(t)^2+x_2(t)+x_2(t-1),\qquad (2.14)$$
$$\dot{x}_2(t)=u(t).$$

This system may be written in the form of (2.10), with

$$L(t,\phi)=\begin{bmatrix}\phi_2(0)+\phi_2(-1)\\ 0\end{bmatrix},\qquad f_1(t,x)=\begin{bmatrix}x_1^2\\ 0\end{bmatrix}$$
$$C(t)=\begin{bmatrix}0\\ 1\end{bmatrix}.$$

The linear approximation to (2.14) is

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

This system is null controllable on $[t_0, t_0+2+\varepsilon]$ for any $\varepsilon>0$. In fact, a control u steering $x_{t_0}=\psi$ to $0\in X$ at time $t_0+2+\varepsilon$ can be calculated directly by assuming $x_2(t)=a(t-t_0)^2+b(t-t_0)+\psi_2(0)$ for $t_0\leqq t\leqq t_0+\varepsilon$ and requiring $x_2(t)=0$ for $t_0+\varepsilon\leqq t\leqq t_0+2+\varepsilon$ and $x_1(t)=0$ for $t_0+1+\varepsilon\leqq t\leqq t_0+2+\varepsilon$. It now follows from our remarks above concerning system (2.10) that (2.14) is locally null controllable on $[t_0, t_1]$ for any $t_1>t_0+2$.

2.2. The underlying concepts. The main result of this subsection is Theorem 2.7. Although the hypotheses are somewhat abstract, the significance of this theorem can be appreciated by seeing how the theorem is applied to the special cases in Theorems 2.1, 2.8 and 2.9. Also, Theorem 2.7 can itself be applied directly to systems more general than those covered by these particular corollaries, as we shall see by the example following Theorem 2.9. Before presenting Theorem 2.7, we develop some preliminary results.

For $\|u\|$ and $\|x_{t_0}\|$ sufficiently small, the existence of a solution x on $[t_0-r, t_1]$ to (2.1) is established by Proposition 2.5. More generally, we consider the existence of a solution on $[t_0-r, t_1]$ to the initial value problem

$$\dot{x}(t)=F(t,x_t,u(t)),\qquad (2.15)$$
$$x_{t_0}=\phi,$$

where $F: J\times C([-r,0],N)\times R^m\to R^n$ satisfies (A1)–(A4).

DEFINITION 2.2. Let $F: J\times C([-r,0],N)\times R^m\to R^n$ satisfy (A1)–(A4), let $A(t,\cdot)=D_2F(t,0,0)$ and let $B(t)=D_3F(t,0,0)$, and consider the initial value problem (2.15) and its linear approximation:

$$\dot{x}(t)=A(t,x_t)+B(t)u(t),\qquad (2.16)$$
$$x_{t_0}=\phi.$$

Define functions p and p_L on subsets of $X\times L^2(J,R^m)$ as follows. Let $p(\phi,u)\in C([t_0-r,t_1],R^n)$ be the (necessarily unique) solution of (2.15) on $[t_0-r,t_1]$ for any pair (ϕ,u) such that that solution exists, and let $p_L(\phi,u)\in C([t_0-r,t_1],R^n)$ be the solution of (2.16) on $[t_0-r,t_1]$ for any pair (ϕ,u) such that that solution exists.

Although p and p_L depend on F, we have not indicated such dependence in the notation because in any particular context where p and p_L are needed, F will be fixed.

It will be convenient to treat (2.15) in integrated form and consider it as an operator equation. Define $Z: X \to C([t_0-r, t_1], R^n)$ by

$$(Z\phi)(t)=\begin{cases}\phi(t-t_0) & \text{if } t_0-r \leqq t < t_0,\\ \phi(0) & \text{if } t_0 \leqq t \leqq t_1.\end{cases} \tag{2.17}$$

Then x is a solution of (2.15) on $[t_0-r, t_1]$ if and only if it satisfies

$$x(t)=\int_{t_0}^{\alpha(t)} F(s, x_s, u(s))\,ds+(Z\phi)(t) \qquad (t_0-r\leqq t\leqq t_1). \tag{2.18}$$

(Recall that we use the notation $\alpha(t)=\max\{t_0, t\}$.)

LEMMA 2.3. *Let* $F: J\times C([-r, 0], N)\times R^m \to R^n$ *satisfy* (A1)–(A4), *and define* $H: C([t_0-r, t_1], N)\times L^2(J, R^m)\to C([t_0-r, t_1], R^n)$ *by*

$$H(x, u)(t)=\int_{t_0}^{\alpha(t)} F(s, x_s, u(s))\,ds \qquad (t_0-r\leqq t\leqq t_1). \tag{2.19}$$

Then H is continuously differentiable, and

$$(D_1H(0, 0)x)(t)=\int_{t_0}^{\alpha(t)} D_2F(s, 0, 0)x_s\,ds, \tag{2.20}$$

$$(D_2H(0, 0)u)(t)=\int_{t_0}^{\alpha(t)} D_3F(s, 0, 0)u(s)\,ds \tag{2.21}$$

hold for all $x\in C([t_0-r, t_1], R^n)$, $u\in L^2(J, R^m)$ *and* $t\in[t_0-r, t_1]$.

Since most of the computations involved in proving this lemma are quite standard, we shall not give a complete proof. The formulas for D_1H and D_2H are

$$(D_1H(\bar{x}, \bar{u})x)(t)=\int_{t_0}^{\alpha(t)} D_2F(s, \bar{x}_s, \bar{u}(s))x_s\,ds$$

and

$$(D_2H(\bar{x}, \bar{u})u)(t)=\int_{t_0}^{\alpha(t)} D_3F(s, \bar{x}_s, \bar{u}(s))u(s)\,ds,$$

where $t\in[t_0-r, t_1]$, $\bar{x}\in C([t_0-r, t_1], N)$, $x\in C([t_0-r, t_1], R^n)$, and u, $\bar{u}\in L^2(J, R^m)$. The use of the dominated convergence theorem involved in proving these two formulas is justified by (A3). At certain points in the proof of the lemma, arguments involving extraction of subsequences are necessary. For example, in proving the continuity of D_1H, one shows that $\lim_{(x,u)\to(\bar{x},\bar{u})} D_1H(x, u)=D_1H(\bar{x}, \bar{u})$ by showing that

$$\lim_{(x,u)\to(\bar{x},\bar{u})}\int_{t_0}^{t_1} \|D_2F(s, x_s, u(s))-D_2F(s, \bar{x}_s, \bar{u}(s))\|\,dx=0. \tag{2.22}$$

To prove (2.22), suppose by way of contradiction that there existed an $\varepsilon>0$ and sequences $x^j\to\bar{x}$, $u^j\to\bar{u}$ such that

$$\int_{t_0}^{t_1} \|D_2F(s, x_s^j, u^j(s))-D_2F(s, \bar{x}_s, \bar{u}(s))\|\,ds\geqq\varepsilon$$

for $j=1, 2, \cdots$. Then there exists a subsequence $\{\tilde{u}^j\}$ of $\{u^j\}$ and a corresponding subsequence $\{\tilde{x}^j\}$ of $\{x^j\}$ such that $\tilde{u}^j(t)\to\bar{u}(t)$ a.e. and $\{\tilde{u}^j\}$ is dominated by a function $v\in L^2(J, R)$ (see [10, p. 66]). Then by the dominated convergence theorem (justified by

(2.2)), we have

$$\lim_{j\to\infty}\int_{t_0}^{t_1}\|D_2F(s, \tilde{x}_s^j, \tilde{u}^j(s))-D_2F(s, \bar{x}_s, \bar{u}(s))\|\,ds=0$$

giving the desired contradiction and establishing (2.22). The continuity of D_2H is established similarly. We also note that the measurability of all integrands involved in the proof can be verified by standard techniques (see, e.g., [10]).

LEMMA 2.4. *Let* $A: J\times X\to R^n$ *satisfy* (A5)–(A7), *and define* $K\in\mathcal{B}(C([t_0-r, t_1], R^n))$ *by*

$$(Kx)(t)=\int_{t_0}^{\alpha(t)}A(s, x_s)\,ds \qquad (t_0-r\leqq t\leqq t_1).$$

Then $(I-K)^{-1}$ *exists.*

Proof. It follows from (2.4) (with A in place of L) that $\{Kx: x\in C([t_0-r, t_1], R^n), \|x\|\leqq 1\}$ is equicontinuous. Hence K is a compact operator. For fixed $\phi\in X$, any solution of the initial value problem

$$\dot{x}(t)=A(t, x_t),$$
$$x_{t_0}=\phi$$

is unique. Therefore, the unique solution of $x=Kx$ is $x=0$. It now follows by the Fredholm alternative [11, p. 103], that $(I-K)^{-1}$ exists.

If A is as in Lemma 2.4 and $B: J\to\mathcal{M}_{nm}$ is square-integrable, then it follows from Lemma 2.4 that, for any $\phi\in X$ and any $u\in L^2(J, R^m)$, the initial value problem

$$\dot{x}(t)=A(t, x_t)+B(t)u(t),$$
$$x_{t_0}=\phi$$

has a unique solution on $[t_0-r, t_1]$.

If F and H are as in Lemma 2.3 and p_L is as in Definition 2.2, then it follows from Lemma 2.4 that $(I-D_1H(0, 0))^{-1}$ exists and that p_L is a continuous bilinear function on $X\times L^2(J, R^m)$.

Using the operator H defined in Lemma 2.3, (2.18) can be rewritten as

$$x=H(x, u)+Z\phi. \tag{2.23}$$

Note that by (A4), $H(0, 0)=0$.

PROPOSITION 2.5. *Suppose* $F: J\times C([-r, 0], N)\times R^m\to R^n$ *satisfies* (A1)–(A4), *let p be as given in Definition* 2.2, *and let* $H: C([t_0-r, t_1], N)\times L^2(J, R^m)\to C([t_0-r, t_1], R^n)$ *be given by*

$$H(x, u)(t)=\int_{t_0}^{\alpha(t)}F(s, x_s, u(s))\,ds \qquad (t_0-r\leqq t\leqq t_1).$$

Then there exist open neighborhoods N_1 *and* N_2 *of the origin in* X *and* $L^2(J, R^m)$, *respectively, such that, for each* $\phi\in N_1$ *and each* $u\in N_2$, (2.23) *has the unique solution* $x=p(\phi, u)$, *and such that p is continuously differentiable on* $N_1\times N_2$.

Proof. Define $\tilde{H}(x, u, \phi)=x-H(x, u)-Z\phi$. Then $\tilde{H}(0, 0, 0)=0$, $\tilde{H}$ is continuously differentiable on $C([t_0-r, t_1], N)\times L^2(J, R^m)\times X$, and $D_1\tilde{H}(0, 0, 0)=I-D_1H(0, 0)$. Since $(I-D_1H(0, 0))^{-1}$ exists, it follows by the implicit function theorem that there exists a continuously differentiable function $\tilde{p}$ defined on an open neighborhood $\tilde{N}$ of the origin in $X\times L^2(J, R^m)$ such that $\tilde{H}(\tilde{p}(\phi, u), u, \phi)=0$ for all

$(\phi, u) \in \tilde{N}$. For each such pair (ϕ, u), $x = \tilde{p}(\phi, u)$ satisfies (2.23), or equivalently x satisfies

$$\text{(2.24)} \qquad \begin{aligned} \dot{x}(t) &= F(t, x_t, u(t)) \qquad (t \in J \text{ a.e.}) \\ x_{t_0} &= \phi. \end{aligned}$$

By uniqueness of solutions to (2.24), we conclude that $x = \tilde{p}(\phi, u)$ is the unique solution of (2.23) and that $\tilde{p}(\phi, u) = p(\phi, u)$ for all $(\phi, u) \in \tilde{N}$. Choosing N_1 and N_2 such that $N_1 \times N_2 \subset \tilde{N}$ completes the proof.

For ordinary linear control systems, it is well known [6, p. 92] that if the system is null controllable, then a control which "does the job" can be chosen depending on the initial condition in a bounded, linear manner. We now prove an analogous result pertaining to system (2.25) below.

LEMMA 2.6. *Let* $A: J \times X \to R^n$ *satisfy* (A5)–(A7), *let* $B: J \to \mathcal{M}_{nm}$ *be square-integrable, and let* p *be as in Definition* 2.2, *where* $F(t, \phi, w) = A(t, \phi) + B(t)w$. *Suppose the system*

$$\text{(2.25)} \qquad \dot{x}(t) = A(t, x_t) + B(t)u(t)$$

is null controllable on $[t_0, \bar{t}]$ *for some* $\bar{t} \in (t_0 + r, t_1]$, *and let*

$$U = \{u \in L^2(J, R^m) : u(t) = 0 \text{ for } \bar{t} \leqq t \leqq t_1\}.$$

Then there exists a bounded, linear mapping $S: X \to U$ *such that for each* $\phi \in X$ *and for* $\bar{t} - r \leqq t \leqq t_1$, $p(\phi, S\phi)(t) = 0$.

Proof. For all $x \in C([t_0 - r, t_1], R^n)$, $u \in L^2(J, R^m)$, $t \in [t_0 - r, t_1]$, define $(Kx)(t) = \int_{t_0}^{\alpha(t)} A(s, x_s)\, ds$ and $(Gu)(t) = \int_{t_0}^{\alpha(t)} B(s)u(s)\, ds$. Then $K \in \mathcal{B}(C([t_0 - r, t_1], R^n))$, $G \in \mathcal{B}(L^2(J, R^m), C([t_0 - r, t_1], R^n))$ and the integrated form of (2.25) with initial condition $x_{t_0} = \phi$ is

$$\text{(2.26)} \qquad x = Kx + Gu + Z\phi$$

where Z is given by (2.17). The operator $(I - K)^{-1}$ exists, by Lemma 2.4, and $p(\phi, u) = (I - K)^{-1}(Gu + Z\phi)$. Note that U is closed in $L^2(J, R^m)$. Define $P \in \mathcal{B}(X)$ by $(P\phi)(\theta) = ((I - K)^{-1}Z\phi)(\bar{t} + \theta)$, $-r \leqq \theta \leqq 0$, and define $Q \in \mathcal{B}(U, X)$ by $(Qu)(\theta) = ((I - K)^{-1}Gu)(\bar{t} + \theta)$, $-r \leqq \theta \leqq 0$. The statement that (2.25) is null controllable on $[t_0, \bar{t}]$ is equivalent to the statement that for every $\phi \in X$ there exists a $u \in U$ such that $Qu + P\phi = 0$. This in turn is equivalent to the statement that

$$\text{(2.27)} \qquad P(X) \subset Q(U).$$

Thus, (2.27) holds by hypothesis.

Let $\mathcal{N}$ be the null space of Q, denote the orthogonal complement of $\mathcal{N}$ in U by $\mathcal{N}^\perp$ and let $Q_0: \mathcal{N}^\perp \to Q(U)$ be the restriction of Q to $\mathcal{N}^\perp$. Then Q_0^{-1} exists and Q_0^{-1} is linear but not necessarily bounded (since $Q(U)$ is not necessarily closed in X). Referring to (2.27), we define $S: X \to U$ by $S\phi = -Q_0^{-1}P\phi$. Then $p(\phi, S\phi)(\bar{t} + \theta) = (QS\phi + P\phi)(\theta) = 0$, $\phi \in X$, $-r \leqq \theta \leqq 0$. Since $S\phi \in U$, it follows that $p(\phi, S\phi)(t) = 0$ for $\bar{t} - r \leqq t \leqq t_1$. To see that S is bounded, let $\{\phi_n\}$ be a convergent sequence in X such that $\{S\phi_n\}$ converges in U, and let $\phi = \lim_{n\to\infty} \phi_n$, $u = \lim_{n\to\infty} S\phi_n$. Since $\mathcal{N}^\perp$ is closed in U, $u \in \mathcal{N}^\perp$ and $Qu + P\phi = \lim_{n\to\infty} (QS_n + P\phi_n) = 0$. Thus, $u = -Q_0^{-1}P\phi = S\phi$, and therefore by the closed graph theorem, S is bounded, and this completes the proof.

In Theorem 2.1, certain criteria were given regarding the structure of f. The following two hypotheses, (H1) and (H2), generalize these criteria. Here $A(t, \cdot) = D_2f(t, 0, 0)$, $B(t) = D_3f(t, 0, 0)$ and $\bar{t} \in (t_0 + r, t_1]$.

(H1) For any $u \in L^2(J, R^m)$ satisfying $u(t)=0$ for $\bar{t} \leqq t \leqq t_1$ and any $y \in C([t_0-r, t_1], N)$ satisfying $y(t)=0$ for $\bar{t}-r \leqq t \leqq t_1$, there exists no solution z of

$$\dot{z}(t)=f(t, y_t+z_t, u(t))-A(t, y_t)-B(t)u(t) \tag{2.28}$$

on $[t_0-r, t_1]$ which satisfies both $z(t_1)=0$ and $z_{t_1} \neq 0$.

(H2) The only solution $z \in C([t_0-r, t_1], R^n)$ of

$$\begin{aligned} \dot{z}(t) &= A(t, z_t) \qquad (t_0 \leqq t \leqq t_1 \text{ a.e.}) \\ z(t_1) &= 0 \end{aligned} \tag{2.29}$$

which is constant on $[t_0-r, t_0]$ is $z=0$.

Note that for fixed $y \in C([t_0-r, t_1], N)$, for fixed $u \in L^2(J, R^m)$ and for $z \in C([t_0-r, t_1], R^n)$ of sufficiently small norm, the right-hand side of (2.28) is an integrable function of t on $[t_0, t_1]$.

THEOREM 2.7. *Let* (A1)–(A8) *hold. Let* $A(t, \cdot)=D_2f(t, 0, 0)$ *and* $B(t)=D_3f(t, 0, 0)$, *so that the linear approximation to* (2.1) *is*

$$\dot{x}(t)=A(t, x_t)+L(t, x_t)+(B(t)+C(t))u(t). \tag{2.30}$$

Let $\bar{t} \in (t_0+r, t_1]$, *suppose* (H1) *holds for this* $\bar{t}$ *and suppose* (H2) *holds. Then the null controllability of* (2.30) *on* $[t_0, \bar{t}]$ *implies the local null controllability of* (2.1) *on* $[t_0, t_1]$.

Proof. Let (2.30) be null controllable. Define $F: J \times C([-r, 0], N) \times R^m \to R^n$ by $F(t, \phi, w)=f(t, \phi, w)+L(t, \phi)+C(t)w$, and for this F, let p and p_L be as given in Definition 2.2. Let N_1 and N_2 be open neighborhoods of the origin in X and $L^2(J, R^m)$, respectively, as in Proposition 2.5, and let $U=\{u \in L^2(J, R^m): u(t)=0 \text{ for } \bar{t} \leqq t \leqq t_1\}$. By Lemma 2.6, there exists an operator $S \in \mathscr{B}(X, U)$ such that, for all $\phi \in X$ and for $\bar{t}-r \leqq t \leqq t_1$, $p_L(\phi, S\phi)(t)=0$. Define $T\phi=p_L(\phi, S\phi)$. Then $T \in \mathscr{B}(X, C([t_0-r, t_1], R^n))$. Define $M: \tilde{N}_1 \times \tilde{N}_2 \to C([t_0-r, t_1], R^n)$ by

$$\begin{aligned} M(z, \psi)(t)=\int_{t_1}^{\alpha(t)} & [f(s, (T\psi)_s+z_s, (S\psi)(s))-A(s, (T\psi)_s) \\ & -B(s)(S\psi)(s)]\, ds \qquad (t_0-r \leqq t \leqq t_1), \end{aligned} \tag{2.31}$$

where $\tilde{N}_1$ and $\tilde{N}_2$ are open balls around the origin in $C([t_0-r, t_1], R^n)$ and X, respectively, such that $\tilde{N}_1+T(\tilde{N}_2) \subset C([t_0-r, t_1], N)$. We claim that M is continuously differentiable and that $D_2M(0, 0)=0$.

To verify this claim, define $H: C([t_0-r, t_1], N) \times L^2(J, R^m) \to C([t_0-r, t_1], R^n)$ by $H(x, u)(t)=\int_{t_0}^{\alpha(t)} f(s, x_s, u(s))\, ds$, $t_0-r \leqq t \leqq t_1$, and let $K=D_1H(0, 0)$ and $G=D_2H(0, 0)$. Then referring to (2.20) and (2.21), we find that (2.31) can be written as

$$M(z, \psi)(t)=(H(T\psi+z, S\psi)-KT\psi-GS\psi)(t)-(H(T\psi+z, S\psi)-KT\psi-GS\psi)(t_1),$$

so the continuous differentiability of M now follows from Lemma 2.3 and the chain rule. Also, for all $\psi \in X$, $t \in [t_0-r, t_1]$,

$$\begin{aligned} (D_2M(0, 0)\psi)(t) &= (D_1H(0, 0)T\psi+D_2H(0, 0)S\psi-KT\psi-GS\psi)(t) \\ &\quad -(D_1H(0, 0)T\psi+D_2H(0, 0)S\psi-KT\psi-GS\psi)(t_1)=0, \end{aligned}$$

which shows that $D_2M(0, 0)=0$.

The operator $D_1M(0, 0)$ is given by $(D_1M(0, 0)z)(t)=\int_{t_1}^{\alpha(t)} A(s, z_s)\, ds$, so $D_1M(0, 0)$ is a compact operator, by the same type of argument as used in the proof of Lemma 2.4. It follows from (H2) that $z=D_1M(0, 0)z$ has only the solution $z=0$. Thus, by the Fredholm alternative, $(I-D_1M(0,0))^{-1}$ exists. Since $M(0,0)=0$, we can now apply the implicit function theorem to solve the equation $z=M(z,\psi)$: there exists an open ball N_3 about the origin in X and a continuously differentiable function $\zeta\colon N_3\to C([t_0-r,t_1],R^n)$ such that $\zeta(0)=0$ and $\zeta(\psi)=M(\zeta(\psi),\psi)$ for all $\psi\in N_3$. We may and shall assume that N_3 is chosen so that for any $\psi\in N_3$ we have $S\psi\in N_2$ and $\psi+\zeta(\psi)_{t_0}\in N_1$.

For fixed $\psi\in N_3$, let $z=\zeta(\psi)$, $u=S\psi$ and $y=T\psi$. Since $z=M(z,\psi)$, (2.28) holds for $t_0\leqq t\leqq t_1$ a.e., z is constant on $[t_0-r,t_0]$ and $z(t_1)=0$. By (H1), $z_{t_1}=0$. Let $\eta(t,\theta)$ be as in (A8) and denote by $\mathscr{Z}$ the intersection of the null spaces of $\eta(t,\theta)$, $t\in J$, $-r\leqq\theta\leqq 0$. By (A8), the ranges of f, A and B are contained in $\mathscr{Z}$. Since $z(t_1)=0$, it follows from (2.28) that $z(t)\in\mathscr{Z}$ for all $t\in J$, and since z is constant on $[t_0-r,t_0]$, $z(t)\in\mathscr{Z}$ for $t_0-r\leqq t\leqq t_1$. Thus, from (2.28) we get

$$\dot z(t)=f(t,y_t+z_t,u(t))+L(t,z_t)-A(t,y_t)-B(t)u(t)\qquad (t_0\leqq t\leqq t_1\text{ a.e.}).\tag{2.32}$$

Since $y=p_L(\psi,u)$, we have

$$\dot y(t)=A(t,y_t)+L(t,y_t)+(B(t)+C(t))u(t)\qquad (t_0\leqq t\leqq t_1\text{ a.e.}).\tag{2.33}$$

Letting $x=y+z$ and adding (2.32) and (2.33) gives (2.1). Since $x_{t_0}=y_{t_0}+z_{t_0}=\psi+\zeta(\psi)_{t_0}$, we conclude that $x=p(\psi+\zeta(\psi)_{t_0},S\psi)$. Thus, $S\psi$ steers $\psi+\zeta(\psi)_{t_0}$ to $x_{t_1}=y_{t_1}+z_{t_1}=0$ at time t_1 for the system (2.1).

Define $\omega\colon N_3\to X$ by

$$\omega(\psi)=\psi+\zeta(\psi)_{t_0}.\tag{2.34}$$

We claim that the range of ω covers a neighborhood of the origin in X. To see this, differentiate $\zeta(\psi)=M(\zeta(\psi),\psi)$ at $\psi=0$, giving $D\zeta(0)=D_1M(0,0)D\zeta(0)+D_2M(0,0)$. Since $D_2M(0,0)=0$ and since $(I-D_1M(0,0))^{-1}$ exists, we get $D\zeta(0)=0$. Now by (2.34) and the chain rule, $D\omega(0)=I$. Furthermore $\omega(0)=0$, so by the implicit function theorem, the claim follows. For each $\psi\in N_3$, $S\psi$ steers $\omega(\psi)$ to $0\in X$ at time t_1 for (2.1), and this proves the theorem.

Proof of Theorem 2.1. For system (2.9), equation (2.28) tākes the form

$$\begin{aligned}\dot z^1(t)&=f_1(t,y^1(t)+z^1(t),y_t^2+z_t^2,u(t))-A_1(t)y^1(t)-Q(t,y_t^2)-B_1(t)u(t),\\ \dot z^2(t)&=f_2(t,y^2(t)+z^2(t),u(t))-A_2(t)y^2(t)-B_2(t)u(t)\end{aligned}\tag{2.35}$$

and (2.29) takes the form

$$\begin{aligned}\dot z^1(t)&=A_1(t)z^1(t)+Q(t,z_t^2)\qquad (t_0\leqq t\leqq t_1\text{ a.e.})\\ \dot z^2(t)&=A_2(t)z^2(t)\\ z(t_1)&=0\end{aligned}\tag{2.36}$$

where $A_1(t)=D_2f_1(t,0,0,0)$, $Q(t,\cdot)=D_3f_1(t,0,0,0)$, $B_1(t)=D_4f_1(t,0,0,0)$, $A_2(t)=D_2f_2(t,0,0)$, and $B_2(t)=D_3f_2(t,0,0)$.

Clearly the only solution of (2.36) on $[t_0-r,t_1]$ which is constant on $[t_0-r,t_0]$ is $z(t)\equiv 0$, so (H2) holds. If $u(t)=0$ for $t_1-2r\leqq t\leqq t_1$ and $y(t)=0$ for $t_1-3r\leqq t\leqq t_1$, then the second equation of (2.35) shows that if $z(t_1)=0$ then $z^2(t)=0$ for $t_1-2r\leqq t\leqq t_1$. We can then conclude from the first equation of (2.35) that $z^1(t)=0$ for $t_1-r\leqq t\leqq t_1$. Hence (H1) holds, with $\bar t=t_1-2r$. By Theorem 2.7, the null controllability of (2.7) on $[t_0,t_1-2r]$ implies the local null controllability of (2.1) on $[t_0,t_1]$, completing the proof.

In the above proof, note that we never used the fact that $y(t)=0$ for $t_1-3r\leqq t<t_1-2r$. The only reason for requiring null controllability of (2.7) on $[t_0,t_1-2r]$ instead of

just on $[t_0, t_1 - r]$ was so that we could take $u(t) = 0$ on $[t_1 - 2r, t_1]$ in the second equation of (2.35). Clearly, if f_2 is independent of its third argument, then we need only have $u(t) = 0$ on $[t_1 - r, t_1]$ and we can take $\bar{t} = t_1 - r$. Thus, we have the following theorem.

THEOREM 2.8. *Let the hypotheses of Theorem* 2.1 *hold, except require only that* $t_1 - t_0 > 2r$, *and assume that* f_2 *is independent of its third argument. Then the null controllability of* (2.7) *on* $[t_0, t_1 - r]$ *implies the local null controllability of* (2.1) *on* $[t_0, t_1]$.

The following theorem relates to systems such as (2.10).

THEOREM 2.9. *Let* $L: J \times X \to R^n$ *satisfy* (A5)–(A7) *and let* $C: J \to \mathcal{M}_{nm}$ *be square-integrable. Let* $g: J \times N \times R^m \to R^n$ *satisfy the following*:

(i) $g(t, \cdot, \cdot)$ *is continuously differentiable for each* $t \in J$.

(ii) $g(\cdot, x, w)$ *is measurable for all* $(x, w) \in N \times R^m$.

(iii) *For each compact set* $K \subset N$ *there exists an integrable function* $M_1: J \to [0, \infty)$, *and square integrable functions* $M_i: J \to [0, \infty)$, $i = 2, 3$, *such that*

$$\|D_2 g(t, x, w)\| \leqq M_1 + M_2(t)|w|,$$
$$\|D_3 g(t, x, w)\| \leqq M_3(t)$$

for all $t \in J$, $w \in R^m$, $x \in K$.

(iv) $g(t, 0, 0) = 0$ *for all* $t \in J$.

Let $\eta(t, \theta)$ *be as in* (A8) *and assume that the range of g is contained in the null space of* $\eta(t, \theta)$ *for each* $t \in J$, $\theta \in [-r, 0]$. *Let* $A(t) = D_2 g(t, 0, 0)$ *and* $B(t) = D_3 g(t, 0, 0)$. *Then the system*

$$\dot{x}(t) = g(t, x(t), u(t)) + L(t, x_t) + C(t)u(t) \tag{2.37}$$

is locally null controllable on $[t_0, t_1]$ *if either of the following two conditions holds*:

(a) $t_1 - t_0 > 2r$ *and the system*

$$\dot{x}(t) = A(t)x(t) + L(t, x_t) + (B(t) + C(t))u(t) \tag{2.38}$$

is null controllable on $[t_0, t_1 - r]$, *or*

(b) *g is independent of its third argument and* (2.38) *is null controllable on* $[t_0, t_1]$.

Proof. Note that system (2.37) is in fact of the form (2.1), with $f(t, x_t, u(t)) = g(t, x(t), u(t))$. For system (2.37), equation (2.29) is equivalent to a linear ordinary differential equation, and it follows easily that (H2) is satisfied. Equation (2.28) reduces to

$$\dot{z}(t) = g(t, y(t) + z(t), u(t)) - A(t)y(t) - B(t)u(t).$$

Assuming that $y(t) = 0$ and $u(t) = 0$ for $t_1 - r \leqq t \leqq t_1$ and that $z(t_1) = 0$, we see that $z(t) = 0$ for $t_1 - r \leqq t \leqq t_1$. Hence, when condition (a) holds, (H1) is satisfied with $\bar{t} = t_1 - r$. If g is independent of its third argument, it is necessary only to assume that $y(t) = 0$ for $t_1 - r \leqq t \leqq t_1$ and that $z(t_1) = 0$ in order to conclude that $z(t) = 0$ for $t_1 - r \leqq t \leqq t_1$. Thus, (H1) is satisfied under condition (b) with $\bar{t} = t_1$. Appealing to Theorem 2.7, we now get the desired local null controllability of (2.37) under either condition (a) or (b).

In the following example, Theorem 2.7 is applied directly. None of Theorems 2.1, 2.8 or 2.9 is applicable.

Example 3. Consider the system

$$\begin{aligned} \dot{x}_1(t) &= x_1(t)^2 + x_2(t) + x_2(t-1)^2, \\ \dot{x}_2(t) &= \sin x_3(t-1) + x_2(t)^2 + u(t), \\ \dot{x}_3(t) &= x_3(t) + x_3(t-1) + u(t). \end{aligned} \tag{2.39}$$

This system is of the form (2.1), with

$$L(t,\phi)=\begin{bmatrix}0\\0\\\phi_3(0)+\phi_3(-1)\end{bmatrix},\qquad f(t,\phi,w)=\begin{bmatrix}\phi_1(0)^2+\phi_2(0)+\phi_2(-1)^2\\\sin\phi_3(-1)+\phi_2(0)^2\\0\end{bmatrix}.$$

Clearly (A1)–(A8) are satisfied. Equations (2.28) and (2.29) take the form

$$\begin{aligned}&\dot z_1(t)=(y_1(t)+z_1(t))^2+z_2(t)+(y_2(t-1)+z_2(t-1))^2,\\ &\dot z_2(t)=\sin(y_3(t-1)+z_3(t-1))+(y_2(t)+z_2(t))^2-y_3(t-1),\\ &\dot z_3(t)=0\end{aligned}\tag{2.40}$$

and

$$\begin{aligned}&\dot z_1(t)=z_2(t),\\ &\dot z_2(t)=z_3(t-1),\\ &\dot z_3(t)=0\end{aligned}\tag{2.41}$$

respectively.

Suppose $[t_0, t_1]$ is of length greater than 3, and let $\bar t = t_1 - 2$. The linear approximation to (2.39) is

$$\begin{aligned}&\dot x_1(t)=x_2(t),\\ &\dot x_2(t)=x_3(t-1)+u(t),\\ &\dot x_3(t)=x_3(t)+x_3(t-1)+u(t)\end{aligned}$$

which is the same as the linear approximation in Example 1. As pointed out there, this system is null controllable on $[t_0, \bar t]$ for any $\bar t > t_0+1$. Hence, we need only verify that (H1) and (H2) hold for (2.39).

It is obvious from (2.41) that (H2) holds. To check (H1), suppose $y \in C([t_0-1, t_1], R^3)$ satisfies $y(t)=0$ for $\bar t - 1 \leqq t \leqq t_1$, and suppose z is a solution of (2.40) on $[t_0-1, t_1]$ satisfying $z(t_1)=0$. Then $z_3(t)=0$ for $t_0 \leqq t \leqq t_1$, so the second equation of (2.40) shows that $z_2(t)=0$ for $\bar t \leqq t \leqq t_1$. Now the first equation of (2.40) shows that $z_1(t)=0, \bar t+1 \leqq t \leqq t_1$. Hence $z(t)=0$ for $\bar t+1 \leqq t \leqq t_1$, which can be rewritten as $z_{t_1}=0$. Thus, (H1) holds, and we can conclude by Theorem 2.7 that (2.39) is locally null controllable on $[t_0, t_1]$.

System (2.39) can be modified slightly to illustrate a technique of using the control to "blot out" certain terms, thus reducing a system to a simpler one. This is shown in the next example. Such a technique is helpful in various situations. It will be used again in § 3 to reduce system (3.5) to system (3.3). This same technique is essentially the one used in the proof of [1, Prop. 3.1].

Example 4. Consider the system

$$\begin{aligned}&\dot x_1(t)=x_1(t)^2+x_2(t)+x_2(t-1)^2,\\ &\dot x_2(t)=\sin x_3(t-1)+x_2(t)^2+x_1(t-1)^2+u(t),\\ &\dot x_3(t)=x_3(t)+x_3(t-1)+x_1(t-1)^2+u(t).\end{aligned}\tag{2.42}$$

If the control $\hat u$ steers $\phi \in X$ to $0 \in X$ at time t_1 for the system (2.39) and if $\tilde x$ is the corresponding trajectory, then (2.42) is satisfied, with $x=\tilde x$ and $u(t)=\hat u(t)-\tilde x_1(t-1)^2$. Hence the control u steers ϕ to $0 \in X$ at time t_1 for system (2.42). By the results of Example 3, therefore, we conclude that (2.42) is locally null controllable on any interval $[t_0, t_1]$ of length greater than 3.

3. Cascading techniques. Application of Theorem 2.7 involves verification of null controllability of the linear approximation to the system being considered. In making such verification, Corollary 3.3 in [1] is often useful (e.g. Example 1 of § 2.1). In [1], "null controllability" is defined to mean null controllability on arbitrary intervals of length greater than r. With this definition, [1] gives the definitive result on the null controllability of systems of the form

$$\dot{x}(t)=\sum_{i=0}^{k} A_i x(t-r_i)+Bu(t). \tag{3.1}$$

In the present paper, however, we are dealing with null controllability on a fixed interval $[t_0, t_1]$ rather than with null controllability as defined in [1]. Hence, additional results are desirable, besides those of [1], for determining null controllability of linear systems like (3.1).

One such additional result, which applies to the case where $m=1$ and $b\in R^n$, is originally due to Kirillova and Churakova [7]. For the system

$$\dot{x}(t)=Ax(t-r)+bu(t) \tag{3.2}$$

to be null controllable on every interval $[t_0, t_1]$ of length greater than nr, it is necessary and sufficient that the matrix $[b, Ab, \cdots, A^{n-1}b]$ have full rank. An important special case, for which the stated matrix does have full rank, is the system

$$\begin{aligned}\dot{x}_i(t)&=x_{i+1}(t-r) \qquad\qquad (i=1,\cdots,n-1)\\ \dot{x}_n(t)&=u(t).\end{aligned} \tag{3.3}$$

In spite of the simplicity of (3.3), the null controllability of (3.3) actually implies the sufficiency part of Kirillova and Churakova's result regarding the general system (3.2). We shall demonstrate this below. Theorem 3.1, the main result of this section, can be applied to system (3.3). Hence, sufficiency in Kirillova and Churakova's result follows as a special case of our results. Another linear system, more general than (3.3) (and not included in the systems considered by Kirillova and Churakova), to which Theorem 3.1 will apply is

$$\begin{aligned}\dot{x}_i(t)&=a_{ii}x_i(t)+a_{i,i+1}x_{i+1}(t-q)+a_{i,i+2}x_{i+2}(t-2q)+\cdots+a_{in}x_n(t-(n-i)q)\\ &\qquad\qquad\qquad\qquad (i=1,\cdots,n-1)\\ \dot{x}_n(t)&=a_{nn}x_n(t)+u(t).\end{aligned} \tag{3.4}$$

Theorem 3.1 applies to nonlinear systems as well as linear. In some cases, null controllability of a nonlinear system can be determined by using Theorem 3.1 directly, rather than by first taking the linear approximation. See Example 1 below.

As mentioned above, the null controllability of (3.2) can be checked by using the fact that (3.3) is null controllable. To show this, let $A\in\mathcal{M}_{nn}$, let $b\in R^n$, suppose $t_1>t_0+nr$ and suppose $[b, Ab, \cdots, A^{n-1}b]$ has full rank. By a suitable linear change of coordinates in R^n, (3.2) may be brought to the form

$$\dot{z}(t)=\begin{bmatrix}0&1&0&\cdots&0\\0&0&1&\cdots&0\\ \vdots&&&&\\0&0&0&\cdots&1\\ \alpha_1&\alpha_2&\alpha_3&\cdots&\alpha_n\end{bmatrix}z(t-r)+\begin{bmatrix}0\\0\\ \vdots\\0\\1\end{bmatrix}u(t) \tag{3.5}$$

where $x(t)=Pz(t)$ for some invertible $P\in\mathcal{M}_{nn}$. This follows from the assumed rank condition [8, pp. 90–91]. Suppose $\phi\in X$. Then there exists a $\hat{u}\in L^2(J, R)$ steering ϕ to $0\in X$ at time t_1 for the system (3.3). Let z be the corresponding trajectory. Then $u(t)=\hat{u}(t)-\alpha_1 z_1(t-r)-\cdots-\alpha_n z_n(t-r)$ steers ϕ to $0\in X$ at time t_1 for the system (3.5). Hence (3.5) is null controllable on $[t_0, t_1]$, and we draw the same conclusion for (3.2).

Theorem 3.1 extends the situation of system (3.4) in two ways. First, a nonlinear analogue of (3.4) is considered, and second, the control is allowed to occur in other than just the final equation. This second extension will be made possible by introducing lags in the control.

In § 2, the definition of null controllability was given for the system

$$\dot{x}(t)=F(t, x_t, u(t)) \tag{3.6}$$

where F, defined on an appropriate domain, satisfied (A1)–(A4). For a system with lags in the control, we shall need to modify this definition slightly. Let k be a positive integer, let N be an open convex neighborhood of the origin in R^n and let $F\colon J\times C([-r, 0], N)\times R^{km}\to R^n$ satisfy (A1)–(A4). Let $h\geqq 0$ and let $h_1, \cdots, h_k\in[0, h]$, with $h_i\neq h_j$ for $i\neq j$. For any $u\in L^2(J, R^m)$, define $\tilde{u}$ to be the extension of u to $[t_0-h, t_1]$ which is zero on $[t_0-h, t_0)$. We generalize (3.6) to the control system

$$\dot{x}(t)=F(t, x_t, (\Gamma u)(t)), \tag{3.7}$$

where $\Gamma\colon L^2(J, R^m)\to L^2(J, R^{km})$ is the operator given by

$$(\Gamma u)(t)=\begin{bmatrix}\tilde{u}(t-h_1)\\ \vdots\\ \tilde{u}(t-h_k)\end{bmatrix} \qquad (t_0\leqq t\leqq t_1). \tag{3.8}$$

Just as in the case of (3.6), we say that a control $u\in L^2(J, R^m)$ *steers* $\phi\in X$ to $\psi\in X$ at some time $\bar{t}\in(t_0, t_1]$ for the system (3.7) if there exists a trajectory x of (3.7) corresponding to u, with $x_{t_0}=\phi$ and $x_{\bar{t}}=\psi$. Define

$$U_{\bar{t}}=\{u\in L^2(J, R^m): u(t)=0 \text{ for } t\in J\cap[\bar{t}-h, t_1]\}. \tag{3.9}$$

For $\bar{t}\in(t_0, t_1]$, system (3.7) will be called *null controllable* on $[t_0, \bar{t}]$ if for every $\phi\in X$ there exists a control $u\in U_{\bar{t}}$ which steers ϕ to $0\in X$ at time $\bar{t}$. System (3.7) will be called *locally null controllable* on $[t_0, \bar{t}]$ if such a u exists for each ϕ in some open neighborhood of the origin in X. Note that in these definitions, if $u(t)=0$ were only required for $\bar{t}<t\leqq t_1$, we could not necessarily guarantee that the trajectory corresponding to u and ϕ would remain zero on $(\bar{t}, t_1]$, due to the lags in the control. This is the reason for giving the definitions as we have.

In proving Theorem 3.1, we shall need an extension of Proposition 2.5 applicable to the system (3.7). Define $p(\phi, u)\in C([t_0-r, t_1], R^n)$ to be the solution of

$$\dot{x}(t)=F(t, x_t, (\Gamma u)(t)),$$
$$x_{t_0}=\phi$$

on $[t_0-r, t_1]$ for any pair $(\phi, u)\in X\times L^2(J, R^m)$ such that that solution exists. Then $p(\phi, u)$ exists for all pairs (ϕ, u) in some open neighborhood of the origin in $X\times L^2(J, R^m)$, and on this neighborhood, p is continuously differentiable. The proof of this result is constructed exactly as is the proof of Proposition 2.5, with the operator H

given by

$$H(x, u)(t)=\int_{t_0}^{\alpha(t)} F(s, x_s, (\Gamma u)(s))\, ds.$$

THEOREM 3.1. *Let $q>0$. For each $i=1, \cdots, n$, let $r_i \in \{0, \cdots, n-i\}$, let $N_i \subset R^{r_i+1}$ be an open convex set containing the origin, and let $g_i\colon N_i \times R^m \to R$ be continuously differentiable and satisfy $g_i(0, 0)=0$. Assume that $r_i>0$ for at least one value of i. Let $r=\max_i r_i q$, suppose $t_1-t_0>r+(n-1)q$, and suppose that the ordinary differential control system*

$$\dot{y}_i=g_i(y_i, y_{i+1}, \cdots, y_{i+r_i}, u) \qquad (i=1, \cdots, n) \tag{3.10}$$

is locally null controllable (in the pointwise sense usual for such systems) on $[t_0, t_1-r-(n-1)q]$. Then the control system

$$\dot{x}_i(t)=g_i(x_i(t), x_{i+1}(t-q), \cdots, x_{i+r_i}(t-r_iq), \tilde{u}(t-(n-i)q)) \qquad (i=1, \cdots, n) \tag{3.11}$$

is locally null controllable on $[t_0, t_1]$ in the sense defined for system (3.7).

Note. It should be clear that for some F satisfying (A1)–(A4) defined on the appropriate domain and for some Γ of the form (3.8), system (3.11) is indeed of the form (3.7). In particular, r is as defined in the statement of the theorem, $h=(n-1)q$ and $k=n$.

Proof. Let $\bar{t}=t_1-r-(n-1)q$. Let N_1 be an open neighborhood of the origin in R^n with the property that for any $\eta \in N_1$ there exists a $u \in L^2(J, R^m)$ steering $y(t_0)=\eta$ to $y(\bar{t})=0$ for the system (3.10). The solution $x=p(\phi, u)$ of (3.11) satisfying $x_{t_0}=\phi$ exists on $[t_0-r, t_1]$ for all (ϕ, u) in some open neighborhood of the origin in $X \times L^2(J, R^m)$, and x depends continuously on (ϕ, u) in this neighborhood. In particular, there exists an open neighborhood N_2 of the origin in X such that for all $\phi \in N_2$, $x=p(\phi, 0)$ exists and $x(t) \in N_1$ for $t_0-r \leqq t \leqq t_1$. Fix $\phi \in N_2$ and let $\hat{x}=p(\phi, 0)$. We shall assume (without loss of generality) that $N_1=\{x \in R^n\colon |x_i|<\varepsilon, i=1, \cdots, n\}$ for some $\varepsilon>0$.

Let $\eta=(\hat{x}_1(t_0+(n-1)q, \hat{x}_2(t_0+(n-2)q), \cdots, \hat{x}_n(t_0))$, let $u \in L^2(J, R^m)$ steer $y(t_0)=\eta$ to $y(\bar{t})=0$ for the system (3.10) and satisfy $u(t)=0$ for $\bar{t}<t \leqq t_1$, and let $y \in C([t_0, t_1], R^n)$ be the corresponding trajectory. Note that $y(t)=0$ for $\bar{t} \leqq t \leqq t_1$. We shall show that u steers ϕ to $0 \in X$ at time t_1 for the system (3.11). This will complete the proof, since it is readily seen that $u \in U_{t_1}$ and since $\phi \in N_2$ is arbitrary.

Define $x \in C([t_0-r, t_1], R^n)$ by

$$x_i(t)=\begin{cases} \hat{x}_i(t) & \text{if } t_0-r \leqq t \leqq t_0+(n-i)q, \\ y_i(t-(n-i)q) & \text{if } t_0+(n-i)q \leqq t \leqq t_1, \end{cases} \tag{3.12}$$

$i=1, \cdots, n$. Then for $t_0 \leqq t \leqq t_0+(n-i)q$ a.e., we have

$$\begin{aligned} \dot{x}_i(t) &= \dot{\hat{x}}_i(i) \\ &= g_i(\hat{x}_i(t), \cdots, \hat{x}_{i+r_i}(t-r_iq), 0) \\ &= g_i(x_i(t), \cdots, x_{i+r_i}(t-r_iq), \tilde{u}(t-(n-i)q)), \end{aligned}$$

and for $t_0+(n-i)q \leqq t \leqq t_1$ a.e., we have

$$\begin{aligned} \dot{x}_i(t) &= \dot{y}_i(t-(n-i)q) \\ &= g_i(y_i(t-(n-i)q), y_{i+1}(t-q-(n-(i+1))q), \cdots, y_{i+r_i}(t-r_iq \\ &\quad -(n-(i+r_i))q), \tilde{u}(t-(n-i)q)) \\ &= g_i(x_i(t), x_{i+1}(t-q), \cdots, x_{i+r_i}(t-r_iq), \tilde{u}(t-(n-i)q)). \end{aligned}$$

Thus, (3.11) holds for $t_0 \leqq t \leqq t_1$ a.e. Furthermore, $x_{t_0} = \hat{x}_{t_0} = \phi$, and for $t_1 - r \leqq t \leqq t_1$, $x_i(t) = y_i(t-(n-i)q) = 0$, $i = 1, \cdots, n$. Hence u steers ϕ to $0 \in X$ at time t_1 for the system (3.11).

Example 1. Consider the system

$$\begin{aligned} \dot{x}_1(t) &= -x_1(t) - x_2(t-1)^3, \\ \dot{x}_2(t) &= -x_2(t) + u(t). \end{aligned} \tag{3.13}$$

Local null controllability of this system cannot be established by using Theorem 2.7, due to the fact that the linear approximation is not null controllable. However, we can compare (3.13) to the system

$$\begin{aligned} \dot{y}_1 &= -y_1 - y_2^3, \\ \dot{y}_2 &= -y_2 + u, \end{aligned} \tag{3.14}$$

which is locally null controllable on $[t_0, \bar{t}]$ for any $\bar{t} > t_0$. This can be shown by a modification of the argument given in [8, p. 365]. (In [8] it is shown only that there *exists* a $\bar{t} > t_0$ such that (3.14) is locally null controllable on $[t_0, \bar{t}]$.) We can therefore conclude that (3.13) is locally null controllable on $[t_0, t_1]$ for any $t_1 > t_0 + 2$. (Note that in applying Theorem 3.1 to (3.13), we must take $h = 1$, while the natural h is $h = 0$. This is no drawback, however, since the conclusion of Theorem 3.1 gives local null controllability in the sense defined for system (3.7) with $h = 1$. This is a stronger result than local null controllability with $h = 0$.)

At the beginning of this section, we discussed how a result such as Theorem 3.1 could be used in conjunction with Theorem 2.7. Even more can be gained in this regard by extending Theorem 2.7 to include systems with lags in the controls. We consider the extension of (2.1) given by

$$\dot{x}(t) = f(t, x_t, (\Gamma u)(t)) + L(t, x_t) + C(t)(\Gamma u)(t) \tag{3.15}$$

and its linear approximation

$$\dot{x}(t) = A(t, x_t) + L(t, x_t) + (B(t) + C(t))(\Gamma u)(t), \tag{3.16}$$

where f, L and C are as in § 2, except that $C \in L^2(J, \mathcal{M}_{n,km})$ and f is defined on $J \times C([-r, 0], N) \times R^{km}$, where $N \subset R^n$ is an open convex set containing the origin. The appropriate modification of (H1) here is as follows:

(H1′) For any $u \in U_{\bar{t}}$ and any $y \in C([t_0 - r, t_1], N)$ satisfying $y(t) = 0$ for $\bar{t} - r \leqq t \leqq t_1$, there exists no solution z of

$$\dot{z}(t) = f(t, y_t + z_t, (\Gamma u)(t)) - A(t, y_t) - B(t)(\Gamma u)(t) \tag{3.17}$$

on $[t_0 - r, t_1]$ which satisfies both $z(t_1) = 0$ and $z_{t_1} \neq 0$.

Hypothesis (H2) is as in § 2. We restate it here for convenience.

(H2) The only solution $z \in C([t_0 - r, t_1], R^n)$ of

$$\begin{aligned} \dot{z}(t) &= A(t, z_t) \qquad (t_0 \leqq t \leqq t_1 \text{ a.e.}) \\ z(t_1) &= 0 \end{aligned} \tag{3.18}$$

which is constant on $[t_0 - r, t_0]$ is $z = 0$.

THEOREM 3.2. *Let* (A1)–(A8) *hold for the system* (3.15). *Let* $A(t, \cdot) = D_2 f(t, 0, 0)$ *and* $B(t) = D_3 f(t, 0, 0)$. *Let* $\bar{t} \in (t_0 + r, t_1]$, *suppose* (H1′) *holds for this* $\bar{t}$ *and suppose* (H2) *holds. Then the null controllability of* (3.16) *on* $[t_0, \bar{t}]$ (*in the sense defined for systems* (3.7)) *implies the local null controllability of* (3.15) *on* $[t_0, t_1]$ (*also in that sense*).

The details of the proof will be omitted. We point out only that the space U occurring in the proof of Theorem 2.7 must be replaced here by $U_{\bar{t}}$ defined by (3.9) and that Lemma 2.6 can be extended in a natural way to establish the existence of an operator $S \in \mathscr{B}(X, U_{\bar{t}})$ such that for any $\phi \in X$ the solution y of

$$\begin{aligned}&\dot{y}(t) = A(t, y_t) + L(t, y_t) + (B(t) + C(t))(\Gamma S\phi)(t),\\ &y_{t_0} = \phi\end{aligned}$$

satisfies $y(t) = 0$ for $\bar{t} - r \leqq t \leqq t_1$. Using this operator S, one proceeds in a manner analogous to the proof of Theorem 2.7.

Note that in applying Theorem 3.2, the same value of r must be used for both (3.15) and (3.16), since the state space for both systems must be the same.

Before presenting the next example, we recall that for linear systems global null controllability and local null controllability are equivalent. Hence, for linear systems Theorem 3.1 holds with the word "locally" deleted.

Example 2. Assume $t_1 > t_0 + 10$, and consider the system

$$\begin{aligned}&\dot{x}_1(t) = x_1(t) + x_2(t-1) + x_3(t-2)^2 + u(t-2).\\ &\dot{x}_2(t) = x_2(t) + u(t-1),\\ &\dot{x}_3(t) = x_2(t-3)^3 + u(t).\end{aligned} \tag{3.19}$$

This is a system of the form (3.15), with

$$L(t, \phi) = \begin{bmatrix}\phi_2(-1)\\ \phi_2(0)\\ 0\end{bmatrix}, \qquad (\Gamma u)(t) = \begin{bmatrix}\tilde{u}(t-2)\\ \tilde{u}(t-1)\\ \tilde{u}(t)\end{bmatrix},$$

$$f(t, \phi, w) = \begin{bmatrix}\phi_1(0) + \phi_3(-2)^2\\ 0\\ \phi_2(-3)^3\end{bmatrix}, \qquad C = I.$$

It is readily seen that (A1)–(A8) hold.

The linear approximation to (3.19) is

$$\begin{aligned}&\dot{x}_1(t) = x_1(t) + x_2(t-1) + u(t-2),\\ &\dot{x}_2(t) = x_2(t) + u(t-1),\\ &\dot{x}_3(t) = u(t).\end{aligned} \tag{3.20}$$

By Theorem 3.1, (3.20) is null controllable on $[t_0, \bar{t}]$ for any $\bar{t} > t_0 + 3$. However, in this application of Theorem 3.1, $r = 1$ is used, and the maximum delay in system (3.19) is $r = 3$. For $r = 3$, $h = 2$ (so that the state space is $C([-3, 0], R^3)$) we can conclude only that (3.20) is null controllable on $[t_0, \bar{t}]$ for any $\bar{t} > t_0 + 5$. In particular, we take $\bar{t} = t_1 - 5$. We shall show that (H1′) holds for this $\bar{t}$ and that (H2) holds.

Let $A(t, \cdot) = D_2 f(t, 0, 0)$ and $B(t) = D_3 f(t, 0, 0)$. Then (3.18) reduces to an initial value problem for an ordinary differential equation, so (H2) is obviously satisfied. Equation (3.17) becomes

$$\begin{aligned}&\dot{z}_1(t) = z_1(t) + (y_3(t-2) + z_3(t-2))^2,\\ &\dot{z}_2(t) = 0,\\ &\dot{z}_3(t) = (y_2(t-3) + z_2(t-3))^3.\end{aligned} \tag{3.21}$$

Assume $y \in C([t_0 - r, t_1], R^3)$ satisfies $y(t) = 0$ for $\bar{t} - 3 \leqq t \leqq t_1$ and assume z is a solution of (3.21) on $[t_0 - r, t_1]$ with $z(t_1) = 0$. To show that (H1′) holds, we must show that $z(t) = 0$ for $t_1 - 3 \leqq t \leqq t_1$. Clearly $z_2(t) = 0$ for $t_0 \leqq t \leqq t_1$, and so the third equation of (3.21) shows that $z_3(t) = 0$ for $\bar{t} \leqq t \leqq t_1$. Thus, for $\bar{t} + 2 \leqq t \leqq t_1$, we have $z_3(t-2) = y_3(t-2) = 0$. Thus, the first equation of (3.21) gives $z_1(t) = 0$ for $\bar{t} + 2 \leqq t \leqq t_1$. We conclude that $z(t) = 0$ for $\bar{t} + 2 \leqq t \leqq t_1$, which shows that (H1′) holds. (Here one sees that it was necessary to have $t_1 > t_0 + 10$.) We can now apply Theorem 3.2 to see that (3.19) is locally null controllable on $[t_0, t_1]$.

The above two examples and the other examples throughout this paper illustrate the fact that there are many nonlinear control systems involving functional differential equations which are indeed locally null controllable. However, the lack of a theorem analogous to Theorem 1 of Chapter 6 in [8] makes the possibility of a comprehensive theory of null controllability doubtful for such systems. The most that one can reasonably hope to find are various techniques, such as the ones gives in this paper, which are applicable to systems of specific types.

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. C. DELFOUR AND S. K. MITTER, *Controllability, observability and optimal feedback control of affine hereditary differential systems*, this Journal, 10 (1972), pp. 298–328.

[3] J. DIEUDONNÉ, *Foundations of Modern Analysis*, Academic Press, New York, 1969.

[4] N. DUNFORD AND J. SCHWARTZ, *Linear Operators*, part I, Interscience, New York, 1958.

[5] J. HALE, *Functional Differential Equations*, Springer-Verlag, New York, 1971.

[6] H. HERMES AND J. P. LA SALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.

[7] F. M. KIRILLOVA AND S. V. CHURAKOVA, *The controllability problem for linear systems with aftereffect*, Differentsialnye Uravneniya, 3 (1967), pp. 436–445.

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

[9] A. MANITIUS AND R. TRIGGIANI, *New results on functional controllability of time-delay systems*, Technical report, Centre de Recherches Mathématiques, Université de Montréal, 1976.

[10] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.

[11] W. RUDIN, *Functional Analysis*, McGraw-Hill, New York, 1966.

[12] L. WEISS, *On the controllability of delay-differential systems*, this Journal, 5 (1967), pp. 575–587.

[13] R. B. ZMOOD, *The Euclidean space controllability of control systems with delay*, this Journal, 12 (1974), pp. 609–623.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 17, No. 6, November 1979

0363-0129/79/1706-0008 $01.00/0

EQUILIBRIUM POINTS IN NONZERO-SUM n-PERSON SUBMODULAR GAMES*

DONALD M. TOPKIS†

Abstract. A submodular game is a finite noncooperative game in which the set of feasible joint decisions is a sublattice and the cost function of each player has properties of submodularity and antitone differences. Examples of submodular games include 1) a game version of a system with complementary products; 2) an extension of the minimum cut problem to a situation where players choose from different sets of nodes and perceive different capacities, with special cases being a game with players choosing whether or not to participate in available economic activities and a game version of the selection problem; 3) the pricing problem of competitors producing substitute products; 4) a game version of the facility location problem; and 5) a game with players determining their optimal usage of available products. A fixed point approach establishes the existence of a pure equilibrium point for certain submodular games. Two algorithms which correspond to fictitious play in dynamic games generate sequences of feasible joint decisions converging monotonically to a pure equilibrium point. Bounds show these algorithms to be very efficient when the set of feasible decisions is finite. An optimal decision for each player is an isotone function of the decisions of other players.

Introduction. Consider a noncooperative n-person game with the players indicated by $i=1,\cdots,n$. The decision of player i is an m_i-vector x_i. The joint decision is $x=(x_1,\cdots,x_n)$. The set of feasible joint decisions is a subset S of E^m where $m=\sum_{i=1}^n m_i$. The feasible decisions for a given player may depend upon the decisions chosen by the other players. (By assigning a very large or infinite cost to each m-vector not in S one could embed such a game into a game in which any m-vector is considered feasible, but that approach is not convenient here because it is not easy to transform subsequent assumptions about the players' cost functions into equivalent properties for such extended costs.) Let $x\sim x_i=(x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_n)$ be the vector of decisions of all players except player i. Let $(x;y_i)=(x_1,\cdots,x_{i-1},y_i,x_{i+1},\cdots,x_n)$ be the vector of joint decisions for all n players where y_i is the decision of player i and $x\sim x_i$ is the vector of decisions of the other $n-1$ players. The set of feasible decisions for player i given $x\sim x_i$ is $S_i(x)=\{y_i:(x;y_i)\in S\}$. The vectors $x\sim x_i$ and $(x;y_i)$ and the set $S_i(x)$ do not depend on x_i. Define $T_i=\{x\sim x_i:S_i(x)\text{ is nonempty}\}$ and $S_i=\bigcup_{x\sim x_i\in T_i}S_i(x)$. As a result of a joint decision $x\in S$, player i incurs the cost $f_i(x)$ where $f_i(x)$ is a real-valued function on S for $i=1,\cdots,n$. A feasible joint decision $x\in S$ is an *equilibrium point* if $f_i(x)\leqq f_i(x;y_i)$ for all $y_i\in S_i(x)$ and $i=1,\cdots,n$.

If x and y are real numbers, then $x\vee y=\max\{x,y\}$ and $x\wedge y=\min\{x,y\}$. If x and y are n-vectors, then $x\vee y=(x_1\vee y_1,\cdots,x_n\vee y_n)$ and $x\wedge y=(x_1\wedge y_1,\cdots,x_n\wedge y_n)$. If a subset L of E^n has the property that $x\in L$ and $y\in L$ imply that $x\vee y\in L$ and $x\wedge y\in L$, then L is a *sublattice* of E^n. If $f(x)$ is a real-valued function on a sublattice L of E^n and if $f(x\wedge y)+f(x\vee y)\leqq f(x)+f(y)$ for all $x\in L$ and $y\in L$, then $f(x)$ is *submodular* on L. If $f(x)$ is a function from $L\subseteq E^n$ into E^m and if $x\leqq y$ in L implies $f(x)\leqq f(y)(f(y)\leqq f(x))$, then $f(x)$ is *isotone* (*antitone*) on L. If $f(x,y)$ is a real-valued function on $L\subseteq E^{n+m}$ where $x\in E^n$ and $y\in E^m$ and if $f(x,w)-f(x,y)$ is antitone in x on L for all $w\geqq y$, then $f(x,y)$ has *antitone differences in* (x,y).

A game is a *submodular game* if S is a nonempty sublattice of E^m, $f_i(x)$ is submodular in x_i on S_i for each $x\sim x_i\in T_i$ and each i, and $f_i(x)$ has antitone differences in $(x\sim x_i,x_i)$ on $T_i\times S_i$ for each i. This paper presents some examples of submodular

* Received by the editors October 10, 1978 and in revised form April 17, 1979.

† Bell Laboratories, Holmdel, New Jersey 07733. This research was supported in part by the National Science Foundation under Grant GP-25081.

games, gives conditions for the existence of an equilibrium point in a submodular game, and gives two algorithms for finding or approximating an equilibrium point in a submodular game.

Section 1 provides further definitions and background relevant to the consideration of submodular games in subsequent sections.

Section 2 presents a variety of examples of submodular games. Example (a) is a game version of a system with complementary products. Example (b) extends the minimum cut problem to a situation where n players choose from among different subsets of nodes and perceive different capacities. An application of the minimum cut game involves players choosing whether or not to participate in certain economic activities, with a special case being a game version of the selection problem. Example (c) considers the pricing problem of n competitors producing substitute products. Example (d) is a game version of the problem of locating new facilities in the plane. Example (e) is a game with the players determining their optimal usage of available products.

Section 3 uses a fixed point approach to establish the existence of an equilibrium point in certain submodular games. Actually, a greatest and a least equilibrium point exist.

Section 4 gives two algorithms for seeking an equilibrium point in a submodular game. The operation of the algorithms corresponds to fictitious play in certain dynamic games. Under conditions slightly stronger than required for the existence result of § 3, each algorithm generates a sequence of feasible joint decisions which converges monotonically to an equilibrium point. Each algorithm has an inherent check which determines whether or not an equilibrium point has been attained. When S contains a finite number of elements there are bounds on the number of iterations required by each algorithm to find an equilibrium point, and these bounds show the algorithms to be very efficient. In a submodular game an optimal decision x_i for player i is an isotone function of the decisions $x \sim x_i$ of the other players.

1. Further definitions and background. The partially overlapping results of Bergman [4] and Topkis [35] characterize the structure of sublattices of E^n. See also Baker and Pixley [1].

A subset L of E^n is a *chain* if $x \in L$ and $y \in L$ imply that either $x \leqq y$ or $y \leqq x$.

Suppose that L is a subset of E^{n+m} and elements of L are denoted (x, y) where $x \in E^n$ and $y \in E^m$. The *section* of L at $y \in E^m$ is $L_y = \{x : (x, y) \in L\}$. The *projection* of L on the last m coordinates is $\{y : L_y \text{ is nonempty}\}$. The sections and projections of a sublattice are also sublattices.

In terms of the game, $S_i(x)$ is the section of the feasible set S at $x \sim x_i$, T_i is the projection of S on the coordinates of $x \sim x_i$, and S_i is the projection of S on the coordinates of x_i. If S is a sublattice of E^m, then each $S_i(x)$ and S_i is a sublattice of E^{m_i} and each T_i is a sublattice of E^{m-m_i}.

If $x \in L \subseteq E^n$ and $y \leqq x (x \leqq y)$ for each $y \in L$, then x is the *greatest* (*least*) element of L. A topological result of Birkhoff [5] implies that a nonempty compact sublattice of E^n has a greatest element and a least element. See also Frink [13] and Topkis [36].

Applying a result of Birkhoff [5], Tarski's fixed point theorem [30] states in part that if L is a nonempty compact sublattice of E^n and $f(x)$ is an isotone function from L into L, then there exists a fixed point for $f(x)$ in L.

If X and Y are nonempty sublattices of E^n, then $X \leqq^p Y$ if $x \in X$ and $y \in Y$ imply that $x \wedge y \in X$ and $x \vee y \in Y$. Veinott [personal communication] introduced the relation $\leqq^p$. The collection of all nonempty sublattices of E^n together with the relation $\leqq^p$ is a partially ordered set [37]. If $\{L_y\}_{y \in Y}$ is a collection of nonempty sublattices of E^n for

$Y \subseteq E^m$ and if $y \leqq w$ in Y implies $L_y \leqq^p L_w$, then L_y is *ascending* in y on Y. The property that L_y is ascending is the same as the property that L_y is isotone with respect to the ordering relation $\leqq^p$. Veinott [personal communication] showed that the section of a sublattice is ascending in its argument. If L is a sublattice of E^n then $L_y = \{x : x \in L, x \leqq y\}$ and $L_y = \{x : x \in L, y \leqq x\}$ are ascending in y on $\{y : L_y \text{ is nonempty}\}$. If L_y is ascending in y on $Y \subseteq E^m$ and each L_y has a greatest (least) element $\bar{x}_y(\underline{x}_y)$, then $\bar{x}_y(\underline{x}_y)$ is isotone in y on Y [37].

If $f(x)$ is a real-valued function on $L \subseteq E^n$ and has antitone differences in (x_j, x_k) for all $j \neq k$ when each x_i is held fixed for $i \neq j$ and $i \neq k$, then $f(x)$ has *antitone differences on* L. Let u^i denote the ith unit vector in E^n. A function $f(x)$ has antitone differences on E^n if and only if $f(x + \varepsilon u^i) - f(x)$ is antitone in x_j for all $i \neq j$, $\varepsilon > 0$, and x. If $f(x)$ is differentiable on E^n, then $f(x)$ has antitone differences on E^n if and only if $\partial f(x)/\partial x_i$ is antitone in x_j for all $i \neq j$ and x. If $f(x)$ is twice differentiable on E^n, then $f(x)$ has antitone differences on E^n if and only if $\partial^2 f(x)/\partial x_i \partial x_j \leqq 0$ for all $i \neq j$ and x.

Theorem 1.1 shows an equivalence between antitone differences and submodularity. The property of antitone differences is more meaningful economically and is often easier to recognize, while the property of submodularity is more convenient to use mathematically. If $L_i \subseteq E^1$ for $i = 1, \cdots, n$ then $\times_{i=1}^n L_i$ is a *product set* in E^n.

THEOREM 1.1 [37]. *If $f(x)$ is submodular on the sublattice L of E^n, then $f(x)$ has antitone differences on L. If L is a product set in E^n and $f(x)$ has antitone differences on L, then $f(x)$ is submodular on L.*

Consider the collection of optimization problems

$$\text{(1)} \qquad \text{minimize } f(x, y) \quad \text{subject to} \quad x \in L_y \subseteq E^n,$$

where both the constraint set and the objective function depend on the parameter y for $y \in Y \subseteq E^m$. Let L_y^* be the set of optimal solutions for (1) given $y \in Y$, and let $Y^* = \{y : L_y^* \text{ is nonempty}\}$.

THEOREM 1.2 [37]. *If L is a sublattice of E^n, $L_y \subseteq L$ is ascending in y on $Y \subseteq E^m$, $f(x, y)$ is submodular in x on L for each $y \in Y$, and $f(x, y)$ has antitone differences in (x, y) on $L \times Y$, then L_y^* is ascending in y on Y^*. If, in addition, L_y is compact and $f(x, y)$ is lower semicontinuous in x on L_y for each y, then each L_y^* is a nonempty compact sublattice which has a greatest element $\bar{x}_y^*$ and a least element $\underline{x}_y^*$ and both $\bar{x}_y^*$ and $\underline{x}_y^*$ are isotone in y on Y.*

The conditions on $f(x, y)$ in Theorem 1.2 hold by Theorem 1.1 if $f(x, y)$ is twice differentiable on a convex product set containing $L \times Y$ and if $\partial^2 f(x, y)/\partial x_i \partial x_j \leqq 0$ for all $i \neq j$ and all (x, y) and $\partial^2 f(x, y)/\partial x_i y_k \leqq 0$ for all i and k and all (x, y).

The above results and concepts can be extended to include the case where the variable is a subset chosen from a given finite set N instead of being an n-vector. This can be done by defining an indicator vector with one component corresponding to each element of N so that this component takes on the value 1 if the corresponding element is in the chosen subset and 0 otherwise. The above results and concepts would then apply when the variables are subsets if and only if they apply when the variables are the corresponding indicator vectors.

A broader and more general discussion of the problem of minimizing a submodular function on a sublattice appears in [37]. Topkis [31], [33], [34], [37] and Veinott [personal communiction] have developed other applications on the theory of [37]. Maschler, Peleg, and Shapley [17] and Shapley [29] have analyzed n-person cooperative games in which minus the characteristic function is submodular. An earlier version of parts of § 3 and § 4 appeared in [32].

2. Examples of submodular games. This section gives various examples of classes of submodular games. Under the additional regularity conditions required in the next two sections, § 3 shows that an equilibrium point exists for each of these games and § 4 gives two algorithms for finding or approximating an equilibrium point for each of these games. Examples (b) and (d) note existing algorithms for solving those games for the special case $n = 1$ or, equivalently, for solving the problem of any player i where the decisions of the other $n-1$ players are given, and these algorithms can effectively solve the subproblems at each iteration for the algorithms of § 4.

Example (a). *Games with complementary products.* Antitone differences is a well-known condition [26] for a cost function to be that of a system of complementary products. Suppose that $f(x)$ is the cost function (or minus the utility function) for a system of n products whose levels are $x = (x_1, \cdots, x_n)$. Then $f(x+\varepsilon u^i)-f(x)$ is the additional cost for an additional $\varepsilon > 0$ units of product i. Antitone differences for $f(x)$ is equivalent to the property that the net additional cost for additional product i will not increase if there is more of product j where $j \neq i$; that is, the desirability of more product i will never decrease if there is an increase in the level of product j.

Now consider the n-player game in which each player i chooses an m_i-vector x_i of products, the vector of all products chosen $x = (x_1, \cdots, x_n)$ must be in a subset S of E^m where $m = \sum_{i=1}^n m_i$, and the choice of x results in a cost $f_i(x)$ for player i. Suppose that S is a sublattice of E^n, L is a product set in E^m with $S \subseteq L$, and each $f_i(x)$ has antitone differences in each pair (z_h, z_k) such that z_h is a component of x_i and z_k is a component of x other than z_h. The latter assumption means that from each player i's point of view (that is, with respect to that player's cost function) each product chosen by that player is complementary with all other products chosen by that player and by any other player. This is a submodular game by Theorem 1.1.

Example (b). *Minimum cut games.* Consider a network with a source s, a sink t, and a set N of m other nodes. The set N is divided into n disjoint sets of nodes $N_1, N_2, \cdots, N_n$, where N_i has m_i elements. Thus $N = \bigcup_{i=1}^n N_i$, $N_i \cap N_j = \emptyset$ for all $i \neq j$, $\sum_{i=1}^n m_i = m$, and the network has $m+2$ nodes. If $X \subseteq N$ then $X \cup \{s\}$ is a *cut.*

If there is a real-valued nonnegative capacity function $c(w, z)$ defined on each pair of nodes (w, z) from among $N \cup \{s\} \cup \{t\}$ then the associated cut capacity function is $f(X) = \sum_{w \in X \cup \{s\}} \sum_{z \notin X \cup \{s\}} c(w, z)$ where $X \subseteq N$. If $\hat{X}$ is optimal for the problem of minimizing $f(X)$ over $X \subseteq N$, then $\hat{X} \cup \{s\}$ is a *minimum cut.*

There are n players $i = 1, \cdots, n$. Each player is interested in optimizing that player's own minimum cut problem in this same network with nodes $N \cup \{s\} \cup \{t\}$. The different players are faced with different problems. Player i is only able to choose a set of nodes X_i from N_i. The other players choose the nodes $\bigcup_{j \neq i} X_j$ for the cut from $\bigcup_{j \neq i} N_j$. Player i has no control over those nodes from $\bigcup_{j \neq i} N_j$ in the cut and sees them as fixed. Each player i has a capacity function $c_i(w, z)$, and the associated cut capacity function is $f_i(X)$ for $X \subseteq N$. The problem of player i is to minimize $f_i(X_i \cup (\bigcup_{j \neq i} X_j))$ over subsets X_i of N_i where X_j is a predetermined subset of N_j for each $j \neq i$. This is the *minimum cut game.*

Ore [personal communication] pointed out that a cut capacity function is submodular on the power set of N. A proof is in [31], [33]. It follows directly from this result that $f_i(X_i \cup (\bigcup_{j \neq i} X_j))$ is submodular in X_i for $X_i \subseteq N_i$ and has antitone differences in (X_i, X_j) for $X_i \subseteq N_i$, $X_j \subseteq N_j$, and $j \neq i$. The hypotheses of the n-person submodular game thus hold for the minimum cut game.

Minimizing $f_i(X_i \cup (\bigcup_{j \neq i} X_j))$ over $X_i \subseteq N_i$ is equivalent to finding a minimum cut in a capacitated network with $m_i + 2$ nodes consisting of a source s', a sink t', and the nodes N_i, where the capacity function $c_i'(w, z)$ is such that $c_i'(w, z) = c_i(w, z)$ for

$w \in N_i$ and $z \in N_i$, $c'_i(s', z) = c_i(s, z) + \sum_{w \in \bigcup_{j \neq i} X_j} c_i(w, z)$ for $z \in N_i$, $c'_i(w, t') = c_i(w, t) + \sum_{z \in \bigcup_{j \neq i}(N_j \sim X_j)} c_i(w, z)$ for $w \in N_i$, and $c'_i(w, z) = 0$ otherwise. The minimum cut problem is dual to the maximum flow problem [11], [12]. Efficient algorithms are available for solving the maximum flow problem [9], [10], [15]. After finding a maximum flow, it is a simple matter to find the greatest minimum cut and the least minimum cut [12, pp. 11–14].

As an example of the minimum cut game, consider an n-person game in which player i chooses whether or not to participate in the m_i economic activities contained in the set N_i. The sets N_i for $i = 1, \cdots, n$ are disjoint and so $N = \bigcup_{i=1}^n N_i$ has $m = \sum_{i=1}^n m_i$ elements. The set of all activities N is divided into two sets S and T with $S \cup T = N$ and $S \cap T = \emptyset$. The activities of S are not in themselves profitable, so if player i chooses activity $w \in N_i \cap S$ then i incurs a net cost $c_w \geqq 0$. The activities of T are in themselves profitable, so if player i chooses activity $w \in N_i \cap T$ then i receives a net profit $v_w > 0$. Each activity available to a player is complementary with each activity available to that player or to any other player. Such complementarity could arise where an activity of one player may benefit from the use of transportation facilities, distribution outlets, or product inputs that become available as a result of other activities undertaken by the same player or other players. The effect of this complementarity between activities is that player i incurs an additional cost $b(w, z) \geqq 0$ if i chooses activity $w \in N_i$ and if the player j for whom $z \in N_j$ does not choose activity z, where j may or may not equal i. This cost structure might induce a player i to choose an unprofitable activity from $N_i \cap S$ in order to reduce the costs associated with choosing profitable activities from $N_i \cap T$. If player j has chosen some $X_j \subseteq N_j$ for all $j \neq i$, then the problem for player i is to minimize $h_i(X) = \sum_{w \in X_i \cap S} c_w - \sum_{w \in X_i \cap T} v_w + \sum_{w \in X_i} \sum_{z \notin X} b(w, z)$ over $X_i \subseteq N_i$ where $X = \bigcup_{j=1}^n X_j$. It is possible to express this game as a minimum cut game as follows. Construct a network with a source s, a sink t, and m other nodes corresponding to each of the activities of N. For each i define a nonnegative capacity function $c_i(w, z)$ on each pair of nodes (w, z) from among $N \cup \{s\} \cup \{t\}$ so that $c_i(s, z) = v_z$ if $z \in N_i \cap T$, $c_i(w, t) = c_w$ if $w \in N_i \cap S$, $c_i(w, z) = b(w, z)$ if $w \in N_i$ and $z \in N$, and $c_i(w, z) = 0$ otherwise. For player i the associated cut capacity function for $X \subseteq N$ is $f_i(X) = \sum_{w \in X \cup \{s\}} \sum_{z \notin X \cup \{s\}} c_i(w, z) = \sum_{z \in (N_i \sim X_i) \cap T} v_z + \sum_{w \in X_i \cap S} c_w + \sum_{w \in X_i} \sum_{z \in N \sim X} b(w, z) = h_i(X) + \sum_{z \in N_i \cap T} v_z$.

A special case of the above game dealing with choosing among economic activities involves an n-person game version of the selection problem. Player i selects a set W_i from a finite collection of items S_i. The cost to player i of choosing item w from S_i is $c_w > 0$, so i incurs the cost $\sum_{w \in W_i} c_w$ for selecting a set W_i. The sets S_i and S_j are disjoint for $i \neq j$, and $S = \bigcup_{i=1}^n S_i$. There is a finite collection of projects T_i potentially available to player i. The sets T_i and T_j are disjoint for $i \neq j$, and $T = \bigcup_{i=1}^n T_i$. Project $z \in T_i$ has a value $v_z \geqq 0$ to player i and its value can only be realized if a subset D_z of the items S are available. For a project z and $W \subseteq S$, define $u_z(W)$ to be 1 if $D_z \subseteq W$ and 0 otherwise. Thus $u_z(W) = 1$ if and only if it is possible to undertake project z with the items of W. The value to player i of possible projects given items W is $\sum_{z \in T_i} v_z u_z(W)$. Given that player j has chosen $W_j \subseteq S_j$ for all $j \neq i$, the problem of player i is to minimize $\sum_{w \in W_i} c_w - \sum_{z \in T_i} v_z u_z(W)$ over $W_i \subseteq S_i$ where $W = \bigcup_{j=1}^n W_j$. Let M be an arbitrarily large positive number. Because it is the object of each player i to minimize that player's net cost, the game is effectively unchanged if the requirement that the items D_z be available in order to undertake project z is replaced by having player i incur the cost M if i undertakes project z without all the items D_z being available. Define $N_i = S_i \cup T_i$ for each i and $N = S \cup T = \bigcup_{i=1}^n N_i$. The set N is the collection of all economic activities, both items and projects. For w and z in N, define $b(z, w) = M$ if $z \in T$ and $w \in D_z$ and $b(z, w) = 0$ otherwise. Player i incurs the cost $b(z, w)$ by choosing activity $z \in N_i$ if player j for whom $w \in N_j$ does not choose activity w, where j may or may not equal i. This transforms the game version of the selection problem into an equivalent game which is a special case of the above game involving choosing economic activities, where the latter game is itself an example of the minimum cut game. Each player's problem in the game version of the selection problem involves solving a minimum cut problem in a bipartite network. Balinski [2] and Rhys [23] considered the selection problem and showed that it can be solved by solving a minimum cut problem in a bipartite network. The game version of the selection problem degenerates to their selection problem when $n = 1$.

Example (c). *Competitive pricing with substitute products.* Consider a system with n competitors $i = 1, \cdots, n$. Each competitor produces a single product. The n products are substitutes for one another. The products might be virtually identical items bearing different labels as with competing brands of gasoline, or they might be different items with similar purposes such as beef and chicken. The problem for each producer i is to determine the price p_i for product i. The vector of prices is $p = (p_1, \cdots, p_n)$. The price p_i must come from a set of possibilities S_i. Furthermore, each producer i requires that the price p_i be within a given range of the prices of the other products, and for some real numbers a_{ij} and b_{ij} for $j \neq i$ this requirement is expressed as $p \in W_i = \{p: a_{ij} \leqq p_i - p_j \leqq b_{ij}$ for all $j \neq i\}$. The set of feasible price vectors is thus $S = (\times_{i=1}^{n} S_i) \cap (\cap_{i=1}^{n} W_i)$. The set S is a sublattice of E^n. The demand for product i depends on the price vector p according to a known function $D_i(p)$. The total revenue for product i given p is $p_i D_i(p)$. There is a unit production cost c_i for product i, so the total production cost for product i is $c_i D_i(p)$. Assume that $S_i \subseteq [c_i, \infty)$, so no feasible price is below the production cost. The net profit for product i is $(p_i - c_i)D_i(p)$. The net cost $f_i(p)$ for product i is minus the net profit, so $f_i(p) = -(p_i - c_i)D_i(p)$. Given the prices p_j for the other competitors $j \neq i$, producer i is interested in choosing p_i to minimize $f_i(p)$ over feasible p_i. This competitive situation is an n-person nonzero-sum game.

Consider the following two hypotheses:

(A) $D_i(p)$ is an isotone function of p_j for each $j \neq i$;

and

(B) $-D_i(p)$ has antitone differences in (p_i, p_j) for all $j \neq i$.

Conditions (A) and (B) have natural economic interpretations. Condition (A) states that increasing the price of product j will increase the demand for product i, and this is a standard definition for substitute products. Condition (B) states that cutting the price p_i for product i will result in a greater increase in the demand for product i if the price p_j for product j happens to be lower; that is, the demand for product i is more sensitive to its price when another product j is more competitive by virtue of its lower price. By (B), an increase in the price of beef would cause a greater reduction in the demand for beef if the price of chicken happens to be lower. (The discussion of a price cut for product i leading to an increase in its demand is for illustration. These conditions do not preclude an anomaly in which demand for a product could increase with its price.) Condition (B) always holds in the case where $D_i(p)$ is a separable function of p.

Conditions (A) and (B) together imply that $f_i(p)$ has antitone differences in (p_i, p_j) for all $j \neq i$. This is therefore a submodular game.

In the case for which $S = E^n$ and each $D_i(p)$ is an affine function, Levitan and Shubik [16] considered the above model and found a closed-form algebraic expression for an equilibrium point.

A similar model can be developed where the demands depend on advertising expenditure instead of price.

Example (d). *The location problem game.* Consider the problem of locating n new facilities in the plane E^2 where there are already p existing facilities. The 2-vector x_i indicates the location chosen for new facility i for $i = 1, \cdots n$, and each x_i must be in a set S_i. The 2-vector w_k indicates the location of existing facility k for $k = 1, \cdots, p$. Let $d(a, b)$ be some symmetric measure of the distance between vectors a and b in E^2. For each pair of new facilities i and j there is a cost $c_{ij}d(x_i, x_j)$ where c_{ij} is nonnegative and symmetric. For each new facility i and old facility k there is a cost $h_{ik}d(x_i, w_k)$ where h_{ik} is nonnegative. The coefficient c_{ij} may represent the flow between new facility i and new facility j times the cost per unit flow per unit distance between i and j. The coefficient h_{ik} may represent the flow between new facility i and old facility k times the cost per unit flow per unit distance between i and k. The problem is to locate the n new facilities in the feasible region so as to minimize the sum of all costs. That problem is to minimize $(1/2) \sum_{i=1}^n \sum_{j=1}^n c_{ij}d(x_i, x_j) + \sum_{i=1}^n \sum_{k=1}^p h_{ik}d(x_i, w_k)$ subject to $x_i \in S_i$ for each i.

Now consider a version of the location problem where n decision-makers independently choose locations for different collections of facilities. For $i = 1, \cdots, n$ player i controls the location of m_i new facilities. The new facilities are indexed $j = 1, \cdots, m$ where $m = \sum_{i=1}^n m_i$. The set Q_i consists of the indices of the m_i facilities located by i. Player i chooses the location x_j of new facility $j \in Q_i$ from a set $S_j \subseteq E^2$. The 2-vector w_k indicates the location of existing facility k for $k = 1, \cdots, p$. The function $d(a, b)$ is a symmetric measure of the distance between a and b in E^2. For each pair of new facilities j and k both located by player i, there is a cost $c_{jk}d(x_j, x_k)$ to player i where c_{jk} is nonnegative and symmetric. For each pair of new facilities j and k with player i locating j and some other player locating k, there is a cost $c_{jk}d(x_j, x_k)$ to player i where c_{jk} is nonnegative. For each new facility j located by player i and each existing facility k, there is a cost $h_{jk}d(x_j, w_k)$ to player i where h_{jk} is nonnegative. The coefficients c_{jk} and h_{jk} may have interpretations as in the 1-player problem described above. Player i wants to choose feasible locations for facilities Q_i to minimize total related costs. The problem of player i is to minimize $f_i(x) = (1/2) \sum_{j \in Q_i} \sum_{k \in Q_i} c_{jk}d(x_j, x_k) + \sum_{j \in Q_i} \sum_{k \in \cup_{e \neq i} Q_e} c_{jk}d(x_j, x_k) + \sum_{j \in Q_i} \sum_{k=1}^p h_{jk}d(x_j, w_k)$ subject to $x_j \in S_j$ for all $j \in Q_i$ where $x = (x_1, \cdots, x_m) \in E^{2m}$. If each S_j is a sublattice of E^2 and if $d(a, b)$ is submodular on E^4, then this is a submodular game.

If $d(a, b) = |a_1 - b_1| + |a_2 - b_2|$, then the distance measure is *rectilinear*. This $d(a, b)$ is submodular on E^4 by Theorem 1.1. The rectilinear distance would seem an appropriate distance measure where travel between facilities must be on perpendicular city streets or along perpendicular aisles in a machine shop. Picard and Ratliff [21] and a number of other authors cited in [21] have considered the rectilinear distance facility location problem corresponding to the 1-player game with each $S_j = E^2$, and they have given algorithms which solve this problem.

The distance measure $d(a, b) = (a_1 - b_1)^2 + (a_2 - b_2)^2$ is submodular on E^4 by Theorem 1.1. White [40] considered and solved the facility location problem corresponding to the 1-player game with this distance measure and each $S_j = E^2$.

By Theorem 1.1, the Euclidean distance measure $d(a, b) = ((a_1 - b_1)^2 + (a_2 - b_2)^2)^{1/2}$ is not submodular on E^4.

Example (e). *A game with optimal product usage.* Consider a situation in which n players $i = 1, \cdots, n$ each choose a subset of p products $k = 1, \cdots, p$ to use. The decision of player i is denoted by a p-vector $x_i = (x_{i1}, \cdots, x_{ip})$ such that each x_{ik} equals either 0 or 1 where $x_{ik} = 1$ indicates that player i chooses to use product k and $x_{ik} = 0$ indicates that player i chooses not to use product k. Player i may only choose certain combinations of products, so x_i is restricted to a subset S_i of $\times_{k=1}^p \{0, 1\}$. If player i chooses to use product k then the usage is a fixed amount $a_{ik} > 0$. If player i uses product k then i incurs a cost which is a function of the total amount of product k used by all the players. That cost is $c_{ik}(\sum_{j=1}^n a_{jk}x_{jk})$. Player i incurs no cost with respect to product k if i chooses not to use product k. Define $b_{ik}(x_{ik}, z) = c_{ik}(z)$ if $x_{ik} = 1$ and $b_{ik}(x_{ik}, z) = 0$ if $x_{ik} = 0$. The cost to player i for product k is thus $b_{ik}(x_{ik}, \sum_{j=1}^n a_{jk}x_{jk})$. In using the products indicated by a decision x_i, player i receives a reward $v_i(x_i)$. The problem for player i is thus to minimize $f_i(x) = \sum_{k=1}^p b_{ik}(x_{ik}, \sum_{j=1}^n a_{jk}x_{jk}) - v_i(x_i)$ subject to $x_i \in S_i$ where $x = (x_1, \cdots, x_n)$.

This game is a submodular game if each S_i is a sublattice of E^p, each $c_{ik}(z)$ is antitone in z on $[a_{ik}, \infty)$, and each $-v_i(x_i)$ is submodular on S_i. The property that $c_{ik}(z)$ be antitone implies that $b_{ik}(x_{ik}, \sum_{j=1}^n a_{jk}x_{jk})$ has antitone differences in (x_{ik}, x_{jk}) on $\{0, 1\} \times \{0, 1\}$ for each $j \neq i$.

The set S_i is a sublattice if $S_i = \times_{k=1}^p \{0, 1\}$. The set S_i is also a sublattice if it includes constraints that player i can choose certain products only if i chooses certain other products. The inequality $x_{ik} - x_{ir} \leqq 0$ defines a sublattice, and that inequality requires that player i can choose product k only if i also chooses product r.

Suppose that $C_k(z)$ is the cost of producing z units of product k for the use of the n players, where $C_k(0) = 0$ and $C_k(z)$ is concave on $[0, \infty)$. Thus there are increasing returns to scale in the production of product k. The cost of producing product k is allocated to the n players in proportion to their use of product k. The cost to player i as a result of a decision to use product k is thus $c_{ik}(z) = (a_{ik}/z)C_k(z)$. The properties that $C_k(0) = 0$ and that $C_k(z)$ is concave on $[0, \infty)$ imply that $C_k(z)/z$ is antitone in z on $(0, \infty)$ and therefore $c_{ik}(z)$ is antitone in z on $[a_{ik}, \infty)$.

The condition that $-v_i(x_i)$ is submodular indicates that the k products are complementary from the point of view of the reward function of player i. That condition holds if $v_i(x_i)$ is separable.

Rosenthal [25] showed the existence of an equilibrium point in a game related to that described above. His game model required that each $a_{ik} = 1$, $c_{ik}(z)$ does not depend on i, and $v_i(x_i) = 0$, while it did not require that $c_{ik}(z)$ be antitone and it permitted S_i to be any subset of $\times_{k=1}^p \{0, 1\}$.

3. Existence of an equilibrium point. Throughout this section and § 4, assume that the game under consideration is a submodular game.

Let $S(x) = (\times_{i=1}^n S_i(x)) \cap S$, $g(x, y) = \sum_{i=1}^n f_i(x; y_i)$ for $x \in S$ and $y \in S(x)$, and $Y(x) = \{y : y \in S(x), g(x, y) = \inf_{z \in S(x)} g(x, z)\}$ for $x \in S$. Ponstein [22] introduced the mapping $Y(x)$ to generalize the earlier work of Nash [18], [19]. Ponstein [22] proved that if $x \in Y(x)$ then x is an equilbrium point, and the converse statement is also clearly true. Thus finding an equilibrium point is equivalent to finding a fixed point for the point to set mapping $Y(x)$ on S.

LEMMA 3.1. *If S is compact and $f_i(x)$ is lower semicontinuous in x_i on $S_i(x)$ for all x in S and each i, then $Y(x)$ is ascending in x on S and for each x in S the set $Y(x)$ has a greatest element $\bar{y}(x)$ and a least element $\underline{y}(x)$ such that $\bar{y}(x)$ and $\underline{y}(x)$ are isotone functions from S into S.*

Proof. Since S is compact it follows that $S_i(x)$ is a compact subset of E^{m_i} for $i = 1, \cdots, n$ and all x, and so $S(x)$ is a compact subset of E^m for all x in S. If x is in S then x is in $S(x)$ and so $S(x)$ is nonempty. By hypothesis, $S(x)$ is ascending in x on S, $g(x, y)$ is lower semicontinuous in y on $S(x)$ and submodular in y on S for each x in S, and $g(x, y)$ has antitone differences in (x, y) on $S \times S$. This result then follows by applying Theorem 1.2 to the problem of minimizing $g(x, y)$ over y in $S(x)$ for x in S. □

Algorithm II of § 4 uses the isotone functions $\bar{y}(x)$ and $\underline{y}(x)$ to construct an equilibrium point, but their use in the existence result of Theorem 3.1 is not constructive.

THEOREM 3.1. *If S is compact and $f_i(x)$ is lower semicontinuous in x_i on $S_i(x)$ for all x in S and each i, then the set of equilibrium points is nonempty and a greatest and a least equilibrium point exist.*

Proof. By Lemma 3.1, $\bar{y}(x)$ exists and is an isotone function from the compact sublattice S into itself. It follows from Tarski's fixed point theorem [30] that $\bar{y}(x)$ has a greatest fixed point which takes the form $\bar{x} = \sup\{x : x \in S, x \leqq \bar{y}(x)\}$. Since $\bar{x} = \bar{y}(\bar{x}) \in Y(\bar{x})$, $\bar{x}$ is an equilibrium point.

Pick any equilibrium point $\hat{x}$. Then $\hat{x} \in Y(\hat{x})$ so $\hat{x} \leqq \bar{y}(\hat{x})$. Thus $\hat{x} \leqq \sup\{x : x \in S, x \leqq \bar{y}(x)\} = \bar{x}$ and $\bar{x}$ is the greatest equilibrium point.

The existence of a least equilibrium point follows dually. □

When $n = 1$ the problem of finding an equilibrium point degenerates to the mathematical programming problem of minimizing $f_1(x)$ subject to x in S. When $n = 1$ and the hypotheses of Theorem 3.1 hold, Theorem 1.2 implies that the set of equilibrium points is a compact sublattice of E^m. The following examples show, however, that when $n > 1$ and the hypotheses of Theorem 3.1 hold, the set of equilibrium points need not be compact and it need not be a sublattice. (Even without the submodular game's special assumptions about S and the cost functions, the set of equilibrium points is compact under the hypotheses stated in Theorem 4.2 by a proof similar to the first paragraph of the proof of Theorem 4.2. In the next example, $S_i(x)$ is not a lower semicontinuous mapping.)

Let $n = 2$, $m_1 = m_2 = 1$, $S = \{x : x_1, x_2 \in [0, 1]\} \cup \{x : x_1 = x_2 \in (1, 2]\}$, $f_1(x) = x_1$, and $f_2(x) = x_2$. The set of equilibrium points is $\{(z, z) : z \in \{0\} \cup (1, 2]\}$ which is not closed.

Let $n = 3$, $m_1 = m_2 = m_3 = 1$, $S = \times_{i=1}^{3} [0, 1]$, and $f_1(x) = f_2(x) = f_3(x) = -x_1x_2x_3$. Then $(1, 0, 0)$ and $(0, 1, 0)$ are equilibrium points but $(1, 1, 0) = (1, 0, 0) \vee (0, 1, 0)$ is not an equilibrium point, so the set of equilibrium points is not a sublattice of E^3.

A joint decision x in S is a *strong equilibrium point* [20] if there does not exist a subset $N \subseteq \{1, \cdots, n\}$ of the n players such that the sum of the costs to these players can be strictly decreased by finding a new feasible joint decision which leaves unchanged the decisions of the players not in N. Even if $n = 2$, $m_1 = m_2 = 1$, and $f_1(x)$ and $f_2(x)$ are submodular and convex, a strong equilibrium point may not exist. If $S = [0, 1] \times [0, 1]$, $f_1(x) = x_1^2 - x_2$, and $f_2(x) = x_2^2 - x_1$, then $(0, 0)$ is the unique equilibrium point but players 1 and 2 can decrease the sum of their costs by choosing $(1/2, 1/2)$.

4. Algorithms for approximating an equilibrium point.

Based on a suggestion by Brown [6], Robinson [24] proved that one can approximate a *mixed* strategy equilibrium point for a static zero-sum two-person finite game as a result of fictitious play in a sequential game corresponding to a natural behavioral process. At a given iteration each player assumes that the other player will choose the mixed strategy determined by the relative frequency of the pure strategies that the other player has already chosen, and each player either in turn or simultaneously chooses the best individual pure strategy. With new pure strategies having been chosen, each player's relative frequency of pure strategies is updated and the next iteration begins. See also Danskin [7]. An example of Shapley [28] shows that this procedure need not generally succeed for nonzero-sum two-person finite games.

Algorithms I and II below approximate a *pure* equilibrium point for a static nonzero-sum n-person submodular game as a result of fictitious play in sequential games corresponding to natural behavior processes. Algorithm I corresponds to the iterative decision-making process by which the n players take turns with each player successively minimizing that player's own cost function with respect to feasible decisions while the decisions of the other $n - 1$ players are held fixed. In the case where each player's set of feasible decisions is independent of the decisions of the other players, Algorithm II corresponds to the iterative decision-making process by which each of the n players concurrently and individually chooses the next decision by minimizing that player's own cost function under the assumption that the other $n-1$ players will hold their decisions unchanged. A new joint decision is put together by combining these n individually determined decisions, and the next iteration then begins.

ALGORITHM I. This algorithm proceeds by starting with $x^{0,0}=x^0$, where x^0 is the least element of S. Given $x^{k,i}$ in S where k and i are nonnegative integers with $i<n$, the next point $x^{k,i+1}=(x^{k,i};\bar{y}^k_{i+1})$ is generated by picking $\bar{y}^k_{i+1}$ to be the least y_{i+1} to minimize $f_{i+1}(x^{k,i};y_{i+1})$ over y_{i+1} in $S_{i+1}(x^{k,i})$. When $x^{k,n}$ has been generated for some k, set $x^{k+1,0}=x^{k,n}$ and continue.

Given the decisions $x\sim x_i$ of the other $n-1$ players, the problem of player i is to minimize $f_i(x)$ over x_i in $S_i(x)$. Theorem 1.2 implies that the optimal decision x_i for player i is an isotone function of the decisions $x\sim x_i$ of the other players. Lemma 4.1 uses that result to show that $x^{k,i}$ is isotone in k and i.

LEMMA 4.1. *If S is compact and $f_i(x)$ is lower semicontinuous in x_i on $S_i(x)$ for all x and each i, then Algorithm* I *generates a sequence $x^{k,i}$ which is isotone in k and i for $i=0,1,\cdots,n$ and $k=0,1\cdots$. Hence there exists $\bar{x}$ in S such that* $\lim_{k\to\infty}x^{k,i}=\bar{x}$ *for $i=0,1,\cdots,n$.*

Proof. By Theorem 1.2, Algorithm I is well-defined and generates an infinite sequence $\{x^{k,i}\}$.

Since $x^{0,0}=x^0$, $x^{0,i}\leqq x^{0,i+1}$ for $i=0,\cdots,n-1$. Now suppose that $x^{k,i}\leqq x^{k,i+1}$ for all $k<K$ and $i=0,\cdots,n-1$ and that $x^{K,i}\leqq x^{K,i+1}$ for $i=0,\cdots,I-1$ where $1\leqq K$ and $0\leqq I\leqq n-1$. Since this supposition holds for $K=1$ and $I=0$, it suffices to show that $x^{K,I}\leqq x^{K,I+1}$ for the proof to follow by induction. Because $\bar{y}^{K-1}_{I+1}$ is the least minimizing point of $f_{I+1}(x^{K-1,I};y_{I+1})$ over y_{I+1} in $S_{I+1}(x^{K-1,I})$, $\bar{y}^K_{I+1}$ is the least minimizing point of $f_{I+1}(x^{K,I};y_{I+1})$ over y_{I+1} in $S_{I+1}(x^{K,I})$, and $x^{K-1,I}\leqq x^{K,I}$, Theorem 1.2 implies that $\bar{y}^{K-1}_{I+1}\leqq\bar{y}^K_{I+1}$ and hence

$$x^{K,I}=(x^{K,I};\bar{y}^{K-1}_{I+1})\leqq(x^{K,I};\bar{y}^K_{I+1})=x^{K,I+1}. \qquad \square$$

Lemma 4.1 reduces the problem of finding $x^{k,i+1}$ given $x^{k,i}$ for $i<n$ from a minimization problem over $S_{i+1}(x^{k,i})$ to a minimization problem over $S_{i+1}(x^{k,i})\cap[x^{k,i}_{i+1},\infty)$.

The following result shows that there is a check inherent in Algorithm I which indicates whether or not an equilibrium point has been attained.

LEMMA 4.2. *If a point appears n successive times in the sequence $\{x^{k,i}: k\geqq 0, 1\leqq i\leqq n\}$ generated by Algorithm* I, *then that point is an equilibrium point. If Algorithm* I *generates an equilibrium point at some iteration, then that point will be generated at all subsequent iterations.*

Proof. The first part follows directly from the definition of an equilibrium point.

Suppose $x^{k,i}$ is an equilibrium point where $i<n$. To establish the second part it suffices to show that $x^{k,i+1}=x^{k,i}$ or, equivalently, that $x^{k,i+1}_{i+1}=x^{k,i}_{i+1}$. Since $x^{k,i}$ is an equilibrium point, $x^{k,i}_{i+1}$ minimizes $f_{i+1}(x^{k,i};y_{i+1})$ over y_{i+1} in $S_{i+1}(x^{k,i})$ and so by construction $x^{k,i+1}_{i+1}\leqq x^{k,i}_{i+1}$. However, $x^{k,i}_{i+1}\leqq x^{k,i+1}_{i+1}$ by Lemma 4.1, and so $x^{k,i+1}_{i+1}=x^{k,i}_{i+1}$. □

THEOREM 4.1. *If S has a finite number of elements and no chain contained in S_i has more than q_i elements for $i=1,\cdots,n$, then Algorithm* I *generates an equilibrium point in no more than $(n-1)(\sum_{i=1}^n q_i)-n^2+n+1$ iterations.*

Proof. By Lemma 4.1, $x_j^{k,i}$ is isotone in k and i for $i = 0, 1, \cdots, n$, $k = 0, 1, \cdots$ and fixed $j = 1, \cdots, n$. Thus as Algorithm I proceeds $x_j^{k,i}$ can change its value no more than $q_j - 1$ times and hence $x^{k,i}$ can change its value no more than $\sum_{j=1}^{n} (q_j - 1)$ times and the sequence $\{x^{k,i}\}$ can contain no more than $\sum_{j=1}^{n} (q_j - 1) + 1$ distinct elements. Since $x^{k,i}$ is isotone in k and i and S is finite, Algorithm I must eventually generate some point at some iteration such that the same point is generated at all subsequent iterations. By Lemma 4.2, that last distinct point generated must be an equilibrium point. Since $x^{k,i}$ is isotone in k and i all appearances of a point in this sequence must be consecutive, and by Lemma 4.2 no point except the last distinct one generated can appear in $\{x^{k,i}: k \geqq 0, 1 \leqq i \leqq n\}$ more than $n - 1$ times. Thus Algorithm I must attain an equilibrium point in no more than $(n-1) \sum_{j=1}^{n} (q_j - 1) + 1$ iterations. □

Theorem 4.1 indicates that Algorithm I is quite efficient computationally when S is finite. If S is finite, $S_i(x) = S_i$ for all x in S and $i = 1, \cdots, n$, and S_i has p_i elements, then adding one element to S_j will add $\prod_{i \neq j} p_i$ elements to S but the bound on the number of iterations required by Algorithm I will increase by either $n - 1$ or 0. If S is finite and $q_i = \bar{q}$ for $i = 1, \cdots, n$ then the bound on the number of iterations required by Algorithm I is $(n^2 - n)(\bar{q} - 1) + 1$ which varies with n^2, while there are at least $\bar{q}^n$ feasible joint decisions if $S_i(x) = S_i$ for all i and all x in S. For example, if $n = 20$, $S_i = \{1, \cdots, 10\}$ for $i = 1, \cdots, 20$, and $S = \times_{i=1}^{n} S_i$, then S contains 10^{20} feasible joint decisions but Algorithm I will generate an equilibrium point in no more than $n(n-1)(\bar{q} - 1) + 1 = (20)(19)(9) + 1 = 3421$ iterations. Since each iteration requires looking at no more than 10 joint decisions and making at most 9 comparisons, one can find an equilibrium point for this problem with 10^{20} feasible joint decisions by looking at no more than 34,210 joint decisions and making no more than $9(3421) = 30{,}789$ comparisons.

A point to set mapping Γ from $A \subseteq E^m$ into nonempty subsets of E^r is a *lower semicontinuous mapping* [3] if $\{x^k : k = 1, 2, \cdots\} \subseteq A$, $\lim_{k \to \infty} x^k = \bar{x} \in A$, and $\bar{y} \in \Gamma(\bar{x})$ imply that there exist $y^k \in \Gamma(x^k)$ for $k = 1, 2, \cdots$ such that $\lim_{k \to \infty} y^k = \bar{y}$. If S contains a finite number of elements or if $S = \times_{i=1}^{n} S_i$ where $S_i \subseteq E^{m_i}$ for $i = 1, \cdots, n$ (so that $S_i(x) = S_i$ for all $x \sim x_i$ in T_i and each i), then $S_i(x)$ is a lower semicontinuous mapping from T_i into subsets of E^{m_i} for each i and $S(x)$ is a lower semicontinuous mapping from S into subsets of E^m.

THEOREM 4.2. *If S is compact, $S_i(x)$ is a lower semicontinuous mapping from T_i into subsets of E^{m_i} for each i, and $f_i(x)$ is continuous on S for each i, then the limit point $\bar{x}$ of $\{x^{k,i}\}$ generated by Algorithm* I *is an equilibrium point. Furthermore, $\bar{x}$ is the least equilibrium point.*

Proof. Pick any i, $1 \leqq i \leqq n$, and any $\bar{y}_i$ in $S_i(\bar{x})$. Since $\lim_{k \to \infty} x^{k,i} = \bar{x}$ and $S_i(x)$ is a lower semicontinuous mapping, there exists y_i^k in $S_i(x^{k,i})$ for $k = 0, 1, \cdots$ such that $\lim_{k \to \infty} y_i^k = \bar{y}_i$. By the construction of Algorithm I, $f_i(x^{k,i}) \leqq f_i(x^{k,i}; y_i^k)$ for all k. Then by the continuity of $f_i(x)$, $f_i(\bar{x}) \leqq f_i(\bar{x}; \bar{y}_i)$. Because $f_i(\bar{x}) \leqq f_i(\bar{x}; \bar{y}_i)$ for each i and all $\bar{y}_i$ in $S_i(\bar{x})$, $\bar{x}$ is an equilibrium point.

Let $\hat{x}$ be an equilibrium point. Since $x^{0,0} = x^0$, $x^{0,0} \leqq \hat{x}$. Suppose $x^{k,i} \leqq \hat{x}$ for some $k \geqq 0$ and $0 \leqq i \leqq n - 1$. Since $x_{i+1}^{k,i+1}$ is the least minimizing point of $f_{i+1}(x^{k,i}; y_{i+1})$ over y_{i+1} in $S_{i+1}(x^{k,i})$, $\hat{x}_{i+1}$ minimizes $f_{i+1}(\hat{x}; y_{i+1})$ over y_{i+1} in $S_{i+1}(\hat{x})$, and $x^{k,i} \leqq \hat{x}$, Theorem 1.2 implies that $x_{i+1}^{k,i+1} \leqq \hat{x}_{i+1}$ and hence

$$x^{k,i+1} = (x^{k,i}; x_{i+1}^{k,i+1}) \leqq (\hat{x}; \hat{x}_{i+1}) = \hat{x}.$$

Thus by induction $x^{k,i} \leqq \hat{x}$ for all k and i and so $\bar{x} = \lim_{k \to \infty} x^{k,i} \leqq \hat{x}$. □

When $f_i(x) = f(x)$ for all x in S and each i then under the assumptions of Theorem 4.2, $f(\bar{x}) \leqq f(\bar{x}; y_i)$ for all y_i in $S_i(\bar{x})$ and each i, where $\bar{x}$ is the limit point of the sequence generated by Algorithm I. If also $S = \times_{i=1}^{n} S_i$ where $S_i \subseteq E^{m_i}$ for each i (so $S_i(x) = X_i$ for all $x \sim x_i$ in T_i and each i), S_i is convex for each i, and $f(x)$ is differentiable on S, then $(y - \bar{x}) \cdot \nabla f(\bar{x}) \geqq 0$ for all y in S. If $f(x)$ is also pseudo-convex on S, then $\bar{x}$ is the least minimizing point of $f(x)$ on S. When $m_i = 1$ for each i, Veinott [38] established that $\{x^{k,i}\}$ converges monotonically to an optimum for the nonlinear programming problem of minimizing $f(x)$ subject to $x \in S$ under all the above assumptions. This nonlinear programming algorithm falls into the class of coordinate descent algorithms [8], [14], [27], [39], [41]. Indeed, even if the assumptions that $x^{0,0} = x^0$, that $f(x)$ be submodular, and that the minimizing point chosen at each iteration be the least minimizing point are deleted, it still follows by applying the techniques of Zadeh [41] (who considered a slightly less general case) that every accumulation point of $\{x^{k,i}\}$ is optimal for this particular nonlinear programming problem.

ALGORITHM II. The second algorithm proceeds by starting with x^0. Given x^k in S, the next point x^{k+1} is the least y to minimize $g(x^k, y)$ over y in $S(x^k)$. Thus $x^{k+1} = \underline{y}(x^k)$.

LEMMA 4.3. *If S is compact and $f_i(x)$ is lower semicontinuous in x_i on $S_i(x)$ for all x in S and each i, then Algorithm* II *generates a sequence x^k which is isotone in k for $k = 0, 1, \cdots$. Hence there exists $\bar{x}$ in S such that $\lim_{k\to\infty} x^k = \bar{x}$.*

Proof. By Theorem 1.2, Algorithm II is well-defined and generates an infinite sequence $\{x^k\}$.

Clearly $x^0 \leqq x^1$. Suppose $x^{k-1} \leqq x^k$ for some $k \geqq 1$. Then by Lemma 3.1 $x^k = \underline{y}(x^{k-1}) \leqq \underline{y}(x^k) = x^{k+1}$, and by induction the proof is complete. □

Lemma 4.3 reduces the problem of finding x^{k+1} given x^k from a minimization problem over $S(x^k)$ to a minimization problem over $S(x^k) \cap [x^k, \infty)$.

The following result shows that there is a check inherent in Algorithm II which indicates whether or not an equilibrium point has been attained.

LEMMA 4.4. *A point x^k generated by Algorithm* II *is an equilibrium point if and only if $x^k = x^{k+1}$.*

Proof. If $x^k = x^{k+1}$ then $x^k = \underline{y}(x^k) \in Y(x^k)$ and so x^k is an equilibrium point.

If x^k is an equilibrium point then $x^k \in Y(x^k)$ and so $x^k \geqq \underline{y}(x^k) = x^{k+1}$. However, $x^k \leqq x^{k+1}$ by Lemma 4.3, so $x^k = x^{k+1}$. □

THEOREM 4.3 *If S has a finite number of elements and no chain contained in S has more than q elements, then Algorithm* II *generates an equilibrium point in no more than $q - 1$ iterations.*

Proof. If any element appears more than once in the sequence $\{x^k\}$ then by Lemma 4.3 it must appear at least two times consecutively in the sequence and by Lemma 4.4 it is an equilibrium point and is the last distinct element in $\{x^k\}$. Thus at most one point can appear more than once in $\{x^k\}$ and such a point is an equilibrium point. Since S is finite at least one point must appear more than once in $\{x^k\}$. Therefore if $\{x^k\}$ contains K distinct elements, then the elements $x^0, \cdots, x^{K-2}$ will appear exactly once, $x^k = x^{K-1}$ for all $k \geqq K - 1$, x^{K-1} is an equilibrium point, and so an equilibrium point is attained in $K - 1$ iterations. By Lemma 4.3 the distinct elements of $\{x^k\}$ form a chain so $K \leqq q$ and an equilibrium point is attained in no more than $q - 1$ iterations. □

If S is finite, $S_i(x) = S_i$ for all $x \sim x_i$ in T_i and each i, and S_i has p_i elements, then adding one element to S_j will add $\prod_{i \neq j} p_i$ elements to S but the bound on the number of iterations required by Algorithm II will increase by either 1 or 0. If S is finite and q_i is defined as in Theorem 4.1 then $q - 1 \leqq \sum_{i=1}^{n} (q_i - 1)$ so a bound, $\sum_{i=1}^{n} (q_i - 1)$, exists for the number of iterations required for Algorithm II to attain an equilibrium point such that this bound is proportional to the number of players when $q_i = \bar{q}$ for all i. If $a_j \leqq b_j$ are integers for $j = 1, \cdots, m$ and $S \subseteq \{z : a_j \leqq z_j \leqq b_j, z_j \text{ integer for } 1 \leqq j \leqq m\}$, then $q \leqq \sum_{j=1}^{m} (b_j - a_j) + 1$ while S may have as many as $\prod_{j=1}^{m} (b_j - a_j + 1)$ elements. Similar bounds with additive rather than multiplicative components can be found for q_i for use in the bound of Theorem 4.1.

THEOREM 4.4. *If S is compact, $S(x)$ is a lower semicontinuous mapping from S into subsets of E^m, and $f_i(x)$ is contintinous on S for each i, then the limit point $\bar{x}$ of $\{x^k\}$ generated by Algorithm* II *is an equilibrium point. Furthermore, $\bar{x}$ is the least equilibrium point.*

Proof. Pick any $\bar{y}$ in $S(\bar{x})$. Since $\lim_{k\to\infty} x^k = \bar{x}$ and $S(x)$ is a lower semicontinuous mapping on S, there exists y^k in $S(x^k)$ for $k = 0, 1, \cdots$ such that $\lim_{k\to\infty} y^k = \bar{y}$. By the construction of Algorithm II, $g(x^k, x^{k+1}) \leqq g(x^k, y^k)$ for all k. By continuity, $g(\bar{x}, \bar{x}) \leqq g(\bar{x}, \bar{y})$. Because $g(\bar{x}, \bar{x}) \leqq g(\bar{x}, \bar{y})$ for all $\bar{y} \in S(\bar{x})$, $\bar{x} \in Y(\bar{x})$ and so $\bar{x}$ is an equilibrium point.

Let $\hat{x}$ be any equilibrium point. Clearly $x^0 \leqq \hat{x}$. Suppose $x^k \leqq \hat{x}$ for some k. By Lemma 3.1, $x^{k+1} = \underline{y}(x^k) \leqq \underline{y}(\hat{x})$. Since $\hat{x}$ is an equilibrium point, $\hat{x} \in Y(\hat{x})$ and so $\underline{y}(\hat{x}) \leqq \hat{x}$. Thus $x^{k+1} \leqq \hat{x}$ and by induction $x^k \leqq \hat{x}$ for all k. Therefore $\bar{x} = \lim_{k\to\infty} x^k \leqq \hat{x}$. □

The conditions required to establish the existence of an equilibrium point in Theorem 3.1 are weaker than those assumed in this section to prove that Algorithms I and II actually approximate such an equilibrium point. The additional assumptions are that each $S_i(x)$ be a lower semicontinuous mapping for Algorithm I, that $S(x)$ be a lower semicontinuous mapping for Algorithm II, and that for each algorithm $f_i(x)$ be continuous in x on S for each i rather than that $f_i(x)$ be lower semicontinuous in x_i for each i. The following two examples show that these stronger assumptions cannot be dispensed with entirely.

Let $n = 2$, $m_1 = m_2 = 1$, $S = \{x : x = (1, 1) - (1/k, 1/k)$ for $k = 1, 2, \cdots\} \cup \{x : x = (1, 1) - (1/(k+1), 1/k)$ for $k = 1, 2, \cdots\} \cup \{(1, 1)\} \cup \{(1, 2)\}$, $f_1(x) = -x_1$, and $f_2(x) = -x_2$. This example satisfies the conditions of Theorem 3.1. The unique equilibrium point is $(1, 2)$. However, if $z^k = (1, 1) - (1/k, 1/k)$ for $k = 1, 2, \cdots$ and $\bar{z} = (1, 1)$ then $\lim_{k\to\infty} z^k = \bar{z}$ and $2 \in S_2(\bar{z})$ but $S_2(z^k) = (1 - 1/k, 1 - 1/(k-1))$ for $k \geqq 2$ so there does not exist $\{y_2^k\}$ with $y_2^k \in S_2(z^k)$ for $k = 1, 2, \cdots$ and $\lim_{k\to\infty} y_2^k = 2$. Thus $S_2(x)$ is not a lower semicontinuous mapping on T_2 and hence $S(x)$ is not a lower semicontinuous mapping on S. For this problem Algorithm I will generate $x^{k,0} = (1, 1) - (1/(k+1), 1/(k+1))$ and $x^{k,1} = (1, 1) - (1/(k+2), 1/(k+1))$ for $k = 0, 1, \cdots$, and Algorithm II will generate $x^k = x^{k/2,0}$ for even k and $x^k = x^{(k-1)/2,1}$ for odd k. However, since $\lim_{k\to\infty} x^{k,0} = \lim_{k\to\infty} x^{k,1} = \lim_{k\to\infty} x^k = (1, 1) \neq (1, 2)$, neither Algorithm I nor Algorithm II approximates an equilibrium point here.

Let $n = 2$, $m_1 = m_2 = 1$, $S = [-1, 0] \times [-1, 1]$, $f_1(x) = (2x_1 - x_2)^2$ for $x \in S$, $f_2(x) = (2x_2 - x_1)^2$ for $x \in S$ and $x \notin \{0\} \times [0, 1]$, and $f_2(x) = -x_2^2$ for each $x \in \{0\} \times [0, 1]$. The set S is a compact sublattice. Both $f_1(x)$ and $f_2(x)$ have antitone differences because their derivatives with respect to x_2 exist and are an antitone function of x_1. The function $f_1(x)$ is continuous, while $f_2(x)$ is lower semicontinuous but not continuous. The unique equilibrium point is $(0, 1)$. However, starting with $x^0 = (-1, -1)$ both Algorithms I and II generate sequences of points which converge upwards to $(0, 0)$.

If Algorithms I and II begin with the greatest instead of the least element of S and "least" is replaced by "greatest" in the statements of the algorithms, then by dual arguments all results in this section will hold with "isotone" replaced by "antitone" and "least" replaced by "greatest" and so each algorithm will generate a sequence of points converging downwards to the greatest equilibrium point.

Both Algorithm I and Algorithm II can be modified so that at each iteration the solution chosen for the minimization problem can be any optimal solution (not necessarily the least one) that assures that the sequence generated is isotone.

REFERENCES

[1] K. BAKER AND A. PIXLEY, *Polynomial interpolation and the Chinese remainder theorem*, Math. Z., 143 (1975), pp. 165–174.

[2] M. BALINSKI, *On a selection problem*, Management Sci., 17 (1970), pp. 230–231.

[3] C. BERGE, *Topological Spaces*, E. M. Patterson, trans., Macmillan Company, New York, 1963.

[4] G. BERGMAN, *On the existence of subalgebras with prescribed d-fold projections*, Algebra Universalis, 7 (1977), pp. 341–356.

[5] G. BIRKHOFF, *Lattice theory*, American Mathematical Society Colloquium Publications, Vol. XXV, third edition, Providence, RI, 1967.

[6] G. BROWN, *Iterative solution of games by fictitious play*, Activity Analysis of Production and Allocation, T. Koopmans, ed., Wiley, New York, 1951, pp. 374–376.

[7] J. DANSKIN, *Fictitious play for continuous games*, Naval Res. Logist. Quart., 1 (1954), pp. 313–320.

[8] D. D'ESOPO, *A convex programming procedure*, Ibid., 6 (1959), pp. 33–42.

[9] E. DINIC, *Algorithm for solution of a problem of maximum flow in a network with power estimation*, Soviet Math. Dokl., 11 (1970), pp. 1277–1280.

[10] J. EDMONDS AND R. KARP, *Theoretical improvements in algorithmic efficiency for network flow problems*, J. Assoc. Comput. Mach., 19 (1972), pp. 248–264.

[11] L. FORD AND D. FULKERSON, *Maximal flow through a network*, Canad. J. Math., 8 (1956), pp. 399–404.

[12] ———, *Flows in Networks*, Princeton University Press, Princeton, NJ, 1962.

[13] O. FRINK, *Topology in lattices*, Trans. Amer. Math. Soc., 51 (1942), pp. 569–582.

[14] V. IVANOV, *A general approximation method for solving linear problems*, Soviet Math. Dokl., 3 (1962), pp. 415–418.

[15] A. KARZANOV, *Determining the maximal flow in a network by the method of preflows*, Ibid., 15 (1974), pp. 434-437.

[16] R. LEVITAN AND M. SHUBIK, *Noncooperative equilibrium and strategy spaces in an oligopolistic market*, Differential Games and Related Topics, H. Kuhn and G. Szegö, eds., North-Holland, Amsterdam, 1971, pp. 429–447.

[17] M. MASCHLER, B. PELEG AND L. SHAPLEY, *The kernel and bargaining set for convex games*, Internat. J. Game Theory, 1 (1972), pp. 73–93.

[18] J. NASH, *Equilibrium points in N-person games*, Proc. Nat. Acad. Sci. U.S.A., 36 (1950), pp. 48–49.

[19] ———, *Non-cooperative games*, Ann. of Math., 54 (1951), pp. 286–295.

[20] G. OWEN, *Game Theory*, W. B. Saunders, Philadelphia, 1968.

[21] J.-C. PICARD AND H. RATLIFF, *A cut approach to the rectilinear distance facility location problem*, Management Sci., 26 (1978), pp. 422–433.

[22] J. PONSTEIN, *Existence of equilibrium points in nonproduct spaces*, SIAM J. Appl. Math., 14 (1966), pp. 181–190.

[23] J. RHYS, *A selection problem of shared fixed costs and network flows*, Management Sci., 17 (1970), pp. 200–207.

[24] J. ROBINSON, *An iterative method of solving a game*, Ann. of Math., 54 (1951), pp. 296–301.

[25] R. ROSENTHAL, *A class of games possessing pure-strategy Nash equilibria*, Internat. J. Game Theory, 2 (1973), pp. 65–67.

[26] P. SAMUELSON, *Foundations of Economic Analysis*, Atheneum, New York, 1947.

[27] S. SCHECTER, *Iterative methods for nonlinear programming*, Trans. Amer. Math. Soc., 104 (1962), pp. 179–189.

[28] L. SHAPLEY, *Some topics in two-person games*, Advances in Game Theory, Annals of Mathematics Studies, No. 52, M. Dresher, L. Shapley and A. Tucker, eds., Princeton University Press, Princeton, NJ, 1964, pp. 1–28.

[29] ———, *Cores of convex games*, Internat. J. Game Theory, 1 (1971), pp. 11–26.

[30] A. TARSKI, *A lattice-theoretical fixpoint theorem and its applications*, Pacific J. Math., 5 (1955), pp. 285-309.

[31] D. TOPKIS, *Ordered optimal solutions*, doctoral dissertation, Stanford Univ., Stanford, CA, 1968.

[32] ———, *Equilibrium points in nonzero-sum n-person subadditive games*, Univ. of California Operations Res. Center Tech. Rep. ORC 70-38, Berkeley, CA, 1970.

[33] ———, *Monotone minimum node-cuts in capacitated networks*, Univ. of California Operations Res. Center Tech. Rep. ORC 70-39, Berkeley, CA, 1970.

[34] ———, *Applications of minimizing a subadditive function on a lattice*, Tech. Rep., 1976.

[35] ———, *The structure of sublattices of the product of n lattices*, Pacific J. Math. 65 (1976), pp. 525-532.

[36] ———, *Topology and subcomplete sublattices*, Tech. Rep., 1977.

[37] ———, *Minimizing a submodular function on a lattice*, Operations Res., 26 (1978), pp. 305–321.

[38] A. VEINOTT, JR., *Notes on nonlinear programming*, unpublished, Stanford University, Stanford, CA, 1964.

[39] J. WARGA, *Minimizing certain convex functions*,SIAM J. Appl. Math., 11 (1963), pp. 588–593.

[40] J. WHITE, *A quadratic facility location problem*, AIIE Trans., 3 (1971), pp. 156–157.

[41] N. ZADEH, *A note on the cyclic coordinate ascent method*, Management Sci., 16 (1970), pp. 642–644.

ERRATUM: ALGEBRAIC THEORY OF LINEAR TIME-VARYING SYSTEMS*

EDWARD W. KAMEN† AND KHALED M. HAFEZ‡

On page 509, the correct copy for line 6 was inadvertently replaced by a duplication of the copy appearing on line 12. Line 6 of this page should read:

Case 1. If T is the empty set, we can take $u_i = 0$ and $u_j = \bar{u}_j$ for $j = 0, 1, \cdots, i-1$, .

*This journal, 17 (1979), pp. 500-510.

†School of Electrical Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332.

‡Department of Mathematical Sciences, Memphis State University, Memphis, Tennessee 38153.